В.В.Вьюгин

МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ МАШИННОГО ОБУЧЕНИЯ И ПРОГНОЗИРОВАНИЯ

МОСКВА
2013

УДК 005.519.8(075.8)
ББК 65.290-2в6я73

Вьюгин В.В. “Математические основы теории машинного обучения и прогнозирования” М.: 2013. - 390 с.

Предназначено для первоначального знакомства с математическими основами современной теории машинного обучения (Machine Learning) и теории игр на предсказания. В первой части излагаются основы статистической теории машинного обучения, рассматриваются задачи классификации и регрессии с опорными векторами, теория обобщения и алгоритмы построения разделяющих гиперплоскостей. Во второй и третьей частях рассматриваются задачи адаптивного прогнозирования в нестохастических теоретико-игровой и сравнительной постановках: игры с предсказаниями и предсказания с использованием экспертных стратегий (Prediction with Expert Advice).

Для студентов и аспирантов, специализирующихся в области машинного обучения и искусственного интеллекта.
© МЦМНО, 2013

Лаборатория структурных методов анализа данных в предсказательном моделировании МФТИ

Оглавление

Введение **9**

I Статистическая теория машинного обучения **19**

1 Элементы теории классификации **20**

1.1. Задача классификации 21

1.1.1. Байесовский классификатор 21

1.1.2. Постановка задачи классификации 23

1.1.3. Линейные классификаторы: персептрон 28

1.2. Теория обобщения 34

1.2.1. Верхние оценки вероятности ошибки классификации 34

1.2.2. VC-размерность 46

1.3. Теория обобщения для задач классификации с помощью пороговых решающих правил 57

1.3.1. Пороговая размерность и ее приложения . . . 57

1.3.2. Покрытия и упаковки 64

1.4. Средние по Радемахеру 72

1.5. Средние по Радемахеру и другие меры емкости класса функций 81

1.6. Задачи и упражнения 86

2 Метод опорных векторов 90

2.1. Оптимальная гиперплоскость 90

2.2. Алгоритм построения оптимальной гиперплоскости 95

2.3. Оценка вероятности ошибки обобщения через число опорных векторов 98

2.4. SVM-метод в пространстве признаков 99

2.5. Ядра 104

2.6. Случай неразделимой выборки 117

2.6.1. Вектор переменных мягкого отступа 117

2.6.2. Оптимизационные задачи для классификации с ошибками 119

2.7. Среднее по Радемахеру и оценка ошибки классификации 128

2.8. Задача многомерной регрессии 133

2.8.1. Простая линейная регрессия 133

2.8.2. Гребневая регрессия 136

2.9. Регрессия с опорными векторами 139

2.9.1. Решение задачи регрессии с помощью SVM . 139

2.9.2. Гребневая регрессия в двойственной форме . 145

2.10. Нелинейная оптимизация 149

2.11. Конформные предсказания 154

2.12. Задачи и упражнения 157

2.13. Лабораторные работы 160

II Прогнозирование индивидуальных последовательностей 164

3 Универсальные предсказания 165

3.1. Последовательные вероятностные предсказания . . . 165

3.1.1. Смешивающий предсказатель 166
3.1.2. Правило Лапласа 167
3.1.3. Оптимальный минимаксный предсказатель . . 171
3.1.4. Приложения 173
3.2. Калибруемость прогнозов 177
3.3. Алгоритм вычисления калибруемых прогнозов 182
3.4. Прогнозирование с произвольным ядром 187
3.5. Универсальная алгоритмическая торговая стратегия 194
3.5.1. Калибруемость с дополнительной информацией 200
3.5.2. Доказательство теоремы 3.6 212
3.6. Задачи и упражнения 216
3.7. Лабораторные работы 217

4 Предсказания с использованием экспертных стратегий 219

4.1. Алгоритм взвешенного большинства 220
4.2. Алгоритм оптимального распределения потерь в режиме онлайн 224
4.3. Усиление простых классификаторов – бустинг 230
4.4. Алгоритм следования за возмущенным лидером . . . 239
4.5. Алгоритмы экспоненциального смешивания: общий подход 251
4.5.1. Экспоненциально смешанные потери и их свойства 251
4.5.2. Анализ алгоритма Hedge с переменным параметром обучения 256
4.6. Алгоритмы экспоненциального взвешивания экспертных прогнозов 261
4.6.1. Детерминированные прогнозы 261
4.6.2. Рандомизированные прогнозы 264

4.7. Алгоритм отслеживания наилучшей комбинации экспертов Fixed-Share 270
4.7.1. Составные эксперты и алгоритмы для них . . 270
4.7.2. Теорема об эквивалентности двух алгоритмов 277
4.7.3. Регрет алгоритма Fixed-Share 279
4.8. Схемы смешивания апостериорных распределений экспертов 281
4.9. Предсказания с экспертами в условиях частичного мониторинга 289
4.10. Задачи и упражнения 296
4.11. Лабораторные работы 298

5 Агрегирующий алгоритм Вовка 300
5.1. Смешиваемые функции потерь 300
5.2. Конечное множество экспертов 307
5.3. Бесконечное множество экспертов 313
5.4. Произвольная функция потерь 316
5.5. Логарифмическая функция потерь 317
5.6. Простая игра на предсказания 321
5.7. Игра с квадратичной функцией потерь 323
5.8. Метод защитных предсказаний и агрегирующий алгоритм 327
5.9. Универсальный портфель 333
5.10. Многомерная онлайн регрессия 341
5.10.1. Многомерная регрессия с помощью агрегирующего алгоритма 341
5.10.2. Переход к ядерной многомерной регрессии . . 349
5.10.3. Ядерная форма гребневой регрессии 352
5.11. Задачи и упражнения 352

5.12. Лабораторные работы 354

6 Выпуклая онлайн оптимизация 356

6.1. Алгоритмы следования за лидером FTL и FTRL . . 356

6.2. Онлайн градиентный спуск 365

6.3. Онлайн зеркальный спуск и связанные с ним вопросы 369

6.4. Задачи и упражнения 378

III Игры и предсказания 380

7 Элементы теории игр 381

7.1. Антагонистические игры двух игроков 381

7.2. Достаточное условие существования седловой точки 384

7.3. Смешанные расширения матричных игр 387

7.3.1. Минимаксная теорема 387

7.3.2. Чистые стратегии 389

7.3.3. Решение матричной игры типа $(2 \times M)$ 392

7.3.4. Решение игры типа $(N \times M)$ 395

7.3.5. Конечная игра между K игроками 398

7.4. Задачи и упражнения 404

8 Теоретико-игровая интерпретация теории вероятностей 405

8.1. Теоретико-игровой закон больших чисел 405

8.2. Теоретико-игровая вероятность 411

8.3. Игры на универсальные предсказания 417

8.4. Рандомизированные калибруемые предсказания . . . 423

8.5. Задачи и упражнения 429

9 Повторяющиеся игры **432**

9.1. Бесконечно повторяющиеся игры двух игроков с нулевой суммой 433

9.2. Теорема Блекуэлла о достижимости 438

9.3. Калибруемые предсказания 445

9.4. Калибруемые предсказания и коррелированное равновесие 450

9.5. Задачи и упражнения 457

9.6. Генеративно-состязательные сети (GANs) 457

IV Вспомогательные утверждения **465**

10 Некоторые замечательные неравенства **466**

Литература **474**

Введение

Основная задача науки и реальной жизни – получение правильных предсказаний о будущем поведении сложных систем на основании их прошлого поведения.

Многие задачи, возникающие в практических приложениях, не могут быть решены заранее известными методами или алгоритмами. Это происходит по той причине, что нам заранее не известны механизмы порождения исходных данных или же известная нам информация недостаточна для построения модели источника, генерирующего поступающие к нам данные. Как говорят, мы получаем данные из "черного ящика". В этих условиях ничего не остается, как только изучать доступную нам последовательность исходных данных и пытаться строить предсказания, совершенствуя нашу схему в процессе предсказания. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, называется методом *машинного обучения* (Machine Learning).

Машинное обучение – чрезвычайно широкая и динамически развивающаяся область исследований, использующая огромное число теоретических и практических методов. Данная книга ни в какой мере не претендует на какое-либо исчерпывающее изложение содержания данной области. Наша цель – познакомить читателя с некоторыми современными математическими проблемами данной области и их решениями, основной из которых является проблема построения и оценки *предсказаний* будущих исходов.

С данным подходом тесно связана задача *универсального предсказания.* В том случае, когда мы не имеем достаточной информации для того чтобы построить модель источника, генерирующего наблюдаемые данные, нам приходится учитывать как можно более широкие классы таких моделей и строить методы, которые предсказывают “не хуже”, чем любая модель из данного класса. Понятие универсального предсказания, которое первоначально возникло в теории предсказаний стационарных источников, в настоящее время вышло далеко за рамки этой теории.

Первая часть книги – “Статистическая теории машинного обучения” – использует методы теории вероятностей и математической статистики. В основе данного подхода лежит предположение о том, что наблюдаемые исходы генерируются вероятностным источником, возможно, с неизвестными параметрами.

В рамках статистической теории машинного обучения мы рассматриваем задачи классификации и регрессии. Процесс обучения заключается в выборе функции классификации или регрессии из заранее заданного широкого класса таких функций.

Отметим два способа машинного обучения. При первом способе часть совокупности данных – *обучающая выборка* – выделяется только для обучения. После того как метод предсказания определяется по обучающей выборке, более он не изменяется и в дальнейшем используется для решения задачи предсказания.

При втором способе обучение никогда не прекращается, как говорится, оно происходит в *режиме онлайн*, т.е. предсказания и обучение происходят постоянно в процессе поступления новых данных.

Методы машинного обучения первого типа будут рассмотрены в первой части, которая посвящена статистической теории машинного обучения, методы второго типа будут изучаться во второй и третьей частях книги.

После того как схема предсказания определена, нам необходимо оценить ее предсказательные возможности, т.е. качество ее предсказаний. Предварительно напомним, как оцениваются модели предсказания в классической статистической теории. В статистической теории предсказания мы предполагаем, что последова-

тельность исходных данных (или *исходов*) является реализацией некоторого стационарного стохастического процесса. Параметры этого процесса оцениваются на основании прошлых наблюдений, а на основании уточненного стохастического процесса строится правило предсказания. В этом случае *функция риска* данного правила предсказания определяется как среднее значение некоторой функции потерь, измеряющей различие между предсказаниями и исходами. Среднее значение вычисляется по "истинному вероятностному распределению", которое лежит в основе модели генерации данных. Различные правила предсказания сравниваются по значениям своих функций риска.

В статистической теории машинного обучения также используется стохастическая модель генерации данных, а именно, используется предположение о том, что поступающие данные независимо и одинаково распределены. Вероятность ошибочной классификации или регрессии называется *ошибкой обобщения*. Первый шаг в сторону от классической постановки заключается в том, что распределение, генерирующее данные, нам может быть неизвестно и мы не можем и не будем оценивать его параметры, так как они не используются в оценках ошибок классификации или регрессии. Мы заранее не знаем, какой из методов классификации или регрессии будет построен по наблюдаемой части данных в процессе обучения; нам задан целый класс таких методов – например, это может быть класс разделяющих гиперповерхностей в многомерном пространстве. Оценки ошибки обобщения при классификации или регрессии должны быть равномерными по всем таким вероятностным распределениям и применяемым методам. Иными словами, эти оценки не зависят от распределения, генерирующего данные, а также от функции классификации или регрессии. Впервые данный подход был реализован в работах Вапника и Червоненкиса (см. [2]).

Для оценки предсказательной способности схемы классификации или регрессии используется *теория обобщения*. В рамках этой теории даются оценки вероятности ошибки классификации будущих данных при условии, что обучение проведено на случайной обучающей выборке достаточно большого размера и в его резуль-

тате функция классификации (регрессии) согласована с обучающей выборкой. Важнейшим параметром такой оценки является *сложность (емкость)* класса функций классификации (регрессии). Обычно в оценке вероятности ошибки конкурируют длина выборки и сложность класса гипотез – при заданной величине ошибки чем больше длина обучающей выборки, тем больший по сложности класс гипотез можно использовать.

Методы вычисления ошибок обобщения и теория размерности классов функций излагаются в главе 1.

Сложность классов функций будет измеряться тремя способами. Первый из них – функция роста и связанная с ней размерность Вапника–Червоненкиса (VC-размерность) известны с середины 60-х годов 20-го века. Позже были введены числа покрытия и упаковки и связанная с ними пороговая размерность (fat-размерность), которые дают более точные верхние оценки ошибки обобщения в том случае, когда разделение данных производится с заданным порогом. Еще один способ измерения сложности класса функций – средние Радемахера – также изучается в этой главе. Последние два способа измерения емкости класса функций в отличие от VC-размерности не зависят от размерности пространства объектов.

Глава 2 посвящена построению алгоритмов классификации и регрессии. В основном это алгоритмы, использующие метод опорных векторов. Рассматриваются методы распознавания образов на основе построения разделяющих гиперплоскостей или гиперповерхностей в пространствах признаков, построенных с помощью ядерных методов. Излагаются основы теории функциональных гильбертовых пространств, порожденных воспроизводящим ядром (Reproducing Kernel Hilbert Space – RKHS), и их применение для построения разделяющих гиперповерхностей и для получения оценок ошибки классификации.

Во второй части – Предсказания индивидуальных последовательностей Ч вообще не используются никакие гипотезы о стохастических механизмах, генерирующих данные. Наблюдаемые исходы могут генерироваться совершенно неизвестным нам механизмом, который может быть как детерминированным, так и сто-

хастическим или даже “адаптивно враждебным” к нашим предсказаниям (т.е. может использовать наши прошлые предсказания при генерации очередного исхода).

При этом возникает естественный вопрос – как в этом случае оценивать предсказательную способность метода. В отсутствие вероятностной модели функция риска в виде математического ожидания не может быть определена.

В теории последовательного предсказания (глава 3) используются два подхода для оценки качества предсказаний. В разделе 3.1 рассматриваются максимально широкие классы моделей, описывающих возможные варианты поведения источника, генерирующего данные. Строится метод, собирающий все возможности этих моделей в одну агрегирующую модель. Производится сравнение предсказательной способности этой модели с предсказательными способностями всех моделей класса. В разделе 3.1 в качестве примера агрегирующих моделей рассматривается смешивающий предсказатель и правило Лапласа, а также оптимальный минимаксный предсказатель и его приложения – универсальное кодирование, игра Келли.

В разделе 3.2 приведен метод построения хорошо калибруемых предсказаний. Для оценки качества предсказаний используются тесты, оценивающие рассогласованность между предсказаниями и соответствующими исходами. Эти тесты выбираются исходя из тех задач, для решения которых будут использоваться предсказания.

Один из видов таких тестов – тесты на калибруемость. Цель алгоритма – выдавать такие предсказания, которые выдерживают все тесты на калибруемость.

Тесты на калибруемость строятся в зависимости от того, как мы планируем использовать хорошо калибруемые предсказания. В разделе 3.5 специальные тесты и соответствующие им хорошо калибруемые предсказания будут использованы при построении универсальной алгоритмической стратегии для торговли на финансовом рынке. Алгоритм такой стратегии автоматически покупает и продает акции. Мы докажем, что доход при такой торговле будет не меньше, чем доход любой стационарной алгоритмической

стратегии.

Основные принципы сравнительной (или соревновательной) теории предсказания рассматриваются в главе 4. Эффективность алгоритма предсказания оценивается в форме сравнения с предсказаниями некоторого набора экспертных методов, или просто экспертов. В теории предсказаний с учетом экспертов вводится класс предсказателей, называемых экспертами. Класс экспертов может быть конечным или бесконечным, может иметь мощность континуума. В качестве экспертов могут рассматриваться различные методы предсказания, стохастические теории, методы регрессии и т.д. Эксперты предоставляют свои прогнозы, прежде чем будет представлен соответствующий исход. Агрегирующий алгоритм, при вычислении предсказания, может использовать эти прогнозы, а также кумулятивные потери экспертов. Когда очередной исход становится известным, каждый эксперт вычисляет свои потери. С ростом числа исходов и предсказаний потери каждого эксперта суммируются и образуют его кумулятивные потери. Наилучшим называется эксперт несущий минимальные потери. Агрегирующий алгоритм при вычислении своих прогнозов может использовать прогнозы и потери экспертов в прошлом.

В разделах 4.1 и 4.2 будут построены адаптивные алгоритмы предсказания, которые несут потери не большие, чем потери наилучшего эксперта с точностью до некоторой ошибки обучения, называемой “регретом”. В разделе 4.2 приводится алгоритм распределения потерь в режиме онлайн Фройнда и Шапире, который применяется в разделе 4.3 для усиления слабых алгоритмов классификации. Слабый алгоритм классификации делает лишь незначительно меньшее число ошибок, чем простое случайное угадывание. В разделе 4.3 излагается алгоритм усиления слабых классификаторов – бустинг (Boosting). Приводится алгоритм AdaBoost, решающий эту задачу. Алгоритм AdaBoost усиливает слабый алгоритм классификации до алгоритма, который с некоторого момента в процессе обучения начинает делать как угодно малое число ошибок. Приведены достаточные условия эффективности этого алгоритма.

В разделе 4.4 приводится вероятностный алгоритм Ханнана

следования за возмущенным лидером, в котором принимается решения следовать за экспертом несущим наименьшие потери в прошлом. Прямое следование этой идее может привести к неприемлемым потерям предсказателя, однако введение рандомизации позволяет построить алгоритм с достаточно малым регретом.

В разделе 4.5 мы рассмотрим общие свойства алгоритмов экспоненциального смешивания. При анализе этих алгоритмов основную роль играет так называемые экспоненциально смешанные потери. Кумулятивный вариант этой величины асимптотически близок к потерям каждого из вариантов алгоритмов экспоненциального смешивания. Поэтому анализ свойств этих алгоритмов, в основном, сводится к анализу экспоненциально смешанных потерь.

В разделе 4.6 рассматривается более детальная постановка, при которой потери экспертов и агрегирующего алгоритма возникают в результате принятых ими решений. В процессе прогнозирования выдаются исходы, а функция потерь сопоставляет решения (прогнозы) с этими исходами и выдает соответствующие потери. В разделе 4.7 рассматривается алгоритм отслеживания наилучшей комбинации экспертов Fixed-Share. В предыдущих разделах мы сравнивали потери агрегирующего алгоритма с потерями наилучшего эксперта. Другой возможный подход состоит в следующем: серия шагов, на которых производится обучение, делится на сегменты. Каждому сегменту ставится в соответствие свой эксперт, последовательность сегментов и соответствующих экспертов называется составным экспертом. Цель агрегирующего алгоритма также несколько меняется – теперь он должен предсказывать так, чтобы его потери были бы не больше чем потери составного эксперта (с точностью до некоторого регрета). Дальнейшие обобщения схем смешивания рассматриваются в разделе 4.8.

В разделе 4.9 рассматривается постановка задачи предсказания с экспертами в условиях частичного мониторинга окружающей среды. В этом случае, предсказатель в результате своего действия получает только собственные потери и не имеет доступа к потерям других экспертов. В мировой литературе эта постановка называется “проблема многоруких бандитов”(multi-armed bandit

problem).

В главе 5 мы вернемся к задаче предсказания с использованием экспертных стратегий. Будет рассмотрен агрегирующий алгоритм Вовка, который имеет значительно меньшую ошибку предсказания для логарифмической, квадратичной и некоторых других функций потерь, чем метод экспоненциального смешивания, использованный в главе 4. Будет также построен соответствующий алгоритм многомерной регрессии в режиме онлайн, основанный на применении агрегирующего алгоритма.

В главе 6 задача предсказания с использованием экспертных стратегий рассматривается в терминах теории онлайн выпуклой оптимизации. В разделе 6.2 приведен и изучен алгоритма онлайн градиентного спуска, а в разделе 6.3 рассматривается онлайн оптимизация общего вида основанная на методе зеркального спуска, частными случаями которого являются онлайн градиентный спуск и алгоритмы оптимального смешивания.

Предсказания в режиме онлайн тесно связаны с теорией игр. Третья часть – *Игры и предсказания* – посвящена изложению теории предсказаний на языке теории игр.

Основные понятия теории игр рассматриваются в главе 7. Мы рассмотрим матричную игру двух лиц с нулевой суммой и докажем для нее минимаксную теорему Дж. фон Неймана. Доказательство минимаксной теоремы проведено в стиле теории машинного обучения с использованием метода экспоненциального смешивания. В этой главе также вводятся понятия равновесия Нэша и коррелированного равновесия Аумана.

В главе 8 рассматривается новый теоретико-игровой подход к теории вероятностей, предложенный Вовком и Шейфером [49]. В рамках этого подхода формулируются игры с предсказаниями, на траекториях которых, при определенных условиях, выполнены различные законы теории вероятностей. Примеры таких законов – закон больших чисел, закон повторного логарифма, центральная предельная теорема и т.д. Вводится понятие теоретико-игровой вероятности, которое определяется для подобных игр.

В рамках этого подхода также наиболее естественным образом формулируется задача построения универсальных предсказаний,

рассмотренная в главе 3.

В главе 9 будут рассматриваться более сложные вопросы теории игр. В основе излагаемой теории находится знаменитая теорема Блекуэлла о достижимости (Blackwell approachability theorem). Эта теорема является обобщением минимаксной теоремы для игр двух лиц с произвольными векторнозначными функциями выигрыша. Теорема Блекуэлла служит основой для построения калибруемых предсказаний для случая произвольного конечного числа исходов.

В свою очередь, в этой же главе будет показано, что использование калибруемых предсказаний позволяет построить стратегии, при которых совместное частотное распределение ходов всех игроков сходится к коррелированному равновесию Аумана.

Данная книга представляет собой краткий обзор идей и математических методов современной теории машинного обучения и тесно связанных с ней теории предсказания с использованием экспертных стратегий, нестохастических теоретико-игровых методов предсказания, теоретико-игровых основ теории вероятностей и теории универсальных предсказаний. По мнению автора, все эти темы представляют собой необходимый минимум теоретических знаний для студентов и аспирантов, специализирующихся в области машинного обучения и искусственного интеллекта.

Работа над этой книгой частично проводилась в лаборатории структурных методов анализа данных в предсказательном моделировании при Московском физико-технического институте (МФТИ, государственный угиверситет).

Материал данной книги использовался в качестве основы курсов лекций, прочитанных автором в 2008–2013 годах на базовых кафедрах Института проблем передачи информации (ИППИ РАН) на факультете управления и прикладной математики МФТИ и на факультете компьютерных наук НИУ Высшая школа экономики. Данная книга является существенным расширением учебного пособия [3].

Главы 1 и 2 могут послужить основой для курса лекций “Статистическая теория машинного обучения”. Главы 3, 4 и 5 могут послужить основой для курса лекций “Универсальные предсказа-

ния” и, наконец, на основе глав 3, 7, 8 и 9 можно составить курс “Игры и предсказания”.

С рядом идей и постановок задач, представленных в данной книге, автор познакомился во время краткосрочных визитов в департамент компьютерных наук Ройал Холловей колледжа Лондонского университета. Автор с благодарностью вспоминает многолетнее общение с сотрудниками и аспирантами этого департамента: Александром Гаммерманом, Владимиром Вовком, Юрием Калнишканом, Михаилом Вьюгиным, Ильей Нуретдиновым, Алексеем Черновым, Федором Ждановым, Дмитрием Адамским и Лео Гордоном.

Автор благодарен Юрию Калнишкану и Алексею Чернову за предоставленные материалы.

Автор особенно благодарен Владимиру Вовку за многочисленные беседы и разъяснения по многим вопросам, затронутым в данной работе.

Второе издание содержит следующие дополнительные разделы. В главе 3 добавлен раздел 3.1. В главе 4 добавлены разделы 4.5– 4.9. В главу 5 добавлен раздел 5.8. Новой также является глава 6.

Автор благодарен А.В.Гасникову, Г.А.Кабатянскому, С.М.Карпенко, А.П.Кулешову, А.В.Назину, а также студентам МФТИ и ВШЭ, указавшим на опечатки и неточности первого издания этой книги.

Работа над дополнениями ко второму изданию книги частично поддержана грантами РФФИ: проекты 16-01-00576 А и 16-29-09649 офи-м.

Часть I

Статистическая теория машинного обучения

Глава 1

Элементы теории классификации

Как было замечено во введении, теория машинного обучения решает задачи предсказания будущего поведения сложных систем в том случае, когда отсутствуют точные гипотезы о механизмах, управляющих поведением таких систем.

Имеется ряд категорий машинного обучения: контролируемое обучение или "обучение с учителем" (supervised learning), неконтролируемое обучение (unsupervised learning) (в частности, кластеризация), обучение с подкреплением (reinforcement learning). В этой и следующей главах нас будет интересовать первый тип машинного обучения – контролируемое обучение. Мы начинаем с обучающей выборки, которая представляет собой примеры – пары вида "вход – выход". Целью обучения является – восстановление зависимости между элементами этих пар с целью предсказания будущего выхода по заданному входу.

В основе статистической теории машинного обучения лежит гипотеза о существовании стохастического механизма генерирующего такие пары. В этом случае мы можем оценивать вероятность ошибки классификации будущих примеров. При этом делаются минимальные предположения о виде вероятностного источника, генерирующего данные. Теория обобщения предоставляет оценки таких ошибок, равномерные относительно максимально широких

классов вероятностных распределений генерирующих данные.

Мы рассмотрим два основных класса задач: *задачи классификации* и *задачи регрессии*. В данной главе рассматривается задача классификации, в которой выход это метка класса, к которому принадлежит вход.

1.1. Задача классификации

1.1.1. Байесовский классификатор

Предварительно рассмотрим один простейший метод классификации. Рассмотрим пару случайных переменных (X, Y), принимающую значения в множестве $\mathcal{X} \times \{-1, 1\}$. Предполагаем, что этой паре соответствует распределение вероятностей P и определены апостериорные вероятности принадлежности объекта $x \in \mathcal{X}$ к первому и второму классам: $P\{Y = 1|X = x\}$ и $P\{Y = -1|X = x\}$.

Легко построить оптимальный классификатор, если распределение вероятностей P, генерирующее пары "вход-выход", известно.

Обозначим условную вероятность того, что объект x принадлежит первому классу

$$\eta(x) = P\{Y = 1|X = x\}.$$

Для произвольного классификатора $g : \mathcal{X} \to \{-1, 1\}$ вероятность ошибки классификации равна

$$\mathrm{err}_P(g) = P\{g(X) \neq Y\}.$$

Байесовский классификатор определяется как

$$h(x) = \begin{cases} 1, \text{ если } \eta(x) > \frac{1}{2}, \\ -1 \text{ в противном случае.} \end{cases}$$

Следующая лемма показывает, что байесовский классификатор минимизирует вероятность ошибки $\mathrm{err}_P(h)$, которая в данном случае называется *байесовской ошибкой*.

Лемма 1.1. *Для любого классификатора* $g : \mathcal{X} \to \{-1, 1\}$

$$P\{h(X) \neq Y\} \leqslant P\{g(X) \neq Y\}. \tag{1.1}$$

Доказательство. Для произвольного классификатора g условная вероятность ошибки классификации при $X = x$ выражается в виде

$$\begin{aligned} P\{g(X) \neq Y | X = x\} = \\ = 1 - P\{g(X) = Y | X = x\} = \\ = 1 - (P\{Y = 1, g(X) = 1 | X = x\} + \\ + P\{Y = -1, g(X) = -1 | X = x\}) = \\ = 1 - (1_{g(x)=1} P\{Y = 1 | X = x\} + \\ + 1_{g(x)=-1} P\{Y = -1 | X = x\}) = \\ = 1 - (1_{g(x)=1} \eta(x) + 1_{g(x)=-1}(1 - \eta(x))\}), \end{aligned}$$

где для любого условия $R(x)$ будет $1_{R(x)} = 1$, если $R(x)$ выполнено, и $1_{R(x)} = 0$, в противном случае.

Аналогичное равенство выполнено для классификатора $h(x)$.

Заметим, что $1_{g(x)=-1} = 1 - 1_{g(x)=1}$ для любой функции классификации g. Таким образом, для каждого $x \in \mathcal{X}$

$$\begin{aligned} P\{g(X) \neq Y | X = x\} - P\{h(X) \neq Y | X = x\} = \\ = \eta(x)(1_{h(x)=1} - 1_{g(x)=1}) + \\ + (1 - \eta(x))(1_{h(x)=-1} - 1_{g(x)=-1}) = \\ = (2\eta(x) - 1)(1_{h(x)=1} - 1_{g(x)=1}) \geqslant 0 \end{aligned}$$

по определению байесовского классификатора h.

Интегрируем обе части этого неравенства по x. Получим неравенство леммы. $\triangle$

Байесовский классификатор служит эталоном для оценки качества алгоритмов классификации.

Обозначим посредством $\mathcal{D}$ множество всех измеримых функций классификаторов типа $g : \mathcal{X} \to \{-1, 1\}$. Условие (1.1) можно записать в виде

$$\mathrm{err}_P(h) = P\{h(X) \neq Y\} = \inf_{g \in \mathcal{D}} P\{g(X) \neq Y\}.$$

Пусть некоторый классификатор $g_l \in \mathcal{D}$ построен некоторым алгоритмом $\mathcal{A}$ по случайной выборке $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, сгенерированной распределением вероятностей P. Алгоритм классификации $\mathcal{A}$ называется *состоятельным* для распределения P, если случайная величина $\mathrm{err}_P(g_l)$ сходится к $\mathrm{err}_P(h)$ по вероятности P, т.е. для любого $\epsilon > 0$

$$P\{|\mathrm{err}_P(g_l) - \mathrm{err}_P(h)| > \epsilon\} \to 0 \tag{1.2}$$

при $l \to \infty$.

Алгоритм классификации $\mathcal{A}$ называется *универсально состоятельным*, если условие (1.2) имеет место для любого распределения P.

Недостатком байесовского классификатора является то, что он использует для вычисления значений функции $h(x)$ вероятностное распределение P, генерирующее пары (x, y). Прежде чем использовать байесовский классификатор, надо решить задачу восстановления вероятностного распределения P по его реализациям. На практике такое вероятностное распределение часто неизвестно и его трудно восстановить. Обычно для получения достоверного результата требуется довольно много реализаций случайной величины (X, Y).

Основные проблемы статистической теории классификации связаны с тем, что при построении классификаторов $h(x)$ мы не можем использовать распределения вероятностей, генерирующие пары (x, y).

Таким образом, в дальнейшем будут рассматриваться классификаторы, не зависящие от вероятностного распределения, генерирующего данные.

Байесовский классификатор служит для сравнения предсказательной способности других алгоритмов классификации.

1.1.2. Постановка задачи классификации

Важная проблема, возникающая в статистической теории машинного обучения, заключается в том, как много случайных примеров необходимо использовать при обучении для того, чтобы гаранти-

ровать достаточно малую ошибку классификации с заданной степенью достоверности.

Сначала напомним основные идеи PAC-теории машинного обучения (Probably Approximately Correct-learning), предложенную Валиантом [55].

В данном случае формальная постановка задачи основана на вероятностных предположениях. Мы предполагаем, что все примеры, представленные для обучения или проверки, независимо и одинаково распределены согласно некоторому фиксированному неизвестному распределению вероятностей P на множестве $\mathcal{X}$ этих примеров, имеющем структуру вероятностного пространства.

Предполагаем, что каждый пример $x \in \mathcal{X}$ имеет метку – признак принадлежности к некоторому классу. Метки классов образуют множество D и задаются с помощью неизвестной нам функции $c \in C$ типа $c : \mathcal{X} \to D$, которая называется *концептом*: $c(x)$ – метка x.

Допустим, что по некоторой случайной выборке

$$S = ((x_1, c(x_1)), \ldots, (x_l, c(x_l)),$$

порожденной распределением P, мы построили гипотезу $h = h_S$, которая выражает принадлежность объектов x к классам (подмножествам $\mathcal{X}$), порожденным неизвестным нам концептом c. Ошибка гипотезы h определяется как

$$\mathrm{err}_P(h) = P\{h(x) \neq c(x)\}.$$

Ошибка $\mathrm{err}_P(h)$ является случайной величиной, так как функция $h = h_S$ зависит от S.

Рассмотрим задачу: найти такую гипотезу h, для которой вероятность события, заключающегося в том, что ошибка $\mathrm{err}_P(h)$ велика, является малой. Другими словами, мы хотели бы утверждать, что гипотеза h *вероятно приблизительно верна* (probably approximately correct).

Степень “приблизительности” количественно будет выражаться с помощью параметра ϵ: мы будем требовать выполнения неравенства $\mathrm{err}_P(h) \leqslant \epsilon$.

Степень “вероятности” будет измеряться с помощью параметра уровня доверия δ. Мы хотим получить хорошую аппроксимацию концепта $c \in C$ с высокой вероятностью. В частности, мы требуем, чтобы неравенство $\mathrm{err}_P(h) \leqslant \epsilon$ выполнялось бы с вероятностью не меньшей чем $1 - \delta$.

Все эти соображения приводят к следующей точной формулировке PAC-теории машинного обучения.

Алгоритм A восстанавливает класс концептов C с помощью класса гипотез H, если для любого концепта $c \in C$, для любого распределения вероятностей P на примерах x, а также для любых $\epsilon \in (0, 1/2)$ и $\delta \in (0, 1/2)$ выполнено следующее:

- алгоритм A получает на вход обучающую выборку, состоящую из случайных пар $(x, c(x))$, независимо и одинаково распределенных согласно P, число которых полиномиально зависит от $1/\epsilon$ и $1/\delta$;

- алгоритм A выдает в качестве результата функцию h, для которой $\mathrm{err}_P(h) \leqslant \epsilon$ с вероятностью не менее $1 - \delta$.

В этом случае говорим, что класс концептов C является PAC-изучаемым (Probably Approximately Correct learnable).

В данной работе будет рассматриваться постановка задачи, несколько отличная от классической постановки задачи PAC-теории машинного обучения. Мы не используем понятие концепта, вместо этого мы просто предполагаем, что пары (x, y) объектов x и их меток y одинаково и независимо распределены согласно некоторому неизвестному вероятностному распределению P на множестве $\mathcal{X} \times D$. Подобная постановка принята в современной *статистической теории машинного обучения*. В остальном все идеи PAC-теории сохраняются.

Мы предполагаем, что выборка $S = ((x_1, y_1), \ldots, (x_l, y_l))$ генерируется (порождается) некоторым источником. Основное предположение об источнике, порождающем выборку S, заключается в том, что на парах (x, y), т.е. на пространстве $\mathcal{X} \times D$, задано распределение вероятностей P, а пары (x_i, y_i), образующие выборку S, одинаково и независимо распределены.

Соответственно на множестве $(\mathcal{X} \times D)^l$ задано распределение вероятностей $P^l = P \times P \times \cdots \times P$.

Правило, или функция (гипотеза), классификации – это функция типа $h : \mathcal{X} \to D$, которая разбивает элементы $x_i \in \mathcal{X}$ на несколько классов. Мы будем также называть функцию h классификатором, или решающим правилом.

В дальнейшем у нас всегда будет рассматриваться случай бинарной классификации $D = \{-1, 1\}$, а функция $h : \mathcal{X} \to D$ будет называться индикаторной.

В этом случае вся выборка S разбивается на две подвыборки: $S^+ = ((x_i, y_i) : y_i = 1)$ – положительные примеры (или первый класс) и $S^- = ((x_i, y_i) : y_i = -1)$ – отрицательные примеры (или второй класс).

В некоторых случаях индикаторная функция классификации h задается с помощью некоторой вещественной функции f и числа $r \in \mathcal{R}$:

$$h(x) = \begin{cases} 1, \text{ если } f(x) \geqslant r, \\ -1 \text{ в противном случае.} \end{cases}$$

Предсказательная способность произвольной функции классификации h будет оцениваться по *ошибке классификации*, которая определяется как вероятность неправильной классификации

$$\mathrm{err}_P(h) = P\{h(x) \neq y\} = P\{(x, y) : h(x) \neq y\}.$$

Функция $\mathrm{err}_P(h)$ также называется *риск-функционалом.*

Основная цель при решении задачи классификации – для заданного класса функций классификации H построить оптимальный классификатор, т.е. такую функцию классификации $h \in H$, при которой ошибка классификации $\mathrm{err}_P(h)$ является наименьшей в классе H.

В этой главе в основном будет рассматриваться задача классификации n-мерных векторов – элементов множества $\mathcal{R}^n$, где $\mathcal{R}$ – множество всех действительных чисел. Далее D – множество классов этих векторов – конечное множество с небольшим числом элементов. Размерность n евклидового пространства $\mathcal{R}^n$ обычно велика по сравнению с числом классов.

Далее элементы $\mathcal{R}^n$ будем обозначать подчеркнутыми сверху буквами: $\bar{x}, \bar{y}, \ \ldots \ \in \mathcal{R}^n$; в координатах – $\bar{x} = (x_1, \ \ldots, \ x_n)$. Будут рассматриваться операции сложения векторов

$$\bar{x} + \bar{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \ldots \\ x_n + y_n \end{pmatrix}$$

умножения на вещественное число

$$\alpha\bar{x} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \ldots \\ \alpha x_n \end{pmatrix},$$

где $\bar{x} = (x_1, \ \ldots, \ x_n)'$ и $\bar{y} = (y_1, \ \ldots, \ y_n)'$.

С помощью штриха мы уточняем форму представления вектора в виде матрицы – простую или транспонированную, но только в тех случаях когда это имеет существенное значение.

На векторах из $\mathcal{R}^n$ также определено их скалярное произведение $(\bar{x} \cdot \bar{y}) = x_1 y_1 + \cdots + x_n y_n$. Норма (длина) вектора $\bar{x}$ определяется как $\|\bar{x}\| = \sqrt{(\bar{x} \cdot \bar{x})} = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$. При решении задачи классификации мы исходим из *обучающей выборки* $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{X}$ – вектор евклидового пространства $\mathcal{R}^n$ размерности n (например, это может быть цифровой образ какого-либо изображения), y_i – это элемент конечного множества D с небольшим числом элементов (метка класса), например, $y_i \in \{-1, 1\}$. Элементы $y_i \in D$ определяют классы объектов $\bar{x}_i$.

При решении задачи многомерной регрессии также рассматривается обучающая выборка $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, при этом элементы y_i обычно являются вещественными числами, т.е. $D = \mathcal{R}$. Задача регрессии будет рассмотрена в разделах 2.8, 2.9, а также в разделе 5.10.

1.1.3. Линейные классификаторы: персептрон

Рассмотрим один из наиболее старых алгоритмов классификации – персептрон.

Персептрон представляет собой некоторую техническую модель восприятия. Модель имеет два слоя. Первый, рецепторный слой подает сигнал на входы пороговых элементов – нейронов преобразующего слоя.

Математическая модель *персептрона* будет задаваться следующим образом. Задано пространство $\mathcal{X}$ исходных описаний объекта. Преобразование $\bar{y} = \bar{\varphi}(\bar{x})$, которое в координатном виде записывается как $y_i = \varphi_i(\bar{x})$, $i = 1, \ldots, n$, ставит исходному описанию $\bar{x} = (x_1, \ldots, x_m) \in \mathcal{X}$ объекта преобразованное описание объекта $\bar{y} = (y_1, \ldots, y_n) \in \mathcal{Y}$. Предполагаем, что $\mathcal{X} \subseteq \mathcal{R}^m$ и $\mathcal{Y} \subseteq \mathcal{R}^n$ для некоторых m, n.

Персептрон задается однородной линейной функцией

$$L(\bar{x}) = (\Lambda \cdot \bar{\varphi}(\bar{x})) = \sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) = \sum_{i=1}^{n} \lambda_i y_i,$$

где действительные числа λ_i интерпретируются как веса, приписываемые преобразованным признакам y_i. Здесь $(\Lambda \cdot \bar{\varphi}(\bar{x}))$ обозначает скалярное произведение двух векторов $\Lambda = (\lambda_1, \ldots, \lambda_n)$ и $\bar{\varphi}(\bar{x}) = (\varphi_1(\bar{x}), \ldots, \varphi_n(\bar{x}))$ в евклидовом пространстве $\mathcal{R}^n$.

Будем связывать с персептроном *функцию активации*:

$$f(\bar{x}) = \sigma\left(\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x})\right).$$

Примеры функций активации:

$$\sigma(t) = \operatorname{sign}(t),$$
$$\sigma(t) = \frac{1}{1 + e^{-t}},$$
$$\sigma(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}},$$

где

$$\operatorname{sign}(t) = \begin{cases} 1, \text{ если } t \geqslant 0, \\ -1, \text{ если } t < 0. \end{cases}$$

В дальнейшем для простоты будем использовать бинарную функцию активации $\sigma(t) = \operatorname{sign}(t)$, которая определяет следующий классификатор. Считаем, что вектор $\bar{x}$ принадлежит первому классу, если

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) \geqslant 0.$$

В противном случае вектор $\bar{x}$ принадлежит второму классу.

Геометрически это означает, что в пространстве признаков $\mathcal{X}$ задана гиперповерхность

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\bar{x}) = 0, \tag{1.3}$$

которая делит пространство $\mathcal{X}$ на два полупространства. Объекты первого класса относятся к одному полупространству, объекты второго класса относятся ко второму полупространству. Подобная гиперповерхность называется *разделяющей*.

Каждой разделяющей гиперповерхности (1.3) соответствует разделяющая гиперплоскость

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

в пространстве преобразованных признаков $\mathcal{Y}$. Пространство $\mathcal{Y}$ также называется *спрямляющим*.

Пусть задана бесконечная обучающая выборка в спрямляющем пространстве

$$S = ((\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2), \ \dots),$$

где ϵ_i обозначает принадлежность объекта $\bar{y}_i = \bar{\varphi}(\bar{x}_i)$ классу $\epsilon_i \in \{-1, 1\}$, $i = 1, 2, \ \dots$

Допустим, что существует гиперплоскость, строго разделяющая выборку S. Пусть $\Lambda = (\lambda_1, \ \dots, \ \lambda_n)$ – вектор коэффициентов этой разделяющей гиперплоскости. По определению гиперплоскость строго разделяет выборку, если выполнены неравенство

$$\inf_i \epsilon_i (\Lambda \cdot \bar{y}_i) > 0. \tag{1.4}$$

Для удобства преобразуем обучающую выборку следующим образом. Рассмотрим последовательность векторов $\tilde{y}_1, y_2, \ \dots$, где

$$\tilde{y}_i = \begin{cases} \bar{y}_i, \text{ если } \epsilon_i = 1, \\ -\bar{y}_i, \text{ если } \epsilon_i = -1, \end{cases}$$

для всех i. Тогда условие строгого разделения (1.4) запишется в виде

$$\inf_i (\Lambda \cdot \tilde{y}_i) > 0.$$

Обозначим

$$\rho(\Lambda) = \inf_i \frac{(\Lambda \cdot \tilde{y}_i)}{\|\Lambda\|},$$
$$\rho_0 = \sup_{\Lambda \neq \bar{0}} \rho(\Lambda), \tag{1.5}$$

где $\|\Lambda\| = \sqrt{\sum\limits_{i=1}^{n} \lambda_i^2}$ – длина вектора Λ в пространстве $\mathcal{R}^n$.

Условие строгой разделимости выборки S может быть записано в виде $\rho_0 > 0$.

Алгоритм Розенблатта построения разделяющей гиперплоскости.

Пусть задана произвольная бесконечная обучающая выборка

$$(\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2), \ \dots$$

и пусть существует гиперплоскость, проходящая через начало координат $(\Lambda^* \cdot \bar{y}) = 0$, строго разделяющая эту выборку, т.е. такая, что

$$\inf_i (\Lambda^* \cdot \tilde{y}_i) > 0.$$

Считаем, что $\|\Lambda^*\| = 1$.

Пусть задан порог разделения – число $\rho_0 > 0$ такое, что

$$(\Lambda^* \cdot \tilde{y}_i) > \rho_0 \tag{1.6}$$

для всех i. Также предполагаем, что векторы $\bar{y}_i$ равномерно ограничены по норме

$$\sup_i \|\bar{y}_i\| = D < \infty.$$

Полагаем $\Lambda_0 = (0,\ \ldots,\ 0)$.
FOR $t\ =\ 1,\ \ldots,\ T$
Если $(\Lambda_{t-1} \cdot \tilde{y}_t) > 0$, то полагаем $\Lambda_t = \Lambda_{t-1}$.
(т.е. если очередной вектор классифицируется правильно, то текущая гиперплоскость не изменяется).
Если $(\Lambda_{t-1} \cdot \tilde{y}_t) \leqslant 0$ (очередной вектор классифицируется неправильно), то производим корректировку вектора гиперплоскости $\Lambda_t = \Lambda_{t-1} + \tilde{y}_t$, назовем эту операцию также *исправлением ошибки.*
ENDFOR

Рис. 1.1: Алгоритм Розенблатта построения разделяющей гиперплоскости

Обучение персептрона заключается в изменении координат вектора весов Λ на каждом шаге алгоритма, приведенного на рис. 1.1.

Пусть $\Lambda_t = (\lambda_{1,t},\ \ldots,\ \lambda_{n,t})$ – текущий вектор коэффициентов гиперплоскости, вычисленный на шаге t алгоритма, $t = 1, 2,\ \ldots$.

Алгоритм использует преобразованную последовательность векторов $\tilde{y}_1, \tilde{y}_2,\ \ldots$.

Следующая теорема, принадлежащая А.А. Новикову, утверждает, что в том случае, когда существует гиперплоскость разделяющая выборку с положительным порогом, алгоритм Розенблатта после многократного предъявления обучающей последовательности, составленной из элементов выборки, построит за конечное число шагов гиперплоскость, строго разделяющую всю выборку.

Теорема 1.1. *Если существует гиперплоскость, разделяющая бесконечную выборку*

$$(\bar{y}_1, \epsilon_1), (\bar{y}_2, \epsilon_2),\ \ldots$$

с положительным порогом, то в алгоритме Розенблатта исправление ошибки происходит не более чем

$$\left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$$

раз.[1] *Это значит, что неравенство* $\Lambda_t \neq \Lambda_{t-1}$ *выполнено для не более чем* $\left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$ *различных* t.

После этого, разделяющая гиперплоскость стабилизируется и будет безошибочно делить всю бесконечную оставшуюся часть последовательности.

Доказательство. Если на шаге t происходит изменение вектора Λ_t, то

$$\|\Lambda_t\|^2 = \|\Lambda_{t-1}\|^2 + 2(\Lambda_{t-1} \cdot \tilde{y}_t) + \|\tilde{y}_t\|^2.$$

Так как $(\Lambda_{t-1} \cdot \tilde{y}_t) \leqslant 0$ (классификация t-го вектора неправильная) и $\|\tilde{y}_t\| \leqslant D$, получаем

$$\|\Lambda_t\|^2 \leqslant \|\Lambda_{t-1}\|^2 + D^2.$$

Если до шага T включительно произошло k таких исправлений, то получаем

$$\|\Lambda_t\|^2 \leqslant kD^2. \tag{1.7}$$

По условию разделимости (1.6) существует единичный вектор Λ^* такой, что

$$(\Lambda^* \cdot \tilde{y}_i) \geqslant \rho_0$$

для всех i.

Оценим величину $(\Lambda_t \cdot \Lambda^*)$. По определению $(\Lambda_0 \cdot \Lambda^*) = 0$. Если на шаге t алгоритм производит исправление, то

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*) + (\Lambda^* \cdot \tilde{y}_t) \geqslant (\Lambda_{t-1} \cdot \Lambda^*) + \rho_0.$$

Если на шаге t исправления не происходит, то

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*).$$

Таким образом, если к шагу t алгоритм произвел k исправлений, то

$$(\Lambda_t \cdot \Lambda^*) \geqslant k\rho_0.$$

[1] Здесь и далее, для любого вещественного числа r через $\lfloor r \rfloor$ обозначается максимальное целое число не больше r, а $\lceil r \rceil$ обозначает минимальное целое число не меньше r.

По неравенству Коши–Буняковского

$$(\Lambda_t \cdot \Lambda^*) \leqslant \|\Lambda_t\| \cdot \|\Lambda^*\| = \|\Lambda_t\|.$$

Поэтому имеет место неравенство

$$\|\Lambda_t\| \geqslant k\rho_0. \tag{1.8}$$

Объединяем неравенства (1.7) и (1.8), получаем

$$k \leqslant \frac{D^2}{\rho_0^2}.$$

Таким образом, число исправлений не превосходит

$$k \leqslant \left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor.$$

Теорема доказана. $\triangle$

По теореме 1.1, какова бы ни была бесконечная разделимая с положительным порогом выборка, алгоритм Розенблатта, сделав конечное число исправлений, не превосходящее $\lfloor \frac{D^2}{\rho_0^2} \rfloor$, найдет какую-либо гиперплоскость, строго разделяющую всю выборку.

В некоторых случаях в персептроне рассматривается бинарное спрямляющее пространство, т.е. $\mathcal{Y} = \{-1, 1\}^n$.

В этом случае ясно, что $D^2 \leqslant n$. Тогда оценка теоремы 1.1 имеет вид

$$k \leqslant \lfloor \frac{n}{\rho_0^2} \rfloor,$$

т.е. число коррекций алгоритма обучения персептрона растет линейно с размерностью пространства.

В этом разделе была рассмотрена двухуровневая модель персептрона. На первом уровне определяется отображение $\bar{y} = \bar{\phi}(\bar{x})$ исходного пространства описаний объектов $\mathcal{X}$ в спрямляющее пространство $\mathcal{Y}$. На втором уровне реализуется алгоритм обучения – построение разделяющей гиперплоскости в пространстве $\mathcal{Y}$ на основе обучающей последовательности. Основное требование к отображению $\bar{\phi}$ диктует вторая часть модели, а именно, множества векторов – образов $\bar{y}$, принадлежащих к различным классам должны быть разделимы гиперплоскостью.

Многослойная нейронная сеть. Персептроны можно комбинировать в виде *многослойных нейронных* сетей. В каждой вершине ν такой сети располагается некоторая функция

$$f^{\nu}(\bar{x}) = \sigma((\bar{w}^{\nu} \cdot \bar{x}) + b^{\nu}),$$

σ – функция активации; на место аргумента в ней подставлено значение персептрона.

Рассмотрим *сеть* вершин, состоящую из l слоев. Заданы натуральные числа $n_1, \ldots, n_l$ – размеры слоев (число вершин в слое), причем, самый верхний слой состоит из одной вершины: $n_l = 1$.

С каждой j-й вершиной i-го слоя сети ассоциируется функция $f_{i,j}(\bar{x}) = \sigma((\bar{w}^{i,j} \cdot \bar{x}) + b^{i,j})$, где $\bar{w}^{i,j}, \bar{x} \in \mathcal{R}^{n_{i-1}}$ и $b^{i,j} \in \mathcal{R}$.

Нейронная сеть может быть представлена в виде набора векторнозначных функций

$$f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i},$$

$i = 1, \ldots, l$, где $f_i = (f_{i,1}, \ldots, f_{i,n_{i-1}})$.

Выход нейронной сети задается одномерной функцией – композицией

$$f_l \circ f_{l-1} \circ \cdots \circ f_2 \circ f_1.$$

Векторы $\bar{w}^{i,j}$ называются весами, которые приписаны вершинам (i, j) нейронной сети.

1.2. Теория обобщения

1.2.1. Верхние оценки вероятности ошибки классификации

В теории обобщения вычисляются вероятности ошибки классификации на *тестовой выборке*, после того как функция классификации определена по *обучающей выборке*, т.е. проведено обучение алгоритма классификации.

В этом разделе мы приведем основные положения статистической теории обобщения.

Статистическая теория машинного обучения использует гипотезу о том, что пары (x_i, y_i) генерируются некоторым *неизвестным* нам распределением вероятностей, при этом, как правило, рассматривается очень широкий класс таких распределений. Используется только предположение о том, что данные независимо и одинаково распределены.

В статистической теории машинного обучения исходят из *обучающей выборки*, по которой определяется функция классификации или регрессии, наилучшим образом описывающая эту выборку. Класс функций классификации может быть очень широк – от разделяющих гиперплоскостей в n-мерном пространстве до произвольных многообразий, которые отображаются с помощью ядерных методов в гиперплоскости, расположенные в пространствах большей размерности $m > n$. Никакие распределения вероятностей не используются в алгоритмах, вычисляющих значения функции классификации.

Функция классификации проверяется на *тестовой выборке*. Задача теории обобщения состоит в том, чтобы оценить вероятность ошибки классификации на произвольной тестовой выборке.

Теория обобщения Вапника–Червоненкиса позволяет вычислить вероятность ошибки классификации или регрессии (относительно распределения вероятностей, генерирующего данные, возможно неизвестного нам) для согласованной по обучающей выборке функции классификации или регрессии на любых будущих данных.

Такая вероятность зависит от размера обучающей выборки и размерности или емкости класса функций, описывающих данные.

Емкость класса функций не зависит от числа параметров этих функций или от аналитического способа их задания. Она зависит от геометрических свойств класса – максимального размера проекций функций этого класса на выборки заданной длины.

В этом разделе будут даны равномерные верхние оценки вероятности ошибки в зависимости от длины обучающей выборки и размерности класса функций классификации.

Критерий выбора функции классификации основан на минимизации верхней оценки вероятности ошибки обобщения.

Пусть $S = ((x_1, y_1),\ \ldots, (x_l, y_l))$ – обучающая выборка. Здесь $x_i \in \mathcal{X}$ и $y_i \in \{-1, 1\}$ при $1 \leqslant i \leqslant l$. Элементы $\mathcal{X}$ называются *объектами*, а элементы D называются *метками*.

В приложениях обычно $\mathcal{X} \subseteq \mathcal{R}^n$, где $\mathcal{R}^n$ – n-мерное евклидово векторное пространство.

Предполагаем, что на множестве $\mathcal{X} \times \mathcal{D}$ задана структура вероятностного пространства с распределением P. Посредством P^l обозначаем вероятностное распределение на выборках S длины l, которое есть произведение l экземпляров распределения P.

Рассмотрим общий случай. Задан класс H функций классификации и функция потерь $\lambda(y, y')$, где $y, y' \in D$. Для произвольной функции классификации $h \in H$ и выборки $S = ((x_1, y_1),\ \ldots, (x_l, y_l))$ определяется выборочная (эмпирическая) ошибка

$$L_S(h) = \frac{1}{l} \sum_{i=1}^{l} \lambda(h(x_i), y_i).$$

Предполагаем, что пары (x_i, y_i) независимо и одинаково распределены по P, обозначаем это $(x_i, y_i) \sim P$. Обозначим математическое ожидание потерь при использовании классификатора h

$$E_P(h) = E_{(x,y)\sim P}(\lambda(h(x), y))$$

Из определения $E_{S\sim P^l}(E_S(h)) = E_P(h)$.

Если функция потерь ограниченная, то имеет место усиленный закон больших чисел: для фиксированного классификатора h, с вероятностью единица,

$$L_S(h) = \frac{1}{l} \sum_{i=1}^{l} \lambda(h(x_i), y_i) \to E_P(h)$$

при $l \to \infty$. Соответствующее неравенство больших уклонений имеет вид

$$P^l\{S : L_P(h) - L_S(h) > t\} \leqslant \Delta(h, l, t),$$

где $\Delta(h, l, t) \to 0$ при $l \to \infty$.

Функция классификации $h \in H$ будет определяться по случайной выборке S, поэтому нам необходимо неравенство больших уклонений с не зависящей от h оценкой вида:

$$P^l\{S : \sup_h (L_P(h) - L_S(h)) > t\} =$$
$$= P^l\{S : (\exists h \in H)(L_P(h) - L_S(h)) > t\} \leqslant \Delta(l, t). \qquad (1.9)$$

Обозначим $\delta = \Delta(l, t)$ и допустим, что мы можем выразить $t = t(l, \delta)$. Тогда PAC-формулировка оценки (1.9) выглядит следующим образом: для произвольного $\delta > 0$, с вероятностью $\geqslant 1 - \delta$, для каждого $h \in H$ выполнено

$$L_P(h) \leqslant L_S(h) + t(l, \delta). \qquad (1.10)$$

Основная задача классификации – построение классификатора $h \in H$, который минимизирует риск-функционал (математическое ожидание ошибки) $E_P(h)$. Так как мы не можем прямо вычислить величину $E_P(h)$, а имеем только оценку (1.10), будем решать две задачи: (1) минимизация выборочного среднего функции потерь $E_S(h)$ при заданной случайной выборке S; (2) нахождения нервенств больших уклонений с минимальными оценками $\Delta(l, t)$ и $t(l, \delta)$.

Предварительно рассмотрим частный случай – задачу оценки ощибки классификации для случая, когда функциональный класс H содержит гипотезу классификации, согласованную с выборкой.

Пусть $D = \{-1, 1\}$ и $\lambda(y, y') = 1_{y \neq y'}$. Обозначаем

$$\mathrm{err}_S(h) = \frac{1}{l}|\{i : h(x_i) \neq y_i, 1 \leqslant i \leqslant l\}|$$

относительное число ошибок классификации h на выборке S. Здесь $|A|$ – число элементов множества A. Имеем

$$L_S(h) = \mathrm{err}_S(h) = \frac{1}{l}\sum_{i=1}^{l} 1_{h(x_i) \neq y_i} = \frac{1}{l}|\{i : h(x_i) \neq y_i\}|.$$

Пусть задано правило классификации $h : \mathcal{X} \to \{-1, 1\}$. Тогда

$$\mathrm{err}_P(h) = E_{(x,y)\sim P}(1_{h(x) \neq y}) = P\{(x, y) : h(x) \neq y\}.$$

Эта величина равна вероятности неправильной классификации.

Гипотеза классификации h согласована с выборкой

$$S = ((x_1, y_1),\ \ldots,\ (x_l, y_l)),$$

если $h(x_i) = y_i$ для всех $1 \leqslant i \leqslant l$. Другими словами, гипотеза классификации h согласована с выборкой S, если $\mathrm{err}_S(h) = 0$.

Для произвольной гипотезы классификации h и $\epsilon > 0$ имеем

$$\begin{aligned} P^l\{S : \mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon\} = \\ = \prod_{i=1}^{l} P\{h(x_i) = y_i\} = \\ = \prod_{i=1}^{l} (1 - P\{h(x_i) \neq y_i\}) = \\ = (1 - \mathrm{err}_P(h))^l \leqslant e^{-l\epsilon}. \end{aligned} \tag{1.11}$$

Здесь мы использовали независимость ошибок на элементах выборки.

Пусть H – некоторый класс гипотез классификации. Если класс H конечный, то из (1.11) получаем оценку

$$P^l\{S : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant |H| e^{-l\epsilon}. \tag{1.12}$$

Интерпретация (1.12) заключается в следующем.

Пусть задан критический уровень $\delta > 0$ принятия ошибочной гипотезы классификации $h \in H$, согласованный с обучающей выборкой S. Тогда по (1.12) мы можем утверждать, что с вероятностью не меньше $1 - \delta$ гипотеза классификации $h_S \in H$, построенная по случайной обучающей выборке S и согласованная с ней, будет иметь ошибку классификации [2] $\mathrm{err}_P(h) \leqslant \epsilon = \frac{1}{l} \ln \frac{|H|}{\delta}$.

Другими словами, всякая гипотеза классификации h, имеющая ошибку $\mathrm{err}_P(h) > \epsilon$, с вероятностью не меньше $1 - |H| e^{-l\epsilon}$, не будет согласована со случайной выборкой длины l.

[2] В дальнейшем $\ln r$ обозначает натуральный логарифм положительного числа r, а $\log r$ будет обозначать логарифм r по основанию 2.

В случае бесконечного семейства функций H аналогичные оценки на ошибку классификации дает теория обобщения Вапника–Червоненкиса. Сложность класса H оценивается с помощью функции роста

$$B_H(l) = \max_{(x_1, x_2, \ \ldots, \ x_l)} |\{(h(x_1), h(x_2), \ \ldots, \ h(x_l)) : h \in H\}|.$$

Свойства этой функции будут изучаться далее.

Далее мы приводим две оценки ошибки обобщения для случая обучения с ошибками – теорема 1.2 для общего случая, когда допускаются ошибки при обучении классификатора – теорема 1.3.

Обучение без ошибок. Имеет место теорема – аналог соотношения (1.12) для бесконечного H.

Теорема 1.2. *При $l > 2/\epsilon$ имеет место оценка*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant 2B_H(2l)e^{-\epsilon l/4}.$$

Доказательство теоремы. Пусть $1_{h(\bar{x}) \neq y}$ есть величина, равная 1, если $h(x) \neq y$ и равная 0, в противном случае. Тогда

$$E1_{h(x) \neq y} = \mathrm{err}_P(h),$$

где E – математическое ожидание по мере P. По определению

$$\mathrm{err}_S(h) = \frac{1}{l} \sum_{i=1}^{l} 1_{h(x_i) \neq y_i}$$

– частота ошибок классификации на выборке S.

Утверждение теоремы будет следовать из следующих двух лемм.

Лемма 1.2. *Пусть задан класс H функций классификации. Рассматриваются две случайные выборки S, S' длины l. Тогда для любого $\epsilon > 0$ при $l > 2/\epsilon$ имеет место неравенство*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant$$
$$\leqslant\ 2P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon)\}, \qquad (1.13)$$

где SS' – двойная выборка, составленная из элементов выборки S и из следующих за ними элементов выборки S'.

Доказательство. Легко видеть, что неравенство (1.13) эквивалентно неравенству

$$P^l\{S : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_P(h) > \epsilon\} \leqslant$$
$$\leqslant 2P^{2l}\{SS' : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon\}. \qquad (1.14)$$

Докажем (1.14). Для каждой выборки S из множества левой части неравенства (1.14) обозначим посредством h_S какую-нибудь функцию из класса H, для которой выполняются равенство $\mathrm{err}_S(h_S) = 0$ и неравенство $\mathrm{err}_P(h_S) > \epsilon$. Это случайная величина, зависящая от выборки.

Имеет место следующее неравенство между случайными величинами [3]

$$1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_P(h_S)\,>\,\epsilon} 1_{\mathrm{err}_P(h_S)\,-\,\mathrm{err}_{S'}(h_S)\,\leqslant\,\frac{1}{2}\epsilon} \leqslant$$
$$\leqslant 1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_{S'}(h_S)>\frac{1}{2}\epsilon}. \qquad (1.15)$$

Возьмем математическое ожидание по второй выборке S' от обеих частей неравенства (1.15). Получим неравенство для случайных величин, зависящих от первой выборки S :

$$1_{\mathrm{err}_S(h_S)\,=\,0\&\mathrm{err}_P(h_S)\,>\,\epsilon} P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \leqslant \frac{1}{2}\epsilon\} \leqslant$$
$$\leqslant P^l\{S' : \mathrm{err}_S(h_S) = 0\&\mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\}. \qquad (1.16)$$

Используя свойства биномиального распределения получаем

$$P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \leqslant \frac{1}{2}\epsilon\} =$$
$$= P^l\{S' : \mathrm{err}_{S'}(h_S) \geqslant \mathrm{err}_P(h_S) - \frac{1}{2}\epsilon\} =$$
$$= \sum_{\{k:k/l\geqslant p-\epsilon/2\}} \binom{l}{k} p^k (1-p)^{l-k} > \frac{1}{2} \qquad (1.17)$$

[3] Здесь h_S – произвольная и $1_{\mathrm{err}_S(h_S)=0\&\mathrm{err}_P(h_S)>\epsilon}(S) = 0$, если S не лежит в множестве из левой части неравенства (1.14).

при $l > 2/\epsilon$. Здесь $p = \mathrm{err}_P(h_S)$.

Действительно, при $l > 2/\epsilon$ будет $p - \epsilon/2 < p - 1/l$. Поэтому достаточно доказать, что

$$\sum_{\{k:k/l \geqslant p-1/l\}} \binom{l}{k} p^k (1-p)^{n-k} = \sum_{\{k:k \geqslant lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} > \frac{1}{2}.$$

Это неравенство эквивалентно неравенству

$$\sum_{\{k:k < lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} < \frac{1}{2},$$

которое следует из того, что медиана биномиального распределения равна целому числу ближайшему к числу lp.

Подставляя неравенство (1.17) в (1.16), получим

$$\begin{aligned} 1_{\mathrm{err}_S(h_S)=0 \& \mathrm{err}_P(h_S)>\epsilon} \leqslant \\ \leqslant 2P^l\{S' : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\}. \end{aligned} \tag{1.18}$$

Возьмем среднее по S и получим

$$\begin{aligned} P^l\{S : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_P(h_S) > \epsilon\} \leqslant \\ \leqslant 2P^{2l}\{SS' : \mathrm{err}_S(h_S) = 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon\} \leqslant \\ \leqslant 2P^{2l}\{SS' : \sup_{h:\mathrm{err}_S(h)=0} \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon\}. \end{aligned} \tag{1.19}$$

Отсюда получаем (1.14). Лемма доказана. $\triangle$

Лемма 1.3. *Вероятность того, что на двух случайных выборках S и S' длины l некоторая функция классификации $h \in H$ согласована с первой из них и совершает более ϵl ошибок на второй выборке ограничена величиной*

$$P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon)\} \leqslant B_H(2l)e^{-\epsilon l/2}.$$

Доказательство. Поскольку для фиксированной функции h ошибки классификации описываются бернуллиевским распределением с вероятностью ошибки $p = P\{h(x) \neq y\}$ вероятность того,

что на некоторой двойной выборке SS' все ошибки сосредоточены на второй половине этой выборки, оценивается сверху:

$$\frac{\binom{l}{m}}{\binom{2l}{m}} = \frac{l!}{(l-m)!m!} \cdot \frac{(2l-m)!m!}{(2l)!} =$$
$$= \frac{(2l-m)\dots(l-m+1)}{2l\dots(l+1)} \leqslant$$
$$\leqslant \left(1-\frac{m}{2l}\right)^l \leqslant \left(1-\frac{\epsilon}{2}\right)^l < e^{-\epsilon l/2}, \tag{1.20}$$

при $m \geqslant \epsilon l$.

Число всех проекций функций классификации из H на множество всех объектов $\{x_1, \dots, x_{2l}\}$ равно числу элементов множества

$$\{(h(x_1), h(x_2), \dots, h(x_{2l})) : h \in H\},$$

Оценку числа таких наборов дает *функция роста* семейства индикаторных функций H :

$$B_H(l) = \max_{(x_1, x_2, \dots, x_l)} |\{(h(x_1), h(x_2), \dots, h(x_l)) : h \in H\}|.$$

Ясно, что $B_H(l) \leqslant 2^l$. Точные оценки функции роста различных семейств классификаторов будут даны в разделе 1.2.2.

Из определения функции роста следует, что число всех ограничений функций классификации из H на выборках длины $2l$ не превосходит $B_H(2l)$.

Поэтому вероятность того, что некоторая функция классификации из класса H делает более ϵl ошибок на двойной выборке все они сосредоточены на второй половине этой выборки, ограничена сверху

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon)\} =$$
$$P^{2l}\{SS' : (\exists\, h \in H_{x_1,\dots x_{2l}})(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \epsilon)\} \leqslant$$
$$\leqslant B_H(2l)e^{-\epsilon l/2}.$$

Здесь $x_1, \dots x_{2l}$ обозначают объекты выборки SS', $H_{x_1,\dots,x_{2l}}$ – конечное множество функций, которые являются проекциями функций из класса H на множество этих объектов. Также использовано неравенство $|H_{x_1,\dots,x_{2l}}| \leqslant B_H(2l)$. $\triangle$

Теорема 1.2 непосредственно следует из лемм 1.2 и 1.3.

Из теоремы 1.2 следует, что с вероятностью не меньше $1 - 2B_H(2l)e^{-\epsilon l/4}$ всякая гипотеза классификации h, имеющая ошибку $\mathrm{err}_P(h) > \epsilon$, не будет согласована со случайной выборкой длины $l > 2/\epsilon$, т.е. будет отвергнута как ошибочная.

Обозначим $\delta = 2B_H(2l)e^{-\epsilon l/4}$. Тогда при $0 < \delta < 1$ будет выполнено $l\epsilon > 2$, т.е. условие теоремы 1.2 выполнено. Отсюда получаем следствие.

Следствие 1.1. *Допустим, что класс H функций классификации имеет конечную VC-размерность* [4] *d.*

Пусть задан критический уровень $0 < \delta < 1$ принятия ошибочной гипотезы классификации $h \in H$, согласованной с обучающей выборкой S.

Тогда при $l \geqslant d$ с вероятностью не ниже $1-\delta$ гипотеза классификации $h_S \in H$, построенная по случайной обучающей выборке S и согласованная с ней, будет иметь ошибку классификации

$$\mathrm{err}_P(h_S) \leqslant \frac{4}{l}\left(d \ln \frac{2el}{d} + \ln \frac{2}{\delta}\right).$$

Обучение с ошибками. Все эти результаты можно распространить на случай обучения с ошибками. При этом, оценки вероятности ошибки ослабляются. Аналогичным образом доказываются следующие две леммы 1.4 и 1.5, а также их следствие – теорема 1.3.

Лемма 1.4. *Пусть задан класс H функций классификации. Рассматриваются две случайные выборки S, S' длины l. Тогда для любого $\epsilon > 0$ при $l > 2/\epsilon^2$ имеет место неравенство*

$$P^l\{S : \exists\, h \in H)(\mathrm{err}_P(h) - \mathrm{err}_S(h) > \epsilon\} \leqslant$$
$$\leqslant 2P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \frac{1}{2}\epsilon)\}. \qquad (1.21)$$

Доказательство. Доказательство этой леммы аналогично доказательству леммы 1.3. Из определения супремума для каждой

[4] Определение VC-размерности дано в разделе 1.2.2. Там же получена оценка $B_H(l) \leqslant \left(\frac{el}{d}\right)^d$ при $l \geqslant d$.

S существует функция $h_S \in H$ так, что можно переписать неравенство (1.21) в виде

$$P^l\{S : \mathrm{err}_P(h_S) - \mathrm{err}_S(h_S) > \epsilon)\} \leqslant$$
$$\leqslant 2P^{2l}\{SS' : \mathrm{err}_{S'}(h_S) - \mathrm{err}_S(h_S) > \frac{1}{2}\epsilon\}. \qquad (1.22)$$

Вседа имеет место неравенство между случайными величинами

$$1_{\mathrm{err}_P(h_S)-\mathrm{err}_S(h_S)>\epsilon} \cdot 1_{\mathrm{err}_P(h_S)-\mathrm{err}_{S'}(h_S)<\frac{\epsilon}{2}} \leqslant$$
$$1_{\mathrm{err}_{S'}(h_S)-\mathrm{err}_S(h_S)>\frac{\epsilon}{2}}. \qquad (1.23)$$

Применим математическое ожидание $E_{S'\sim P^l}$ (возьмем интеграл) к обеим частям неравенства (1.23) и получим

$$1_{\mathrm{err}_P(h_S)-\mathrm{err}_S(h_S)>\epsilon} \cdot P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) < \frac{\epsilon}{2}\} \leqslant$$
$$P^l\{S' : \mathrm{err}_{S'}(h_S) - \mathrm{err}_S(h_S) > \frac{\epsilon}{2}\}. \quad (1.24)$$

Из неравенства Чебышева для бернуллиевских случайных величин с вероятностью исхода, равной $p = \mathrm{err}_P(h_S)$, будет

$$P\{\mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \geqslant \frac{\epsilon}{2}\} \leqslant \frac{1}{l\epsilon^2} \leqslant \frac{1}{2}$$

при $l > \frac{2}{\epsilon^2}$. Тогда (1.24) превращается в

$$1_{\mathrm{err}_P(h_S)-\mathrm{err}_S(h_S)>\epsilon} \leqslant 2P^l\{S' : \mathrm{err}_{S'}(h_S) - \mathrm{err}_S(h_S) > \frac{\epsilon}{2}\}. \quad (1.25)$$

Применяем $E_{S\sim P^l}$ к обеим частям (1.25) и получаем (1.21). $\triangle$

Лемма 1.5. *Вероятность того, что на двух случайных выборках S и S' длины l частоты ошибок некоторой функции классификации $h \in H$ различаются более чем на $\epsilon > 0$, ограничена величиной*

$$P^{2l}\{SS' : \sup_{h\in H}(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant B_H(2l)e^{-\frac{\epsilon^2 l}{8}}.$$

Доказательство. Доказательство этой леммы аналогично доказательству леммы 1.5.

$$P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant$$

$$P^{2l}\{SS' : (\exists h \in H_{x_1,\ldots,x_{2l}})(\mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant \tag{1.26}$$

$$B_H(2l)P^{2l}\{SS' : \mathrm{err}_{S'}(h) - \mathrm{err}_S(h) > \epsilon\} \leqslant \tag{1.27}$$

$$B_H(2l)P^{2l}\{SS' : \mathrm{err}_{S'}(h) - \mathrm{err}_P(h) > \frac{\epsilon}{2} \vee$$

$$\mathrm{err}_S(h) - \mathrm{err}_P(h) > \frac{\epsilon}{2}\} \leqslant 2B_H(2l)e^{-\frac{l\epsilon^2}{2}}. \tag{1.28}$$

Здесь в (1.26) $H_{x_1,\ldots,x_{2l}}$ – конечное множество функций, которые являются проекциями функций из класса H на множество объектов из выборки SS'. Переход от (1.26) к (1.27) использует неравенство $|H_{x_1,\ldots,x_{2l}}| \leqslant B_H(2l)$. Оценки в (1.27) основаны на неравенстве Хефдинга для фиксированной $h \in H$

$$P\{\mathrm{err}_P(h) - \mathrm{err}_S(h) > t\} \leqslant e^{-2lt^2}.$$

$\triangle$

Следующая теорема дает оценку вероятности отклонения риск-функционала от среднего числа ошибок на обучающей выборке.

Теорема 1.3. *Имеет место оценка*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_P(h) - \mathrm{err}_S(h) > \epsilon)\} \leqslant 4B_H(2l)e^{-\frac{\epsilon^2 l}{8}}$$

при $l > 2/\epsilon^2$.

Отсюда, а также из теоремы 1.4, приведенной ниже, получаем следствие, связывающее вероятность ошибки обобщения и среднее число ошибок классификатора на обучающей выборке.

Следствие 1.2. *Допустим, что класс H функций классификации имеет конечную VC-размерность d, $0 < \delta < 1$ и $l \geqslant d$. Тогда с вероятностью не менее $1 - \delta$ для каждого $h \in H$ выполнено*

$$\mathrm{err}_P(h) \leqslant \mathrm{err}_S(h) + \sqrt{\frac{8}{l}\left(d \ln \frac{2el}{d} + \ln \frac{4}{\delta}\right)}.$$

Данная оценка ошибки классификации заметно хуже соответствующей оценки вида $O(d \ln d/l)$ для случая, когда существует гипотеза классификации согласованная с выборкой.

Следует отметить, что оценки теорем 1.2 и 1.3, а также следствий 1.1 и 1.2 имеют в основном теоретическое значение, так как на практике VC-размерность d может быть сравнимой с длиной выборки l. Ближе к практике находятся оценки, основанные на понятиях размерности классов функций, которые не зависят от размерности пространства объектов.

1.2.2. VC-размерность

В этом разделе мы рассмотрим определение и свойства размерности Вапника–Червоненкиса – VC-размерности, которая характеризует "сложность" бесконечного класса функций классификации.

Пусть $\mathcal{X}$ – множество объектов и H – произвольный класс функций классификации на $\mathcal{X}$. Рассмотрим функцию $h \in H$ и произвольный набор – *выборку* элементов $(x_1, \ldots, x_l)$ из $\mathcal{X}$.

Бинарный набор $(h(x_1), \ldots, h(x_l))$, состоящий из элементов множества $\{-1, 1\}$, определяет разделение множества $\{x_1, \ldots, x_l\}$ на два подмножества $\{x_i : h(x_i) = 1\}$ – положительные примеры и $\{x_i : h(x_i) = -1\}$ – отрицательные примеры.

Выборка $(x_1, \ldots, x_l)$ *полностью разделена* функциями из H (shattered by H), если

$$\{(h(x_1), \ldots, h(x_l)) : h \in H\} = \{-1, 1\}^l.$$

Функция роста семейства классификаторов H определяется как максимальное число различных разбиений выборок длины l на два подмножества, которые можно осуществить с помощью функций из класса H:

$$B_H(l) = \max_{(x_1, x_2, \ldots, x_l)} |\{(h(x_1), h(x_2), \ldots, h(x_l) : h \in H\}|.$$

Ясно, что $B_H(l) \leqslant 2^l$, а если существует полностью разделимая (функциями из класса H) выборка из l элементов, то $B_H(l) = 2^l$.

Приведем основную теорему теории VC-размерности (лемму Сауэра).[5]

Теорема 1.4. *Для любого класса индикаторных функций H реализуется одна из двух возможностей:*

1) $B_H(l) = 2^l$ *для всех* l, *т.е. для любого* l *существует полностью разделимая выборка размера* l.

2) *Существует полностью разделимая выборка максимального размера* d; *в этом случае* $B_H(l) = 2^l$ *при* $l \leqslant d$ *и*

$$B_H(l) \leqslant \sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{el}{d}\right)^d \tag{1.29}$$

при $l > d$.

Другими словами, функция $G_H(l) = \ln B_H(l)$ *– линейная или начиная с некоторого значения ограничена логарифмической функцией* $O(d \ln l)$ *(например, она не может иметь вид* $O(\sqrt{l})$*).*

Число d *называется размерностью Вапника–Червоненкиса или VC-размерностью класса функций* H. *Обозначаем* $d = VCdim(H)$. *Если реализуется первый случай, то VC-размерность класса* H *бесконечная.*

Доказательство. Допустим, что VC-размерность некоторого класса индикаторных функций H равна d. Тогда по определению $B_H(l) = 2^l$ при всех $l \leqslant d$.

Мы докажем неравенство (1.29) математической индукцией по величине $l + d$. Для $l = d = 1$ это неравенство верно, так как обе его части равны 2.

Допустим, что это неравенство верно для любой суммы, меньшей $l + d$, в частности, для $l - 1$ и d, а также для $l - 1$ и $d - 1$.

Докажем его для случая, когда размер выборки равен l, а VC-размерность класса функций равна d.

Введем обозначение

$$h(l, d) = \sum_{i=0}^{d} \binom{l}{i}.$$

[5]Это утверждение было также независимо получено Вапником и Червоненкисом [2].

Тогда нам надо доказать, что для любого класса функций H с VC-размерностью не выше d будет выполняться $B_H(l) \leqslant h(l,d)$ для всех l.

Из свойства биномиальных коэффициентов:

$$\binom{l}{i} = \binom{l-1}{i} + \binom{l-1}{i-1},$$

получаем соответствующее свойство, связывающее значения функции h :

$$h(l,d) = h(l-1,d) + h(l-1,d-1).$$

Пусть H – произвольный класс функций, VC-размерность которого равна d и пусть $X_1 = (x_1, x_2, \ldots,\ x_l)$ – выборка размера l, $X_2 = (x_2, \ldots,\ x_l)$ – она же, но без первого элемента.

Рассмотрим ограничения $H_1 = H_{|X_1}$ функций из класса H на элементы выборки X_1 и $H_2 = H_{|X_2}$ – ограничения функций из класса H на элементы выборки X_2. Заметим, что

$$|H_1| = |\{(h(x_1), h(x_2), \ldots,\ h(x_l)) : h \in H\}|,$$
$$B_H(l) = \max_{(x_1, x_2, \ldots,\ x_l)} |\{(h(x_1), h(x_2), \ldots,\ h(x_l)) : h \in H\}|.$$

Пусть теперь $X_1 = (x_1, x_2, \ldots,\ x_l)$ – выборка, на которой этот максимум достигается. Тогда $|H_1| = B_H(l)$.

При ограничении функций из класса H_1 на выборку X_2 некоторые пары функций, которые различались только на элементе x_1, слились в одну функцию из класса H_2.

Пусть класс функций H_3 состоит из тех функций f класса H_2, для каждой из которых найдется функция $f' \in H$ такая, что $f'(x_1) = -f(x_1)$ и $f'(x_i) = f(x_i)$ при $i = 2,\ \ldots,\ l$. Поэтому $|H_3|$ равно числу этих функций f'.

Отсюда следует, что $|H_1| = |H_2| + |H_3|$, поскольку класс H_2 отличается от класса H_1 тем, что двум индикаторным функциям f и f' из класса H_1, различающимся на объекте x_1 (если такие функции существуют), соответствует только одна функция из класса H_2.

Также заметим, что VC-размерность класса H_2 не превосходит d, поскольку это подкласс класса H_1.

VC-размерность класса H_3 не превосходит $d-1$, поскольку если этот класс функций полностью разделяет некоторое множество объектов мощности d, то существуют расширения этих функций из класса H_1 на элемент x_1, которые полностью разделят это же множество с добавленным к нему элементом x_1 (так как для x_1 и любой функции из класса H_3 найдутся два их расширения из H_1, принимающие противоположные значения на этом элементе). В этом случае класс H_1 также полностью разделяет эту расширенную выборку, а его размерность больше или равна $d+1$. Противоречие.

По предположению индукции

$$|H_2| \leqslant B_{H_2}(l-1) \leqslant h(l-1,d),$$
$$|H_3| \leqslant B_{H_3}(l-1) \leqslant h(l-1,d-1).$$

Поэтому

$$|H_1| = |H_2| + |H_3| \leqslant h(l-1,d) + h(l-1,d-1) = h(l,d).$$

Отсюда в силу выбора X получаем

$$B_H(l) = |H_1| \leqslant h(l,d) = \sum_{i=0}^{d} \binom{l}{i}.$$

Неравенство (1.29) доказано. Оценка

$$B_H(l) \leqslant \sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{el}{d}\right)^d$$

при $l > d$ следует из следующей цепочки неравенств:

$$\sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{l}{d}\right)^d \sum_{i=0}^{d} \binom{l}{i} \left(\frac{d}{l}\right)^i \leqslant$$
$$\leqslant \left(\frac{l}{d}\right)^d \sum_{i=0}^{l} \binom{l}{i} \left(\frac{d}{l}\right)^i =$$
$$= \left(\frac{l}{d}\right)^d \left(1+\frac{d}{l}\right)^l < \left(\frac{l}{d}\right)^d e^d = \left(\frac{el}{d}\right)^d. \qquad (1.30)$$

Теорема доказана. $\triangle$

Мы приведем оценку VC-размерности для класса $\mathcal{L}$ всех линейных классификаторов на $\mathcal{R}^n$, т.е. всех индикаторных функций вида $h(\bar{x}) = \mathrm{sign}(L(\bar{x}))$, где $L(\bar{x})$ – линейная функция. Здесь $\mathrm{sign}(r) = 1$, если $r \geqslant 0$, $\mathrm{sign}(r) = -1$ в противном случае. Иногда классификатором будем называть соответствующую функцию $L(\bar{x})$.

Пусть теперь $\mathcal{X} = \mathcal{R}^n$ – n-мерное евклидово пространство. Линейная функция – это функция вида $L(\bar{x}) = (\bar{a} \cdot \bar{x}) + b$, где $\bar{x} \in \mathcal{R}^n$ – переменная, $\bar{a} \in \mathcal{R}^n$ – вектор весов, b – константа.

Если $b = 0$, то линейный классификатор $\mathrm{sign}(L(\bar{x})) = \mathrm{sign}(\bar{a} \cdot \bar{x})$ называется однородным.

Заметим, что в случае линейных (и однородных) классификаторов выборка разделима тогда и только тогда, когда она строго разделима.

Теорема 1.5. 1) *VC-размерность класса всех линейных функций классификации над $\mathcal{R}^n$ равна $n + 1$.*

2) *VC-размерность класса всех линейных однородных классификаторов над $\mathcal{R}^n$ равна n.*

3) *Для класса всех линейных однородных функций классификации над $\mathcal{R}^n$ выполнено*

$$G_{\mathcal{L}}(l) = \ln H_{\mathcal{L}}(l) = \ln \left(2 \sum_{i=1}^{n-1} \binom{l-1}{i} \right) <$$
$$< (n-1)(\ln(l-1) - \ln(n-1) + 1) + \ln 2 \qquad (1.31)$$

при $l > n$.

Доказательство. Сначала докажем второе утверждение. Набор n векторов $\bar{e}_1 = (1, 0, \ldots, 0), \ldots, \bar{,} e_n = (0, 0, \ldots, 1)$ является полностью строго разделимым, так как для любого подмножества $\bar{e}_{i_1}, \ldots, \bar{e}_{i_k}$ этого набора существует линейный однородный классификатор $h(\bar{x}) = \mathrm{sign}(L(\bar{x}))$, где $L(\bar{x}) = a_1 x_1 + \cdots + a_n x_n$, который отделяет векторы этого подмножества от остальных векторов набора. Определим коэффициенты гиперплоскости $L(\bar{x})$, проходящей через начало координат, следующим образом: $a_{i_j} = 1$ при

$1 \leqslant j \leqslant k$ и $a_i = -1$ для всех остальных i. Тогда $L(\bar{e}_{i_j}) = 1$ при $1 \leqslant j \leqslant k$ и $L(\bar{e}_{i_j}) = -1$ для всех остальных j.

Допустим, что некоторые $n+1$ векторов $\bar{u}_1, \ \ldots, \ \bar{u}_n, \bar{u}_{n+1}$ могут быть полностью строго разделимыми.

Тогда существуют 2^{n+1} весовых векторов $\bar{a}_1, \ \ldots, \ \bar{a}_{2^{n+1}}$ таких, что в матрице Z, образованной числами $z_{i,j} = (\bar{a}_j \cdot \bar{u}_i)$, $i = 1, \ldots, n+1$, $j = 1, \ \ldots, \ 2^{n+1}$,

$$Z = \begin{pmatrix} z_{1,1}, \ \ldots, \ \mathbf{z_{1,j}}, \ \ldots, \ z_{1,2^{n+1}} \\ \ldots \\ z_{i,1}, \ \ldots, \ \mathbf{z_{i,j}}, \ \ldots, \ z_{i,2^{n+1}} \\ \ldots \\ z_{n+1,1}, \ \ldots, \ \mathbf{z_{n+1,j}}, \ \ldots, \ z_{n+1,2^{n+1}} \end{pmatrix}$$

знаки элементов j-го столбца (выделенных черным цветом) соответствуют j-му классификационному классу, поэтому знаки элементов столбцов образуют все 2^{n+1} бинарных последовательностей длины $n+1$.

Векторы $\bar{u}_1, \ \ldots, \ \bar{u}_n, \bar{u}_{n+1}$ расположены в n-мерном пространстве и поэтому линейно зависимы, т.е. для некоторой их нетривиальной ($\lambda_i \neq 0$ хотя для одного i) линейной комбинации

$$\lambda_1 \bar{u}_1 + \cdots + \lambda_n \bar{u}_n + \lambda_{n+1} \bar{u}_{n+1} = 0.$$

Домножаем это равенство на $\bar{a}_j$, $j = 1, \ \ldots, \ 2^{n+1}$, и получаем равенство нулю линейной комбинации с вещественными коэффициентами $\lambda_1, \ \ldots, \ \lambda_{n+1}$ элементов произвольного j-го столбца,

$$\lambda_1 (\bar{a}_j \cdot \bar{u}_1) + \cdots + \lambda_{n+1} (\bar{a}_j \cdot \bar{u}_{n+1}) = 0.$$

Заметим, что разделяющую гиперплоскость $L(\bar{x})$ всегда можно выбрать так что $L(\bar{u}_i) \neq 0$ для всех i.

Среди столбцов имеется хотя бы один, знаки элементов которого совпадают со знаками набора $\lambda_1, \ \ldots, \ \lambda_{n+1}$. Одно из чисел λ_i не равно нулю. Поэтому сумма попарных произведений для одного из столбцов положительна. Полученное противоречие доказывает второе утверждение теоремы.

Докажем первое утверждение теоремы. Заметим, что набор из $n+1$ векторов

$$\bar{e}_0 = (0, 0, \ \ldots, \ 0)', \bar{e}_1 = (1, 0, \ \ldots, \ 0)', \ \ldots, \ \bar{e}_n = (0, 0, \ \ldots, \ 1)'$$

является полностью строго разделимым с помощью линейных классификаторов. Для доказательства этого утверждения, для любого подмножества $\{\bar{e}_{i_1}, \ \ldots, \ \bar{e}_{i_k}\}$ данного набора рассмотрим линейный классификатор $h(\bar{x}) = \mathrm{sign}(L(\bar{x}))$, где

$$L(\bar{x}) = a_1 x_1 + \cdots + a_n x_n + b, \ \bar{x} = (x_1, \ \ldots, \ x_n).$$

Коэффициенты гиперплоскости, отделяющей это подмножество от всех остальных векторов набора определяются следующим образом: $a_{i_j} = 1$ при $1 \leqslant j \leqslant k$ и $a_i = -1$ для всех остальных i, $b = \frac{1}{2}$, если вектор $\bar{e}_0$ входит в подмножество, и $b = -\frac{1}{2}$ в противном случае. Тогда выполнено $L(\bar{e}_{i_j}) > 0$ при $1 \leqslant j \leqslant k$ и $L(\bar{e}_{i_j}) < 0$ для всех остальных j.

Допустим, что существует выборка из $n+2$ векторов n-мерного пространства, полностью строго разделимая линейными классификаторами. Пусть это векторы

$$\bar{x}_1 = (x_{1,1}, \ \ldots, \ x_{1,n})', \ \ldots, \ \bar{x}_{n+2} = (x_{n+2,1}, \ \ldots, \ x_{n+2,n})'.$$

Покажем, что соответствующая выборка, состоящая из $n+2$ векторов

$$\bar{x}'_1 = (x_{1,1}, \ \ldots, \ x_{1,n}, 1)', \ \ldots, \ \bar{x}'_{n+2} = (x_{n+2,1}, \ \ldots, \ x_{n+2,n}, 1)',$$

лежащих в $(n+1)$-мерном пространстве, полностью разделима однородными классификаторами. Рассмотрим произвольное подмножество выборки $\bar{x}'_{i_1}, \ \ldots, \ \bar{x}'_{i_k}$, а также соответствующее подмножество $\bar{x}_{i_1}, \ \ldots, \ \bar{x}_{i_k}$ исходной выборки. Пусть некоторая гиперплоскость

$$L(\bar{x}) = a_1 x_1 + \cdots + a_n x_n + b$$

отделяет подмножество $\bar{x}_{i_1}, \ \ldots, \ \bar{x}_{i_k}$ от остальных векторов исходного набора, т.е. $L(\bar{x}_{i_j}) > 0$ при $j = 1, \ \ldots, \ k$ и $L(\bar{x}_i) < 0$ для остальных i.

Рассмотрим соответствующий линейный однородный классификатор в $n+1$-мерном пространстве

$$L'(\bar{x}) = a_1x_1 + \cdots + a_nx_n + bx_{n+1}.$$

Тогда $L'(\bar{x}'_i) = L(\bar{x}_i)$ при $i = 1, \ \ldots, \ n+2$. Поэтому линейный однородный классификатор $L'(\bar{x}')$ отделяет соответствующее подмножество $\bar{x}'_{i_1}, \ \ldots, \ \bar{x}'_{i_k}$ от всех остальных элементов выборки

$$\bar{x}'_1 = (x_{1,1}, \ \ldots, \ x_{1,n}, 1)', \ \ldots, \ \bar{x}'_{n+2} = (x_{n+2,1}, \ \ldots, \ x_{n+2,n}, 1)'.$$

Таким образом, некоторая выборка $n+1$-мерного пространства, состоящая из $n+2$ векторов, оказалась полностью разделимой однородными классификаторами. Это противоречит второму утверждению теоремы. Данное противоречие доказывает первое утверждение теоремы.

Докажем третье утверждение теоремы. Пусть даны l векторов $\bar{x}_1, \ \ldots, \ \bar{x}_l$. Мы рассматриваем все возможные разбиения этих векторов на два подкласса. Такие разбиения производятся с помощью гиперплоскостей $L(\bar{u}) = (\bar{u} \cdot \bar{x})$, где $\bar{u}$ – весовой вектор, задающий гиперплоскость, а $\bar{x}$ – переменный вектор.

По определению $\mathcal{R}^n(u) = \mathcal{R}^n(x) = \mathcal{R}^n$. Для удобства мы выделяем переменную, пробегающую по этому множеству.

Каждому вектору $\bar{u} \in \mathcal{R}^n(u)$ соответствует гиперплоскость $L(\bar{x}) = (\bar{u} \cdot \bar{x})$ в $\mathcal{R}^n(x)$.

Заметим, что можно рассмотреть двойственный вариант. Вектору $\bar{x} \in \mathcal{R}^n(x)$ соответствует гиперплоскость $L(\bar{u}) = (\bar{x} \cdot \bar{u})$ в $\mathcal{R}^n(u)$, а l векторам $\bar{x}_1, \ \ldots, \ \bar{x}_l$ из $\mathcal{R}^n(x)$ соответствуют l гиперплоскостей $X_1, \ldots, X_l$ в пространстве $\mathcal{R}^n(u)$, проходящих через начало координат.

Пусть $\bar{u} \in \mathcal{R}^n(u)$ – вектор, соответствующий некоторой гиперплоскости $L(\bar{u}) = (\bar{u} \cdot \bar{x})$ в $\mathcal{R}^n(x)$, разделяющей $\bar{x}_1, \ \ldots, \ \bar{x}_l$.

Если непрерывно двигать эту гиперплоскость в $\mathcal{R}^n(x)$, так что разделение векторов $\bar{x}_1, \ \ldots, \ \bar{x}_l$ не нарушается, соответствующий вектор $\bar{u}$ заметает компоненту в пространстве $\mathcal{R}^n(u)$. Компонента – это множество векторов (точек) пространства $\mathcal{R}^n(u)$, ограниченное гиперплоскостями $X_1, \ \ldots, \ X_l$, образованными в $\mathcal{R}^n(u)$ векторами $\bar{x}_1, \ \ldots, \ \bar{x}_l$, которые в данном случае рассматриваются как

весовые. Заметим, что таким образом каждая такая компонента соответствует одному варианту разбиения векторов $\bar{x}_1, \ \ldots, \ \bar{x}_l$ на два класса.

Тогда максимальное число вариантов разбиения векторов $\bar{x}_1, \ \ldots, \ \bar{x}_l$ на два класса гиперплоскостями, проходящими через начало координат в $\mathcal{R}^n(x)$, равно максимальному числу компонент, на которые l гиперплоскостей $X_1, \ \ldots, \ X_l$ делят n-мерное пространство $\mathcal{R}^n(u)$.

Пусть $\Phi(n, l)$ – максимальное число компонент, на которые l гиперплоскостей $X_1, \ldots, X_l$ разделяют n-мерное пространство $\mathcal{R}^n(u)$.

Имеем $\Phi(1, l) = 2$, так как функция $L(x) = ux$ может разделить l точек на прямой только на два класса. Также $\Phi(n, 1) = 2$, так как одна гиперплоскость может разделить точки $\mathcal{R}^n(u)$ только на две компоненты.

Пусть теперь заданы $l - 1$ векторов $\bar{x}_1, \ \ldots, \ \bar{x}_{l-1}$ в пространстве $\mathcal{R}^n(x)$. Им соответствуют $l-1$ гиперплоскостей $X_1, \ \ldots, \ X_{l-1}$ в пространстве $\mathcal{R}^n(u)$, которые разделяют это пространство как максимум на $\Phi(n, l - 1)$ компонент.

Добавим новый вектор $\bar{x}_l$ к ранее рассмотренным векторам $\bar{x}_1, \ \ldots, \ \bar{x}_{l-1}$ в пространстве $\mathcal{R}^n(x)$. Ему соответствует новая гиперплоскость X_l в пространстве $\mathcal{R}^n(u)$.

Если эта гиперплоскость X_l пересекает одну из компонент, она делит ее на две части. Появляется новая компонента.

В то же время эта новая компонента добавляет новое разделение гиперплоскости X_l – новую компоненту внутри гиперплоскости X_l.

Таким образом, число новых компонент, которые образует гиперплоскость X_l, равно числу новых частей, на которые гиперплоскости $X_1, \ \ldots, \ X_{l-1}$ делят гиперплоскость X_l.

Так как размерность X_l равна $n - 1$, число этих делений не превосходит $\Phi(n - 1, l - 1)$.

Отсюда получаем рекуррентное соотношение на максимум числа компонент

$$\Phi(n, l) = \Phi(n, l - 1) + \Phi(n - 1, l - 1) \tag{1.32}$$

с начальными условиями $\Phi(1, l) = 2$, $\Phi(n, 1) = 2$.

Доказать в виде задачи, что рекуррентное соотношение (1.32) имеет решение:

$$\Phi(n,l) = \begin{cases} 2^l \text{ если } l \leqslant n, \\ 2\sum\limits_{i=1}^{n-1} \binom{l-1}{i} \text{ если } l > n. \end{cases}$$

Для получения последнего неравенства из (1.31) (а также из неравенства (1.29)) мы используем оценку $\sum\limits_{i=0}^{n} \binom{l}{i} \leqslant \left(\frac{el}{n}\right)^n$, которая имеет место при любом $n \leqslant l$. Эта оценка следует из цепочки неравенств (1.30). Доказательство теоремы закончено. $\triangle$

Получим верхнюю оценку VC-размерности класса всех многослойных нейронных сетей заданного размера для случая функции активации $\sigma(t) = \mathrm{sign}(t)$.

Пусть $\mathcal{F}$ – некоторый класс индикаторных функций, определенных на $\mathcal{R}^n$. Эти функции могут быть векторнозначными.

Функцию роста класса $\mathcal{F}$ можно записать в виде

$$B_{\mathcal{F}}(m) = \max_{X \subset \mathcal{R}^n, |X|=m} |\mathcal{F}_{|X}|,$$

где $\mathcal{F}_{|X}$ – множество всех функций, которые получаются ограничением функций из класса $\mathcal{F}$ на конечное множество X.

Необходимая оценка будет следовать из следующего утверждения.

Предложение 1.1. *Пусть $\mathcal{F}^1$ и $\mathcal{F}^2$ – два класса функций и $\mathcal{F} = \mathcal{F}^1 \times \mathcal{F}^2$ – их декартово произведение, а $\mathcal{G} = \mathcal{F}^1 \circ \mathcal{F}^2$ – класс функций, которые являются композициями функций из этих классов. Тогда для произвольного m*

1. $B_{\mathcal{F}}(m) \leqslant B_{\mathcal{F}_1}(m) \cdot B_{\mathcal{F}_2}(m)$;
2. $B_{\mathcal{G}}(m) \leqslant B_{\mathcal{F}_1}(m) \cdot B_{\mathcal{F}_2}(m)$

Доказательство. Для доказательства части (1) заметим, что для любого X такого, что $|X| = m$ выполнено

$$|\mathcal{F}_{|X}| \leqslant |\mathcal{F}^1_{|X}| \cdot |\mathcal{F}^2_{|X}| \leqslant B_{\mathcal{F}^1} \cdot B_{\mathcal{F}^2}.$$

Доказательства части (2) предоставляется читателю. $\triangle$

Как было замечено в разделе 1.1.3, нейронная сеть может быть представлена в виде набора векторнозначных функций

$$f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i},$$

где n_i – натуральные числа, $i = 1, \ \ldots, \ l$, $f_i = (f_{i,1}, \ \ldots, \ f_{i,n_{i-1}})$ – набор одномерных функций типа $\mathcal{R}^{n_{i-1}} \to \mathcal{R}$.

Выход нейронной сети задается одномерной функцией – композицией

$$f = f_l \circ f_{l-1} \circ \cdots \circ f_2 \circ f_1.$$

Пусть $\mathcal{F}$ есть класс всех таких функций f, которые вычисляются с помощью нейронной сети, $\mathcal{F}^i$ – класс векторнозначных функций $f_i : \mathcal{R}^{n_{i-1}} \to \mathcal{R}^{n_i}$ и $\mathcal{F}^{i,j}$ – класс функций, которые образуют j-ю компоненту этих композиций.

Заметим также, что функции, ассоциированные с вершинами i-го слоя, являются линейными пороговыми функциями. Поэтому VC-размерность класса $\mathcal{F}^{i,j}$ равна $n_{i-1} + 1$ для каждого j.

По предложению 1.1, а также по лемме Сауэра выполнены следующие неравенства:

$$\begin{aligned} B_{\mathcal{F}}(m) &\leqslant \prod_{i=1}^{l} B_{\mathcal{F}^i}(m) \leqslant \\ &\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} B_{\mathcal{F}^{i,j}}(m) \leqslant \\ &\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} \left(\frac{me}{n_{i-1}+1} \right)^{n_{i-1}+1} = \\ &= \prod_{i=1}^{l} \left(\frac{me}{n_{i-1}+1} \right)^{n_i(n_{i-1}+1)} \leqslant (me)^N, \end{aligned}$$

где

$$N = \sum_{i=1}^{l} n_i(n_{i-1}+1)$$

– общее число параметров нейронной сети.

Оценим теперь VC-размерность класса $\mathcal{F}$. Пусть m – размер максимального по числу элементов множества, которое полностью разделимо функциями из класса $\mathcal{F}$. Тогда $2^m \leqslant (me)^N$. Это неравенство выполнено при $m = O(N \log N)$. Таким образом, VC-размерность класса $\mathcal{F}$ оценивается как $O(N \log N)$.

1.3. Теория обобщения для задач классификации с помощью пороговых решающих правил

В предыдущем разделе показано, что VC-размерность класса всех линейных функций классификации зависит от размерности пространства объектов. На практике длина выборки может быть меньше, чем размерность пространства, поэтому оценки теоремы 1.2 и следствия 1.1, зависящие от VC-размерности, имеют в основном теоретическое значение.

Подобные недостатки VC-размерности проистекают из того, что при разделении выборки с помощью вещественных функций объекты, принадлежащие различным классам, могут быть разделены с как угодно малым порогом. Кроме этого, мы не ограничиваем распределение таких векторов в пространстве.

В этом разделе будем требовать, чтобы объекты из различных классов разделялись функциями с заранее заданным положительным порогом. Мы также ограничим область определения классификационных функций. Будет рассмотрено понятие размерности класса функций, не зависящее от размерности пространства. Все это приведет к оценкам вероятности ошибки обобщения, которые уже могут иметь практическое применение.

1.3.1. Пороговая размерность и ее приложения

Пусть $\mathcal{F}$ – класс вещественных функций с областью определения $\mathcal{X}$. Каждой функции $f \in \mathcal{F}$ сопоставим индикаторную функцию классификации

$$h_f(x) = \begin{cases} 1, \text{ если } f(x) \geqslant 0, \\ -1 \text{ в противном случае.} \end{cases}$$

Пусть $S = ((x_1, y_1),\ \ldots,\ (x_l, y_l))$ – выборка длины l, где $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$.

Границей ошибки при классификации примера (x_i, y_i) с помощью вещественной функции f называется величина $\gamma_i = y_i f(x_i)$. Заметим, что $\gamma_i > 0$ означает, что классификация с помощью f является правильной: $h_f(x_i) = y_i$. Кроме этого, будем рассматривать величину

$$m_S(f) = \min_{i=1,\ldots,\ l} \gamma_i$$

– границу ошибки классификации с помощью функции f на выборке S. По определению $m_S(f) > 0$ тогда и только тогда, когда функция f классифицирует S без ошибок и с положительным порогом.

Пусть $\epsilon > 0$. Назовем *ϵ-покрытием* множества функций $\mathcal{F}$ относительно множества $X = \{x_1,\ \ldots,\ x_l\}$ любое конечное множество функций $\mathcal{B}$ такое, что для любой функции $f \in \mathcal{F}$ существует функция $g \in \mathcal{B}$ такая, что $|f(x_i) - g(x_i)| < \epsilon$ при $i = 1,\ \ldots,\ l$.

Пусть $\mathcal{N}(\epsilon, \mathcal{F}, X)$ – размер наименьшего по мощности покрытия $\mathcal{F}$ относительно множества X (минимальное покрытие).

Супремум этих величин по всем множествам X мощности l

$$\mathcal{N}(\epsilon, \mathcal{F}, l) = \sup_{|X|=l} \mathcal{N}(\epsilon, \mathcal{F}, X)$$

называется *числом покрытия* класса $\mathcal{F}$.

По определению

$$\mathrm{err}_S(f) = \frac{1}{l} |\{(x_i, y_i) : h_f(x_i) \neq y_i\}|$$

– доля ошибок классификации с помощью функции f на выборке $S = ((x_1, y_1),\ \ldots,\ (x_l, y_l))$.

Пусть P – распределение вероятностей на $\mathcal{X} \times \{-1, 1\}$, генерирующее элементы выборки S. Рассмотрим вероятность ошибочной классификации с помощью функции f :

$$\mathrm{err}_P(f) = P\{h_f(x) \neq y\}.$$

Имеет место теорема – аналог теоремы 1.2.

Теорема 1.6. *Для произвольных* $\epsilon > 0$ *и* $\gamma > 0$

$$P^l\{S : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_P(f) > \epsilon)\} \leqslant \\ \leqslant 2\mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}$$

при $l > 2/\epsilon$.

Доказательство теоремы 1.6 аналогично доказательству теоремы 1.2. Надо только к равенству $\mathrm{err}_S(f) = 0$ добавить более сильное условие $m_S(f) \geqslant \gamma$ (в правой части условия (1.13) леммы 1.4). Аналогичная лемма утверждает, что

Лемма 1.6. *При* $l > 2/\epsilon$

$$P^l\{S : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_P(f) > \epsilon)\} \leqslant \\ \leqslant 2P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_{\hat{S}}(f) > \frac{\epsilon}{2})\}.$$

Доказательство этой леммы почти полностью повторяет доказательство леммы 1.2.

Вторая лемма аналогична лемме 1.3:

Лемма 1.7. *При* $l > 2/\epsilon$

$$P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_{\hat{S}}(f) > \frac{\epsilon}{2})\} \leqslant \\ \leqslant \mathcal{N}(\gamma/2, \mathcal{F}, 2l)e^{-\epsilon l/4}.$$

Для доказательства леммы рассмотрим $\gamma/2$-покрытие $\mathcal{B}$ множества $\mathcal{F}$ относительно двойной выборки $S\hat{S}$. Пусть $g \in \mathcal{B}$ приближает функцию $f \in \mathcal{F}$ с точностью до $\gamma/2$. Если $m_S(f) \geqslant \gamma$, то $m_S(g) > \gamma/2$. Кроме этого, если $\mathrm{err}_S(f) = 0$ и $m_S(f) \geqslant \gamma$, то $\mathrm{err}_S(g) = 0$.

Если функция f ошибочно классифицирует объект x_i, т.е. $y_i f(x_i) \leqslant 0$, то $y_i g(x_i) < \gamma/2$. Пусть $\mathrm{err}_{\hat{S}}(\gamma/2, g)$ обозначает долю тех i, для которых $y_i g(x_i) < \gamma/2$, где x_i находится во второй части $\hat{S}$ двойной выборки. Отсюда следует неравенство

$$P^{2l}\{S\hat{S} : (\exists\ f \in \mathcal{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geqslant \gamma \& \mathrm{err}_{\hat{S}}(f) > \frac{\epsilon}{2})\} \leqslant \\ \leqslant P^{2l}\{S\hat{S} : (\exists\ g \in \mathcal{B})(\mathrm{err}_S(g) = 0 \& m_S(g) \geqslant \frac{\gamma}{2} \& \mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2})\}.$$

Далее рассуждения аналогичны комбинаторной части доказательства леммы 1.5. Здесь мы оцениваем долю вариантов, при которых некоторая функция $g \in \mathcal{B}$ разделяет первую часть выборки S без ошибок: $\mathrm{err}_S(g) = 0$, и более того, с порогом $m_S(g) \geqslant \gamma/2$, а на второй половине выборки либо делает ошибки, либо имеет порог разделения не выше $\gamma/2$ в доле $\mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2}$ примеров. Оценка такая же, как и в лемме 1.3.

В результате получаем оценку

$$P^{2l}\{S\hat{S} : (\exists\, g \in \mathcal{B})(\mathrm{err}_S(g) = 0 \& m_S(g) \geqslant \frac{\gamma}{2} \& \mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2})\} \leqslant \\ \leqslant \mathcal{N}(\gamma/2, \mathcal{F}, 2l) e^{-\epsilon l/4}.$$

Из оценок лемм 1.6 и 1.7 получаем оценку теоремы 1.6. $\triangle$

Обозначим $\delta = 2\mathcal{N}(\gamma/2, \mathcal{F}, 2l)2^{-\epsilon l/4}$. Из теоремы 1.6 и неравенства $e^{-\epsilon l/4} < 2^{-\epsilon l/4}$ получаем

Следствие 1.3. *Заданы класс $\mathcal{F}$ вещественных функций и число $\gamma > 0$. Тогда для любого распределения вероятностей P на $\mathcal{X} \times \{-1, 1\}$, любых l и $0 < \delta < 1$ с вероятностью $1 - \delta$ произвольная функция $f \in \mathcal{F}$, которая классифицирует S с границей ошибки $m_S(f) > \gamma$, имеет ошибку классификации*

$$\mathrm{err}_P(f) \leqslant \frac{4}{l}\left(\log \mathcal{N}(\gamma/2, \mathcal{F}, 2l) + \log \frac{2}{\delta}\right).$$

Пороговая размерность. Каждый класс функций $\mathcal{F}$ порождает так называемую пороговую размерность, или fat-размерность (fat-shattering dimension). Пусть $\gamma > 0$. Множество $X = \{x_1, \ldots, x_l\}$ называется *γ-разделимым*, если существуют вещественные числа $r_1, \ldots, r_l$ (свидетели) такие, что для любого подмножества $E \subseteq X$ существует функция $f_E \in \mathcal{F}$ такая, что $f_E(x_i) \geqslant r_i + \gamma$, если $x_i \in E$, и $f_E(x_i) \leqslant r_i - \gamma$, если $\bar{x}_i \notin E$.

По определению *пороговая размерность* $\mathrm{fat}_\gamma(\mathcal{F})$ класса $\mathcal{F}$ равна размеру самого большого по количеству элементов γ-разделимого множества X. По определению пороговая размерность класса $\mathcal{F}$ – это невозрастающая функция от $\gamma > 0$. Класс $\mathcal{F}$ имеет бесконечную пороговую размерность, если существуют как угодно большие γ-разделимые выборки.

Следующая теорема является прямым следствием теоремы 1.10, которая будет доказана в разделе 1.3.2.

Теорема 1.7. *Пусть $\mathcal{F}$ – класс функций типа $\mathcal{X} \to [a, b]$, где $a < b$. Выберем $0 < \gamma < 1$ и обозначим $d = \mathrm{fat}_{\gamma/4}(\mathcal{F})$. Тогда*

$$\log \mathcal{N}(\gamma, \mathcal{F}, l) \leqslant 1 + d \log \frac{2el(b-a)}{d\gamma} \log \frac{4l(b-a)^2}{\gamma^2}.$$

Теорема 1.7 вместе со следствием 1.3 влечет такое следствие.

Следствие 1.4. *Пусть $\mathcal{F}$ – класс вещественных функций со значениями в отрезке $[-1, 1]$, $\gamma > 0$, $\delta > 0$ и P – распределение вероятностей, генерирующее выборку S. Тогда с вероятностью $1 - \delta$ любая гипотеза $f \in \mathcal{F}$, для которой $m_S(f) \geqslant \gamma$, имеет ошибку классификации*

$$\mathrm{err}_P(f) \leqslant \frac{4}{l} \left(d \log \frac{16el}{d\gamma} \log \frac{128l}{\gamma^2} + \log \frac{2}{\delta} \right),$$

где $d = \mathrm{fat}_{\gamma/8}(\mathcal{F})$.

Для класса всех (однородных) линейных функций с ограниченной областью определения имеет место не зависящая от размерности пространства объектов верхняя оценка пороговой размерности.

Теорема 1.8. *Пусть $X = \{\bar{x} : \|\bar{x}\| \leqslant R\}$ – шар радиуса R в n-мерном евклидовом пространстве: и $\mathcal{F}$ – класс линейных однородных функций $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, где $\|\bar{w}\| \leqslant 1$ и $\bar{x} \in X$. Тогда*

$$\mathrm{fat}_\gamma(\mathcal{F}) \leqslant \left(\frac{R}{\gamma} \right)^2.$$

Доказательство. Допустим, что множество $Y = \{\bar{x}_1, \ \ldots, \ \bar{x}_l\}$ является γ-разделимым с помощью линейных однородных функций из класса $\mathcal{F}$ и свидетелями $r_1, \ldots, r_l$.

Рассмотрим произвольное подмножество $\hat{Y} \subseteq Y$. Допустим, что

$$\sum_{\bar{x}_i \in \hat{Y}} r_i \geqslant \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i.$$

По определению найдется весовой вектор $\bar{w}$, $\|\bar{w}\| \leqslant 1$, такой, что $(\bar{w} \cdot \bar{x}_i) > r_i + \gamma$ для $\bar{x}_i \in \hat{Y}$ и $(\bar{w} \cdot \bar{x}_i) \leqslant r_i - \gamma$ для $\bar{x}_i \notin \hat{Y}$. Тогда

$$\sum_{\bar{x}_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) \geqslant \sum_{\bar{x}_i \in \hat{Y}} r_i + |\hat{Y}|\gamma$$

и

$$\sum_{\bar{x}_i \in Y \setminus \hat{Y}} (\bar{w} \cdot \bar{x}_i) \leqslant \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i - |Y \setminus \hat{Y}|\gamma.$$

Разность этих двух сумм оценивается снизу

$$\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \in Y \setminus \hat{Y}} (\bar{w} \cdot \bar{x}_i) \geqslant \gamma l. \tag{1.33}$$

Из неравенства Коши–Буняковского для евклидовой нормы получаем

$$\begin{aligned} &\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \in Y \setminus \hat{Y}} (\bar{w} \cdot \bar{x}_i) = \\ &= \left(\bar{w} \cdot \left(\sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i \right) \right) \leqslant \\ &\leqslant \left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i \right\| \cdot \|\bar{w}\|. \end{aligned} \tag{1.34}$$

Из (1.33), (1.34) и из неравенства $|\bar{w} \leqslant 1$ получаем нижнюю оценку

$$\left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i \right\| \geqslant \gamma l. \tag{1.35}$$

Допустим теперь, что

$$\sum_{\bar{x}_i \in \hat{Y}} r_i < \sum_{\bar{x}_i \in Y \setminus \hat{Y}} r_i.$$

Поменяем множества $\hat{Y}$ и $Y \setminus \hat{Y}$ местами и аналогичным образом получим

$$\left\| \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i - \sum_{x_i \in \hat{Y}} \bar{x}_i \right\| \geqslant \gamma l.$$

Таким образом (1.35) выполнено в обоих случаях.

Продолжим доказательство теоремы.

Пусть $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ – случайный равномерно распределенный бинарный вектор длины l; $\xi_i \in \{-1, 1\}$ при $i = 1 \ , \ldots, \ l$.

Вектор $\bar{\xi}$ естественным образом определяет разбиение множества Y на два подмножества $\hat{Y}$ и $Y \setminus \hat{Y}$. Вычислим математическое ожидание квадрата нормы разности (1.35) относительно случайного разбиения множества Y, определяемого вектором $\bar{\xi}$. Имеем

$$E \left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i \right\|^2 = E \left\| \sum_{i=1}^{l} \xi_i \bar{x}_i \right\|^2 =$$

$$= E \sum_{i=1}^{l} \xi_i^2 \|\bar{x}_i\|^2 + 2E \sum_{i,j=1, i \neq j}^{l} \xi_i \xi_j (\bar{x}_i \cdot \bar{x}_j) =$$

$$= E \sum_{i=1}^{l} \|\bar{x}_i\|^2 \leqslant R^2 l.$$

Найдется хотя бы одно подмножество $\hat{Y}$, для которого значение нормы разности не больше ее среднего значения:

$$\left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \in Y \setminus \hat{Y}} \bar{x}_i \right\| \leqslant R\sqrt{l}.$$

Вместе с неравенством (1.35) это влечет $R\sqrt{l} \geqslant \gamma l$. Отсюда получаем $l \leqslant (R/\gamma)^2$. Это означает, что $\mathrm{fat}_\gamma(\mathcal{F}) \leqslant (R/\gamma)^2$. $\triangle$

Теорема 1.8 доказана для класса однородных функций. Это не является существенным ограничением, так как мы можем увеличить размерность векторов на единицу и рассмотреть выборку

$S' = ((\bar{x}'_1, y_1),\ \ldots,\ (\bar{x}'_l, y_l))$, где $\bar{x}'_i = (\bar{x}_i, 1)$ при $1 \leqslant i \leqslant l$. В этом случае однородная функция $f'(\bar{x}') = (\bar{w}' \cdot \bar{x}')$, где $\bar{w}' = (\bar{w}, b)$, разделяет выборку S' с тем же порогом, с которым функция $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$ разделяет выборку S.

Подставляем оценку теоремы 1.8 в оценку следствия 1.4 и получаем следующую итоговую теорему.

Теорема 1.9. *Рассмотрим задачу классификации с помощью линейных однородных функций $f(\bar{x}) = (\bar{w} \cdot \bar{x}) \in \mathcal{L}$, где $\bar{x} \in \mathcal{R}^n$, $\|\bar{w}\| \leqslant 1$. Заданы числа $\gamma > 0$ и $0 < \delta < 1$.*

Для произвольного распределения вероятностей P, сконцентрированного в шаре радиуса R с центром в начале координат и генерирующего выборку $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$, с вероятностью $1 - \delta$ произвольная гипотеза $f \in \mathcal{L}$ с границей ошибки $m_S(f) \geqslant \gamma$ имеет верхнюю оценку ошибки классификации

$$\mathrm{err}_P(f) \leqslant \frac{4}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{4R}\log\frac{128Rl}{\gamma^2} + \log\frac{2}{\delta}\right). \tag{1.36}$$

Для получения этой оценки мы использовали то, что в неравенстве (1.45) теоремы 1.8 вместо d можно взять любую верхнюю оценку числа $d' = Sdim(\mathcal{F}^{\alpha/2})$. В данном случае удобно взять $d = \frac{64R^2}{\gamma^2}$.

Теорема 1.9 также будет иметь место для функций из произвольного гильбертового пространства, порожденного воспроизводящим ядром. Такие пространства будут введены в разделе 2.5.

Оценки теорем 1.8 и 1.9 послужат основой для получения не зависящих от размерности пространства оценок вероятности ошибки обобщения для метода опорных векторов в теореме 2.5 из раздела 2.6.1.

1.3.2. Покрытия и упаковки

Рассмотрим содержание предыдущего раздела с более общих позиций.

Пусть $(\mathcal{X}, d)$ – некоторое метрическое пространство, $d(x, y)$ – расстояние между элементами $x, y \in \mathcal{X}$.

Пусть $A \subseteq \mathcal{X}$ и $B \subseteq A$ и $\alpha > 0$. Говорим, что множество B является *α-покрытием* множества A, если для любого $a \in A$ существует $b \in B$ такое, что $d(a, b) < \alpha$. *Числом покрытия* множества A называется функция

$$\mathcal{N}_d(\alpha, A) = \min\{|B| : B \text{ является } \alpha\text{-покрытием } A\}. \qquad (1.37)$$

Говорим, что множество $B \subseteq \mathcal{X}$ является *α-отделимым*, если для любых $a, b \in B$ таких, что $a \neq b$, будет $d(a, b) \geqslant \alpha$. *Числом упаковки* множества A называется функция

$$\mathcal{M}_d(\alpha, A) = \max\{|B| : B \subseteq A \text{ является } \alpha\text{-отделимым}\}. \qquad (1.38)$$

Основные соотношения между числом покрытия и числом упаковки даются в следующей лемме.

Лемма 1.8. *Для любых $A \subseteq \mathcal{X}$ и $\alpha > 0$*

$$\mathcal{M}_d(2\alpha, A) \leqslant \mathcal{N}_d(\alpha, A) \leqslant \mathcal{M}_d(\alpha, A).$$

Доказательство. Пусть M – 2α-отделимое подмножество A и N – α-покрытие A. По определению множества N для каждого $a \in M$ найдется $b \in N$ такое, что $d(a, b) < \alpha$. Если $a, a' \in M$ различные и $b, b' \in N$ им таким образом соответствуют, то b и b' также различные, так как иначе было бы $b = b'$ и $d(a, a') \leqslant d(a, b) + d(b, a') < 2\alpha$. Это противоречит тому, что любые два различные элемента M находятся на расстоянии не меньшем 2α. Отсюда заключаем, что $|M| \leqslant |N|$. Первое неравенство доказано.

Пусть M – максимальное по включению α-отделимое подмножество A. Докажем, что M является α-покрытием множества A. Допустим, это не так. Тогда найдется элемент $x \in A$ такой, что нет ни одного элемента из M в окрестности x радиуса α. Добавим x к M и получим строго большее по включению подмножество $M \cup \{x\}$ множества A, которое также α-отделимо. Получаем противоречие с выбором M. Данное противоречие доказывает второе неравенство из условия леммы. $\triangle$

Основная цель данного раздела – доказательство теоремы 1.7, которая будет следствием теоремы 1.10 (см. ниже). Для его проведения нам потребуется развитие теории размерности для функций с конечным числом значений.

Пусть $\mathcal{X}$ – некоторое множество и $B = \{0, 1, \ldots, b\}$ – конечное множество. Рассмотрим некоторый класс $\mathcal{F} \subseteq B^{\mathcal{X}}$ функций, определенных на множестве $\mathcal{X}$ и принимающих конечное число значений из множества B. Рассмотрим метрику на $\mathcal{F}$

$$l(f, g) = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

Две функции $f, g \in \mathcal{F}$ *отделены* (2-*отделены*), если $l(f, g) \geqslant 2$. Иными словами, существует $x \in \mathcal{X}$ такое, что $|f(x) - g(x)| \geqslant 2$. Класс $\mathcal{F}$ *отделим*, если любые две функции $f, g \in \mathcal{F}$ отделены.

Пусть $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ – некоторое множество с заданным линейным порядком на его элементах (выборка) и $\mathcal{F} \subseteq B^{\mathcal{X}}$. По определению класс $\mathcal{F}$ *строго разделяет* множество X, если существует набор $s = (s_1, \ldots, s_n\}$ элементов B такой, что для любого $E \subseteq X$ существует функция $f_E \in \mathcal{F}$ такая, что

$$x_i \in E \Longrightarrow f_E(x_i) \geqslant s_i + 1,$$
$$x_i \notin E \Longrightarrow f_E(x_i) \leqslant s_i - 1$$

для любого i.

Говорят также, что $\mathcal{F}$ строго разделяет множество X относительно набора s. Размер максимального множества X, строго разделимого с помощью класса функций $\mathcal{F}$, называется *строгой размерностью* $\mathcal{F}$ и обозначается $Sdim(\mathcal{F})$.

Рассмотрим простую дискретизацию, переводящую произвольную вещественнозначную функцию $f : \mathcal{X} \to [0, 1]$ в функцию, принимающую конечное число значений. Для произвольного вещественного $\alpha > 0$ определим

$$f^{\alpha}(x) = \left[\frac{f(x)}{\alpha}\right]$$

для всех x, где $[r]$ обозначает ближайшее к r, сверху или снизу, целое число; таким образом, $|r - [r]| \leqslant 0.5$. Если число r находится строго в середине интервала между целыми числами, то можно округлять с недостатком. Определим также $\mathcal{F}^{\alpha} = \{f^{\alpha} : f \in \mathcal{F}\}$.

Очевидно, что функция f^{α} принимает значения в множестве $\{0, 1, \ldots, [1/\alpha]\}$.

Далее, рассмотрим связь между комбинаторными размерностями и числами покрытия классов функций $\mathcal{F}$ и $\mathcal{F}^\alpha$.

Напомним, что число покрытия $\mathcal{N}_d(\alpha, A)$ и число упаковки $\mathcal{M}_d(\alpha, A)$ определены согласно (1.37) и (1.38).

Определим специальную метрику на $\mathcal{F}$, связанную с множеством $X = \{x_1, \ldots, x_n\}$, следующим образом:

$$l_X(f, g) = \max_{1 \leqslant i \leqslant n} |f(x_i) - g(x_i)|.$$

Рассмотрим соответствующие числа покрытия и упаковки:

$$\mathcal{N}(\alpha, \mathcal{F}, X) = \mathcal{N}_{l_X}(\alpha, \mathcal{F}),$$
$$\mathcal{M}(\alpha, \mathcal{F}, X) = \mathcal{M}_{l_X}(\alpha, \mathcal{F}),$$

а также соответствующие величины

$$\mathcal{N}(\alpha, \mathcal{F}, n) = \sup_{|X|=n} \mathcal{N}(\alpha, \mathcal{F}, X),$$
$$\mathcal{M}(\alpha, \mathcal{F}, n) = \sup_{|X|=n} \mathcal{M}(\alpha, \mathcal{F}, X).$$

Связь между строгой размерностью дискретизированного класса функций и пороговой размерностью исходного класса представлена в следующей лемме.

Лемма 1.9. *Пусть $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ и $\alpha > 0$. Тогда*

$$Sdim(\mathcal{F}^\alpha) \leqslant fat_{\alpha/2}(\mathcal{F}), \tag{1.39}$$

$$\mathcal{M}(\alpha, \mathcal{F}, X) \leqslant \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X). \tag{1.40}$$

Доказательство леммы предлагается в качестве задачи.

Следующая лемма представляет техническую часть основного результата этого раздела.

Лемма 1.10. *Пусть $\mathcal{X}$ – конечное множество, $|\mathcal{X}| = n$ и $B = \{0, 1, \ldots, b\}$. Пусть также $\mathcal{F} \subseteq B^{\mathcal{X}}$ и $d = Sdim(\mathcal{F})$. Тогда* [6]

$$\mathcal{M}(2, \mathcal{F}, n) \leqslant 2(n(b+1)^2)^{\lceil \log y \rceil},$$

где $y = \sum\limits_{i=1}^{d} \binom{n}{i} b^i$.

[6] Здесь $\lceil r \rceil$ обозначает наименьшее целое большее или равное r.

Доказательство. Определим функцию $t(h,n)$, где h и n – целые числа. Предполагаем, что $n \geqslant 1$ и $h \geqslant 2$. Значение $t(h,n)$ этой функции определяется следующим образом.

Рассматриваются все отделимые подклассы F класса функций $\mathcal{F}$, содержащие по h элементов. Понятие отделимости класса функций было определено ранее в этом разделе. Каждый такой класс функций F может строго разделять некоторые множества $X \subseteq \mathcal{X}$ относительно некоторых последовательностей целых чисел s. Когда мы говорим, что F строго разделяет пару (X,s), мы имеем в виду, что F строго разделяет X относительно последовательности s.

Функцию $t(h,n)$ зададим формальным условием:

$$\begin{aligned} t(h,n) = \max\{k : \forall F \subseteq \mathcal{F}, |F| = h, F \text{ отделимо} \\ \Rightarrow F \text{ строго разделяет не менее } k \text{ пар } (X,s)\}. \end{aligned}$$

Если не существует ни одного отделимого множества F размера h, то полагаем $t(h,n) = \infty$.

Лемма 1.11. *Если $t(h,n) > y$ и $Sdim(\mathcal{F}) \leqslant d$, то $\mathcal{M}_l(2,\mathcal{F}) < h$, где $y = \sum\limits_{i=0}^{d} \binom{n}{i} b^i$.*

Доказательство. Допустим, что $\mathcal{M}_l(2,\mathcal{F}) \geqslant h$. Это значит, что существует отделимое множество $F \subseteq \mathcal{F}$ размера не менее h. Так как $t(h,n) \geqslant y$, F строго разделяет по-крайней мере y пар (X,s).

Так как $Sdim(\mathcal{F}) \leqslant d$, если F строго разделяет пару (X,s), то $|X| \leqslant d$. Подмножество X размера i можно выбрать $\binom{n}{i}$ способами, кроме того, имеется менее b^i возможных последовательностей s длины i (из-за строгой разделимости X элементами s не могут быть 0 или b). Таким образом, F строго разделяет менее чем

$$\sum_{i=1}^{d} \binom{n}{i} b^i = y$$

(X,s)-пар. Полученное противоречие доказывает лемму. $\triangle$

Из леммы 1.11 следует, что для того, чтобы доказать лемму 1.10, достаточно доказать, что

$$t\left(2(n(b+1)^2)^{\lceil \log y \rceil}, n\right) \geqslant y, \tag{1.41}$$

где $y = \sum\limits_{i=1}^{d} \binom{n}{i} b^i$.

Для того, чтобы доказать неравенство (1.41) предварительно докажем следующее утверждение.

Лемма 1.12. *Для функции t выполняются неравенства*

$$t(2,n) \geqslant 1 \text{ при } n \geqslant 1, \tag{1.42}$$

$$t(2mn(b+1)^2, n) \geqslant 2t(2m, n-1) \text{ при } n \geqslant 2, m \geqslant 1. \tag{1.43}$$

Доказательство. Для любых двух отделимых функций f и g, $|f(x)-g(x)| \geqslant 2$ хотя бы для одного x, т.е. эти функции разделяют одноэлементное множество $\{x\}$. Таким образом, $t(2,n) \geqslant 1$, т.е. неравенство (1.42) выполнено.

Для доказательства (1.43) рассмотрим множество F, содержащее по-крайней мере $2mn(b+1)^2$ попарно отделимых функций. Если такого множества не существует, то $t(2mn(b+1)^2, n) = \infty$ и неравенство (1.43) автоматически выполнено. Разделим произвольным образом функции из F на пары $\{f,g\}$. Всего таких пар не менее чем $mn(b+1)^2$. Пусть S обозначает это множество пар. Для произвольной пары $\{f,g\} \in S$ пусть $\chi(f,g)$ равно одному из тех x, для которых $|f(x)-g(x)| \geqslant 2$.

Для $x \in \mathcal{X}$, $i,j \in B$, $j \geqslant i+2$, определим

$$bin(x,i,j) = \{\{f,g\} \in S : \chi(f,g) = x, \{f(x), g(x)\} = \{i,j\}\}.$$

Общее число таких множеств не превосходит

$$n\binom{b+1}{2} < n(b+1)^2/2.$$

Напомним, что по условию леммы 1.10 выполнено $|\mathcal{X}| = n$.

Так как число всех пар (f,g) равно $mn(b+1)^2$, должны существовать x^*, i^* и $j^* > i^*+1$ такие, что

$$|bin(x^*, i^*, j^*)| \geqslant 2m.$$

Определим два множества функций

$$F_1 = \{f \in \cup bin(x^*, i^*, j^*) : f(x^*) = i^*\},$$
$$F_2 = \{g \in \cup bin(x^*, i^*, j^*) : g(x^*) = j^*\}.$$

Здесь $\cup A$, где множество A состоит из пар, есть множество, состоящее из элементов всех таких пар.

Ясно, что $|F_1| = |F_2| \geqslant 2m$. Класс функций F_1 отделим, если рассматривать эти функции на множестве $\mathcal{X} \setminus \{x^*\}$. Действительно, класс F, а значит, и класс F_1, отделим на $\mathcal{X}$, поэтому для любых двух функций $f, f' \in F_1$ будет $|f(x') - f'(x')| \geqslant 2$ для некоторого x', причем $x' \in \mathcal{X} \setminus \{x^*\}$, так как на x^* значения этих функций совпадают. Аналогичным образом, класс функций F_2 также отделим на области определения $\mathcal{X} \setminus \{x^*\}$.

Следовательно, должны существовать два множества U и V, каждое размера не менее $t(2m, n-1)$, состоящие из пар (X, s) такие, что F_1 строго разделяет пары из U и F_2 строго разделяет пары из V.

Очевидно, что любая пара из $U \cup V$ строго разделяется классом F.

Пусть $(X, s) \in U \cap V$. Тогда пара $(\{x^*\} \cup X, (\lfloor \frac{i^*+j^*}{2} \rfloor, s))$ также строго разделяется посредством F. Это так, поскольку любые функции $f \in F_1$ и $g \in F_2$, строго разделяющие X, удовлетворяют также условиям $f(x^*) = i^*$ и $g(x^*) = j^*$, причем $j^* \geqslant i^* + 2$. Поэтому $g(x^*) = j^* \geqslant \frac{i^*+j^*}{2} + 1$ и $f(x^*) = i^* \leqslant \frac{i^*+j^*}{2} - 1$. Поэтому одна из этих функций строго разделяет выбранное подмножество.

Действительно, пусть $E \subseteq \{x^*\} \cup X = \{x^*, x_1, \ldots, x_{|X|}\}$ и $f(x_i) \geqslant s_i + 1$ при $x_i \in E$, $f(x_i) \leqslant s_i - 1$ при $x_i \notin E$ для некоторого набора $s = (s_1, \ldots, s_{n-1})$. Аналогично, пусть $g(x_i) \geqslant s'_i + 1$ при $x_i \in E$, $g(x_i) \leqslant s'_i - 1$ при $x_i \notin E$ для некоторого набора $s' = (s'_1, \ldots, s'_{n-1})$. При этом $f \in F_1$ и $g \in F_2$. Тогда f разделяет E относительно набора $s_1 = (\lfloor \frac{i^*+j^*}{2} \rfloor, s_1, \ldots, s_{n-1})$, если $x^* \notin E$, или g разделяет E относительно набора $s_2 = (\lfloor \frac{i^*+j^*}{2} \rfloor, s'_1, \ldots, s'_{n-1})$, если $x^* \in E$.

Следовательно, класс F строго разделяет

$$|U \cup V| + |U \cap V| = |U| + |V| \geqslant 2t(2m, n-1)$$

пар (X, s). Неравенство (1.43) и лемма 1.12 доказаны. $\triangle$

Переходим теперь к доказательству леммы 1.10. Применяя неравенства (1.42) и (1.43) рекурсивным образом, получим

$$t(2(n(b+1)^2)^r, n) \geqslant 2^r t(2, n-r) \geqslant 2^r \tag{1.44}$$

при $n > r \geqslant 1$.

Если $\lceil \log y \rceil < n$, то полагаем $r = \lceil \log y \rceil$ в (1.44) и получаем неравенство (1.10).

Если $\lceil \log y \rceil \geqslant n$, то величина

$$2(n(b+1)^2)^{\lceil \log y \rceil} > (b+1)^n$$

превышает число всех функций со значениями в B и областью определения $\mathcal{X}$, $|\mathcal{X}| = n$. В этом случае ни одного отделимого множества F размера $2(n(b+1)^2)^{\lceil \log y \rceil}$ не существует и

$$t(2(n(b+1)^2)^{\lceil \log y \rceil}, n) = \infty.$$

Таким образом, лемма (1.10) доказана. $\triangle$

Теперь мы можем сформулировать и доказать основное утверждение этого раздела – теорему Алона, Бен-Давида, Чеза-Бьянки и Хауслера [8].

Теорема 1.10. *Пусть* $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ *и* $\alpha \in [0,1]$. *Обозначим* $d = \mathrm{fat}_{\alpha/4}(\mathcal{F})$. *Тогда*

$$\mathcal{N}(\alpha, \mathcal{F}, n) \leqslant 2\left(n\left(\frac{2}{\alpha}+1\right)^2\right)^{\lceil d \log(\frac{2en}{d\alpha})\rceil}.$$

Доказательство. Используя то, что число упаковки не превосходит числа покрытия, неравенство (1.40) леммы 1.9, а также лемму 1.10, получим следующую цепочку неравенств

$$\begin{aligned}\mathcal{N}(\alpha, \mathcal{F}, n) = \sup_{|X|=n} \mathcal{N}(\alpha, \mathcal{F}, X) \leqslant \\ \leqslant \sup_{|X|=n} \mathcal{M}(\alpha, \mathcal{F}, X) \leqslant \\ \leqslant \sup_{|X|=n} \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X) = \mathcal{M}(2, \mathcal{F}^{\alpha/2}, n) \leqslant \\ \leqslant 2(n(b+1)^2)^{\lceil \log y \rceil},\end{aligned}$$

где $b = [\frac{1}{\alpha}]$, $y = \sum\limits_{i=1}^{d'} \binom{n}{i} b^i$, $d' = Sdim(\mathcal{F}^{\alpha/2})$.

Заметим, что класс функций $\mathcal{F}^{\alpha/2}$ удовлетворяет условию леммы 1.10 при $b = [\frac{2}{\alpha}]$.

Из неравенства (1.39) леммы 1.9 получаем $d' \leqslant fat_{\alpha/4}(\mathcal{F}) = d$. Отсюда

$$y = \sum_{i=1}^{d'} \binom{n}{i} b^i \leqslant \sum_{i=1}^{d} \binom{n}{i} b^i \leqslant b^d \sum_{i=1}^{d} \binom{n}{i} \leqslant b^d \left(\frac{en}{d}\right)^d. \qquad (1.45)$$

В частности, $\log y \leqslant d \log(\frac{ben}{d})$.

Теорема доказана. $\triangle$

Теорема 1.7 из раздела 1.3 является переформулировкой этой теоремы с небольшим ослаблением оценки.

1.4. Средние по Радемахеру

В этом разделе будет изучаться еще одно понятие емкости класса функций – средние Радемахера. Это понятие позволяет получать новые верхние оценки ошибки обобщения.[7]

Пусть $\mathcal{X}$ – некоторое множество, на котором задана структура вероятностного пространства с вероятностным распределением P. Пусть также задан некоторый класс $\mathcal{F}$ измеримых функций, определенных на множестве $\mathcal{X}$.

Пусть $z^l = (z_1, \ldots, z_l)$ – некоторая выборка, состоящая из элементов множества $\mathcal{X}$. Мы предположим, что элементы выборки независимо и одинаково распределены согласно распределению P.

Пусть $\sigma_1, \ldots, \sigma_l$ – независимые бернуллиевские величины, принимающие два значения $+1$ и -1 с равными вероятностями: $B_{1/2}(\sigma_i = 1) = B_{1/2}(\sigma_i = -1) = 1/2$ для любого $1 \leqslant i \leqslant l$. Эти случайные величины называются случайными Радемахера.

Обозначим посредством $\sigma = B_{1/2}^l$ распределение всего набора $\sigma_1, \ldots, \sigma_l$ длины l.

Выборочным средним Радемахера класса $\mathcal{F}$ называется услов-

[7] См. источники [12], [16], а также лекции Какаде и Тевари [35].

ное математическое ожидание [8]

$$\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma \left(\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(z_i) \right).$$

Вероятностное распределение P на элементах выборки индуцирует вероятностное распределение P^l на выборках $z^l = (z_1, \ \ldots, \ z_l)$ длины l.

Средним по Радемахеру класса $\mathcal{F}$ называется число

$$\mathcal{R}_l(\mathcal{F}) = E_{P^l}(\tilde{\mathcal{R}}_l(\mathcal{F})) = E_{P^l} E_\sigma \left(\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(z_i) \right).$$

Согласно определению, среднее по Радемахеру равно среднему значению выборочного среднего по Радемахеру относительно распределения P^l. Эта величина, наряду с функцией роста и числом покрытия, отражает сложность (емкость) класса функций $\mathcal{F}$.

Приведем ряд свойств средних по Радемахеру, которые будут использоваться при получении верхней оценки ошибки обобщения в разделе 2.7.

По определению выборочное среднее функции f на выборке z^l равно

$$\tilde{E}_{z^l}(f) = \frac{1}{l} \sum_{i=1}^{l} f(z_i).$$

Напомним, что элементы выборки $z^l = (z_1, \ \ldots, \ z_l)$ независимо друг от друга генерируются с помощью вероятностного распределения P.

Математическое ожидание функции f равно $E_P(f) = \int f(z) dP$.

В следующем утверждении приводится оценка разности между математическим ожиданием и выборочным средним, равномерная по всем функциям из класса $\mathcal{F}$.

Теорема 1.11. *Имеет место неравенство*

$$E_{z^l \sim P^l}(\sup_{f \in \mathcal{F}}(E_P(f) - \tilde{E}_{z^l}(f))) \leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.46}$$

[8] Это случайная величина, зависящая от случайной выборки $z^l = (z_1, \ \ldots, \ z_l)$.

Доказательство. Пусть $\tilde{z}^l = (\tilde{z}_1, \ \ldots, \ \tilde{z}_l)$ – случайные величины, распределенные так же, как случайные величины $z^l = (z_1, \ \ldots, \ z_l)$. Кроме этого, предполагаем, что

$$\tilde{z}_1, \ \ldots, \ \tilde{z}_l, z_1, \ \ldots, \ z_l$$

есть последовательность независимых случайных величин.

Имеет место следующая цепочка равенств и неравенств:

$$\begin{aligned}
E_{z^l \sim P^l} \left(\sup_{f \in \mathcal{F}} \left(E_P(f) - \frac{1}{l} \sum_{i=1}^{l} f(z_i) \right) \right) = \\
= E_{z^l \sim P^l} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} E_{\tilde{z}_i \sim P}(f(\tilde{z}_i)) - f(z_i)) \right) \right) \leqslant \\
\leqslant E_{z^l \sim P^l} \left(E_{\tilde{z}^l \sim P^l} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} (f(\tilde{z}_i) - f(z_i)) \right) \right) \right) = \\
= E_{z^l \tilde{z}^l \sim P^{2l}} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} (f(\tilde{z}_i) - f(z_i)) \right) \right) = \\
= E_{z^l \tilde{z}^l \sim P^{2l}} E_{\sigma \sim B_{1/2}} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} \sigma_i (f(\tilde{z}_i) - f(z_i)) \right) \right) \leqslant \\
= E_{\tilde{z}^l \sim P^l} E_{\sigma \sim B_{1/2}} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} \sigma_i f(\tilde{z}_i) \right) \right) + \\
+ E_{z^l \sim P^l} E_{\sigma \sim B_{1/2}} \left(\sup_{f \in \mathcal{F}} \left(\frac{1}{l} \sum_{i=1}^{l} \sigma_i f(z_i) \right) \right) = \\
= 2\mathcal{R}_l(\mathcal{F}). \quad (1.47)
\end{aligned}$$

Переход от 2-й строки к 3-й происходит по свойству: супремум суммы не превосходит суммы супремумов. Появление в 5-й строке σ_i не изменило супремум, так как математическое ожидание супремума инвариантно относительно перестановок переменных z_i и $\tilde{z}_i$; по этой же причине мы можем вставить в 6-й строке символ среднего $E_{\sigma \sim B_{1/2}}$.

Неравенство (1.46) доказано. $\triangle$

Приведем два следствия из теоремы 1.11.

Во-первых, неравенство (1.46) можно обратить:

Следствие 1.5. *Имеет место неравенство*

$$E_{P^l}(\sup_{f\in\mathcal{F}}(\tilde{E}_{z^l}(f) - E_P(f))) \leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.48}$$

Неравенство (1.48) прямо следует из неравенства (1.46) и очевидного равенства $\mathcal{R}_l(\mathcal{F}) = \mathcal{R}_l(-\mathcal{F})$, где $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$.

Для доказательства второго следствия нам потребуется следующая лемма, которую мы приводим без доказательства. Эта лемма также будет использована в дальнейшем.

Лемма 1.13. *Пусть $f : \mathcal{Z}^l \to \mathcal{R}$ – функция, удовлетворяющая условию*

$$\begin{gathered}|f(z_1,\ \ldots,\ z_{i-1}, z_i, z_{i+1},\ \ldots,\ z_l) - \\ -f(z_1,\ \ldots,\ z_{i-1}, z_i', z_{i+1},\ \ldots,\ z_l)| \leqslant c_i\end{gathered}$$

для любого i и для всех $z_1,\ \ldots,\ z_l, z_i' \in \mathcal{Z}$, где $c_1,\ \ldots,\ c_l$ – некоторые константы.

Пусть также $\tilde{z}_1,\ \ldots,\ \tilde{z}_l$ – независимые одинаково распределенные (согласно вероятностному распределению P) случайные величины, принимающие значения в множестве $\mathcal{Z}$.

Тогда имеет место неравенство:

$$\begin{aligned}P^l\{f(\tilde{z}_1,\ \ldots,\ \tilde{z}_l) - E_{P^l}(f(\tilde{z}_1,\ \ldots,\ \tilde{z}_l)) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right),\end{aligned} \tag{1.49}$$

где E_{P^l} – символ математического ожидания по распределению P^l на выборках длины l.

Доказательство этой леммы можно найти в работе [45] и в монографии [51].

Так как условие леммы выполнено при замене f на $-f$, выполнено также неравенство:

$$P^l\{E_{P^l}(f(z_1,\ \ldots,\ z_l)) - f(z_1,\ \ldots,\ z_l) \geqslant t\} \leqslant$$
$$\leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right). \qquad (1.50)$$

Следующее следствие дает равномерную по функциям из класса $\mathcal{F}$ оценку разности между математическим ожиданием функции и выборочным средним этой же функции. Эти величины отличаются на удвоенное среднее по Радемахеру класса функций.

Следствие 1.6. *Допустим, что значения функций из класса $\mathcal{F}$ лежат в интервале $[0,1]$. Тогда для произвольного $\delta > 0$ с вероятностью $1-\delta$ выполнено:*

$$E_P(f) \leqslant \tilde{E}_{z^l}(f) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2l}} \leqslant$$
$$\leqslant \tilde{E}_{z^l}(f) + 2\tilde{\mathcal{R}}_l(\mathcal{F}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}} \qquad (1.51)$$

для всех $f \in \mathcal{F}$, где $z^l = z_1,\ \ldots,\ z_l$.

Доказательство. Для любой функции f имеет место очевидное неравенство

$$E_P(f) \leqslant \tilde{E}_{z^l}(f) + \sup_{h\in\mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h)). \qquad (1.52)$$

Применим неравенство (1.49) леммы 1.13 ко второму члену в правой части (1.52).

Так как значения функции f ограничены единицей, можно взять $c_i = 1/l$ при $1 \leqslant i \leqslant l$. Подставляем эти значения в правую часть неравенства (1.49) и приравниваем ее $\delta/2$. Получаем

$$\exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right) = e^{-2t^2 l} = \delta/2.$$

Отсюда $t = \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}$. Из неравенства (1.49) следует, что с вероятностью $1 - \delta/2$ выполнено

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h)) \leqslant E_{P^l}(\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h))) + \\ + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \qquad (1.53)$$

Неравенство (1.46) утверждает, что

$$E_{P^l}(\sup_{f \in \mathcal{F}}(E_P(f) - \tilde{E}_{z^l}(f))) \leqslant 2\mathcal{R}_l(\mathcal{F}).$$

Отсюда и из (1.53) получаем

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}_{z^l}(h)) \leqslant 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \qquad (1.54)$$

Отсюда следует, что с вероятностью $1 - \delta/2$ выполнено

$$E_P(f) \leqslant \tilde{E}_{z^l}(f) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}} \qquad (1.55)$$

для любой функции $f \in \mathcal{F}$.

Таким образом выполнено первое неравенство (1.51) следствия.

Аналогичным образом, с помощью неравенства (1.50) леммы 1.13 получаем, что с вероятностью $1 - \delta/2$ выполнено

$$\mathcal{R}_l(\mathcal{F}) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F}) + \sqrt{\frac{1}{2l} \ln \frac{2}{\delta}}. \qquad (1.56)$$

Из неравенств (1.55) и (1.56) получаем, что с вероятностью $1 - \delta$ выполнено второе неравенство (1.51) следствия. $\triangle$

В следующей теореме дается оценка среднего по Радемахеру класса $\phi \circ \mathcal{F} = \phi(\mathcal{F}) = \{\phi(f) : f \in \mathcal{F}\}$ композиций функций из $\mathcal{F}$ с заданной функцией ϕ.

Теорема 1.12. *Пусть функция ϕ удовлетворяет условию Липшица с константой L:*

$$|\phi(x) - \phi(y)| \leqslant L|x - y|$$

для всех x, y. Тогда выборочное среднее и среднее по Радемахеру классов $\mathcal{F}$ и $\phi \circ \mathcal{F}$ связаны неравенствами:

$$\tilde{\mathcal{R}}_l(\phi(\mathcal{F})) \leqslant L\tilde{\mathcal{R}}_l(\mathcal{F}), \tag{1.57}$$

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant L\mathcal{R}_l(\mathcal{F}). \tag{1.58}$$

Доказательство. Пусть $z^l = (z_1, \ \ldots, \ z_l)$ – случайная выборка элементов из области определения функций из класса $\mathcal{F}$, распределенная согласно мере P, $\sigma_1, \ \ldots, \ \sigma_l$ – набор независимых бернуллиевских случайных величин со значениями из множества $\{+1, -1\}$, и пусть σ – соответствующее распределение на наборах этих величин.

Нижеследующие преобразования верны при $E = E_\sigma$, а также при $E = E_{P^l}E_\sigma$ – соответствующие математические ожидания по распределениям на этих наборах. Таким образом, мы сразу докажем оба неравенства (1.57) и (1.58).

По определению (выборочное) среднее по Радемахеру класса функций $\phi(\mathcal{F})$ равно

$$\mathcal{R}_l(\phi(\mathcal{F})) = E\left(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma_i \phi(f(z_i))\right). \tag{1.59}$$

Для простоты рассуждений предполагаем, что $L = 1$. Для этого можно заменить функцию ϕ на ϕ/L. Нам необходимо доказать неравенство

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant \mathcal{R}_l(\mathcal{F}) = E\left(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma_i f(z_i)\right). \tag{1.60}$$

Мы осуществим переход от (1.59) к (1.60) с помощью цепочки неравенств по шагам. На каждом шаге рассматривается последовательность вспомогательных функций $(\phi_1, \ \ldots, \ \phi_l)$, где каждая функция ϕ_i есть функция ϕ или тождественная функция I.

На первом шаге все функции $\phi_i = \phi$, на последнем шаге все эти функции – тождественные: $\phi_i = I$. Мы также предполагаем, что на каждом шаге, кроме последнего, $\phi_1 = \phi$. При переходе к следующему шагу очередная функция $\phi_i = \phi$ будет заменяться на тождественную функцию: $\phi_i' = I$. При этом будет выполнена сле-

дующая цепочка равенств и неравенств:

$$
\begin{aligned}
E(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i(f(z_i))) = \\
\frac{1}{2l}E(\sup_{f\in\mathcal{F}}(\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)))+ \\
+\sup_{f\in\mathcal{F}}(-\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)))) = \\
=\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(\phi(f(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))- \\
-\phi(f'(z_1))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) \leqslant \\
\leqslant\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(|f(z_1)-f'(z_1)|+ \\
+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) = \\
=\frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(f(z_1)-f'(z_1)+ \\
+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) = \\
=\frac{1}{2l}E(\sup_{f\in\mathcal{F}}(f(z_1)+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))+ \\
\sup_{f'\in\mathcal{F}}(-f'(z_1)+\sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))) = \\
=E(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i'(f(z_i))),
\end{aligned}
\tag{1.61}
$$

где набор функций $\phi_1',\ \ldots,\ \phi_l'$ имеет на единицу большее число тождественных функций, чем набор $\phi_1,\ \ldots,\ \phi_l$.

В цепочке (1.61) при переходе от 1-й строки к 2-й и 3-й было взято математическое ожидание по σ_1; после этого можно попрежнему рассматривать E как математическое ожидание по всему набору σ, так как теперь переменная σ_1 отсутствует. При переходе от 4-й и 5-й строки к 6-й и 7-й было использовано замечание, что верхняя грань достигается при неотрицательном значении разности $\phi(f(z_1)) - \phi(f'(z_1))$, поэтому можно заменить эту разность на ее абсолютную величину, после чего, использовать условие Липшица с $L = 1$. Аналогичное замечание было использовано при переходе от 6-й и 7-й строки к 8-й и 9-й. При переходе от 8-й и 9-й строки к 10-й строке было использовано то же соображение, что и при переходе от 1-й строки к 2-й и 3-й.

Применяя несколько раз цепочку преобразований (1.61), мы получим выражение

$$E(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma_i \phi_i'(f(z_i))), \tag{1.62}$$

в котором все ϕ_i' являются тождественными функциями, т.е. сумма (1.62) равна $\mathcal{R}_l(\mathcal{F})$.

Первая строка цепочки (1.61) равна $\mathcal{R}_l(\phi(\mathcal{F}))$ при $E = E_{P^l}E_\sigma$ или $\tilde{\mathcal{R}}_l(\phi(\mathcal{F}))$ при $E = E_\sigma$. Таким образом, неравенства (1.57) и (1.58) выполнены и теорема доказана. $\triangle$

1.5. Средние по Радемахеру и другие меры емкости класса функций

Укажем связь среднего по Радемахеру с другими известными мерами емкости классов функций – функцией роста $B_{\mathcal{F}}(l)$ и числом покрытия $\mathcal{N}(\alpha, \mathcal{F}, l)$.

Связь с функцией роста. Нам потребуется следующее вспомогательное утверждение – лемма Массара:

Лемма 1.14. *Пусть A – конечное подмножество $\mathcal{R}^l$ и $\sigma_1, \ldots, \sigma_l$*

– независимые бернуллиевские случайные величины. Тогда

$$E_\sigma\left(\sup_{a\in A}\frac{1}{l}\sum_{i=1}^{l}\sigma_i a_i\right)\leqslant\sup_{a\in A}\|a\|\frac{\sqrt{2\ln|A|}}{l},$$

где $a=(a_1,\ \ldots,\ a_l)$.

Доказательство. Имеет место следующая ниже цепочка равенств и неравенств. Обозначим $E=E_\sigma$. При переходе от первой строки ко второй используется выпуклость экспоненты. При переходе от 7-й строки к 8-й используется неравенство $e^x+e^{-x}\leqslant 2e^{x^2/2}$. Остальные переходы очевидны:

$$\begin{aligned}
&\exp\left(\lambda E\left(\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right)\right)\leqslant\\
&\leqslant E\left(\exp\left(\lambda\sup_{a\in A}\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=E\left(\sup_{a\in A}\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)\leqslant\\
&\leqslant E\left(\sum_{a\in A}\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=\sum_{a\in A}E\left(\exp\left(\lambda\sum_{i=1}^{l}\sigma_i a_i\right)\right)=\\
&=\sum_{a\in A}\prod_{i=1}^{l}E(\exp(\lambda\sigma_i a_i))=\\
&=\sum_{a\in A}\prod_{i=1}^{l}\frac{e^{\lambda a_i}+e^{-\lambda a_i}}{2}\leqslant\\
&\leqslant\sum_{a\in A}\prod_{i=1}^{l}e^{\lambda^2\|a\|^2/2}\leqslant\\
&\leqslant|A|e^{\lambda^2 r^2/2},
\end{aligned}$$

где $r = \sup\limits_{a \in A} \|a\|$.

Логарифмируем первую и последнюю части этой цепочки и получаем неравенство

$$E\left(\sup_{a \in A} \sum_{i=1}^{l} \sigma_i a_i\right) \leqslant \frac{\ln |A|}{\lambda} + \frac{\lambda r^2}{2}. \tag{1.63}$$

Легко проверяется, что правая часть неравенства (1.63) достигает своего минимума при $\lambda = \sqrt{2 \ln |A| / r^2}$. Подставляем это значение λ в правую часть неравенства (1.63) и получаем

$$E\left(\sup_{a \in A} \sum_{i=1}^{l} \sigma_i a_i\right) \leqslant r\sqrt{2 \ln |A|}.$$

Лемма доказана. $\triangle$

Связь среднего по Радемахеру с функцией роста устанавливается в следующей теореме.

Теорема 1.13. *Пусть $\mathcal{F}$ – класс индикаторных функций, т.е. функций принимающих бинарные значения из множества $\{-1, +1\}$. Тогда*

$$\mathcal{R}_l(\mathcal{F}) \leqslant \sqrt{\frac{2 \ln B_{\mathcal{F}}(l)}{l}}$$

для всех l.

Доказательство. Пусть $E = E_{P^l}$ и бинарная строка $a = (a_1, \ldots, a_l)$ представляет значения $(f(z_1), \ldots, f(z_l))$, $z^l = (z_1, \ldots, z_l)$. Имеет место следующая цепочка неравенств:

$$\begin{aligned}
\mathcal{R}_l(\mathcal{F}) = EE_\sigma\left(\sup_a \frac{1}{l} \sum_{i=1}^{l} \sigma_i a_i\right) \leqslant \\
\leqslant E\left(\sup_a \|a\| \frac{\sqrt{2 \ln |\mathcal{F}_{z^l}|}}{l}\right) \leqslant \\
\leqslant \sqrt{l} \frac{\sqrt{2 \ln B_{\mathcal{F}}(l)}}{l} = \\
= \sqrt{\frac{2 \ln B_{\mathcal{F}}(l)}{l}}.
\end{aligned}$$

При переходе от 1-й строки ко 2-й была использована лемма 1.14, при переходе от 2-й строке к 3-й было использовано значение евклидовой нормы бинарного вектора $\|a\| = \sqrt{l}$. Здесь же было использовано определение функции роста семейства. Теорема доказана. $\triangle$

Связь с числом покрытия. Пусть $\mathcal{F}$ – класс функций с областью определения $\mathcal{X}$ и с областью значений $[-1, 1]$. На множестве $\mathcal{X}$ задана некоторая вероятностная мера. Пусть $x^l = (x_1, \ldots, x_l)$ – случайная выборка из элементов $\mathcal{X}$.

Рассмотрим норму $l_{x^l}(f, g) = \sup_{1 \leqslant i \leqslant l} |f(x_i) - g(x_i)|$ на выборке x^l и число покрытия $\mathcal{N}(\alpha, \mathcal{F}, x^l)$ относительно этой выборки, которое равно размеру наименьшего по числу элементов множества $B \subseteq \mathcal{F}$ такого, что для любого $f \in \mathcal{F}$ найдется $g \in B$, так что $l_{x^l}(f, g) < \alpha$.

Теорема 1.14. *Для выборочного среднего по Радемахеру имеет место неравенство*

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \inf_{\alpha} \left(\sqrt{\frac{2 \ln \mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha \right). \tag{1.64}$$

Доказательство. Пусть B – минимальное покрытие класса $\mathcal{F}$ относительно выборки x^l. Можно считать, что область определения функций из B есть $\{x_1, \ldots, x_l\}$.

Пусть также

$$B_\alpha(g) = \{f \in \mathcal{F} : l_{x^l}(f, g) < \alpha\}.$$

Из определения покрытия имеем $\cup_{g\in B} B_\alpha(g) = \mathcal{F}$. Поэтому

$$\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right)\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right)\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i) + \frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right)\right) \leqslant$$

$$\leqslant E_\sigma\left(\sup_{g\in B}\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i)\right) +$$

$$+E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right). \quad (1.65)$$

Для последнего среднего из (1.65) выполнено неравенство

$$E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right) =$$

$$= E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\left|\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right|\right) \leqslant$$

$$\leqslant E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i|f(x_i)-g(x_i)|\right) \leqslant \alpha. \quad (1.66)$$

По лемме 1.14 получаем

$$E_\sigma\left(\sup_{g\in B}\frac{1}{l}\sum_{i=1}^{l}\sigma_i g(x_i)\right) \leqslant$$

$$\leqslant \sup_{g\in B}\|g\|\frac{\sqrt{2\ln|B|}}{l} \leqslant$$

$$\leqslant \sqrt{\frac{2\ln|B|}{l}} = \sqrt{\frac{2\ln\mathcal{N}(\alpha,\mathcal{F},x^l)}{l}}. \quad (1.67)$$

Здесь $\|g\| = \sqrt{\sum\limits_{i=1}^{l} g^2(x_i)} \leqslant \sqrt{l}$, так как размер области определения функции g равен l, а значения по абсолютной величине ограничены единицей.

Соединяем вместе неравенства (1.66) и (1.67) и получаем неравенство

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \left(\sqrt{\frac{2\ln \mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha \right). \tag{1.68}$$

Так как неравенство (1.68) выполнено для любого $\alpha > 0$, оно выполнено и для нижней грани по $\alpha > 0$. Отсюда получаем неравенство (1.64). Теорема доказана. $\triangle$

Из теоремы 1.14 очевидным образом вытекает аналогичное неравенство между средним по Радемахеру и числом покрытия.

Следствие 1.7.

$$\mathcal{R}_l(\mathcal{F}) \leqslant \inf_{\alpha} \left(\sqrt{\frac{2\ln \mathcal{N}(\alpha, \mathcal{F}, l)}{l}} + \alpha \right).$$

Более подробно о средних Радемахера см. в [12], [13], [41].

1.6. Задачи и упражнения

1. Провести полное доказательство лемм 1.4 и 1.5.

2. Пусть Z – некоторое бесконечное множество,

$$\mathcal{P}_k(Z) = \{A : A \subseteq Z \& |A| \leqslant k\}$$

– множество всех его подмножеств, содержащих не более k элементов, f_A – характеристическая функция подмножества A, т.е. функция, равная 1 на элементах A и 0 на элементах его дополнения. Пусть $\mathcal{H}_Z$ – класс всех характеристических функций. Доказать, что функция роста $B_{\mathcal{H}_Z}(l)$ удовлетворяет соотношениям

$$B_{\mathcal{H}_Z}(l) = 2^l$$

при $l \leqslant k$, и

$$B_{\mathcal{H}_Z}(l) = \sum_{i=0}^{k} \binom{l}{i}$$

при $l > k$.

3. Рассматривается множество всех точек на плоскости. Найти значения функции роста $B_H(2)$, $B_H(3)$, $B_H(4)$, $\ldots$, где
a) H – множество классификаторов, порожденных всеми прямыми на плоскости проходящими через начало координат;
b) H – множество классификаторов, порожденных всеми прямыми на плоскости;
c) H – множество классификаторов, порожденных всеми кривыми 2-го порядка, 3-го порядка на плоскости и т.д.

4. Привести примеры классов функций-классификаторов, для которых VC-размерность равна ∞. (*Указание*: Рассмотреть класс функций $\mathcal{F} = \{\mathrm{sign}(\sin(tx)) : t \in \mathcal{R}\}$. Для произвольного l пусть $x_i = 2\pi 10^{-i}$, $i = 1, \ldots, l$, и $\delta_1, \ldots, \delta_l$ – произвольный набор элементов из $\{0, 1\}$, задающий принадлежность элементов x_i двум классам разбиения. Доказать, что при

$$t = \frac{1}{2}\left(\sum_{i=1}^{l}(1-\delta_i)10^i + 1\right)$$

будет выполнено $\mathrm{sig}(\sin(tx_i)) = \delta_i$ для всех i, где $sig(r) = 1$ при $r \geqslant 0$ и $sig(r) = 0$ при $r < 0$ (см. [56])).

5. Рассматривается множество всех точек на плоскости.

a) Проверить, что VC-размерность класса $\mathcal{F}$ индикаторных функций, заданных всеми выпуклыми многоугольниками на плоскости, равна ∞;

b) Проверить, что VC-размерность класса $\mathcal{F}$ характеристических функций всех прямоугольников, стороны которых параллельны осям координат, равна 4 (такая функция равна 1 внутри прямоугольника; как изменится VC-размерность, если рассматривать индикаторные функции, которые могут быть равны как +1 так и -1 внутри прямоугольника).

c) Найти VC-размерность класса $\mathcal{F}$ всех характеристических

функций выпуклых многоугольников на плоскости с d вершинами, где $d \geqslant 3$.

6. Получить оценку 3) из теоремы 1.5 для класса всех линейных функций классификации.

7. Доказать, что рекуррентное соотношение (1.32) имеет решение:

$$\Phi(n,l) = \begin{cases} 2^l \text{ если } l \leqslant n, \\ 2\sum\limits_{i=1}^{n-1} \binom{l-1}{i} \text{ если } l > n. \end{cases}$$

8. Пусть $\mathcal{G}$ – k-мерное векторное пространство функций, определенных на $\mathcal{R}^n$, и задано функциональное прстранство:

$$\mathcal{F} = \{f : f(\bar{x}) = \operatorname{sign}(g(\bar{x})) : g \in \mathcal{G}\}.$$

Доказать, что VC-размерность класса $\mathcal{F}$ не превосходит k. (*Указание*: Пример такого пространства $\mathcal{F}$ – класс всех линейных однородных функций на $\mathcal{R}^n$.)

9. Пусть $\mathcal{L}$ – класс всех линейных однородных функций $f(\bar{x}) = (\bar{w} \cdot \bar{x})$ таких, что $\|\bar{w}\| = \sqrt{(\bar{w} \cdot \bar{w})} \leqslant A$ и $\|\bar{x}\| \leqslant R$. Доказать, что среднее Радемахера этого класса имеет верхнюю оценку:

$$\mathcal{R}_l(\mathcal{L}) \leqslant \frac{AR}{\sqrt{l}}.$$

10. Пусть $\mathcal{F}$ – класс функций, отображающих некоторое множество $\mathcal{X}$ в $\mathcal{R}$ и предположим, что задано подмножество $X = \{x_1, \ldots x_l\}$, где $x_i \in \mathcal{X}$. Тогда X называется *псевдоразделимым* с помощью класса $\mathcal{F}$, если существуют действительные числа $r_1, \ldots, r_l$ такие, что для каждого $E \subseteq X$ существует функция $f_E \in \mathcal{F}$, для которой $f_E(x_i) \geqslant r_i$ при $x_i \in E$ и $f_E(x_i) < r_i$ при $x_i \notin E$, $1 \leqslant i \leqslant l$. Говорим, что $r_1, \ldots, r_l$ – свидетели разделения. По определению *псевдоразмерность* $Pdim(\mathcal{F})$ класса $\mathcal{F}$ равна размеру самого большого по количеству элементов множества X, псевдоразделимого с помощью класса функций $\mathcal{F}$. Доказать следующие свойства псевдоразмерности:

(a) Определим $B_f(x,y) = \operatorname{sign}(f(x) - y)$, где $f \in \mathcal{F}$, $x \in \mathcal{X}$ и $y \in \mathcal{R}$. Пусть $\mathcal{B}_{\mathcal{F}} = \{B_f : f \in \mathcal{F}\}$. Доказать, что $VCdim(\mathcal{B}_{\mathcal{F}}) = Pdim(\mathcal{F})$.

(b) Пусть $\sigma : \mathcal{R} \to \mathcal{R}$ – произвольная неубывающая функция и $\sigma(\mathcal{F}) = \{\sigma \cdot f : f \in \mathcal{F}\}$. Тогда $Pdim(\sigma(\mathcal{F})) \leqslant Pdim(\mathcal{F})$.

(c) Для всех $\gamma > 0$ $\mathrm{fat}_{\mathcal{F}}(\gamma) \leqslant Pdim(\mathcal{F})$.

(d) Если конечное множество X псевдо-разделимо, то существует γ_0 такое, что для всех $\gamma < \gamma_0$, X является γ-разделимым (*Указание*: Допустим, что конечное множество $X = \{x_1, \ldots, x_l\}$ псевдоразделимо с помощью класса $\mathcal{F}$ со свидетелями $r_1, \ldots, r_l$. Определим

$$\gamma_0 = \frac{1}{2} \min\{r_i - f_E(x_i); 1 \leqslant i \leqslant l, E \subseteq X, f_E(x_i) < r_i\}$$

Тогда для любого $\gamma < \gamma_0$ множество X является γ-разделимым.)

e) Допустим, что класс вещественных функций $\mathcal{F}$ замкнут относительно умножения на константы. Тогда для всех положительных чисел γ будет $\mathrm{fat}_{\mathcal{F}}(\gamma) = Pdim(\mathcal{F})$; см. [9].

(f) $Pdim(\mathcal{F}) = \lim\limits_{\gamma \to 0} \mathrm{fat}_{\mathcal{F}}(\gamma)$.

(g) Пусть $\mathcal{F}$ – класс всех линейных однородных функций $f(\bar{w}) = (\bar{w} \cdot \bar{x})$, где $\bar{x}, \bar{w} \in \mathcal{R}^n$. Тогда $Pdim(\mathcal{F}) = n$.

11. Доказать лемму 1.9:

$$Sdim(\mathcal{F}^{\alpha}) \leqslant fat_{\alpha/2}(\mathcal{F}),$$
$$\mathcal{M}(\alpha, \mathcal{F}, X) \leqslant \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X),$$

где $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ и $\alpha > 0$.

Глава 2

Метод опорных векторов

Задача классификации и регрессии с помощью метода опорных векторов – Support Vector Machines (SVM) имеет целью разработку алгоритмически эффективных методов построения оптимальной разделяющей гиперплоскости в пространстве признаков высокой размерности. Оптимальность понимается в смысле минимизации верхних оценок вероятности ошибки обобщения.

2.1. Оптимальная гиперплоскость

Предварительно рассмотрим случай полностью разделимой выборки, т.е. случай, когда обучение возможно провести без ошибок.

Выборка $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ \ldots, \ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{R}^n$ и $y_i \in \{-1, 1\}$, $i = 1, \ \ldots, \ l$, называется *разделимой (отделимой)* с помощью гиперплоскости $(\bar{w} \cdot \bar{x}) - c = 0$, если существуют вектор $\bar{w}$ единичной длины ($|\bar{w}| = 1$) и число c такие, что

$$\begin{aligned} (\bar{w} \cdot \bar{x}_i) - c > 0 \text{ при } y_i = 1, \\ (\bar{w} \cdot \bar{x}_i) - c < 0 \text{ при } y_i = -1. \end{aligned} \tag{2.1}$$

В том случае, когда разделяющая гиперплоскость $(\bar{w} \cdot \bar{x}_i) - c = 0$ существует, определим

$$\begin{aligned} c_1(\bar{w}) = \min_{y_i=1} (\bar{w} \cdot \bar{x}_i), \\ c_2(\bar{w}) = \max_{y_i=-1} (\bar{w} \cdot \bar{x}_i). \end{aligned} \tag{2.2}$$

По определению $c_1(\bar{w}) > c_2(\bar{w})$. Кроме того, $c_1(\bar{w}) > c > c_2(\bar{w})$, если гиперплоскость $(\bar{w} \cdot \bar{x}_i) - c = 0$ разделяет выборку.

Определим

$$\rho(\bar{w}) = \frac{c_1(\bar{w}) - c_2(\bar{w})}{2}.$$

Тогда $\rho(\bar{w}) = \frac{1}{2}((c_1(\bar{w}) - c) + (c - c_2(\bar{w}))$ равно половине суммы расстояний от ближайших сверху и снизу точек до разделяющей гиперплоскости $(\bar{w} \cdot \bar{x}) - c = 0$.

Допустим, что что выборка S разделима, т.е. существует c такое, что выполнено условие (2.1).

Максимум непрерывной функции $\rho(\bar{w})$ на компакте $\{\bar{w} : \|\bar{w}\| \leqslant 1\}$ существует. Пусть максимум достигается при $\bar{w} = \bar{w}_0$.

Лемма 2.1. *Пусть указанный выше максимум $\rho(\bar{w})$ достигается при $\bar{w} = \bar{w}_0$. Тогда гиперплоскость $(\bar{w}_0 \cdot \bar{x}) - c_0 = 0$, где $c_0 = \frac{1}{2}(c_1(\bar{w}_0) + c_2(\bar{w}_0))$, отделяет выборку S и находится точно в середине между ближайшими сверху и снизу точками положительной и отрицательной частей выборки.*

Доказательство. Действительно, при $y_i = 1$

$$(\bar{w}_0 \cdot \bar{x}_i) - c_0 \geqslant c_1(\bar{w}_0) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = \\ = \frac{c_1(\bar{w}_0) - c_2(\bar{w}_0)}{2} > 0. \tag{2.3}$$

При $y_i = -1$

$$(\bar{w}_0 \cdot \bar{x}_i) - c_0 \leqslant c_2(\bar{w}_0) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = \\ = -\frac{c_1(\bar{w}_0) - c_2(\bar{w}_0)}{2} < 0. \tag{2.4}$$

Оставшаяся часть леммы предоставляется читателю в качестве задачи. $\triangle$

Назовем гиперплоскость $(\bar{w}_0 \cdot \bar{x}) - c_0 = 0$ *оптимальной*. Для этой гиперплоскости сумма расстояний от ближайших к ней (сверху и снизу) точек выборки максимальна среди всех разделяющих S гиперплоскостей.

Лемма 2.2. *Оптимальная гиперплоскость – единственная гиперплоскость для которой сумма расстояний от ближайших к ней (сверху и снизу) точек выборки максимальна среди всех разделяющих S гиперплоскостей, расположенных на равных от них расстояниях.*

Доказательство. Максимум $\bar{w}_0$ непрерывной функции $\rho(\bar{w})$ на компакте $\{\bar{w} : \|\bar{w}\| \leqslant 1\}$ достигается на границе, так как в противном случае при $\bar{w}^* = \frac{\bar{w}_0}{\|\bar{w}_0\|}$ было бы $\|\bar{w}^*\| = 1$ и $\rho(\bar{w}^*) = \frac{\rho(\bar{w}_0)}{\|\bar{w}_0\|} > \rho(\bar{w}_0)$.

Этот максимум единственный, так как функция $\rho(\bar{w})$ вогнутая; если бы ее максимум достигался в двух точках, лежащих на границе компакта, то он достигался бы и во внутренней точке, что противоречит только что доказанному. $\triangle$

Докажем, что функция $\rho(\bar{w})$ вогнутая. Для этого надо проверить, что

$$\rho(\lambda\bar{w} + (1-\lambda)\bar{u}) \geqslant \lambda\rho(\bar{w}) + (1-\lambda)\lambda(\bar{u}) \tag{2.5}$$

для всех $0 \leqslant \lambda \leqslant 1$ и $\bar{w}$, $\bar{u}$, лежащих в единичном шаре.

Имеют место неравенства

$$\begin{aligned} &\min_{i\in I}(f(i) + g(i)) \geqslant \min_{i\in I} f(i) + \min_{i\in I} g(i), \\ &\max_{i\in I}(f(i) + g(i)) \leqslant \max_{i\in I} f(i) + \max_{i\in I} g(i) \end{aligned} \tag{2.6}$$

для произвольных функций f и g и множества I.

По определению

$$\rho(\bar{w}) = \frac{1}{2}(\min_{y_i=1}(\bar{w}\cdot\bar{x}_i) - \max_{y_i=-1}(\bar{w}\cdot\bar{x}_i)),$$

Из (2.6) при $f(i) = (\bar{w}\cdot\bar{x}_i)$ и $g(i) = (\bar{u}\cdot\bar{x}_i)$ имеем

$$\begin{aligned} &\min_{y_i=1}((\lambda\bar{w} + (1-\lambda)\bar{u})\cdot\bar{x}_i) = \\ &= \min_{y_i=1}(\lambda(\bar{w}\cdot\bar{x}_i) + (1-\lambda)(\bar{u}\cdot\bar{x}_i)) \geqslant \\ &\geqslant \lambda\min_{y_i=1}(\bar{w}\cdot\bar{x}_i) + (1-\lambda)\min_{y_i=1}(\bar{u}\cdot\bar{x}_i). \end{aligned}$$

Аналогичное неравенство имеет место для максимумов.

Вычитанием соответствующих неравенств получаем неравенство (2.5). $\triangle$

Рассмотрим эквивалентное определение оптимальной разделяющей гиперплоскости. На основе этого определения будет разработан *алгоритмически эффективный* метод построения оптимальной гиперплоскости в виде задачи квадратичного программирования. Точный алгоритм, построенный по этому методу, будет приведен в следующем разделе.

Найдем вектор $\bar{w}_0$ и число b_0 так, чтобы было

$$\begin{aligned} (\bar{w}_0 \cdot \bar{x}_i) + b_0 \geqslant 1 \text{ при } y_i = 1, \\ (\bar{w}_0 \cdot \bar{x}_i) + b_0 \leqslant -1 \text{ при } y_i = -1, \end{aligned} \tag{2.7}$$

где $i = 1, \ \ldots, \ l$, и величина $\|\bar{w}_0\|$ была бы минимальна при этих ограничениях.

Теорема 2.1. *Вектор $\bar{w}_0$, удовлетворяющий условиям* (2.7) *и имеющий минимальную норму, определяет оптимальную разделяющую гиперплоскость с весовым вектором $\bar{w}_0^* = \frac{\bar{w}_0}{\|\bar{w}_0\|}$. При этом*

$$\rho(\bar{w}_0^*) = \max_{\|\bar{w}\|=1} \left(\frac{1}{2} (\min_{y_i=1} (\bar{w} \cdot \bar{x}_i) - \max_{y_i=-1} (\bar{w} \cdot \bar{x}_i)) \right) = \frac{1}{\|\bar{w}_0\|}.$$

Доказательство. Имеем

$$\rho(\bar{w}_0^*) = \frac{1}{2} \left(c_1 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) - c_2 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) \right) \geqslant \frac{1}{\|\bar{w}_0\|},$$

так как по (2.7):

$$\begin{aligned} c_1 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) \geqslant \frac{1 - b_0}{\|\bar{w}_0\|}, \\ c_2 \left(\frac{\bar{w}_0}{\|\bar{w}_0\|} \right) \leqslant \frac{-1 - b_0}{\|\bar{w}_0\|}. \end{aligned}$$

Остается доказать, что $\rho(\bar{w}_0^*) > \frac{1}{\|\bar{w}_0\|}$ невозможно. Допустим противное. Рассмотрим вектор $\bar{w}_1 = \frac{\bar{w}_0^*}{\rho(\bar{w}_0^*)}$. Для него имеем неравенство

$$\|\bar{w}_1\| = \frac{\|\bar{w}_0^*\|}{\rho(\bar{w}_0^*)} < \|\bar{w}_0\|,$$

так как $\|\bar{w}_0^*\| = 1$.

Докажем, что вектор $\bar{w}_1$ удовлетворяет условию (2.7) при $b_0 = -\frac{c_1(\bar{w}_1)+c_2(\bar{w}_1)}{2}$. Имеем при $y_i = 1$:

$$(\bar{w}_1 \cdot \bar{x}_i) - \frac{c_1(\bar{w}_1) + c_2(\bar{w}_1)}{2} =$$
$$= \frac{1}{\rho(w_0^*)}(\bar{w}_0^* \cdot \bar{x}_i) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2\rho(w_0^*)} \geqslant$$
$$\geqslant \frac{c_1(\bar{w}_0^*)}{\frac{1}{2}(c_1(\bar{w}_0^*) - c_2(\bar{w}_0^*))} - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{c_1(\bar{w}_0^*) - c_2(\bar{w}_0^*)} = 1.$$

Случай $y_i = -1$ разбирается аналогичным образом.

Отсюда получаем противоречие, так как вектор $\bar{w}_1^*$ имеет меньшую норму, чем вектор $\bar{w}_0^*$. Поэтому $\rho(\bar{w}_0^*) = \frac{1}{\|\bar{w}_0\|}$. $\triangle$

По выбору $\bar{w}_0^*$ величина $\rho(\bar{w}_0^*)$ максимальна

$$\rho(\bar{w}_0^*) = \max_{\|\bar{w}\|=1} \rho(\bar{w}) = \frac{1}{\|\bar{w}_0\|}.$$

По теореме 2.1 величина $\rho(\bar{w}_0^*) = \frac{1}{\|\bar{w}_0\|}$ равна расстоянию от ближайших точек положительной и отрицательной частей выборки до оптимальной гиперплоскости

$$(\bar{w}_0^* \cdot \bar{x}) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2} = 0,$$

которая расположена на равных расстояниях между гиперплоскостями

$$(\bar{w}_0^* \cdot \bar{x}) - \frac{c_1(\bar{w}_0^*) + c_2(\bar{w}_0^*)}{2} = \pm 1,$$

оптимально ограничивающими точки положительной и отрицательной частей выборки.

Мы называем величину $\rho(\bar{w}_0^*)$ *геометрическим краем ошибки.*

Уравнение оптимальной гиперплоскости также можно записать в виде

$$(\bar{w}_0 \cdot \bar{x}) - \frac{c_1(\bar{w}_0) + c_2(\bar{w}_0)}{2} = 0.$$

2.2. Алгоритм построения оптимальной гиперплоскости

В этом разделе мы приведем алгоритм построения оптимальной гиперплоскости.

Две группы условий (2.7) запишем в виде

$$y_i((\bar{w} \cdot \bar{x}_i) + b) \geqslant 1 \tag{2.8}$$

при $i = 1, \ \ldots, \ l$.

Согласно результатам предыдущего раздела, для нахождения оптимальной гиперплоскости мы должны минимизировать норму весового вектора $\|\bar{w}\|$ при ограничениях (2.8).

В разделе 2.10 (ниже) указано, что для решения квадратичной задачи оптимизации

$$(\bar{w} \cdot \bar{w}) = \sum_{i=1}^{l} w_i^2 \to \min$$

при ограничениях (2.7) (или эквивалентных им ограничениях (2.8)) составим лагранжиан

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) - \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1), \tag{2.9}$$

где $\alpha_i \geqslant 0$ – множители Лагранжа.

Для того, чтобы найти седловую точку лагранжиана (2.9), необходимо минимизировать его по переменным $\bar{w}$ и b, а после этого максимизировать по множителям Лагранжа при условиях $\alpha_i \geqslant 0$, $i = 1, \ \ldots, \ l$.

Необходимое условие минимума лагранжиана имеет вид

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i = \bar{0}, \tag{2.10}$$

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.11}$$

Из (2.10) – (2.11) следует, что

$$\bar{w} = \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i, \tag{2.12}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.13}$$

Подставим (2.12) в (2.9) и положим $W(\bar{\alpha}) = L(\bar{w}, b, \bar{\alpha})$. С учетом (2.13) получим

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{x}_i \cdot \bar{x}_j). \tag{2.14}$$

Для нахождения оптимальной гиперплоскости нам надо максимизировать функцию (2.14) при условиях (2.13) и $\alpha_i \geqslant 0$, где $i = 1, \ \ldots, \ l$.

Пусть максимум достигается при $\alpha_i = \alpha_i^0$, $i = 1, \ \ldots, \ l$. Тогда решение задачи поиска оптимальной гиперплоскости имеет вид

$$\bar{w}_0 = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{x}_i. \tag{2.15}$$

При этом

$$b_0 = \frac{\min\limits_{y_i=1}(\bar{w}_0 \cdot \bar{x}_i) + \max\limits_{y_i=-1}(\bar{w}_0 \cdot \bar{x}_i)}{2}.$$

Оптимальные решения $\bar{w}_0$ и b_0 должны удовлетворять условиям Каруша–Куна–Таккера

$$\alpha_i^0 (y_i((\bar{w}_0 \cdot \bar{x}_i) + b) - 1) = 0 \tag{2.16}$$

при $i = 1, \ \ldots, \ l$.

Отсюда следует, что $\alpha_i^0 > 0$ может быть только для тех i, для которых $y_i((\bar{w}_0 \cdot \bar{x}_i) \ + \ b_0) - 1 = 0$, т.е. для тех векторов, которые лежат на гиперплоскостях $(\bar{w}_0 \cdot \bar{x}_i) + b_0 = \pm 1$. Такие векторы называются *опорными* векторами (support vectors). Вектор весов $\bar{w}_0$

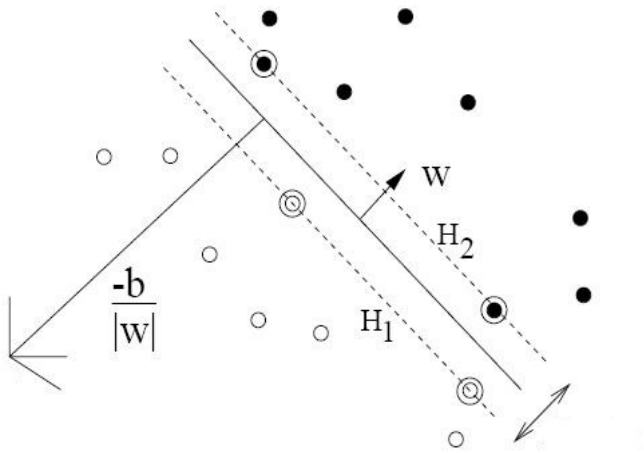

Рис. 1.1. Опорные векторы расположены на граничных гиперплоскостях H_1 и H_2

представляет собой линейную комбинацию опорных векторов $\bar{x}_{i_s}$, $s = 1, \ \ldots, \ k$, где k – число опорных векторов:

$$\bar{w}_0 = \sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} \bar{x}_{i_s}.$$

Оптимальная гиперплоскость имеет вид

$$\sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} (\bar{x}_{i_s} \cdot \bar{x}) + b_0 = 0. \tag{2.17}$$

Остальные, не опорные, векторы, можно не принимать во внимание, например, их можно изменить, при этом оптимальная гиперплоскость не изменится.

Приведем также некоторые соотношения с опорными векторами:

$$\|\bar{w}_0\|^2 = (\bar{w}_0 \cdot \bar{w}_0) = \sum_{s,q=1}^{k} \alpha_{i_s}^0 \alpha_{i_q}^0 y_{i_s} y_{i_q} (\bar{x}_{i_s} \cdot \bar{x}_{i_q}), \tag{2.18}$$

а также

$$W(\bar{\alpha}^0) = \sum_{s,q=1}^{k} \alpha_{i_s}^0 - \frac{1}{2}\|\bar{w}_0\|^2.$$

Суммируя (2.16), получим

$$\sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} (\bar{w}_0 \cdot \bar{x}_{i_s}) + b_0 \sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} = \sum_{s=1}^{k} \alpha_{i_s}^0.$$

По (2.11) второе слагаемое этой суммы равно 0. Отсюда, используя (2.18), получаем

$$\sum_{s=1}^{k} \alpha_{i_s}^0 = \sum_{s,q=1}^{k} \alpha_{i_s}^0 \alpha_{i_q}^0 y_{i_s} y_{i_q} (\bar{x}_{i_s} \cdot \bar{x}_{i_q}) = \|\bar{w}_0\|^2.$$

Поэтому $W(\bar{\alpha}^0) = \frac{1}{2}\|\bar{w}_0\|^2$. Имеем также

$$\|\bar{w}_0\| = \frac{1}{\sqrt{\sum\limits_{s=1}^{k} \alpha_{i_s}^0}}.$$

2.3. Оценка вероятности ошибки обобщения через число опорных векторов

Выше было показано, что оптимальная разделяющая гиперплоскость определяется не всеми векторами выборки S, а только опорными векторами.

Можно рассматривать переход от выборки S к разделяющей гиперплоскости $\rho(S)$ как схему сжатия информации, содержащейся в выборке S.

Небольшое число опорных векторов и их признаков $\hat{S}$ определяет ту же гиперплоскость, что и вся выборка S, т.е. $\rho(\hat{S}) = \rho(S)$.

Допустим, что размер $\hat{S}$ равен d. Всего имеется $\binom{l}{d}$ подмножеств множества элементов выборки. Каждое такое подмножество определяет функцию классификации $h_{\hat{S}}$. Для каждой такой функции классификации $h_{\hat{S}}$ вероятность того, что она согласована с остальными $l-d$ точками, но имеет ошибку обобщения больше ϵ, ограничена величиной $(1-\epsilon)^{l-d} \leqslant \exp(-\epsilon(l-d))$. Тогда вероятность того, что какая-нибудь функция классификации $h_{\hat{S}}$, построенная с помощью схемы сжатия по подмножеству размера d, согласована с l векторами и имеет ошибку обобщения больше чем ϵ, ограничена величиной

$$\binom{l-d}{d} \exp(-\epsilon(l-d)).$$

Таким образом, мы доказали теорему

Теорема 2.2. *Пусть задана некоторая схема сжатия информации $\rho(S)$. Тогда для любого распределения вероятностей P на $X \times \{-1, 1\}$ и любого $0 < \delta < 1$ с вероятностью $1-\delta$ на случайной выборке S размера l функция классификации $h_{\hat{S}}$, построенная с помощью схемы сжатия по подмножеству выборки размера d, имеет ошибку обобщения не более*

$$\mathrm{err}_P(h_{\hat{S}}) \leqslant \frac{1}{l-d}\left(d \ln\left(\frac{el}{d}\right) + \ln\left(\frac{l}{\delta}\right)\right).$$

Из этой теоремы следует, что при $d > 2$ и при достаточно больших l

$$\mathrm{err}_P(h_{\hat{S}}) \leqslant \frac{d \ln l}{l-d},$$

где d – число опорных векторов.

2.4. SVM-метод в пространстве признаков

Задана выборка $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ \ldots, \ (\bar{x}_l, y_l))$. Метод SVM основан на следующей идее. Векторы выборки $\bar{x}_1, \ldots, \bar{x}_l$, принадлежащие пространству $\mathcal{R}^n$, отображаются в пространство более

высокой размерности – пространство признаков (feature space) с помощью некоторого нелинейного отображения, выбранного априори:

$$\bar{x} = (x_1, \ \ldots, \ x_n) \to \bar{\phi}(\bar{x}) = (\phi_1(\bar{x}), \ \ldots, \ \phi_N(\bar{x})). \qquad (2.19)$$

Получаем векторы $\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$ в пространстве признаков $\mathcal{R}^N$.

Заметим, что отображение (2.19) может быть необратимым. Исходное пространство $\mathcal{R}^n$ переходит при отображении $\bar{x} \to \bar{\phi}(\bar{x})$ в некоторое подмножество пространства признаков $\mathcal{R}^N$.

В пространстве $\mathcal{R}^N$ будет строиться оптимальная гиперплоскость, разделяющая векторы $\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$.

Пример. Допустим, что для классификации данных в n-мерном пространстве используется полиномы второй степени от n переменных. Тогда можно рассмотреть следующую конструкцию.

Вводим новые переменные в пространстве признаков

$$\begin{gathered} z_1 = x_1, \ \ldots, \ z_n = x_n, \\ z_{n+1} = x_1^2, \ \ldots, \ z_{2n} = x_n^2, \\ z_{2n+1} = x_1 x_2, \ \ldots, \ z_N = x_n x_{n-1}. \end{gathered}$$

Всего имеется $N = 2n + \frac{n(n-1)}{2}$ таких переменных. Таким образом, мы построили нелинейное отображение

$$\bar{x} = (x_1, \ \ldots, \ x_n) \to \bar{\phi}(\bar{x}) = \bar{z} = (z_1, \ \ldots, \ z_N)$$

пространства $\mathcal{R}^n$ в пространство $\mathcal{R}^N$.

Прообразом разделяющей гиперплоскости в пространстве признаков $Z = \mathcal{R}^N$

$$(\bar{w} \cdot \bar{z}) + b = 0$$

при отображении $\bar{x} \to \bar{\phi}(\bar{x}) = \bar{z}$ является поверхность второго порядка в исходном пространстве $\mathcal{R}^n$:

$$(\bar{w} \cdot \bar{\phi}(\bar{x})) + b = \sum_{i=1}^{N} w_i z_i + b =$$

$$= \sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{2n} w_i x_i^2 + \sum_{i=2n+1}^{N} w_i x_{j_i} x_{k_i} + b = 0,$$

где (j_i, k_i) – пара натуральных чисел с номером i в какой-нибудь взаимно однозначной нумерации всех пар натуральных чисел не больше n.

Рассмотрим теперь общий случай. Пусть задано отображение (2.19)

$$\bar{\phi}(\bar{x}) = (\phi_1(\bar{x}),\ \ldots,\ \phi_N(\bar{x}))$$

исходного пространства $\mathcal{R}^n$ в пространство признаков

$$\mathcal{R}^N = \{\bar{z} = (z_1,\ \ldots,\ z_N) : z_i \in \mathcal{R}, i = 1,\ \ldots,\ N\}.$$

В координатах это отображение записывается в виде $z_j = \phi_j(\bar{x})$, $j = 1,\ \ldots,\ N$.

Элементы выборки $\bar{x}_1,\ \ldots,\ \bar{x}_l$ исходного пространства $\mathcal{R}^n$ переходят в вектора $\bar{\phi}(\bar{x}_1),\ \ldots,\ \bar{\phi}(\bar{x}_l)$ пространства признаков $\mathcal{R}^N$.

Используя метод построения разделяющей гиперплоскости, изложенный в разделе 2.2, построим гиперплоскость в пространстве признаков $\mathcal{R}^N$:

$$\sum_{j=1}^{N} w_j z_j + b = 0, \tag{2.20}$$

разделяющую векторы $\bar{\phi}(\bar{x}_1),\ \ldots,\ \bar{\phi}(\bar{x}_l)$.

Эта гиперплоскость имеет своим прообразом в пространстве $\mathcal{R}^n$ в общем случае нелинейную поверхность

$$\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = 0. \tag{2.21}$$

Используя представление функции классификации в двойственной форме, представим вектор весов разделяющей гиперплоскости в пространстве признаков в виде линейной комбинации опорных векторов из множества $\{\bar{\phi}(\bar{x}_i) : \alpha_i^0 > 0\}$:

$$\bar{w} = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{\phi}(\bar{x}_i).$$

В координатах это представление имеет вид

$$w_j = \sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \tag{2.22}$$

при $j = 1, \ldots, N$. Число слагаемых в этой сумме не зависит от размерности пространства признаков.

Подставим (2.22) в (2.21) и получим выражение для нелинейной поверхности, которая является прообразом в пространстве $\mathcal{R}^n$ разделяющей гиперплоскости, построенной в пространстве признаков $\mathcal{R}^N$:

$$\begin{aligned}
\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = \\
= \sum_{j=1}^{N} \left(\sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \right) \phi_j(\bar{x}) + b = \\
= \sum_{i=1}^{l} \alpha_i^0 y_i \sum_{j=1}^{N} \phi_j(\bar{x}) \phi_j(\bar{x}_i) + b = \\
= \sum_{i=1}^{l} \alpha_i^0 y_i (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}_i)) + b = \\
= \sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}, \bar{x}_i) + b = 0,
\end{aligned} \tag{2.23}$$

где

$$K(\bar{x}_i, \bar{x}) = (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x})). \tag{2.24}$$

Таким образом, все рассуждения для гиперплоскостей в пространстве $\mathcal{R}^n$ годятся и для нелинейных поверхностей в том же пространстве, если мы заменим скалярное произведение $(\bar{x}_i \cdot \bar{x})$ в двойственном представлении оптимальной гиперплоскости (2.17) :

$$\sum_{s=1}^{k} \alpha_i^0 y_i (\bar{x}_i \cdot \bar{x}) + b = 0$$

на функцию $K(\bar{x}_i, \bar{x})$, которая задается выражением (2.24) и которая будет называться *ядром*.

Отметим, что вычисление нелинейной функции

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}_i, \bar{x}) + b, \tag{2.25}$$

соответствующей гиперплоскости (2.23) требует всего l операций и не зависит от размерности N пространства признаков. Из этой формулы также видно, что для построения нелинейного классификатора в исходном пространстве $\mathcal{R}^n$ с помощью линейного классификатора в пространстве признаков нам не нужно знать отображение $\bar{x} \to \bar{\phi}(\bar{x})$, а достаточно только знать ядро $K(\bar{x}_i, \bar{x})$.

Для решения прямой задачи классификации формально мы строим в пространстве $\mathcal{R}^N$ гиперплоскость, разделяющую образы

$$\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$$

векторов $\bar{x}_1, \ \ldots, \ \bar{x}_l$ выборки. При решении задачи построения оптимальной гиперплоскости нам надо решить двойственную задачу – максимизировать функцию $W(\alpha)$, заданную выражением (2.14).

Эта функция, с учетом определения ядра, упрощается следующим образом:

$$\begin{aligned} W(\alpha) &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) = \\ &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\bar{x}_i, \bar{x}_j). \end{aligned} \tag{2.26}$$

Таким образом, для нахождения оптимальной гиперповерхности (2.25), разделяющей выборку $((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ в пространстве $\mathcal{R}^n$, нам надо максимизировать нелинейную функцию (2.26) при условиях (2.13) и $\alpha_i \geqslant 0$, $i = 1, \ \ldots, \ l$. При этом нам не требуется знание N-мерных векторов $\bar{\phi}(\bar{x}_1), \ \ldots, \ \bar{\phi}(\bar{x}_l)$, достаточно знать, что их попарные скалярные произведения $K(\bar{x}_i, \bar{x}_j)$ вычисляются с помощью ядра.

На практике подбираются ядра, для которых соответствующая поверхность наилучшим образом разделяет обучающую выборку.

2.5. Ядра

В этом разделе рассмотрим свойства ядер более подробно. Пусть X – произвольное множество. В общем случае под ядром мы понимаем произвольную функцию $K(x, y)$, отображающую $X \times X$ в множество всех действительных чисел $\mathcal{R}$, которая может быть представлена в виде скалярного произведения

$$K(x, y) = (\phi(x) \cdot \phi(y)), \tag{2.27}$$

где ϕ – отображение множества X в некоторое пространство признаков, снабженное скалярным произведением.

Разберем некоторые примеры ядер, которые применяются в практических приложениях.

Первый пример: $K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y})^d$ или $K(\bar{x}, \bar{y}) = ((\bar{x} \cdot \bar{y}) + c)^d$ – полиномиальные ядра.

Пример. Рассмотрим отображение из $\mathcal{R}^n$ в $\mathcal{R}^{\frac{n(n+1)}{2}}$:

$$\bar{\phi}(\bar{x}) = \bar{\phi}(x_1, \ldots, x_n) =$$
$$= (1, x_1^2, \ \ldots, \ x_n^2, \sqrt{2}x_1x_2, \ \ldots, \ \sqrt{2}x_{n-1}x_n).$$

Тогда

$$K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{y})) = 1 + \sum_{i=1}^{n} x_i^2 y_i^2 + \sum_{i,j=1, i<j}^{n} 2x_i x_j y_i y_j = (1 + \bar{x} \cdot \bar{y})^2.$$

Получаем $K(\bar{x}, \bar{y}) = (1 + (\bar{x} \cdot \bar{y}))^2$ – полиномиальное ядро второго порядка.

Функция классификации (2.25), соответствующая оптимальной разделяющей гиперплоскости в пространстве признаков, в этом случае имеет вид

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i (1 + (\bar{x}_i \cdot \bar{x}))^2 + b. \tag{2.28}$$

Другой вид ядер определяется функциями, которые имеют вид $K(\bar{x}, \bar{y}) = K(\bar{x} - \bar{y})$. Такая функция инвариантна относительно прибавления к $\bar{x}$ и $\bar{y}$ одного и того же вектора.

Пусть размерность равна 1 и функция $K(x)$ определена на $[0, 2\pi]$. В таком случае ее можно доопределить до периодической функции и разложить в равномерно сходящийся ряд Фурье:

$$K(x) = \sum_{n=0}^{\infty} a_n \cos(nx).$$

Тогда

$$K(x, y) = K(x - y) =$$
$$= a_0 + \sum_{n=0}^{\infty} a_n \sin(nx) \sin(ny) + \sum_{n=0}^{\infty} a_n \cos(nx) \cos(ny).$$

Это ядро соответствует отображению

$$x \to (1, \sin x, \cos x, \sin 2x, \cos 2x, \ \ldots, \ \sin(nx), \cos(nx), \ \ldots)$$

исходного пространства в пространство признаков.

В задачах регрессии широко используется гауссово ядро

$$K(\bar{x}, \bar{y}) = \exp(-\|\bar{x} - \bar{y}\|^2/\sigma^2).$$

Гауссово ядро может быть получено трансформацией экспоненциального ядра. Экспоненциальные ядра – это ядро вида

$$K(\bar{u}, \bar{v}) = \exp((\bar{u} \cdot \bar{v})/\sigma^2),$$

где $\sigma > 0$ – параметр.

Экспоненциальное ядро можно разложить в ряд Тейлора в виде бесконечной суммы полиномиальных ядер:

$$\exp((\bar{u} \cdot \bar{v})) = \sum_{k=0}^{\infty} \frac{(\bar{u} \cdot \bar{v})^k}{k!}.$$

Экспоненциальное ядро трансформируется в гауссово ядро следующим образом:

$$\frac{K(\bar{u},\bar{v})}{\sqrt{K(\bar{u},\bar{u})K(\bar{v},\bar{v})}} =$$
$$= \frac{\exp((\bar{u}\cdot\bar{v})/\sigma^2)}{\sqrt{\exp((\bar{u}\cdot\bar{u})/\sigma^2)\exp((\bar{v}\cdot\bar{v})/\sigma^2)}} =$$
$$= \exp(-\|\bar{u}-\bar{v}\|^2/2\sigma^2).$$

В задачах распознавания текстов используются ядра, определенные на дискретных множествах. Приведем пример такого ядра и соответствующего пространства признаков.

Пусть Ξ – конечный алфавит. Слово s в этом алфавите – это произвольная конечная последовательность букв $s = s_1\ s_2\ \ldots\ s_n$; Ξ^* – множество всех слов в алфавите Ξ, включая пустое слово. $|s| = n$ – длина слова $s \in \Xi^*$; длина пустого слова равна 0.

Пусть Ξ^n – множество всех слов (последовательностей) длины n. По определению $\Xi^* = \cup_{n=0}^{\infty}\Xi^n$.

Пусть также, st – это слово, полученное конкатенацией слов s и t, $s[i:j] = s_i\ s_{i+1}\ \ldots\ s_j$.

Слово u является подсловом (подпоследовательностью) слова s, если существует последовательность индексов $\bar{i} = (i_1,\ \ldots,\ i_{|u|})$ такая, что $1 \leqslant i_1 < \cdots < i_{|u|} \leqslant |s|$ и $u_j = s_{i_j}$ для всех $j = 1,\ \ldots,\ |u|$; Это также обозначаем $u = s[\bar{i}]$. Длиной подпоследовательности u в s называется число $l(\bar{i}) = i_{|u|} - i_1 + 1$.

Мы предполагаем, что на всех словах задан линейный порядок: всем словам меньшей длины предшествуют слова большей длины, а все слова одной длины упорядочены лексикографически. Тогда можно рассмотреть пространство признаков $F_n = \mathcal{R}^{\Xi^n}$ – множество всех векторов действительных чисел, индексами которых являются все слова длины n.

Определим отображение из множества всех слов в пространство признаков

$$\bar{\phi}^n(s) = (\phi_u^n(s) : u \in \Xi^n),$$

где

$$\phi_u^n(s) = \sum_{\bar{i}:u=s[\bar{i}]} \lambda^{l(\bar{i})}$$

при $0 < \lambda \leqslant 1$, которое представляет собой числа всех вхождений подпоследовательностей из n букв в последовательность s, взвешенные в соответствии с длинами этих вхождений в s.

Тогда соответствующее скалярное произведение вычисляется следующим образом:

$$\begin{aligned} K_n(s,t) &= \sum_{u\in\Xi^n} (\phi_u^n(s)\cdot\phi_u^n(t)) = \\ &= \sum_{u\in\Xi^n} \sum_{\bar{i}:u=s[\bar{i}]} \lambda^{l(\bar{i})} \sum_{\bar{j}:u=s[\bar{j}]} \lambda^{l(\bar{j})} = \\ &= \sum_{u\in\Xi^n} \sum_{\bar{i}:u=s[\bar{i}]} \sum_{\bar{j}:u=s[\bar{j}]} \lambda^{l(\bar{i})+l(\bar{j})}. \end{aligned}$$

При таком задании вычисление ядра $K_n(s,t)$ требует большого числа вычислительных операций. Существуют рекурсивные схемы для вычисления подобных сумм за приемлемое полиномиальное время (см. [24]).

Ядра такого типа используются при классификации текстов.

Более подробно о ядрах см. монографию Шолькопфа и Смолы [52].

Положительно определенные ядра.

Мы будем изучать ядра специального типа – положительно определенные ядра. По каждому такому ядру можно построить некоторое каноническое гильбертово пространство признаков. Предварительно рассмотрим пример из раздела 2.4.

Пусть функция $K(\bar{x},\bar{y}) = (\bar{\phi}(\bar{x})\cdot\bar{\phi}(\bar{x}))$ задана некоторым отображением $\bar{\phi}$ из евклидова пространства $\mathcal{R}^n$ в евклидово пространство признаков $\mathcal{R}^N$.

По определению функция $K(\bar{x},\bar{y})$ является симметричной, т.е. $K(\bar{x},\bar{y}) = K(\bar{y},\bar{x})$ для всех $\bar{x}$ и $\bar{y}$. Кроме этого, выполняется еще одно важное свойство: для любой последовательности элементов $\bar{x}_1, \ldots, \bar{x}_l$ и любой последовательности вещественных чисел

$\alpha_1, \ldots, \alpha_l$

$$\sum_{i,j=1}^{l} \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) = \sum_{i,j=1}^{l} \alpha_i \alpha_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) =$$

$$= \left(\sum_{i=1}^{l} \alpha_i \bar{\phi}(\bar{x}_i) \cdot \sum_{i=1}^{l} \alpha_i \bar{\phi}(\bar{x}_i) \right) = \left\| \sum_{i=1}^{l} \alpha_i \bar{\phi}(\bar{x}_i) \right\|^2 \geqslant 0. \qquad (2.29)$$

В общем случае сформулируем свойство (2.29) в качестве определения. Пусть X – произвольное множество. Функция $K : X \times X \to \mathcal{R}$ называется *положительно определенной*, если для любого набора элементов $x_1, \ldots, x_l$ и любого набора вещественных чисел $\alpha_1, \ldots, \alpha_l$ выполнено

$$\sum_{i,j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

Согласно (2.29) функция $K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}))$ является положительно определенной.

Матрица $(K(x_i, x_j)_{i,j=1}^{l}$ называется матрицей Грама.

Гильбертово пространство, порожденное воспроизводящим ядром. Гильбертово пространство $\mathcal{H}$ – это линейное векторное пространство над $\mathcal{R}$ с заданным скалярным произведением $(x \cdot y)$ на его элементах $x, y \in \mathcal{H}$ и полное относительно метрики, определяемой нормой $\|x\| = \sqrt{(x \cdot x)}$. Полнота означает, что каждая фундаментальная последовательность элементов $\mathcal{H}$ имеет предел.

Пусть X – некоторое множество и $\mathcal{F}$ – гильбертово пространство функций $f : X \to \mathcal{R}$.

Гильбертово пространство $\mathcal{F}$ называется *пространством, порожденным воспроизводящим ядром* (RKHS – Reproducing Kernel Hilbert Space), если для каждого $x \in X$ линейный функционал

$$f \to f(x)$$

является непрерывным. По теореме Рисса–Фишера каждый непрерывный линейный функционал на гильбертовом пространстве может быть представлен в виде скалярного произведения на некоторый элемент этого пространства. Поэтому в RKHS для каждого

x существует функция $K_x \in \mathcal{F}$ такая, что $f(x) = (f \cdot K_x)$ для всех $f \in \mathcal{F}$. Таким образом, определено отображение $\Phi : X \to \mathcal{F}$, где $\Phi(x) = K_x$ для всех x.

Имеет место свойство *воспроизводимости* – любая функция $f \in \mathcal{F}$ представлена в виде:

$$f(x) = (f \cdot \Phi(x))$$

для всех x. *Воспроизводящее ядро* определяется так:

$$K(x, y) = (\Phi(x) \cdot \Phi(y)) = (K_x \cdot K_y) = K_x(y) = (K_y \cdot K_x) = K_y(x).$$

Простейший пример RKHS можно построить с помощью отображения $\phi : \mathcal{R}^n \to \mathcal{R}^N$. Пусть $\mathcal{F}$ – пространство функций вида $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$, где $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$ и $(\bar{w} \cdot \bar{w}')$ – скалярное произведение в $\mathcal{R}^N$. Норма функции f определяется как $\|f\| = \|\bar{w}\|$, а скалярное произведение функций $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$ и $g(\bar{x}) = (\bar{w}' \cdot \phi(\bar{x}))$ определяется как $(f \cdot g) = (\bar{w} \cdot \bar{w}')$.

Тогда легко видеть, что отображение $f \to f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$ является непрерывным относительно такой нормы f. Воспроизводящее ядро есть $K(\bar{x}, \bar{y}) = (\phi(\bar{x}) \cdot \phi(\bar{y}))$. Кроме этого, из $f(\bar{y}) = (\bar{w} \cdot \phi(\bar{y}))$ и $K_{\bar{x}}(\bar{y}) = K(\bar{x}, \bar{y}) = (\phi(\bar{x}), \phi(\bar{y}))$ по определению скалярного произведения следует $(f \cdot K_{\bar{x}}) = (\bar{w} \cdot \phi(\bar{x})) = f(\bar{x})$.

Для произвольного RKHS имеет место неравенство $|f(x)| \leqslant \sqrt{(K(x, x))}\|f\|$. Действительно, по неравенству Коши–Буняковского имеем

$$|f(x)| = |(f \cdot K_x)| \leqslant \sqrt{(K_x \cdot K_x)}\|f\| = \sqrt{(K(x, x))}\|f\|.$$

Также выполнено неравенство

$$|f_1(x) - f_2(x)| \leqslant \sqrt{(K(x, x))}\|f_1 - f_2\|$$

для любых $f_1, f_2 \in \mathcal{F}$ и $x \in X$. Отсюда следует, что если последовательность функций f_n сходится к некоторой функции f по норме RKHS $\mathcal{F}$, то имеет место и поточечная сходимость: $f_n(x) \to f(x)$ для каждого $x \in X$.

С другой стороны, по каждой симметричной положительно определенной функции $K(x, y)$ можно определить некоторое *каноническое* функциональное гильбертово пространство $\mathcal{F}$, для которого она является воспроизводящим ядром.

Определим отображение $\Phi : X \to \mathcal{R}^X$ из множества X в множество всех функций из X в $\mathcal{R}$:

$$\Phi(x) = K(x, \cdot) = K_x.$$

По определению K_x – это функция, для которой $K_x(y) = K(x, y)$ для всех y. Определим линейное пространство $\mathcal{F}_1$ функций, порожденное всеми линейными комбинациями

$$f = \sum_{i=1}^{n} \alpha_i K_{x_i}, \tag{2.30}$$

где n, $\alpha_i \in \mathcal{R}$ и $x_i \in X$ – произвольные. Операции сумма и умножение на константу определяются стандартным образом. Определим скалярное произведение двух функций $f = \sum\limits_{i=1}^{n} \alpha_i K_{x_i}$ и $g = \sum\limits_{j=1}^{n'} \beta_j K_{x'_j}$ в виде

$$(f \cdot g) = \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j K(x_i, x'_j). \tag{2.31}$$

Легко проверить, что выражение (2.31) можно представить в виде $(f \cdot g) = \sum\limits_{j=1}^{n'} \beta_j f(x'_j)$ или $(f \cdot g) = \sum\limits_{i=1}^{n} \alpha_i g(x_i)$. Отсюда следует, что выражение (2.31) определено однозначно и не зависит от представления функций f и g в виде линейных комбинаций. Отсюда также следует, что функция $(f \cdot g)$ является билинейной по f и g. Она также симметричная: $(f \cdot g) = (g \cdot f)$ для всех $f, g \in \mathcal{F}_1$. Она также является положительно определенной.

Предварительно заметим, что

$$(f \cdot f) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

Учитывая это свойство, получаем, что для любого набора функций $f_1, \ldots, f_n$ и набора $\alpha_1, \ldots, \alpha_n \in \mathcal{R}$ коэффициентов будет выполнено условие положительной определенности функции $(f \cdot g)$:

$$\sum_{i,j=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (f_i \cdot f_j) = \left(\sum_{i=1}^{n} \alpha_i f_i \cdot \sum_{j=1}^{n} \alpha_j f_j \right) \geqslant 0.$$

По свойству (2.31) выполнено $f(x) = (f \cdot K_x)$ и, в частности, $(K_x \cdot K_y) = K(x, y)$ для всех x и y.

Из этих свойств и из свойства (2.30) также следует, что

$$|f(x)|^2 = |(f \cdot K_x)|^2 \leqslant K(x, x)(f \cdot f)$$

для $f \in \mathcal{F}_1$. В частности, из $(f \cdot f) = 0$ следует, что $f(x) = 0$ для всех x.

Аналогично

$$|f_1(x) - f_2(x)| \leqslant \sqrt{K(x, x)} \|f_1 - f_2\| \tag{2.32}$$

для $f_1, f_2 \in \mathcal{F}_1$.

Функция $\|f\| = \sqrt{(f \cdot f)}$ является нормой, так как она определена по скалярному произведению.

Построенное линейное нормированное пространство $\mathcal{F}_1$ всех линейных комбинаций (2.30) не является полным. Рассмотрим пополнение этого пространства относительно его нормы до полного метрического пространства $\mathcal{F}$. По свойству (2.32) для любой фундаментальной последовательности функций $f_1, f_2, \cdots \in \mathcal{F}_1$ и любого x выполнено

$$|f_n(x) - f_m(x)| \leqslant \sqrt{K(x, x)} \|f_n - f_m\|$$

для всех m, n. Поэтому для любого x последовательность чисел $f_1(x), f_2(x), \cdots \in \mathcal{R}$ также является фундаментальной и мы можем определить функцию $f(x) = \lim\limits_{n \to \infty} f_n(x)$ для всех $x \in X$. Таким образом, мы получаем расширение пространства $\mathcal{F}_1$ до пространства $\mathcal{F}$.

Также определим скалярное произведение на $\mathcal{F}$

$$(f \cdot g) = \lim_{n \to \infty} (f_n \cdot g_n),$$

где $f_1, f_2, \ldots$ и $g_1, g_2, \ldots$ – некоторые фундаментальные последовательности из $\mathcal{F}_1$, сходящиеся к $f, g \in \mathcal{F}$, соответственно.

Докажем, что предел существует и не зависит от выбора фундаментальных последовательностей $f_1, f_2, \ldots$ и $g_1, g_2, \ldots$. Пусть $f(x) = \lim\limits_{n\to\infty} f_n(x)$ и $g(x) = \lim\limits_{n\to\infty} g_n(x)$.

Докажем, что последовательность $\{(f_n \cdot g_m)\}_{n,m}$ – фундаментальная в $\mathcal{R}^1$. Имеем

$$|(f_{n_1} \cdot g_{m_1}) - (f_{n_2} \cdot g_{m_2})| \leqslant$$
$$|(f_{n_1} \cdot g_{m_1}) - (f_{n_1} \cdot g_{m_2})| + |(f_{n_1} \cdot g_{m_2}) - (f_{n_2} \cdot g_{m_2})| \leqslant$$
$$\|f_{n_1}\|_{\mathcal{F}_1} \cdot \|g_{m_1} - g_{m_2}\|_{\mathcal{F}_1} + \|g_{m_2}\|_{\mathcal{F}_1} \cdot \|f_{n_1} - f_{n_2}\|_{\mathcal{F}_1} \to 0$$

при $n_1, n_2, m_1, m_2 \to \infty$. В частности, $\lim\limits_{n\to\infty}(f_n \cdot g_n)$ также существует.

Докажем корректность определения скалярного произведения.

Лемма 2.3. *Пусть $f, g \in \mathcal{F}$. Тогда значение $(f \cdot g)$ не зависит от выбора последовательностей f_n и g_n.*

Доказательство. Пусть $f(x) = \lim\limits_{n\to\infty} f_n(x)$ и $g(x) = \lim\limits_{n\to\infty} g_n(x)$ для фундаментальных в $\mathcal{F}_1$ последовательностей $f_1, f_2, \ldots$ и $g_1, g_2, \ldots$. Пусть также $f(x) = \lim\limits_{n\to\infty} f'_n(x)$ и $g(x) = \lim\limits_{n\to\infty} g'_n(x)$ для фундаментальных в $\mathcal{F}_1$ последовательностей $f'_1, f'_2, \ldots$ и $g'_1, g'_2, \ldots$.

Докажем, что $\lim\limits_{n\to\infty}(f_n \cdot g_n) = \lim\limits_{n\to\infty}(f'_n \cdot g'_n)$. Пусть n – произвольное и $g_n = \sum\limits_{i=1}^{k} \alpha_i K_{z_i}$. Тогда

$$(f_m - f'_m \cdot g_n) = \left(f_m - f'_m \cdot \sum_{i=1}^{k} \alpha_i K_{z_i}\right) =$$
$$\sum_{i=1}^{k} \alpha_i (f_m \cdot K_{z_i}) - \alpha_i (f'_m \cdot K_{z_i}) = \sum_{i=1}^{k} \alpha_i (f_m(z_i) - f'_m(z_i) \to 0$$

при $m \to \infty$. Аналогично, $(g_n - g'_n \cdot f'_m) \to 0$ при $n \to \infty$.

Тогда

$$(f_n \cdot g_n) - (f'_m \cdot g'_m) =$$
$$(f_m - f'_m \cdot g_n) + (f'_m \cdot g_n - g'_n) \to 0$$

при $n, m \to \infty$. Это значит, что у этих последовательностей один и тот же предел. $\triangle$

Докажем, что пространство $\mathcal{F}$ является полным в норме, порожденной скалярным произведением. Предварительно докажем лемму.

Лемма 2.4. *Пусть $\{g_n\}$ – последовательность фундаментальная в $\mathcal{F}_1$ и $\lim_{n\to\infty} g_n(x) = g(x)$. Тогда $\|g - g_n\|_{\mathcal{F}} \to 0$.*

Доказательство. Из определения $\|g-g_n\|_{\mathcal{F}} = \sqrt{(g - g_n \cdot g - g_n)} = \sqrt{\lim_{m\to\infty}(g_m - g_n \cdot g_m - g_n)} = \sqrt{\lim_{m\to\infty} \|g_m - g_n\|_{\mathcal{F}_1}} \to 0$ при $n \to \infty$. $\triangle$

Докажем полноту пространства $\mathcal{F}$.

Теорема 2.3. *Всякая фундаментальная в пространстве $\mathcal{F}$ последовательность имеет предел в $\mathcal{F}$.*

Доказательство. Пусть $\{f_n\}$ – фундаментальная в $\mathcal{F}$, т.е. $\|f_n - f_m\|_{\mathcal{F}} \to 0$ при $n, m \to \infty$.

Для каждого n существует фундаментальная в $\mathcal{F}_1$ последовательность $g_{n,i} \in \mathcal{F}_1$ такая, что $f_n(x) = \lim_{i\to\infty} g_{n,i}(x)$. По демме 2.4 $\|f_n - g_{n,i}\|_{\mathcal{F}} \to 0$ при $i \to \infty$. Выберем $g_n = g_{n,i_n}$ так, чтобы $\|f_n - g_n\| \leqslant 2^{-n}$ для всех n. Тогда

$$\|g_n - g_m\|_{\mathcal{F}_1} = \|g_n - g_m\|_{\mathcal{F}} \leqslant$$
$$\|g_n - f_n\|_{\mathcal{F}} + \|g_m - f_m\|_{\mathcal{F}} + \|f_n - f_m\|_{\mathcal{F}} \to 0$$

при $n, m \to \infty$, так как $\|g_n - f_n\|_{\mathcal{F}} \leqslant 2^{-n}$ и $\|g_m - f_m\|_{\mathcal{F}} \leqslant 2^{-m}$, кроме того, $\|f_n - f_m\|_{\mathcal{F}} \to 0$ из фундаментальности.

Следовательно, последовательность $\{g_n\}$ – фундаментальнаяв $\mathcal{F}_1$ и поэтому для любого x существует предел $g(x) = \lim_{n\to\infty} g_n(x)$. Выполнено также

$$\|f_n - g\|_{\mathcal{F}} \leqslant \|f_n - g_n\|_{\mathcal{F}} + \|g_n - g\|_{\mathcal{F}} \to 0$$

при $n \to \infty$, т.е. g – предел последовательности $\{f_n\}$ по норме $\|\cdot\|_{\mathcal{F}}$. Здесь $\|f_n - g_n\|_{\mathcal{F}} \leqslant 2^{-n}$ и $\|g_n - g\|_{\mathcal{F}} \to 0$ по лемме 2.4. $\triangle$

Свойство воспроизводимости также имеет место из определения скалярного произведения в $\mathcal{F}$:

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} (f_n \cdot K_x) = (f \cdot K_x).$$

Более подробное изложение теории RKHS можно найти на вебстранице [38] а также в работе [10].

Полученное пространство $\mathcal{F}$ называется *каноническим гильбертовым пространством*, порожденным воспроизводящим ядром $K(x, y)$.

Другие свойства RKHS представлены в виде задач в разделе 2.12.

Гауссово ядро $K(\bar{x}, \bar{y}) = \exp(-\|\bar{x} - \bar{y}\|^2/\sigma^2)$ является положительно определенным и поэтому по нему можно определить каноническое гильбертово пространство RKHS и соответствующее отображение в это пространство.

Теорема о представителе. Теорема о представителе (Representer theorem) показывает, что решения широкого класса оптимизационных задач можно представить в виде линейных комбинаций значений ядер в точках обучающей выборки. Эта теорема была доказана Кимельдорфом и Вахбой [39]. См. также [52].

Теорема 2.4. *Пусть X – некоторое множество объектов и $S = ((x_1, y_1)\ , \ldots,\ (x_l, y_l))$ – обучающая выборка, где $(x_i, y_i) \in X \times \mathcal{R}$. Пусть $K(x, x')$ – положительно определенное ядро на $X \times X$ и $\mathcal{F}$ – соответствующее каноническое гильбертово пространство RKHS с нормой $\|\cdot\|$.*

Заданы также функция потерь $c : (X \times \mathcal{R}^2)^l \to \mathcal{R} \cup \{\infty\}$ и некоторая строго монотонно возрастающая функция Ω на множестве всех неотрицательных вещественных чисел.

Тогда любая функция $f \in \mathcal{F}$, минимизирующая регуляризованный риск функционал

$$c((x_1, y_1, f(x_1))\ , \ldots,\ (x_l, y_l, f(x_l))) + \Omega(\|f\|), \tag{2.33}$$

может быть представлена в виде

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) \tag{2.34}$$

для некоторых чисел $\alpha_1\ , \ldots,\ \alpha_l$.

Пример такого риск-функционала для задачи регрессии в пространстве признаков $f \in \mathcal{F}$:

$$c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_l, f(x_l))) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda \|f\|^2,$$

где $\lambda > 0$.

Доказательство. Напомним, что $K_x = K(x, \cdot)$ – функция, порожденная ядром. Любая функция $f \in \mathcal{F}$ представляется в виде $f(x) = (f \cdot K_x)$ для всех x.

Рассмотрим разложение линейного пространства $\mathcal{F}$ в прямую сумму конечномерного пространства, порожденного всеми линейными комбинациями функций K_{x_i}, $i = 1 \ , \ldots , \ l$, и его ортогонального дополнения. Тогда любая функция $f \in \mathcal{F}$ представляется в виде:

$$f = \sum_{i=1}^{l} \alpha_i K_{x_i} + f_*,$$

где $(f_* \cdot K_{x_i}) = 0$ для всех $i = 1 \ , \ldots , \ l$.

Вычислим значения $f(x_j)$ для всех $j = 1 \ , \ldots , \ l$:

$$f(x_j) = (f \cdot K_{x_j}) =$$
$$= \left(\left(\sum_{i=1}^{l} \alpha_i K_{x_i} + f_* \right) \cdot K_{x_j} \right) =$$
$$= \sum_{i=1}^{l} \alpha_i (K_{x_i} \cdot K_{x_j}).$$

Здесь важно, что значение функции $f(x_j)$ не зависит от элемента f_* из ортогонального дополнения. Таким образом, значение главной части $c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_l, f(x_l)))$ регуляризованного функционала (2.33) не зависит от f_*.

Так как f_* ортогонально элементу $\sum\limits_{i=1}^{l} \alpha_i K_{x_i}$ и функция Ω яв-

ляется строго монотонной, выполнено

$$\Omega(\|f\|) = \Omega\left(\|\sum_{i=1}^{l}\alpha_i K_{x_i} + f_*\|\right) =$$

$$= \Omega\left(\sqrt{\|\sum_{i=1}^{l}\alpha_i K_{x_i}\|^2 + \|f_*\|^2}\right) \geqslant$$

$$\geqslant \Omega\left(\|\sum_{i=1}^{l}\alpha_i (K_{x_i}\|\right),$$

причем равенство достигается тогда и только тогда, когда $f_* = 0$. Поэтому в точке минимума функционала (2.33) должно быть $f_* = 0$. Отсюда решение задачи минимизации функционала (2.33) должно иметь вид (2.34):

$$f(x) = \sum_{i=1}^{l}\alpha_i K_{x_i}.$$

Теорема доказана. $\triangle$

Теорема 2.4 показывает, что для решения задачи (2.33) функциональной минимизации в произвольном RKHS (которое может оказаться бесконечномерным) достаточно решить задачу минимизации в конечномерном пространстве $\mathcal{R}^n$.

Пример риск-функционала, соответствующего оптимизационной задаче SVM:

$$c((x_1, y_1, f(x_1)) \ , \ldots, \ (x_l, y_l, f(x_l))) =$$

$$= \frac{1}{\lambda}\sum_{i=1}^{l}\max\{0, 1 - y_i f(x_i)\} + \|f\|^2,$$

где $x_i \in \mathcal{R}^n$ и $y_i \in \{-1, +1\}$ при $i = 1 \ , \ldots, \ l$. Соответствующее пространство признаков $\mathcal{F}$ порождается ядром $K(x, x')$. Функция $f \in \mathcal{F}$, минимизирующая функционал (2.33), имеет вид

$$f = \sum_{i=1}^{l}\alpha_i K_{x_i}.$$

2.6. Случай неразделимой выборки

Предварительно приведем верхнюю оценку ошибки обобщения для случая, когда выборка не полностью разделена функцией классификации. Эта оценка послужит основой для постановки соответствующей оптимизационной задачи построения функции классификации.

2.6.1. Вектор переменных мягкого отступа

Рассмотрим теперь задачу классификации неразделимой выборки. Задачи такого типа характерны для практических приложений.

Задан класс $\mathcal{F}$ функций типа $\mathcal{X} \to \mathcal{R}$, с помощью которых производится классификация. Область определения $\mathcal{X}$ функций из $\mathcal{F}$ является подмножеством $\mathcal{R}^n$. По каждой функции $f \in \mathcal{F}$ определим индикаторную функцию классификации

$$h(\bar{x}) = \begin{cases} 1, \text{ если } f(\bar{x}) \geqslant 0, \\ -1 \text{ в противном случае.} \end{cases}$$

Задана выборка $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$. Пусть $\gamma_i = y_i f(\bar{x}_i)$ – граница ошибки примера $(\bar{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ относительно функции $f \in \mathcal{F}$. Заметим, что $\gamma_i > 0$ означает, что классификация с помощью функции f является правильной.

Распределение ошибок на выборке $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ определяется вектором $M_S(f) = (\gamma_1,\ \ldots,\ \gamma_l)$. Пусть

$$m_S(f) = \min_{i=1,\ldots,l} \gamma_i$$

– граница ошибки классификации выборки S посредством f. Величина $m_S(f) > 0$ тогда и только тогда, когда f строго разделяет S без ошибок.

Пусть $\gamma > 0$. Переменная мягкого отступа (margin slack variable) примера $(\bar{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ для пороговой функции f и границы ошибки γ определяется как

$$\xi_i = \max\{0, \gamma - y_i f(\bar{x}_i)\}.$$

Величина ξ_i равна превышению значения функции f по сравнению с заданной границей ошибки γ для примера $(\bar{x}_i, y_i)$. Мы назовем γ *функциональной границей* ошибки.

В рассматриваемой ранее оптимизационной задаче построения разделяющей гиперплоскости функциональная граница была равна 1 и она соответствовала геометрической границе $1/\|\bar{w}\|$, которая равна половине расстояния между граничными гиперплоскостями $(\bar{w} \cdot \bar{x}) + b = \pm 1$.

Заметим, что из $\xi_i > \gamma$ следует, что классификация примера $(\bar{x}_i, y_i)$ является ошибочной.

Вектор $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ называется вектором переменных мягкого отступа для выборки $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$.

По определению $y_i f(\bar{x}_i) + \xi_i \geqslant \gamma$ для всех i.

Роль вектора переменных мягкого отступа надо понимать следующим образом.

Если $\xi_i > \gamma$, то $y_i f(\bar{x}_i) < 0$, т.е. классификация примера $(\bar{x}_i, y_i)$ с помощью f является ошибочной. В этом случае величина ξ_i отражает степень удаленности примера $(\bar{x}_i, y_i)$ от разделяющей гиперплоскости – она тем больше, чем больше ошибка классификации.

Далее $\xi_i = 0$ тогда и только тогда, когда $y_i f(\bar{x}_i) \geqslant \gamma$; в этом случае классификация правильная и даже с некоторым запасом.

Случай $0 < \xi_i \leqslant \gamma$ является промежуточным, в этом случае классификация $0 < y_i f(\bar{x}_i) \leqslant \gamma$ – правильная, но с очень маленьким порогом, например, это может быть вследствие наличия шума в исходных данных.

В целом норма вектора ошибок $\bar{\xi}$ отражает размер ошибок классификации, а также роль шума в обучающей выборке. В дальнейшем величина $\|\bar{\xi}\|$ будет входить в верхние оценки вероятности неправильной классификации.

Если норма вектора $\bar{\xi}$ положительна, то выборка не разделима классификатором $f(\bar{x})$ с порогом $\gamma > 0$ и теорема 1.9 в этом случае прямо не применима. Однако в случае линейного классификатора можно сделать выборку разделимой, если перейти к эквивалентной задаче в пространстве большей размерности. Этот результат, принадлежащий Шо-Тейлору и Кристианини [50], представлен без

доказательства в следующей ниже теореме. Предполагаем, что примеры $(\bar{x}, y)$ генерируются некоторым непрерывным распределением вероятностей, при котором мера каждой такой пары равна 0.

Теорема 2.5. *Пусть $\mathcal{L}$ – класс всех линейных однородных функций вида $L(\bar{x}) = (\bar{w} \cdot \bar{x})$ с единичным весовым вектором $\|\bar{w}\| = 1$. Пусть $\gamma > 0$. Тогда для произвольного $\delta > 0$ и непрерывного распределения вероятностей P на $\mathcal{X} \times \{-1, 1\}$ с носителем внутри шара радиуса R и с центром в начале координат с вероятностью $1 - \delta$ произвольная функция $f \in \mathcal{L}$ с границей ошибки $\|\xi\|$ имеет на случайной выборке S длины l ошибку классификации* [1]

$$\mathrm{err}_P(f) \leqslant \frac{c}{l}\left(\frac{R^2 + \|\bar{\xi}\|^2}{\gamma^2}\log^2 l + \log\frac{2}{\delta}\right), \tag{2.35}$$

где $c > 0$ – константа.

2.6.2. Оптимизационные задачи для классификации с ошибками

Случай квадратичной нормы. В случае, когда выборка $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ не разделимая, рассматривается задача оптимизации с переменными мягкого отступа ξ_i, $i = 1,\ \ldots,\ l$.

Найдем векторы $\bar{w}$, $\bar{\xi}$ и число b так, чтобы

$$(\bar{w} \cdot \bar{w}) + C\sum_{i=1}^{l}\xi_i^2 \to \min, \tag{2.36}$$

$$y_i(\bar{w} \cdot \bar{x}_i) \geqslant 1 - \xi_i, \tag{2.37}$$

$$\xi_i \geqslant 0 \tag{2.38}$$

при $i = 1,\ \ldots,\ l$. Константа C определяет баланс между двумя частями функционала.

На практике константа C подбирается так, чтобы разделяющая гиперплоскость разделяла элементы обучающей выборки с

[1] Имеется в виду шар в пространстве $\mathcal{R}^n$, которому принадлежат классифицируемые элементы $\bar{x}_1,\ \ldots,\ \bar{x}_l$ выборки S.

минимальным значением нормы вектора разделяющих граничных переменных.

Заметим, что условие $\xi_i \geqslant 0$ можно опустить, так как оптимальное решение $\bar{w}$, ξ, b, где некоторые $\xi_i < 0$, является оптимальным и при $\xi_i = 0$.

Лагранжиан задачи (2.36) – (2.38) имеет вид

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + \frac{C}{2}\sum_{i=1}^{l} \xi_i^2 -$$

$$- \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i), \tag{2.39}$$

где $\alpha_i \geqslant 0$ – множители Лагранжа.

Соответствующая двойственная задача формулируется путем дифференцирования лагранжиана

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0}, \tag{2.40}$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{\xi}} = C\bar{\xi} - \bar{\alpha} = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0,$$

а также подстановкой этих соотношений в (2.39) :

$$W(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) +$$

$$+ \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}) - \frac{1}{C}(\bar{\alpha} \cdot \bar{\alpha}) =$$

$$= \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) - \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}). \tag{2.41}$$

Таким образом, мы должны максимизировать по $\bar{\alpha}$ величину

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j ((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C}\delta_{ij}) \tag{2.42}$$

при условиях $\alpha_i \geqslant 0$, $i = 1, 2, \ldots, l$, где $\delta_{ij} = 1$ при $i = j$ и $\delta_{ij} = 0$ при $i \neq j$. Соответствующие условия Каруша–Куна–Таккера имеют вид

$$\alpha_i(y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

при $i = 1, \ldots, l$.

Согласно (2.40) вектор весов выражается в виде линейной комбинации опорных векторов:

$$\bar{w} = \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i.$$

Из условий Каруша–Куна–Таккера следует, что $\alpha_i = 0$, если $y_i((\bar{w} \cdot \bar{x}_i) + b) > 1$, при этом $\xi_i = 0$. Эти векторы правильно классифицируются и лежат с внешней стороны относительно граничных гиперплоскостей. Опорными являются те векторы $\bar{x}_i$, для которых выполнено $\alpha_i = C\xi_i > 0$. Для них также выполнено $y_i((\bar{w} \cdot \bar{x}_i) + b) < 1$ и $\xi_i > 0$.

Сформулируем задачу оптимизации для пространства признаков, заданного некоторым ядром $K(\bar{x}_i, \bar{x}_j)$.

Теорема 2.6. *Даны пространство признаков, определенное ядром $K(\bar{x}_i, \bar{x}_j)$, и обучающая выборка $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$. Пусть вектор параметров $\bar{\alpha}$ является решением задачи оптимизации:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (K(\bar{x}_i, \bar{x}_j) + \frac{1}{C}\delta_{ij}) \to \max \tag{2.43}$$

$$\textit{при условиях} \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$\alpha_i \geqslant 0 \; i = 1, \ldots, l. \tag{2.44}$$

Тогда соответствующая разделяющая поверхность имеет вид

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\bar{x}_i, \bar{x}) + b,$$

где b находится из условия $y_i f(\bar{x}_i) = 1 - \alpha_i/C$ для произвольного i такого, что $\alpha_i \neq 0$.

Функция классификации $\mathrm{sign}(f(\bar{x}))$ *разделяет элементы выборки так же, как соответствующая гиперплоскость, полученная в результате решения задачи оптимизации* (2.36) – (2.38) *в пространстве признаков, определенном ядром $K(\bar{x}, \bar{z})$, где переменные мягкого отступа определяются для геометрического края ошибки:*

$$\gamma = \left(\sum_{j \in sv} \alpha_j - \frac{1}{C}(\bar{\alpha} \cdot \bar{\alpha}) \right)^{-1/2}.$$

Для вычисления b^* используем равенства $\alpha_i = C\xi_i$, а также условия Каруша–Куна–Таккера:

$$\alpha_i(y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

при $i = 1, \ \ldots, \ l$.

Геометрический край ошибки $\rho(\bar{w}) = \frac{1}{|\bar{w}|}$, равный половине расстояния между гиперплоскостями $(\bar{w} \cdot \bar{x}_i) + b) = \pm 1$, в ядерном случае определяется следующим образом:

$$\begin{aligned}
(\bar{w} \cdot \bar{w}) &= \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} y_j \alpha_j \sum_{i \in sv} y_i \alpha_i K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} \alpha_j (1 - \xi_j - y_j b) = \\
&= \sum_{j \in sv} \alpha_j - \sum_{j \in sv} \alpha_j \xi_j = \\
&= \sum_{j \in sv} \alpha_j - \frac{1}{C}(\bar{\alpha} \cdot \bar{\alpha}).
\end{aligned}$$

Верхняя оценка (2.35) вероятности ошибки классификации при обобщении не зависит от размерности пространства, что позволяет применять ядра $K(\bar{x}, \bar{z})$, порождающие пространства при-

знаков высокой размерности. Увеличение размерности пространства признаков приводит к разделению обучающей выборки гиперплоскостью с меньшей нормой вектора ξ, что уменьшает вероятность ошибки классификации при обобщении.

Применим теорему 2.5. Оценка вероятности ошибки (2.35) имеет место для пороговых линейных функций с единичным весовым вектором $\bar{w}$. Для того чтобы ее применить к задаче (2.36) – (2.38), поделим обе части неравенства (2.37) на $\|\bar{w}\|$, где $\bar{w}$ – оптимальное решение задачи. Тогда в (2.35) надо взять в качестве ξ_i величину $\xi_i/\|\bar{w}\|$, а $\gamma = 1/\|\bar{w}\|$. Получим новую версию вероятности ошибки (2.35) :

$$\mathrm{err}_P(f) \leqslant \frac{c}{l}((\|\bar{w}\|^2 R^2 + \|\bar{\xi}\|^2)\log^2 l + \log\frac{2}{\delta}). \tag{2.45}$$

Неравенство (2.45) показывает, что для минимизации верхней оценки вероятности ошибки обобщения нам действительно необходимо минимизировать величину (2.36).

Случай линейной нормы. На практике чаще рассматривается задача оптимизации, в которой вместо квадратичной нормы вектора переменных мягкого отступа $\bar{\xi}$ используется линейная норма. В этом случае формулируется следующая задача оптимизации.

Находим векторы $\bar{w}$, $\bar{\xi}$ и число b так, чтобы

$$(\bar{w}\cdot\bar{w}) + C\sum_{i=1}^{l}\xi_i \to \min, \tag{2.46}$$

$$y_i((\bar{w}\cdot\bar{x}_i) + b) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0 \tag{2.47}$$

при $i = 1, \ldots, l$. Константа C определяет баланс между двумя частями функционала.

Соответствующий лагранжиан имеет вид

$$W(\bar{w}, b, \xi, \alpha, \bar{r}) = \frac{1}{2}(\bar{w}\cdot\bar{w}) + C\sum_{i=1}^{l}\xi_i -$$

$$-\sum_{i=1}^{l}\alpha_i(y_i((\bar{w}\cdot\bar{x}_i) + b) - 1 + \xi_i) - \sum_{i=1}^{l} r_i\xi_i,$$

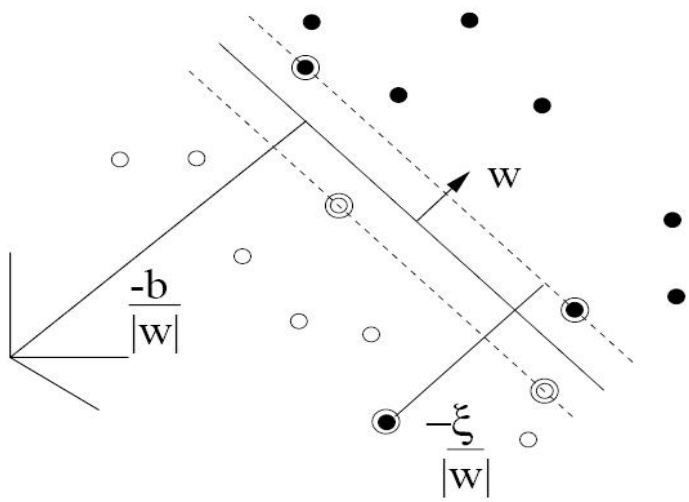

Рис. 1.2. Опорные векторы расположены на граничных гиперплоскостях или же неправильно ими классифицируются (случай линейной нормы)

где $\alpha_i \geqslant 0$, $r_i \geqslant 0$ при $i = 1, \ \ldots, \ l$.

Соответственная двойственная задача получается путем приравнивания к нулю производных:

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial \xi_i} = C - \alpha_i - r_i = 0,$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0.$$

Подставляем решения этих уравнений в прямую задачу и получаем двойственное представление задачи в виде задачи максими-

зации функционала:

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}, \bar{r}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j)$$

при $\alpha_i \geqslant 0$, $i = 1, \ldots, l$, который совпадает с функционалом для случая разделения без ошибок. Отсюда следует, что весовой вектор разделяющей гиперплоскости однозначно определяется по решению этой задачи в виде $\bar{w} = \sum\limits_{i=1}^{l} \alpha_i y_i \bar{x}_i$.

Отличие от задачи с квадратичной нормой заключается в том, что условие $C - \alpha_i - r_i = 0$ вместе с условием $r_i \geqslant 0$ вынуждает неравенство $\alpha_i \leqslant C$. В то же время $\xi_i > 0$ выполнено только при $r_i = 0$. Отсюда следует, что $\alpha_i = C$ для всех таких i. Таким образом, условия Каруша–Куна–Таккера имеют вид

$$\alpha_i(y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0, \ i = 1, \ \ldots, \ l,$$
$$\xi_i(\alpha_i - C) = 0, \ i = 1, \ \ldots, \ l.$$

Согласно этим условиям переменная мягкого отступа ξ_i отлична от нуля только при $\alpha_i = C$.

Из условий Каруша–Куна–Таккера и граничных условий оптимизационной задачи следует, что

- если $y_i((\bar{w} \cdot \bar{x}_i) + b) > 1$, то $\alpha_i = 0$ и $\xi_i = 0$ (классификация правильная);
- если $y_i((\bar{w} \cdot \bar{x}_i) + b) < 1$, то $\xi_i > 0$ и $\alpha_i = C$ (ошибка классификация, вектор является опорным);
- если $y_i((\bar{w} \cdot \bar{x}_i) + b) = 1$, то $\xi_i = 0$ и $0 \leqslant \alpha_i \leqslant C$.

Можно показать, что в оптимизационной задаче разделения с ошибками в линейной норме для любой нетривиальной выборки $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$ (т.е. такой, что $y_i = 1$ и $y_j = -1$ для каких-нибудь $1 \leqslant i, j \leqslant l$) обязательно найдутся число b и вектор $\bar{x}_i$, лежащий на одной из граничных гиперплоскостей: $y_i((\bar{w} \cdot \bar{x}_i) + b) = 1$, для которого $\alpha_i > 0$ (см. задачу из раздела 2.12).

Опорные векторы – это те векторы $\bar{x}_i$, где $\alpha_i > 0$ Это те векторы, которые неправильно классифицируются или лежат на граничных гиперплоскостях. Легко видеть, что расстояние от такого вектора до соответствующей граничной гиперплоскости равно $-\frac{\xi_i}{\|\bar{w}\|}$ (см. рис. 1.2).

Таким образом, для произвольного ядра получаем следующее утверждение.

Теорема 2.7. *Даны обучающая выборка* $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$ *и пространство признаков, определенное ядром* $K(\bar{x}_i, \bar{x}_j)$. *Пусть вектор параметров* $\bar{\alpha}$ *является решением задачи оптимизации:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) \to \max \tag{2.48}$$

$$\text{при условиях } \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$C \geqslant \alpha_i \geqslant 0 \ i = 1, \ \ldots, \ l. \tag{2.49}$$

Тогда соответствующая разделяющая поверхность имеет вид

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\bar{x}_i, \bar{x}) + b,$$

где b *находится из условия* $y_i f(\bar{x}_i) = 1$ *для произвольного* i *такого, что вектор* $\bar{x}_i$ *находится на граничной гиперплоскости и* $\alpha_i > 0$.

Соответствующая функция классификации $\operatorname{sign}(f(\bar{x}))$ *разделяет элементы выборки так же, как соответствующая гиперплоскость, полученная в результате решения задачи оптимизации* (2.46) – (2.47) *в пространстве признаков, определенном ядром* $K(\bar{x}, \bar{z})$, *где переменные мягкого отступа определены для геометрического края ошибки:*

$$\gamma = \left(\sum_{i,j} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) \right)^{-1/2}.$$

Таким образом, задача оптимизации (2.46) – (2.47) эквивалентна задаче оптимизации (2.36) – (2.38) с одним дополнительным условием, что $\alpha_i \leqslant C$. По этой причине эти ограничения называются квадратными (box constraints), так как они требуют, чтобы каждое α_i находилось внутри квадрата со стороной C, расположенного в положительном октанте.

Параметр C контролирует соотношение между точностью регуляризации и величиной коэффициентов α_i. В частности, чем меньше параметр C, тем меньше значения α_i, т.е. меньше влияние примеров, находящихся далеко от разделяющей гиперплоскости.

Задача для классификации с ошибками в форме задачи линейного программирования. Можно сформулировать предыдущую задачу как задачу линейного программирования, в которой вместо квадратичной нормы вектора $\bar{w}$ минимизируется сумма коэффициентов α_i, которые характеризуют степень участия примеров в построении разделяющей гиперплоскости.

Оценка вероятности ошибки обобщения через число опорных векторов, приведенная в теореме 2.2 , может служить обоснованием применимости этого метода.

В этом случае рассматривается задача оптимизации:

$$\sum_{i=1}^{l} \alpha_i + C \sum_{i=1}^{l} \xi_i \to \min$$

$$\text{при условиях } y_i \left(\sum_{j=1}^{l} \alpha_i (\bar{x}_i \cdot \bar{x}_j) + b \right) \geqslant 1 - \xi_i,$$

$$\alpha_i \geqslant 0, \ \xi_i \geqslant 0,$$

где $i = 1, \ \ldots, \ l$. Константа C определяет баланс между двумя частями функционала.

Преимущество данной постановки заключается в том, что здесь решается задача линейного программирования вместо задачи квадратичного программирования.

2.7. Среднее по Радемахеру и оценка ошибки классификации

В этом разделе будут получены оценки вероятности ошибки обобщения для функции классификации, заданной пороговой функцией из некоторого RKHS. [2]

Пусть $\mathcal{F}$ – гильбертово пространство функций, определенных на некотором множестве $\mathcal{X}$. Мы также предполагаем, что это пространство является пространством, порожденным воспроизводящим ядром (RKHS), $K(x, y)$ – соответствующее ядро. Произвольная функция $f \in \mathcal{F}$ представляется в виде скалярного произведения $f(x) = (f \cdot \phi(x))$, где $\phi(x) = K(x, \cdot)$.

Важный для метода опорных векторов пример такого RKHS можно построить с помощью отображения $\phi : \mathcal{R}^n \to \mathcal{R}^N$. Пусть $\mathcal{F}$ – пространство функций вида $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$, где $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$ и $(\bar{w} \cdot \bar{w}')$ – скалярное произведение в $\mathcal{R}^N$. Норма функции f определяется как $\|f\| = \|\bar{w}\|$, а скалярное произведение функций f и $g(\bar{x}) = (\bar{w}' \cdot \phi(\bar{x}))$ определяется как $(f \cdot g) = (\bar{w} \cdot \bar{w}')$. Функция $K(x, y) = (\phi(x) \cdot \phi(y))$ является соответствующим ядром.

По каждой функции $f \in \mathcal{F}$ определим индикаторную функцию классификации

$$h(x) = \begin{cases} 1, \text{ если } f(x) \geqslant 0, \\ -1 \text{ в противном случае.} \end{cases}$$

Пусть $\mathcal{F}_1 = \{f \in \mathcal{F} : \|f\| \leqslant 1\}$. В приведенном выше примере $\mathcal{F}_1$ – класс функций $f(x) = (\bar{w} \cdot \phi(\bar{x}))$ с ограниченным весовым вектором $\|\bar{w}\| \leqslant 1$.

Задана обучающая выборка $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, где $x_i \in \mathcal{X}$ и $y_i \in \{-1, 1\}$.

Пусть $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^{l}$ – матрица Грама значений ядра на элементах для выборки S; $tr(\mathbf{K}) = \sum_{i=1}^{l} K(x_i, x_i)$ – след матрицы $\mathbf{K}$.

Приведем оценку выборочного среднего Радемахера для класса $\mathcal{F}_1$ относительно обучающей выборки S.

[2] Материал данного раздела использует результаты из монографии Шо-Тейлора и Кристианини [51].

Теорема 2.8. *Выборочное среднее Радемахера класса $\mathcal{F}_1$ относительно обучающей выборки $S = ((x_1, y_1),\ \ldots,\ (x_l, y_l))$ удовлетворяет неравенству*

$$\tilde{\mathcal{R}}_l(\mathcal{F}_1) \leqslant \frac{1}{l}\sqrt{tr(\mathbf{K})}. \tag{2.50}$$

Доказательство. Имеет место следующая цепочка равенств и неравенств:

$$\begin{aligned}
\tilde{\mathcal{R}}_l(\mathcal{F}_1) &= E_\sigma\left(\sup_{f\in\mathcal{F}_1}\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right) = \\
&= E_\sigma\left(\sup_{\|f\|\leqslant 1}\left(f\cdot\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi(x_i)\right)\right) \leqslant \\
&\leqslant \frac{1}{l}E_\sigma\left(\left\|\sum_{i=1}^{l}\sigma_i\phi(x_i)\right\|\right) = \\
&= \frac{1}{l}E_\sigma\left(\left(\sum_{i=1}^{l}\sigma_i\phi(x_i)\cdot\sum_{i=1}^{l}\sigma_i\phi(x_i)\right)^{1/2}\right) \leqslant \\
&\leqslant \frac{1}{l}\left(E_\sigma\left(\sum_{i,j=1}^{l}\sigma_i\sigma_j K(x_i,x_j)\right)\right)^{1/2} = \\
&= \frac{1}{l}\left(\sum_{i=1}^{l}K(x_i,x_i)\right)^{1/2}.
\end{aligned}$$

Здесь при переходе от 2-й строки к 3-й мы использовали неравенство Коши–Буняковского, при переходе от 3-й строки к 4-й было использовано определение нормы вектора. При переходе от 4-й строки к 5-й было использовано неравенство Йенсена, при переходе от 5-й строки к 6-й мы использовали независимость случайных величин σ_i, а также $E(\sigma_i^2) = 1$ и $E(\sigma_i\sigma_j) = E(\sigma_i)E(\sigma_j) = 0$ при $i \neq j$. Теорема доказана. $\triangle$

Напомним, что число $\gamma_i = y_i f(x_i)$ называется границей ошибки примера $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ относительно функции $f \in \mathcal{F}$.

Заметим, что $\gamma_i > 0$ означает, что классификация с помощью функции f является правильной.

Задана выборка $S = ((x_1, y_1), \ldots, (x_l, y_l))$. Пусть задано число $\gamma > 0$. Переменная мягкого отступа примера (x_i, y_i) для пороговой функции f и границы ошибки γ определяется как

$$\xi_i = \max\{0, \gamma - y_i f(x_i)\}. \tag{2.51}$$

Вектор $\bar{\xi} = (\xi_1, \ldots, \xi_l)$ называется вектором переменных мягкого отступа для выборки $S = ((x_1, y_1), \ldots, (x_l, y_l))$.

Определим вспомогательную функцию $f(x, y) = -yf(x)$. Обозначим класс функций от двух переменных, определенных на $\mathcal{X} \times \{-1, 1\}$:

$$\mathcal{F}_2 = \{f(x, y) : f(x, y) = -yf(x), f \in \mathcal{F}_1\}.$$

Пусть

$$\chi(x) = \begin{cases} 1, \text{ если } x \geqslant 0, \\ 0 \text{ в противном случае.} \end{cases}$$

Предполагаем, что элементы выборки S генерируются независимо друг от друга с помощью вероятностного распределения P. Легко проверить, что

$$P\{(x, y) : y \neq \operatorname{sign}(f(x))\} \leqslant E_P(\chi(-yf(x))) = E_P(\chi(f(x, y))).$$

Пусть $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ – матрица значений ядра на элементах для выборки S.

В следующей теореме дается верхняя оценка ошибки обобщения классификатора из класса $\mathcal{F}_1$.

Теорема 2.9. *Для произвольного $\delta > 0$ с вероятностью $1 - \delta$ для любой функции $f \in \mathcal{F}_1$ выполнено*

$$P\{y \neq \operatorname{sign}(f(x))\} \leqslant \frac{1}{l\gamma} \sum_{i=1}^{l} \xi_i + \frac{2}{l\gamma} \sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{1}{2l} \ln \frac{2}{\delta}}. \tag{2.52}$$

Заметим, что правая часть неравенства (2.52) является случайной величиной. Матрица $\mathbf{K}$ построена по случайной выборке

S. Переменные мягкого отступа ξ_i также зависят от элементов выборки x_i и поэтому также являются случайными величинами.

Доказательство. Напомним, что $\gamma > 0$ – порог функции классификации. Определим вспомогательную функцию $g : \mathcal{R} \to [0, 1]$:

$$g(r) = \begin{cases} 1, \text{ если } r > 0, \\ 1 + r/\gamma, \text{ если } -\gamma \leqslant r \leqslant 0, \\ 0 \text{ иначе.} \end{cases}$$

Из определения этой функции следует, что $g(r) \geqslant \chi(r)$. Отсюда и по следствию 1.6 с вероятностью $1 - \delta$ выполнено

$$E_P(\chi(f(x, y))) \leqslant E_P(g(f(x, y))) \leqslant$$
$$\leqslant \tilde{E}_S(g \circ f) + 2\tilde{\mathcal{R}}_l(g \circ \mathcal{F}_2) + 3\sqrt{\frac{1}{2l} \ln \frac{2}{\delta}}. \tag{2.53}$$

По определению переменной мягкого отступа (2.51) выполнено

$$\frac{\xi_i}{\gamma} = \begin{cases} 1 - \frac{y_i f(x_i)}{\gamma}), \text{ если } \frac{y_i f(x_i)}{\gamma}) < 1, \\ 0 \text{ в противном случае.} \end{cases}$$

Отсюда получаем неравенство

$$g(-y_i f(x_i)) \leqslant \xi_i/\gamma$$

при $1 \leqslant i \leqslant l$.

Оценим выборочное среднее Радемахера класса функций $\mathcal{F}_2$:

$$\tilde{\mathcal{R}}_l(\mathcal{F}_2) = E_\sigma \left(\sup_{f \in \mathcal{F}_2} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(x_i, y_i) \right) =$$
$$= E_\sigma \left(\sup_{f \in \mathcal{F}_1} \frac{1}{l} \sum_{i=1}^{l} \sigma_i y_i f(x_i) \right) =$$
$$= E_\sigma \left(\sup_{f \in \mathcal{F}_1} \frac{1}{l} \sum_{i=1}^{l} \sigma_i f(x_i) \right) =$$
$$= \tilde{\mathcal{R}}_l(\mathcal{F}_1) \leqslant \frac{1}{l} \sqrt{tr(K)}.$$

Так как функция g удовлетворяет условию Липшица с $L = 1/\gamma$, по теореме 1.12 имеем

$$\tilde{\mathcal{R}}_l(g \circ \mathcal{F}_2) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F}_2)/\gamma = \tilde{\mathcal{R}}_l(\mathcal{F}_1)/\gamma.$$

По определению для $f \in \mathcal{F}_2$ будет

$$\tilde{E}_S(g \circ f)) = \frac{1}{l}\sum_{i=1}^{l} g(-y_i f(\bar{x}_i)) \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l} \xi_i.$$

Отсюда и из неравенств (2.53) и (2.50) теоремы 2.8 следует, что с вероятностью $1-\delta$

$$E_P(\chi(f(x,y))) \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l} \xi_i + \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{1}{2l}\ln\frac{2}{\delta}}. \qquad (2.54)$$

Теорема доказана. $\triangle$

В частности, если функция $f(x)$ разделяет выборку S без ошибок, то имеет место оценка:

Следствие 2.1. *Для произвольного $\delta > 0$ с вероятностью $1-\delta$ для любой функции $f \in \mathcal{F}_1$, разделяющей выборку S с порогом γ без ошибок, выполнено*

$$P\{y \neq \operatorname{sign}(f(x))\} \leqslant \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}.$$

Рассмотрим полученные оценки для случая семейства функций вида $f(\bar{x}) = (\bar{w} \cdot \phi(\bar{x}))$, где $\bar{x} \in \mathcal{R}^n$, $\bar{w} \in \mathcal{R}^N$. В отличие от оценок (1.36) и (2.35), полученных в теории пороговой размерности, оценка (2.52) имеет лучшие константы и не требует предварительного знания радиуса шара, в котором находятся векторы обучающей выборки – вместо этого в оценке присутствует след матрицы Грама.

С другой стороны, если радиус шара, содержащего векторы выборки известен, то оценка (2.52) по порядку уступает аналогичным оценкам, полученным с помощью пороговой размерности.

Пусть $\|\bar{x}_i\| \leqslant R$ для всех $1 \leqslant i \leqslant l$. Тогда для малых значений порядок величины

$$\frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} \leqslant \frac{2}{l\gamma}\sqrt{lR^2} = 2\sqrt{\frac{R^2}{l\gamma^2}}$$

значительно превышает порядок $O\left(\frac{R^2}{l\gamma^2}\right)$ главного члена оценки (1.36) из теоремы 1.9 и порядок главного члена оценки (2.35) теоремы 2.5.

2.8. Задача многомерной регрессии

2.8.1. Простая линейная регрессия

Пусть задана обучающая выборка $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{R}^n$, $y_i \in \mathcal{R}$ при $i = 1,\ \ldots,\ l$.

Задача линейной регрессии заключается в нахождении линейной функции

$$f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b,$$

наилучшим образом интерполирующей элементы выборки S. Геометрически данная функция представляет собой гиперплоскость, которая приближает точки y_i на аргументах $\bar{x}_i$ при $i = 1,\ \ldots,\ l$. Далее называем $f(\bar{x})$ функцией регрессии.

Данная задача была решена Гауссом и Лежандром еще в XVIII веке при помощи минимизации суммы квадратов разностей значений функции $f(\bar{x}_i)$ и точек y_i при $i = 1,\ \ldots,\ l$. Теория обобщения для данного метода хорошо представлена в математической статистике для случая линейной модели генерации данных с гауссовским случайным шумом.

В дальнейшем произвольный вектор $\bar{x}$ также будет рассмат-

риваться как матрица типа $n \times 1$, т.е. как вектор-столбец

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}.$$

Этот же вектор, записанный в виде строки $(x_1, \ \ldots, \ x_n)$, т.е. в виде транспонированной матрицы типа $1 \times n$, будет записываться как $\bar{x}'$. Произведение двух матриц A и B обозначается AB без точки между ними. Мы часто будем отождествлять скалярное произведение векторов $(\bar{x} \cdot \bar{z})$ и матрицу $\bar{x}'\bar{z}$:

$$\bar{x}'\bar{z} = (x_1, \ \ldots, \ x_n) \begin{pmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{pmatrix} = (x_1 z_1 + \cdots + x_n z_n)$$

типа (1×1) с единственным элементом, равным этому скалярному произведению.

Согласно методу наименьших квадратов, минимизируем *квадратичную функцию потерь*:

$$L(\bar{w}, b) = \sum_{i=1}^{l} (y_i - (\bar{w} \cdot \bar{x}_i) - b)^2. \tag{2.55}$$

Обозначим через $\tilde{w}$ расширенный вектор-столбец весовых коэффициентов и свободного члена

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix}.$$

Аналогичным образом, обозначим через $\tilde{x}$ расширенный вектор-столбец переменных

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ 1 \end{pmatrix}.$$

В новых расширенных переменных функция регрессии имеет однородный вид без свободного члена:

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}). \tag{2.56}$$

Рассмотрим матрицу типа $l \times (n+1)$, строками которой являются расширенные вектор-строки $\tilde{x}_i' = (\bar{x}_i', 1)$ переменных:

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \tilde{x}_l' \end{pmatrix} = \begin{pmatrix} x_{11}, \ \ldots, \ x_{1n}, 1 \\ x_{21}, \ \ldots, \ x_{2n}, 1 \\ \cdot \\ \cdot \\ \cdot \\ x_{l1}, \ \ldots, \ x_{ln}, 1 \end{pmatrix}.$$

Вводим также l-мерный вектор-столбец значений интерполяции

$$\bar{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_l \end{pmatrix}.$$

Разности $|f(\bar{x}_i) - y_i|$ (а также $y_i - f(\bar{x}_i)$ и $f(\bar{x}_i) - y_i$) называются *остатками*. Вектор-столбец остатков имеет вид $\bar{y} - (\tilde{X} \cdot \tilde{w})$, а функционал (2.55) можно записать в матричном виде как квадрат нормы вектор-столбца остатков:

$$L(\tilde{w}) = \|\tilde{X}\tilde{w} - \bar{y}\|^2 = (\bar{y} - \tilde{X}\tilde{w})'(\bar{y} - \tilde{X}\tilde{w}).$$

Здесь и далее для произвольной матрицы A посредством A' обозначаем транспонированную матрицу A.

Теперь задача регрессии может быть записана в виде задачи минимизации квадрата нормы вектора остатков:

$$L(\tilde{w}) = \|\tilde{X}\tilde{w} - \bar{y}\|^2 \to \min. \tag{2.57}$$

Геометрически это может интерпретироваться так же, как поиск проекции наименьшей длины вектора $\bar{y}$ на подпространство (гиперплоскость), порожденное вектор-столбцами матрицы $\tilde{X}$.

Для поиска минимума приравниваем частные производные этого функционала (по переменным $w_1, \ \ldots, \ w_n, b$) к нулю. Получим систему из $n+1$ уравнений

$$\frac{\partial L(\tilde{w})}{\partial \tilde{w}} = -2\tilde{X}'\bar{y} + 2\tilde{X}'\tilde{X}\tilde{w} = \bar{0}.$$

Преобразуем эту систему к виду

$$\tilde{X}'\tilde{X}\tilde{w} = \tilde{X}'\bar{y}.$$

Если матрица $X'\tilde{X}$ обратима, получаем решение этой системы:

$$\tilde{w} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\bar{y}.$$

2.8.2. Гребневая регрессия

Другой метод, обеспечивающий численную устойчивость, – *гребневая регрессия (ridge regression)*, был рассмотрен Хоэрлом и Кеннардом.

Напомним, что для того, чтобы избавиться от свободного члена в уравнении регрессии, мы рассматриваем задачу регрессии с весовой переменной $\tilde{w}$ – расширенным вектор-столбцом весовых коэффициентов и свободного члена

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix},$$

а также с расширенным вектор-столбцом переменных

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ 1 \end{pmatrix}.$$

В новых переменных функция регрессии имеет однородный вид без свободного члена

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}).$$

Рассматривается функция потерь следующего вида:

$$L(\tilde{w}) = \lambda(\tilde{w} \cdot \tilde{w}) + \sum_{i=1}^{l} (y_i - (\tilde{w} \cdot \tilde{x}_i))^2 = \\ = \lambda \|\tilde{w}\|^2 + \|\tilde{X}\tilde{w} - \bar{y}\|^2. \qquad (2.58)$$

Параметр λ контролирует баланс между квадратичными потерями и нормой весового вектора. Норма весового вектора отражает сложность регрессионной гипотезы. Обсуждение роли параметра λ см. ниже.

Решение задачи гребневой регрессии в прямой форме. Для нахождения экстремума приравниваем к нулю частные производные $L(\tilde{w})$ по w_i, $i = 1, \ldots, n+1$:

$$\lambda\tilde{w} - \sum_{i=1}^{l} ((y_i - (\tilde{w} \cdot \tilde{x}_i))\tilde{x}_i = \tilde{0}.$$

В матричной форме это уравнение имеет вид

$$\lambda\tilde{w} - \tilde{X}'\tilde{y} + \tilde{X}'\tilde{X}\tilde{w} = \tilde{0}.$$

Решение записывается в матричной форме:

$$\tilde{w} = (\lambda I + \tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y},$$

где I – единичная матрица.

Матрицы $\tilde{X}'\tilde{X}$, I и $\lambda I + \tilde{X}'\tilde{X}$ имеют размер $(n+1) \times (n+1)$. Матрица $\tilde{X}'\tilde{X}$ является положительно определенной, т.е.

$$\tilde{z}'(\tilde{X}'\tilde{X})\tilde{z} \geqslant 0$$

для любого вектора $\tilde{z}$. Это следует из равенства

$$\tilde{z}'(\tilde{X}'\tilde{X})\tilde{z} = (\tilde{X}\tilde{z})'(\tilde{X}\tilde{z}) = \|\tilde{X}\tilde{z}\|^2 \geqslant 0.$$

При добавлении к матрице $\tilde{X}'\tilde{X}$ матрицы λI при $\lambda > 0$ новая матрица становится строго положительно определенной

$$\tilde{z}'(\lambda I + \tilde{X}'\tilde{X})\tilde{z} = \lambda\|\tilde{z}\|^2 + \|\tilde{X}\tilde{z}\|^2 > 0$$

при $\tilde{z} \neq \tilde{0}$. Известно, что любая положительно определенная матрица обратима. Поэтому решение задачи гребневой регрессии всегда существует при $\lambda > 0$.

При $\lambda = 0$ матрица $\tilde{X}'\tilde{X}$ может оказаться необратимой. В этом случае решение задачи регрессии не является единственным. Поэтому гребневая регрессия при $\lambda > 0$ численно существенно проще, чем простая регрессия. Кроме того, параметр λ играет роль штрафа за большую норму вектора весов $\tilde{w}$.

Если λ приближается к нулю, матрица $\lambda I + \tilde{X}'\tilde{X}$ может становиться все ближе к необратимой. В этом случае алгоритм обращения этой матрицы становится все более нестабильным. Большие значения λ делают процесс вычисления обратной матрицы более стабильным.

С другой стороны, при больших значениях λ матрица λI начинает преобладать над матрицей $\tilde{X}'\tilde{X}$ и поэтому остатки регрессии становятся большими и найденное уравнение регрессии теряет свои предсказательные возможности. Поэтому значение λ должно иметь тот же порядок, как и элементы матрицы $\tilde{X}'\tilde{X}$.

Задача гребневой регрессии в двойственной форме и ее обобщение на нелинейный случай будут рассмотрены в разделе 2.9.2.

2.9. Регрессия с опорными векторами

2.9.1. Решение задачи регрессии с помощью SVM

Метод опорных векторов также применяется к задаче регрессии. При этом так же, как и в задаче классификации, нелинейным разделяющим функциям соответствуют линейные разделяющие функции в пространстве признаков, т.е. применяется техника ядер.

Линейная ϵ-нечувствительная функция потерь – это функция вида

$$L^{\epsilon}(\bar{x}, y, f) = |y - f(\bar{x})|_{\epsilon} = \max\{0, |y - f(\bar{x})| - \epsilon\}, \tag{2.59}$$

где f – произвольная функция типа $\mathcal{R}^n \to \mathcal{R}$.

Пусть $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ – вектор переменных мягкого отступа, где $\xi_i = L^{\epsilon}(\bar{x}_i, y_i, f)$, $i = 1, \ \ldots, \ l$.

Аналогично, ϵ-нечувствительная квадратичная функция потерь определяется так:

$$L_2^{\epsilon}(\bar{x}, y, f) = (|y - f(\bar{x})|_{\epsilon})^2. \tag{2.60}$$

Задача минимизации в случае квадратичной функции потерь.

В случае квадратичной функции потерь минимизируется величина

$$\|\bar{w}\|^2 + C \sum_{i=1}^{l} L_2^{\epsilon}(\bar{x}_i, y_i, f),$$

где C – положительная константа. Далее $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$.

Мы будем минимизировать функцию потерь при фиксированном $\epsilon > 0$. Для этого в оптимизационную задачу вводятся переменные ξ_i и $\hat{\xi}_i$, с помощью которых контролируется отклонение остатков регрессии в большую или меньшую сторону от заданной границы ϵ.

Параметр C вводится для учета баланса между сложностью регрессионной гипотезы и суммой величин квадратичных остатков для этой гипотезы.

Прямая задача минимизации в случае квадратичной функции потерь (2.60) при фиксированных значениях параметров C и ϵ формулируется следующим образом:

$$\|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) \to \min$$

$$\text{при условиях } ((\bar{w}\cdot\bar{x}_i) + b) - y_i \leqslant \epsilon + \xi_i,$$

$$y_i - ((\bar{w}\cdot\bar{x}_i) + b) \leqslant \epsilon + \hat{\xi}_i,$$

$$\xi_i, \hat{\xi}_i \geqslant 0,\ i = 1,\ \ldots,\ l. \tag{2.61}$$

На практике параметр C подбирается путем процедуры типа перебора для данной обучающей выборки.

Лагранжиан прямой задачи запишем в виде

$$L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = \frac{1}{2}\|\bar{w}\|^2 + \frac{1}{2}C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) +$$

$$+\sum_{i=1}^{l}\alpha_i((\bar{w}\cdot\bar{x}_i) + b - y_i - \epsilon - \xi_i) +$$

$$+\sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w}\cdot\bar{x}_i) - b - \epsilon - \hat{\xi}_i),$$

где $\bar{\alpha} = (\alpha_1,\ \ldots,\ \alpha_l)$ и $\hat{\alpha} = (\hat{\alpha}_1,\ \ldots,\ \hat{\alpha}_l)$.

Заметим, что так же, как в задаче классификации, условия $\xi_i \geqslant 0$, $\hat{\xi}_i \geqslant 0$ можно опустить, так как всякое решение, где $\xi_i < 0$ или $\hat{\xi}_i < 0$, можно преобразовать в решение $\xi_i = 0$ или $\hat{\xi}_i = 0$.

Для поиска минимума приравниваем к нулю частные производные лагранжиана по переменным $\bar{w}$, b, $\bar{\xi}$, $\hat{\xi}$. Из равенства нулю производной по $\bar{w}$ получаем выражение для весового вектора

$$\bar{w} = \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i)\bar{x}_i. \tag{2.62}$$

Из равенства нулю производной по b получаем

$$\sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0.$$

Из равенства нулю производных по ξ_i и $\hat{\xi}_i$ получаем $\xi_i = \frac{1}{C}\alpha_i$ и $\hat{\xi}_i = \frac{1}{C}\hat{\alpha}_i$ для всех i.

Заметим, что для любого допустимого решения задачи (2.61) выполнено $\xi_i\hat{\xi}_i = 0$ для всех i. Поэтому для двойственной задачи будет $\alpha_i\hat{\alpha}_i = 0$.

Соответствующая двойственная задача формулируется следующим образом:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon\sum_{i=1}^{l}(\hat{\alpha}_i + \alpha_i) -$$
$$-\frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C}\delta_{ij}) \to \max$$
$$\text{при условиях } \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0,$$
$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0,\ i = 1,\ \ldots,\ l, \qquad (2.63)$$

где $\delta_{ij} = 1$ тогда и только тогда, когда $i = j$.

Условия Каруша–Куна–Таккера следующие:

$$\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \epsilon - \xi_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \epsilon - \hat{\xi}_i) = 0,\ i = 1,\ \ldots,\ l. \qquad (2.64)$$

Из условий Каруша–Куна–Таккера (2.64) следует, что для всех векторов выборки $\bar{x}_i$, попавших в слой размера ϵ вокруг гиперплоскости регрессии, выполнено $\alpha_i = \hat{\alpha}_i = 0$. Поэтому в сумме (2.62) соответствующие слагаемые отсутствуют ("опорных векторов" становится меньше), и решение задачи нахождения максимума в двойственной задаче упрощается.

Заметим, что для опорных векторов $\bar{x}_i$ выполнено $(\bar{w} \cdot \bar{x}_i) + b < y_i - \epsilon$ или $(\bar{w} \cdot \bar{x}_i) + b > y_i + \epsilon$.

Насколько уменьшается число параметров $\alpha_i, \hat{\alpha}_i$, зависит от взаимного расположения векторов выборки. Обычно такое уменьшение происходит при увеличении ϵ до определенного предела, при этом увеличивается точность регрессии на тестовой выборке. При дальнейшем увеличении ϵ эта точность падает.

Ядерная версия задачи регрессии. Поскольку векторы выборки входят в оптимизационную задачу только через скалярные произведения, можно использовать отображение в пространство признаков и перейти к ядерной версии.

Ядерная версия результата формулируется следующим образом.

Теорема 2.10. *Задана выборка* $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, *где* $\bar{x}_i \in \mathcal{X}$ *и* $y_i \in \mathcal{R}$. *Используется пространство признаков, задаваемое ядром* $K(\bar{x}, \bar{z})$.

Пусть $\bar{\alpha}$, $\bar{\alpha}$, *где* $\bar{\alpha} = (\alpha_1, \ \ldots, \ \alpha_l)$ *и* $\hat{\alpha} = (\hat{\alpha}_1, \ \ldots, \ \hat{\alpha}_l)$, *являются решениями квадратичной оптимизационной задачи:*

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(K(\bar{x}_i, \bar{x}_j) + \frac{1}{C}\delta_{ij}) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0,$$

$$\alpha_i \geqslant 0, \ \hat{\alpha}_i \geqslant 0 \ \text{при } i = 1, \ \ldots, \ l. \qquad (2.65)$$

Пусть также $f(\bar{x}) = \sum\limits_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) K(\bar{x}_i, \bar{x}) + b$, *где* b *выбирается так, чтобы* $f(\bar{x}_i) - y_i = -\epsilon - (\hat{\alpha}_i - \alpha_i)/C$ *для произвольного* i, *для которого* $\alpha_i > 0$ *или* $\hat{\alpha}_i > 0$.

Тогда функция $f(\bar{x})$ *эквивалентна гиперплоскости в пространстве признаков, определяемом ядром* $K(\bar{x}_i, \bar{x})$, *которая решает задачу оптимизации* (2.61).

Задача минимизации в случае линейной функции потерь. Аналогичным образом рассматривается задача регрессии при линейной функции потерь (2.59). Минимизируем величину

$$\|\bar{w}\|^2 + C \sum_{i=1}^{l} L^{\epsilon}(\bar{x}_i, y_i, f),$$

где $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$, C – положительный параметр, контролирующий баланс между сложностью регрессионной гипотезы и суммой величин линейных остатков для этой гипотезы.

Прямая задача минимизации в случае линейной функции потерь (2.59) при фиксированных значениях параметров C и ϵ формулируется следующим образом:

$$\begin{aligned} \|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) \to \min \\ \text{при условиях } ((\bar{w} \cdot \bar{x}_i) + b) - y_i \leqslant \epsilon + \xi_i, \\ y_i - ((\bar{w} \cdot \bar{x}_i) + b) \leqslant \epsilon + \hat{\xi}_i, \\ \xi_i, \hat{\xi}_i \geqslant 0,\ i = 1,\ \ldots,\ l. \end{aligned} \tag{2.66}$$

На практике параметр C подбирается путем процедуры типа перебора для данной обучающей выборки.

Лагранжиан прямой задачи имеет вид

$$\begin{aligned} L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}, \bar{r}, \hat{r}) = \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) + \\ + \sum_{i=1}^{l}\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \epsilon - \xi_i) + \\ + \sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \epsilon - \hat{\xi}_i) - \\ - \sum_{i=1}^{l} r_i\xi_i - \sum_{i=1}^{l}\hat{r}_i\hat{\xi}_i, \end{aligned} \tag{2.67}$$

где $\alpha_i \geqslant 0$, $\hat{\alpha}_i \geqslant 0$ и $r_i \geqslant 0$, $\hat{r}_i \geqslant 0$.

Соответствующая прямой задаче (2.66) двойственная задача получается и анализируется так же, как в разделе 2.6.2 (случай линейной нормы).

Для поиска минимума приравниваем к нулю частные производные лагранжиана по переменным $\bar{w}$, b, $\bar{\xi}$, $\hat{\xi}$ и получаем выра-

жение для весового вектора

$$\bar{w} = \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i)\bar{x}_i. \tag{2.68}$$

Из равенства нулю производной по b получаем

$$\sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0.$$

Кроме этого, получаем условия $C - \alpha_i - r_i = 0$ и $C - \hat{\alpha}_i - \hat{r}_i = 0$ при $i = 1, \ldots, l$.

Подставляем эти условия в лагранжиан (2.67) и получаем формулировку двойственной задачи:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) -$$
$$-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(\bar{x}_i \cdot \bar{x}_j) \to \max$$
$$\text{при условиях } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0,$$
$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C, \ i = 1, \ \ldots, \ l. \tag{2.69}$$

Условия Каруша–Куна–Таккера имеют вид:

$$\begin{gathered}
\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \epsilon - \xi_i) = 0, \\
\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \epsilon - \hat{\xi}_i) = 0, \\
(\alpha_i - C)\xi_i = 0, \ (\hat{\alpha}_i - C)\hat{\xi}_i = 0, \\
\xi_i\hat{\xi}_i = 0, \ \alpha_i\hat{\alpha}_i = 0, \ i = 1, \ \ldots, \ l.
\end{gathered} \tag{2.70}$$

Опорные векторы – это $\bar{x}_i$, для которых $\alpha_i > 0$ или $\hat{\alpha}_i > 0$. Если y_i находится вне слоя размера ϵ вокруг оптимальной гиперплоскости (гиперповерхности), то $\alpha_i = C$ или $\hat{\alpha}_i = C$.

Неравенства $0 < \alpha_i < C$ или $0 < \hat{\alpha}_i < C$ могут быть выполнены только для векторов $\bar{x}_i$ расположенных на границе слоя.

Векторы $\bar{x}_i$, у которых значения y_i расположены внутри слоя, заведомо не являются опорными. Для них $\alpha_i = 0$ и $\hat{\alpha}_i = 0$, так как в этом случае выполнены неравенства $(\bar{w} \cdot \bar{x}_i) + b - y_i < \epsilon$ и $y_i - ((\bar{w} \cdot \bar{x}_i) + b) < \epsilon$ и при этом $\xi_i = \hat{\xi}_i = 0$.

Весовой вектор – линейная комбинация опорных векторов

$$\bar{w} = \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i)\bar{x}_i.$$

Функция линейной регрессии имеет вид:

$$f(\bar{x}) = \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i)(\bar{x}_i \cdot \bar{x}) + b.$$

Для пространства с ядром двойственная задача имеет вид:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) K(\bar{x}_i, \bar{x}_j) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0,$$

$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C, \ i = 1, \ \ldots, \ l.$$

Функция регрессии для ядерной версии имеет вид:

$$f(\bar{x}) = \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) K(\bar{x}_i, \bar{x}) + b.$$

2.9.2. Гребневая регрессия в двойственной форме

Гребневая регрессия может быть представлена как частный случай регрессии с опорными векторами при ϵ-нечувствительной квадратичной функции потерь (2.60), где $\epsilon = 0$.

Проиллюстрируем решение этой задачи как частный случай регрессии с опорными векторами в случае $\epsilon = 0$, независимо от результатов раздела 2.8.2.

Вводим переменные мягкого отступа, как это было сделано разделе 2.8.1; при этом мы будем использовать те же обозначения расширенных переменных.

Тогда функция регрессии с расширенной переменной $\tilde{x}$ будет иметь однородный вид

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}).$$

Рассматривается прямая задача минимизации

$$\lambda \|\tilde{w}\|^2 + \sum_{i=1}^{l} \xi_i^2 \to \min$$

$$\text{при условии } y_i - (\tilde{w} \cdot \tilde{x}_i) = \xi_i, \ i = 1, \ \ldots, \ l.$$

В этом случае лагранжиан имеет вид

$$L(\tilde{w}, \bar{\xi}, \bar{\alpha}) = \lambda \|\tilde{w}\|^2 + \sum_{i=1}^{l} \xi_i^2 + \sum_{i=1}^{l} \alpha_i (y_i - (\tilde{w} \cdot \tilde{x}_i) - \xi_i). \qquad (2.71)$$

Приравниваем к нулю частные производные лагранжиана (2.71) по w_j и ξ_j :

$$\frac{\partial L(\tilde{w}, \bar{\xi}, \bar{\alpha})}{\partial \tilde{w}} = 2\lambda \tilde{w} - \sum_{i=1}^{l} \alpha_i \tilde{x}_i = 0,$$

$$2\xi_i - \alpha_i = 0$$

при $i = 1, \ \ldots, \ l$. Отсюда выражаем весовой вектор функции регрессии и переменные мягкого отступа через переменные двойственной задачи:

$$\tilde{w} = \frac{1}{2\lambda} \sum_{i=1}^{l} \alpha_i \tilde{x}_i,$$

$$\xi_i = \frac{\alpha_i}{2}.$$

Предварительно вычислим

$$\lambda(\tilde{w}\cdot\tilde{w}) = \frac{1}{4\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j),$$

$$\sum_{i=1}^{l}\alpha_i(\tilde{w}\cdot\tilde{x}_i) = \frac{1}{2\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j).$$

Подставим эти выражения в (2.71), получим задачу в двойственной форме

$$W(\bar{\alpha}) = \sum_{i=1}^{l}y_i\alpha_i - \frac{1}{4\lambda}\sum_{i,j=1}^{l}\alpha_i\alpha_j(\tilde{x}_i\cdot\tilde{x}_j) - \\ -\frac{1}{4}\sum_{i=1}^{l}\alpha_i^2 \to \max. \qquad (2.72)$$

Эту задачу можно переписать в векторной форме

$$W(\bar{\alpha}) = \bar{y}'\bar{\alpha} - \frac{1}{4\lambda}\bar{\alpha}'K\bar{\alpha} - \frac{1}{4}\bar{\alpha}'\bar{\alpha} \to \max,$$

где K – матрица Грама, элементами которой являются попарные скалярные произведения векторов $K_{i,j} = (\tilde{x}_i\cdot\tilde{x}_j)$.

Приравнивая к нулю частные производные $W(\bar{\alpha})$ в выражении (2.72) по α_i, получим систему уравнений, записанную в векторном виде

$$-\frac{1}{2\lambda}K\bar{\alpha} - \frac{1}{2}\bar{\alpha} + \bar{y} = \bar{0}.$$

Решение этого уравнения в векторном виде записывается так:

$$\bar{\alpha} = 2\lambda(K+\lambda I)^{-1}\bar{y}. \qquad (2.73)$$

Получаем уравнение регрессии в двойственной форме.

Представим скалярное произведение расширенного весового вектора и вектора расширенных переменных

$$(\tilde{w}\cdot\tilde{x}) = \frac{1}{2\lambda}\sum_{i=1}^{l}\alpha_i(\tilde{x}_i\cdot\tilde{x}) = \frac{1}{2\lambda}(\bar{\alpha}'\cdot\bar{k}),$$

где $\bar{\alpha} = (\alpha_1, \ \ldots, \ \alpha_l)$, $\bar{k} = (k_1, \ \ldots, \ k_l)$ при $k_i = (\tilde{x}_i \cdot \tilde{x})$.

Матрица K является симметрической, поэтому $K' = K$. По свойству транспонирования произведения матриц $(AB)' = B'A'$ и (2.73) имеем

$$\bar{\alpha}' = 2\lambda \bar{y}'(K + \lambda I)^{-1}.$$

Отсюда функция регрессии имеет вид

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}) = \bar{y}'(K + \lambda I)^{-1}\bar{k}. \tag{2.74}$$

Отметим один недостаток данной постановки. Поскольку число параметров α_i равно числу l элементов выборки, размер обращаемой матрицы $K + \lambda I$ равен $l \times l$, и мы не можем использовать для обучения слишком большую выборку.

В случае больших выборок можно разделять данные на группы и строить гиперплоскость регрессии для каждой группы отдельно. При этом возникают проблемы стыковки на границах групп.

Нелинейная гребневая регрессия с помощью ядер. Двойственная форма регрессии служит основой для обобщения линейной регрессии до нелинейной ее формы в пространстве признаков. Для этого вводится ядро $K(\tilde{x}, \tilde{y})$.

Рассмотрим схему построения нелинейной регрессии в пространстве признаков более подробно. Рассмотрим отображение $\tilde{x} \to \bar{\phi}(\tilde{x})$ в пространство признаков большей размерности $\mathcal{R}^N$. Тогда скалярное произведение в $\mathcal{R}^N$ порождает ядро $K(\tilde{x}_i, \tilde{x}_j) = (\bar{\phi}(\tilde{x}_i) \cdot \bar{\phi}(\tilde{x}_j))$.

Матрица Грама K имеет вид

$$K = \begin{pmatrix} K(\tilde{x}_1, \tilde{x}_1), & \ldots, & K(\tilde{x}_1, \tilde{x}_l) \\ K(\tilde{x}_2, \tilde{x}_1), & \ldots, & K(\tilde{x}_2, \tilde{x}_l) \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ K(\tilde{x}_l, \tilde{x}_1), & \ldots, & K(\tilde{x}_l, \tilde{x}_l) \end{pmatrix}.$$

Вектор $\bar{z}$ будет иметь вид

$$\bar{z} = \begin{pmatrix} K(\tilde{x}_1, \tilde{x}) \\ K(\tilde{x}_2, \tilde{x}) \\ \cdot \\ \cdot \\ \cdot \\ K(\tilde{x}_l, \tilde{x}) \end{pmatrix}.$$

Прообразом гиперплоскости (2.74), построенной в пространстве признаков $\mathcal{R}^N$, по образам векторов $\bar{\phi}(\tilde{x}_1), \ \ldots, \ \bar{\phi}(\tilde{x}_l)$ при отображении $\bar{\phi}$ является нелинейная поверхность

$$f(\tilde{x}) = \bar{y}'(\lambda I + K)^{-1}\bar{z}, \tag{2.75}$$

где $\bar{z} = (z_1, \ \ldots, \ z_l)$, $z_i = K(\tilde{x}, \tilde{x}_i)$ при $i = 1, \ \ldots, \ l$.

Для построения нелинейной поверхности (2.75) совсем не обязательно знать конкретный вид отображения в пространство признаков, достаточно знать ядро $K(\tilde{x}, \tilde{z})$.

Другой, более прямой, метод получения формулы (2.75) приведен в разделе 5.10.3.

Основной проблемой при решении таких задач является подбор ядра, наилучшим образом подходящего для разделения исходных данных. Другая проблема заключается в удачном подборе нормализующего параметра λ. Для ее решения разработаны специальные алгоритмы.

Вероятностный аналог изложенного выше способа построения регрессии с произвольным ядром называется *кригингом (Kriging)*. При вероятностной постановке векторы $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$ являются случайными величинами, при этом задан вид ковариационной функции $R(\tilde{x}_i, \tilde{x}_j) = E(\tilde{x}_i \cdot \tilde{x}_j)$. Обычно предполагается, что вид вероятностного распределения, генерирующего векторы $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$, известен с точностью до небольшого числа параметров.

2.10. Нелинейная оптимизация

Основные преимущества метода опорных векторов связаны с использованием двойственной задачи оптимизации. Двойственная

задача оптимизации не только упрощает граничные условия в задаче оптимизации, но и дает представление весовых коэффициентов разделяющей гиперплоскости (поверхности) через опорные векторы. Это представление не зависит от размерности пространства. Оно представляет собой метод сжатия информации, содержащейся в обучающей выборке.

В этом разделе рассматриваются постановки прямой и двойственной задач оптимизации, приведены основные их свойства.

Прямая задача оптимизации. Заданы вещественные функции $f(\bar{w})$, $g_i(\bar{w})$, $h_i(\bar{w})$, $i = 1, \ldots, m$, определенные на $\mathcal{R}^n$, $\bar{w} \in \mathcal{R}^n$. Необходимо найти

$$\inf_{\bar{w}} f(\bar{w}) \text{ при условиях}$$

$$g_i(\bar{w}) \leqslant 0,\ i = 1, \ \ldots, \ m, \tag{2.76}$$

$$h_i(\bar{w}) = 0,\ i = 1, \ \ldots, \ m. \tag{2.77}$$

Последние два условия можно записать в векторном виде $\bar{g}(\bar{w}) \leqslant \bar{0}$ и $\bar{h}(\bar{w}) = \bar{0}$. Пусть

$$\Omega = \{\bar{w} \in \mathcal{R}^n : \bar{g}(\bar{w}) \leqslant \bar{0}, \bar{h}(\bar{w}) = 0\}$$

– *область допустимости решений.*

Решение задачи оптимизации – это такой вектор $\bar{w}^*$, что $\bar{w}^* \in \Omega$ и не существует $\bar{w} \in \mathcal{R}^n$ такого, что $f(\bar{w}) < f(\bar{w}^*)$. Иными словами, на векторе $\bar{w}^*$ достигается *глобальный* минимум функции f. Если данное свойство верно в некоторой окрестности $\bar{w}^*$, то получаем определение *локального* минимума. Функция f называется *целевой* функцией.

Если $f(\bar{w})$ – квадратичная функция от координат $\bar{w}$, а $\bar{g}$ и $\bar{h}$ – линейные функции, то такая задача оптимизации называется задачей *квадратичного программирования.*

Функция f называется выпуклой, если для всех $\bar{w}, \bar{u} \in \mathcal{R}^n$ и $0 \leqslant \lambda \leqslant 1$ выполнено

$$f(\lambda\bar{w} + (1-\lambda)\bar{u}) \leqslant \lambda f(\bar{w}) + (1-\lambda) f(\bar{u}).$$

Теория Лагранжа – это случай, когда имеются только условия $\bar{h}(\bar{w}) = \bar{0}$. Лагранжиан имеет вид

$$L(\bar{w}, \bar{\beta}) = f(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}).$$

Необходимое условие минимума

$$\frac{\partial L(\bar{w}, \bar{\beta})}{\partial \bar{w}} = \bar{0},$$
$$\frac{\partial L(\bar{w}, \bar{\beta})}{\partial \bar{\beta}} = \bar{0}.$$

Это условие является достаточным, если функция L выпуклая.

При общей постановке задачи (2.77) лагранжиан имеет вид

$$L(\bar{w}, \bar{\alpha}, \bar{\beta}) = f(\bar{w}) + \sum_{i=1}^{m} \alpha_i g_i(\bar{w}) + \sum_{i=1}^{m} \beta_i h_i(\bar{w}) =$$
$$= f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}).$$

Двойственная задача оптимизации. Двойственная задача оптимизации часто проще, чем прямая, так как у нее более простые граничные условия. Пусть

$$\Theta(\bar{\alpha}, \bar{\beta}) = \inf_{\bar{w}} L(\bar{\alpha}, \bar{\beta}, \bar{w}).$$

Двойственная задача оптимизации заключается в том, чтобы найти

$$\max_{(\bar{\alpha}, \bar{\beta})} \Theta(\bar{\alpha}, \bar{\beta}) \text{ при условиях}$$
$$\alpha_i \geqslant 0,\ i = 1,\ \ldots,\ m. \tag{2.78}$$

Ниже приводится слабая теорема двойственности.

Теорема 2.11. *Пусть вектор $\bar{w}$ удовлетворяет условиям* (2.76) *и* (2.77) *прямой задачи оптимизации* (*в частности, он может быть решением прямой задачи*), *а* $(\bar{\alpha}, \bar{\beta})$ *– решение двойственной задачи* (2.78). *Тогда* $f(\bar{w}) \geqslant \Theta(\bar{\alpha}, \bar{\beta})$.

Доказательство. Имеем

$$\Theta(\bar{\alpha}, \bar{\beta}) = \inf_{\bar{u}} L(\bar{u}, \bar{\alpha}, \bar{\beta}) \leqslant L(\bar{w}, \bar{\alpha}, \bar{\beta}) =$$
$$= f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}) \leqslant f(\bar{w}). \tag{2.79}$$

Здесь $\bar{\alpha}\bar{g}(\bar{w}) \leqslant 0$, так как $\bar{\alpha} \geqslant \bar{0}$ и $\bar{g}(\bar{w}) \leqslant \bar{0}$, $\bar{h}(\bar{w}) = \bar{0}$. $\triangle$

Непосредственно из теоремы получаем

Следствие 2.2. *Значение решения двойственной задачи не превосходит значения решения прямой задачи:*

$$\sup\{\Theta(\bar{\alpha}, \bar{\beta}) : \bar{\alpha} \geqslant \bar{0}\} \leqslant \inf\{f(\bar{w}) : \bar{g}(\bar{w}) \leqslant \bar{0}, \bar{h}(\bar{w}) = \bar{0}\}.$$

Еще одно следствие из этой теоремы дает достаточное условие для того, чтобы значения решений прямой и двойственной задач совпадали.

Следствие 2.3. *Если* $f(\bar{w}^*) = \Theta(\bar{\alpha}^*, \bar{\beta}^*)$, *где* $\bar{\alpha}^* \geqslant \bar{0}$, $\bar{g}(\bar{w}^*) \leqslant \bar{0}$, $\bar{h}(\bar{w}^*) = \bar{0}$, *то* $\bar{w}^*$ *и* $(\bar{\alpha}^*, \beta^*)$ *– решения прямой и двойственной задач соответственно.*

В этом случае также $\bar{\alpha}^* \bar{g}(\bar{w}^*) = 0$.

Доказательство. В условиях следствия в неравенстве (2.79) два крайних члена равны, поэтому оно является равенством. В частности, $f(\bar{w}^*) = \inf\limits_{\bar{u}} L(\bar{u}, \bar{\alpha}^*, \bar{\beta}^*)$ и $\bar{\alpha}^* \bar{g}(\bar{w}^*) = 0$. $\triangle$

Достаточным условием существования решения прямой и двойственной задачи является существование *седловой* точки лагранжиана. Для седловой точки $(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)$ должны выполняться неравенства

$$L(\bar{w}^*, \bar{\alpha}, \bar{\beta}) \leqslant L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*) \leqslant L(\bar{w}, \bar{\alpha}^*, \bar{\beta}^*)$$

для всех $\bar{w}$, $\bar{\alpha}$, $\bar{\beta}$.

Другое достаточное условие равенства значений решений прямой и двойственной задачи дано в следующей теореме.

Теорема 2.12. *Пусть область допустимости задачи* Ω *– выпуклое подмножество* $\mathcal{R}^n$, *функции* $\bar{h}$, $\bar{g}$ *– аффинные (т.е.* $h_i(\bar{w})$, $g_i(\bar{w})$ *имеют вид* $A_i\bar{w} + \bar{b}_i$, *где* A_i *– некоторая матрица). Тогда значения решений прямой и двойственной задач совпадают.*

Теперь сформулируем основную теорему выпуклой оптимизации – *Теорему Куна–Таккера.*

Теорема 2.13. *Пусть область допустимости задачи* Ω *– выпуклое подмножество* $\mathcal{R}^n$, *функция* f *– выпуклая, функции* $\bar{h}$, $\bar{g}$ *– аффинные (т.е.* $h_i(\bar{w})$, $g_i(\bar{w})$ *имеют вид* $A_i\bar{w} + \bar{b}_i$, *где* A_i *– некоторая матрица).*

Тогда вектор $\bar{w}^*$ является решением прямой задачи

$$\inf f(\bar{w}),\ \bar{w} \in \Omega,\ \textit{при условиях}$$
$$\bar{g}(\bar{w}) \leqslant \bar{0},$$
$$\bar{h}(\bar{w}) = \bar{0}$$

тогда и только тогда, когда существует пара $(\bar{\alpha}^*, \bar{\beta}^*)$ такая, что

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{w}} = \bar{0},$$
$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{\beta}} = \bar{0}, \tag{2.80}$$
$$\alpha_i^* g_i(\bar{w}^*) = 0, \tag{2.81}$$
$$g_i(\bar{w}^*) \leqslant 0,$$
$$\alpha_i^* \geqslant 0,\ i = 1,\ \ldots,\ m.$$

Условия достижения максимума по $\bar{\beta}$ линейной по $\bar{\alpha}$ и $\bar{\beta}$ функции $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ задаются условиями (2.80); они эквивалентны совокупности условий: $h_i(\bar{w}^*) = 0$, $i = 1,\ \ldots,\ k$.

Условия максимума функции $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ по α_i^* содержатся в условиях (2.81), так как при $\alpha_i^* > 0$, каждое такое условие превращается в условие $g_i(\bar{w}^*) = 0$ (что эквивалентно равенству нулю производной $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ по α_i), а при $g_i(\bar{w}^*) < 0$ в точке максимума функции $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ должно быть $\alpha_i^* = 0$.

Условия (2.81) называются *условиями Каруша–Куна–Таккера.* Они означают, что если решение задачи оптимизации достигается на границе i-го условия, то $\alpha_i^* \geqslant 0$, в противном случае $\alpha_i^* = 0$.

Квадратичное программирование. Рассмотрим задачу квадратичного программирования

$$\frac{1}{2}\bar{w}' Q \bar{w} - \bar{k}\bar{w} \to \min$$
$$\text{при условии } X\bar{w} \leqslant \bar{c}, \tag{2.82}$$

где Q – $(n \times n)$-положительно определенная матрица, $\bar{k}$ – n-вектор, $\bar{c}$ – m-вектор, $\bar{w}$ – n-вектор неизвестных, X – $(m \times n)$-матрица.

Допускаем, что условия определяют непустое множество векторов. Тогда задача может быть переписана в виде: найти максимум

$$\max_{\bar{\alpha}} \left(\min_{\bar{w}} \left(\frac{1}{2} \bar{w}' Q \bar{w} - \bar{k} \bar{w} + \bar{\alpha}' (X \bar{w} - \bar{c}) \right) \right) \qquad (2.83)$$

$$\text{при условии } \bar{\alpha} \geqslant \bar{0}.$$

Минимум по $\bar{w}$ в (2.83) достигается при

$$\bar{w} = Q^{-1}(\bar{k} - X' \bar{\alpha}).$$

Подставляем это выражение в (2.82), получим двойственную задачу

$$-\frac{1}{2} \bar{\alpha}' P \bar{\alpha} - \bar{\alpha}' \bar{d} - \frac{1}{2} \bar{k}' Q \bar{k} \to \max$$

$$\text{при условии } \bar{\alpha} \geqslant \bar{0}, \qquad (2.84)$$

где $P = XQ^{-1}X'$, $\bar{d} = \bar{c} - XQ^{-1}\bar{k}$.

Двойственная задача также является квадратичной, но ее граничные условия проще, чем у прямой задачи.

2.11. Конформные предсказания

Задана выборка $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{R}^n$ и $y_i \in \{-1, +1\}$ при $1 \leqslant i \leqslant l$.

При решении задачи классификации с помощью разделяющей гиперповерхности различные примеры из выборки классифицируются с разной степенью качества. Мера качества классификации примера $(\bar{x}_i, y_i)$ – мера неконформности – была введена Вовком и Гаммерманом [62]. Мера неконформности применяется для повышения эффективности известных алгоритмов на основе новых способов оценки уровня доверия к результатам их работы. Эти способы оценки носят общий характер и приводят к состоятельным алгоритмам при очень общих вероятностных предположениях о механизмах генерации данных.

Мы определим меру неконформности для классификации с помощью SVM. Напомним основные положения метода построения машин на опорных векторах (SVM). В методе SVM исходные векторы $\bar{x}_i$ отображаются в векторы $\bar{\phi}(\bar{x}_i)$ в пространстве признаков, определенном ядром $K(\bar{x}, \bar{x}')$. После этого, строится разделяющая гиперплоскость в пространстве признаков. Согласно (2.40) вектор весов разделяющей гиперплоскости выражается в виде линейной комбинации образов опорных векторов:

$$\bar{w} = \sum_{i=1}^{l} y_i \alpha_i \bar{\phi}(\bar{x}_i),$$

где коэффициенты Лагранжа α_i вычисляются в результате решения двойственной задачи оптимизации.

По теореме 2.6 в исходном пространстве соответствующая разделяющая поверхность имеет вид:

$$f(\bar{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\bar{x}_i, \bar{x}) + b.$$

Решаем оптимизационную задачу построения SVM по выборке S, при этом будут вычислены коэффициенты Лагранжа α_i. Возьмем в качестве *меры неконформности* примера $(\bar{x}_i, y_i)$ значение коэффициента Лагранжа α_i.

Это определение обосновывается следующим образом. Из условий Каруша–Куна–Таккера следует, что $\alpha_i = 0$, если векторы $\bar{\phi}(\bar{x}_i)$ правильно классифицируются и лежат с внешней стороны относительно граничных гиперплоскостей. Опорными являются те векторы $\bar{x}_i$, для которых $\alpha_i > 0$. Это те векторы, образы которых лежат на граничных гиперплоскостях или же неправильно ими классифицируются. Таким образом?, примеры с $\alpha_i = 0$ имеют высшую степень согласованности с выборкой, по которой построена разделяющая гиперплоскость (или гиперповерхность), а примеры с положительными значениями α_i имеют различную степень согласованности с выборкой, которая тем хуже, чем больше значение α_i.

Пусть введенная мера неконформности применяется к примеру $(\bar{x}_i, y_i)$. Определяется *p-тест (p-value)*:

$$p_i = \frac{|\{j : \alpha_j \geqslant \alpha_i\}|}{l}.$$

По определению $0 \leqslant p_i \leqslant 1$. Малое значение p_i означает, что пример $(\bar{x}_i, y_i)$ имеет одну из самых больших мер неконформности среди примеров выборки.

На основе введенного понятия p-теста в [62] предложен мета-алгоритм для вычисления конформных предсказаний с использованием SVM.

Пусть дана выборка $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ и вектор $\bar{x}_{l+1}$, которому надо приписать метку класса $y_{l+1} \in \{-1, +1\}$. Задан также уровень доверия $\epsilon > 0$.

Мета-алгоритм:

Для каждого $y \in \{-1, +1\}$ решаем оптимизационную задачу построения SVM по выборке $S' = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l), (\bar{x}_{l+1}, y))$, находим значения коэффициентов Лагранжа α_i, $1 \leqslant i \leqslant l+1$ и вычислем значение p-теста

$$p(y) = \frac{|\{j : \alpha_j \geqslant \alpha_{l+1}\}|}{l+1}.$$

Результат работы алгоритма:

- если $p(y) < \epsilon$ для всех y, то алгоритм не выдает никакого результата;
- если $p(y) \geqslant \epsilon$ для некоторого y, то выдаем в качестве результата то значение y, для которого величина $p(y)$ принимает максимальное значение:

$$y_{l+1} = \arg\max_y p(y).$$

Подобный порядок действий обосновывается вероятностным результатом, который утверждает, что при некоторых вероятностных предположениях о механизме генерации примеров выборки p-тест удовлетворяет естественному условию: $P\{p_i \leqslant \epsilon\} \leqslant \epsilon$, где

P – некоторая мера на наборах α_i инвариантная относительно их перестановок (подробнее см. в [62]).

Меры неконформности строятся исходя из специфики моделей данных. В монографии [62] построены меры некомформности для алгоритмов ближайшего соседа, SVM, будстрепа, нейронных сетей, решающих деревьев, гребневой регрессии и алгоритма Байеса.

Приведем пример меры некомформности для алгоритма классификации методом ближайшего соседа.

Идея алгоритма k-ближайших соседей заключается в следующем. Для того чтобы предсказать метку нового объекта $\bar{x}$, находятся k ближайших по расстоянию соседей этого объекта. В задаче классификации "методом голосования" объекту приписывается метка, которая наиболее часто встречается у ближайших k объектов, а в методе регрессии можно взять медиану их меток.

Рассмотрим примеры $(\bar{x}, y)$, где $\bar{x} \in \mathcal{R}^n$, $y \in D$, D – конечное множество меток. Допустим, что $\{\bar{x}_1, \ldots, \bar{x}_k\}$ – множество k ближайших к $\bar{x}$ объектов и $\{y_1, \ldots, y_k\}$ – их метки.

Пусть $(\bar{x}, y)$ – некоторый пример. Определим меру неконформности этого примера в виде отношения минимального расстояния от объекта $\bar{x}$ до объектов $\bar{x}_i$ с той же меткой $y_i = y$ к минимальному расстоянию от этого объекта $\bar{x}$ до объектов $\bar{x}_i$ с другими метками $y_i \neq y$:

$$\alpha_{(\bar{x},y)} = \frac{\min_{1 \leqslant j \leqslant k, y_j = y} d(\bar{x}, \bar{x}_j)}{\min_{1 \leqslant j \leqslant k, y_j \neq y} d(\bar{x}, \bar{x}_j)}.$$

Под расстоянием $d(\bar{x}, \bar{x}')$ понимается обычное евклидово расстояние между двумя векторами.

Чем больше величина $\alpha_{(\bar{x},y)}$, тем ближе расположен объект $\bar{x}$ к другим объектам, отмеченным метками отличными от y, т.е. тем больше степень неконформности примера $(\bar{x}, y)$.

2.12. Задачи и упражнения

1. Доказать оставшуюся часть утверждения леммы 2.1.

2. Построить отображения $\mathcal{R}^n$ в пространства признаков и соответствующие полиномиальные ядра для полиномов общего вида и более высокого порядка ($k = 3, 4, \ \dots$), а также соответствующие функции классификации вида (2.25).

3. Для любой симметричной положительно определенной функции $K(x, y)$ выполнено неравенство типа Коши–Буняковского:

$$K(x_1, x_2) \leqslant \sqrt{K(x_1, x_1)K(x_2, x_2)} \text{ для всех } x_1, x_2 \in X. \qquad (2.85)$$

(Указание: из положительной определенности (2×2)-матрицы $K(x_i, x_j)$ следует, что ее собственные значения неотрицательные. Поэтому то же верно и для определителя.)

4. Докажите, что для любой симметричной положительно определенной функции $K(x, y)$ будет:
(i) $K(x, x) \geqslant 0$ для всех x.
(ii) Если $K(x, x) = 0$ для всех x, то $K(x, y) = 0$ для всех x и y.

5. Рассматривается гильбертово пространство $\mathcal{F}$ функций на X, которое обладает следующим свойством: функционал $f \to f(x)$ является непрерывным линейным функционалом. По теореме Рисса–Фишера для каждого $x \in X$ существует элемент $K_x \in \mathcal{F}$ такой, что $f(x) = (K_x \cdot f)$. Воспроизводящее ядро определяется так: $K(x, y) = (K_x \cdot K_y)$.

Доказать, что функция $K(x, y) = (K_x \cdot K_y)$ является симметричной и положительно определенной.

6. Пусть $K_1(x, y), K_2(x, y), \ \dots$ – симметричные положительно определенные ядра на X. Доказать, что следующие их комбинации также являются симметричными положительно определенными ядрами:
(i) $\alpha_1 K_1(x, y) + \alpha_2 K_2(x, y)$, где $\alpha_1, \alpha_2 \geqslant 0$;
(ii) $K(x, y) = \lim\limits_{n \to \infty} K_n(x, y)$;
(iii) $K_1(x, y)K_2(x, y)$ (Указание: использовать представление положительно определенной матрицы (Грама) в виде $K = MM'$);
(iv) $K(A, B) = \sum\limits_{x \in A, y \in B} K(x, y)$, где A, B – конечные подмножества X (это ядро на множестве всех конечных подмножеств X).

Указать соответствующие отображения в пространства признаков.

7. Доказать, что функции $K_1(\bar{x},\bar{y}) = \exp\left(\frac{(\bar{x}\cdot\bar{y})}{\sigma^2}\right)$ и $K_2(\bar{x},\bar{y}) = \exp\left(-\frac{\|\bar{x}-\bar{y}\|_2^2}{\sigma^2}\right)$ являются положительно определенными (здесь $\bar{x},\bar{y} \in \mathcal{R}^n$, $\|\cdot\|_2$ – эвклидова норма). Привести соответствующие представления этих функций через скалярные произведения в пространстве признаков.

(*Указание.* Использовать свойства (i-iv) предыдущей задачи.

7. При определении в разделе 2.5 канонического гильбертова пространства, порожденного ядром, было определено скалярное произведение на пополнении $\mathcal{F}$ множества линейных комбинаций $\mathcal{F}_1$: $(f\cdot g) = \lim\limits_{n\to\infty}(f_n\cdot g_n)$, где $f_1, f_2, \ldots$ и $g_1, g_2, \ldots$ – некоторые фундаментальные последовательности из $\mathcal{F}_1$, сходящиеся к $f, g \in \mathcal{F}$.
(i) Доказать, что этот предел существует и не зависит от выбора фундаментальных последовательностей $f_1, f_2, \ldots$ и $g_1, g_2, \ldots$.
(ii) Доказать также, что каноническое гильбертово пространство $\mathcal{F}$ является полным в норме, порожденной этим скалярным произведением.
(iii) Доказать. что функционал $f \to (f \cdot K_x)$ является непрерывным на $\mathcal{F}$ (см. [38]).

8. Возможно другое определение: RKHS – это гильбертово пространство $\mathcal{F}$ функций определенных на множестве X, для которого существует функция $K(x,y)$ такая, что

- $K(x,\cdot) \in \mathcal{F}$ для любого $x \in X$;
- $f(x) = (f \cdot K(x,\cdot))$ для любых $f \in \mathcal{F}$ и $x \in X$.

(i) Доказать, что это определение эквивалентно определению RKHS, данному в разделе 2.5: RKHS – это гильбертово пространство $\mathcal{F}$ функций на X, которое обладает следующим свойством: функционал $f \to f(x)$ является непрерывным линейным функционалом.
(ii) Доказать, что функция $K(x,y)$ из этого определения единственная для заданного RKHS $\mathcal{F}$.
(iii) Доказать, что функция $K(x,y)$, обладающая заданными двумя свойствами, определяет единственное с точностью до изоморфизма RKHS (см. доказательство в [38]).

9. Пусть X конечное и пространство H состоит из всех вещественных функций, определенных на X. Тогда каждый элемент H

можно представить в виде конечномерного вектора. На этих векторах рассматривается евклидова норма и соответствующее скалярное произведение. Построить соответствующее ядро $K(x, y)$, элемент K_x и функционал $f \to f(x)$. Проверить, что пространство H есть RKHS.

10. Доказать, что гильбертово пространство $L_2([0,1], P)$ всех измеримых функций $f : [0,1] \to \mathcal{R}$, для которых $\int f^2 dP < \infty$, со скалярным произведением $(f \cdot g) = \int fg dP$ не является RKHS.

11. Доказать, что в оптимизационной задаче (2.65) значение b не зависит от i.

12. Доказать, что в оптимизационной задаче разделения с ошибками в линейной норме для любой нетривиальной выборки $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$ (т.е., такой, что $y_i = 1$ и $y_j = -1$ для каких-нибудь $1 \leqslant i, j \leqslant l$) обязательно найдутся число b и вектор $\bar{x}_i$, лежащий на одной из граничных гиперплоскостей, $y_i((\bar{w} \cdot \bar{x}_i) + b) = 1$, для которого $\alpha_i > 0$.

13. Доказать, что матрица Грама $K_{i,j} = (\tilde{x}_i \cdot \tilde{x}_j)_{i,j=1}^n$ обратима тогда и только тогда, когда векторы $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$ линейно независимы.

14. (i) Найти максимум объема параллелепипеда при заданной площади поверхности.

(ii) Найти максимум энтропии $H(p_1, \ \ldots, \ p_n) = -\sum\limits_{i=1}^{n} p_i \ln p_i$ при условиях $\sum p_i = 1$, $\sum c_i p_i = e$.

15. Провести все необходимые выкладки для получения решения (2.84) квадратичной задачи.

16. Вывести соотношения (2.65) для двойственной задачи регрессии.

2.13. Лабораторные работы

В этом разделе предлагаются стандартные лабораторные работы для решения задачи классификации с помощью SVM.

Выполнение работы включает следующие процедуры:

- Загрузить исходные данные с соответствующего сайта. Как правило, исходные данные – это набор векторов большой

размерности, в которых уже указан класс объекта.

- Разделить данные на обучающую и тестовую выборки. Класс объекта используется в обучающей выборке для проведения обучения, а в тестовой выборке – для проверки правильности классификации. После проведения классификации требуется подсчитать долю правильных ответов.

- Перевод данных в формат, допускаемый программным обеспечением SVM.

- Провести калибровку (шкалирование) исходных данных.

  Шкалирование данных помогает избежать потери точности из-за слишком малых или слишком больших значений некоторых признаков. В частности, это важно при использовании гауссова ядра. Рекомендуется нормировать численное значение каждого признака так, чтобы оно попадало в интервал типа $[-1, 1]$ или $[0, 1]$.

- Выбрать ядро, наилучшим образом классифицирующее обучающую выборку. Как правило, стандартные программы SVM используют следующие ядра:

  1) линейное ядро $K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y})$,

  2) полиномиальное ядро $K(\bar{x}, \bar{y}) = (\gamma(\bar{x} \cdot \bar{y}) + r)^d$, где $\gamma > 0$,

  3) гауссово ядро $K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x}-\bar{y}\|^2}{\sigma^2}}$,

  4) сигмоидное ядро $K(\bar{x}, \bar{y}) = th(\gamma(\bar{x} \cdot \bar{y}) + r)$.

  Рекомендуется первоначально выбрать гауссово ядро. Имеются случаи, когда гауссово ядро дает неудовлетворительные результаты. Например, это может происходить в случае, когда размерность пространства объектов очень большая. В этом случае хорошие результаты может давать линейное ядро.

- Провести перекрестную проверку для нахождения наилучших значений параметров C и γ. Заметим, что недостаточно подобрать значения параметров, которые дают наилучшую

точность только на обучающей выборке. Простейший способ – разделить выборку на две части, найти наилучшие значения параметров при обучении на первой части и использовать результаты классификации на второй части в качестве оценки качества обучения.

Имеется более сложная процедура перекрестной проверки (cross-validation), при которой обучающая выборка разделяется на N равных частей. Последовательно выбирается одно из подмножеств, после этого классификатор обучается на объединении $N-1$ оставшихся подмножеств и проверяется на выбранном подмножестве. Выбираются значения параметров, дающие наибольшую точность на одном из таких подмножеств.

Подбор параметров C и γ может производиться простым перебором по некоторому дискретному подмножеству – решетке. Недостатком этого метода является большое время вычисления. Применяются различные эвристические методы перебора C и γ.

- Выбрать для использования параметры C и γ, которые дают наилучшую точность классификации.

- Провести классификацию на тестовой выборке. Оформить результаты с сопоставлением точности классификации на обучающей и тестовой выборках.

Имеется ряд сайтов, содержащих программное обеспечение SVM и соответствующие примеры для проведения экспериментальных расчетов. Отметим некоторые из них.

Программное обеспечение SVM можно найти на сайтах:
http://www.csie.ntu.edu.tw, www.support-vector.net

Там же можно найти инструкции по практическому применению программ SVM и подготовке исходных данных. Там же приведены примеры.

На сайте http://archive.ics.uci.edu содержатся исходные данные для решения задач классификации и регрессии.

Лабораторная работа 1

Провести обучение и классификацию рукописных цифр. Данные в формате MATLAB можно найти по адресу:
http://www.cs.toronto.edu

В частности, по этому адресу имеются данные из базы USPS, содержащие цифровые образы различных написаний рукописных цифр.

Лабораторная работа 2

Провести обучение и классификацию по данным из указанных выше сайтов. Выбрать набор данных, провести обучение SVM и тестирование на тестовой выборке.

Лабораторная работа 3

Провести обучение и классификацию на предыдущих данных с помощью персептрона и алгоритма Розенблатта. Провести сравнение времени работы.

Часть II

Прогнозирование индивидуальных последовательностей

Глава 3

Универсальные предсказания

3.1. Последовательные вероятностные предсказания

Рассматривается следующая задача прогнозирования: предсказатель получает в режиме онлайн некоторую числовую последовательность исходов $\omega_1,\ \omega_2,\ \ldots,\ \omega_{n-1},\ \ldots$ При этом предсказателю не известно распределение вероятностей источника, генерирующего эту последовательность. Задачей предсказателя является вычисление оценок вероятностей p_n будущих исходов ω_n по уже известным $n-1$ исходам $\omega_1,\ \omega_2,\ \ldots,\ \omega_{n-1}$.

Число p_n может рассматриваться как вероятность события $\omega_n = 1$ в том случае, когда ω_i принимают значения 0 или 1. Легко видеть, что в этом случае число p_n также является математическим ожиданием случайной величины ω_n.

В случае конечного числа исходов величина p_n может быть вектором вероятностей всех возможных исходов.

Задано множество $\Omega = \{1, 2, \ldots, m\}$ исходов, $\mathcal{P}(\Omega)$ – множество всех распределений вероятностей на Ω. *Предсказательная стратегия* (или просто стратегия) – это бесконечная последовательность $f_t(\cdot|y^{t-1}) \in \mathcal{P}(\Omega)$ условных распределений вероятностей на Ω. Здесь $y^t = y_1, \ldots, y_t,\ y_t \in \Omega,\ t = 1, 2, \ldots$. Соответству-

ющее распределение вероятностей на Ω^n определяется $f(y^n) = \prod_{t=1}^n f_i(y_t|y^{t-1})$, $f(y^0) = 1$. Легко проверить, что $\sum_{y^n} f(y^n) = 1$. С другой стороны, по каждому распределению вероятностей f на Ω^n можно восстановить условные распределения

$$f_i(y_t|y^{t-1}) = \frac{f(y^t)}{f(y^{t-1})}.$$

для всех $1 \leqslant t \leqslant n$.

В общем случае, предсказательная стратегия определяется функцией f определенной на $\Omega^* = \cup_n \Omega^n$. В дальнейшем индекс t в f_t опускаем, обозначим $f(y_t|y^{t-1}) = f_t(y_t|y^{t-1})$.

Будем рассматривать логарифмическую функцию потерь. Исчисляем потери предсказательной стратегии f на последовательности исходов $y^n = y_1, \ldots, y_n$ в виде

$$L_n^f(y^n) = \sum_{t=1}^n \ln \frac{1}{f(y_t|y^{t-1})} = -\ln f(y^n).$$

Пусть $\mathcal{F}$ – некоторое множество предсказательных стратегий, $p \in \mathcal{F}$. Оцениваем эффективность предсказательной стратегии p на последовательности исходов $y^n = y_1 \ldots y_n$ с помощью *регрета*

$$R_n^p(y^n) = L_n^p(y^n) - \inf_{f \in \mathcal{F}} L_n^f(y^n) = \sup_{f \in \mathcal{F}} \ln \frac{f(y^n)}{p(y^n)}.$$

3.1.1. Смешивающий предсказатель

В дальнейшем мы рассмотрим два типа оптимальных предсказательных стратегий. Первая из них – смешивающая предсказательная стратегия (смешивающий предсказатель).

Для простоты изложения предполагаем, что класс $\mathcal{F}$ – конеч-

ный.[1] Определим *смешивающий предсказатель*

$$p(y_n|y^{n-1}) = \frac{\sum\limits_{f\in\mathcal{F}} f(y_n|y^{n-1})e^{-L^f_{n-1}(y^{n-1})}}{\sum\limits_{f\in\mathcal{F}} e^{-L^f_{n-1}(y^{n-1})}} = \\ = \frac{\sum\limits_{f\in\mathcal{F}} f(y_n|y^{n-1})f(y^{n-1})}{\sum\limits_{f\in\mathcal{F}} f(y^{n-1})} = \\ = \frac{\sum\limits_{f\in\mathcal{F}} f(y^n)}{\sum\limits_{f\in\mathcal{F}} f(y^{n-1})}.$$

Обозначим $W_n = \sum\limits_{f\in\mathcal{F}} f(y^n)$. $W_0 = \sum\limits_{f\in\mathcal{F}} f(y^0) = |\mathcal{F}|$. Тогда

$$p(y^n) = \prod_{i=1}^{n} p(y_i|y^{i-1}) = \prod_{i=1}^{n} \frac{W_i}{W_{i-1}} = \frac{W_n}{W_0} = \\ = \frac{\sum\limits_{f\in\mathcal{F}} f(y^n)}{|\mathcal{F}|}.$$

Регрет смешивающего предсказателя оценивается сверху

$$R^p_n(y^n) = \sup_{f\in\mathcal{F}} \ln \frac{f(y^n)}{\frac{1}{|\mathcal{F}|}\sum\limits_{f\in\mathcal{F}} f(y^n)} \leqslant \ln|\mathcal{F}|.$$

Правило Лапласа (см. раздел 3.1.2) является частным случаем смешивающего предсказателя для случая бесконечного множества $\mathcal{F}$.

3.1.2. Правило Лапласа

Исторически первой процедурой универсального прогнозирования является правило Лапласа. [2] Эта процедура использует гипо-

[1] В противном случае надо заменить суммы на соответствующие интегралы.

[2] Лаплас рассматривал задачу вычисления вероятности события, которое заключается в том, что солнце взойдет завтра, если известно, что оно всхо-

тезу о том, что исходы ω_i порождаются некоторым источником, который генерирует их независимо друг от друга с одной и той же вероятностью единицы, равной p.

Пусть исходы ω_i принадлежат множеству $\{0,1\}$. Мы также предполагаем, что в каждый момент времени $i = 1, 2, \ldots$ исход ω_i порождается независимо от предыдущих исходов с неизвестными нам постоянными вероятностями $p = P\{\omega_i = 1\}$ и $q = P\{\omega_i = 0\} = 1 - p$. Необходимо оценивать эти вероятности в режиме онлайн на основе статистики предыдущих исходов.

Пусть мы наблюдаем исходы $\omega^n = \omega_1, \ldots \omega_n$, в которых имеется n_1 единиц и n_2 нулей, $n_1 + n_2 = n$. Вероятность получить такую последовательность исходов равна $p^{n_1}(1-p)^{n_2}$, если вероятность единицы равна p. Так как истинная вероятность p неизвестна, рассмотрим байесовскую смесь вероятностей последовательности длины n по всем возможным p :

$$P(\omega^n) = \int\limits_0^1 p^{n_1}(1-p)^{n_2} dp.$$

Значение этого интеграла легко вычислить.

Лемма 3.1.

$$\int\limits_0^1 p^{n_1}(1-p)^{n_2} dp = \frac{1}{(n+1)\binom{n}{n_1}}.$$

Доказательство. Проверим это равенство обратной индукцией по n_1. При $n_1 = n$ имеем $\int\limits_0^1 p^n dp = \frac{1}{(n+1)}$.

дило каждый день последние 5000 лет (1826251 день). Эта задача находится на границе применимости частотной интерпретации вероятности, так как не выполнено условие повторяемости данного опыта. Кроме того, известная последовательность состоит из одних восходов. Тем не менее, для того, чтобы выразить степень нашей субъективной неопределенности о значении p этой вероятности, мы считаем все значения p равновозможными. Согласно формуле, приведенной далее, вероятность рассматриваемого события будет равна $\frac{1826252}{1826253}$.

Предположим, что

$$\int\limits_0^1 p^{n_1+1}(1-p)^{n_2-1}dp = \frac{1}{(n+1)\binom{n}{n_1+1}}.$$

Интегрируя по частям, получим

$$\int\limits_0^1 p^{n_1}(1-p)^{n_2}dp = \frac{n-n_1}{n_1+1}\int\limits_0^1 p^{n_1+1}(1-p)^{n_2-1}dp =$$
$$= \frac{n-n_1}{n_1+1}\frac{1}{(n+1)\binom{n}{n_1+1}} = \frac{1}{(n+1)\binom{n}{n_1}}.$$

Лемма доказана. $\triangle$

Условная вероятность события $\omega_{n+1} = 1$ при известных исходах $\omega^n = \omega_1, \ \ldots, \ \omega_n$ равна

$$P\{\omega_{n+1} = 1|\omega^n\} = \frac{P(\omega^n 1)}{P(\omega^n)} = \frac{\frac{1}{(n+2)\binom{n+1}{n_1+1}}}{\frac{1}{(n+1)\binom{n}{n_1}}} = \frac{n_1+1}{n+2}.$$

Таким образом, получаем правило Лапласа:

$$P\{\omega_{n+1} = 1|\omega^n\} = \frac{n_1+1}{n+2},$$
$$P\{\omega_{n+1} = 0|\omega^n\} = \frac{n_2+1}{n+2}.$$

Качество такой процедуры прогнозирования можно оценивать с помощью какой-нибудь функции потерь. Пример – логарифмическая функция потерь:

$$L_p(\omega^n) = -\ln(p^{n_1}(1-p)^{n_2}).$$

Из теории информации известно, что эта величина с точностью до 1 совпадает со средним количеством двоичных битов, необходимых для кодирования последовательностей ω^n, состоящих из

n_1 единиц и n_2 нулей и порождаемых источником с вероятностью единицы, равной p. Нетрудно проверить, что

$$\sup_{0\leqslant p\leqslant 1} p^{n_1}(1-p)^{n_2} = \left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}.$$

Для правила Лапласа

$$L(\omega^n) = -\ln P(\omega^n) = -\ln \int\limits_0^1 p^{n_1}(1-p)^{n_2}dp.$$

Допустим, что последовательность ω^n порождена источником с вероятностью $\omega_n = 1$, равной p_0. Тогда для произвольной последовательности ω^n выполнено

$$L(\omega^n) - L_{p_0}(\omega^n) \leqslant$$

$$L(\omega^n) - \inf_{0\leqslant p\leqslant 1} L_p(\omega^n) = \ln \frac{\sup\limits_{0\leqslant p\leqslant 1} p^{n_1}(1-p)^{n_2}}{\int\limits_0^1 p^{n_1}(1-p)^{n_2}dp} =$$

$$= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{(n+1)\binom{n}{n_1}}} \leqslant \ln(n+1).$$

Таким образом, используя для кодирования вероятности, вычисленные по правилу Лапласа, мы истратим $\ln(n+1)$ дополнительных битов по сравнению с длиной оптимального кода, т.е. кода, построенного на основе истинной вероятности p_0 источника, порождающего исходы ω_i.

Другой, более точный, метод прогнозирования был предложен Кричевским и Трофимовым. Рассматривается байесовская смесь вероятностей последовательности длины n по всем возможным p с плотностью $1/(\pi\sqrt{p(1-p)})$:

$$P(\omega^n) = \int\limits_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}}dp.$$

В этом случае условная вероятность появления единицы после n наблюдений $\omega^n = \omega_1, \ \ldots, \ \omega_n$ равна

$$P(1|\omega^n) = \frac{n_1 + 1/2}{n+1}.$$

Имеет место оценка:

$$\int\limits_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp \geqslant \frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}.$$

Эти утверждения предлагается далее в разделе 3.6 в виде задач 1 и 2.

Отсюда получаем оценку на дополнительное число битов при кодировании с использованием прогнозирования по методу Кричевского и Трофимова:

$$L(\omega^n) - \inf_{0 \leqslant p \leqslant 1} L_p(\omega^n) = \ln \frac{\sup\limits_{0 \leqslant p \leqslant 1} p^{n_1}(1-p)^{n_2}}{\int_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp} \leqslant$$

$$\leqslant \ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}} \leqslant \ln(2\sqrt{n}) = \frac{1}{2} \ln n + \ln 2.$$

В этой оценке ошибка асимптотически в два раза меньше, чем в соответствующей оценке для метода Лапласа.

3.1.3. Оптимальный минимаксный предсказатель

Другой тип предсказателя – оптимальный минимаксный предсказатель Далее мы не предполагаем, что $\mathcal{F}$ – конечный класс.

Определим *минимаксный регрет*

$$V_n(\mathcal{F}) = \inf_{p \in \mathcal{F}} V_n(p, \mathcal{F}),$$

где

$$V_n(p, \mathcal{F}) = \sup_{y^n} (L_n^p(y^n) - \inf_f L_n^f(y^n)) =$$

$$= \sup_{y^n} \ln \frac{\sup_f f(y^n)}{p(y^n)}.$$

Определим

$$p^*(y^n) = \frac{\sup_f f(y^n)}{\sum\limits_{x^n} \sup_f f(x^n)}. \tag{3.1}$$

Из определения следует, что $p^*(\cdot)$ – распределение вероятностей на Ω^n. *Минимаксный предсказатель* определяется

$$p^*(y_i|y^{i-1}) = \frac{p^*(y^i)}{p^*(y^{i-1})}.$$

Теорема 3.1. *Пусть p^* – минимаксный предсказатель для класса предсказательных стратегии $\mathcal{F}$. Тогда имеет место свойство оптимальности $V_n(p^*, \mathcal{F}) = V_n(\mathcal{F})$. Кроме этого,*

$$V_n(\mathcal{F}) = \ln \sum_{x^n} \sup_f f(x^n).$$

Доказательство. Из определения

$$\ln \frac{\sup_f f(y^n)}{p^*(y^n)} = \sum_{x^n} \sup_f f(x^n). \tag{3.2}$$

Правая сумма из (3.2) не зависит от y^n, поэтому левая часть равенства также не зависит от y^n и

$$V_n(p^*, \mathcal{F}) = \sup_{y^n} \ln \frac{\sup_f f(y^n)}{p^*(y^n)} = \sum_{x^n} \sup_f f(x^n).$$

Докажем оптимальность предсказателя p^*, т.е. $V_n(p^*, \mathcal{F}) = V_n(\mathcal{F})$. Рассмотрим произвольный предсказатель $p \neq p^*$. Тогда, так как $\sum\limits_{x^n} p(x^n) = \sum\limits_{x^n} p^*(x^n) = 1$, существует y^n такое, что $p(y^n) < p^*(y^n)$. Отсюда

$$\ln \frac{\sup_f f(y^n)}{p(y^n)} > \ln \frac{\sup_f f(y^n)}{p^*(y^n)} =$$
$$= \sup_{x^n} \ln \frac{\sup_f f(x^n)}{p^*(x^n)} = V_n(p^*, \mathcal{F}).$$

Отсюда $V_n(p, \mathcal{F}) > V_n(p^*, \mathcal{F})$ при $p \neq p^*$. Следовательно,

$$V_n(p^*, \mathcal{F}) = \inf_p \sup_{y^n} \ln \frac{\sup_f f(y^n)}{p(y^n)} = V_n(\mathcal{F}).$$

Теорема доказана. $\triangle$

Для конечного класса $\mathcal{F}$ имеет место оценка

$$\begin{aligned} V_n(\mathcal{F}) &= \ln \sum_{x^n} \sup_f f(x^n) \leqslant \\ &\leqslant \ln \sum_{x^n} \sum_{f \in \mathcal{F}} f(x^n) = \\ &= \ln \sum_{f \in \mathcal{F}} \sum_{x^n} f(x^n) \leqslant \ln |\mathcal{F}|. \end{aligned}$$

3.1.4. Приложения

Рассмотрим некоторые конкретные примеры применения минимаксных предсказателей.

Избыточность кода. Под кодом (методом кодирования) понимаем последовательность функций $c_n : \Omega^n \to \{0,1\}^*$, $c_n(y^n)$ – кодовое слово состоящее из 0 и 1. Код называется префиксно-корректным, если для любого n кодовые слова $c_n(y^n)$ не являются продолжениями друг друга. В этом случае, легко заметить, что выполнено неравенство Крафта $\sum\limits_{y^n} 2^{-|c_n(y^n)|} \leqslant 1$, где $|x|$ – длина слова x. С другой стороны, по теореме Крафта (см. [23]) для любой предсказательной стратегии f существует код c_n такой, что $|c(y^n)| = \lceil -\log_2 f(y^n) \rceil$.

Задача универсального кодирования заключается в построении такого кодового распределения (стратегии) p^*, чтобы *избыточность кода*

$$\begin{aligned} &\sup_{y^n} (-\log_2 p_n^*(y^n) - \inf_{f \in \mathcal{F}} (-\log_2 f(y^n))) = \\ &\qquad = \sup_{y^n} \log_2 \frac{\sup_{f \in \mathcal{F}} f(y^n)}{p^*(y^n)} \end{aligned}$$

была бы минимальной, т.е.

$$V_n(p^*, \mathcal{F}) = V_n(\mathcal{F}) = \ln 2 \inf_{p \in \mathcal{F}} \sup_{y^n} \log_2 \frac{\sup_{f \in \mathcal{F}} f(y^n)}{p^*(y^n)}.$$

По теореме 3.1 соответствующее кодовое распределение задается формулой (3.1)

$$p^*(y^n) = \frac{\sup_f f(y^n)}{\sum\limits_{x^n} \sup_f f(x^n)}.$$

Код, соответствущий распределению p^*, называется универсальным для класса кодов, определяемых распределениями из класса $\mathcal{F}$. В данном случае избыточность кода с точностью до множителя $\ln 2$ совпадает с регретом соответствующего распределения, а длины кодовых слов, с точностью до регрета, не превосходят длины кодовых слов любого кода из $\mathcal{F}$ (пример см. в следствии 3.1 далее).

Впервые данный способ универсального кодирования предложен в работе Ю.М.Штарькова [5].

Игра Келли. Классическая модель игры Келли заключается в следующем. На каждом раунде i производится забег m лошадей, одна из которых выигрывает. Пусть $y^{i-1} = y_1, \ldots, y_{i-1}$ – номера выигрывших лошадей на раундах $1, \ldots, i-1$. Игрок на каждом раунде i вкладывает все свои денежные средства в ставки на лошадей в пропорциях, которые представлены в виде координат вектора стратегии $(f(1|y^{i-1}), \ldots, f(m|y^{i-1}))$. Сумма координат равна единице, а сами координаты есть неотрицательные вещественные числа. Функция выигрыша имеет вид

$$k(y|y_1, \ldots, y_{i-1}) = \begin{cases} k_i, \text{ если } y = y_i, \\ 0 \text{ в противном случае.} \end{cases}$$

Пусть начальный капитал игрока равен 1. Тогда выигрыш игрока применяющего стратегию вложения f за n раундов равен

$$K_n(f, y^n) = \prod_{i=1}^{n} f(y_i|y^{i-1}) k(y_i|y^{i-1}).$$

Заметим, что отношение выигрышей стратегий f и p

$$\frac{K_n(f,y^n)}{K_n(p,y^n)} = \frac{\prod_{i=1}^n f(y_i|y^{i-1})k(y_i|y^{i-1})}{\prod_{i=1}^n p(y_i|y^{i-1})k(y_i|y^{i-1})} = \frac{f(y^n)}{p(y^n)}$$

не зависит от функции выигрыша. По определению $V_n(p,\mathcal{F}) = \sup_{y^n}\sup_f \ln\frac{f(y^n)}{p(y^n)}$ и $V_n(\mathcal{F}) = \inf_p V_n(p,\mathcal{F})$. Называем оптимальной такую стратегию вложения p^*, для которой $V_n(p^*,\mathcal{F}) = V_n(\mathcal{F})$.

Во всех этих примерах возникает задача оценки величины $V_n(\mathcal{F})$ для соответствующих классов стратегий $\mathcal{F}$.

Стратегии не зависящие от истории. Оценим величину $V_n(\mathcal{F})$ для одного простого класса стратегий. Пусть класс $\mathcal{F}$ состоит из распределений вероятностей $f(y)$ на множестве $\Omega = \{1,2\}$. Тогда $\mathcal{F} = \{(q,1-q) : 0 \leqslant q \leqslant 1\}$, $q = f(1)$.

Теорема 3.2. *$V_n(\mathcal{F}) = \frac{1}{2}\ln n + \frac{1}{2}\ln\frac{\pi}{2} + \epsilon_n$, где $\epsilon_n \to 0$ при $n \to \infty$.*

Доказательство. По теореме 3.1

$$V_n(\mathcal{F}) = \ln\sum_{y^n}\sup_{f\in\mathcal{F}} f(y^n).$$

здесь $y^n \in \{1,2\}$, пусть n_1 – число 1ц в y^n и n_2 – число 2ек, $n_2 = n - n_1$. Тогда $f(y^n) = q^{n_1}(1-q)^{n_2}$, $\sup_f f(y^n) = \left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}$ и

$$V_n(\mathcal{F}) = \ln\sum_{y^n}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2},$$

где $n_i = n_i(y^n)$. Можно представить

$$V_n(\mathcal{F}) = \ln\sum_{n_1=1}^{n-1}\binom{n}{n_1}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}.$$

Из формулы Стирлинга

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+\epsilon}},$$

где $\epsilon > 0$ – константа. Тогда

$$\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{n_1n_2}}e^{\frac{1}{12n}} \leqslant \binom{n}{n_1}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2} \leqslant \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{n_1n_2}}e^{\frac{1}{12n+1}}.$$

Отсюда, в частности, получаем верхнюю оценку

$$V_n(\mathcal{F}) \leqslant \ln \sqrt{\frac{n}{2\pi}} e^{\frac{1}{12n+1}} \sum_{n_1=1}^{n-1} \frac{1}{\sqrt{n_1 n_2}}.$$

Сумма оценивается через интеграл

$$\sum_{n_1=1}^{n-1} \frac{1}{\sqrt{n_1 n_2}} =$$

$$= \sum_{n_1=1}^{n-1} \frac{1}{n} \frac{1}{\sqrt{\frac{n_1}{n}(1-\frac{n_1}{n})}} \approx$$

$$\approx \int_0^1 \frac{dx}{\sqrt{x(1-x)}} = \pi.$$

Отсюда получаем

$$V_n(\mathcal{F}) \leqslant \ln \left((1+o(1)) \sqrt{\frac{n\pi}{2}} \right) = \frac{1}{2} \ln n + \frac{1}{2} \ln \frac{\pi}{2} + \epsilon_n,$$

где $\epsilon_n \to 0$ при $n \to \infty$. Нижняя оценка может быть получена аналогичным образом. Теорема доказана. $\triangle$

Теорема 3.2 может быть сформулирована также в виде:

Следствие 3.1. *Можно построить оптимальную стратегию $p^* \in \mathcal{F}$ такую, что для любой стратегии $f \in \mathcal{F}$ и для каждой последовательности y^n выполнено*

$$L_n^{p^*}(y^n) \leqslant L_n^f(y^n) + \frac{1}{2} \ln n + \frac{1}{2} \ln \frac{\pi}{2} + \epsilon_n,$$

где $\epsilon_n \to 0$ при $n \to \infty$.

В частности, избыточность оптимального кода для класса кодовых распределений $\mathcal{F}$ ограничена величиной $\frac{1}{2}\log_2 n + \frac{1}{2}\log_2 \frac{\pi}{2} + \epsilon_n$.

Выигрыш игрока в игре Келли, применяющего стратегию p^*, удовлетворяет неравенству

$$K_n(p^*, y^n) \geqslant (1-\epsilon_n) n^{-\frac{1}{2}} \sup_{f \in \mathcal{F}} K_n(f, y^n),$$

где $\epsilon_n \to 0$ при $n \to \infty$.

Более подробное изложение материалов этого раздела можно найти в монографиях [23] и [43].

3.2. Калибруемость прогнозов

В том случае, когда отсутствует гипотеза о механизме порождения исходов ω_i, для оценки качества прогнозов используются целевые функционалы (функции потерь), которые выбираются исходя из конкретных задач, для решения которых производится прогнозирование.

Типичным примером задачи на прогнозирование является задача предсказания погоды на завтра, например, событие $\omega_n = 1$ может интерпретироваться как дождь в n-й день, а число p_n – как его вероятность, вычисленная на основе наблюдений погоды $\omega_1, \omega_2, \ \ldots, \ \omega_{n-1}$ за предыдущие $n-1$ дней.

Предсказатель погоды считается хорошо калибруемым, если дождь случается так же часто, как он прогнозируется предсказателем. Например, если дождь случается в 80% всех дней, для которых предсказатель давал прогноз $p_n = 0.8$, и т.д. Величина среднего отклонения частоты исходов ω_n от прогнозов p_n, где $p_n \approx p^*$, для различных значений p^* может использоваться как тест для выявления "плохих" предсказателей.

В предыдущем примере $p_n \in [0,1]$. Можно рассматривать последовательности данных более общего характера. Например, пусть $\omega_n = S_n$ – цена некоторого финансового инструмента в некоторые последовательные моменты времени $n = 1, 2, \ \ldots$. Цена имеет стохастический характер изменения. В практических приложениях иногда трудно восстановить параметры модели, управляющей изменением цены. Кроме этого, эти параметры могут изменяться со временем. Число p_n рассматривается как прогноз "среднего значения" этой величины на шаге n.

В разделе 3.3 мы будем рассматривать как бинарные исходы $\omega_n \in \{0,1\}$, так и вещественные исходы, лежащие в единичном интервале: $\omega_n \in [0,1]$; число p_n лежит в единичном отрезке $[0,1]$.

Если бы задача восстановления истинных значений вероятностей p_i решалась традиционными статистическими методами, то

мы бы предполагали, что исходы генерируются с помощью некоторой вероятностной меры P, т.е.

$$p_n = P(\omega_n = 1 | \omega_1, \omega_2, \ \ldots, \ \omega_{n-1}) \text{ при } n = 1, 2, \ \ldots$$

– условная вероятность события $\omega_n = 1$ при известных значениях $\omega_1, \omega_2, \ \ldots, \ \omega_{n-1}$. В этом случае предсказатель должен был бы решать классическую задачу математической статистики – восстановление вероятностной меры P по наблюдениям. Обычно при этом класс возможных мер сильно ограничивается на основе некоторой априорной информации об источнике. Например, предполагается, что распределение принадлежит заданному параметрическому классу и мы должны по наблюдениям восстановить некоторый неизвестный параметр этой меры.

Однако на практике мы часто имеем дело с единственной исторической последовательностью исходов $\omega_1, \omega_2, \ \ldots, \ \omega_{n-1}, \ \ldots$ и не имеем представления о механизмах, генерирующих эту последовательность. Мы даже можем не знать, являются ли эти механизмы стохастическими.

В данной главе предположение о наличии такой меры P не используется. В условиях отсутствия меры возникает естественная трудность – неизвестно, каким образом оценивать качество прогнозов. Требуются критерии качества, использующие только последовательность данных, получаемую предсказателем в режиме онлайн.

Тем не менее, можно указать метод прогнозирования произвольной последовательности $\omega_1, \omega_2, \ \ldots, \ \omega_{n-1}$, удовлетворяющий так называемым тестам на *калибруемость*.

Приведенное выше правило проверки предсказателя погоды можно записать в следующем виде: для любого действительного числа $p^* \in [0, 1]$ выполнено

$$\frac{\sum_{i=1 \& p_i \approx p^*}^{n} \omega_i}{\sum_{i=1 \& p_i \approx p^*}^{n} 1_{p_i \approx p^*}} \approx p^*, \tag{3.3}$$

если знаменатель отношения (3.3) стремится к бесконечности при $n \to \infty$. Здесь мы использовали символ $\approx$ приближенного равенства, потому что на практике число p^* можно задавать только с

FOR $n = 1, 2, \ \ldots$
Предсказатель анонсирует прогноз $p_n \in [0, 1]$.
Природа анонсирует исход $\omega_n \in \{0, 1\}$.
ENDFOR

Рис. 3.1: Протокол игры с детерминированными предсказаниями

некоторой точностью. Условие $p_i \approx p^*$ требует дальнейшего уточнения.

Уточним схему предсказания в виде протокола игры с участием двух игроков *Предсказателя* и *Природы*. Протокол представлен на рис. 3.1.

Игроки – *Предсказатель* и *Природа* – могут использовать всю информацию, которая известна на момент его или ее действия. В частности, на шаге n *Природа* может использовать прогноз p_n, анонсированный *Предсказателем*; *Предсказатель* не знает исход ω_n, так как к моменту выдачи прогноза p_n *Природа* еще не анонсировала свой исход.

Приведем точное определение калибруемости, предложенное Дейвидом [25]. Рассмотрим произвольные подынтервалы I вида $[a, b]$, $(a, b]$, $[a, b)$, (a, b) интервала $[0, 1]$ и их характеристические функции

$$I(p) = \begin{cases} 1, \text{ если } p \in I, \\ 0 \text{ в противном случае.} \end{cases}$$

Последовательность прогнозов $p_1, p_2, \ldots$ *калибруется* на бесконечной последовательности $\omega_1, \omega_2, \ldots$, если для характеристической функции $I(p)$ каждого подынтервала $[0, 1]$ *калибровочная ошибка* стремится к нулю, т.е.

$$\frac{\sum_{i=1}^{n} I(p_i)(\omega_i - p_i)}{\sum_{i=1}^{n} I(p_i)} \longrightarrow 0, \tag{3.4}$$

если знаменатель отношения (3.4) стремится к бесконечности при $n \to \infty$. Характеристическая функция $I(p_i)$ определяет некоторое правило выбора, которое определяет те номера исходов i, для которых мы вычисляем отклонение прогноза p_i от соответствующего исхода ω_i.

Дейвид рассматривал задачу построения “универсального предсказателя” – алгоритма, который для любой последовательности исходов выдает калибруемые прогнозы.

Простые соображения показывают, что никакой алгоритм, вычисляющий прогнозы на основании прошлых исходов, не может всегда выдавать калибруемые прогнозы. А именно, для произвольного такого алгоритма f можно определить последовательность $\omega = \omega_1, \omega_2, \ldots$ так, что

$$\omega_i = \begin{cases} 1, \text{ если } p_i < \frac{1}{2}, \\ 0 \text{ в противном случае,} \end{cases}$$

где $p_i = f(\omega_1, \ldots, \omega_{i-1})$, $i = 1, 2, \ \ldots$. Легко видеть, что для интервала $I = [0, \frac{1}{2})$ или для интервала $I = [\frac{1}{2}, 1)$ условие калибруемости (3.4) нарушается.

Данная последовательность $\omega = \omega_1, \omega_2, \ldots$ является простейшим примером реализации “адаптивно враждебной” стратегии *Природы*. “Адаптивность” заключается в том, что на каждом шаге i *Природа* выбирает исход p_i в зависимости от прогноза p_i, анонсированного *Предсказателем*. “Враждебность” заключается в том, что исход ω_i выбирается так, чтобы $|\omega_i - p_i| \geqslant \frac{1}{2}$.

Подобные трудности построения универсального предсказателя оказались преодолимыми с помощью понятия рандомизированного прогноза. Пусть $\mathcal{P}[0, 1]$ – множество всех вероятностных мер на отрезке $[0, 1]$.

Протокол игры с детерминированными предсказаниями, приведенный на рис. 3.1, заменяется на протокол игры с рандомизированными предсказаниями, см. рис. 3.2. При этом вводится вспомогательный игрок – *Генератор случайных чисел*. *Природа* при выборе очередного исхода ω_n может использовать распределение вероятностей, анонсированное *Предсказателем*, но прогноз p_n скрыт от нее. Поэтому ход *Генератора случайных чисел* помещен в протоколе после хода *Природы*.[3]

Обозначаем $\omega^{n-1} = \omega_1, \ldots, \omega_{n-1}$. Вероятностные распределения P_n, в общем случае зависящие от ω^{n-1}, порождают распре-

[3] Чисто игровая формулировка без использования понятия вероятности будет приведена в разделе 8.3.

FOR $n = 1, 2, \ldots$
Предсказатель анонсирует распределение вероятностей $P_n \in \mathcal{P}([0,1])$.
Природа анонсирует исход $\omega_n \in \{0,1\}$.
Генератор случайных чисел анонсирует случайное число $\tilde{p}_n$, распределенное согласно мере P_n.
ENDFOR

Рис. 3.2: Протокол игры с рандомизированными предсказаниями

деление вероятностей P на множестве всех бесконечных последовательностей прогнозов $\tilde{p}_1, \tilde{p}_2, \ldots$, где $\tilde{p}_i \in [0,1]$, $i = 1, 2, \ldots$. Бесконечная последовательность $\omega = \omega_1, \omega_2, \ldots$ является параметром этого распределения.

Заметим, что такая мера P существует и в гораздо более общем случае, а именно, когда каждый исход ω_n является измеримой функцией от последовательности $\tilde{p}_1, \ldots, \tilde{p}_{n-1}$ для всех [4] n.

В этом случае, для любого n определено семейство вероятностных мер $P_n(d\tilde{p}_n; \tilde{p}_1, \ldots, \tilde{p}_{n-1})$ (вероятностных ядер), зависящих от параметров $\tilde{p}_1, \ldots, \tilde{p}_{n-1} \in [0,1]$. В частности, определена начальная мера $P_1(d\tilde{p}_1)$.

По этому семейству для любого n определяется вероятностная мера Q_n на $[0,1]^n$ следующим образом. Для любого борелевского множества $A \subseteq [0,1]^n$ определим

$$Q_n(A) =$$
$$= \int P_1(d\tilde{p}_1) \int P_2(d\tilde{p}_2; p_1) \ldots \int P_n(d\tilde{p}_n; \tilde{p}_1, \ldots, \tilde{p}_{n-1}) 1_A(\tilde{p}_1, \ldots, \tilde{p}_n),$$

где

$$1_A(\tilde{p}_1, \ldots, \tilde{p}_n) = \begin{cases} 1, \text{ если } (\tilde{p}_1, \ldots, \tilde{p}_n) \in A, \\ 0 \text{ в противном случае.} \end{cases}$$

По теореме Ионеско–Тульчи [4] о продолжении меры существует вероятностная мера P на множестве $[0,1]^\infty$ всех бесконечных

[4] Согласно протоколу, представленному на рис. 3.2, *Природа* наблюдает на шаге n значения прогнозов $\tilde{p}_1, \ldots, \tilde{p}_{n-1}$, выдаваемых *Генератором случайных чисел.* на прошлых шагах. При этом значение прогноза $\tilde{p}_n$ ей недоступно.

траекторий $\tilde{p}_1, \tilde{p}_2, \ \ldots$ такая, что

$$Q_n(A) = P(A \times [0,1]^\infty)$$

для всех n и всех борелевских $A \subseteq [0,1]^n$.

В частности, можно для любого подынтервала $I \subseteq [0,1]$ рассматривать вероятность P события (3.4).

Фостер и Вохра [28] построили алгоритм для вычисления калибруемых рандомизированных предсказаний для случая, когда исходы ω_i принимают конечное число значений.

Приведем этот результат для случая, когда исходы ω_i принимают только два значения 0 и 1: для произвольного значения параметра $\Delta > 0$ алгоритм генерирует рандомизированные предсказания так, что

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \right| \leqslant \Delta$$

выполнено с вероятностью 1, где траектории прогнозов $\tilde{p}_1, \tilde{p}_2, \ \ldots$ распределены по вероятностной мере P, а $I(p)$ – характеристическая функция произвольного подынтервала $[0,1]$.

Версия этого алгоритма для случая $\omega_i \in [0,1]$ будет приведена в следующем разделе 3.3.

3.3. Алгоритм вычисления калибруемых прогнозов

Приведем некоторый модернизированный вариант рандомизированного алгоритма Какаде и Фостера вычисления калибруемых предсказаний, предложенный Какаде и Фостером [34].

Пусть $\omega_1, \omega_2, \ldots$ – произвольная последовательность элементов $\{0,1\}$ или $[0,1]$, поступающая в режиме онлайн. Построим алгоритм для вычисления случайной величины, выдающей прогноз $\tilde{p}_n \in [0,1]$ будущего значения ω_n по начальному фрагменту $\omega_1, \ldots, \omega_{n-1}$. Основное требование к таким прогнозам: они должны с вероятностью 1 удовлетворять условию калибруемости. Соответствующее распределение вероятностей является внутренним по отношению к алгоритму и строится в процессе конструкции.

Предварительно разобьем интервал значений прогнозов $[0,1]$ на равные части длины $\Delta = 1/K$ с помощью рациональных точек $v_i = i\Delta$, где $i = 0, 1, \ \ldots, \ K$. Пусть V обозначает множество всех этих точек. Любое число $p \in [0,1]$ представляется в виде линейной комбинации граничных точек подынтервала разбиения, содержащего p :

$$p = \sum_{v \in V} w_v(p) v = w_{v_{i-1}}(p) v_{i-1} + w_{v_i}(p) v_i,$$

где $p \in [v_{i-1}, v_i]$, $i = \lfloor p/\Delta + 1 \rfloor$ и

$$w_{v_{i-1}}(p) = 1 - \frac{p - v_{i-1}}{\Delta}, \ w_{v_i}(p) = 1 - \frac{v_i - p}{\Delta}.$$

Полагаем $w_v(p) = 0$ для всех остальных значений $v \in V$.

В дальнейшем детерминированный прогноз p, выдаваемый алгоритмом, приведенным далее, будет округляться до v_{i-1} с вероятностью $w_{v_{i-1}}(p)$ и до v_i с вероятностью $w_{v_i}(p)$.

Сначала построим алгоритм, выдающий детерминированные прогнозы.

Пусть прогнозы $p_1, \ \ldots, \ p_{n-1}$ уже определены (пусть $p_1 = 0$). Вычислим прогноз p_n.

Рассмотрим вспомогательную величину

$$\mu_{n-1}(v) = \sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i).$$

Имеем

$$(\mu_n(v))^2 = (\mu_{n-1}(v))^2 + 2 w_v(p_n) \mu_{n-1}(v)(\omega_n - p_n) + \\ + (w_v(p_n))^2 (\omega_n - p_n)^2. \qquad (3.5)$$

Суммируем (3.5) по v :

$$\sum_{v \in V} (\mu_n(v))^2 = \sum_{v \in V} (\mu_{n-1}(v))^2 + \\ + 2(\omega_n - p_n) \sum_{v \in V} w_v(p_n) \mu_{n-1}(v) + \\ + \sum_{v \in V} (w_v(p_n))^2 (\omega_n - p_n)^2. \qquad (3.6)$$

Изменим порядок суммирования в сумме вспомогательных величин

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) =$$
$$= \sum_{v\in V} w_v(p) \sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i) =$$
$$= \sum_{i=1}^{n-1} (\sum_{v\in V} w_v(p) w_v(p_i))(\omega_i - p_i) =$$
$$= \sum_{i=1}^{n-1} (\bar{w}(p) \cdot \bar{w}(p_i))(\omega_i - p_i) =$$
$$= \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i),$$

где

$$\bar{w}(p) = (w_1,\ \ldots,\ w_{v_K}) = (0,\ \ldots,\ w_{v_{i-1}}(p), w_{v_i}(p),\ \ldots,\ 0)$$

– вектор вероятностей округления, $p \in [v_{i-1}, v_i]$ и

$$K(p, p_i) = (\bar{w}(p) \cdot \bar{w}(p_i)) \tag{3.7}$$

– скалярное произведение соответствующих векторов (ядро). По определению $K(p, p_i)$ – непрерывная функция.

Второй член правой части равенства (3.6) при подходящем значении p_n всегда можно сделать меньшим или равным нулю. Действительно, в качестве p_n берем корень $p_n = p$ уравнения

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i) = 0, \tag{3.8}$$

если он существует. В противном случае если левая часть уравнения (3.8) (которая является непрерывной по p функцией) больше нуля для всех значений p_n, то полагаем $p_n = 1$, а если она меньше нуля, то полагаем $p_n = 0$. Определенное таким образом значение p_n выдаем в качестве детерминированного прогноза.

Третий член (3.6) ограничен числом 1. Действительно, так как $|\omega_i - p_i| \leqslant 1$ для всех i, имеем для произвольного n

$$\sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2 \leqslant \sum_{v\in V} w_v(p_n) = 1.$$

Отсюда и по (3.6), если последовательно выбирать прогнозы p_i согласно указанному правилу, получим

$$\sum_{v\in V}(\mu_n(v))^2 \leqslant \sum_{i=1}^{n}\sum_{v\in V}(w_v(p_i))^2(\omega_i - p_i)^2 \leqslant n. \tag{3.9}$$

Пусть теперь $\tilde{p}_i$ – случайная величина, принимающая значения $v \in V$ с вероятностями $w_v(p_i)$ (на самом деле, для каждого p ненулевыми являются только значения $w_v(p)$ для двух соседних границ подынтервала разбиения, содержащего детерминированный прогноз p_i). Пусть также $I(p)$ – характеристическая функция произвольного подынтервала $[0, 1]$. Для любого i математическое ожидание случайной величины $I(\tilde{p}_i)(\omega_i - \tilde{p}_i)$ равно [5]

$$E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) = \sum_{v\in V} w_v(p_i)I(v)(\omega_i - v). \tag{3.10}$$

Согласно усиленному мартингальному закону больших чисел (см. следствие 10.6 ниже) с вероятностью 1 :

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) - \frac{1}{n}\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))\right| \to 0 \tag{3.11}$$

при $n \to \infty$.

По определению детерминированного прогноза p_i и функции $w_v(p)$

$$\left|\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v) - \sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i)\right| < \Delta \tag{3.12}$$

[5] В случае когда ω_i есть функция от значений прошлых прогнозов, здесь имеется в виду условное математическое ожидание относительно случайных величин $\tilde{p}_1, \ldots, \tilde{p}_{i-1}$.

для каждого i.

Применяем неравенство Коши–Буняковского к векторам $\{\mu_n(v) : v \in V\}$ и $\{I(v) : v \in V\}$, учитываем (3.12) и получаем

$$\left|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i)\right| =$$
$$= \left|\sum_{v\in V} I(v)\sum_{i=1}^{n} w_v(p_i)(\omega_i - p_i)\right| \leqslant$$
$$\leqslant \sqrt{\sum_{v\in V}(\mu_n(v))^2}\sqrt{\sum_{v\in V} I(v)} \leqslant$$
$$\leqslant \sqrt{Kn}, \tag{3.13}$$

где $K = 1/\Delta$ - число подынтервалов разбиения.

Используя (3.12) и (3.13), получаем верхнюю оценку для абсолютной величины суммы математических ожиданий (3.10) :

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))\right| =$$
$$= \left|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v)\right| \leqslant \tag{3.14}$$
$$\leqslant \Delta n + \sqrt{n/\Delta} \tag{3.15}$$

для всех n.

Из (3.14) и (3.11) получаем, что с вероятностью 1:

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leqslant \Delta. \tag{3.16}$$

Сформулируем результаты этого раздела в виде следующей теоремы.

Теорема 3.3. *Для каждого $\Delta > 0$ можно построить алгоритм, выдающий рандомизированные прогнозы, такой, что для любого*

подынтервала $I \subseteq [0,1]$ *неравенство*

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \right| \leqslant \Delta$$

выполнено почти всюду, где $I(p)$ *– характеристическая функция этого подынтервала.*

Если в процессе конструкции в определенные моменты времени n_s, $s = 1, 2, \ldots$, изменять $\Delta = \Delta_s$ так, чтобы $\Delta_s \to 0$ при $s \to \infty$, можно достичь асимптотически точного результата:

Теорема 3.4. *Можно построить алгоритм, выдающий рандомизированные прогнозы, такой, что для любого подынтервала* $I \subseteq [0,1]$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) = 0$$

почти всюду.

Детали конструкции приведены в разделе 3.5.1.

Прогнозы, удовлетворяющие условию теоремы 3.4, будут называться *хорошо калибруемыми.*

3.4. Прогнозирование с произвольным ядром

Существуют два подхода к универсальному прогнозированию:

- универсальное прогнозирование, при котором в качестве прогноза выдается распределение вероятностей на множестве возможных прогнозов (в частности, в случае бинарных последовательностей множество возможных прогнозов состоит из двух элементов, как в предыдущем разделе); при этом в качестве правил выбора используются произвольные подынтервалы единичного интервала;

- универсальное прогнозирование, при котором прогноз является детерминированным, однако в качестве правил выбора разрешается использовать только гладкие приближения к характеристическим функциям подынтервалов единичного интервала.

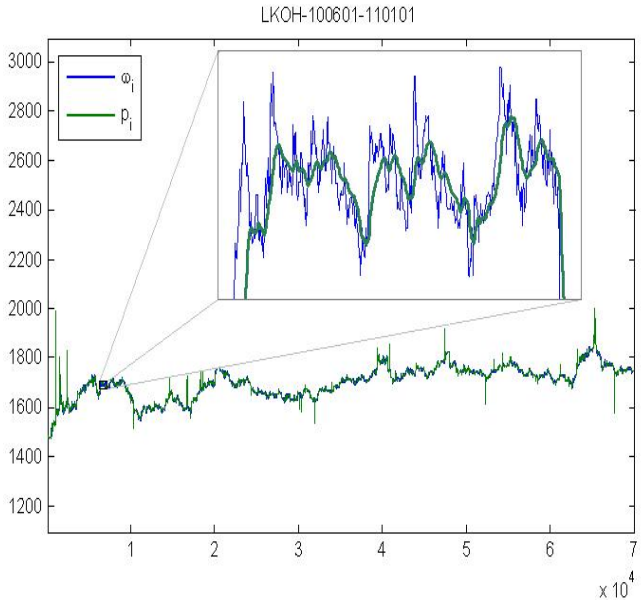

Рис. 2.1. Пример последовательности данных $\omega_1, \omega_2, \ldots$ и последовательности калибруемых прогнозов $p_1, p_2, \ldots$

В первом случае последовательность прогнозов удовлетворяет условию калибруемости с вероятностью единица. Во втором случае условие калибруемости просто выполнено для последовательности детерминированных прогнозов с гладкими весами.

Можно показать, что оба эти подхода по существу эквивалентны (см. [34]).

Второй метод прогнозирования будет рассмотрен в этом разделе.

Метод построения алгоритмов универсального прогнозирования, предложенный в работах Фостера и Вохры [28], а также Какаде и Фостера [34], был обобщен В. Вовком на случай произвольных ядер в работах [64] и [68]. В этом разделе мы представим основную идею этого обобщения.

Предварительно сформулируем задачу прогнозирования в ви-

FOR $n = 1, 2, \ \ldots$
Природа анонсирует сигнал $\bar{x}_n \in X$.
Скептик анонсирует непрерывную по p функцию $S_n : [0,1] \to \mathcal{R}$.
Предсказатель анонсирует прогноз $p_n \in [0,1]$.
Природа анонсирует исход $y_n \in \{0,1\}$.
Скептик вычисляет свой выигрыш на шаге n игры
$\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$.
ENDFOR

Рис. 3.3: Алгоритм для вычисления детерминированных предсказаний в случае произвольного ядра

де некоторой абстрактной игры между игроками: *Природа*, *Предсказатель* и *Скептик*. Игрок *Скептик* будет определять цель игры. На каждом шаге n *Скептик* анонсирует непрерывную функцию $S_n(p)$ и после того как *Предсказатель* и *Природа* объявили свои ходы подсчитывает свой выигрыш как это указано на рис. 3.3.

Игра регулируется протоколом, приведенным на рис. 3.3.

В этой игре прогнозы будут детерминированными, подобно прогнозам p_i, которые вычисляются в виде корней уравнений типа (3.8) в разделе 3.3. Мы рассмотрим более общую постановку, а именно, введем дополнительную информацию – *сигналы*.

Задано множество сигналов $X \subseteq \mathcal{R}^m$ – множество m-мерных векторов $\bar{x} = (x_1, \ \ldots, \ x_m)$, на котором рассматривается обычная m-мерная евклидова норма

$$\|\bar{x}\| = \sqrt{\sum_{i=1}^{m} x_i^2}.$$

Сигналы можно интерпретировать как дополнительную информацию, которая поступает *Предсказателю* в режиме онлайн.

Полагаем начальный выигрыш *Скептика* $\mathcal{K}_0 = 1$.

Следующая теорема показывает, что *Предсказатель* имеет стратегию, при которой выигрыш *Скептика* не возрастает как бы *Природа* не выбирала свои ходы.

Теорема 3.5. *Предсказатель имеет стратегию для вычисления прогнозов p_n, при которой $\mathcal{K}_0 \geqslant \mathcal{K}_1 \geqslant \ldots \mathcal{K}_n \geqslant \ldots$*

Доказательство. Стратегия *Предсказателя* заключается в следующем.

На произвольном шаге n игры *Предсказатель* вычисляет свой прогноз p_n следующим образом.

Пусть $S_n(p)$ – непрерывная по p функция, анонсированная *Скептиком* согласно протоколу на рис. 3.3. Если $S_n(p)$ положительно для всех $p \in [0,1]$, то полагаем $p_n = 1$. Если $S_n(p)$ отрицательно для всех $p \in [0,1]$, то полагаем $p_n = 0$. В противном случае из теоремы о промежуточных значениях следует, что уравнение

$$S_n(p) = 0, \tag{3.17}$$

рассматриваемое относительно p, имеет корень. В этом случае *Предсказатель* выбирает в качестве p_n один из таких корней.

Легко видеть, что при таком выборе p_n выигрыш *Скептика* всегда не возрастает, как бы он не выбирал непрерывную по p функцию $S_n(p)$ и как бы *Природа* не выбирала свои ходы, т.е. всегда выполнено

$$\mathcal{K}_0 \geqslant \mathcal{K}_1 \geqslant \ldots \mathcal{K}_n \geqslant \ldots$$

для всех n. $\triangle$

Мы будем использовать ядро $K((p,\bar{x}),(p',\bar{x}'))$ - симметричную положительно определенную вещественную гладкую функцию на $([0,1] \times X)^2$. Пример ядра – гауссово ядро:

$$K((p,\bar{x}),(p',\bar{x}')) = \\ = \exp\left(-\frac{(p-p_i)^2}{\sigma_1^2} - \frac{\|\bar{x}-\bar{x}'\|^2}{\sigma_2^2}\right), \tag{3.18}$$

где σ_1, σ_2 – параметры ядра.

Рассмотрим следующую стратегию *Скептика*, которая будет вынуждать *Предсказателя* делать на каждом шаге n “хорошо калибруемые” прогнозы. Пусть в начале шага n прогнозы $p_1, \ldots, p_{n-1}$ уже известны. Рассмотрим функцию

$$S_n(p) = \sum_{i=1}^{n-1} K((p,\bar{x}_n),(p_i,\bar{x}_i))(y_i - p_i).$$

Пусть *Предсказатель* использует стратегию, описанную в теореме 3.5. Тогда выигрыш *Скептика* за N шагов игры удовлетворяет соотношениям

$$\mathcal{K}_N - \mathcal{K}_0 = \sum_{n=1}^{N} S_n(p_n)(y_n - p_n) =$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{n-1} K((p_n, \bar{x}_n), (p_i, \bar{x}_i))(y_i - p_i)(y_n - p_n) =$$

$$= \frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{N} K((p_n, \bar{x}_n), (p_i, \bar{x}_i))(y_i - p_i)(y_n - p_n) -$$

$$-\frac{1}{2}\sum_{n=1}^{N} K((p_n, \bar{x}_n), (p_n, \bar{x}_n))(y_n - p_n)^2. \qquad (3.19)$$

Согласно результатам раздела 2.5 существуют гильбертово пространство признаков $\mathcal{H}$ и отображение $\Phi : [0,1] \times X \to \mathcal{H}$ такое, что

$$K(a, b) = (\bar{\Phi}(a) \cdot \bar{\Phi}(b))$$

при $a, b \in [0,1] \times X$, где точкой обозначено скалярное произведение в пространстве $\mathcal{H}$ (далее $\|\cdot\|$ – соответствующая норма).

Величина $c_{\mathcal{H}} = \sup\limits_{a} \|\Phi(a)\|$ называется константой вложения (embedding constant). Мы рассматриваем ядра, для которых соответствующая величина конечна: $c_{\mathcal{H}} < \infty$.

Перепишем (3.19) в виде

$$\mathcal{K}_N - \mathcal{K}_0 = \frac{1}{2}\left|\left|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right|\right|^2 -$$

$$-\frac{1}{2}\sum_{n=1}^{N} ||\bar{\Phi}((p_n, \bar{x}_n)(y_n - p_n)||^2. \qquad (3.20)$$

По предположению

$$c_{\mathcal{H}} = \sup_{p,\bar{x}} ||\bar{\Phi}(p, \bar{x})|| < \infty.$$

По теореме 3.5 неравенство $\mathcal{K}_N - \mathcal{K}_0 \leqslant 0$ выполнено для всех N. Тогда из (3.20) следует

$$\frac{1}{2}\left|\left|\sum_{n=1}^{N}\bar{\Phi}(p_n,\bar{x}_n)(y_n-p_n)\right|\right|^2 \leqslant \frac{1}{2}NC^2. \tag{3.21}$$

Неравенство (3.21) перепишем в виде

$$\left|\left|\sum_{n=1}^{N}\bar{\Phi}(p_n,\bar{x}_n)(y_n-p_n)\right|\right| \leqslant \sqrt{N}C. \tag{3.22}$$

Иными словами, средняя ошибка алгоритма предсказания ограничена

$$\frac{1}{N}\left|\left|\sum_{n=1}^{N}\bar{\Phi}(p_n,\bar{x}_n)(y_n-p_n)\right|\right| \leqslant \frac{C}{\sqrt{N}}.$$

Используя полученную оценку средней ошибки алгоритма предсказания, получим результат о калибруемости, аналогичный результату из раздела 3.3. Для этого возьмем в качестве ядра какое-нибудь семейство гладких приближений к характеристическим функциям одноэлементных множеств $\{(p^*,\bar{x}^*)\}$, где $(p^*,\bar{x}^*) \in [0,1]\times X$, т.е. семейство функций вида

$$K((p^*,\bar{x}^*),(p,\bar{x})) = I_{(p^*,\bar{x}^*)}(p,\bar{x}). \tag{3.23}$$

Примером такого семейства $I_{p^*}(p)$ является семейство гауссовых ядер типа (3.18).

Для прогнозов p_i будет выполнено

Следствие 3.2.

$$\left|\frac{1}{N}\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)\right| \leqslant \frac{C^2}{\sqrt{N}} \tag{3.24}$$

для каждой точки $(p^*,\bar{x}^*) \in [0,1]\times X$.

Доказательство. По свойству ядра существует такая функция $\Phi(p, \bar{x})$ со значениями в некотором гильбертовом пространстве признаков $\mathcal{H}$, что

$$K((p^*, \bar{x}^*), (p, \bar{x})) = I_{(p^*, \bar{x}^*)}(p, \bar{x}) = (\bar{\Phi}(p^*, \bar{x}^*) \cdot \bar{\Phi}(p, \bar{x})).$$

Применим неравенство Коши–Буняковского к неравенству (3.22) и получим

$$\left| \sum_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n)(y_n - p_n) \right| =$$

$$= \left| \left(\left(\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n) \right) \cdot \bar{\Phi}(p^*, \bar{x}^*) \right) \right| \leqslant$$

$$\leqslant \left| \left| \sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n) \right| \right| ||\bar{\Phi}(p^*, \bar{x}^*)|| \leqslant C^2 \sqrt{N}.$$

Отсюда получаем (3.24). $\triangle$

Величина

$$\sum_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n)$$

является гладким аналогом числа пар (p_n, x_n), находящихся в "мягкой" окрестности пары $(p^*, \bar{x}^*)$.

Неравенство (3.24) можно переписать в виде

$$\left| \frac{\sum\limits_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n)(y_n - p_n)}{\sum\limits_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n)} \right| \leqslant \frac{C^2 \sqrt{N}}{\sum\limits_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n)}. \qquad (3.25)$$

Оценка (3.25) имеет смысл при

$$\sum_{n=1}^{N} I_{(p^*, \bar{x}^*)}(p_n, \bar{x}_n) \gg \sqrt{N},$$

т.е. сходимость частот к прогнозам будет иметь место только в подпоследовательностях "статистически значимой" длины.

Представленный в этом разделе алгоритм универсального прогнозирования можно легко реализовать в виде компьютерной программы.

3.5. Универсальная алгоритмическая торговая стратегия

В этом разделе мы рассмотрим финансовое приложение метода построения хорошо калибруемых предсказаний. [6] Алгоритмическая торговля (Algorithmic trading) – это формализованный процесс совершения торговых операций на финансовых рынках по заданному алгоритму с использованием специализированных компьютерных систем (торговых роботов).

Теоретико-информационный подход к построению универсальных торговых стратегий был основан Т.Ковером [21], [22], [23], который предложил алгоритм для построения "универсального" динамического портфеля акций.

Свойство универсальности этого алгоритма заключается в том, что он является асимптотически наилучшим в целом классе торговых стратегий. Алгоритм Ковера динамически перераспределяет деньги между несколькими акциями в зависимости от их текущей доходности так, что получаемый им доход оказывается асимптотически не меньшим, чем доход от любой постоянной стратегии вложения в эти акции. При этом не делается никаких стохастических предположений о поведении временных рядов цен акций. Полученные результаты имеют смысл при любом характере изменения цен акций.

Точная постановка и алгоритм Ковера будут приведены в разделе 5.9 в рамках более общей теории, а в этом разделе мы приведем другой алгоритм подобного рода для алгоритмической торговли с одной акцией. Это алгоритм также будет давать асимптотически наибольший доход по сравнению с любой "не слишком сложной" торговой стратегией.

Данный алгоритм будет использовать хорошо калибруемые

[6] Этот раздел может быть пропущен при первом чтении.

прогнозы цен акции. Метод предсказания будет основан на комбинации метода построения рандомизированных предсказаний из раздела 3.3 и метода предсказаний с использованием ядер из раздела 3.4. В основе всего этого подхода лежит понятие калибруемости предсказаний, предложенное Дэвидом и ранее рассмотренное в разделе 3.2. Мы обобщим его для более широкого класса правил выбора.

Основной результат этого раздела – теорема 3.6 – утверждает, что предлагаемая в этом разделе торговая стратегия является универсальной – она является асимптотически наилучшей в классе всех стратегий, представленных непрерывными функциями от входной информации. Подробно этот метод и соответствующие численные эксперименты представлены в [72] и [73].

Предположим, что вещественные значения $S_1, S_2, \ldots$, которые мы будем интерпретировать как цены некоторой акции, поступают в режиме онлайн. Мы также предполагаем, что они ограничены и нормированы так, что $0 \leqslant S_i \leqslant 1$ для всех i.

В процессе торговли на финансовом рынке трейдеры покупают и продают акции. Мы считаем, число C покупаемых или продаваемых единиц финансового инструмента может принимать любое вещественное значение. Допускается возможность покупать отрицательное число акций: при $C > 0$, покупка $-C$ акций эквивалентна продаже C акций, а продажа $-C$ акций эквивалентна покупке C акций. Мы также допускаем, что каждый трейдер может одалживать деньги в неограниченном количестве.

Под стратегией мы понимаем некоторый алгоритм (может быть рандомизированный), который в начале каждого раунда i игры выдает число C_i единиц финансового инструмента, которое необходимо купить (если это число положительное или равное нулю) или продать (если оно отрицательное) по цене S_{i-1}. В конце этого же раунда трейдер продает купленные единицы или покупает проданные в том же количестве, соответственно, по цене S_i. Таким образом, за i-й раунд игры капитал трейдера увеличивается (уменьшается) на величину $C_i(S_i - S_{i-1})$.

Проведем сравнение двух типов торговых стратегий на финансовом рынке в виде протокола игры с двумя типами трейдеров:

на шаге i *Трейдер* M использует рандомизированную стратегию – он покупает случайное число акций $\tilde{M}_i$, которое вычисляется некоторым рандомизированным алгоритмом; *Трейдер* D – представитель широкого класса трейдеров, он покупает $D(x_i)$ акций, где D – произвольная непрерывная функция, определенная на единичном отрезке $[0,1]$, а x_i – число из отрезка $[0,1]$, в котором закодирована входная информация и которое выдается трейдерам прежде, чем они применят свои методы. Таким образом, C_i равно $\tilde{M}_i$ для *Трейдера* M и равно $D(x_i)$ для *Трейдеров* D второго типа.

Вещественное число x_i, как и в разделе 3.4, будет также называться *сигналом* или *дополнительной информацией*. Число x_i принадлежит $[0,1]$ и в нем может быть закодирована любая числовая информация. Например, это может быть даже будущая цена акции S_i.

Каждый *Трейдер* D использует на шаге i только информацию x_i – он покупает (или продает) $D(x_i)$ единиц акции. Стратегия этого типа будет называться *стационарной*.

Для *Трейдера* M данная игра является игрой с полной информацией. Алгоритм *Трейдера* M будет использовать все значения S_{j-1} и x_j при $j \leqslant i$.

Прошлые цены акции, сигналы и сделанные предсказания также можно закодировать в сигнале x_i, поэтому *Трейдер* D может использовать эту информацию. Здесь имеется ограничение – функция D должна быть непрерывной.

Рандомизация. Стратегия *Трейдера* M является рандомизированной. Напомним метод рандомизации из раздела 3.3. Пусть K – произвольное натуральное число. Разбиваем интервал $[0,1]$ на K равных подынтервалов длины $\Delta = 1/K$ с помощью рациональных точек $v_i = i\Delta$, где $i = 0, 1, \ldots, K$. Пусть V – множество этих точек. Любое число $p \in [0,1]$ может быть представлено в виде линейной комбинации двух граничных точек подынтервала, содержащего p:

$$p = \sum_{v \in V} w_v(p) v = w_{v_{i-1}}(p) v_{i-1} + w_{v_i}(p) v_i, \tag{3.26}$$

где $p \in [v_{i-1}, v_i]$, $i = \lfloor p^1/\Delta + 1 \rfloor$, $w_{v_{i-1}}(p) = 1 - (p - v_{i-1})/\Delta$ и

$w_{v_i}(p) = 1 - (v_i - p)/\Delta$. Определим $w_v(p) = 0$ для всех остальных $v \in V$.

Пусть $\tilde{p}$ равно v_{i-1} с вероятностью $w_{v_{i-1}}(p)$ и равно v_i с вероятностью $w_{v_i}(p)$.

Обозначим через $\bar{w}(p) = (w_v(p) : v \in V)$ вектор этих вероятностей. Говорим, что мы округляем число p до v_{i-1} с вероятностью $w_{v_{i-1}}(p)$ и до v_i с вероятностью $w_{v_i}(p)$.

В дальнейшем мы будем рассматривать переменную точность округления, а именно, будет задана последовательность параметров $\Delta_1 > \Delta_2 > \cdots \to 0$. На каждом шаге конструкции мы будем округлять числа вышеуказанным способом с точностью до одного из таких Δ_i. Такой способ случайного округления будет называться *последовательной рандомизацией*.

Универсальная торговая стратегия. Определим универсальную торговую стратегию в виде последовательности случайных величин $\tilde{M}_i$, $i = 1, 2, \ldots$. Для того, чтобы построить такую стратегию, на каждом раунде i вычисляем прогноз p_i будущего значения цены согласно алгоритму, который будет приведен на рис. 3.5 в разделе 3.5.1 ниже. Рандомизирует это число, т.е., определим соответствующую случайную величину $\tilde{p}_i$.

Подчеркнем, что в протоколе, представленном на рис. 3.4, реализация случайной величины $\tilde{M}_i$ скрыта от *Рынка*, когда он анонсирует цену S_i. *Рынок* может знать метод рандомизации и вычислять вероятности событий $\tilde{M}_i = 1$ и $\tilde{M}_i = -1$.

Мы также рандомизируем предыдущее (уже известное трейдерам) значение цены S_{i-1}, т.е., определим соответствующую случайную величину $\tilde{S}_{i-1}$.

После всех этих приготовлений, определим $\tilde{M}_i$ согласно протоколу, приведенному на рис. 3.4.

Говорим, что в случае $\tilde{M}_i > 0$ *Трейдер M* переходит на длинную позицию; он переходит на короткую позицию, в противном случае.[7] То же самое относится к *Трейдеру D*.

[7] Длинная позиция на финансовом рынке – это приобретение ценных бумаг в начале раунда игры с целью их продажи в конце раунда в расчете на увеличение их стоимости. Короткая позиция – это продажа одолженных (у другого трейдера) акций в начале раунда игры с целью их покупки и возвращения в конце раунда игры. Переходя на короткую позицию, трейдер рассчитывает

Полагаем $\mathcal{K}_0^D = 0$ и $\mathcal{K}_0^M = 0$.
FOR $i = 1, 2, \ldots$
Рынок анонсирует сигнал $x_i \in [0, 1]$.
Определим точность случайного округления на шаге i: $\Delta = \Delta_s$, где $n_s < i \leqslant n_{s+1}$ (последовательности n_s и Δ_s, $s = 1, 2, \ldots$, определены ниже после неравенства (3.48) в доказательстве Теоремы 3.7 далее).
Вычисляем прогноз p_i с помощью алгоритма, приводимого ниже на рис. 3.5, с входным параметром Δ.
Получаем рандомизированное значение прогноза: $\tilde{p}_i$.
Получаем рандомизированное значение цены: $\tilde{S}_{i-1}$.
Трейдер M покупает $\tilde{M}_i$ акций по цене S_{i-1} за каждую, где

$$\tilde{M}_i = \begin{cases} 1, \text{ если } \tilde{p}_i > \tilde{S}_{i-1}, \\ -1, \text{ в противном случае.} \end{cases}$$

Трейдер D покупает $D(x_i)$ акций по цене S_{i-1} за каждую, где D – произвольная непрерывная функция на $[0, 1]$.
Рынок объявляет цену S_i акции.
Трейдер M продает $\tilde{M}_i$ акций по цене S_i и обновляет значение своего кумулятивного выигрыша: $\mathcal{K}_i^M = \mathcal{K}_{i-1}^M + \tilde{M}_i(S_i - S_{i-1})$.
Трейдер D продает $D(x_i)$ акций по цене S_i и обновляет значение своего кумулятивного выигрыша: $\mathcal{K}_i^D = \mathcal{K}_{i-1}^D + D(x_i)(S_i - S_{i-1})$.
ENDFOR

Рис. 3.4: Протокол игры

Мы допускаем, что трейдеры могут одалживать деньги и акции в неограниченном количестве. По окончании серии раундов игры мы оценим выигрыш (или долг) каждого из трейдеров.

Центральным моментом стратегии $\tilde{M}_i$ является рандомизированный алгоритм для вычисления предсказаний $\tilde{p}_i$. Этот алгоритм будет представлен на рис. 3.5 в разделе 3.5.1.

Трейдер M может покупать или продавать только одну акцию. Поэтому для сравнения эффективности стратегий трейдеров необходимо нормировать стратегию *Трейдера D*. Рассмотрим норму $\|D\|_\infty = \sup\limits_{0\leqslant x\leqslant 1} |D(x)|$, где D – произвольная непрерывная функция, определенная на единичном отрезке $[0,1]$. Далее мы будем использовать величину $\|D\|_+ = \max\{1, \|D\|_\infty\}$ в качестве нормирующего множителя.

Предполагаем, что значения $S_1, S_2, \dots \in [0,1]$ и $x_1, x_2, \dots \in [0,1]$ подаются последовательно в режиме онлайн в рамках протокола, представленного на рис 3.4.

Основной результат этого раздела представим в виде теоремы 3.6, которая утверждает, что, с вероятностью 1, средний выигрыш универсальной торговой стратегии асимптотически не меньше чем средний выигрыш любой стационарной торговой стратегии, отнесенный на одну акцию:

Теорема 3.6. *Стратегия Трейдера M является асимптотически оптимальной: для любой непрерывной функции D асимптотическое соотношение*

$$\liminf_{n\to\infty} \frac{1}{n}\left(\mathcal{K}_n^M - \|D\|_+^{-1}\mathcal{K}_n^D\right) \geqslant 0 \tag{3.27}$$

выполнено почти всюду относительно распределения вероятностей, порожденного методом рандомизации.

Заметим, что требование (3.27) для всех D эквивалентно требованию выполнения неравенства:

$$\liminf_{n\to\infty} \frac{1}{n}\left(\mathcal{K}_n^M - \mathcal{K}_n^D\right) \geqslant 0$$

на уменьшение цены в конце раунда игры.

для всех D таких, что $\|D\|_\infty \leqslant 1$.

Поскольку условие (3.27) асимптотической оптимальности *Трейдера M* выполнено и относительно тривиальной стратегии $D(x) = 0$ для всех x, эта стратегия является также и *асимптотически безрисковой.*

Следствие 3.3. *Универсальная стратегия Трейдера M является асимптотически безрисковой:*

$$\liminf_{n\to\infty} \frac{\mathcal{K}_n^M}{n} \geqslant 0$$

выполнено почти всюду.

Доказательство теоремы 3.6 приведено в разделе 3.5.2. Там мы рассмотрим соответствующую торговую стратегию, которая, в свою очередь, использует хорошо калибруемые предсказания, определяемые в разделе 3.5.1.

3.5.1. Калибруемость с дополнительной информацией

В этом разделе мы приведем и изучим алгоритм для вычисления хорошо калибруемых предсказаний, которые будут использоваться для определения торговой стратегии $\tilde{M}_i$.

Мы рассмотрим правила выбора подпоследовательностей более общего вида чем те, которые рассматривались в разделах 3.2 и 3.3. Для произвольного подмножества $S \subseteq [0,1]^2 = [0,1] \times [0,1]$ определим

$$I_S(p,x) = \begin{cases} 1, & \text{если } (p,x) \in S, \\ 0, & \text{в противном случае,} \end{cases}$$

где $p, x \in [0,1]$. В разделе 3.5.2 мы будем использовать множества типа $S = \{(p,x) : p > x\}$ или $S = \{(p,x) : p \leqslant x\}$.

Гильбертовы пространства, порожденные воспроизводящие ядром. В качестве первого этапа доказательства теоремы 3.6 мы покажем в разделе 3.5.2, что для произвольного гильбертова пространства RKHS можно определить стратегию $\tilde{M}_i$,

которая является универсальной для класса всех стационарных стратегий, задаваемых функциями из этого пространства. После этого, используя универсальное пространство RKHS, мы перенесем свойство универсальности на произвольные стратегии, задаваемые непрерывными функциями.

Напомним, что гильбертово пространство $\mathcal{F}$ функций, определенных на множестве X называется RKHS на X, если функционал $f \to f(x)$ является непрерывным для каждого $x \in X$. Обозначаем $\|\cdot\|_{\mathcal{F}}$ норму на пространстве $\mathcal{F}$. Также определяется $c_{\mathcal{F}}(x) = \sup\limits_{\|f\|_{\mathcal{F}} \leqslant 1} |f(x)|$. Константа вложения пространства $\mathcal{F}$ определяется как $c_{\mathcal{F}} = \sup\limits_{x} c_{\mathcal{F}}(x)$.

Мы будем рассматривать RKHS $\mathcal{F}$ на $X = [0,1]$ с конечной константой вложения: $c_{\mathcal{F}} < \infty$. Важный пример такого пространства RKHS – соболевское пространство $\mathcal{F} = H^1([0,1])$, которое состоит из абсолютно непрерывных функций $f : [0,1] \to \mathcal{R}$ с $\|f\|_{\mathcal{F}} < \infty$, где $\|f\|_{\mathcal{F}} = \sqrt{\int_0^1 (f(t))^2 dt + \int_0^1 (f'(t))^2 dt}$. Для такого пространства $c_{\mathcal{F}} = \sqrt{\operatorname{cth} 1} = \sqrt{\frac{1+e^{-2}}{1-e^{-2}}}$ (см. [65]).

Пусть $\mathcal{F}$ – некоторое пространство RKHS на X со скалярным произведением $(f \cdot g)$ для $f, g \in \mathcal{F}$. По теореме Рисса–Фишера для любого $x \in X$ существует $k_x \in \mathcal{F}$ такое, что $f(x) = (k_x \cdot f)$ для всех $f \in \mathcal{F}$. Тогда воспроизводящее ядро определяется как $K(x,y) = (k_x \cdot k_y)$.

В обратную сторону, любое ядро $K(x,y)$ определяет некоторое *каноническое* пространство RKHS $\mathcal{F}$, а также отображение $\Phi : X \to \mathcal{F}$, так что $K(x,y) = (\Phi(x) \cdot \Phi(y))$.

Пусть $\mathcal{F}$ – некоторое гильбертово пространство RKHS на $[0,1]$ с конечной константой вложения $c_{\mathcal{F}}$, $\|\cdot\|_{\mathcal{F}}$ – соответствующая норма и $R(x,x')$ – ядро на $[0,1]$.

Предполагаем, что последовательность $S_1, S_2, \dots \in [0,1]$ вещественных чисел и последовательность $x_1, x_2, \dots \in [0,1]$ сигналов поступают последовательно согласно протоколу, представленному на рис. 3.4.

Теорема 3.7. *Пусть $\epsilon > 0$. Можно построить алгоритм для вычисления прогнозов $p_1, p_2, \dots$ такой, что выполнены следую-*

щие условия:

- *для любого $\delta > 0$, для любого подмножества $S \subseteq [0,1]^2$ и для любого n с вероятностью не менее $1-\delta$*

$$\left|\sum_{i=1}^{n} I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$
$$\leqslant 18 n^{3/4+\epsilon}(c_{\mathcal{F}}^2 + 1)^{1/4} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}} \tag{3.28}$$

  где $\tilde{p}_1, \tilde{p}_2, \ldots$ – соответствующие рандомизации $p_1, p_2, \ldots$ и $\tilde{z}_1, \tilde{z}_2, \ldots$ – рандомизации чисел $z_1, z_2, \ldots$, где $z_i = S_{i-1}$, $i = 1, 2, \ldots$;

- *для любого $D \in \mathcal{F}$*

$$\left|\sum_{i=1}^{n} D(x_i)(S_i - p_i)\right| \leqslant \|D\|_{\mathcal{F}}\sqrt{(c_{\mathcal{F}}^2 + 1)n} \tag{3.29}$$

  для всех n;

- *для любого подмножества $S \subseteq [0,1]^2$ с вероятностью 1*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) = 0. \tag{3.30}$$

Доказательство. План доказательства следующий. Сначала по заданному $\Delta > 0$ мы модифицируем рандомизированный алгоритм из раздела 3.3 и объединим его с алгоритмом из раздела 3.4. Прогнозы модифицированного алгоритма будут калиброваться относительно расширенных правил выбора на последовательности $S_1, S_2, \ldots$ с точностью до Δ. После этого, мы применим этот алгоритм в условиях переменной точности округления: $\Delta \to 0$, так что будут выполнены условия теоремы.

Предложение 3.1. *При предположениях теоремы 3.7 можно построить алгоритм для вычисления прогнозов такой, что выполнено неравенство (3.29) для всех D из RKHS $\mathcal{F}$ и для всех n.*

Также для любого $\delta > 0$, любого $S \subseteq [0,1]^2$ и любого n с вероятностью не менее $1-\delta$ будет выполнено:

$$\left|\sum_{i=1}^{n} I_S(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$
$$\Delta n + 2\sqrt{\frac{n(c_{\mathcal{F}}^2+1)}{\Delta}} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}.$$

Доказательство. Предположим, что детерминированные прогнозы $p_1, \ldots, p_{n-1}$ уже определены (полагаем $p_1 = 1/2$). Вычислим теперь детерминированный прогноз p_n и случайно округлим его до $\tilde{p}_n$.

Разбиение $V = \{v_0, \ldots, v_K\}$ и вероятности случайного округления были определены по (3.26). Далее, мы округляем детерминированный прогноз p_n до v_{i-1} с вероятностью $w_{v_{i-1}}(p_n)$ и до v_i с вероятностью $w_{v_i}(p_n)$. Мы также случайным образом округляем $z_n = S_{n-1}$ до v_{s-1} с вероятностью $w_{v_{s-1}}(z_n)$ и до v_s с вероятностью $w_{v_s}(z_n)$, где $z_n \in [v_{s-1}, v_s]$.

Пусть $W_v(p_n, z_n) = w_{v^1}(p_n) w_{v^2}(z_n)$, где $v = (v^1, v^2)$, $v^1, v^2 \in V$, и $W(p_n, z_n) = (W_v(p_n, z_n) : v \in V^2)$ – вектор, задающий вероятностное распределение на $V^2 = V \times V$. Определим соответствующее ядро: $K(p, z, p', z') = (W(p, z) \cdot W(p', z'))$.

По определению ядро $R(x, x')$ можно представить в виде скалярного произведения в пространстве признаков: $R(x, x') = (\Phi(x) \cdot \Phi(x'))$.

Алгоритм для вычисления детерминированных предсказаний $p_1, p_2, \ldots$ представлен на рис. 3.5. Продолжим доказательство предложения 3.1. Пусть прогнозы $p_1, p_2, \ldots$ вычислены с помощью этого алгоритма (см. рис. 3.5).

По определению прогноза p_n в алгоритме, представленном на рис. 3.5, для любого n и исхода S_n выполнено $U(p_n)(S_n - p_n) \leqslant 0$.

Определим $p_1 = 1/2$. Инициализируем параметр Δ.
FOR $n = 1, 2 \ldots$
Определим $U_n(p) = \sum\limits_{i=1}^{n-1} (K(p, z_n, p_i, z_i) + R(x_n, x_i))(S_i - p_i)$.
Если $U_n(p) > 0$ для всех $p \in [0, 1]$, то определим $p_n = 1$;
Если $U_n(p) < 0$ для всех $p \in [0, 1]$, то определим $p_n = 0$.
В противном случае, пусть p_n есть какой-либо корень уравнения $U_n(p) = 0$ (такой корень существует по теореме о промежуточном значении).
ENDFOR

Рис. 3.5: Алгоритм для вычисления детерминированных предсказаний

Тогда, как легко видеть, для любого N:

$$0 \geqslant \sum_{n=1}^{N} U_n(p_n)(S_n - p_n) =$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{n-1} (K(p_n, z_n, p_i, z_i) + R(x_n, x_i)) \times$$

$$\times (S_i - p_i)(S_n - p_n) =$$

$$= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} K(p_n, z_n, p_i, z_i)(S_i - p_i)(S_n - p_n) -$$

$$-\frac{1}{2} \sum_{n=1}^{N} (K(p_n, z_n, p_n, z_n)(S_n - p_n))^2 +$$

$$+\frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} R(x_n, x_i)(S_i - p_i)(S_n - p_n) -$$

$$-\frac{1}{2} \sum_{n=1}^{N} (R(x_n, x_n)(S_n - p_n))^2 = \tag{3.31}$$

$$= \frac{1}{2} \left\| \sum_{n=1}^{N} W(p_n, z_n)(S_n - p_n) \right\|^2 -$$

$$-\frac{1}{2} \sum_{n=1}^{N} \|W(p_n, z_n)\|^2 (S_n - p_n)^2 + \tag{3.32}$$

$$+\frac{1}{2} \left\| \sum_{n=1}^{N} \Phi(x_n)(S_n - p_n) \right\|_{\mathcal{F}}^2 -$$

$$-\frac{1}{2} \sum_{n=1}^{N} \|\Phi(x_n)\|_{\mathcal{F}}^2 (S_n - p_n)^2. \tag{3.33}$$

Здесь в строке (3.32) $\|\cdot\|$ – эвклидова норма, а в строке (3.33) $\|\cdot\|_{\mathcal{F}}$ – норма на $\mathcal{F}$.

Так как $(S_n - p_n)^2 \leqslant 1$ для всех n и

$$\|W(p_n, z_n)\|^2 = \sum_{v \in V^2} (W_v(p_n, z_n))^2 \leqslant$$
$$\leqslant \sum_{v \in V^2} W_v(p_n, z_n) = 1,$$

вычитаемая сумма в строке (3.32) ограничена сверху числом N.

Так как $\|\Phi(x_n)\|_{\mathcal{F}} = c_{\mathcal{F}}(x_n)$ и $c_{\mathcal{F}}(x) \leqslant c_{\mathcal{F}}$ для всех x, вычитаемая сумма в строке (3.33) ограничена сверху величиной $c_{\mathcal{F}}^2 N$. В результате получаем:

$$\left\| \sum_{n=1}^{N} W(p_n, z_n)(S_n - p_n) \right\| \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)N}, \tag{3.34}$$

$$\left\| \sum_{n=1}^{N} \Phi(x_n)(S_n - p_n) \right\|_{\mathcal{F}} \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)N} \tag{3.35}$$

для всех N. Обозначим: $\bar{\mu}_n = \sum\limits_{i=1}^{n} W(p_i, z_i)(S_i - p_i)$. По (3.34), $\|\bar{\mu}_n\| \leqslant \sqrt{(c_{\mathcal{F}}^2 + 1)n}$ для всех n.

Пусть $\bar{\mu}_n = (\mu_n(v) : v \in V^2\}$. По определению для любого v:

$$\mu_n(v) = \sum_{i=1}^{n} W_v(p_i, z_i)(S_i - p_i). \tag{3.36}$$

Вставляем величину $I(v)$ в сумму (3.36), где I – характеристическая функция произвольного подмножества $S \subseteq [0,1]^2$, суммируем по $v \in V^2$ и изменяем порядок суммирования. Применяем неравенство Коши–Буняковского для векторов $\bar{I} = (I(v) : v \in V^2)$,

$\bar{\mu}_n = (\mu_n(v) : v \in V^2)$ в евклидовой норме и получаем:

$$\left| \sum_{i=1}^{n} \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - p_i) \right| =$$

$$= \left| \sum_{v \in V^2} I(v) \sum_{i=1}^{n} W_v(p_i, z_i)(S_i - p_i) \right| =$$

$$= |(\bar{I} \cdot \bar{\mu}_n)| \leqslant \|\bar{I}\| \cdot \|\bar{\mu}_n\| \leqslant \sqrt{|V^2|(c_{\mathcal{F}}^2 + 1)n} \tag{3.37}$$

для всех n, где $|V^2| = (1/\Delta + 1)^2 \leqslant 4/\Delta^2$ – число элементов разбиения.

Пусть $\tilde{p}_i$ – случайная величина, принимающая значения $v \in V$ с вероятностями $w_v(p_i)$. Напомним, что $\tilde{z}_i$ – случайная величина, принимающая значения $v \in V$ с вероятностями $w_v(z_i)$.

Пусть $S \subseteq [0,1]^2$ и I – характеристическая функция этого множества. Для любого i математическое ожидание случайной величины $I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)$ равно

$$E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) =$$

$$= \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1), \tag{3.38}$$

где $v = (v^1, v^2)$.

Неравенство Хефдинга–Адзумы утверждает, что для любого $\gamma > 0$

$$Pr\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} V_i \right| > \gamma \right\} \leqslant 2e^{-2n\gamma^2}$$

для всех n, где V_i – ограниченные мартингал-разности (см. лемму 10.2 и следствие 10.4 из раздела 10 далее).

Далее мы полагаем $V_i = I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) - E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))$ и $\gamma = \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$, где $\delta > 0$. Отсюда получаем, что для любых $\delta > 0$,

S и n, с вероятностью $1-\delta$:

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i) - \sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant$$
$$\leqslant \sqrt{\frac{n}{2} \ln \frac{2}{\delta}}. \qquad (3.39)$$

По определению детерминированного прогноза, суммы:

$$\sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - p_i) \text{ и } \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1)$$

различаются не более чем на Δ для всех i, где $v = (v^1, v^2)$. Суммируя (3.38) по $i = 1, \ldots, n$ и используя неравенство (3.37), получаем:

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| =$$
$$= \left|\sum_{i=1}^{n} \sum_{v \in V^2} W_v(p_i, z_i) I(v)(S_i - v^1)\right| \leqslant$$
$$\leqslant \Delta n + 2\sqrt{(c_{\mathcal{F}}^2 + 1) n / \Delta^2} \qquad (3.40)$$

для всех n.

По (3.39) и (3.40) для любых S и n имеем с вероятностью $1-\delta$

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)\right| \leqslant$$
$$\leqslant \Delta n + 2\sqrt{(c_{\mathcal{F}}^2 + 1) n / \Delta^2} + \sqrt{\frac{n}{2} \ln \frac{2}{\delta}}. \qquad (3.41)$$

По неравенству Коши–Буняковского:

$$\left|\sum_{n=1}^{N} D(x_n)(S_n - p_n)\right| =$$
$$= \left|\sum_{n=1}^{N} (S_n - p_n)(D \cdot \Phi(x_n))\right| =$$
$$= \left|\left(\sum_{n=1}^{N} (S_n - p_n)\Phi(x_n) \cdot D\right)\right| \leqslant$$
$$\leqslant \left\|\sum_{n=1}^{N} (S_n - p_n)\Phi(x_n)\right\|_{\mathcal{F}} \cdot \|D\|_{\mathcal{F}} \leqslant$$
$$\leqslant \|D\|_{\mathcal{F}} \sqrt{(c_{\mathcal{F}}^2 + 1)N}.$$

Предложение доказано. $\triangle$

Переходим теперь к заключительному этапу доказательства теоремы 3.7.

Выражение $\Delta n + 2\sqrt{(c_{\mathcal{F}}^2 + 1)n/\Delta^2}$ из оценок (3.40) и (3.41) принимает свое минимальное значение при $\Delta = \sqrt{2}(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n^{-\frac{1}{4}}$. В этом случае правая часть неравенства (3.40) равна

$$\Delta n + 2\sqrt{n(c_{\mathcal{F}}^2 + 1)/\Delta^2} = 2\Delta n = 2\sqrt{2}(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n^{\frac{3}{4}}. \qquad (3.42)$$

Далее будем использовать верхнюю оценку $2\Delta n$ для (3.40).

Для получения оценки (3.28) выберем монотонную последовательность чисел $\Delta_1 > \Delta_2 > \ldots$, так что $\Delta_s \to 0$ при $s \to \infty$. Также будем использовать возрастающую последовательность натуральных чисел $n_1 < n_2 < \ldots$ Для произвольного s на шагах $n_s \leqslant n < n_{s+1}$ мы будем использовать для рандомизации разбиение интервала $[0, 1]$ на подынтервалы длины Δ_s. Начинаем наши последовательности с $n_1 = 1$ и $\Delta_1 = 1$.

Определим числа $n_2, n_3, \ldots$ так, что неравенство

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant 4(s+1)\Delta_s n \qquad (3.43)$$

выполнено для всех $n_s \leqslant n \leqslant n_{s+1}$ и для всех $s \geqslant 1$.

Мы определим эту последовательность с помощью метода математической индукции по s. Допустим, что n_s $(s \geqslant 1)$ уже определено так, что неравенство

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant 4s\Delta_{s-1}n \tag{3.44}$$

выполнено для всех $n_{s-1} \leqslant n \leqslant n_s$. Допустим, что также выполнено дополнительное неравенство

$$\left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant 4s\Delta_s n_s. \tag{3.45}$$

Определим n_{s+1}. Рассмотрим предсказания p_i ($\tilde{p}_i$ – их рандомизированные значения), вычисленные с помощью определенного выше алгоритма при дискретизации $\Delta = \Delta_{s+1}$. Первые n_s из этих предсказаний не будем использовать (точнее, они появляются только в оценках (3.46) и (3.47); обозначим их $\mathbf{p_1}, \ldots, \mathbf{p_{n_s}}$). Добавим все предсказания p_i для $i > n_s$ к предсказаниям, полученным на предыдущих шагах индукции. Пусть n_{s+1} такое, что выполнены неравенства

$$\begin{aligned}\left|\sum_{i=1}^{n_{s+1}} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| &\leqslant \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| + \\ + \left|\sum_{i=n_s+1}^{n_{s+1}} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{z}_i)(S_i - \hat{\mathbf{p}}_\mathbf{i}))\right| &+ \\ + \left|\sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{z}_i)(S_i - \hat{\mathbf{p}}_\mathbf{i}))\right| &\leqslant 4(s+1)\Delta_{s+1}n_{s+1}.\end{aligned} \tag{3.46}$$

Здесь первая сумма из правой части неравенства (3.46) ограничена $4s\Delta_s n_s$ по предположению индукции (3.45). Вторая и третья суммы ограничены величинами $2\Delta_{s+1}n_{s+1}$ и $2\Delta_{s+1}n_s$, соответственно, где $\Delta = \Delta_{s+1}$ определена так, что выполнено (3.42). Это следует из (3.40) и выбора n_s.

Тогда предположение индукции (3.45) выполнено при

$$n_{s+1} \geqslant \frac{2s\Delta_s + \Delta_{s+1}}{\Delta_{s+1}(2s+1)} n_s.$$

Подобным образом получаем

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| +$$
$$+ \left|\sum_{i=n_s+1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{z}_i)(S_i - \hat{\mathbf{p}}_\mathbf{i}))\right| +$$
$$+ \left|\sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_\mathbf{i}, \tilde{z}_i)(S_i - \hat{\mathbf{p}}_\mathbf{i}))\right| \leqslant 4(s+1)\Delta_s n \qquad (3.47)$$

при $n_s < n \leqslant n_{s+1}$. Здесь первая сумма из правой части неравенства также ограничена: $4s\Delta_s n_s \leqslant 4s\Delta_s n$ по предположению индукции (3.45). Вторая и третья суммы ограничены $2\Delta_{s+1}n \leqslant 2\Delta_s n$ и $2\Delta_{s+1}n_s \leqslant 2\Delta_s n$, соответственно. Это следует из (3.40) и выбора Δ_s. Индуктивное предположение (3.44) выполнено.

По (3.43) для произвольного s выполнено

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{z}_i)(S_i - \tilde{p}_i))\right| \leqslant 4(s+1)\Delta_s n \qquad (3.48)$$

для всех $n \geqslant n_s$, если Δ_s удовлетворяет условию $\Delta_{s+1} \leqslant \Delta_s(1 - \frac{1}{s+2})$ для всех s.

Покажем, что последовательности n_s и Δ_s удовлетворяющие этим условиям, существуют.

Пусть $\epsilon > 0$ и $M = \lceil 2/\epsilon \rceil$, где $\lceil r \rceil$, есть наименьшее целое число большее или равное r. Определим $n_s = (s+M)^M$ и $\Delta_s = \sqrt{2}(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n_s^{-\frac{1}{4}}$. Нетрудно проверить, что в этом случае все условия, наложенные на n_s и Δ_s, выполнены для всех достаточно больших s; пусть они будут выполнены при всех $s \geqslant s_0$. Переопределим $n_i = n_{s_0}$ для всех $1 \leqslant i \leqslant s_0$. Заметим, что неравенства (3.44) и (3.45) выполнены для таких n_i тривиальным образом.

В оценке (3.44) для всех $n_s \leqslant n < n_{s+1}$ выполнено

$$\begin{aligned} 4(s+1)\Delta_s n \leqslant 4(s+M)\Delta_s n_{s+1} = \\ = 4\sqrt{2}(c_{\mathcal{F}}^2+1)^{\frac{1}{4}}(s+M)(s+M+1)^M(s+M)^{-\frac{M}{4}} \leqslant \\ \leqslant 18(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n_s^{\frac{3}{4}+2/M} \leqslant \\ \leqslant 18(c_{\mathcal{F}}^2+1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}. \end{aligned}$$

Таким образом, получаем

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i,\tilde{z}_i)(S_i-\tilde{p}_i))\right| \leqslant 18(c_{\mathcal{F}}^2+1)^{1/4} n^{3/4+\epsilon} \tag{3.49}$$

для всех n.

Положим $V_i = I(\tilde{p}_i,\tilde{z}_i)(S_i-\tilde{p}_i) - E(I(\tilde{p}_i,\tilde{z}_i)(S_i-\tilde{p}_i))$ и $\gamma = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$, где $\delta > 0$. Здесь E – символ условного математического ожидания относительно $\tilde{p}_1,\ldots,p_{i-1}$. Случайные величины V_i образуют мартингал-разность относительно последовательности $\tilde{p}_1,\tilde{z}_1,\tilde{p}_2,\tilde{z}_2\ldots$

Согласно неравенству Хефдинга–Адзумы для любого $\gamma > 0$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right| > \gamma\right\} \leqslant 2e^{-2n\gamma^2} \tag{3.50}$$

для всех n (см. лемму 10.2 из раздела 10).

Комбинируя (3.49) с (3.50), получим, что для любых $\delta > 0$, S и n с вероятностью $1-\delta$ выполнено

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i,\tilde{z}_i)(S_i-\tilde{p}_i)\right| \leqslant 18(c_{\mathcal{F}}^2+1)^{1/4} n^{3/4+\epsilon} + \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}.$$

Соотношение (3.30) может быть получено из неравенства (3.28) с помощью леммы Бореля–Кантелли. Доказательство аналогично рассуждению в конце доказательства теоремы 3.6 (см. раздел 3.5.2). Теорема 3.7 доказана. $\triangle$

3.5.2. Доказательство теоремы 3.6

На каждом шаге i вычисляем детерминированный прогноз p_i согласно алгоритму, приведенному в разделе 3.5.1, и после этого переходим к его рандомизации $\tilde{p}_i$ используя параметры: $M = \lceil 2/\epsilon \rceil$, $\Delta = \Delta_s = \sqrt{2}(c_{\mathcal{F}} + 1)^{\frac{1}{4}}(s + M)^{-\frac{M}{4}}$ и $n_s = (s + M)^M$, где $n_s \leqslant i < n_{s+1}$. Пусть также $\tilde{S}_{i-1}$ – рандомизированная прошлая цена S_{i-1} финансового инструмента. В теореме 3.7, $z_i = S_{i-1}$ и $\tilde{z}_i = \tilde{S}_{i-1}$.

При данном методе рандомизации имеют место следующие верхние оценки;

$$\left| \sum_{i=1}^{n} I(\tilde{p}_i > \tilde{S}_{i-1})(\tilde{S}_{i-1} - S_{i-1}) \right| \leqslant \sum_{t=0}^{s} (n_{t+1} - n_t)\Delta_t \leqslant$$
$$\leqslant 4(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n_s^{\frac{3}{4}+\epsilon} \leqslant 4(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}, \qquad (3.51)$$

где $n_s \leqslant n < n_{s+1}$.

Пусть $D(x)$ – произвольная функция из RKHS $\mathcal{F}$. Очевидным образом, оценка (3.51) верна, если мы заменим $I(\tilde{p}_i > \tilde{S}_{i-1})$ на $\|D\|_+^{-1} D(\mathbf{x}_i)$. Для простоты, мы приведем оценки для случая, когда $D(x) \geqslant 0$ для всех x и

$$\tilde{M}_i = \begin{cases} 1, \text{ если } \tilde{p}_i > \tilde{S}_{i-1}, \\ 0, \text{ в противном случае.} \end{cases}$$

Используем обозначения:

$$\nu_1(n) = 4(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} n^{\frac{3}{4}+\epsilon}, \qquad (3.52)$$

$$\nu_2(n) = 18 n^{\frac{3}{4}+\epsilon}(c_{\mathcal{F}}^2 + 1)^{\frac{1}{4}} + \sqrt{\frac{n}{2} \ln \frac{2}{\delta}}. \qquad (3.53)$$

$$\nu_3(n) = \sqrt{(c_{\mathcal{F}}^2 + 1)n} \qquad (3.54)$$

Далее все суммы предполагаются для $i = 1, \ldots, n$. Мы также используем неравенство Хефдинга–Адзумы.

Для произвольного $\delta > 0$ с вероятностью $1 - \delta$ имеет место

следующая цепь равенств и неравенств:

$$\mathcal{K}_n^M = \sum_{i=1}^n \tilde{M}_i(S_i - S_{i-1}) = \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (S_i - S_{i-1}) =$$

$$= \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (S_i - \tilde{p}_i) + \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{p}_i - \tilde{S}_{i-1}) + \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{S}_{i-1} - S_{i-1}) \geqslant \tag{3.55}$$

$$\geqslant \sum_{\tilde{p}_i > \tilde{S}_{i-1}} (\tilde{p}_i - \tilde{S}_{i-1}) - \nu_1(n) - \nu_2(n) \geqslant \tag{3.56}$$

$$\geqslant \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(\tilde{p}_i - \tilde{S}_{i-1}) - \nu_1(n) - \nu_2(n) =$$

$$= \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(p_i - S_{i-1}) + \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(\tilde{p}_i - p_i) -$$

$$-\|D\|_+^{-1} \sum_{i=1}^n D(x_i)(\tilde{S}_{i-1} - S_{i-1}) - \nu_1(n) - \nu_2(n) \geqslant \tag{3.57}$$

$$\geqslant \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(p_i - S_{i-1}) - 3\nu_1(n) - \nu_2(n) = \tag{3.58}$$

$$= \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(S_i - S_{i-1}) -$$

$$-\|D\|_+^{-1} \sum_{i=1}^n D(x_i)(S_i - p_i) - 3\nu_1(n) - \nu_2(n) \geqslant \tag{3.59}$$

$$\geqslant \|D\|_+^{-1} \sum_{i=1}^n D(x_i)(S_i - S_{i-1}) -$$

$$-3\nu_1(n) - \nu_2(n) - \|D\|_+^{-1}\|D\|_{\mathcal{F}}\nu_3(n) =$$

$$= \|D\|_+^{-1}\mathcal{K}_n^D - 3\nu_1(n) - \nu_2(n) - \|D\|_+^{-1}\|D\|_{\mathcal{F}}\nu_3(n).$$

При переходе от (3.55) к (3.56) были использованы неравенство (3.28) из теоремы 3.7 и оценка (3.51); по этой причине были вычтены члены (3.52) и (3.53). При переходе от (3.57) к (3.58) дважды была применена оценка (3.51); по этой причине член (3.51) был вычтен дважды. При переходе от (3.58) к (3.59) было использова-

но неравенство (3.29) из теоремы 3.7; таким образом, был вычтен член (3.54).

Таким образом, существует такая константа $c > 0$, что для любого $\delta > 0$ и для любого n с вероятностью $1 - \delta$ выполнено

$$\mathcal{K}_n^M \geqslant \|D\|_+^{-1}\mathcal{K}_n^D - cn^{\frac{3}{4}+\epsilon} - \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}. \tag{3.60}$$

В частности, при $D(x) = 0$ при всех x, получаем условие на риск возножных потерь нашей универсальной стратегии: для любого $\delta > 0$ и для любого n, с вероятностью $1 - \delta$, выполнено

$$\mathcal{K}_n^M \geqslant -cn^{\frac{3}{4}+\epsilon} - \sqrt{\frac{n}{2}\ln\frac{2}{\delta}}.$$

Неравенство (3.27) будет следовать из (3.60). Для доказательства используем лемму Бореля-Кантелли. Эта лемма утверждает, что если для некоторой последовательности событий A_n ряд $\sum\limits_{n=1}^{\infty} P(A_n)$ сходится, то вероятность того, что событие A_n происходит для бесконечно многих n, равна 0.

Для того, чтобы применить эту лемму, исползуем неравенство 10.5 из раздела 10.

Обозначим $\gamma = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$. Тогда $\delta = 2e^{-2n\gamma^2}$. Перепишем (3.60) в форме:

$$\frac{1}{n}\mathcal{K}_n^M - \|D\|_+^{-1}\frac{1}{n}\mathcal{K}_n^D \geqslant -cn^{-\frac{1}{4}+\epsilon} - \gamma. \tag{3.61}$$

Согласно (3.60), для любых n и $\gamma > 0$ неравенство (3.61) не выполнено с вероятностью $2e^{-2n\gamma^2}$. Так как ряд $\sum\limits_{n=1}^{\infty} e^{-2n\gamma^2}$ сходится при $\gamma > 0$, по лемме Бореля-Кантелли, при фиксированном $\gamma > 0$, неравенство (3.61) может нарушаться не более чем для конечного числа различных n. Отсюда следует, что событие:

$$\liminf_{n\to\infty} \frac{1}{n}\left(\mathcal{K}_n^M - \|D\|_+^{-1}\mathcal{K}_n^D\right) \geqslant 0$$

выполнено почти всюду.

Теорема 3.6 доказана для произвольной неотрицательной функции $D \in \mathcal{F}$. Доказательство утверждения для произвольной неположительной функции $D \in \mathcal{F}$, а также для функций общего вида аналогично.

Мы расширим полученные асимптотические оценки для любой непрерывной функции $D(x)$ с помощью универсального ядра и соответствующего канонического RKHS.

Гильбертово пространство RKHS $\mathcal{F}$ на X называется *универсальным*, если X – компактное метрическое пространство и любая непрерывная функция f на X может быть аппроксимирована в метрике $\|\cdot\|_\infty$ функцией из $\mathcal{F}$ с произвольной точностью: для любого $\epsilon > 0$ существует $D \in \mathcal{F}$ такая, что

$$\sup_{x \in X} |f(x) - D(x)| \leqslant \epsilon$$

(см. Стейнварт [54], Определение 4).

Мы будем использовать $X = [0, 1]$. Соболевское пространство $\mathcal{F} = H^1([0, 1])$ является универсальным RKHS (см. работы: [54], [65]).

Существование универсального RKHS на $[0, 1]$ позволяет получить полную версию теоремы 3.6.

Можно построить алгоритм для вычисления детерминированных предсказаний p_i и метод последовательной рандомизации такие, что соответствующая рандомизированная торговая стратегия $\tilde{M}_i$ выигрывает на одну единицу финансового инструмента не меньше, чем любая нетривиальная непрерывная стационарная стратегия f. Точнее, неравенство

$$\liminf_{n \to \infty} \frac{1}{n} \left(\mathcal{K}_n^M - \|f\|_+^{-1} \mathcal{K}_n^f \right) \geqslant 0 \tag{3.62}$$

выполнено почти всюду относительно распределения вероятностей, порожденного последовательной рандомизацией.

Этот результат непосредственно следует из неравенства (3.60) и возможности как угодно точно аппроксимировать любую непрерывную функцию f на $[0, 1]$ с помощью функции D из универсального (соболевского) RKHS $\mathcal{F}$.

Торговая стратегия $\tilde{M}_i$, удовлетворяющая (3.62), называется *универсально состоятельной.*

Заметим, что свойство универсальной состоятельности (3.62) является асимптотическим и ничего не утверждает о скорости сходимости в случае произвольной непрерывной функции f. Оценка скорости сходимости (3.60) была получена для функций из более специальных пространств RKHS.

3.6. Задачи и упражнения

1. Доказать, что при использовании смешивания по методу Кричевского и Трофимова условная вероятность появления единицы после n бинарных наблюдений $\omega^n = \omega_1, \ \ldots, \ \omega_n$ равна

$$P(\omega_{n+1} = 1|\omega^n) = \frac{n_1 + 1/2}{n+1},$$

где n_1 – число единиц в ω^n.

2. Доказать, что также имеет место оценка:

$$\int\limits_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}}dp \geqslant \frac{1}{2\sqrt{n}}\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}.$$

3. Получить явную формулу для оптимальной стратегии игры Келли не зависящей от истории из раздела 3.1.4. Получить явную оценку выигрыша для этого класса стратегий.

4. Написать выражение для универсального кодового распределения (3.1) относительно класса кодов не зависящих от истории *(Ответ:*

$$p^*(y^n) = \frac{\left(\frac{k(y^n)}{n}\right)^{k(y^n)}\left(\frac{1-k(y^n)}{n}\right)^{n-k(y^n)}}{\sqrt{\frac{2\pi n}{2}}(1+\epsilon_n)}.$$

где $k(y^n) = |\{i : y_i = 1\}|$ – число единиц в последовательности y^n и $\epsilon_n \to 0$ при $n \to \infty$).

5. Оценить минимаксный регрет для класса стратегий не зависящих от истории в случае произвольного конечного множества исходов.

6. Для некоторых последовательностей легко построить калибруемые предсказания. Последовательность $\omega_1, \omega_2, \ldots$, состоящая из 0 и 1, называется стационарной, если предел

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \omega_i$$

существует.

Доказать, что последовательность прогнозов $p_1, p_2, \ldots$, определенная соотношениями $p_1 = 0$ и

$$p_i = \frac{1}{i-1} \sum_{j=1}^{i-1} \omega_j$$

при $i > 1$, калибруется на стационарной последовательности $\omega_1, \omega_2, \ldots$.

7. Объяснить, почему в доказательстве теоремы 3.2 нельзя использовать просто округление с недостатком (или с избытком) прогноза p_n до соответствующей границы интервала разбиения. Привести контрпример.

8. Описать возможные методы вероятностного округления детерминированного прогноза p_n в доказательстве теоремы 3.2.

9. Доказать теорему 3.6 для произвольной неположительной функции $D \in \mathcal{F}$, а также для функций общего вида.

10. Привести контрпример к теореме 3.6 для случая разрывной функции $D(x)$.

3.7. Лабораторные работы

Алгоритм, описанный в разделе 3.3, может быть легко реализован в виде компьютерной программы. При этом для вычисления корня уравнения (3.8) лучше всего использовать гладкое приближение к ядру (3.7) – гауссово ядро $K(p, p') = e^{-\gamma(p-p')^2}$ для некоторого [8] Можно также использовать ядро вида $K(p, p') = \cos(\frac{\pi}{2}(p-p'))$ при $p, p' \in [0, 1]$.

[8] В данном случае сигналы отсутствуют. Для использования сигналов можно использовать ядро (3.18.) $\gamma > 0$.

Различные временные ряды можно загружать с сайта FINAM (http://finam.ru).

Например, можно загрузить поминутные данные цен акций какой-нибудь компании: $S_0, S_1, S_2, \ \ldots, \ S_n$ и откалибровать их так, чтобы $S_i \in [0,1]$ для всех i.

Лабораторная работа 1

Реализовать алгоритм раздела 3.3. Написать программу для вычисления хорошо калибруемых прогнозов $p_1, p_2, \ \ldots, \ p_n$ для двоичной последовательности $\omega_1, \omega_2, \ldots, \omega_n$, где $\omega_i \in \{0,1\}$. Сравнить эти прогнозы с прогнозами по правилу Лапласа.

Двоичную последовательность можно образовать из последовательности приращений $\Delta S_0, \Delta S_1, \ \ldots, \ \Delta S_{n-1}$, где $\Delta S_i = S_i - S_{i-1}$. Это можно сделать следующим образом:

$$\omega_i = \begin{cases} 1, \text{ если } \Delta S_i \geqslant \delta, \\ 0 \text{ в противном случае,} \end{cases}$$

где δ – некоторое положительное число.

Произвести отбор подпоследовательностей, на которых прогноз $p_i > \delta$ для различных положительных значений δ.

Создать графическое представление результатов.

Лабораторная работа 2

Загрузить временной ряд цен какой-либо акции. Нормировать цены акции $S_0, S_1, \ \ldots, S_{n-1}$ так, чтобы $S_i \in [0,1]$.

Написать программу для вычисления хорошо калибруемых прогнозов $p_1, p_2, \ \ldots, \ p_n$ для откалиброванной последовательности чисел $S_0, S_1, \ \ldots, S_{n-1}$.

Произвести отбор подпоследовательностей, на которых прогноз удовлетворяет $p_i > S_{i-1} + \epsilon$ для различных значений $\epsilon > 0$.

Создать графическое представление результатов.

Предложить и реализовать программы-роботы для игры на курсах акций, использующие эти калибруемые прогнозы.

Глава 4

Предсказания с использованием экспертных стратегий

Задача принятия правильного рационального решения является центральной в науке и практике. Решение принимается на основе некоторых наблюдаемых данных. Как и в предыдущей главе, мы будем рассматривать задачу прогнозирования параметров какого-либо процесса. Только теперь мы будем оценивать правильность наших прогнозов, руководствуясь иными принципами. Мы также не будем использовать никаких предположений о природе механизма генерации прогнозируемой последовательности.

Правильный прогноз или правильное решение ведут к меньшим потерям, чем неправильные. При традиционном статистическом подходе мы оцениваем потери при наших прогнозах в сравнении с некоторой идеальной моделью принятия правильных решений, которая обычно основана на некоторой статистической модели, описывающей наблюдаемые данные. При традиционном подходе сначала оцениваются параметры статистической модели на основе наблюдений, а потом производится прогноз на основе этой модели при оцененных параметрах.

При сравнительном подходе вместо одной идеальной модели рассматривается набор возможных моделей, которые называют-

ся конкурирующими экспертными стратегиями, или просто, экспертами. Множество таких экспертных стратегий может быть конечным или бесконечным и даже несчетным. Используя исходы, поступающие в режиме онлайн, экспертные стратегии производят прогнозы будущего исхода. Прогнозирующий алгоритм может наблюдать прогнозы конкурирующих экспертных стратегий и оценивать их эффективность в прошлом. После этого алгоритм делает свой прогноз.

Результаты прогнозов нашего алгоритма сравниваются с результатами прогнозов экспертных алгоритмов. Обычно производится сравнение потерь нашего алгоритма за некоторый период прогнозирования с потерями наилучшего на ретроспективе экспертного алгоритма.

Сравнение может производиться как в наихудшем случае, а так же в среднем, если наш алгоритм использует рандомизацию. Заметим, что распределение вероятностей, которое использует рандомизированный алгоритм, является внутренним вспомогательным распределением алгоритма; оно не имеет никакого отношения к источнику, генерирующему исходы. Мы сами генерируем случайные числа для нашего алгоритма.

Обсудим также типы процессов, генерирующих данные, для которых будут рассматриваться наши методы прогнозирования. Поведение некоторых процессов не зависит от прогнозов предсказателя. Такие процессы часто рассматриваются в классической механике, физике. Например, погода не зависит от предсказателя погоды. Приводимые ниже методы работают так же и в случае, когда характеристики процесса зависят от предсказаний. Это так называемый случай “адаптивно враждебной природы”. Например, данное предположение является естественным при прогнозировании финансового рынка. Рассматриваемые алгоритмы будут эффективно работать во всех этих случаях.

4.1. Алгоритм взвешенного большинства

В этом разделе мы рассмотрим простейшие алгоритмы, выдающие точное предсказание будущего исхода. Имеется два возмож-

ных исхода 0 и 1. Имеются N экспертов (стратегий), которые на каждом шаге t выдают предсказания $p_t^i \in \{0,1\}$, $i = 1, \ldots, N$.

Изучающий алгоритм обозревает в режиме онлайн бинарную последовательность $\omega_1, \ldots, \omega_{t-1}$ и прогнозы экспертов $p_1^i, \ldots, p_t^i$, $i = 1, \ldots, N$, и предсказывает будущий исход $p_t \in \{0,1\}$.

Предварительно рассмотрим случай, когда один из экспертов i точно предсказывает будущий исход: $p_t^i = \omega_t$ для всех t.

Рассмотрим так называемый "Алгоритм большинства". Алгоритм определяет на каждом шаге $t = 1, 2, \ldots$ множество всех экспертов, которые ни разу не сделали ошибку на предыдущих шагах:

$$B_t = \{i : p_j^i = \omega_j \text{ при всех } 1 \leqslant j \leqslant t-1\}.$$

Алгоритм большинства выдает прогноз $p_t = 1$, если большинство ранее ни разу не ошибавшихся экспертов выдают 1 в качестве такого прогноза, в противнов случае $p_t = 0$. Точнее,

$$p_t = \begin{cases} 1, \text{ если } |\{i : i \in B_t, p_t^i = 1\}| \geqslant |B_t|/2, \\ 0, \text{ в противном случае.} \end{cases}$$

Теорема 4.1. *Допустим, что существует эксперт i такой, что $p_t^i = \omega_t$ для всех t. Тогда "Алгоритм большинства" делает не более чем $\lceil \log_2 N \rceil$ ошибок, где N – число экспертов.* [1]

Доказательство. Если "Алгоритм большинства" делает ошибку на шаге t, то число ранее никогда не ошибавшихся экспертов уменьшается по крайней мере вдвое: $|B_{t+1}| \leqslant \lceil |B_t|/2 \rceil$. По предположению $|B_t| \geqslant 1$ для всех t. Отсюда число уменьшений величины $|B_t|$ в два раза не превосходит $\lceil \log_2 N \rceil$. $\triangle$

Рассмотрим теперь случай, когда эксперта, точно угадывающего будущие исходы, не существует. В этом случае рассмотрим рассмотрим "Алгоритм взвешенного большинства", который был предложен Литтлстоуном и Вармутом [42].

Приведем протокол игры с предсказаниями экспертов (см. рис. 4.1). Участники игры: *Эксперт* i, $i = 1, \ldots, N$, *Предсказатель*, *Природа*. Каждому участнику игры в момент его действия

[1] Здесь $\lceil r \rceil$ обозначает наименьшее целое большее или равное r.

FOR $t = 1, 2, \ \ldots, \ T$
Эксперт i выдает прогноз $p_t^i \in \{0, 1\}$, $i = 1, \ \ldots, \ N$
Предсказатель выдает прогноз p_t алгоритма $WMA(\epsilon)$:
IF $\sum\limits_{i:p_t^i=0} w_t^i > \sum\limits_{i:p_t^i=1} w_t^i$
THEN $p_t = 0$
ELSE $p_t = 1$
ENDIF
Природа выдает исход $\omega_t \in \{0, 1\}$
Предсказатель производит пересчет весов экспертов:
Пусть $E_t = \{i : p_t^i \neq \omega_t\}$ – множество всех экспертов i, которые выдали ошибочный прогноз на шаге t.
Уменьшаем веса таких экспертов:

$$w_{t+1}^i = \begin{cases} (1-\epsilon)w_t^i, \text{ если } i \in E_t, \\ w_t^i, \text{ в противном случае.} \end{cases}$$

ENDFOR

Рис. 4.1: Алгоритм $WMA(\epsilon)$

доступна информация о всех действиях других игроков в моменты, предшествующие данному. Говорим, что это игра с полной информацией.

Пусть ϵ – параметр, $0 < \epsilon < 1$. Полагаем начальные веса экспертов $w_1^i = 1$ при всех $1 \leqslant i \leqslant N$.

Пусть $L_T^i = \sum\limits_{t=1}^{T} |p_t^i - \omega_t|$ – число всех ошибок *Эксперта* i, $L_T = \sum\limits_{t=1}^{T} |p_t - \omega_t|$ – число всех ошибок *Предсказателя*, т.е. алгоритма $WMA(\epsilon)$ на T шагах.

Теорема 4.2. *Для любого* i *выполнено*

$$L_T \leqslant \left(\frac{2}{1-\epsilon}\right) L_T^i + \left(\frac{2}{\epsilon}\right) \ln N$$

для всех T.

Доказательство. Пусть $W_t = \sum\limits_{i=1}^{N} w_t^i$. Пусть $m = \min\limits_{1 \leqslant i \leqslant N} L_T^i$ – число ошибок наилучшего эксперта на T шагах. Пусть этот минимум достигается для эксперта i. Тогда вес эксперта i корректировался не более m раз. Тогда

$$W_t > w_t^i \geqslant (1-\epsilon)^m \tag{4.1}$$

для всех t таких, что $1 \leqslant t \leqslant T$.

С другой стороны, если наш алгоритм делает ошибку на шаге t, то

$$\sum_{i \in E_t} w_t^i \geqslant W_t/2.$$

Следовательно,

$$\begin{aligned} W_{t+1} = \sum_{i \in E_t} (1-\epsilon) w_t^i + \sum_{i \notin E_t} w_t^i = \\ = \sum_{i=1}^{N} w_t^i - \epsilon \sum_{i \in E_t} w_t^i \leqslant \\ \leqslant W_t \left(1 - \frac{\epsilon}{2}\right). \end{aligned}$$

По определению $W_{t+1} \leqslant W_t$ для любого t. Отсюда для любого $T > 0$ имеем

$$\frac{W_T}{W_1} = \prod_{t=1}^{T-1} \frac{W_{t+1}}{W_t} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M, \tag{4.2}$$

где $M = L_T$ – общее число ошибок алгоритма $WMA(\epsilon)$ на первых T шагах.

Заметим, что $W_1 = \sum\limits_{i=1}^{N} w_1^i = N$. Из (4.1) и (4.2) следует

$$\frac{(1-\epsilon)^m}{N} < \frac{W_T}{W_1} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M.$$

Вычисляем натуральный логарифм от обоих частей этого неравенства, проводим следующие переходы:

$$
\begin{aligned}
m\ln(1-\epsilon) - \ln N &< M\ln\left(1-\frac{\epsilon}{2}\right)\\
m\ln(1-\epsilon) - \ln N &< -\frac{\epsilon}{2}M\\
m\ln\left(\frac{1}{1-\epsilon}\right) + \ln N &> \frac{\epsilon}{2}M\\
m\left(\frac{2}{\epsilon}\right)\ln\left(\frac{1}{1-\epsilon}\right) + \left(\frac{2}{\epsilon}\right)\ln N &> M\\
\left(\frac{2}{1-\epsilon}\right)m + \left(\frac{2}{\epsilon}\right)\ln N &> M, \qquad (4.3)
\end{aligned}
$$

Вторая строка (4.3) получена из первой с помощью неравенства $\ln(1+x) \leqslant x$, которое имеет место при $x > -1$.

Последняя строка (4.3) получено из предпоследней с помощью неравенства

$$\frac{1}{y}\ln\left(\frac{1}{1-y}\right) \leqslant \frac{1}{1-y}.$$

Это неравенство получается из неравенства $\ln(1+x) \leqslant x$ путем подстановки $x = y/(1-y)$. $\triangle$

Теорема 4.2 показывает, что алгоритм взвешенного большинства WMA ошибается не более чем почти в два раза больше, чем наилучший эксперт.

Исторически, по-видимому, это первый алгоритм такого рода. Он был предложен Литлстоуном и Вармутом в 1989 году и назывался Weighted Majority Algorithm [42]. Несколько позже, в 1990 году, В.Г. Вовк предложил более общий агрегирующий алгоритм (Aggregating Algorithm) и понятие смешиваемой функции потерь, которые работают для игр более общего типа (см. [57]).

4.2. Алгоритм оптимального распределения потерь в режиме онлайн

В этом разделе мы рассмотрим простейшую модель и алгоритм оптимального следования за экспертами в режиме онлайн для то-

го случая, когда нам доступны только величины потерь экспертов на каждом шаге (какая-либо конкретная функция потерь отсутствует). Этот алгоритм бы предложен Фройндом и Шапире [29].

Типичный пример такой задачи: распределитель имеет нескольких друзей, делающих ставки на скачках и выигрывающих или теряющих на каждом шаге некоторые суммы. Распределитель располагает на каждом шаге некоторой суммой, которую он хочет распределить между друзьями с целью получить максимальный выигрыш (или минимальные потери). Естественный критерий оценки успешности стратегии распределителя – его выигрыш, с некоторой точностью, должен быть не меньше чем у наиболее удачливого друга.

Процесс предсказания представим в форме протокола некоторой игры с полной информацией. Участники игры: стратегии или *Эксперты*, $1, 2, \ldots, N$, а также *Распределитель*. Цель *Распределителя* построить такую последовательность предсказаний потери которой были бы не намного больше, чем потери наилучшего эксперта.

На каждом шаге игры $t = 1, 2, \ldots, T$ распределитель определяет вектор распределения стратегий $\bar{p}_t = (p_t^1, \ldots, p_t^N)$, где $p_t^1 + \cdots + p_t^N = 1$ и $p_t^i \geqslant 0$ при $i = 1, 2, \ldots, N$. После этого каждая из стратегий объявляет свои потери на шаге t – число l_t^i, где $i = 1, 2, \ldots, N$. Потери распределителя на шаге t равны смеси потерь экспертов на этом шаге

$$(\bar{p}_t \cdot \bar{l}_t) = \sum_{i=1}^{N} p_t^i l_t^i,$$

где $\bar{l}_t = (l_t^1, \ldots, l_t^N)$ – вектор потерь всех стратегий на шаге t.

Мы будем предполагать, что потери экспертов на каждом шаге ограничены, например, $l_t^i \in [0, 1]$ для всех i и t.

В случае ограниченных на каждом шаге потерь нет принципиальной разницы между алгоритмами, которые добиваются минимальных потерь, и алгоритмами, которые добиваются максимального выигрыша; можно от потерь l_t на каждом шаге перейти к выигрышу $1 - l_t$ и обратно.

Кумулятивные потери *Эксперта* i на шагах $t = 1, 2, \ldots, T$ равны

$$L_T^i = \sum_{t=1}^{T} l_t^i.$$

Соответственно, кумулятивные потери *Распределителя* на шагах $t = 1, 2, \ldots, T$ равны

$$L_T = \sum_{t=1}^{T} (\bar{p}_t \cdot \bar{l}_t).$$

Задача *Распределителя* заключается в выборе таких векторов распределения стратегий $\bar{p}_t$, $t = 1, 2, \ldots, T$, чтобы минимизировать величину

$$R_T = L_T - \min_i L_T^i.$$

Для решения этой задачи рассмотрим алгоритм $Hedge(\beta)$ из работы [29] (см. рис. 4.2). Его параметром является число $\beta \in (0, 1)$, и вектор весов $\bar{w}_1 = (w_1^1, \ldots, w_1^N)$. Предполагаем, что начальные веса всех экспертов удовлетворяют условию $\sum\limits_{i=1}^{N} w_1^i = 1$.

Основной технический результат для алгоритма $Hedge(\beta)$ представлен в следующей лемме.

Лемма 4.1. *Для любой последовательности векторов потерь* $\bar{l}_1, \ldots, \bar{l}_T$ *экспертных стратегий* $1, \ldots, N$ *выполнено неравенство*

$$\ln\left(\sum_{i=1}^{N} w_{T+1}^i\right) \leqslant -(1-\beta)L_T, \tag{4.6}$$

где L_T *– потери алгоритма распределения* $Hedge(\beta)$ *за* T *шагов.*

Доказательство. Из выпуклости экспоненты имеет место неравенство $\beta^r \leqslant 1-(1-\beta)r$ при всех $r \in [0, 1]$ и $0 < \beta < 1$. Используя

FOR $t = 1, 2, \ \ldots, \ T$
Распределитель вычисляет распределение экспертных стратегий:

$$\bar{p}_t = \frac{\bar{w}_t}{\sum\limits_{i=1}^{N} w_t^i}. \tag{4.4}$$

Эксперт i объявляет свои потери l_t^i, $i = 1, 2, \ \ldots, \ N$. Пусть $\bar{l}_t = (l_t^1, \ \ldots, \ l_t^N)$ – вектор потерь всех стратегий на шаге t.
Распределитель подсчитывает свои потери: $l_t = (\bar{p}_t \cdot \bar{l}_t)$.
Распределитель производит пересчет весов экспертных стратегий:

$$w_{t+1}^i = w_t^i \beta^{l_t^i} \tag{4.5}$$

для $i = 1, \ \ldots, \ N$.
ENDFOR

Рис. 4.2: Алгоритм $Hedge(\beta)$

это неравенство и комбинируя (4.4) и (4.5), получаем

$$\begin{aligned} \sum_{i=1}^{N} w_{t+1}^i = \sum_{i=1}^{N} w_t^i \beta^{l_t^i} \leqslant \\ \leqslant \sum_{i=1}^{N} w_t^i (1 - (1-\beta) l_t^i) = \\ = \left(\sum_{i=1}^{N} w_t^i \right) (1 - (1-\beta)(\bar{p}_t \cdot \bar{l}_t)). \end{aligned} \tag{4.7}$$

Последовательно применяя (4.7) при $t = 1, \ldots, T$, получим

$$\sum_{i=1}^{N} w_{T+1}^i \leqslant$$

$$\leqslant \prod_{t=1}^{T} (1 - (1-\beta)(\bar{p}_t \cdot \bar{l}_t)) \leqslant$$

$$\leqslant \exp\left(-(1-\beta)\sum_{t=1}^{T}(\bar{p}_t \cdot \bar{l}_t)\right).$$

Здесь было использовано неравенство $1 + x \leqslant \exp(x)$ для всех x. Мы также использовали свойство $\sum\limits_{i=1}^{N} w_1^i = 1$ для начальных весов.

Отсюда немедленно следует утверждение леммы. $\triangle$

По (4.6) имеем

$$L_T \leqslant \frac{-\ln\left(\sum\limits_{i=1}^{N} w_{T+1}^i\right)}{1-\beta}. \tag{4.8}$$

Из определения весов (4.5) следует

$$w_{T+1}^i = w_1^i \prod_{t=1}^{T} \beta^{l_t^i} = w_1^i \beta^{L_T^i}. \tag{4.9}$$

Отсюда получаем следующую теорему.

Теорема 4.3. *Для любой последовательности векторов потерь $\bar{l}_1, \ldots, \bar{l}_T$ экспертных стратегий $i = 1, \ldots, N$ для произвольных i и T выполнено неравенство*

$$L_T \leqslant \frac{-\ln(w_1^i) - L_T^i \ln\beta}{1-\beta}. \tag{4.10}$$

В случае конечного числа экспертов естественно положить начальные веса экспертных стратегий равными $w_1^i = \frac{1}{N}$ для всех i. Тогда (4.10) можно переписать в виде

$$L_T \leqslant \frac{\ln(1/\beta)}{1-\beta} \min_i L_T^i + \frac{\ln N}{1-\beta}. \tag{4.11}$$

Неравенство (4.11) можно интерпретировать как то, что кумулятивные потери распределительного алгоритма $Hedge(\beta)$ не превосходят потерь наилучшего эксперта, умноженных на константу $\frac{\ln(1/\beta)}{1-\beta}$ плюс ошибка обучения – регрет $\frac{\ln N}{1-\beta}$.

В работе [58] показано, что оценка (4.11) является неулучшаемой. А именно имеет место теорема.

Теорема 4.4. *Пусть B – произвольный алгоритм распределения потерь с произвольным числом экспертов. Допустим, что существуют такие положительные действительные числа a и c, что для произвольного числа N стратегий и для любой последовательности векторов потерь $\bar{l}^1, \ldots, \bar{l}^T$ экспертных стратегий, где $\bar{l}^t = (l_1^t, \ldots, l_N^t)$ при $t = 1, \ldots, T$, выполнено неравенство*

$$L_T(B) \leqslant c \min_i L_T^i + a \ln N.$$

Тогда для всех $\beta \in (0,1)$ будет выполнено одно из неравенств:

$$c \geqslant \frac{\ln(1/\beta)}{1-\beta} \text{ или } a \geqslant \frac{1}{1-\beta}.$$

За счет подбора параметра β можно добиться перераспределения констант так, чтобы мультипликативный множитель в (4.11) стал равным единице за счет увеличения аддитивного множителя.

Лемма 4.2. *Допустим, что $0 \leqslant L \leqslant \tilde{L}$ и $0 \leqslant R \leqslant \tilde{R}$. Пусть также $\beta = g(\tilde{L}/\tilde{R})$, где $g(x) = (1 + \sqrt{2/x})^{-1}$. Тогда*

$$-\frac{\ln \beta}{1-\beta} L + \frac{1}{1-\beta} R \leqslant L + \sqrt{2\tilde{L}\tilde{R}} + R. \tag{4.12}$$

Доказательство. Мы будем использовать следующее неравенство: $-\ln \beta \leqslant \frac{1-\beta^2}{2\beta}$ при $\beta \in (0,1]$. Следующая цепочка преобразо-

ваний приводит к нужному результату:

$$L\frac{-\ln\beta}{1-\beta}+\frac{1}{1-\beta}R\leqslant L\frac{1+\beta}{2\beta}+\frac{1}{1-\beta}R=$$
$$=\frac{1}{2}L\left(1+\frac{1}{\beta}\right)+\frac{1}{1-\beta}R=$$
$$=L+\frac{1}{2}L\sqrt{\frac{2\tilde{R}}{\tilde{L}}}+\frac{1}{1-\frac{1}{1+\sqrt{\frac{2\tilde{R}}{\tilde{L}}}}}R\leqslant$$
$$\leqslant L+\sqrt{\frac{1}{2}\tilde{L}\tilde{R}}+R+R\sqrt{\frac{\tilde{L}}{2\tilde{R}}}\leqslant$$
$$\leqslant L+\sqrt{2\tilde{L}\tilde{R}}+R.$$

Так как мы предполагали, что $0\leqslant l_t^i\leqslant 1$ для всех i и t, кумулятивные потери каждого эксперта ограничены: $L_T^i\leqslant T$ для всех i и T. Поэтому можно в неравенстве (4.12) положить $\tilde{L}=T$. Полагаем также $\tilde{R}=\ln N$. Тогда по лемме 4.2

$$L_T\leqslant\min_i L_T^i+\sqrt{2T\ln N}+\ln N,$$

где L_T – кумулятивные потери алгоритма $Hedge(\beta)$ за T шагов.

Недостатком этой оценки является то, что параметр β зависит от горизонта T. См. также комментарий в конце раздела 4.6.1.

Несколько более точные оценки потерь смешивающего алгоритма будут получены в следующих разделах, где потери экспертов и смешивающего алгоритма на каждом шаге будут вычисляться в виде функции от решения эксперта (предсказания) и исхода природы.

4.3. Усиление простых классификаторов – бустинг

В этом разделе рассматривается метод усиления простых классификаторов, который называется *бустинг* (Boosting). Этот метод основан на комбинировании примитивных “слабых” классификаторов в один “сильный”. Будет изучаться алгоритм AdaBoost

(от английских слов “адаптивность” и “усиление”), предложенный Фройндом и Шапире [29]. Этот алгоритм был успешно использован во многих областях, в частности, в задаче поиска лиц на изображении. Рассматриваемый метод усиления простых классификаторов применяется во многих задачах и до сих пор является объектом множества как прикладных так и теоретических исследований.

Алгоритм AdaBoost. В этом разделе алгоритм оптимального распределения потерь, изложенный в разделе 4.2, будет применен к решению задачи усиления алгоритмов классификации.

Напомним задачу построения классификатора. Предсказатель получает выборку, $S = ((x_1, y_1), \ \ldots, \ (x_l, y_l))$, где $x_i \in X$ и $y_i \in Y$. Мы предполагаем, что $Y = \{0, 1\}$, X – множество. Мы также предполагаем, что для всех i пары (x_i, y_i) одинаково и независимо распределены согласно неизвестному нам распределению вероятностей P на $X \times Y$.

Строгий алгоритм машинного обучения для произвольных $\epsilon, \delta > 0$ при обучении на достаточно большой случайной выборке S с вероятностью $1 - \delta$ выдает гипотезу классификации h_S, которая имеет ошибку обобщения не более ϵ. Кроме этого, время работы такого алгоритма должно полиномиальным образом зависеть от $1/\epsilon$, $1/\delta$ и размера выборки.

Слабый алгоритм машинного обучения по определению должен удовлетворять тем же свойствам, за исключением того, что то же самое выполнено для хотя бы одного $\epsilon \leqslant \frac{1}{2} - \gamma$, где $\gamma > 0$ – константа.

Здесь будет рассматриваться только задача построения гипотезы классификации h_S по обучающей выборке S. Проблема оценки ее предсказательной способности не будет обсуждаться.

Пусть $D(i)$ – произвольное распределение вероятностей на индексах (элементах) выборки. По определению $D(i) \geqslant 0$ для всех i и

$$\sum_{i=1}^{l} D(i) = 1.$$

Естественный пример такого распределения – равномерное распределение на элементах выборки: $D(i) = 1/l$ для всех i.

Ошибка обучения классификатора h на обучающей выборке S относительно распределения D определяется как

$$\epsilon = D\{i : h(x_i) \neq y_i\} = \sum_{i:h(x_i)\neq y_i} D(i).$$

В частности, при распределении $D(i) = 1/l$ ошибка обучения равна доле числа неправильных классификаций объектов:

$$\epsilon = |\{i : h(x_i) \neq y_i\}|/l.$$

Некоторые алгоритмы классификации позволяют использовать распределение $D(i)$ на элементах обучающей выборки в качестве входного параметра. В противном случае, можно использовать ресэмплинг обучающей выборки. Ресэмплинг заключается в том, что мы формируем новую выборку, в которой каждая пара (x_i, y_i) встречается с частотой $D(i)$. Для этого, с помощью генератора случайных чисел, мы выбираем элементы из старой выборки согласно распределению $D(i)$.

В этом разделе мы решаем частный случай общей задачи – мы рассмотрим метод усиления слабого алгоритма классификации на обучающей выборке. На рис. 4.3 приведен алгоритм AdaBoost (предложенный Фройндом и Шапире [29]). Этот алгоритм является *мета-алгоритмом*, он перестраивает произвольный слабый алгоритм классификации, имеющий ошибку обучения $\epsilon \leqslant \frac{1}{2} - \gamma$, в сильный алгоритм, имеющий как угодно малую ошибку обучения ϵ (все ошибки – относительно распределения D).

Приведенный алгоритм представляет собой некоторую версию алгоритма оптимального распределения потерь в режиме онлайн $Hedge(\beta)$ (см. раздел 4.2), в котором параметр β динамически изменяется по шагам алгоритма. Кроме того, рассматривается двойственная версия этого алгоритма. В данном алгоритме веса приписываются не стратегиям, а элементам выборки. Так как теперь потери на шаге t измеряются величиной $l_t^i = 1 - |h_t(x_i) - y_i|$, такие потери равны нулю, если гипотеза h_t неправильно классифицирует объект x_i, и они максимальны (единица), если классификация – правильная. Соответственно, вес неправильной классификации

Вход алгоритма: выборка $S = ((x_1, y_1),\ \ldots,\ (x_l, y_l))$, распределение D на $\{1, \ldots, l\}$, слабый алгоритм классификации WeakLearn. Определим начальные значения весов: $w_1^i = D(i)$ для $i = 1,\ \ldots,\ l$.

FOR $t = 1,\ \ldots,\ T$

1) Полагаем при $i = 1,\ \ldots,\ l$

$$p_t^i = \frac{w_t^i}{\sum\limits_{j=1}^{l} w_t^j}.$$

2) Вызываем алгоритм WeakLearn, в котором $D(i) = p_t^i$ для всех i и который возвращает гипотезу классификации h_t.

3) Вычисляем ошибку обучения классификатора h_t :

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(x_i) - y_i|.$$

4) Полагаем $\beta_t = \epsilon_t/(1 - \epsilon_t)$.

5) Определим адаптированные веса при $i = 1,\ \ldots,\ l$:

$$w_{t+1}^i = w_t^i \beta_t^{1-|h_t(x_i)-y_i|}.$$

ENDFOR

Результат работы алгоритма: выдать гипотезу – индикаторную функцию:

$$h(x) = \begin{cases} 1, \text{ если } f(x) \geqslant \frac{1}{2}, \\ 0 \text{ в противном случае,} \end{cases}$$

где пороговая функция f определяется в виде линейной комбинации гипотез алгоритма WeakLearn

$$f(x) = \sum_{t=1}^{T} q_t h_t(x),$$

с весами

$$q_t = \frac{\ln(1/\beta_t)}{\sum\limits_{t=1}^{T} \ln(1/\beta_t)},$$

при $t = 1, \ldots,\ T$.

Рис. 4.3: Алгоритм AdaBoost

растет, а вес правильной классификации уменьшается. Таким образом, алгоритм AdaBoost выделяет примеры, на которых алгоритм WeakLearn дает неправильные классификации и заставляет его обучаться на этих примерах.

При анализе будет существенно использоваться свойство слабого алгоритма WeakLearn – при любом распределении на элементах выборки его ошибка обучения меньше чем 1/2 на некоторую положительную величину γ.

Результат работы алгоритма AdaBoost оценивается в следующей теореме.

Теорема 4.5. *Предположим, что слабый алгоритм классификации* WeakLearn *при его вызовах алгоритмом* AdaBoost *на шагах* $t = 1, \ldots, T$ *выдает гипотезы с ошибками обучения* $\epsilon_1, \ldots, \epsilon_T$ (*относительно соответствующих распределений, заданных в векторном виде* $\bar{p}_1 = \bar{D}, \bar{p}_2, \ldots, \bar{p}_T$). *Тогда ошибка обучения*

$$\epsilon = D\{h(x_i) \neq y_i\} = \sum_{h(x_i)\neq y_i} D(i)$$

результирующей гипотезы h, *выданной алгоритмом* AdaBoost *после* T *шагов работы, ограничена*

$$\epsilon \leqslant 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}. \tag{4.13}$$

Доказательство. Так же, как в доказательстве лемм 4.1 и 4.2 из раздела 4.2, мы оценим сверху и снизу величину $\sum\limits_{i=1}^{l} w_{T+1}^i$. Имеем верхнюю оценку:

$$\sum_{i=1}^{l} w_{t+1}^i = \sum_{i=1}^{l} w_t^i \beta_t^{1-|h_t(x_i)-y_i|} \leqslant$$

$$\leqslant \sum_{i=1}^{l} w_t^i (1-(1-\beta_t)(1-|h_t(x_i)-y_i|)) =$$

$$= \left(\sum_{i=1}^{l} w_t^i\right)(1-(1-\beta_t)(1-\epsilon_t)). \tag{4.14}$$

Используя (4.14) T раз, получим

$$\sum_{i=1}^{l} w_{T+1}^{i} \leqslant \prod_{t=1}^{T}(1-(1-\beta_t)(1-\epsilon_t)). \tag{4.15}$$

Здесь было использовано определение ошибки обучения ϵ_t алгоритма WeakLearn на шаге t :

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(x_i) - y_i| = \sum_{i=1}^{l} \left(\frac{w_t^i}{\sum\limits_{j=1}^{l} w_t^j} \right) |h_t(x_i) - y_i|.$$

Лемма 4.3. *Результирующий классификатор h делает ошибку на объекте x_i тогда и только тогда, когда*

$$\prod_{t=1}^{T} \beta_t^{-|h_t(x_i)-y_i|} \geqslant \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{4.16}$$

Доказательство. Действительно, это утверждение прямо следует из определения классификатора h в случае, когда $y_i = 0$, так как в таком случае $\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-h_t(x_i)}$ для всех t. По определению равенство $h(x_i) = 1$ может быть тогда и только тогда, когда

$$\sum_{t=1}^{T} \ln(1/\beta_t) h_t(x_i) \geqslant \frac{1}{2} \sum_{t=1}^{T} \ln(1/\beta_t). \tag{4.17}$$

Неравенство (4.17) эквивалентно неравенству (4.16).

Пусть теперь $y_i = 1$. Тогда $h_t(x_i) \leqslant y_i$ для всех t. Поэтому $\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-(1-h_t(x_i))}$ для всех t. В этом случае для всех $1 \leqslant t \leqslant T$

$$\beta_t^{-|h_t(x_i)-y_i|} = \beta_t^{-1+h_t(x_i)}. \tag{4.18}$$

Равенство $h(x_i) = 0$ по определению возможно только при

$$\prod_{t=1}^{T} \beta_t^{-h_t(x_i)} < \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{4.19}$$

Неравенство (4.19) эквивалентно неравенству

$$\prod_{t=1}^{T} \beta_t^{h_t(x_i)} > \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}. \tag{4.20}$$

Неравенство (4.20) с учетом равенства (4.18) эквивалентно неравенству (4.16) леммы. Лемма доказана. $\triangle$

Возвращаясь к доказательству теоремы, заметим, что по определению

$$w_{T+1}^i = D(i) \prod_{t=1}^{T} \beta_t^{1-|h_t(x_i)-y_i|}. \tag{4.21}$$

По лемме 4.3 из (4.16) и (4.21) получаем

$$\begin{aligned} \sum_{i=1}^{l} w_{T+1}^i \geqslant \sum_{i:h(x_i)\neq y_i} w_{T+1}^i \geqslant \\ \geqslant \left(\sum_{i:h(x_i)\neq y_i} D(i) \right) \left(\prod_{t=1}^{T} \beta_t \right)^{1/2} = \\ = \epsilon \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}, \end{aligned} \tag{4.22}$$

где ϵ – ошибка обучения результирующего классификатора h относительно распределения D.

Комбинируя (4.15) и (4.22), получим

$$\epsilon \leqslant \prod_{t=1}^{T} \frac{1-(1-\beta_t)(1-\epsilon_t)}{\sqrt{\beta_t}}. \tag{4.23}$$

Так как элементы произведения (4.23) неотрицательны, можно минимизировать по β_t каждый сомножитель отдельно. Приравниваем к нулю производную по β_t :

$$\frac{d}{d\beta_t} \left(\frac{1-(1-\beta_t)(1-\epsilon_t)}{\sqrt{\beta_t}} \right) = 0$$

и получаем: $\beta_t = \epsilon_t/(1-\epsilon_t)$. Подставляем это выражение для β_t в (4.23) и получаем (4.13). Теорема доказана. $\triangle$

Следствие 4.1. *Ошибка обучения результирующего классификатора h удовлетворяет неравенству*

$$\epsilon = D\{i : h(x_i) \neq y_i\} \leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right), \tag{4.24}$$

где $\epsilon_t = \frac{1}{2} - \gamma_t$, $\gamma_t > 0$ при $t = 1, \ \ldots, \ T$.

В случае, когда $\gamma_t = \gamma$ для всех t, неравенство (4.24) *упрощается до*

$$\epsilon \leqslant \exp(-2T\gamma^2). \tag{4.25}$$

Доказательство. Действительно, в оценке (4.13) теоремы 4.5 при $\epsilon_t = \frac{1}{2} - \gamma_t$ будет

$$2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1-4\gamma_t^2}.$$

Отсюда

$$\epsilon \leqslant \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2} =$$

$$= \exp\left(\sum_{t=1}^{T}\frac{1}{2}\ln(1-4\gamma_t^2)\right) \leqslant$$

$$\leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right). \tag{4.26}$$

Неравенство (4.24) доказано.

Оценка (4.25) представляет собой обычную экспоненциально убывающую оценку ошибки обучения типа неравенства Хефдинга.

Неравенство (4.25) позволяет оценить число итераций алгоритма AdaBoost, необходимых для достижения ошибки обучения $\leqslant \epsilon$ результирующего классификатора h :

$$T \geqslant \frac{1}{2\gamma^2}\ln\frac{1}{\epsilon}.$$

Ошибка обобщения алгоритма AdaBoost. Оценка вероятности ошибки классификатора дана в следствии 1.2. Для его применения необходимо оценить VC-размерность класса, к которому принадлежит классификатор, построенный алгоритмом AdaBoost.

Пусть T – число итераций и все классификаторы h_t, $1 \leqslant t \leqslant T$, выдаваемые алгоритмом AdaBoost, принадлежат классу H, для которого $VCdim(H) = d$. Результирующий классификатор h принадлежит классу

$$\mathcal{L}_T(H) = \{h : h(\mathbf{x}) = \operatorname{sign}(\alpha_1 h_1(\mathbf{x}) + \cdots + \alpha_T h_T(\mathbf{x}) + b)\},$$

где $\alpha_1, \ldots, \alpha_T, b \in \mathcal{R}$.

Пусть $\mathcal{L}_T = \{L : L(\mathbf{y}) = \operatorname{sign}(\alpha_1 y_1 + \cdots + \alpha_T y_T + b)\}$ класс всех линейных классификаторов в эвклидовом пространстве размерности T.

По определению класс $\mathcal{L}_T(H)$ состоит из суперпозиций функций из классов $\mathcal{L}_T$ и H. Оценим его VC-размерность.

Предложение 4.1. $VCdim(\mathcal{L}_T(H)) = O(dT \ln T)$.

Доказательство. Пусть $\mathbf{x}_1, \ldots, \mathbf{x}_m$ – полностью разделимая выборка максимального размера. Тогда

$$2^m = \big|\{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) : h \in \mathcal{L}_T(H)\}\big| \leqslant \sum_{(\mathbf{v}_1,\ldots,\mathbf{v}_T)} \big|\{L(\mathbf{v}_1), \ldots, L(\mathbf{v}_T)) : L \in \mathcal{L}\}\big|. \tag{4.27}$$

В (4.27) суммирование производится по всем наборам $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$ таким, что

$$\begin{aligned} \mathbf{v}_1 &= (h_1(\mathbf{x}_1), h_2(\mathbf{x}_1), \ldots, h_T(\mathbf{x}_1)) \\ \mathbf{v}_2 &= (h_1(\mathbf{x}_2), h_2(\mathbf{x}_2), \ldots, h_T(\mathbf{x}_2)) \\ &\ldots \\ \mathbf{v}_m &= (h_1(\mathbf{x}_m), h_2(\mathbf{x}_m), \ldots, h_T(\mathbf{x}_m)), \end{aligned} \tag{4.28}$$

где $h_1, h_2, \ldots, h_T$ – произвольные функции из класса H.

Так как VC-размерность класса всех линейных классификаторов в эвклидовом пространстве размерности T равна $T+1$, при $m \geqslant T$ для каждого фиксированного набора $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$ будет

$$\big|\{L(\mathbf{v}_1), \ldots, L(\mathbf{v}_T)) : L \in \mathcal{L}\}\big| \leqslant \left(\frac{em}{T+1}\right)^{T+1}.$$

Число всех наборов $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$ не превосходит произведения числа всевозможных столбцов в (4.28), которые получаются при изменении $h_1, h_2, \ldots, h_T \in H$. Число значений одного столбца не превосходит функцию роста класса H, а именно, $B_H(m) \leqslant \left(\frac{em}{d}\right)^d$.

Отсюда $2^m \leqslant \left(\frac{em}{T+1}\right)^{T+1} \left(\left(\frac{em}{d}\right)^d\right)^T$. Отсюда $m = O(dT \ln T)$. $\triangle$

4.4. Алгоритм следования за возмущенным лидером

В этом разделе мы рассмотрим другой общий подход к задаче оптимального распределения потерь – алгоритм следования за возмущенным лидером – “Follow the Perturbed Leader – FPL”. Этот алгоритм еще называется алгоритмом Ханнана по имени его первооткрывателя – см. работу Ханнана [30], а также статью Калаи и Вемпала [36] и монографию Чеза-Бианки и Лугоши [43]. Дальнейшее изучение этого алгоритма проводилось Хуттером и Поландом [31], которые обобщили его на счетный класс экспертов.

При данном подходе мы выбираем наилучшего в прошлом предсказателя – лидера. Для того, чтобы нейтрализовать “враждебные” воздействия природы, мы рандомизируем кумулятивные потери экспертов перед выбором наилучшего эксперта. На каждом шаге алгоритм следования за возмущенным лидером несет те же потери, что и выбранный эксперт. Цель алгоритма – получить кумулятивные потери, которые не превосходят потери наилучшего эксперта с точностью до некоторой ошибки обучения – регрета. Регрет нашего алгоритм имеет тот же порядок, что и алгоритм оптимального распределения потерь или алгоритм взвешенного большинства.

Предсказания с экспертами происходят следующим образом. На каждом шаге t эксперты $i = 1, \dots N$ несут потери s_t^i. Мы предполагаем, что потери экспертов на каждом шаге t ограничены: $0 \leqslant s_t^i \leqslant 1$ для всех i и t.

В начале очередного шага t *Предсказатель* наблюдает кумулятивные потери экспертов $s_{1:t-1}^i = s_1^i + \dots + s_{t-1}^i$ за прошлые шаги $< t$, $i = 1, \dots N$. *Предсказатель* принимает решение следовать за одним из этих экспертов, скажем за экспертом i. В конце шага *Предсказатель* несет те же потери, что и выбранный эксперт i: $s_t = s_t^i$. Кумулятивные потери эксперта исчисляются в виде $s_{1:t} = s_{1:t-1} + s_t = s_{1:t-1} + s_t^i$.

Легко привести пример игры с двумя экспертами, который показывает, что простое следование за наилучшим экспертом может привести к большим потерям *Предсказателя*, значительно превышающим потери каждого из экспертов.

Пусть потери каждого эксперта на шагах $t = 0, 1, \dots, 6$ есть $s_{0,1,2,3,4,5,6}^1 = (\frac{1}{2}, 0, 1, 0, 1, 0, 1)$, $s_{0.1,2,3,4,5,6}^2 = (0, 1, 0, 1, 0, 1, 0)$. Ясно. что в этом случае простой алгоритм "следования за лидером" всегда будет принимать неправильное решение и его кумулятивные потери на каждом шаге будут как минимум в два раза больше чем потери каждого эксперта.

В том случае, когда потери экспертов на каждом шаге ограничены, можно бороться с подобными явлениями путем рандомизации кумулятивных потерь экспертов и только после этого выбирать наилучшего эксперта.

Алгоритм FPL выдает на шаге $t = 1, 2, \dots$ в качестве предсказания номер эксперта i, для которого является минимальной величина

$$s_{1:t-1}^i - \frac{1}{\epsilon}\xi^i,$$

где ϵ это *параметр обучения*, и ξ^i, $i = 1, \dots N$, есть последовательность независимых одинаково распределенных случайных вличин распределенных согласно экспоненциальному закону с плотностью $p(x) = \exp\{-x\}$, $x \geqslant 0$.

Заметим, что можно выбрать эти случайные величины перед процессом обучения алгоритма.

FOR $t = 1, 2 \ldots$
Предсказатель выбирает эксперта, имеющего наименьшие возмущеннные кумулятивные потери на шагах $< t$: [2]

$$I_t = \mathrm{argmin}_{i=1,2,\ldots N} \left\{ s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi^i \right\}.$$

Эксперт i несет потери s^i_t for $i = 1, \ldots, N$.
Предсказатель несет потери $s_t = s^{I_t}_t$.
ENDFOR

Рис. 4.4: Псевдокод FPL алгоритма

Мы будем использовать свойства экспоненциального распределения: $P\{\xi > a\} = e^{-a}$ и $P\{\xi > a + b\} = e^{-b}bP\{\xi > a\}$ для всех неотрицательных значений a и b. Эти и другие свойства экспоненциального распределения предлагаются в виде задач в разделе 4.10.

На шаге t игры каждый из N экспертов несет потери $s^i_t \in [0, 1]$, $i = 1, \ldots N$; кумулятивные потери эксперта i исчисляются

$$s^i_{1:t} = s^i_{1:t-1} + s^i_t.$$

Пусть $\epsilon_t = a/\sqrt{t}$ для всех t, где константа a будет уточнена далее. Мы предполагаем, что $s^i_0 = 0$ для всех i.

Псевдокод FPL алгоритма представлен на рис. 4.4.

Пусть $s_{1:T} = \sum\limits_{t=1}^{T} s^{I_t}_t$ – кумулятивные потери алгоритма FPL на шагах $\leqslant T$.

В следующей ниже теореме дается верхняя оценка среднего значения кумулятивных потерь алгоритма FPL.

Теорема 4.6. *Математическое ожидание кумулятивных потерь алгоритма* FPL *с переменным параметром обучения* $\epsilon_t = \sqrt{\frac{2\ln N}{t}}$ *ограничено сверху кумулятивными потерями наилучшего эксперта плюс регрет:*

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + 2\sqrt{2T\ln N} \tag{4.29}$$

FOR $t = 1, 2 \ldots$
Предсказатель выбирает эксперта, имеющего наименьшие возмущеннные кумулятивные потери на шагах $\leqslant t$:

$$J_t = \mathrm{argmin}_{i=1,2,\ldots N} \left\{ s^i_{1:t} - \frac{1}{\epsilon_t}\xi^i \right\}.$$

Эксперт i несет потери s^i_t for $i = 1, \ldots, N$.
Предсказатель несет потери $s^{J_t}_t$.
ENDFOR

Рис. 4.5: Псевдокод IFPL алгоритма

Доказательство. Анализ оптимальности алгоритма FPL основан на сравнении его потерь с потерями вспомогательного алгоритма IFPL (Infeasible FPL) (see рис. 4.5).

Алгоритм IFPL делает свои предсказания на основе использования величин $s^i_{1:t}$, $i = 1, \ldots N$, которые еще неизвестны *Предсказателю* в начале шага t. По этой причине данный алгоритм физически не реализуем и служит только для анализа потерь алгоритма FPL.

Математическое ожидание одношаговых на шаге t и кумулятивных потерь алгоритмов FPL и IFPL на шаге T обозначим

$$l_t = E(s^{I_t}_t) \text{ и } r_t = E(s^{J_t}_t),$$
$$l_{1:T} = \sum_{t=1}^{T} l_t \text{ и } r_{1:T} = \sum_{t=1}^{T} r_t,$$

соответственно, где $s^{I_t}_t$ – потери алгоритма FPL на шаге t и $s^{J_t}_t$ – потери алгоритма IFPL на шаге t, символ E обозначает математическое ожидание.

Напомним, что $I_t = \mathrm{argmin}_i\{s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi^i\}$ и $J_t = \mathrm{argmin}_i\{s^i_{1:t} - \frac{1}{\epsilon'_t}\xi^i\}$.

Лемма 4.4. *Средние кумулятивные потери алгоритмов* FPL *и*

IFPL *удовлетворяют неравенству*

$$l_{1:T} \leqslant r_{1:T} + \sum_{t=1}^{T} \epsilon_t \tag{4.30}$$

для всех T.

Доказательство. Пусть $c_1, \dots c_N$ – произвольные неотрицательные действительные числа. Для произвольного $1 \leqslant j \leqslant N$ определим числа m_j и m'_j:

$$\begin{aligned} m_j = \min_{i \neq j} \{ s^i_{1:t-1} - \frac{1}{\epsilon_t} c_i \} \leqslant \\ \leqslant \min_{i \neq j} \{ s^i_{1:t-1} + s^i_t - \frac{1}{\epsilon_t} c_i \} = \\ = \min_{i \neq j} \{ s^i_{1:t} - \frac{1}{\epsilon_t} c_i \} = m'_j. \end{aligned}$$

Производим сравнение условных вероятностей:

$$P\{I_t = j | \xi^i = c_i, i \neq j\} \text{ и } P\{J_t = j | \xi^i = c_i i \neq j\}$$

Имеет место следующая цепочка равенств и неравенств:

$$\begin{aligned} P\{I_t = j | \xi^i = c_i, i \neq j\} = \\ = P\{ s^j_{1:t-1} - \frac{1}{\epsilon_t} \xi^j \leqslant m_j | \xi^i = c_i, i \neq j\} = \\ = P\{ \xi^j \geqslant \epsilon_t (s^j_{1:t-1} - m_j) | \xi^i = c_i, i \neq j\} = \\ = e^{\epsilon_t} P\{ \xi^j \geqslant \epsilon_t (s^j_{1:t-1} - m_j + 1) | \xi^i = c_i, i \neq j\} \leqslant \\ \leqslant e^{\epsilon_t} P\{ \xi^j \geqslant \epsilon_t (s^j_{1:t-1} + s^j_t - m_j) | \xi^i = c_i, i \neq j\} \leqslant \\ \leqslant e^{\epsilon_t} P\{ \xi^j \geqslant \epsilon_t (s^j_{1:t} - m'_j) | \xi^i = c_i, i \neq j\} = \\ = e^{\epsilon_t} P\{ s^j_{1:t} - \frac{1}{\epsilon_t} \xi^j \leqslant m'_j | \xi^i = c_i, i \neq j\} = \\ = e^{\epsilon_t} P\{ J_t = j | \xi^i = c_i, i \neq j\}. \end{aligned} \tag{4.31}$$

При переходе от 3-й строки к 4-й мы использовали неравенство $P\{\xi \geqslant a + b\} \leqslant e^{-b} P\{\xi \geqslant a\}$ для случайной величины ξ, распределенной согласно экспоненциальному закону, где a и b – произвольные неотрицательные вещественные числа.

Так как эти оценки имеют место при всех условиях c_i, они также имеют место в безусловном виде:

$$P\{I_t = j\} \leqslant e^{\epsilon_t} P\{J_t = j\}. \tag{4.32}$$

для всех $t = 1, 2, \ldots$ и $j = 1, \ldots N$.

Суммируем (4.32) по $j = 1\ , \ldots,\ N$ и получим неравенство

$$l_t = E(s_t^{I_t}) = \sum_{j=1}^{N} s_t^j P\{I_t = j\} \leqslant e^{\epsilon_t} \sum_{j=1}^{N} s_t^j P\{J_t = j\} = e^{\epsilon_t} r_t.$$

Неравенство $l_t - r_t \leqslant \epsilon_t l_t$ следует из неравенства $r_t \geqslant e^{-r} l_t \geqslant (1-r) l_t$ при $r \leqslant 1$. Суммируем эти неравенства по $t = 1\ , \ldots,\ T$ и берем во внимание $0 \leqslant l_t \leqslant 1$ для всех t. В результате получим

$$l_{1:T} \leqslant r_{1:T} + \sum_{t=1}^{T} \epsilon_t \leqslant r_{1:T} + 2a\sqrt{T}.$$

Лемма доказана. $\triangle$

В следующей лемме мы получим верхнюю границу средних кумулятивных потерь алгоритма IFPL.

Лемма 4.5. *Математическое ожидание кумулятивных потерь алгоритма* IFPL *ограничено сверху*

$$r_{1:T} \leqslant \min_i s_{1:T}^i + \frac{\ln N}{\epsilon_T} \tag{4.33}$$

для всех T.

Доказательство. Введем в этом доказательстве $\mathbf{s}_t = (s_t^1, \ldots, s_t^N)$ – вектор одношаговых потерь экспертов и $\mathbf{s}_{1:t} = (s_{1:t}^1, \ldots, s_{1:t}^N)$ – вектор кумулятивных потерь экспертов. Пусть также $\xi = (\xi^1, \ldots, \xi^N)$ – вектор координатами которого являются экспоненциально распределенные случайные величины.

Рассмотрим вспомогательные векторы:

$$\tilde{\mathbf{s}}_t = \mathbf{s}_t - \xi \left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}} \right) \tag{4.34}$$

$$\tilde{\mathbf{s}}_{1:t} = \mathbf{s}_{1:t} - \frac{1}{\epsilon_t} \xi \tag{4.35}$$

при $t = 1, 2, \ldots$.

Для произвольного вектора $\mathbf{s} = (s^1, \ \ldots, \ s^N)$ и единичного вектора $\mathbf{d} = (0, \ \ldots, \ 1, \ \ldots, \ 0)$ обозначим

$$M(\mathbf{s}) = \mathrm{argmin}_{\mathbf{d} \in D}\{\mathbf{d} \cdot \mathbf{s}\},$$

где $D = \{(0, \ \ldots, \ 1), \ \ldots, \ (1, \ \ldots, \ 0)\}$ – множество, состоящее из N размерности N и "$\cdot$" – скалярное произведение.

По определению $M(\mathbf{s})$ есть единичный вектор, i-я координата которого равна 1, где $s^i = \min\limits_{1 \leqslant j \leqslant N} s^j$. Если имеется более одного такого i, то полагаем $M(\mathbf{s})$ равным наименьшему из них.

По определению $(M(\mathbf{s}) \cdot \mathbf{s}) = \min\limits_{1 \leqslant j \leqslant N} s^j$.

По определению алгоритма IFPL

$$r_{1:T} = E\left(\sum_{t=1}^{T} M(\tilde{s}_{1:t}) s_t\right).$$

Таким образом, нам необходимо оценить сумму под знаком математического ожидания.

Предварительно покажем, что

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \tilde{\mathbf{s}}_t \leqslant M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T}. \tag{4.36}$$

Доказательство проводим методом математической индукции по T. Для $T = 1$ утверждение очевидно. Для того, чтобы сделать шаг индукции от $T - 1$ к T сделаем два замечания.

Имеем $\tilde{\mathbf{s}}_{1:T} = \tilde{\mathbf{s}}_{1:T-1} + \tilde{\mathbf{s}}_T$ по определению, а также

$$M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T-1} \geqslant M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_{1:T-1},$$

так как правая часть этого неравенства равна минмальной координате вектора $\tilde{\mathbf{s}}_{1:T-1}$, тогда как левая его часть равна координате, которая выбиралась по другому критерию.

Соединяем оба эти замечания вместе и получаем утверждение индукции (4.36) для шага T используя предположение индукции

для шага $T-1$:

$$M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{1:T} = M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{T} \geqslant$$
$$\geqslant M(\tilde{\mathbf{s}}_{1:T-1})\cdot\tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{T} \geqslant$$
$$\geqslant \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t})\cdot\tilde{\mathbf{s}}_{t}.$$

Вспоминая определение (4.34) вектора $\tilde{\mathbf{s}}_t$, мы можем переписать (4.36) следующим образом:

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t})\cdot\mathbf{s}_t \leqslant M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{1:T} + \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t})\cdot\xi\left(\frac{1}{\epsilon_t}-\frac{1}{\epsilon_{t-1}}\right). \quad (4.37)$$

Аналогично, используя определение (4.35) вектора $\tilde{\mathbf{s}}_{1:t}$ и то что критерий выбора координаты вновь был изменен, получаем неравенство

$$M(\tilde{\mathbf{s}}_{1:T})\cdot\tilde{\mathbf{s}}_{1:T} \leqslant M(\mathbf{s}_{1:T})\cdot\left(\mathbf{s}_{1:T}-\frac{\xi}{\epsilon_T}\right) =$$
$$= \min_{\mathbf{d}\in D}\{\mathbf{d}\cdot\mathbf{s}_{1:T}\} - \frac{M(\mathbf{s}_{1:T})\cdot\xi}{\epsilon_T}. \quad (4.38)$$

По определению $(M(\mathbf{s}_{1:T})\cdot\xi) = \xi^k$ для некоторого k.

Так как $E(\xi) = 1$ для экспоненциально распределенной случайной величины ξ, математическое ожидание вычитаемого члена в (4.38) равно

$$E\left(\frac{M(\mathbf{s}_{1:T})\cdot\xi}{\epsilon_T}\right) = \frac{1}{\epsilon_T}E(\xi^k) = \frac{1}{\epsilon_T}. \quad (4.39)$$

Второй член (4.37) удовлетворяет

$$\sum_{t=1}^{T}(M(\tilde{\mathbf{s}}_{\mathbf{1:t}})\cdot\xi)\left(\frac{1}{\epsilon_t}-\frac{1}{\epsilon_{t-1}}\right) \leqslant$$
$$\leqslant \sum_{t=1}^{T}\max_{1\leqslant i\leqslant N}\xi^i\left(\frac{1}{\epsilon_t}-\frac{1}{\epsilon_{t-1}}\right) = \frac{1}{\epsilon_T}\max_{1\leqslant i\leqslant N}\xi^i. \quad (4.40)$$

Здесь мы использовали свойство $\epsilon_t < \epsilon_{t-1}$ для всех t.

Мы будем использовать верхнюю оценку для математического ожидания максимума экспоненциально распределенных случайных величин:

$$0 \leqslant E(\max_{1 \leqslant i \leqslant N} \xi^i) \leqslant 1 + \ln N. \tag{4.41}$$

Действительно, для экспоненциально распределенных случайных величин ξ^i, $i = 1, \dots N$, выполнено

$$P\{\max_i \xi^i \geqslant a\} = P\{\exists i(\xi^i \geqslant a)\} \leqslant$$
$$\leqslant \sum_{i=1}^{N} P\{\xi^i \geqslant a\} = N \exp\{-a\}. \tag{4.42}$$

Для произвольной неотрицательной случайной величины η выполнено

$$E(\eta) = \int_0^\infty P\{\eta \geqslant y\} dy. \tag{4.43}$$

Доказательство этого соотношение предоставляется читателю в виде задачи из раздела 4.10. Тогда по (4.42) имеем

$$E(\max_i \xi^i - \ln N) =$$
$$= \int_0^\infty P\{\max_i \xi^i - \ln N \geqslant y\} dy \leqslant$$
$$\leqslant \int_0^\infty N \exp\{-y - \ln N\} dy = 1.$$

Следовательно, $E(\max_i \xi^i) \leqslant 1 + \ln N$. Согласно (4.41) математическое ожидание (4.40) ограничено сверху числом $\frac{1}{\epsilon_T}(1 + \ln N)$.

Комбинируя оценки (4.37)–(4.40) и (4.39), получим

$$r_{1:T} = E\left(\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{\mathbf{1:t}}) \cdot \mathbf{s_t}\right) \leqslant$$
$$\leqslant \min_i s^i_{1:T} + \frac{\ln N}{\epsilon_T}. \tag{4.44}$$

Лемма доказана. $\triangle$.

Завершим доказательство теоремы.

Неравенство (4.30) леммы 4.4 и неравенство (4.33) леммы 4.5 влекут неравенство

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + a\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + \frac{1}{a}\ln N\sqrt{T} \leqslant$$
$$\leqslant \min_i s^i_{1:T} + 2a\sqrt{T} + \frac{1}{a}\ln N\sqrt{T}. \qquad (4.45)$$

для всех T. Минимизируем (4.45) по a, получим оптимальное значение $a = \sqrt{2\ln N}$. Таким образом, мы получили оценку (4.29) теоремы

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + 2\sqrt{2T\ln N}.$$

Теорема доказана. $\triangle$

Мы также получим следствие этой теоремы. В этом следствии, используя варианты неравенства Хефдинга, мы заменим оценку для среднего значения на вероятностную оценку для кумулятивных потерь.

Для этого нам необходимо усложнить рандомизацию, применяемую в алгоритме FPL. Прежде мы на каждом шаге использовали одну и ту же последовательность независимых одинаково распределенных случайных величин $\xi^1, \ldots, \xi^N$. Мы модифицируем алгоритмы FPL и IFPL следующим образом. Рассмотрим бесконечную последовательность серий независимых одинаково распределенных согласно экспоненциальному закону случайных величин $\xi^t_1, \ldots, \xi^t_N$, $t = 1, 2, \ldots$, так, что все эти случайные величины рассматриваемые вместе независимы.

В алгоритме FPL на рис. 4.4 на шаге t мы будем возмущать каждого эксперта с помощью серии случайных величин $\xi^1_t, \ldots, \xi^N_t$. *Предсказатель* выбирает эксперта, имеющего наименьшие возмущенные кумулятивные потери на шагах $< t$:

$$I_t = \mathrm{argmin}_{i=1,2,\ldots N}\left\{s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi^i_t\right\}.$$

Аналогичное изменения вносим в алгоритм IFPL.

В этом случае одношаговые потери s_t, $t = 1, 2, \ldots$, алгоритма FPL будут независимыми случайными величинами.

Доказательства леммы 4.4 остается тем же, доказательство леммы 4.5 изменяется незначительно, надо только в неравенствах (4.37), (4.38) и (4.40) сразу рассмотреть математическое ожидание от обоих их частей и использовать то, что $E(\xi_t^i) = 1$ для всех i и t.

Следствие 4.2. *Для произвольного $\delta > 0$ с вероятностью $1 - \delta$ выполнено неравенство*

$$s_{1:T} \leqslant \min_i s_{1:T}^i + 2\sqrt{2T \ln N} + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}. \tag{4.46}$$

Алгоритм FPL *является асимптотически состоятельным (или состоятельным по Ханнану):*

$$\limsup_{T \to \infty} \frac{1}{T}(s_{1:T} - \min_{i=1,\ldots N} s_{1:T}^i) \leqslant 0 \tag{4.47}$$

с вероятностью 1.

Доказательство. Для доказательства первого утверждения мы используем вариант неравенство Чернова (10.3) из следствия 10.2:

Пусть $X_1, X_2, \ldots$ – последовательность независимых случайных величин таких, что при всех $i = 1, 2, \ldots$ выполнено $0 \leqslant X_i \leqslant 1$. Тогда для любого $\epsilon > 0$,

$$P\left\{\sum_{i=1}^T X_i - E\sum_{i=1}^T X_i > \epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{T}\right). \tag{4.48}$$

Полагаем $\delta = \exp\left(-\frac{2\epsilon^2}{T}\right)$. Отсюда $\epsilon = \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}$. При $X_t = s_t$ из неравенства (4.48) следует, что с вероятностью $1 - \delta$,

$$\sum_{t=1}^T s_t \leqslant E(s_{1:T}) + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}$$

Из этого неравенства и оценки (4.29) теоремы 4.6 получаем неравенство (4.46).

Для доказательства утверждения (4.47) мы применим другой вариант (10.4) неравенства Чернова:

$$P\left\{\left|\frac{1}{T}\sum_{i=1}^{T}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2T\epsilon^2\right). \qquad (4.49)$$

Здесь полагаем $X_t = s_t$. Так как для любого $\epsilon > 0$ ряд экспонент в правой части этого неравенства сходится, по лемме Бореля–Кантелли:

$$\lim_{T\to\infty}\frac{1}{T}(s_{1:T} - E(s_{1:T})) = 0$$

с вероятностью 1. Из этого соотношения и оценки (4.29) теоремы 4.6 получаем неравенство (4.47) для верхнего предела.

При доказательстве теоремы 4.6 и следствия 4.2 мы предполагали, что потери s_t^i экспертов на каждом шаге t не зависят от номеров экспертов $I_{t'}$, выбранных *Предсказателем* на шагах [3] $t' < t$.

Можно показать, что следствие 4.2 верно и в случае экспертов, потери которых на каждом шаге t зависят от значений случайных величин $I_{t'}$ при $t' < t$. В этом случае случайные величины $X_t = s_t = s_t^{I_t}$ не будут независимыми, но величины $X_t - E(X_t|X_1,\ldots,X_{t-1})$ образуют мартингал-разность относительно последовательности случайных величин $X_1, X_2, \ldots,$ и мы можем применить соответствующее неравенство Хефдинга–Адзумы (10.7) и усиленный мартингальный закон больших чисел (10.8). В частности, условие асимптотической состоятельности (4.47) также выполнено и в этом случае.

[3] В данном протоколе это эквивалентно предположению о том, что потери экспертов $s_1^i, s_2^i, \ldots$ при $i = 1,\ldots,N$ заданы заранее и предъявляются *Предсказателю* по шагам согласно протоколу.

4.5. Алгоритмы экспоненциального смешивания: общий подход

В этом разделе мы рассмотрим общие свойства алгоритмов экспоненциального смешивания. При анализе этих алгоритмов основную роль играет так называемые экспоненциально смешанные потери (Mixability loss – Mixloss). Анализ свойств этой величины прост и эффективен. Кумулятивный вариант этой величины асимптотически близок к потерям каждого из вариантов алгоритмов экспоненциального смешивания. Поэтому анализ свойств этих алгоритмов, в основном, сводится к анализу экспоненциально смешанных потерь.

4.5.1. Экспоненциально смешанные потери и их свойства

Алгоритм оптимального распределения потерь Hedge имеет наиболее общую формулировку среди алгоритмов экспоненциального смешивания. Мы уже рассматривали его в разделе 4.2. В этом разделе мы рассмотрим этот алгоритм и его разновидности в свете общей теории. Для удобства на рис. 4.6 еще раз приведен его протокол.

Пусть $L_T^i = \sum\limits_{t=1}^{T} l_t^i$ – кумулятивные потери эксперта i. Из определения следует, что веса, используемые на шаге t, можно записать также в виде $w_{i,t} = w_{i,1} e^{-\eta L_{t-1}^i}$ при $i = 1, \ldots, N$.

Пусть $H_T = \sum\limits_{t=1}^{T} h_t$ – кумулятивные потери алгоритма $Hedge(\eta)$. Обозначим $W_t = \sum\limits_{i=1}^{N} w_{i,t}$ – потенциал на шаге t. Из определения $W_1 = 1$. Из определения весов следует, что потенциал также равен $W_t = \sum\limits_{i=1}^{N} w_{i,1} e^{-\eta L_{t-1}^i}$, а нормированные веса экспертов равны

$$w_{i,t}^* = \frac{w_{i,t}}{W_t}$$

для всех i и t. Определим *экспоненциально смешанные потери*

Алгоритм $Hedge(\eta)$
Заданы нормированные начальные веса экспертов $w_{i,1}$ на первом шаге при $i = 1, \ldots, N$, $\sum\limits_{i=1}^{N} w_{i,1} = 1$; обычно задаются равномерные веса $w_{i,1} = 1/N$. Задан параметр обучения $\eta > 0$. Полагаем $L_0^i = 0$ для всех i.

FOR $t = 1, \ldots, T$
Нормализуем накопленные веса экспертов

$$w_{i,t}^* = \frac{w_{i,t}}{\sum\limits_{i=1}^{N} w_{i,t}} \text{ при } i = 1, \ldots, N.$$

Получаем потери экспертов l_t^i при $i = 1, \ldots, N$.
Вычисляем потери алгоритма $Hedge(\eta)$ на шаге t

$$h_t = \sum_{i=1}^{N} w_{i,t}^* l_t^i.$$

(В терминологии раздела 4.2 это потери *Предсказателя.*)
Переопределяем веса всех экспертов $i = 1, \ldots, N$ для использования на следующем шаге

$$w_{i,t+1} = w_{i,t} e^{-\eta l_t^i}.$$

ENDFOR

Рис. 4.6: Алгоритм Hedge с постоянным параметром обучения

(mixloss) на шаге t

$$m_t = -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,t}^{*} e^{-\eta l_t^i}.$$

Используя введенные выше обозначения можно также записать величину m_t в виде

$$m_t = -\frac{1}{\eta} \ln \frac{1}{W_t} \sum_{i=1}^{N} w_{i,t} e^{-\eta l_t^i} =$$

$$= -\frac{1}{\eta} \ln \frac{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta L_{t-1}^i} e^{-\eta t_t^i}}{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta L_{t-1}^i}} =$$

$$= -\frac{1}{\eta} \ln \frac{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta L_t^i}}{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta L_{t-1}^i}} = -\frac{1}{\eta} \ln \frac{W_{t+1}}{W_t}. \tag{4.50}$$

Пусть $M_T = \sum\limits_{t=1}^{T} m_t$ – кумулятивные смешанные потери за первые T шагов.

Лемма 4.6. *Кумулятивные смешанные потери M_T можно представить в виде*

$$M_T = -\frac{1}{\eta} \ln W_{T+1} = -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,1} e^{-\eta L_T^i}. \tag{4.51}$$

Доказательство. Действительно, согласно (4.50),

$$M_T = \sum_{t=1}^{T} m_t = -\frac{1}{\eta} \sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} =$$

$$= -\frac{1}{\eta} \ln \prod_{t=1}^{T} \frac{W_{t+1}}{W_t} = -\frac{1}{\eta} \ln W_{T+1} =$$

$$= -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,T+1} = -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,1} e^{-\eta L_T^i}.$$

Здесь $W_1 = 1$. $\triangle$

Так как сумма не меньше чем любое слагаемое, получаем следующие полезные неравенства.

Следствие 4.3. *Кумулятивные смешанные потери и потери произвольного эксперта i связаны неравенствами*

$$M_T \leqslant -\frac{1}{\eta} \ln w_{i,T+1} = L_T^i + \frac{1}{\eta} \ln \frac{1}{w_{i,1}}. \tag{4.52}$$

В случае равномерных начальных весов неравенство (4.52) имеет вид

$$M_T \leqslant L_T^i + \frac{\ln N}{\eta}.$$

Кроме этого, из второго равенства (4.51) получаем

$$\min_{1 \leqslant i \leqslant N} L_T^i \leqslant M_T$$

для каждого T.

Введем обозначения m_t^η и M_T^η для соответствующих величин m_t и M_T определенных при значении параметра обучения равном η.

Отметим важное для дальнейшего свойство величины M_T^η – монотонность по параметру η.

Лемма 4.7. *$M_T^\eta \geqslant M_T^\mu$ при $\eta \leqslant \mu$.*

Доказательство. Из выпуклости параболы имеем

$$M_T^\eta = -\frac{1}{\eta} \ln \sum_{i=1}^N w_{i,1} e^{-\eta L_T^i} = -\frac{1}{\eta} \ln \sum_{i=1}^N w_{i,1} \left(e^{-\mu L_T^i}\right)^{\frac{\eta}{\mu}} \geqslant$$

$$\geqslant -\frac{1}{\eta} \ln \left(\sum_{i=1}^N w_{i,1} e^{-\mu L_T^i}\right)^{\frac{\eta}{\mu}} = -\frac{1}{\mu} \ln \sum_{i=1}^N w_{i,1} e^{-\mu L_T^i} = M_T^\mu.$$

$\triangle$

Изучим степень близости экспоненциально смешанных потерь m_t и потерь алгоритма h_t. Из выпуклости функции $-\ln$ имеем

$$m_t = -\ln \left(\sum_{i=1}^N w_{i,t}^* e^{-\eta l_t^i}\right)^{\frac{1}{\eta}} \leqslant \sum_{i=1}^N w_{i,t}^* l_t^i = h_t. \tag{4.53}$$

Определим $\delta_t = h_t - m_t$. Из неравенства (4.53) следует, что $\delta_t \geqslant 0$ для всех t. Определим $\Delta_T = \sum\limits_{t=1}^{T} \delta_t$.

Регрет алгоритма $Hedge(\eta)$ относительно эксперта i определяется $R_T^i = H_T - L_T^i$. Представим эту величину в виде

$$R_T^i = H_T - L_T^i = \sum_{t=1}^{T} m_t - L_T^i + \sum_{t=1}^{T} \delta_t = M_T - L_T^i + \Delta_T.$$

Из неравенства (4.52) следствия 4.3 получаем

$$R_T^i \leqslant L_T^i + \frac{1}{\eta} \ln \frac{1}{w_{i,1}} - L_T^i + \Delta_T = \frac{1}{\eta} \ln \frac{1}{w_{i,1}} + \Delta_T.$$

В случае равномерных начальных весов $w_{i,1} = 1/N$ выполнено

$$R_T^i \leqslant \frac{\ln N}{\eta} + \Delta_T$$

для любого i.

Для оценки величины отклонения $\delta_t = h_t - m_t$ используем неравенство Хефдинга $\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8}$, где X – ограниченная случайная величина, $a \leqslant X \leqslant b$, s – любое число. Это неравенство доказано в разделе 10 ниже.

Предполагаем, что на каждом шаге t значения потерь l_t^i всех экспертов i, $1 \leqslant i \leqslant N$, лежат в некотором интервале вещественных чисел длины L. Для каждого t рассмотрим в качестве X случайную величину, которая при $i = 1, \ldots, N$ принимает значение l_t^i с вероятностью $w_{i,t}^*$. Полагаем $s = -\eta$. Тогда

$$\ln E(e^{-\eta X}) \leqslant -\eta E(X) + \frac{\eta^2 L^2}{8}.$$

Здесь $E(X) = \sum\limits_{i=1}^{N} w_{i,t}^* l_t^i = h_t$ и $\ln E(e^{-\eta X}) = \ln \sum\limits_{i=1}^{N} w_{i,t}^* e^{-\eta l_t^i} = -\eta m_t$. Отсюда получаем неравенство

$$h_t \leqslant m_t + \frac{\eta L^2}{8}.$$

Таким образом, $\delta_t \leqslant \frac{\eta L^2}{8}$ и регрет алгоритма относительно любого эксперта i оценивается как

$$R_T^i = H_T - L_T^i \leqslant M_T - L_T^i + \frac{\eta T L^2}{8} \leqslant \frac{1}{\eta} \ln \frac{1}{w_{i,1}} + \frac{\eta T L^2}{8}$$

для любого i. Сформулируем этот результат в виде теоремы.

Теорема 4.7. *Пусть на каждом шаге t значения потерь l_t^i всех экспертов i, $1 \leqslant i \leqslant N$, лежат в некотором интервале вещественной прямой длины L. Тогда для каждого T и любого эксперта i имеет место оценка*

$$R_T^i \leqslant \frac{1}{\eta} \ln \frac{1}{w_{i,1}} + \frac{\eta T L^2}{8}.$$

Определим $R_T = \max_{1 \leqslant i \leqslant N} R_T^i$ – регрет алгоритма $Hedge(\eta)$. Если используются равномерные начальные веса, то теорема 4.7 принимает следующую форму.

Следствие 4.4. *В условиях теоремы 4.7 и при равномерных начальных весах $w_{i,1} = 1/N$ будет*

$$R_T \leqslant \frac{\ln N}{\eta} + \frac{\eta T L^2}{8}$$

или

$$H_T \leqslant \min_{1 \leqslant i \leqslant N} L_T^i + \frac{\ln N}{\eta} + \frac{\eta L^2 T}{8}$$

для каждого T. При $\eta = \sqrt{\frac{8 \ln N}{L^2 T}}$ в случае равномерных начальных весов будет

$$H_T \leqslant \min_{1 \leqslant i \leqslant N} L_T^i + L\sqrt{\frac{1}{2} T \ln N}. \tag{4.54}$$

4.5.2. Анализ алгоритма Hedge с переменным параметром обучения

Анализ алгоритма Hedge с переменным параметром обучения основан на свойстве монотонности кумулятивных потерь M_T^η по параметру η.

Протокол игры для алгоритма Hedge с переменным параметром обучения приведен на рис. 4.7. Отличие от протокола алгоритма с постоянным параметром обучения, представленном на рис. 4.6, заключается в том как определяются веса $w_{i,t+1}$.

Пусть $H_T = \sum\limits_{t=1}^{T} h_t$ – кумулятивные потери алгоритма Hedge, обозначим $W_t = \sum\limits_{i=1}^{N} w_{i,t}$. Тогда по определению

$$w^*_{i,t} = \frac{w_{i,t}}{W_t}.$$

Введем экспоненциально смешанные потери алгоритма на шаге t

$$m_t = -\frac{1}{\eta_t} \ln \sum_{i=1}^{N} w^*_{i,t} e^{-\eta_t l^i_t}.$$

Используя введенные выше обозначения можно также записать величину m_t в виде

$$m_t = -\frac{1}{\eta_t} \ln \frac{1}{W_t} \sum_{i=1}^{N} w_{i,t} e^{-\eta_t l^i_t} =$$
$$= -\frac{1}{\eta_t} \ln \frac{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta_t L^i_t}}{\sum\limits_{i=1}^{N} w_{i,1} e^{-\eta_t L^i_{t-1}}}.$$

Кумулятивные смешанные потери равны $M_T = \sum\limits_{t=1}^{T} m_t$. Определим $\delta_t = h_t - m_t$. Имеем $\delta_t \geqslant 0$ для всех t. Определим $\Delta_T = \sum\limits_{t=1}^{T} \delta_t$.

Регрет алгоритма Hedge относительно произвольного эксперта i представляется в виде

$$R^i_T = H_T - L^i_T = \sum_{t=1}^{T} m_t - L^i_T + \sum_{t=1}^{T} \delta_t = M_T - L^i_T + \Delta_T.$$

Алгоритм Hedge

Заданы начальные веса экспертов $w_{i,1}$ на первом шаге при $i = 1, \ldots, N$, $\sum\limits_{i=1}^{N} w_{i,1} = 1$. Параметр обучения переменный: на шаге t его значение равно η_t, где $0 < \eta_{t+1} \leqslant \eta_t$ для всех t. Полагаем $L_0^i = 0$ для всех i.

FOR $t = 1, 2, \ldots$

Нормализуем накопленные веса экспертов

$$w_{i,t}^* = \frac{w_{i,t}}{\sum\limits_{i=1}^{N} w_{i,t}} \text{ при } i = 1, \ldots, N.$$

Получаем потери экспертов l_t^i при $i = 1, \ldots, N$.

Вычисляем потери алгоритма Hedge на шаге t

$$h_t = \sum_{i=1}^{N} w_{i,t}^* l_t^i.$$

$L_t = L_{t-1} + l_t^i$ – определяем кумулятивные потери эксперта i за t шагов.

Переоопределяем веса всех экспертов $i = 1, \ldots, N$ для использования на следующем шаге

$$w_{i,t+1} = w_{i,1} e^{-\eta_t L_t^i}.$$

ENDFOR

Рис. 4.7: Алгоритм Hedge с переменным параметром обучения

Обозначим m_t^η, M_t^η – смешанные потери, кумулятивные смешанные потери, соответственно, алгоритма $Hedge(\eta)$ из раздела 4.5.1, который использует на всех шагах $t' \leqslant t$ постоянный параметр обучения $\eta = \eta_t$.

По определению $m_t = M_t^{\eta_t} - M_{t-1}^{\eta_t}$ для всех t. В случае переменного параметра обучения невозможно доказать утверждение аналогичное лемме 4.6. Тем не менее, используя свойство монотонности кумулятивных смешанных потерь по параметру η, можно получить утверждение аналогичное следствию 4.3. Впервые эта идея была использована в работе [37].

Предварительно докажем неравенство, связывающее кумулятивные смешанные потери при переменном параметре обучения и кумулятивные смешанные потери при параметре обучения последнего шага.

Лемма 4.8. *$M_T \leqslant M_T^{\eta_T}$ для всех T.*

Доказательство Используя лемму 4.7 и монотонность $M_{t-1}^{\eta_t}$ по параметру η_t, получаем

$$M_T = \sum_{t=1}^{T} m_t = \sum_{t=1}^{T} M_t^{\eta_t} - M_{t-1}^{\eta_t} \leqslant \sum_{t=1}^{T} M_t^{\eta_t} - M_{t-1}^{\eta_{t-1}} = M_T^{\eta_T}.$$

Здесь $M_0^{\eta_0} = 0$. $\triangle$

Приведем неравенство, связывающее кумулятивные смешанные потери и потери произвольного эксперта i.

Следствие 4.5. *Для произвольного эксперта i*

$$M_T \leqslant L_T^i + \frac{1}{\eta_T} \ln \frac{1}{w_{i,1}}$$

для всех T.

Доказательство. Действительно, из лемм 4.6 и 4.8 получаем

$$M_T \leqslant M_T^{\eta_T} = \sum_{t=1}^{T} m_t^{\eta_T} = -\frac{1}{\eta_T} \ln W_{T+1}^{\eta_T} =$$

$$= -\frac{1}{\eta_T} \ln \sum_{i=1}^{N} w_{1,i} e^{-\eta_T L_T^i} \leqslant L_T^i + \frac{1}{\eta_T} \ln \frac{1}{w_{i,1}}.$$

для всех T. $\triangle$

Как и в теореме 4.7 для сравнения m_t и h_t используем неравенство Хефдинга. Предполагаем, что для каждого t потери всех экспертов l_t^i лежат в интервале длины L. Тогда для любого t

$$\ln E(e^{-\eta_t X}) \leqslant -\eta_t E(X) + \frac{\eta_t^2 L^2}{8}$$

или

$$h_t = E(X) \leqslant -\frac{1}{\eta_t} \ln \sum_{i=1}^{N} w_{i,t}^* e^{-\eta_t l_t^i} + \frac{\eta_t L^2}{8} = m_t + \frac{\eta_t L^2}{8}.$$

Таким образом, $\delta_t = h_t - m_t \leqslant \frac{\eta_t L^2}{8}$. Тогда

$$R_T^i = H_T - L_T^i \leqslant M_T - L_T^i + \sum_{t=1}^{T} \frac{\eta_t L^2}{8} \leqslant \frac{1}{\eta_T} \ln \frac{1}{w_{i,1}} + L^2 \sum_{t=1}^{T} \frac{\eta_t}{8}.$$

Сформулируем этот результат в виде теоремы.

Теорема 4.8. *Пусть для каждого t потери всех экспертов l_t^i лежат в интервале длины L. Тогда для любых последовательностей потерь таких экспертов l_t^i, $1 \leqslant i \leqslant N$, и параметров обучения η_t, $t = 1, \ldots, T$, имеет место верхняя оценка регрета алгоритма Hedge относительно произвольного эксперта i*

$$R_T^i \leqslant \frac{1}{\eta_T} \ln \frac{1}{w_{i,1}} + \frac{1}{8} L^2 \sum_{t=1}^{T} \eta_t. \tag{4.55}$$

В случае равномерных начальных весов имеет место следующая оценка.

Следствие 4.6. *Пусть выполнены условия теоремы 4.8 начальные веса экспертов равномерные: $w_{i,1} = 1/N$ при $1 \leqslant i \leqslant N$. Тогда потери алгоритма ограничены сверху*

$$H_T \leqslant \min_{1 \leqslant i \leqslant N} L_T^i + \frac{\ln N}{\eta_T} + \frac{1}{8} L^2 \sum_{t=1}^{T} \eta_t.$$

В частности, при $\eta_t = \sqrt{\frac{4\ln N}{L^2 t}}$ *выполнено*

$$H_T \leqslant \min_{1\leqslant i\leqslant N} L_T^i + L\sqrt{T\ln N}. \tag{4.56}$$

Для получения оценки (4.56) ищем оптимальное η_t в виде $\eta_t = \frac{a}{\sqrt{t}}$. Для этого оценим сверху сумму ряда в правой части неравенства (4.55) величиной $2a\sqrt{T}$. После этого находим минимум правой части неравенства (4.55) по a. Он достигается при $a = \sqrt{\frac{4\ln N}{L^2}}$ и равен $L\sqrt{T\ln N}$.

4.6. Алгоритмы экспоненциального взвешивания экспертных прогнозов

4.6.1. Детерминированные прогнозы

Рассмотрим теперь более детальную постановку задачи предсказания с использованием экспертных стратегий. В этой постановке потери экспертов и предсказателя возникают в результате принятых ими решений. Третий игрок – природа – выдает исходы, а функция потерь сопоставляет решения (прогнозы) экспертов и предсказателя с этими исходами и выдает соответствующие потери.

Пусть множество исходов Ω есть произвольное множество объектов любой природы, $\Theta = \{1, \ldots, N\}$ – множество экспертов, Γ – множество решений или прогнозов (предсказаний). Предполагаем, что прогнозы принадлежат некоторому линейному пространству $\mathcal{L}$ (например, $\Gamma \subseteq \mathcal{R}^n$), где R Чмножество всех вещественных чисел). Таким образом, их можно складывать и умножать на вещественные числа.

Оценка принятого решения (или предсказания) $\gamma \in \Gamma$ при исходе $\omega \in \Omega$ производится с помощью функции потерь $\lambda(\omega, \gamma)$ определенной на множестве $\Omega \times \Gamma$.

Подмножество Z линейного пространства $\mathcal{L}$ называется *выпуклым*, если для любых точек $z, z' \in Z$ и любого числа $0 \leqslant p \leqslant 1$ точка $pz + (1-p)z' \in Z$.

Полагаем $H_0 = 0$ и $L_0^i = 0$ для всех i, $1 \leqslant i \leqslant N$.
FOR $t = 1, 2, \ldots$
Эксперты $i \in \Theta$ аннонсируют предсказания $\xi_t^i \in \Gamma$.
Предсказатель принимает свое решение $\gamma_t \in \Gamma$.
Природа анонсирует исход $\omega_t \in \Omega$.
Эксперты $i \in \Theta$ вычисляют свои суммарные потери на шаге t игры

$$L_t^i = L_{t-1}^i + \lambda(\omega_t, \xi_t^i).$$

Предсказатель вычисляет свои суммарные потери на шаге t игры

$$H_t = H_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

Рис. 4.8: Протокол игры с детерминированными предсказаниями

Функция $h(z)$, определенная на выпуклом множестве Z, называется выпуклой, если ее надграфик $\{(x, y) : y \geqslant h(x)\}$ – выпуклое множество. Это эквивалентно тому, что если для любых $z, z' \in Z$ и любого числа $0 \leqslant p \leqslant 1$ выполнено неравенство

$$h(pz + (1-p)z') \leqslant ph(z) + (1-p)h(z'). \tag{4.57}$$

Предполагаем, что множество прогнозов Γ – выпуклое подмножество $\mathcal{L}$, а функция потерь $\lambda(\omega, \gamma)$ является выпуклой по прогнозу γ при каждом $\omega \in \Omega$. Кроме этого, предполагаем, что функция потерь ограниченная и ее значения лежат в некотором интервале длины L.

Рассматривается игра с полной информацией между игроками: *Предсказатель*, *Эксперты* i, где $i \in \Theta$, и *Природа*. Игра происходит в соответствии со протоколом, приведенным на рис. 4.8.

Протокол определяет порядок действий (ходы) игроков. Каждый игрок может при определении своего действия использовать всю информацию, которая известна к началу его хода.

Целью *Предсказателя* является выбор такой последовательности прогнозов $\gamma_1, \gamma_2, \ldots$, чтобы для каждого T его суммарные

потери H_T за первые T шагов были бы с некоторой степенью точности не больше чем суммарные потери L_T^i каждого эксперта i.

Природа может быть "враждебной" по отношению к экспертам и *Предсказателю*: выдаваемые ею исходы ω_t могут зависеть от прогнозов и ξ_t^i и γ_t, так как *Природа* выдает исход ω_t тогда, когда прогнозы уже выданы экспертами и *Предсказателем*.

Количественной оценкой метода прогнозирования является кумулятивный регрет относительно эксперта i и последовательности исходов $\omega_1, \ldots, \omega_T$

$$R_T^i = H_T - L_T^i = \sum_{t=1}^{T} (\lambda(\omega_t, \gamma_t) - \lambda(\omega_t, \xi_t^i)). \tag{4.58}$$

Простейший алгоритм взвешивания экспертных прогнозов вычисляет прогноз *Предсказателя* по формуле

$$\gamma_t = \sum_{i=1}^{N} w_{i,t}^* \xi_t^i, \tag{4.59}$$

где $\xi_t^i \Gamma$ – прогноз i-го эксперта на шаге t,

$$w_{i,t}^* = \frac{w_{i,t}}{\sum\limits_{j=1}^{N} w_{j,t}} \tag{4.60}$$

– нормированные веса, $w_{i,t}$, $i = 1, \ldots, N$, – накопленные веса экспертов на шаге t. Так как Γ – выпуклое множество, $\gamma_t \in \Gamma$ для всех t.

В алгоритме экспоненциального взвешивания с переменным параметром обучения в качестве накопленных весов экспертов на шаге t берут величины

$$w_{i,t} = w_{i,1} e^{-\eta_t L_{t-1}^i}, \tag{4.61}$$

$i = 1, \ldots, N$, где L_{t-1}^i – суммарные потери i-го эксперта на шагах от 1 до $t-1$, η_t – параметр обучения.

Оценки регрета алгоритма оптимального взвешивания экспертных решений могут быть получены в качестве частных случаев оценок теоремы 4.8. Рассматривая протокол, представленный

на рис. 4.8, только в части потерь, получим протокол алгоритма Hedge. Обозначим потери экспертов $l_t^i = \lambda(\omega_t, \xi_t^i)$. Из выпуклости функции $\lambda(\omega, \gamma)$ по второму аргументу получим

$$\lambda(\omega_t, \gamma_t) = \lambda\left(\omega_t, \sum_{i=1}^N w_{i,t}^* \xi_t^i\right) \leqslant \sum_{i=1}^N w_{i,t}^* \lambda(\omega_t, \xi_t^i) = \sum_{i=1}^N w_{i,t}^* l_t^i = h_t$$

для всех t.

Применяя теорему 4.8, получаем следующее утверждение.

Теорема 4.9. *Пусть выполнены условия теоремы 4.8 начальные веса экспертов равномерные: $w_{i,1} = 1/N$, при $1 \leqslant i \leqslant N$, параметры обучения $\eta_t = \sqrt{\frac{4 \ln N}{L^2 t}}$.*

Допустим, что функция потерь $\lambda(\omega, \gamma)$ является выпуклой по второму аргументу и принимает значения в интервале длины L. Тогда для любого T потери алгоритма экспоненциального взвешивания удовлетворяет неравенству

$$H_T \leqslant \min_{1 \leqslant i \leqslant N} L_T^i + L\sqrt{T \ln N}.$$

как бы Природа не выдавала исходы $\omega_1, \ \ldots, \ \omega_T \in \Omega$, а эксперты ни представляли свои прогнозы $\xi_1^i, \ldots, \xi_T^i$, $1 \leqslant i \leqslant N$.

4.6.2. Рандомизированные прогнозы

Заданы множества исходов Ω и множество прогнозов Γ. Имеется N экспертов. Задана некоторая функция потерь $\lambda(\omega, \gamma)$. В этом разделе мы не предполагаем, что функция потерь выпуклая по второму аргументу.

Пример. Приведем пример, который показывает, что в некоторых играх с невыпуклой по прогнозу функцией потерь $\lambda(\omega, \gamma)$ любой метод детерминированных предсказаний имеет недопустимо большую ошибку, которая растет линейно с ростом длины периода предсказания.

Рассмотрим простую игру с двумя экспертами 1 и 2. Пространства исходов и прогнозов совпадают: $\Omega = \Gamma = [1, 2]$. Потери при предсказании γ и исходе ω равны: $\lambda(\omega, \gamma) = 1_{\{\omega \neq \gamma\}}$ – характеристическая функция множества $\{(\omega, \gamma) : \gamma \neq \omega\}$.

Легко проверить, что эта функция потерь не является выпуклой по прогнозу. Заметим, что для любой последовательности прогнозов *Предсказателя* $\gamma_1, \gamma_2, \ldots$ существует такая последовательность исходов $\omega_1, \omega_2, \ldots$, что потери *Предсказателя* максимальны, т.е. $H_T = T$ для всех T. Действительно, *Природа* может определить для всех $t = 1, 2, \ldots$:

$$\omega_t = \begin{cases} 2, \text{ если } \gamma_t = 1, \\ 1 \text{ в противном случае.} \end{cases}$$

Рассмотрим двух экспертов, один из которых – эксперт 1, всегда предсказывает $\xi_t^1 = 1$, а другой – эксперт 2, всегда предсказывает $\xi_t^2 = 2$, $t = 1, 2, \ldots$. Пусть L_t^i – потери i-го эксперта, $i = 1, 2$.

Заметим, что *Предсказатель* при прогнозе $\gamma_t = 1$ просто следует решению эксперта 1, а при прогнозе $\gamma_t = 2$ – следует решению эксперта 2.

Легко видеть, что, так как для последовательности исходов $\omega_1, \omega_2, \ldots, \omega_T$ число единиц или число двоек будут не больше чем $T/2$, у одного из экспертов потери будут не более чем $T/2$. Поэтому $\min_{i=1,2} L_t^i \leqslant T/2$ для всех T.

Таким образом, для любой последовательности прогнозов *Предсказателя* "адаптивно враждебная" *Природа* может предоставить последовательность $\omega_1, \omega_2, \ldots$ такую, что

$$H_T - \min_{i=1,2} L_T^i \geqslant T/2$$

для всех T.

Приведенный пример показывает, что для некоторых невыпуклых функциях потерь *Природа* может выдавать последовательность исходов так, что при любых детерминированных действиях *Предсказателя* его регрет $\geqslant T/2$ за любой период игры T.

Данную проблему *Предсказатель* может преодолеть с помощью рандомизации прогнозов. Точнее прогнозами будут смешанные стратегии – распределения вероятностей на множестве всех детерминированных прогнозов. Вместо функции потерь будет рассматриваться ее математическое ожидание, к которому мы применим результаты раздела 4.2.

Пусть $H_0 = 0$, $L_0^i = 0$, $i = 1, \ \ldots, \ N$.
FOR $t = 1, 2, \ \ldots$
Эксперт i анонсирует прогноз: $\xi_t^i \in \Gamma$, $i = 1, \ \ldots, \ N$.
Предсказатель анонсирует распределение вероятностей $w_{1,t}^*, \ \ldots, \ w_{N,t}^*$ на множестве экспертов $\{1, \ \ldots, \ N\}$, где $\sum\limits_{i=1}^{N} w_{i,t}^* = 1$.
Природа анонсирует исход: $\omega_t \in \Omega$.
Генератор случайных чисел анонсирует эксперта $i_t \in \{1, \ \ldots, \ N\}$ с вероятностью $w_{i_t,t}^*$.
Эксперт i вычисляет свои суммарные потери на шаге t игры

$$L_t^i = L_{t-1}^i + \lambda(\omega_t, \xi_t^i),$$

где $i = 1, \ \ldots, \ N$.
Предсказатель вычисляет свои суммарные потери на шаге t игры

$$H_t = H_{t-1} + \lambda(\omega_t, \xi_t^{i_t}).$$

ENDFOR

Рис. 4.9: Протокол игры с рандомизированными предсказаниями

Пусть теперь на каждом шаге t игры *Предсказатель* выдает прогноз в виде смешанной стратегии – распределения вероятностей на множестве экспертов $\{1, \ \ldots, \ N\}$.

Протокол рандомизированной игры представлен на рис. 4.9.

Мы вводим в протокол еще одного игрока – *Генератора случайных чисел*, который будет генерировать номер эксперта из множества $\{1, \ \ldots, \ N\}$ согласно заданному *Предсказателем* распределению вероятностей. *Природа* может использовать это распределение вероятностей, но не может использовать номер эксперта, выбранного согласно этому распределению. Поэтому мы помещаем в протоколе ход *Генератора случайных чисел* после хода *Природы*.

Генератор случайных чисел относительно произвольного распределения вероятностей можно реализовать на основе генератора равномерно распределенных случайных чисел следующим об-

разом. Вводим случайные переменные I_t, так что $I_t = i$ тогда и только тогда, когда

$$U_t \in \left[\sum_{j=1}^{i-1} w_{j,t}^*, \sum_{j=1}^{i} w_{j,t}^*\right],$$

где величины $U_1, U_2, \ldots$ – независимые равномерно распределенные в единичном отрезке случайные величины. Из определения следует, что $P\{I_t = i\} = w_{i,t}^*$ для всех t.

В такой игре потери *Предсказателя* $\lambda(\omega_t, \xi_t^{I_t})$ являются случайной величиной. В этом случае качество предсказания *Предсказателя* измеряется также случайной величиной – случайным регретом

$$H_T - \min_{i=1,\ldots,N} L_T^i = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^{I_t}) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \qquad (4.62)$$

Целью *Предсказателя* является минимизация математического ожидания регрета [4]

$$E(H_T - \min_{i=1,\ldots,N} L_T^i) =$$

$$= \sum_{t=1}^{T} E(\lambda(\omega_t, \xi_t^{I_t})) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) =$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{N} \lambda(\omega_t, \xi_t^i) p_{i,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \qquad (4.63)$$

Будем вычислять распределение вероятностей на множестве экспертов с помощью соотношений (4.60) из раздела 4.6.1.

Переформулируем теорему 4.9 для рандомизированной интерпретации.

Теорема 4.10. *Допустим, что функция потерь $\lambda(\omega, \gamma)$ принимает значения в интервале длины L.*[5]

[4] Мы также учитываем случай, когда предсказания экспертов на шаге t зависят от значений $I_1, \ldots, I_{t-1}$. В этом случае, далее, на шаге t имеется ввиду условное математическое ожидание $E_{I_t \sim w_t^*}(H_T - \min_{i=1,\ldots,N} L_T^i | I_1, \ldots, I_{t-1})$, где $w_t^* = (w_{1,t}^*, \ldots, w_{N,t}^*)$ – текущие нормализованные веса экспертов.

[5] Выпуклость по второму аргументу не предполагается.

Пусть начальное распределение на множестве экспертов – равномерное, а распределение вероятностей генератора случайных чисел определено согласно формулам (4.61) и (4.60), H_T – случайные кумулятивные потери алгоритма, представленного на рис. 4.9.

Тогда для любых последовательностей исходов и прогнозов экспертов имеет место оценка математического ожидания регрета вероятностного алгоритма экспоненциального взвешивания

$$E(H_T - \min_i L_T^i) \leqslant L\sqrt{T \ln N}. \tag{4.64}$$

Далее, аналогично тому как было получено утверждение (4.47), следствия 4.2 мы докажем следующие два следствия.

Следствие 4.7. *Пусть* $0 < \delta < 1$. *Тогда для любой последовательности исходов* $\omega_1, \ldots, \omega_T$ *и любой последовательности прогнозов экспертов* $\xi_1^i, \ldots, \xi_T^i$, $1 \leqslant i \leqslant N$, *с вероятностью* $1 - \delta$ *выполнено неравенство*

$$\sum_{t=1}^{T} \lambda(\omega_t, \xi_t^{I_t}) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) \leqslant$$
$$\leqslant L\sqrt{T \ln N} + \sqrt{\frac{1}{2} T \ln \frac{1}{\delta}}.$$

Доказательство. Для доказательства этого утверждения мы используем вариант неравенства Хефдинга–Адзумы (10.7). Легко видеть, что случайные величины

$$X_t = \lambda(\omega_t, \xi_t^{I_t}) - E(\lambda(\omega_t, \xi_t^{I_t}) | I_1, \ldots, I_{t-1})$$

образуют мартингал-разность относительно последовательности случайных величин $I_1, I_2, \ldots$. Поэтому для их сумм $S_T = \sum\limits_{t=1}^{T} X_t$ выполнено неравенство

$$P\{S_T > c\} \leqslant e^{-\frac{2c^2}{T}},$$

где c – произвольное положительное число. Отсюда для произвольного $\delta > 0$ выполнено неравенство

$$P\left\{S_T > \sqrt{\frac{1}{2}T\ln\frac{1}{\delta}}\right\} \leqslant \delta.$$

Утверждение следствия теперь прямо следует из этого неравенства и неравенств (4.63) и (4.64). $\triangle$

Пусть в игре на предсказания с использованием экспертов $i = 1, \ldots, N$ некоторый предсказатель выдает рандомизированные прогнозы $\xi_1, \xi_2, \ldots$, а i-й эксперт выдает прогнозы $\xi_1^i, \xi_2^i, \ldots$, где $i = 1, \ldots, N$.

Предсказатель называется *состоятельным по Ханнану*, если для любой последовательности исходов и для любой последовательности прогнозов экспертов почти всюду выполнено

$$\limsup_{T\to\infty} \frac{1}{T}\left(\sum_{t=1}^{T}\lambda(\omega_t, \xi_t) - \min_{i=1,\ldots,N}\sum_{t=1}^{T}\lambda(\omega_t, \xi_t^i)\right) \leqslant 0. \qquad (4.65)$$

Следующее следствие доказывается точно также как утверждение (4.47) следствия 4.2.

Следствие 4.8. *Вероятностный алгоритм экспоненциального взвешивания с переменным параметром обучения является состоятельным по Ханнану.*

Поясним, как соотносится пример, приведенный в начале этого раздела, со следствием 4.8. В примере *Предсказатель* при прогнозе $\gamma_t = 1$ на шаге t просто следует решению эксперта $i_t = 1$, а при прогнозе $\gamma_t = 2$ - следует решению эксперта $i_t = 2$. Получаем бесконечную траекторию выбираемых экспертов: $i_1, i_2, \ldots$ При рандомизированном выборе экспертов вероятность выбрать эту траекторию, а также любую другую, на которой нарушается условие (4.65), равна 0.

Сравнение с теоремой 4.2 показывает, что рандомизированный алгоритм, примененный к простой функции потерь, имеет примерно (в зависимости от выбора параметра $\beta = e^{-\eta}$) в два раза меньшую оценку ошибки, чем детерминированный алгоритм взвешенного большинства WMA.

4.7. Алгоритм отслеживания наилучшей комбинации экспертов Fixed-Share

В предыдущих разделах мы сравнивали потери агрегирующего алгоритма с потерями наилучшего эксперта. Другой возможный подход, предложенный Хербстером и Вармутом в [27], состоит в следующем: серия шагов, на которых производится обучение, делится на сегменты. Каждому сегменту ставится в соответствие свой эксперт, последовательность сегментов и соответствующих экспертов называется *составным экспертом*. Цель агрегирующего алгоритма также несколько меняется – теперь он должен предсказывать так, чтобы быть не хуже каждого составного эксперта.

Данное изменение позволяет точнее моделировать условия реальной жизни, когда состояние природы может меняться со временем и разные эксперты могут предсказывать с разной степенью успешности в зависимости от текущего тренда. Данный подход называется “trackng the best expert”.

В этом разделе рассмотрим регрет более общего вида, а именно, будем сравнивать потери алгоритма экспоненциального смешивания с потерями наилучшей комбинации экспертов. В разделе 4.7.1 мы приведем алгоритмы Fxed-Share и CompHedge, в разделе 4.7.2 мы докажем их эквивалентность, а в разделе 4.7.3 мы получим верхнюю оценку регрета алгоритма Fixed-Share с помощью аналогичной оценки регрета алгоритма CompHedge.

4.7.1. Составные эксперты и алгоритмы для них

Будем рассматривать экспертов двух видов. Экспертов в старом смысле называем *элементарными*. Имеется N элементарных экспертов. На шаге t элементарный эксперт i несет потери l_t^i, где $i \in \{1, \ldots, N\}$. За s шагов кумулятивные потери элементарного эксперта i равны $L_s^i = \sum\limits_{t=1}^{s} l_t^i$.

Фиксируем горизонт прогнозирования T. *Составным экспертом* называется произвольная последовательность $(i_1, i_2, \ldots, i_T)$ элементарных экспертов, где $i_j \in \{1, \ldots, N\}$ при $j = 1, \ldots, T$.

Рассматриваем алгоритм $Hedge(\eta)$ для составных экспертов

с постоянным параметром обучения η. Обозначим этот алгоритм $CompHedge(\eta)$. В процессе его работы вычисляются веса составных экспертов.

Протокол игры для алгоритма $CompHedge(\eta)$ приведен на рис. 4.10.

На шаге t составной эксперт $(i_1, \ldots, i_t, \ldots, i_T)$ несет потери $l_t^{i_t}$, где $1 \leqslant t \leqslant T$. За s шагов, $1 \leqslant s \leqslant T$, кумулятивные потери элементарного эксперта $(i_1, \ldots, i_s, \ldots, i_T)$ равны $L_s(i_1, \ldots, i_s, \ldots, i_T) = \sum\limits_{t=1}^{s} l_t^{i_t}$. Заметим, что $L_s(i_1, \ldots, i_T) = L_s(i_1, \ldots, i_s)$. Согласно (4.66) также верно

$$\begin{aligned}\tilde{w}_{t+1}(i_1, \ldots, i_t, \ldots, i_T) = \tilde{w}_t(i_1, \ldots, i_t, \ldots, i_T) e^{-\eta l_t^{i_t}} = \\ = \tilde{w}_1(i_1, \ldots, i_t, \ldots, i_T) e^{-\eta \sum\limits_{s=1}^{t} l_s^{i_s}} = \\ = \tilde{w}_1(i_1, \ldots, i_t, \ldots, i_T) e^{-\eta L_t(i_1, \ldots, i_t, \ldots, i_T)}.\end{aligned}$$

Аналогичным образом определяется потенциал

$$\begin{aligned}\tilde{W}_T = \sum_{i_1, \ldots, i_T} \tilde{w}_t(i_1, \ldots, i_T) = \\ = \sum_{i_1, \ldots, i_T} \tilde{w}_1(i_1, \ldots, i_T) e^{-\eta L_{T-1}(i_1, \ldots, i_T)}. \qquad (4.67)\end{aligned}$$

Смешанные потери алгоритма для составных экспертов равны

$$\begin{aligned}\tilde{m}_t = -\frac{1}{\eta} \ln \frac{1}{\tilde{W}_t} \sum_{i_1, \ldots, i_t, \ldots, i_T} \tilde{w}_t(i_1, \ldots, i_t, \ldots, i_T) e^{-\eta l_t^{i_t}} = \\ = -\frac{1}{\eta} \ln \frac{1}{\tilde{W}_t} \sum_{i_1, \ldots, i_t, \ldots, i_T} \tilde{w}_{t+1}(i_1, \ldots, i_t, \ldots, i_T) = -\frac{1}{\eta} \ln \frac{\tilde{W}_{t+1}}{\tilde{W}_t}.\end{aligned}$$

Кумулятивные смешанные потери составных экспертов за T ша-

Алгоритм $CompHedge(\eta)$
Пусть $\tilde{w}_1(i_1, i_2, \ldots, i_T)$ – начальный вес эксперта $i_1, i_2, \ldots, i_T$. Предполагаем, что эти веса нормированы. Точное выражение для начального веса составного эксперта будет приведено ниже (см. формулу (4.68)).

FOR $t = 1, 2, \ldots, T$
Используя веса составных экспертов, накопленные на предыдущих шагах, вычисляем прогноз распределения потерь среди экспертов

$$\tilde{w}_t^*(i_1, \ldots, i_T) = \frac{\tilde{w}_t(i_1, \ldots, i_T)}{\sum\limits_{i_1, \ldots, i_T} \tilde{w}_t(i_1, \ldots, i_T)} \text{ для всех } (i_1, \ldots, i_T).$$

Получаем потери экспертов l_t^i при $i = 1, \ldots, N$.
Вычисляем потери алгоритма $CompHedge(\eta)$ на шаге t

$$\tilde{h}_t = \sum_{i_1, \ldots, i_t, \ldots, i_T} \tilde{w}_t^*(i_1, \ldots, i_t, \ldots, i_T) l_t^{i_t}.$$

Переопределяем веса всех составных экспертов для использования на следующем шаге

$$\tilde{w}_{t+1}(i_1, \ldots, i_t, \ldots, i_T) = \tilde{w}_t(i_1, \ldots, i_t, \ldots, i_T) e^{-\eta l_t^{i_t}}. \qquad (4.66)$$

ENDFOR

Рис. 4.10: Алгоритм $CompHedge(\eta)$

гов равны

$$\tilde{M}_T = \sum_{t=1}^{T} \tilde{m}_t = -\frac{1}{\eta} \sum_{t=1}^{T} \ln \frac{\tilde{W}_{t+1}}{\tilde{W}_t} = -\frac{1}{\eta} \ln \tilde{W}_{t+1} =$$

$$= -\frac{1}{\eta} \ln \sum_{i_1,\ldots,i_T} \tilde{w}_{t+1}(i_1,\ldots,i_T) =$$

$$= -\frac{1}{\eta} \ln \sum_{i_1,\ldots,i_T} \tilde{w}_1(i_1,\ldots,i_T) e^{-\eta L_T(i_1,\ldots,i_T)}.$$

Алгоритм $CompHedge(\eta)$ для составных экспертов физически нереализуем, так как число этих экспертов экспоненциально растет с ростом T. Поэтому определяют эквивалентный ему по результатам физически реализуемый алгоритм, который работает с кумулятивными весами.

Рассмотрим кумулятивные (маргинальные) веса $\tilde{w}_t(i)$ элементарных экспертов при $1 \leqslant t \leqslant T$ и $i = 1, \ldots, N$. Определим

$$\tilde{w}_t(i) = \sum_{i_1,\ldots,i_{t-1},i_t,\ldots,i_T} \tilde{w}_t(i_1,\ldots,i_{t-1},i,i_t,\ldots,i_T),$$

$$\tilde{w}_t^*(i) = \frac{\tilde{w}_t(i)}{\sum\limits_{j=1}^{N} \tilde{w}_t(j)}$$

при $1 \leqslant t \leqslant T$, а также

$$\tilde{w}_t(i_1,\ldots,i_t) = \sum_{i_{t+1},\ldots,i_T} \tilde{w}_t(i_1,\ldots,i_t,i_{t+1},\ldots,i_T).$$

С учетом (4.68) ниже, легко видеть, что $\tilde{w}_1(i) = 1/N$ для всех $1 \leqslant i \leqslant N$.

Лемма 4.9. *Для маргинальных весов составных экспертов выполнены равенства*

1. $\tilde{W}_t = \sum\limits_{i=1}^{N} \tilde{w}_t(i)$

2. $\tilde{m}_t = -\frac{1}{\eta} \ln \sum\limits_{i=1}^{N} \tilde{w}_t^*(i) e^{-\eta l_t^i}$

3. $\tilde{h}_t = \sum\limits_{i=1}^{N} \tilde{w}_t^*(i) l_t^i$

Доказательство. Прямой проверкой получаем

$$\tilde{W}_t = \sum_{i_1,\ldots,i_T} \tilde{w}_t(i_1,\ldots,i_T) =$$

$$= \sum_{i=1}^{N} \sum_{i_1,\ldots,i_{t-1},i_{t+1},\ldots,i_T} \tilde{w}_t(i_1,\ldots,i,\ldots,i_T) = \sum_{i=1}^{N} \tilde{w}_t(i).$$

Смешанные потери (mixloss) алгоритма для составных экспертов равны

$$\tilde{m}_t = -\frac{1}{\eta} \ln \frac{1}{\tilde{W}_t} \sum_{i_1,\ldots,i_t,\ldots,i_T} \tilde{w}_t(i_1,\ldots,i_t,\ldots,i_T) e^{-\eta l_t^{i_t}} =$$

$$= -\frac{1}{\eta} \ln \frac{1}{\tilde{W}_t} \sum_{i=1}^{N} \sum_{i_1,\ldots,i,\ldots,i_T} \tilde{w}_t(i_1,\ldots,i,\ldots,i_T) e^{-\eta l_t^i} =$$

$$= -\frac{1}{\eta} \ln \frac{1}{\tilde{W}_t} \sum_{i=1}^{N} \tilde{w}_t(i) e^{-\eta l_t^i}.$$

Выражение для $\tilde{h}_t$ получается аналогичным образом. $\triangle$

Определим теперь начальные веса составных экспертов для алгоритма $CompHedge(\eta)$. Они будут неравномерными. Фиксируем параметры $1 > \alpha_1 \geqslant \alpha_2 \geqslant \ldots \geqslant \alpha_T \geqslant 0$. Определим

$$\tilde{w}_1(i_1,\ldots,i_T) = \frac{1}{N} \prod_{i_{t-1} \neq i_t}^{T} \left(\frac{\alpha_t}{N-1} \right) \prod_{i_{t-1}=i_t}^{T} (1-\alpha_t). \qquad (4.68)$$

В работе [43] рассматривается постоянный параметр $\alpha_t = \alpha$ для всех t. Пусть

$$s(i_1,\ldots,i_T) = |\{s : i_{s-1} \neq i_s,\ 2 \leqslant s \leqslant T\}| \qquad (4.69)$$

– число перемен элементарных экспертов в составном эксперте. Тогда определение (4.68) принимает вид

$$\tilde{w}_1(i_1,\ldots,i_T) = \frac{1}{N}\left(\frac{\alpha}{N-1}\right)^{s(i_1,\ldots,i_T)}(1-\alpha)^{T-s(i_1,\ldots,i_T)-1}.$$

Можно заметить, что из определения следует, что $\tilde{w}_1(i_1) = \frac{1}{N}$ и

$$\tilde{w}_1(i_1,\ldots,i_{t+1}) = \tilde{w}_1(i_1,\ldots,i_t)\left(\frac{\alpha_t}{N-1}1_{i_{t+1}\neq i_t} + (1-\alpha_t)1_{i_{t+1}=i_t}\right).$$

Условные вероятности $\tilde{w}_1(i_1,\ldots,i_{t+1})/\tilde{w}_1(i_1,\ldots,i_t)$ задают марковский случайный процесс, порождающий последовательность экспертов $i_1, i_2, \ldots$, при котором $i_{t+1} = i_t$ с вероятностью $1-\alpha_t$, и с вероятностью α_t в качестве i_{t+1} равновероятно выбирается один из оставшихся $N-1$ экспертов.

Алгоритм Fixed-Share с постоянным параметром обучения был предложен в работе [27]. Алгоритм Fixed-Share оперирует с "макровесами" $w_{i,t}$, где $i = 1,\ldots,N$. Протокол игры для алгоритма $FS(\eta)$ приведен на рис. 4.11.

Замечание. Существуют другие варианты алгоритма Fixed-Share. Нетрудно заметить, что можно сформулировать финальную часть алгоритма Fixed-Share (рис. 4.11) в следующем виде:

Loss Update: $w^m_{i,t} = \frac{w_{i,t}e^{-\eta l^i_t}}{\sum\limits_{j=1}^{N} w_{j,t}e^{-\eta l^j_t}}$

Fixed-Share Update: $w_{i,t+1} = \frac{\alpha_t}{N-1} + \left(1 - \frac{N}{N-1}\alpha_t\right) w^m_{i,t}$.

В другом варианте можно часть Fixed-Share Update определить

$$w_{i,t+1} = (1-\alpha_t)w^m_{i,t} + \frac{\alpha_t}{N}\sum_{j=1}^{N} w^m_{j,t} = \frac{\alpha_t}{N} + (1-\alpha_t)w^m_{i,t}.$$

Здесь отобранная часть веса эксперта i распределяется между всеми N экспертами поровну. Соответствующие рассуждения будут аналогичны, приводимым далее (см. задачи из раздела 4.10).

Следующая лемма показывает, что операция Fixed-Share Update не изменяет величину потенциала.

Алгоритм $FS(\eta)$

Полагаем $w_{i,1} = \frac{1}{N}$ при $i = 1, \ldots, N$.

FOR $t = 1, \ldots, T$

Вычисляем прогноз распределения потерь среди экспертов

$$w^*_{i,t} = \frac{w_{i,t}}{\sum\limits_{i=1}^{N} w_{i,t}}, 1 \leqslant i \leqslant N.$$

Получаем потери экспертов l^i_t при $1 \leqslant i \leqslant N$.

Вычисляем потери алгоритма $FS(\eta) : h_t = \sum\limits_{i=1}^{N} w^*_{i,t} l^i_t$.

Вычисляем кумулятивные потери всех экспертов $1 \leqslant i \leqslant N$ и алгоритма.

Переопределяем веса всех экспертов $1 \leqslant i \leqslant N$ в два этапа:

Loss Update

$w^m_{i,t} = w_{i,t} e^{-\eta l^i_t}$

Fixed-Share Update

$w_{i,t+1} = (1 - \alpha_t) w^m_{i,t} + \frac{\alpha_t}{N-1} \sum\limits_{j \neq i} w^m_{j,t}$

Эквивалентное определение для второго этапа:

$Pool = \sum\limits_{i=1}^{N} \alpha_t w^m_{i,t}$.

$w_{i,t+1} = (1 - \alpha_t) w^m_{i,t} + \frac{1}{N-1}(Pool - \alpha_t w^m_{j,t})$.

ENDFOR

Рис. 4.11: Алгоритм Fixed-Share

Лемма 4.10. $\sum\limits_{i=1}^{N} w_{i,t+1} = \sum\limits_{i=1}^{N} w_{i,t}^{m}$.

Доказательство. Сумма не изменяется, так как на шаге t от каждого веса забирается доля α_t, а все забранные доли возвращаются другим весам. $\triangle$

Из леммы 4.10 получаем

$$W_{t+1} = \sum_{i=1}^{N} w_{i,t+1} = \sum_{i=1}^{N} w_{i,t}^{m} = \sum_{i=1}^{N} w_{i,t} e^{-\eta l_t^i}.$$

Определим смешанные потери алгоритма $FS(\eta)$

$$m_t = -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,t}^{*} e^{-\eta l_t^i}.$$

Из предыдущих определений получаем выражения для смешанных потерь

$$m_t = -\frac{1}{\eta} \ln \frac{1}{W_t} \sum_{i=1}^{N} w_{i,t} e^{-\eta l_t^i} =$$

$$= -\frac{1}{\eta} \ln \frac{1}{W_t} \sum_{i=1}^{N} w_{i,t}^{m} = -\frac{1}{\eta} \ln \frac{W_{t+1}}{W_t}.$$

Также

$$M_T = \sum_{1=1}^{T} m_t = \sum_{1=1}^{T} -\frac{1}{\eta} \ln \frac{W_{t+1}}{W_t} = -\frac{1}{\eta} \ln W_{T+1}. \tag{4.70}$$

4.7.2. Теорема об эквивалентности двух алгоритмов

Эквивалентность алгоритмов $CompHedge(\eta)$ и $FS(\eta)$ впервые отмечена в работе [59], полное доказательство приведено в [43].

Мы докажем, что при фиксированном горизонте прогнозирования T маргинальные веса $\tilde{w}_t(i)$ алгоритма $CompHedge(\eta)$ для составных экспертов, в котором начальное распределение весов

определено по формуле (4.68), и веса $w_{i,t}$ алгоритма $FS(\eta)$ равны при $1 \leqslant t \leqslant T$. Отсюда будет следовать равенство соответствующих величин $\tilde{W}_t$, $\tilde{m}_t$, $\tilde{M}_t$ и W_t, m_t, M_t соответственно.

Теорема 4.11. *Пусть оба алгоритма $CompHedge(\eta)$ и $FS(\eta)$ используют один и тот же параметр обучения η. Тогда $\tilde{w}_t(i) = w_{i.t}$ для всех $1 \leqslant i \leqslant N$ и $1 \leqslant t \leqslant T$.*

Доказательство. Доказательство математической индукцией по t. Из определения $\tilde{w}_1(i) = w_{i,1}$ при $1 \leqslant i \leqslant N$. Предположим, что $\tilde{w}_t(i) = w_{i,t}$ при $1 \leqslant i \leqslant N$. Докажем, что $\tilde{w}_{t+1}(i) = w_{i,t+1}$ для всех i. Из определения имеет место следующая цепочка равенств:

$$\tilde{w}_{t+1}(i) = \sum_{i_1,\ldots,i_t,i_{t+2},\ldots,i_T} \tilde{w}_{t+1}(i_1,\ldots,i_t,i,i_{t+2},\ldots,i_T) =$$

$$= \sum_{i_1,\ldots,i_t} e^{-\eta \sum\limits_{s=1}^{t} l_s^{i_s}} \tilde{w}_1(i_1,\ldots,i_t,i) =$$

$$= \sum_{i_1,\ldots,i_t} e^{-\eta \sum\limits_{s=1}^{t} l_s^{i_s}} \tilde{w}_1(i_1,\ldots,i_t) \left(\frac{\alpha_t}{N-1} 1_{i_t \neq i} + (1-\alpha_t) 1_{i_t = i} \right) = \quad (4.71)$$

$$= \sum_{i_t} e^{-\eta l_t^{i_t}} \tilde{w}_t(i_t) \left(\frac{\alpha_t}{N-1} 1_{i_t \neq i} + (1-\alpha_t) 1_{i_t = i} \right) = \quad (4.72)$$

$$= \sum_{i_t} e^{-\eta l_t^{i_t}} w_{i_t,t} \left(\frac{\alpha_t}{N-1} 1_{i_t \neq i} + (1-\alpha_t) 1_{i_t = i} \right) = \quad (4.73)$$

$$= \sum_{i_t} w_{i_t,t}^m \left(\frac{\alpha_t}{N-1} 1_{i_t \neq i} + (1-\alpha_t) 1_{i_t = i} \right) = \quad (4.74)$$

$$= \frac{\alpha_t}{N-1} \sum_{i_t \neq i} w_{i_t,t}^m + (1-\alpha_t) w_{i,t}^m =$$

$$= w_{i,t+1}.$$

Здесь при переходе от (4.71) к (4.72) было использовано равенство

$$\tilde{w}_t(i_t) = \sum_{i_1,\ldots,i_{t-1}} \tilde{w}_t(i_1,\ldots,i_t) = \sum_{i_1,\ldots,i_{t-1}} \tilde{w}_1(i_1,\ldots,i_t) e^{-\eta \sum\limits_{s=1}^{t-1} l_s^{i_s}},$$

при переходе от (4.72) к (4.73) было использовано предположение индукции $\tilde{w}_t(i_t) = w_{i_t,t}$, а при переходе от (4.73) к (4.74) было использовано определение $w^m_{i_t,t}$. Таким образом, предположение индукции выполнено и на шаге $t+1$. $\triangle$

Поскольку все величины $\tilde{W}_t$, $\tilde{m}_t$, $\tilde{M}_t$ определены через $\tilde{w}_t(i)$, а величины W_t, m_t, M_t определены через веса $w_{i,t}$, получаем следствие.

Следствие 4.9. *Пусть оба алгоритма CompHedge и FS используют один и тот же параметр обучения η. Тогда $\tilde{W}_t = W_t$, $\tilde{m}_t = m_t$, $\tilde{M}_t = M_t$ и $\tilde{H}_T = H_T$.*

Здесь $\tilde{H}_T = \sum\limits_{t=1}^{T} \tilde{h}_t$ – кумулятивные потери алгоритма $CompHedge(\eta)$ и $H_T = \sum\limits_{t=1}^{T} h_t$ – кумулятивные потери алгоритма $FS(\eta)$.

4.7.3. Регрет алгоритма Fixed-Share

Оценим величину $M_T = -\frac{1}{\eta}\ln W_{T+1}$ через потери произвольного составного эксперта. Для это воспользуемся эквивалентность алгоритмов $FS(\eta)$ и $CompHedge(\eta)$, точнее, равенством их основных характеристик: $\tilde{W}_T = W_T$ и $\tilde{M}_t = M_t$.

По следствию 4.9 величина $W_{T+1} = \tilde{W}_{T+1}$. Поэтому согласно равенствам (4.70) и (4.67), для любого составного эксперта $i_1, \ldots, i_T$ будет

$$M_T \leqslant L_T(i_1, i_2, \ldots, i_T) - \frac{1}{\eta}\ln \tilde{w}_1(i_1, i_2, \ldots, i_T)$$

Учитывая выражение (4.68) для начального веса эксперта $i_1, \ldots, i_T$ в алгоритме $CompHedge(\eta)$ получим следующую оценку для регрета кумулятивных смешанных потерь алгоритма $FS(\eta)$.

Теорема 4.12. *Для произвольного составного эксперта $i_1, \ldots, i_T$*

$$M_T - L_T(i_1, i_2, \ldots, i_T) \leqslant$$
$$\leqslant -\frac{1}{\eta}\ln\left(\frac{1}{N}\prod_{i_{t-1}\neq i_t}^{T}\left(\frac{\alpha_t}{N-1}\right)\prod_{i_{t-1}=i_t}^{T}(1-\alpha_t)\right). \qquad (4.75)$$

В частности, при постоянном параметре $\alpha_t = \alpha$ неравенство (4.75) превращается в

$$M_T - L_T(i_1, i_2, \ldots, i_T) \leqslant$$
$$\leqslant -\frac{1}{\eta} \ln \left(\frac{1}{N} \left(\frac{\alpha}{N-1} \right)^{s(i_1,\ldots,i_T)} (1-\alpha)^{T-s(i_1,\ldots,i_T)-1} \right), \quad (4.76)$$

где величина $s(i_1, \ldots, i_T)$ определена по (4.69).

Зададим параметр $\alpha_t = \frac{1}{t+1}$ для всех t. Тогда правая часть неравенства (4.76) оценивается сверху величиной

$$\frac{1}{\eta} \left(\ln N + \sum_{i_{t-1} \neq i_t}^{T} (\ln(t+1) + \ln(N-1)) - \sum_{i_{t-1}=i_t}^{T} \ln \left(1 - \frac{1}{t+1} \right) \right) \leqslant$$
$$\leqslant \frac{1}{\eta} \left(\ln N + k \ln(T+1) + k \ln(N-1) + 2 \sum_{t=1}^{T} \frac{1}{t+1} \right) \leqslant$$
$$\leqslant \frac{1}{\eta} \left((k+2) \ln(T+1) + (k+1) \ln N) + 2 \right).$$

Согласно лемме 4.9 кумулятивные потери алгоритма $FS(\eta)$ равны кумулятивным потерям алгоритма $CompHedge(\eta)$: $H_T = \tilde{H}_T$. Для оценки потерь алгоритма $CompHedge(\eta)$ применим неравенство Хефдинга (см. неравенство (4.52) следствия 4.3 и теорему 4.7) к составным экспертам. В результате получаем оценку регрета алгоритма $FS(\eta)$ в случае ограниченных одношаговых потерь.

Теорема 4.13. *Пусть на каждом шаге t значения l_t^i потерь экспертов лежат в интервале длины L вещественной прямой. Тогда для любого m*

$$H_T - \min_{s(i_1,\ldots,i_T) \leqslant m} L_T(i_1, \ldots, i_T) \leqslant$$
$$\leqslant \frac{1}{\eta} \left((m+2) \ln(T+1) + (m+1) \ln N) + 2 \right) + \frac{\eta L^2 T}{8} \quad (4.77)$$

для произвольного T.

При заданном горизонте прогнозирования T и параметре m оптимизируем правую часть этой оценки по η и получаем

$$\eta = \sqrt{\frac{8\left((m+2)\ln(T+1) + (m+1)\ln N) + 2\right)}{L^2 T}}.$$

Неравенство (4.77) при таком значении η принимает вид

$$H_T - \min_{s(i_1,\ldots,i_T)\leqslant m} L_T(i_1,\ldots,i_T) \leqslant$$
$$\leqslant L\sqrt{\frac{1}{2}T\left((m+2)\ln(T+1) + (m+1)\ln N) + 2\right)}.$$

4.8. Схемы смешивания апостериорных распределений экспертов

В этом разделе рассмотрим методы смешивания апостериорных распределений экспертов с использованием схем смешивания общего вида (mixing past posteriors – MPP), предложенный в работе [14].

Предварительно изучим свойство относительной энтропии (расхождения Кульбака–Лейблера)

$$D(p\|q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i},$$

где n – произвольное натуральное число, $p = (p_1,\ldots,p_n)$, $q = (q_1,\ldots,q_n)$ – элементы n-мерного симплекса Δ_n, т.е. распределения вероятностей на конечном множестве из n элементов. Предполагаем, что $0\ln 0 = 0$.

Неравенства $p > q$, $p \geqslant q$, $p \geqslant 0$ на векторах будем понимать покомпонентно; здесь 0 – вектор с нулевыми координатами. В дальнейшем будут использоваться следующие свойства относительной энтропии.

Лемма 4.11. *Для любых $p, q, w \in \Delta_n$ таких, что $q, w > 0$ будет*

$$D(p\|q) \leqslant D(p\|w) + \ln\left(\sum_{i=1}^{n} p_i \frac{w_i}{q_i}\right).$$

Если $q \geqslant \beta w$ для некоторого числа $\beta > 0$, то

$$D(p\|q) \leqslant D(p\|w) + \ln\frac{1}{\beta}.$$

В частности , при $p = w$ будет $D(w\|q) \leqslant \ln\frac{1}{\beta}$ при $q \geqslant \beta w$.

Доказательство. Из вогнутости логарифма будет

$$D(p\|q) - D(p\|w) = \sum_{i=1}^{n} p_i \ln\frac{w_i}{q_i} \leqslant \tag{4.78}$$
$$\leqslant \ln\left(\sum_{i=1}^{n} p_i \frac{w_i}{q_i}\right).$$

Если к тому же $q \geqslant \beta w$, будет $\sum\limits_{i=1}^{n} p_i \frac{w_i}{q_i} \leqslant \sum\limits_{i=1}^{n} p_i \frac{w_i}{\beta w_i} = \frac{1}{\beta}$. $\triangle$

Из леммы получаем важное для дальнейшего следствие.

Следствие 4.10. *Пусть $p \in \Delta_n$ и $q = \sum\limits_{i=1}^{t} \beta_i w_i$, где $w_i \in \Delta_n$, $w_i > 0$ при $1 \leqslant i \leqslant t$, $\beta = (\beta_1, \ldots, \beta_t) \in \Delta_t$ и $\beta > 0$. Тогда*

$$D(p\|q) \leqslant D(p\|w_i) + \ln\frac{1}{\beta_i}$$

для любого i. В частности, при $p = q$ имеем оценку расхождения между произвольным элементом w_i выпуклой комбинацией и вектором выпуклой комбинации:

$$D\left(w_i \| \sum_{i=1}^{t} \beta_i w_i\right) \leqslant \ln\frac{1}{\beta_i}$$

для любого i.

В дальнейшем на каждом шаге $0 \leqslant t \leqslant T$ вектор $w_t^m = (w_{1,t}^m, \ldots, w_{N,t}^m)$ будет представлять накопленные нормированные веса N экспертов на этом шаге. Новое апостериорное распределение $w_{t+1} = (w_{1,t+1}, \ldots, w_{N,t+1})$ экспертов (для использования

на шаге $t+1$) будет определяться в виде выпуклой комбинации $w_{t+1} = \sum\limits_{s=0}^{t} \beta_s^{t+1} w_s^m$ с весами β_s^{t+1}, $0 \leqslant s \leqslant t$, где $w_s^m = (w_{1,s}^m, \ldots, w_{N,s}^m)$ – вектор нормированных весов экспертов $\{1, \ldots, N\}$ на шаге t. В координатах эта операция записывается в виде $w_{i,t+1} = \sum\limits_{s=0}^{t} \beta_s^{t+1} w_{i,s}^m$ при $1 \leqslant i \leqslant N$.

Вектор $\beta^{t+1} = (\beta_0^{t+1}, \ldots, \beta_t^{t+1})$ называется *схемой смешивания*. Для него выполнено $\sum\limits_{s=0}^{t} \beta_s^{t+1} = 1$ и $\beta_s^{t+1} > 0$ при $0 \leqslant s \leqslant t$. Вектор β^{t+1} задает веса, с помощью которых смешиваются прошлые распределения экспертов. Он может заново задаваться на каждом шаге t.

Алгоритм смешивания апостериорных распределений экспертов MPP приведен на рис. 4.12. В отличие от ранее рассмотренных алгоритмов экспоненциального смешивания этот алгоритм использует не только текущие накопленные веса экспертов, но он также смешивает эти веса и все веса, накопленные на прошлых шагах. Это позволяет не забывать прошлые удачные комбинации экспертов даже, если локально на данном шаге эти эксперты несли большие потери. Если в будущем такие эксперты вновь будет нести малые потери, их веса быстро возрастут.

Приведем примеры схем смешивания.

Пример 1. $\beta_t^{t+1} = 1$ и $\beta_s^{t+1} = 0$ при $0 \leqslant s < t$. В этом случае, получаем обычный алгоритм экспоненциального смешивания $Hedge(\eta)$:

$$w_{i,t+1} = w_{i,t}^m = \frac{w_{i,t} e^{-\eta l_t^i}}{\sum\limits_{j=1}^{N} w_{j,t} e^{-\eta l_t^j}}.$$

для всех i и t.

Пример 2. Другие варианты схем смешивания могут быть получены следующим образом. Соответствующие алгоритмы будут называться Fixed-Share. В этих схемах смешивания $\beta_t^{t+1} = 1 - \alpha$ и $\sum\limits_{s=0}^{t-1} \beta_s^{t+1} = \alpha$, при этом,

Алгоритм MPP

Параметр $\eta > 0$. Полагаем $w_{i,1} = w^m_{i,0} = \frac{1}{N}$ при $i = 1, \ldots, N$.
FOR $t = 1, \ldots, T$
Получаем потери экспертов l^i_t при $1 \leqslant i \leqslant N$.
Вычисляем потери алгоритма $MPP : h_t = \sum\limits_{i=1}^{N} w_{i,t} l^i_t$.
Переопределяем веса всех экспертов в два этапа:
Loss Update
$w^m_{i,t} = \frac{w_{i,t} e^{-\eta l^i_t}}{\sum\limits_{j=1}^{N} w_{j,t} e^{-\eta l^j_t}}$ при $1 \leqslant i \leqslant N$.
Mixing Update
Выбираем схему смешивания $\beta^{t+1} = (\beta^{t+1}_0, \ldots, \beta^{t+1}_t)$ и определим $w_{i,t+1} = \sum\limits_{s=0}^{t} \beta^{t+1}_s w^m_{i,s}$ при $1 \leqslant i \leqslant N$.
ENDFOR

Рис. 4.12: Алгоритм MPP смешивания апостериорных распределений экспертов

- (i) $\beta^{t+1}_0 = \alpha$. В этом случае, получаем вариант алгоритма $FS(\eta)$ с постоянным параметром $0 \leqslant \alpha \leqslant 1$:

$$w_{i,t+1} = \frac{\alpha}{N} + (1 - \alpha) w^m_{i,t}$$

  для всех i и t;

- (ii) $\beta^{t+1}_s = \frac{\alpha}{t}$ при $0 \leqslant s < t$ – равномерное влияние прошлого;

- (iii) $\beta^{t+1}_s = \frac{\alpha}{(t-s)^\gamma} \frac{1}{Z_t}$ при $0 \leqslant s < t$, где $Z_t = \sum\limits_{s=0}^{t-1} \frac{1}{(t-s)^\gamma}$ – убывающее влияние прошлого (здесь $\gamma > 0$).

Анализ алгоритма MPP.

Мы несколько расширим понятие регрета. Раньше мы сравнивали потери смешивающего алгоритма с потерями наилучшего за данный период эксперта. Теперь, следуя работе [40], мы будем сравнивать потери смешивающего алгоритма с потерями наилучщей выпуклой комбинации потерь экспертов.

Напомним, что $L_T^i = \sum\limits_{t=1}^{T} l_t^i$ – кумулятивные потери эксперта i, $1 \leqslant i \leqslant N$, $H_T = \sum\limits_{t=1}^{T} h_t$ – кумулятивные потери алгоритма за первые T шагов.

Обозначим $l_t = (l_t^1, \ldots, l_t^N)$ – вектор потерь всех экспертов на шаге t. Также $m_t = -\frac{1}{\eta} \sum\limits_{i=1}^{N} w_{i,t} e^{-\eta l_t^i}$ смешанные потери на шаге t, $M_T = \sum\limits_{t=1}^{T} m_t$ кумулятивные смешанные потери за перве T шагов.

Используя неравенство Хефдинга можно получить $h_t \leqslant m_t + \frac{\eta}{8}$ и $H_t \leqslant M_t + \frac{t\eta}{8}$ для любого t. Предполагаем, что потери экспертов лежат в единичном интервале: $l_t^i \in [0, 1]$ для всех i и t.

Будем рассматривать выпуклые комбинации потерь экспертов на шаге t: $(q_t \cdot l_t) = \sum\limits_{i=1}^{N} q_{i,t} l_t^i$, где $q_t = (q_{1,t}, \ldots, q_{N,t}) \in \Delta_N$ – вектор сравнения потерь.

Полагаем $\beta_0^1 = 1$. Оценим сверху кумулятивные смешанные потери алгоритма MPP.

Теорема 4.14. *Для произвольных t, $0 \leqslant s \leqslant t-1$, и произвольного вектора сравнения $q_t \in \Delta_N$ будет*

$$m_t \leqslant (q_t \cdot l_t) + \frac{1}{\eta}(D(q_t \| w_t) - D(q_t \| w_t^m)) \leqslant \tag{4.79}$$

$$\leqslant (q_t \cdot l_t) + \frac{1}{\eta}\left(D(q_t \| w_s^m) - D(q_t \| w_t^m) + \ln \frac{1}{\beta_s^t}\right). \tag{4.80}$$

Доказательство. Используя (4.78), получим неравенство (4.79)

следующими преобразованиями:

$$m_t = -\frac{1}{\eta}\ln\sum_{i=1}^{N} w_{i,t}e^{-\eta l_t^i} =$$

$$= \sum_{i=1}^{N} q_{i,t}\left(l_t^i + \frac{1}{\eta}\ln e^{-\eta l_t^i} - \frac{1}{\eta}\ln\sum_{j=1}^{N} w_{j,t}e^{-\eta l_t^j}\right) =$$

$$= \sum_{i=1}^{N} q_{i,t}l_t^i + \frac{1}{\eta}\ln\sum_{i=1}^{N} q_{i,t}\ln\frac{w_{i,t}^m}{w_{i,t}} =$$

$$= (q_t\cdot l_t) + \frac{1}{\eta}(D(q_t\|w_t) - D(q_t\|w_t^m)).$$

Неравенство (4.80) следует из (4.79) по следствию 4.10. $\triangle$

Применим теорему 4.14 для некоторых схем смешивания из примеров 1-2.

Пример 1а. $\beta_t^{t+1} = 1$ и $\beta_s^{t+1} = 0$ при $0 \leqslant s < t$. В этом случае, имеем алгоритм экспоненциального смешивания $Hedge(\eta)$: $w_{i,t+1} = w_{i,t}^m = \frac{w_{i,t}e^{-\eta l_t^i}}{\sum\limits_{j=1}^{N} w_{j,t}e^{-\eta l_t^j}}$ для всех i и t.

Рассматриваем постоянную выпуклую комбинацию потерь экспертов на всех шагах. Суммируем неравенство (4.80) при постоянном векторе сравнения $q_t = q$ по $t = 1,\ldots,T$ и получаем

$$M_T \leqslant \sum_{t=1}^{T}(q\cdot l_t) + \frac{1}{\eta}\sum_{t=1}^{T}(D(q\|w_t) - D(q\|w_t^m)) =$$

$$= \sum_{t=1}^{T}(q\cdot l_t) + \frac{1}{\eta}(D(q\|w_1) - D(q\|w_T^m)) \leqslant$$

$$\leqslant \sum_{t=1}^{T}(q\cdot l_t) + \frac{1}{\eta}D(q\|w_1).$$

При переходе от первого равенства ко второму используем равенство $w_t = w_{t-1}^m$, которое имеет место для схемы смешивания β_t из этого примера. Соседние слагаемые вычитаемой суммы сокращаются и остаются только первое и последнее слагаемые. Последнее неравенство имеет место, так как $D(q\|w_T^m) \geqslant 0$.

Таким образом, мы доказали следующее утверждение.

Предложение 4.2. *Для любого $q \in \Delta_N$ для любого T выполнено неравенство*

$$M_T \leqslant \sum_{t=1}^{T}(q \cdot l_t) + \frac{1}{\eta}D(q\|w_1).$$

Кроме этого,

$$H_T \leqslant \sum_{t=1}^{T}(q \cdot l_t) + \frac{1}{\eta}D(q\|w_1) + \frac{T\eta}{8}.$$

Второе неравенство следует из неравенства Хефдинга и предположения $0 \leqslant l_t^i \leqslant 1$ при всех i и t.

Стандартная оценка для алгоритма $Hedge(\eta)$ получается при $q = (0, \ldots, 1, \ldots, 0)$. В этом случае, $w_1 = w_0^m = (\frac{1}{N}, \ldots, \frac{1}{N})$ и $D(q\|w_1) = \ln N$. В результате, получаем оценки $M_T \leqslant L_T^i + \frac{\ln N}{\eta}$ и $H_T \leqslant L_T^i + \frac{\ln N}{\eta} + \frac{T\eta}{8}$ для произвольного i.

Пример 2а.

В части (i) имеем схему смешивания $\beta_t^{t+1} = 1 - \alpha$, $\beta_s^{t+1} = 0$ при $0 < s < t$ и $\beta_0^{t+1} = \alpha$. Это вариант алгоритма $FS(\eta)$: $w_{i,t+1} = \frac{\alpha}{N} + (1-\alpha)w_{i,t}^m$ для всех i и t.

Предположим теперь, что вектор сравнения q_t изменяется k раз при $1 \leqslant t \leqslant T$. Пусть $t_1 < t_2 < \ldots t_k$ – те шаги, на которых происходят изменения, т.е. $q_{t_j} \neq q_{t_{j-1}}$ и $q_t = q_{t-1}$ для всех остальных шагов, где $t > 1$. Полагаем $t_0 = 1$ и $t_{k+1} = T + 1$. Применим теорему 4.14 для распределения β^{t+1} из (i). Напомним, что $w_{i,1} = w_{i,0}^m = \frac{1}{N}$ при $i = 1, \ldots, N$.

Предложение 4.3. *Для любой последовательности T векторов сравнения q_t с k изменениями*

$$M_T \leqslant \sum_{t=1}^{T}(q_t \cdot l_t) + \frac{1}{\eta}\sum_{j=0}^{k}\Big(D(q_{t_j}\|w_1) - D(q_{t_j}\|w_{t_{j+1}-1}^m)\Big) + \\ + \frac{1}{\eta}k\ln\frac{1}{\alpha} + \frac{1}{\eta}(T-k-1)\ln\frac{1}{1-\alpha}. \quad (4.81)$$

Доказательство. Применим на каждом шаге t неравенство (4.80) теоремы 4.14 для подходящего s.

В этом неравенстве, если $q_t = q_{t-1}$, то полагаем $s = t-1$ и используем схему смешивания $\beta_{t-1}^t = 1-\alpha$, т.е.

$$m_t \leqslant (q_t \cdot l_t) + \frac{1}{\eta} D(q_t \| w_{t-1}^m) - \frac{1}{\eta} D(q_t \| w_t^m) + \frac{1}{\eta} \ln \frac{1}{1-\alpha}.$$

При $t = 1$ полагаем $s = 0$ и $\beta_0^1 = 1$. По определению $w_0^m = (\frac{1}{N}, \ldots, \frac{1}{N})$. Получаем

$$m_1 \leqslant (q_1 \cdot l_1) + \frac{1}{\eta} D(q_1 \| w_0^m) - \frac{1}{\eta} D(q_1 \| w_1^m).$$

Для шагов t, где происходили изменения вектора сравнения: $t = t_1, \ldots, t_k$, полагаем $s = 0$, $\beta_0^{t_j} = \alpha$, т.е.

$$m_{t_j} \leqslant (q_{t_j} \cdot l_{t_j}) + \frac{1}{\eta} D(q_{t_j} \| w_0^m) - \frac{1}{\eta} D(q_{t_j} \| w_{t_j}^m) + \frac{1}{\eta} \ln \frac{1}{\alpha}.$$

Складываем все эти неравенства трех типов. Одинаковые по абсолютной величине и разные по знаку слагаемые внутри интервалов сократятся, останутся только начальные и конечные для каждого интервала разбиения слагаемые. Кроме этого, началу каждого интервала соответствует дополнительное слагаемое $\frac{1}{\eta} \ln \frac{1}{\alpha}$, а каждому шагу t, где $q_t = q_{t-1}$, соответствует дополнительное слагаемое $\frac{1}{\eta} \ln \frac{1}{1-\alpha}$. Таких дополнительных слагаемых первого типа – k, а второго типа $T-k-1$. В итоге получаем (4.81). $\triangle$

Пусть векторы сравнения q_t имеют вид $q_t = (0, \ldots 1, \ldots, 0)$ и на шагах $1 \leqslant t \leqslant T$ происходит k изменений таких векторов. Тогда $D(q_{t_j} \| w_0^m) = \ln N$ при $0 \leqslant j \leqslant k$ и оценка предложения 4.3 превращается в оценку алгоритма Fixed-Share при постоянных параметрах η и α.

Следствие 4.11. *Для любого составного эксперта $i_1, \ldots, i_T$ с k изменениями элементарных экспертов будет*

$$M_T \leqslant L_T(i_1, \ldots, i_T) + \frac{1}{\eta}(k+1) \ln N + \frac{1}{\eta} k \ln \frac{1}{\alpha} + \\ + \frac{1}{\eta}(T-k-1) \ln \frac{1}{1-\alpha},$$

а также

$$H_T \leqslant L_T(i_1, \ldots, i_T) + \frac{1}{\eta}(k+1)\ln N + \frac{1}{\eta}k\ln\frac{1}{\alpha} + \\ + \frac{1}{\eta}(T-k-1)\ln\frac{1}{1-\alpha} + \frac{T\eta}{8}.$$

Заметим, что для агрегирующего алгоритма Вовка (см. главу 5) $h_T \leqslant m_T$ и $H_T \leqslant M_T$ и поэтому нет необходимости использовать неравенство Хефдинга и предполагать ограниченность потерь l_t^i экспертов. Для этого алгоритма слагаемое $\frac{T\eta}{8}$ отсутствует.

Оценки регрета алгоритма MPP для схем смешивания из (ii) и (iii) приведены в работе [14]. Численные эксперименты, приведенные в работе [14], показывают более высокую эффективность этих схем смешивания по сравнению со схемой смешивания (i).

4.9. Предсказания с экспертами в условиях частичного мониторинга

Рассмотрим постановку задачи предсказания с экспертами в условиях частичного мониторинга окружающей среды. В этом случае, при принятии решения, предсказатель может не иметь доступа к потерям экспертов на прошлых шагах, а также к своим потерям, понесенным вследствие принятого им решения. Вместо этого, ему предоставляется значение некоторой величины производной от этих потерь. Приведем характерный пример (см. [43]). На очередном раунде t продавец (предсказатель) устанавливает цену некоторого товара в размере γ_t, при этом ему неизвестна максимальная цена $\omega_t \in \{1, \ldots, N\}$, которую готов заплатить покупатель. В этом случае, в конце раунда t продавец несет потери

$$l(\omega_t, \gamma_t) = \frac{(\omega_t - \gamma_t)1_{\gamma_t \leqslant \omega_t} + c1_{\gamma_t > \omega_t}}{N},$$

которые отражают его недополученную прибыль от слишком маленькой (по сравнению с ожиданиями покупателя) цены или потери от того, что товар не был продан. Здесь c – положительная константа. При этом, особенность проблемы заключается в том,

FOR $t = 1, 2, \ldots$
Природа выбирает исход ω_t, но не показывает его предсказателю.
Предсказатель выдает предсказание γ_t (но не знает величину своих потерь $l_t = l(\gamma_t, \omega_t)$).
Предсказатель получает величину отклика $h_t = h(\gamma_t, \omega_t)$.
ENDFOR

Рис. 4.13: Игра с предсказаниями в условиях частичного мониторинга

что продавец никогда не будет знать значение величины ω_t и, тем самым, значения своих потерь. Вместо этого, в конце очередного раунда t ему известен только результат $\gamma_t \leqslant \omega_t$ или $\gamma_t > \omega_t$ (был куплен товар или нет), т.е. значение функции $h(\gamma_t, \omega_t) = 1_{\gamma_t \leqslant \omega_t}$.

Возникает задача построения (рандомизированного) алгоритма, который устанавливал бы в режиме онлайн цены γ_t так, чтобы средние по времени потери продавца были (с вероятностью единица) близки к минимальным:

$$\frac{1}{T}\left(\sum_{t=1}^{T} l(\omega_t, \gamma_t) - \min_{1 \leqslant i \leqslant N} \sum_{t=1}^{T} l(i, \omega_t)\right) \to 0 \text{ при } T \to \infty.$$

Считаем, что $\gamma_t \in \{1, \ldots, N\}$. Таким образом, продавцу известна матрица $\{h(i,j)\}_{i,j=1}^{N}$ всех возможных откликов окружения.

Можно расматривать этот процесс в общем случае. Имеется функция потерь $l(i,j) \in [0,1]$ и функция отклика $h(i,j)$, где $i, j \in \{1, \ldots, N\}$.

Протокол соответствующей игры представлен на рис. 4.13.

Соответствующий алгоритм и оценка его регрета приведены в книге [43] (раздел 6.4). Мы рассмотрим эту проблему в наиболее значимом частном случае.

Частным случаем задачи предсказания с частичным мониторингом является так называемая “проблема многоруких бандитов” (multi-armed bandit problem) (см. обзор [17]).[6]

[6]Эта терминология возникла из реальной жизни – в США называют игральный автомат – “one-armed bandit”. Имеется несколько игральных автома-

В этом случае, предсказатель в результате своего действия получает только собственные потери и не имеет доступа к потерям других экспертов.

Эта задача первоначально была сформулирована Роббинсом [46] в стохастической постановке, при которой предполагается, что потери каждого эксперта генерируются случайным образом независимо от потерь остальных экспертов с помощью фиксированного присущему ему распределения вероятностей. Вероятностные распределения экспертов имеют различные средние и неизвестны предсказателю.

При другой, минимаксной (или адаптивно враждебной), постановке мы ничего не предполагаем о природе источника, предсказатель имеет дело только с исторической последовательности потерь [11]. Эта постановка будет рассмотрена в данном разделе. Про обе постановки см. в обзоре [17].

Переходим к изложению задачи в минимаксной постановке. На каждом шаге t *Предсказатель* выбирает номер I_t эксперта ("ручку" автомата) и несет потери $l_{I_t,t}$. Потери $l_{i,t}$, $i \neq I_t$, $i = 1, \ldots, N$, всех остальных экспертов на шаге t ему недоступны. Регрет алгоритма определяется как

$$R_T = \sum_{t=1}^{T} l_{I_t,t} - \min_i \sum_{t=1}^{T} l_{i,t}.$$

Например, если *Предсказатель* выбрал $I_t = 1$, то $l_{1,t} = 1$ и $l_{2,t} = 0$, а если *Предсказатель* выбрал $I_t = 2$, то $l_{1,t} = 0$ и $l_{2,t} = 1$. Легко видеть, что как бы *Предсказатель* не выбирал I_t будет $R_T \geqslant T/2$ для всех T.

Тем не менее, можно добиться приемлемого среднего регрета, если применять рандомизированные стратегии для выбора I_t.

Номер эксперта I_t может быть случайной величиной (*Предсказатель* применяет рандомизацию по некоторому распределению

тов. На каждом раунде игрок дергает за ручку одного из автоматов и получает выигрыш (или проигрыш). При этом он не знает, какой выигрыш он бы получил, если бы дергал ручки других автоматов. Перед игроком стоит проблема выбора автомата для игры на следующем раунде. Данная постановка актуальна во многих жизненных ситуациях, где надо принимать решения в условиях неопределенности.

Алгоритм Exp3

Заданы параметры обучения $\eta_1 > \eta_2 > \cdots > 0$ и начальное распределение $p_1 = (\frac{1}{N}, \ldots, \frac{1}{N})$ на N экспертах.
FOR $t = 1, 2, \ \ldots$
Предсказатель выбирает $I_t \sim p_t$ и несет потери $l_{I_t,t}$.
Предсказатель вычисляет оценку потерь экспертов $\tilde{l}_{i,t} = \frac{l_{i,t}}{p_{i,t}} 1_{I_t=i}$, а также оценку кумулятивных потерь $\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \tilde{l}_{i,t}$.
Производится пересчет вероятностного распределения p_{t+1}
$p_{i,t+1} = \frac{e^{-\eta_t \tilde{L}_{i,t}}}{\sum\limits_{j=1}^{N} e^{-\eta_t \tilde{L}_{j,t}}}$ при $1 \leqslant i \leqslant N$.
ENDFOR

Рис. 4.14: Алгоритм Exp3

вероятностей p_t, которое вычисляет в качестве прогноза на каждом раунде t игры). В этом случае регрет R_T является случайной величиной, также имеет смысл рассмотреть математическое ожидание регрета

$$E(R_T) = E_{I_t \sim p_t} \left(\sum_{t=1}^{T} l_{I_t,t} - \min_i \sum_{t=1}^{T} l_{i,t} \right).$$

Алгоритм и соответствующая игра представлены на рис. 4.14.

Теорема 4.15. *Пусть $0 \leqslant l_{i,t} \leqslant 1$ для всех i и t, а параметр обучения переменный и имеет вид $\eta_t = \sqrt{\frac{\ln N}{tN}}$. Тогда*

$$E_{I_t \sim p_t}(R_T) \leqslant 2\sqrt{TN \ln N}$$

для каждого T, где $R_T = \sum\limits_{t=1}^{T} l_{I_t,t} - \min_{1 \leqslant i \leqslant N} \sum\limits_{t=1}^{T} l_{i,t}$.

Доказательство. Покажем, что для любой последовательности $\eta_1 > \eta_2 > \cdots > 0$ будет $E_{I_t \sim p_t}(R_T) \leqslant \frac{N}{2} \sum\limits_{t=1}^{T} \eta_t + \frac{\ln N}{\eta_T}$ для каждого T.

Используем обозначения $p_{i,t} = P\{I_t = i\}$, $\tilde{l}_{i,t} = \frac{l_{i,t}}{p_{i,t}} 1_{I_t=i}$. Тогда

$$E_{I_t \sim p_t}(\tilde{l}_{i,t}) = \sum_{j=1}^{N} p_{j,t} \frac{l_{i,t}}{p_{i,t}} 1_{j=i} = l_{i,t}. \tag{4.82}$$

$$E_{i \sim p_t}(\tilde{l}_{i,t} | I_t) = \sum_{i=1}^{N} p_{i,t} \frac{l_{i,t}}{p_{i,t}} 1_{I_t=i} = l_{I_t,t}. \tag{4.83}$$

$$E_{i \sim p_t}(\tilde{l}^2_{i,t} | I_t) = \sum_{i=1}^{N} p_{i,t} \frac{l^2_{i,t}}{p^2_{i,t}} 1_{I_t=i} = \frac{l^2_{I_t,t}}{p_{I_t,t}}. \tag{4.84}$$

$$E_{I_t \sim p_t}\left(\frac{1}{p_{I_t,t}}\right) = \sum_{j=1}^{N} p_{j,t} \frac{1}{p_{j,t}} = N. \tag{4.85}$$

Пусть $k = \operatorname{argmin}_{1 \leqslant i \leqslant N} \sum\limits_{t=1}^{T} l_{i,t}$. Из (4.83) и (4.82) следует, что

$$R_T = \sum_{t=1}^{T} l_{I_t,t} - \sum_{t=1}^{T} l_{k,t} = \sum_{t=1}^{T} E_{i \sim p_t}(\tilde{l}_{i,t} | I_t) - \sum_{t=1}^{T} E_{I_t \sim p_t}(\tilde{l}_{k,t}). \tag{4.86}$$

Свойство (4.82) означает, что оценка $\tilde{l}_{i,t}$ является несмещенной.

Предварительно заметим, что

$$\frac{1}{\eta_t} \ln E_{i \sim p_t}\left(\exp(-\eta_t(\tilde{l}_{i,t} - E_{k \sim p_t}(\tilde{l}_{k,t} | I_t))) | I_t\right) =$$
$$= \frac{1}{\eta_t} \ln \left(\exp(\eta_t E_{k \sim p_t}(\tilde{l}_{k,t} | I_t)) E_{i \sim p_t}\left(\exp(-\eta_t \tilde{l}_{i,t}) | I_t\right)\right) =$$
$$= \frac{1}{\eta_t} \ln E_{i \sim p_t}\left(\exp(-\eta_t \tilde{l}_{i,t}) | I_t\right) + E_{k \sim p_t}(\tilde{l}_{k,t} | I_t). \tag{4.87}$$

Исходя из представления (4.87), запишем слагаемое из первой суммы разности (4.86) в виде

$$E_{i \sim p_t}(\tilde{l}_{i,t} | I_t) =$$
$$= \frac{1}{\eta_t} \ln E_{i \sim p_t}\left(\exp(-\eta_t(\tilde{l}_{i,t} - E_{k \sim p_t}(\tilde{l}_{k,t} | I_t)) | I_t\right) - \tag{4.88}$$
$$- \frac{1}{\eta_t} \ln E_{i \sim p_t}(\exp(-\eta_t \tilde{l}_{i,t})) | I_t). \tag{4.89}$$

Используем неравенства $\ln x \leqslant x-1$ и $e^{-x}-1+x \leqslant \frac{x^2}{2}$ при $x>0$.

Используя (4.84) можно оценить (4.88) следующим образом:

$$\ln E_{i\sim p_t}\left(\exp(-\eta_t(\tilde{l}_{i,t}-E_{k\sim p_t}(\tilde{l}_{k,t}|I_t))|I_t\right) = \tag{4.90}$$

$$= \ln E_{i\sim p_t}\left(\exp(-\eta_t\tilde{l}_{i,t})|I_t\right) + \eta_t E_{k\sim p_t}(\tilde{l}_{k,t}|I_t) \leqslant \tag{4.91}$$

$$\leqslant E_{i\sim p_t}\left(\exp(-\eta_t\tilde{l}_{i,t})|I_t\right) - 1 + \eta_t E_{i\sim p_t}(\tilde{l}_{i,t}|I_t) =$$

$$= E_{i\sim p_t}\left(\exp(-\eta_t\tilde{l}_{i,t}) - 1 + \eta_t\tilde{l}_{i,t}|I_t\right) \leqslant$$

$$\leqslant E_{i\sim p_t}\left(\frac{\eta_t^2\tilde{l}_{i,t}^2}{2}|I_t\right) = \frac{\eta_t^2 l_{I_t,t}^2}{2p_{I_t,t}} \leqslant \frac{\eta_t^2}{2p_{I_t,t}}. \tag{4.92}$$

При переходе от (4.90) к (4.91) мы использовали то, что член $E_{k\sim p_t}(\tilde{l}_{k,t}|I_t)$ не зависит от i. В оценке (4.92) использовали $0 \leqslant l_{i,t} \leqslant 1$ при всех i и t.

Остается оценить величину (4.89). Определим $\Phi_0(\eta)=0$ и

$$\Phi_t(\eta) = \frac{1}{\eta}\ln\frac{1}{N}\sum_{i=1}^{N} e^{-\eta\tilde{L}_{i,t}}.$$

Величина $\Phi_t(\eta)$ равна минус кумулятивным смешанным потерям M_T^η из раздела (4.5.1) (см. лемму 4.6). Имеем

$$-\frac{1}{\eta_t}\ln E_{i\sim p_t}(\exp(-\eta_t\tilde{l}_{i,t})|I_t) =$$

$$-\frac{1}{\eta_t}\ln\frac{\sum\limits_{i=1}^{N} e^{-\eta_t\tilde{L}_{i,t-1}}e^{-\eta_t\tilde{l}_{i,t}}}{\sum\limits_{i=1}^{N} e^{-\eta_t\tilde{L}_{i,t-1}}} = \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t). \tag{4.93}$$

В результате, используя оценку (4.92) величины (4.88) и оценку (4.93) величины (4.89), получаем

$$E_{i\sim p_t}(\tilde{l}_{i,t}|I_t) \leqslant \frac{\eta_t}{2p_{I_t,t}} + \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t). \tag{4.94}$$

Используя (4.94), оценим (4.86)

$$R_T = \sum_{t=1}^{T} l_{I_t,t} - \sum_{t=1}^{T} l_{k,t} =$$

$$= \sum_{t=1}^{T} E_{i\sim p_t}(\tilde{l}_{i,t}|I_t) - \sum_{t=1}^{T} E_{I_t\sim p_t}(\tilde{l}_{k,t}) \leqslant$$

$$\leqslant \sum_{t=1}^{T} \frac{\eta_t}{2p_{I_t,t}} + \sum_{t=1}^{T} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) - \sum_{t=1}^{T} E_{I_t\sim p_t}(\tilde{l}_{k,t}). \qquad (4.95)$$

Переставляя слагаемые, получим [7]

$$\sum_{t=1}^{T} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) = \sum_{t=1}^{T-1} \Phi_t(\eta_{t+1}) - \Phi_t(\eta_t) - \Phi_T(\eta_T) \leqslant -\Phi_T(\eta_T).$$

Здесь мы использовались свойством монотонности: $\Phi_t(\eta) < \Phi_t(\mu)$ при $\eta < \mu$. Далее, для произвольного k

$$-\Phi_T(\eta_T) = -\frac{1}{\eta_T} \ln \frac{1}{N} \sum_{i=1}^{N} e^{-\eta_T \tilde{L}_{i,t}} =$$

$$= \frac{\ln N}{\eta_T} - \frac{1}{\eta_T} \ln \sum_{i=1}^{N} e^{-\eta_T \tilde{L}_{i,t}} \leqslant$$

$$\leqslant \frac{\ln N}{\eta_T} - \frac{1}{\eta_T} \ln e^{-\eta_T \tilde{L}_{k,t}} =$$

$$= \frac{\ln N}{\eta_T} + \sum_{t=1}^{T} \tilde{l}_{k,t}.$$

---

[7] Подробнее $\sum_{t=1}^{T} \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) =$
$= \Phi_0(\eta_1) + (-\Phi_1(\eta_1) + \Phi_1(\eta_2)) + (-\Phi_2(\eta_2) + \Phi_2(\eta_3)) +$
$+(-\Phi_3(\eta_3) + \Phi_3(\eta_4)) + \ldots (-\Phi_{T-1}(\eta_{T-1}) + \Phi_{T-1}(\eta_T)) - \Phi_T(\eta_T).$

Отсюда используя свойства (4.82) и (4.85) получаем из (4.95)

$$E(R_T) \leqslant \frac{N}{2}\sum_{t=1}^{T}\eta_t + \frac{\ln N}{\eta_T} + \sum_{t=1}^{T} E_{I_t\sim p_t}(\tilde{l}_{k,t}) - \sum_{t=1}^{T} E_{I_t\sim p_t}(\tilde{l}_{k,t}) =$$

$$\leqslant \frac{N}{2}\sum_{t=1}^{T}\eta_t + \frac{\ln N}{\eta_T} \leqslant 2\sqrt{TN\ln N}$$

для каждого T, где $\eta_t = \sqrt{\frac{\ln N}{tN}}$.

Остается проверить монотонность $\Phi_t(\eta)$. Из выпуклости параболы имеем при $\eta < \mu$

$$\Phi_t(\eta) = \frac{1}{\eta}\ln\sum_{i=1}^{N}\frac{1}{N}e^{-\eta\tilde{L}_t^i} = \frac{1}{\eta}\ln\sum_{i=1}^{N}\frac{1}{N}\left(e^{-\mu\tilde{L}_t^i}\right)^{\frac{\eta}{\mu}} \leqslant$$

$$\frac{1}{\eta}\ln\left(\sum_{i=1}^{N}\frac{1}{N}e^{-\mu\tilde{L}_t^i}\right)^{\frac{\eta}{\mu}} = \frac{1}{\mu}\ln\sum_{i=1}^{N}\frac{1}{N}e^{-\mu\tilde{L}_t^i} = \Phi_t(\mu).$$

Теорема доказана. $\triangle$

4.10. Задачи и упражнения

1. Построить вариант алгоритма большинства для случая когда имеется эксперт, про которого известно, что он делает не более k ошибок. Получить оценку числа ошибок алгоритма большинства.

2. Рассмотрим протокол игры с предсказаниями экспертов, в котором *Природа* выдает последовательность $0^T(01)^T1^T$. Имеется три эксперта, каждый из которых выдает постоянное предсказание: *Эксперт 1* всегда предсказывает $\xi_t^1 = 0$ для всех $t = 1, \ldots, 4T$, *Эксперт 2* предсказывает $\xi_t^2 = 1$ для всех $t = 1, \ldots, 4T$, *Эксперт 3* предсказывает $\xi_t^3 = 1/2$ для всех $t = 1, \ldots, 4T$. Функция потерь – $\lambda(\omega, \gamma) = |\omega - \gamma|$.

Вычислить для всех $t = 1, \ldots, 4T$:

(i) веса экспертов;

(ii) потери *Распределителя* из алгоритма $Hedge$ и предсказания алгоритма экспоненциального взвешивания.

3. Проверить простейшие свойства экспоненциального распределения с плотностью $p(x) = e^{-x}$: $P\{\xi > a\} = e^{-a}$ и $P\{\xi > a+b\} = e^{-b}P\{\xi > a\}$ для всех неотрицательных значений a и b.

4. Доказать, что для любой неотрицательной случайной величины η с плотностью распределения $p(t)$ выполнено соотношение:

$$E(\eta) = \int\limits_0^\infty P\{\eta \geqslant y\}dy.$$

(*Указание*: Использовать свойство $p(y) = F'(y)$, где $F(y) = \int_0^y p(t)dt = 1 - P\{\eta \geqslant y\}$ – функция распределения случайной величины. После этого, проинтегрировать по частям $E(\eta) = \int_0^\infty tp(t)dt$).

5. Провести доказательство леммы 4.5 для того случая, когда на каждом шаге t в алгоритмах FPL и IFPL для рандомизации используется вся серия случайных величин $\xi_t^1, \ \ldots, \ \xi_t^N$, $t = 1, 2, \ \ldots$.

6. Проведите сравнение рандомизированной версии алгоритма экспоненциального взвешивания для случая простой функции потерь с теоремой 4.2. Покажите, что средние потери этого алгоритма примерно в два раза меньше чем потери детерминированного алгоритма взвешенного большинства. Приведите точную оценку.

7. Показать, что алгоритм оптимального распределения потерь из раздела 4.2 есть частный случай алгоритма экспоненциального взвешивания экспертных решений из раздела 4.6.1.

(*Указание:* Рассмотреть в качестве пространства исходов $\Omega = [0,1]^N$, состоящее из наборов потерь экспертов $\bar{l} = (l^1, \ldots, l^N)$, где N – число экспертов. В качестве пространства решений рассмотреть симплекс Γ вероятностных распределений $\bar{p} = (p^1, \ldots, p^N)$, а в качестве функции потерь рассмотреть функцию $\lambda(\bar{l}, \bar{p}) = (\bar{l} \cdot \bar{p})$ – скалярное произведение векторов $\bar{l} \in [0,1]^N$ и $\bar{p} \in \Gamma$.

Прогноз i-го эксперта есть единичный вектор $\bar{\xi}^i = (0, \ldots, 1, \ldots, 0)$ размерности N. Тогда прогноз алгоритма экспоненциального взвешивания на шаге t, приведенный в разделе 4.6.1, представляется в виде вектора:

$$\bar{\xi}_t = \sum_{i=1}^N \xi_t^i w_{i,t}^* = \bar{p}_t,$$

где $\bar{p}_t = (w^*_{1,t}, \ldots, w^*_{N,t})$ – набор нормированных весов экспертов на шаге t, вычисленных как в алгоритме Hedge (см. [71])).

8. Проверить, что можно сформулировать финальную часть алгоритма Fixed-Share в следующем виде:

Loss Update: $w^m_{i,t} = \frac{w_{i,t}e^{-\eta l^i_t}}{\sum\limits_{j=1}^{N} w_{j,t}e^{-\eta l^j_t}}$

Fixed-Share Update: $w_{i,t+1} = \frac{\alpha_t}{N-1} + \left(1 - \frac{N}{N-1}\alpha_t\right) w^m_{i,t}$.

Доказать, что оценка регрета останется прежней (*Указание.* Показать, что величины $w^*_{i,t}$ остаются теми же, что и в первоначальной версии алгоритма, и поэтому, все рассуждения далее, в том числе оценка регрета, сохраняются прежними.)

10. Как изменится оценка регрета, если часть Fixed-Share Update имеет вид: $w_{i,t+1} = \frac{\alpha_t}{N} + (1-\alpha_t)w^m_{i,t}$.

11. Построить алгоритм Fixed-Share с переменным параметрам обучения $\eta_t = \sqrt{\frac{4\ln N}{t}}$, провести его анализ.

Указание. Использовать монотонность величины $M_T = M^\eta_T$ по η.

11. Доказать, что для любого конечного множества из N экспертов, которые на каждом шаге $t = 1, 2, \ldots$ делают прогнозы $\xi^i_t \in [0,1]$, $1 \leqslant i \leqslant N$, и вычисляют свои потери на каждом шаге с помощью одной из функций потерь: абсолютной, квадратичной или логарифмической (одной для всех экспертов), найдется такая последовательность исходов для которой кумулятивные потери каждого из экспертов за первые T шагов не меньше чем $T - O(\sqrt{T\ln N})$.

12. Привести псевдокод алгоритма ресэмплинга, который по данной выборке генерирует новую выборку элементы которой состоят из элементов исходной выборки и распределены согласно заданному распределению вероятностей.

4.11. Лабораторные работы

Лабораторная работа 1

Загрузить временные ряды цен нескольких акций из вебсайта

www.finam.ru (можно взять данные цен по дням в течении года или более). Составить алгоритмы на основе алгоритма Hedge, Fixed-Share и MPP для перераспределения средств, вложенных в акции, в зависимости от их доходности. Предварительно откалибровать данные так, чтобы они находились в единичном интервале. Переформулировать алгоритм Hedge для случая выигрышей.

Лабораторная работа 2

Провести сравнительный анализ кумулятивных потерь алгоритмов Hedge, Fixed-Share и приводимого ниже алгоритма Variable-Share (см. [27]. Этот алгоритм аналогичен алгоритму Fixed-Share, приведенному на рис. 4.11, за исключением последнего этапа, который теперь имеет вид:

Variable-Share Update

$w_{i,t+1} = (1-(1-\alpha_t)^{l_t^i})w_{t,i}^m + \frac{1}{N-1}\sum\limits_{j\neq i}(1-\alpha_t)^{l_t^i}w_{t,j}^m.$

В двух последних алгоритмах подобрать постоянные значения $\alpha_t = \alpha$.

Лабораторная работа 3

Написать программу алгоритма AdaBoost, использующего в качестве слабого алгоритма классификации WeakLearn готовое программное обеспечение SVM, описанное в разделе 2.13. Провести усиление алгоритма классификации рукописных цифр (из базы USPS, содержащей цифровые образы различных написаний рукописных цифр).

Глава 5

Агрегирующий алгоритм Вовка

Рассмотренные в главе 4 алгоритмы машинного обучения, использующие конкурирующие экспертные стратегии, имели регрет (ошибку обучения) порядка $O(\sqrt{T \ln N})$, где T – длина периода, N – число экспертных стратегий. Для некоторых специальных функций потерь, среди которых – квадратичная и логарифмическая – эту ошибку можно значительно уменьшить до величины порядка $O(\ln N)$. В данной главе будут сформулированы общие требования к подобным функциям потерь и будет описан соответствующий агрегирующий алгоритм, имеющий регрет $O(\ln N)$. Агрегирующий алгоритм Вовка представляет собой метод смешивания экспертных стратегий в некотором смысле эквивалентный методам представленным в главе 4.

5.1. Смешиваемые функции потерь

Рассматриваем простейший случай, когда множество исходов является двухэлементным $\Omega = \{0, 1\}$ и множество предсказаний есть единичный интервал $\Gamma = [0, 1]$. Аналогичным образом рассматривается случай $\Omega = \{-1, 1\}$ и $\Gamma = [-1, 1]$. Мы будем предполагать, что функции потерь $\lambda(\omega, \gamma)$ является неотрицательной и удовлетворяет следующим условиям:

- при каждом ω функция $\lambda(\omega, \gamma)$ непрерывна по γ;
- существует $\gamma \in [0, 1]$ такое, что оба значения $\lambda(0, \gamma)$ и $\lambda(1, \gamma)$ конечные;
- не существует $\gamma \in [0, 1]$ такого, что оба значения $\lambda(0, \gamma)$ и $\lambda(1, \gamma)$ бесконечные.

Для произвольной функции потерь $\lambda(\omega, \gamma)$ рассматривается *множество предсказаний*

$$\Pi_\lambda = \{(x, y) : \exists\, p\ (\lambda(0, p) = x, \lambda(1, p) = y)\} \tag{5.1}$$

и *множество суперпредсказаний*

$$\Sigma_\lambda = \{(x, y) : \exists\, p\ (\lambda(0, p) \leqslant x, \lambda(1, p) \leqslant y)\}. \tag{5.2}$$

Из первого свойства функции потерь и компактности $[0, 1]$ следует, что множество суперпредсказаний замкнуто. Для рассматриваемых ниже функций потерь множество предсказаний (5.1) является границей множества суперпредсказаний (5.2).

Нам будет удобно называть полуплоскость $[0, +\infty)^2$, в которой рассматриваются множества предсказаний и суперпредсказаний, *пространством предсказаний*.

Для произвольного $\eta > 0$ пусть $E_\eta : [0, +\infty)^2 \to (0, 1]^2$ есть гомоморфизм из пространства предсказаний в *экспоненциальное пространство*

$$E_\eta(x, y) = (e^{-\eta x}, e^{-\eta y}) \tag{5.3}$$

для всех $x, y \in [0, +\infty)$.

При этом гомоморфизме множество предсказаний (5.1) переходит в множество

$$E_\eta(\Pi_\lambda) = \{(e^{-\eta\lambda(0,p)}, e^{-\eta\lambda(1,p)}) : p \in \Gamma\},$$

а множество суперпредсказаний (5.2) переходит в множество

$$E_\eta(\Sigma_\lambda) = \{(x, y) : \exists\, p\ (0 \leqslant x \leqslant e^{-\eta\lambda(0,p)}, 0 \leqslant y \leqslant e^{-\eta\lambda(1,p)})\}. \tag{5.4}$$

Функция потерь $\lambda(\omega, \gamma)$ называется *η-смешиваемой*, если множество $E_\eta(\Sigma_\lambda)$ является выпуклым. Функция потерь называется *смешиваемой*, если она является η-смешиваемой для некоторого $\eta > 0$.

Мы будем рассматривать следующие функции потерь: логарифмическую, квадратичную, абсолютную и простую. Первые две будут смешиваемыми.

В случае, когда Ω – конечное, Γ – множество всех распределений вероятностей на Ω, логарифмическая функция потерь определяется: $\lambda(\omega, \gamma) = -\ln \gamma\{\omega\}$, где $\omega \in \Omega$ и $\gamma \in \Gamma$ – вероятностная мера на конечном множестве Ω.

Если $\Omega = \{0, 1\}$, то можно отождествить γ с вероятностью единицы, тогда $1-\gamma$ – это вероятность нуля. В этом случае можно взять $\Gamma = [0, 1]$ и рассмотреть логарифмическую функцию потерь в виде

$$\lambda(\omega, \gamma) = -\ln(\omega\gamma + (1-\omega)(1-\gamma))$$

или, более подробно,

$$\lambda(\omega, \gamma) = \begin{cases} -\ln \gamma, \text{ если } \omega = 1, \\ -\ln(1-\gamma), \text{ если } \omega = 0. \end{cases}$$

Обобщенная логарифмическая функция потерь определяется как

$$\lambda(\omega, \gamma) = -\frac{1}{\eta} \ln(\omega\gamma + (1-\omega)(1-\gamma)), \tag{5.5}$$

где $\eta > 0$ – параметр.

Квадратичная функция потерь определяется как

$$\lambda(\omega, \gamma) = c(\omega - \gamma)^2,$$

где c – некоторая положительная константа. Можно рассмотреть $\Omega = \{0, 1\}$ и $\Gamma = [0, 1]$.

Можно также использовать непрерывное множество исходов – единичный интервал $\Omega = [-1, 1]$ и аналогичное множество предсказаний $\Gamma = [-1, 1]$. Эти множества будут рассматриваться в задаче регрессии.

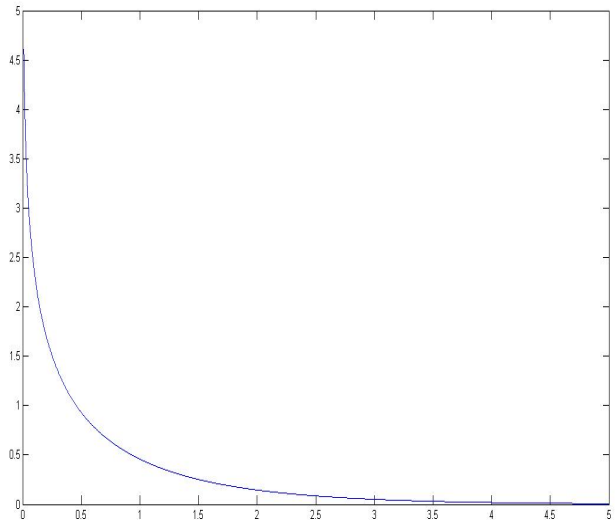

Рис. 6.1. Множество предсказаний и суперпредсказаний логарифмической функции потерь

Абсолютная функция потерь это

$$\lambda(\omega, \gamma) = c|\omega - \gamma|,$$

где c – некоторая положительная константа. Для этой функции потерь используются те же множества исходов и предсказаний, что и для квадратичной функции потерь.

Простая игра на предсказание (простая функция потерь) рассматривается в случае $\Omega = \Gamma = \{0, 1\}$. Функция потерь совпадает с абсолютной функцией потерь (при $c = 1$) и удовлетворяет свойству

$$\lambda(\omega, \gamma) = \begin{cases} 0, \text{ если } \omega = \gamma, \\ 1 \text{ в противном случае.} \end{cases}$$

Обсудим геометрические свойства смешиваемых функций потерь. Здесь обобщенная логарифмическая функция потерь играет

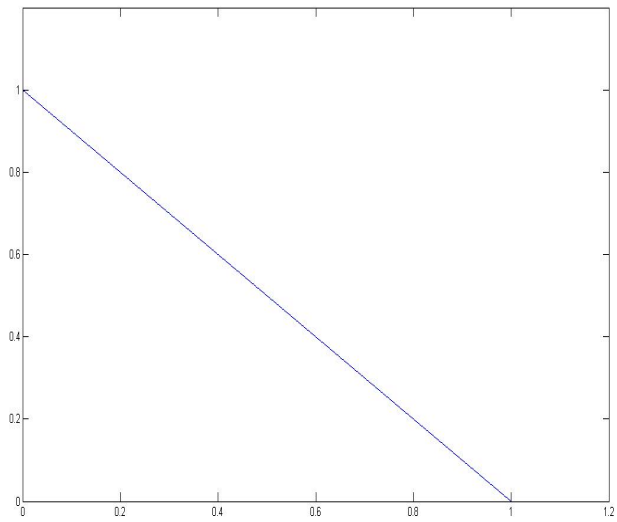

Рис. 6.2. Образы множества предсказаний и суперпредсказаний логарифмической функции потерь в экспоненциальном пространстве

особую роль.

Легко видеть, что множество предсказаний (5.1) обобщенной логарифмической функции потерь (5.5) есть кривая:

$$\{(x, y) : e^{-\eta x} + e^{-\eta y} = 1\}. \tag{5.6}$$

Мы будем рассматривать параллельные сдвиги кривой (5.6) в плоскости суперпредсказаний, т.е. кривые вида

$$\{(x, y) : e^{-\eta(x-\alpha)} + e^{-\eta(y-\beta)} = 1\}, \tag{5.7}$$

для произвольного вектора (α, β).

Говорим, что точка плоскости (x_1, y_1) находится "северо-восточнее", чем точка плоскости (x_2, y_2), если $x_1 \geqslant x_2$ и $y_1 \geqslant y_2$.

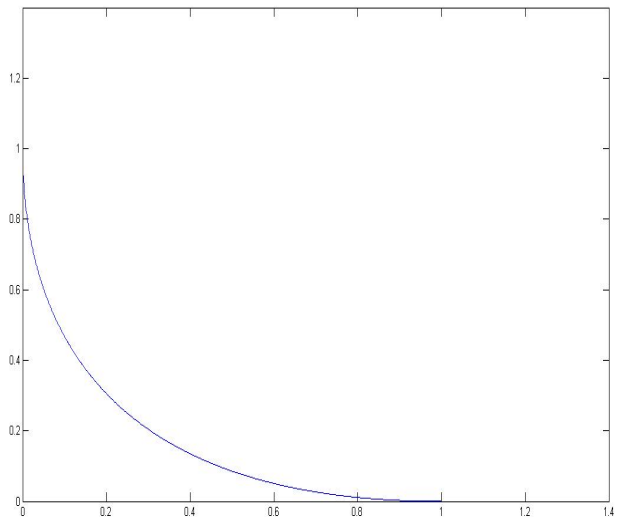

Рис. 6.3. Множество предсказаний и суперпредсказаний квадратичной функции потерь (при подходящем β)

Множество $A \subseteq \mathcal{R}^2$ находится северо-восточнее некоторого параллельного сдвига кривой (5.6), если каждая его точка находится северо-восточнее некоторой точки, лежащей на этом сдвиге (5.7).

Заметим, что прообразами всех прямых вида $ax + by = c$, где $a > 0$ и $b > 0$, рассматриваемых в экспоненциальном пространстве, при гомоморфизме (5.3) являются все параллельные сдвиги кривой $e^{-\eta x} + e^{-\eta y} = 1$, рассматриваемой в пространстве суперпредсказаний. Действительно, легко проверить, что прообраз прямой $ax + by = c$ при гомоморфизме E_η есть кривая

$$ae^{-\eta x} + be^{-\eta y} = c,$$

т.е. параллельный сдвиг кривой $e^{-\eta x} + e^{-\eta y} = 1$ на вектор

$$\left(-\frac{1}{\eta}\ln\frac{a}{c}, -\frac{1}{\eta}\ln\frac{b}{c}\right).$$

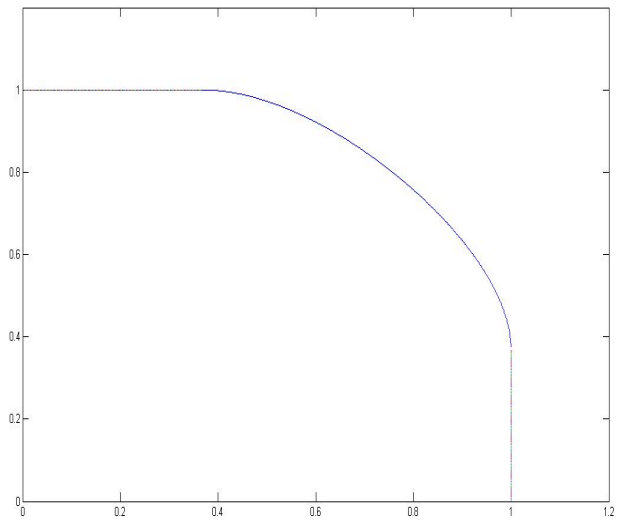

Рис. 6.4. Образы множества предсказаний и суперпредсказаний квадратичной функции потерь в экспоненциальном пространстве

Таким образом, имеется взаимно-однозначное соответствие между такими прямыми $ax+by=c$ в экспоненциальном пространстве и параллельными сдвигами кривой $e^{-\eta x}+e^{-\eta y}=1$ в пространстве суперпредсказаний.

Легко видеть, что образ $E_\eta(\Sigma_\lambda)$ множества суперпредсказаний в экспоненциальном пространстве является выпуклым тогда и только тогда, когда для любой точки его границы существует прямая, проходящая через эту точку такая, что весь этот образ множества суперпредсказаний находится по одну сторону от этой прямой.

Переводя это свойство из экспоненциального пространства в пространство суперпредсказаний, получим следующее характеристическое свойство смешиваемости функции потерь.

Предложение 5.1. *Функция потерь является η-смешиваемой тогда и только тогда, когда для любой точки (a, b), лежащей на границе множества суперпредсказаний, существует параллельный сдвиг $e^{-\eta(x-\alpha)} + e^{-\eta(y-\beta)} = 1$ кривой $e^{-\eta x} + e^{-\eta y} = 1$, проходящий через точку (a, b), и такой, что все множество суперпредсказаний лежит северо-восточнее этого сдвига.*

В следующих разделах мы будем рассматривать смешиваемые функции потерь. Оказывается, что при некоторых интервалах значений параметра η логарифмическая и квадратичная функции потерь оказываются η-смешиваемыми, абсолютная функция потерь этим свойством не обладает.

Для смешиваемых функций потерь чрезвычайно эффективным является так называемый агрегирующий алгоритм, который бы предложен в 1990 году В.Г. Вовком [57]. Этот алгоритм был исторически одним из первых алгоритмов подобного рода. Он является обобщением более простого алгоритма взвешенного большинства, который был предложен в 1989 году Литлстоуном и Вармутом [42]. Агрегирующий алгоритм Вовка имеет регрет, который зависит только от числа экспертов и не зависит от длины последовательности.

5.2. Конечное множество экспертов

В разделах 4.6.1 и 4.6.2 алгоритмы предсказания имели регрет $O(\sqrt{T \ln N})$, где T – длина периода, а N – число экспертов. Алгоритмы и результаты этих разделов относились к функциям потерь произвольного вида (в разделе 4.6.1 дополнительно требовалась выпуклость функции потерь по прогнозам).

В этом разделе приведем алгоритм смешивания прогнозов, который является оптимальным для смешиваемых функций потерь (логарифмической, квадратичной).

Приводимый ниже алгоритм имеет регрет не зависящий от длины периода T. Эта ошибка имеет вид $O(\ln N)$, где N – число экспертов.

В дальнейшем мы построим стратегию предсказателя, для которой ее кумулятивные потери L_T на шаге T связаны с кумуля-

Пусть $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ldots, N$.
FOR $t = 1, 2, \ldots$
Эксперт i анонсирует прогноз $\xi_t^i \in \Gamma$, $i = 1, \ldots, N$.
Предсказатель анонсирует прогноз $\gamma_t \in \Gamma$.
Природа анонсирует исход $\omega_t \in \Omega$.
Эксперт i вычисляет свои суммарные потери на шаге t игры:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i).$$

Предсказатель вычисляет свои суммарные потери на шаге t игры:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

Рис. 5.1: Протокол игры с использованием экспертных прогнозов

тивными потерями $L_T(\theta)$ эксперта θ следующим образом:

$$L_T \leqslant c(\eta) \inf_\theta L_T(\theta) + a(\eta) \ln N$$

для всех T, где в общем случае может быть $c(\eta) > 1$ для прозвольного значения параметра обучения $\eta \in (0, \infty)$.

В случае η-смешиваемой функции потерь $c(\eta) = 1$.

Предварительно рассмотрим схему алгоритма в случае множества исходов $\Omega = \{0, 1\}$ и конечного множества экспертов $\Theta = \{1, 2, \ldots, N\}$. Прогнозы могут принимать любые значения из множества Γ. Задана функция потерь $\lambda(\omega, \gamma)$, где $\omega \in \Omega$ и $\gamma \in \Gamma$.

В последующих разделах будут рассматриваться бесконечные (и даже несчетные) пространства экспертов Θ. При этом результаты существенно не изменятся, надо только ввести меры на экспертах и суммы по экспертам заменить на интегралы по θ.

Напомним протокол игры с предсказаниями с использованием экспертных прогнозов (рис. 5.1).

Анализ данного алгоритма аналогичен анализу подобных алгоритмов, приведенному в разделе 4.6. Для полноты изложения повторим часть этого анализа.

Фиксируем параметр обучения $\eta > 0$ (learning rate), полагаем $\beta = e^{-\eta}$. Введем некоторое априорное распределения $P_0(i)$ на множестве экспертов Θ. Естественно брать равномерное априорное распределение на экспертах $P_0(i) = 1/N$ для всех $i \in \Theta$, где N – число экспертов.

На шагах $t = 1, 2, \ldots$ *Предсказатель* перестраивает веса экспертов $i = 1, \ldots, N$ согласно формуле

$$P_t(i) = \beta^{\lambda(\omega_t, \xi_t^i)} P_{t-1}(i). \tag{5.8}$$

Таким образом, вес эксперта, имеющего большие потери, уменьшается.

Веса экспертов (5.8) нормируем:

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)}, \tag{5.9}$$

чтобы сумма нормированных весов стала равной 1.

Введем вспомогательную функцию, которая называется "псевдопредсказанием":

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i). \tag{5.10}$$

Алгоритм, выдающий псевдопредсказания, вычисленные по формуле (5.10), обозначаем APA (Aggregating Pseudo Algorithm). Обозначим суммарные потери алгоритма APA за T шагов на последовательности исходов $\omega_1, \ldots, \omega_T$:

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{5.11}$$

Следующая лемма представляет суммарные потери алгоритма APA в более простом и ясном виде.

Лемма 5.1. *Суммарные потери обобщенного алгоритма за T шагов могут быть представлены в виде*

$$L_T(APA) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i).$$

Доказательство. Из (5.8) следует, что

$$P_T(i) = \beta^{\sum\limits_{t=1}^{T} \lambda(\omega_t, \xi_t^i)} P_0(i) = \beta^{L_T(i)} P_0(i).$$

Из определения имеют место следующие равенства

$$\log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) - \log_\beta \sum_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i) =$$

$$= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_T(i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} =$$

$$= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i) + \lambda(\omega_T, \xi_T^i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} =$$

$$= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{\lambda(\omega_T, \xi_T^i)} P_{T-1}(i)}{\sum\limits_{i=1}^{N} P_{T-1}(i)} =$$

$$= \log_\beta \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P_{T-1}^*(j) = g_T(\omega_T). \qquad (5.12)$$

Последнее равенство следует из определения (5.10). Поскольку (5.12) имеет место для всех T, получаем утверждение леммы $L_T(APA) = \sum\limits_{t=1}^{T} g_t(\omega_t) = \log_\beta \sum\limits_{i=1}^{N} \beta^{L_T(i)} P_0(i)$. $\triangle$

Псевдопредсказание $g_t(\omega)$ представляет собой некоторые усредненные потери и не дает самого предсказания $\gamma \in \Gamma$, для которого предназначены эти потери.

В некоторых случаях можно перевести псевдопредсказание в обычное предсказание. *Функцией подстановки* называется функция $\gamma_t = \Sigma(g_t)$, такая, что $\lambda(\omega, \Sigma(g_t)) \leqslant g_t(\omega)$ для всех ω.

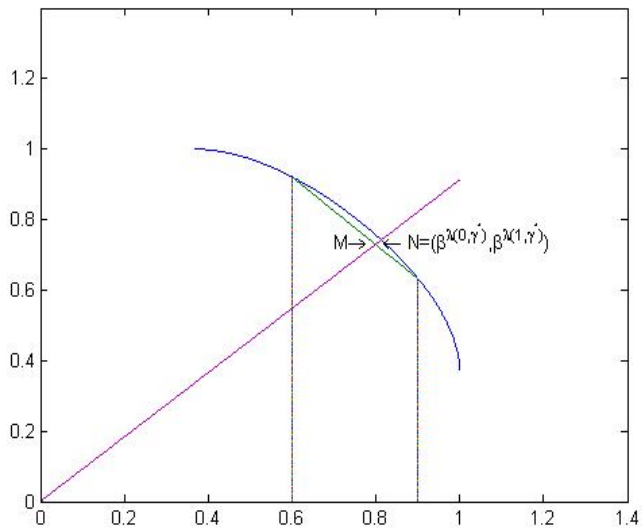

Рис. 6.5. Пример определения предсказания γ^*. Прямая, проходящая через точку M, отмечает точку $N = (\beta^{\lambda(0,\gamma^*)}, \beta^{\lambda(1,\gamma^*)})$ на кривой, по которой вычисляется предсказание γ^*

Мы покажем, что функция подстановки существует, если функция потерь $\lambda(\omega, \gamma_t)$ является смешиваемой.

Предложение 5.2. *Если функция потерь является смешиваемой, то функция подстановки существует.*

Доказательство. Пусть функция потерь $\lambda(\omega, \gamma)$ является η-смешиваемой и пусть $\beta = e^{-\eta}$. Из выпуклости образа

$$E_\eta(\Sigma_\lambda) = \{(x, y) : \exists\ p\ (0 \leqslant x \leqslant \beta^{\lambda(0,p)} 0 \leqslant y \leqslant \beta^{\lambda(1,p)})\}$$

в экспоненциальном пространстве множества суперпредсказаний

функции $\lambda(\omega, \gamma)$ следует, что существует $\gamma^* \in \Gamma$ такая, что

$$\beta^{\lambda(\omega_T, \gamma^*)} \geqslant \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P_{T-1}^*(j) \tag{5.13}$$

для всех $\omega_T \in \{0, 1\}$. Неравенство (5.13) означает, что абсцисса и ордината точки

$$\left(\beta^{\lambda(0,\gamma^*)}, \beta^{\lambda(1,\gamma^*)}\right)$$

больше или равны чем абсцисса и ордината точки

$$\left(\sum_{j=1}^{N} \beta^{\lambda(0,\xi_T^j)} P_{T-1}^*(j), \sum_{j=1}^{N} \beta^{\lambda(1,\xi_T^j)} P_{T-1}^*(j)\right).$$

Полагаем $\Sigma(g_t) = \gamma^*$. Условие $\lambda(\omega, \Sigma(g_t)) \leqslant g_t(\omega)$ будет выполнено для всех ω, т.е, функция $\Sigma(\cdot)$ является функцией подстановки.

Если функция потерь $\lambda(\omega, \gamma)$ вычислима некоторым алгоритмом, то также существует алгоритм, который на шаге t выдает предсказание $\gamma_t = \Sigma(g_t)$. Этот алгоритм называется *агрегирующим алгоритмом* (AA-алгоритм, Aggregating Algorithm).

Для некоторых функций потерь может существовать много различных γ^*, удовлетворяющих неравенству (5.13). В последующих разделах будут представлены конкретные аналитические выражения для функции $\Sigma(g_t)$ в случае логарифмической и квадратичной функций потерь.

В том случае, когда $\Sigma(g_t)$ существует, из леммы 5.1 следует, что будет иметь место неравенство

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant$$
$$\leqslant L_T(APA) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i). \tag{5.14}$$

Припишем каждому эксперту одинаковый начальный вес $P_0(i) = 1/N$. Тогда из (5.14) следует, что для произвольного $i \in \Theta$, для

всех T

$$L_T(AA) \leqslant \log_\beta \left(\frac{1}{N} \sum_{i=1}^{N} \beta^{L_T(i)} \right) \leqslant$$

$$\leqslant \log_\beta \left(\frac{1}{N} \beta^{L_T(i)} \right) = L_T(i) + \frac{\ln N}{\eta}. \tag{5.15}$$

Оценка (5.15) означает, что суммарные потери агрегирующего алгоритма AA не превосходят потери любого эксперта, в том числе и наилучшего, т.е., имеющего наименьшие потери среди всех экспертов, плюс некоторый регрет (ошибка обучения), который зависит только от числа экспертов и параметра и, что очень важно, не зависит от длины периода предсказания, как это было в алгоритме экспоненциального взвешивания.

5.3. Бесконечное множество экспертов

Повторим схему алгоритма в случае бесконечного множества экспертов Θ. Мы предполагаем, что на Θ задана структура вероятностного пространства – сигма алгебра борелевских множеств. Это позволяет рассматривать меры на Θ. В этом случае суммы по экспертам $i = 1, \ldots, N$ заменяются на интегралы по этим мерам на Θ.

По-прежнему $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$. Задана функция потерь $\lambda(\omega, \gamma)$, где $\omega \in \Omega$ и $\gamma \in \Gamma$, $\eta > 0$ – параметр обучения, $\beta = e^{-\eta}$.

Пусть задано некоторое априорное вероятностное распределение $P_0(d\theta)$ на множестве экспертов Θ. Предполагаем, что прогноз ξ_t^θ эксперта θ на шаге t есть измеримая функция от θ.

На шаге $t = 1, 2, \ldots$ *Предсказатель* перестраивает веса экспертов в соответствии с формулой

$$P_t(d\theta) = \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(d\theta). \tag{5.16}$$

Таким образом, вес эксперта, имеющего большие потери, уменьшается. По определению задание весов (5.16) эквивалентно способу вычисления вероятностей событий E по формуле

$$P_t(E) = \int_E \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(d\theta).$$

Веса (5.16) нормируем:

$$P_t^*(d\theta) = \frac{P_t(d\theta)}{P_t(\Theta)}. \tag{5.17}$$

Нормированные веса представляют собой вероятностную меру; для нее $P_t^*(\Theta) = 1$.

Аналогичным образом введем "псевдопредсказание"

$$g_t(\omega) = \log_\beta \int_\Theta \beta^{\lambda(\omega,\xi_t^\theta)} P_{t-1}^*(d\theta). \tag{5.18}$$

Алгоритм, выдающий псевдопредсказания, также обозначается APA, а суммарные потери алгоритма APA за T шагов равны

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{5.19}$$

Из (5.16) следует, что

$$P_T(d\theta) = \beta^{\sum\limits_{t=1}^{T} \lambda(\omega_t,\xi_t^\theta)} P_0(d\theta) = \beta^{L_T(\theta)} P_0(d\theta),$$

$$P_T^*(d\theta) = \frac{\beta^{L_T(\theta)}}{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)} P_0(d\theta).$$

Тогда равенство (5.18) переписывается в виде

$$g_T(\omega) = \log_\beta \int_\Theta \frac{\beta^{\lambda(\omega,\xi_T^\theta)+L_{T-1}(\theta)}}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} P_0(d\theta). \tag{5.20}$$

Имеет место аналог леммы 5.1.

Лемма 5.2. *Суммарные потери обобщенного алгоритма за T шагов могут быть представлены в виде*

$$L_T(APA) = \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{5.21}$$

Доказательство. Доказательство леммы аналогично доказательству леммы 5.1. Из (5.8) следует, что

$$P_t(d\theta) = \beta^{\sum\limits_{t=1}^{T} \lambda(\omega_t, \xi_t^\theta)} P_0(d\theta) = \beta^{L_T(\theta)} P_0(d\theta). \tag{5.22}$$

Из определения имеют место следующие равенства:

$$\log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta) - \log_\beta \int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta) =$$

$$= \log_\beta \frac{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} =$$

$$= \log_\beta \frac{\int_\Theta \beta^{L_{T-1}(\theta) + \lambda(\omega_T, \xi_T^\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} =$$

$$= \log_\beta \frac{\int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}(d\theta)}{\int_\Theta P_{T-1}(d\theta)} =$$

$$= \log_\beta \int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}^*(d\theta) = g_T(\omega_T). \tag{5.23}$$

Последнее равенство следует из определения (5.18).

Поскольку (5.23) имеет место для всех T, получаем утверждение леммы. $\triangle$

Для смешиваемой функции потерь нетрудно показать, что функция подстановки $\Sigma(g_t)$ также существует и в случае бесконечного пространства экспертов Θ. Действительно, интегралы по $d\theta$ приближаются конечными суммами, которые соответствуют конечным множествам экспертов. Предсказания, соответствующие этим конечным множества экспертов, имеют предельную точку γ^*, так как множество предсказаний компактно. Так как функция потерь $\lambda(\omega, \gamma)$ непрерывна по γ, эта предельная точка будет удовлетворять условию

$$\lambda(\omega, \gamma^*) \leqslant g_t(\omega)$$

для всех ω, где $g_t(\omega)$ определено по (5.18). Полагаем $\Sigma(g_t) = \gamma^*$.

В этом случае по лемме 5.2 будет иметь место неравенство

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \qquad (5.24)$$

5.4. Произвольная функция потерь

В последующих разделах мы покажем, что логарифмическая и квадратичная функции потерь являются смешиваемыми.

В общем случае, когда функция потерь не является смешиваемой, определяется *кривая смешиваемости* (mixability curve) $c(\eta)$:

$$c(\eta) = \inf \left\{ c : \forall\ P\ \exists\ \delta \in \Gamma\ \forall\ \omega \left(\lambda(\omega, \delta) \leqslant c \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} dP(\gamma) \right) \right\}.$$

При некоторых естественных предположениях на исходные множества функция $c(\eta)$ является непрерывной и невозрастающей.

В этом случае функция подстановки определяется как функция, удовлетворяющая

$$\forall\ \omega : \lambda(\omega, \Sigma_\eta(g)) \leqslant c(\eta) g(\omega) \qquad (5.25)$$

для любого псевдопредсказания

$$g(\omega) = \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} dP(\gamma)$$

и вероятностного распределения P на Γ.

Можно определить *минимаксную* функцию подстановки.

$$\Sigma_\eta(g) \in \arg\min_{\gamma\in\Gamma} \sup_{\omega\in\Omega} \frac{\lambda(\omega,\gamma)}{g(\omega)}. \qquad (5.26)$$

По определению любая минимаксная функция подстановки $\Sigma_\eta(g)$, удовлетворяющая (5.26), удовлетворяет и неравенству (5.25).

Заметим, что могут существовать другие – не минимаксные, функции подстановки такие, что выполнено условие (5.25). Часто их проще вычислить.

В общем случае, для произвольной функции потерь, не обязательно смешиваемой, вместо (5.24) имеем

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma_\eta(g_t)) \leqslant c(\eta) \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \quad (5.27)$$

Все аналогичные неравенства будут верными, если ввести в них множитель $c(\eta)$.

В случае конечного числа экспертов неравенство (5.15) переходит в неравенство

$$L_T(AA) \leqslant c(\eta) \log_\beta \left(\frac{1}{N} \sum_{i=1}^{N} \beta^{L_T(i)} \right) \leqslant$$

$$\leqslant c(\eta) \log_\beta \left(\frac{1}{N} \beta^{L_T(i)} \right) = c(\eta) L_T(i) + c(\eta) \frac{\ln N}{\eta}$$

для всех T и всех $i = 1, \ \ldots, \ N$.

5.5. Логарифмическая функция потерь

Пусть множество всех исходов Ω и множество всех экспертов Θ – конечные, множество всех прогнозов $\Gamma = \mathcal{P}(\Omega)$ – множество всех вероятностных распределений на Ω. При $\gamma \in \Gamma$ и $\omega \in \Omega$, величина $\gamma(\omega) = \gamma(\{\omega\})$ равна вероятности элемента ω. *Логарифмическая функция потерь* определяется $\lambda(\omega, \gamma) = -\ln \gamma(\omega)$.

Возьмем $\eta = 1$, тогда $\beta = e^{-1}$. В этом случае

$$\beta^{\lambda(\omega,\gamma)} = \gamma(\omega),$$

т.е. равно вероятности, которую эксперт или *Предсказатель* приписывает исходу ω. В этом случае агрегирующий алгоритм совпадает с алгоритмом экспоненциального взвешивания.

Прогноз эксперта i на шаге t – это распределение вероятностей $\xi_t^i = \xi_t^i(\cdot) \in \Gamma$ на пространстве исходов Ω.

С каждым экспертом $i \in \Theta$ на шаге t будем связывать распределение вероятностей Q_i на Ω, определяемое условными вероятностями:

$$Q_i(\omega | \omega_1, \ \ldots, \ \omega_{t-1}) = \xi_t^i(\omega) \in \Gamma. \quad (5.28)$$

Такое распределение можно интерпретировать как субъективное условное распределение эксперта i на t-м шаге. Величина (5.28) равна условной вероятности, которую i-й эксперт приписывает будущему исходу ω, после того как он наблюдал исходы $\omega_1, \ldots, \omega_{t-1}$.

Тогда субъективная вероятность, которая приписывается на шаге t экспертом i последовательности исходов $\omega_1, \ldots, \omega_t$ равна произведению

$$Q_i(\omega_1, \ldots, \omega_t) = \xi_1^i(\omega_1)\xi_2^i(\omega_2) \cdot \ \ldots \ \cdot \xi_t^i(\omega_t). \tag{5.29}$$

Веса экспертов перестраиваются согласно (5.8). В данном случае вес i-го эксперта переопределяется на шаге t :

$$\begin{aligned} P_t(i) = \beta^{\lambda(\omega_t, \xi_t^i)} P_{t-1}(i) = \\ = \xi_1^i(\omega_1)\xi_2^i(\omega_2) \cdot \ \ldots \ \cdot \xi_t^i(\omega_t) P_0(i) = \\ = Q_i(\omega_1, \ldots, \omega_t) P_0(i). \end{aligned} \tag{5.30}$$

Веса (5.30) экспертов нормируются как

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)} = \frac{Q_i(\omega_1, \ldots, \omega_t) P_0(i)}{\sum\limits_{j=1}^{N} Q_j(\omega_1, \ldots, \omega_t) P_0(j)}. \tag{5.31}$$

Вероятность $P_t^*(i)$ представляет собой апостериорную вероятность эксперта i после наблюдения исходов $\omega_1, \ldots, \omega_t$.

Так как $\beta^{\lambda(\omega, \xi_t^i)} = \xi_t^i(\omega)$, псевдопредсказание (5.10) превращается в логарифм байесовской смеси распределений, предлагаемых на шаге t экспертами,

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \xi_t^i(\omega) P_{t-1}^*(i). \tag{5.32}$$

Возьмем в качестве предсказания $\Sigma(g_t)$ алгоритма AA распределение вероятностей, которое представляет собой байесовскую смесь распределений, предлагаемых на шаге t экспертами.

Распределение вероятностей – предсказание алгоритма AA, определяется как

$$\gamma_t(\omega) = \Sigma(g_t) = \sum_{i=1}^{N} \xi_t^i(\omega) P_{t-1}^*(i).$$

Тогда значение логарифмической функции потерь на исходе ω_t при предсказании *Предсказатель*, равном распределению γ_t, просто равно псевдопредсказанию

$$\lambda(\omega_t, \gamma_t) = -\ln \gamma_t(\omega_t) = \log_\beta \sum_{i=1}^{N} \xi_t^i(\omega_t) P_{t-1}^*(i) = g_t(\omega_t).$$

Разьясним данный метод и его связь с *байесовским правилом* на примере первых двух шагов: $t = 1, 2$.

Каждому эксперту i на шаге $t = 1$ соответствует его прогноз – распределение вероятностей $\xi_1^i = \xi_1^i(\cdot) \in \Gamma$ на Ω. На первом шаге предсказание алгоритма AA –

$$\gamma_1(\omega) = \sum_{i=1}^{N} \xi_1^i(\omega) P_0(i),$$

представляет собой байесовскую смесь вероятностных распределений экспертов относительно априорного распределения P_0 на множестве всех экспертов.

После того как появился первый исход ω_1, *Предсказатель* перестраивает априорное распределение на множестве экспертов. Сначала он определяет веса экспертов:

$$P_1(i) = \beta^{\lambda(\omega_1, \xi_1^i)} P_0(i) = \xi_1^i(\omega_1) P_0(i).$$

После этого путем нормирования весов вычисляются апостериорные вероятности экспертов i после наблюдения исхода ω_1 :

$$P_1^*(i) = \frac{\xi_1^i(\omega_1) P_0(i)}{\sum\limits_{j=1}^{N} \xi_1^j(\omega_1) P_0(j)}.$$

Нетрудно заметить, что данная формула представляет собой формулу Байеса для вычисления апостериорной вероятности $P_1^*(i)$ эксперта i после наблюдения исхода ω_1.

Аналогичным образом поступаем на шаге $t = 2$.

Каждому эксперту i на шаге $t = 2$ соответствует его прогноз – распределение вероятностей $\xi_2^i(\cdot)$ на Ω. Предсказание алгоритма АА

$$\gamma_2(\omega) = \sum_{i=1}^{N} \xi_2^i(\omega) P_1^*(i)$$

представляет собой байесовскую смесь вероятностных распределений экспертов относительно апостериорного распределения P_1^* на множестве всех экспертов, построенного на основе исхода, полученного на предыдущем шаге.

После того как появился второй исход ω_2, *Предсказатель* перестраивает апостериорное распределение на множестве экспертов. Сначала он переопределяет веса экспертов

$$P_2(i) = \beta^{\lambda(\omega_2, \xi_2^i)} P_1(i) = \xi_2^i(\omega_2) P_1(i) = \xi_1^i(\omega_1) \xi_2^i(\omega_2) P_0(i).$$

После этого путем нормирования весов вычисляются апостериорные вероятности экспертов i после наблюдения исходов ω_1, ω_2 :

$$P_2^*(i) = \frac{\xi_1^i(\omega_2) P_1^*(i)}{\sum\limits_{j=1}^{N} \xi_2^j(\omega_2) P_1^*(j)}.$$

Вновь нетрудно заметить, что последняя часть равенства представляет собой формулу Байеса для вычисления апостериорной вероятности $P_2^*(i)$ эксперта i после наблюдения исхода ω_2 на основе предыдущих апостериорных вероятностей $P_1^*(i)$, вычисленных на предыдущем шаге.

Таким образом, в случае логарифмической функции потерь алгоритм АА представляет собой последовательное применение байесовского правила в режиме онлайн.

Потери i-го эксперта за T шагов равны

$$\begin{aligned} L_T(i) = \sum_{i=1}^{T} \lambda(\omega_t, \xi_t^i) = \\ = -\ln(\xi_1^i(\omega_1) \cdot \ \ldots \ \cdot \xi_T^i(\omega_T)) = \\ = -\ln Q_i(\omega_1, \ldots, \ \omega_T). \end{aligned} \tag{5.33}$$

Это равенство использует определение субъективной вероятности (5.28), которую *Предсказатель* приписывает экспертам на шаге t.

Потери *Предсказателя*, использующего алгоритм AA, за T шагов равны

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) =$$
$$= \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) =$$
$$= \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \ldots, \omega_T) P_0(i). \tag{5.34}$$

Таким образом, потери *Предсказателя* за T шагов, использующего алгоритм AA, равны минус логарифму от байесовской смеси всех вероятностей, которые эксперты приписывают последовательности исходов $\omega_1, \ldots, \omega_T$ длины T.

Неравенство (5.15) превращается в неравенство

$$L_T(AA) = \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \ldots, \omega_T) P_0(i) \leqslant$$
$$\leqslant -\ln Q_k(\omega_1, \ldots, \omega_T) - \ln P_0(k) \tag{5.35}$$

для всех T и $k = 1, \ldots, N$.

5.6. Простая игра на предсказания

Напомним, что простая игра на предсказания рассматривается в случае, когда пространство исходов и пространство прогнозов двухэлементные и совпадают: $\Omega = \Gamma = \{0, 1\}$. Задача предсказания заключается в том, чтобы точно предсказать будущий исход. Функция потерь определяется

$$\lambda(\omega, \gamma) = \begin{cases} 0, \text{ если } \omega = \gamma, \\ 1 \text{ в противном случае.} \end{cases}$$

Таким образом, кумулятивные потери эксперта равны числу ошибок при предсказании будущего исхода.

Имеется N экспертов; эксперт i делает на шаге t предсказание $\xi_t^i \in \{0,1\}$.

Для анализа этой игры каждое псевдопредсказание

$$g(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega,\xi_t^i)} P_{t-1}^*(i) \tag{5.36}$$

представляется в виде точки $(g(0), g(1))$ на координатной плоскости $\mathcal{R}^2$. Эта точка имеет вид

$$(\log_\beta(\beta p + (1-p)),\ \log_\beta(p + \beta(1-p))), \tag{5.37}$$

где $0 < \beta < 1$ – параметр смешивания, $p = \sum\limits_{\xi_t^i=1} P_{t-1}^*(i)$ – суммарный вес экспертов, предсказывающих 1 на шаге t, при этом $1-p = \sum\limits_{\xi_t^i=0} P_{t-1}^*(i)$ – суммарный вес экспертов, предсказывающих 0 на шаге t.

Все точки типа (5.37) образуют выпуклую кривую, соединяющую точки $(1,0)$ и $(0,1)$, которые соответствуют $p=0$ и $p=1$.

По определению $1/c(\beta)$ равно абсциссе (ординате) точки пересечения прямой $y=x$ и этой кривой. При $p=\frac{1}{2}$ из (5.37) получаем

$$\frac{1}{c(\beta)} = \log_\beta \left(\frac{1+\beta}{2}\right),$$

или

$$c(\beta) = \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}}. \tag{5.38}$$

Применим алгоритм AA к этой игре.

Определим функцию подстановки $\gamma = \Sigma(g)$ следующим образом: $\Sigma(g) = 0$, если точка $(g(0), g(1))$, вычисленная по (5.37), лежит выше прямой $y=x$, $\gamma = \Sigma(g) = 1$, если точка $(g(0), g(1))$ лежит ниже или на прямой $y=x$.

Эта функция подстановки удовлетворяет условию (5.25), так как при $\gamma = 0$ будет абсцисса $g(0) \geqslant \lambda(0,0) = 0$ и ордината $g(1)$ больше ординаты $\frac{1}{c(\beta)}$ точки пересечения биссектрисы координатного угла и кривой (5.37). Таким образом, $g(1) \geqslant \frac{1}{c(\beta)} = \frac{1}{c(\beta)}\lambda(1,0)$. Поэтому $\lambda(\omega,0) \leqslant c(\beta)g(\omega)$ при всех $\omega \in \{0,1\}$.

Аналогичным образом получаем неравенство $\lambda(\omega,1) \leqslant c(\beta)g(\omega)$ при всех $\omega \in \{0,1\}$.

Заметим, что если точка $(g(0), g(1))$ лежит выше прямой $y = x$, то абсцисса меньше ординаты, т.е. $g(0) < g(1)$, или

$$\log_\beta(\beta p + (1-p)) < \log_\beta(p + \beta(1-p)),$$

что эквивалентно $p < \frac{1}{2}$. В этом случае алгоритм предсказывает $\gamma = 0$.

В противном случае, т.е. если точка $(g(0), g(1))$ лежит ниже (или на) прямой $y = x$, то

$$\log_\beta(\beta p + (1-p)) \geqslant \log_\beta(p + \beta(1-p)),$$

что эквивалентно $p \geqslant \frac{1}{2}$. В этом случае алгоритм предсказывает $\gamma = 1$.

Это означает, что алгоритм AA предсказывает 1, если суммарный вес экспертов, предсказывающих 1, больше суммарного веса экспертов, предсказывающих 0; алгоритм AA предсказывает 0 в противоположном случае. Таким образом, алгоритм AA предсказывает как взвешенное большинство экспертов. Этот алгоритм был описан в разделе (4.1).

В этом случае для любого эксперта $\theta \in \Theta$ будет иметь место неравенство

$$L_T(AA) \leqslant \left(\frac{\ln\frac{1}{\beta}}{\ln\frac{2}{1+\beta}}\right) L_T(\theta) + \ln\left(\frac{1}{\ln\frac{2}{1+\beta}}\right)\ln N. \tag{5.39}$$

5.7. Игра с квадратичной функцией потерь

Изучим игру с квадратичной функцией потерь в простейшем случае, когда пространство исходов двухэлементное: $\Omega = \{-1, 1\}$,

пространство прогнозов это все действительные числа из $[-1,1]$. Функция потерь – квадрат разности между исходом и прогнозом $\lambda(\omega,\gamma)=(\omega-\gamma)^2$.

Мы рассматриваем случай $\Omega=\{-1,1\}$, поскольку доказательства в этом случае проще. Все приведенные ниже утверждения также верны и для случая $\Omega=[-1,1]$.

Лемма 5.3. *Квадратичная функция потерь является η-смешиваемой тогда и только тогда, когда $\eta\leqslant\frac{1}{2}$.*

Доказательство. Представим псевдопредсказание $(g(-1),g(1))$ точкой в экспоненциальном пространстве:

$$(e^{-\eta g(-1)},e^{-\eta g(1)}).$$

Множеству всех предсказаний $\gamma\in[-1,1]$ соответствует параметризованная кривая в экспоненциальном пространстве

$$(x(\gamma),y(\gamma))=(e^{-\eta(-1-\gamma)^2},e^{-\eta(1-\gamma)^2}).$$

Функция потерь будет η-смешиваемой, если образ множества суперпредсказаний в экспоненциальном пространстве является выпуклым множеством, т.е. тогда и только тогда, когда ограничивающая его кривая поворачивает налево при возрастании γ (при этом абсцисса уменьшается). Это будет в случае, если выполнено условие вогнутости кривой: $\frac{d^2y}{d^2x}\leqslant 0$.

Вычислим вторую производную параметрически заданной кривой

$$\frac{d^2y}{d^2x}=\frac{d\gamma}{dx}\frac{x'(\gamma)y''(\gamma)-x''(\gamma)y'(\gamma)}{(x'(\gamma))^2}. \tag{5.40}$$

При возрастании параметра γ величина $x(\gamma)$ убывает, поэтому $\frac{d\gamma}{dx}<0$. Игра будет η-перемешиваемой тогда и только тогда, когда $\frac{d^2y}{d^2x}\leqslant 0$, что равносильно условию

$$x'(\gamma)y''(\gamma)-x''(\gamma)y'(\gamma)\geqslant 0.$$

Вычислим производные по параметру:

$$\begin{aligned} x'(\gamma) &= -2\eta(1+\gamma)e^{-\eta(1+\gamma)^2}, \\ x''(\gamma) &= 2\eta(-1+2\eta(1+\gamma)^2)e^{-\eta(1+\gamma)^2}, \\ y'(\gamma) &= 2\eta(1-\gamma)e^{-\eta(1-\gamma)^2}, \\ y''(\gamma) &= 2\eta(-1+2\eta(1-\gamma)^2)e^{-\eta(1-\gamma)^2}. \end{aligned} \tag{5.41}$$

Тогда условие η-смешиваемости требует, чтобы для всех значений $\gamma \in [-1, 1]$

$$\begin{aligned} -(1+\gamma)(-1+2\eta(1-\gamma)^2) - \\ -(1-\gamma)(-1+2\eta(1+\gamma)^2) \geqslant 0, \\ \eta(1-\gamma^2) \leqslant \frac{1}{2}, \\ \eta \leqslant \frac{1}{2}. \end{aligned} \tag{5.42}$$

Лемма доказана. $\triangle$

Найдем теперь вид какой-нибудь функции подстановки $\Sigma(g)$ в случае $\Omega = \{-1, 1\}$ и конечного числа экспертов $\Theta = \{1, \ \ldots, \ N\}$. Пусть $\eta = \frac{1}{2}$, $\beta = e^{-\frac{1}{2}}$.

Произвольное псевдопредсказание

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i) \tag{5.43}$$

задается точкой

$$\begin{aligned} (e^{-\frac{1}{2}g_t(-1)}, e^{-\frac{1}{2}g_t(1)}) = \\ = \left(\sum_{i=1}^{N} \beta^{\lambda(-1, \xi_t^i)} P_{t-1}^*(i), \sum_{i=1}^{N} \beta^{\lambda(1, \xi_t^i)} P_{t-1}^*(i) \right), \end{aligned} \tag{5.44}$$

которая расположена под вогнутой кривой

$$\left(\beta^{\lambda(-1,\gamma)}, \beta^{\lambda(1,\gamma)} \right), \tag{5.45}$$

при $\gamma \in [-1.1]$.

Проведем прямую, проходящую через начало координат и точку (5.44). Коэффициент наклона этой прямой равен

$$k = \frac{\beta^{g_t(1)}}{\beta^{g_t(-1)}} = e^{\frac{1}{2}g_t(-1)-\frac{1}{2}g_t(1)}. \tag{5.46}$$

Точка пересечения $\left(\beta^{\lambda(-1,\gamma^*)}, \beta^{\lambda(1,\gamma^*)}\right)$ этой прямой и кривой (5.45) имеет абсциссу и ординату по величине не меньше, чем абсцисса и ордината точки (5.44) :

$$\begin{aligned} \beta^{\lambda(-1,\gamma^*)} &\geqslant \beta^{g_t(-1)}, \\ \beta^{\lambda(1,\gamma^*)} &\geqslant \beta^{g_t(1)}. \end{aligned} \tag{5.47}$$

Эквивалентная запись (5.47) имеет вид

$$\begin{aligned} \lambda(-1,\gamma^*) &\leqslant g_t(-1), \\ \lambda(1,\gamma^*) &\leqslant g_t(1). \end{aligned} \tag{5.48}$$

Вычислим предсказание γ^*. Значение γ^* находим из уравнения

$$\frac{\beta^{g_t(1)}}{\beta^{g_t(-1)}} = \beta^{g_t(1)-g_t(-1)} = \frac{\beta^{\lambda(1,\gamma^*)}}{\beta^{\lambda(-1,\gamma^*)}} = \beta^{\lambda(1,\gamma^*)-\lambda(-1,\gamma^*)}. \tag{5.49}$$

Остается найти корень уравнения

$$\lambda(1,\gamma^*) - \lambda(-1,\gamma^*) = (1-\gamma^*)^2 - (-1-\gamma^*)^2 = g_t(1) - g_t(-1),$$

который равен

$$\gamma^* = \frac{1}{4}(g_t(-1) - g_t(1)). \tag{5.50}$$

Более детально на шаге t выбирается предсказание

$$\gamma_t^* = \frac{1}{4}\left(\log_\beta \sum_{i=1}^{N} \beta^{\lambda(-1,\xi_t^i)} P_{t-1}^*(i) - \log_\beta \sum_{i=1}^{N} \beta^{\lambda(1,\xi_t^i)} P_{t-1}^*(i)\right)$$

или

$$\gamma_t^* = \frac{1}{2}\ln\left(\frac{\sum\limits_{i=1}^{N} e^{-\frac{1}{2}(1-\xi_t^i)^2} P_{t-1}^*(i)}{\sum\limits_{i=1}^{N} e^{-\frac{1}{2}(1+\xi_t^i)^2} P_{t-1}^*(i)}\right).$$

Аналогичные свойства и утверждения имеют место и для множества $\Omega = [-1, 1]$ (см. [61]).

Для бесконечного множества исходов Ω геометрическое определение смешиваемой функции потерь не имеет смысла. В этом случае можно ввести более общее (прямое) определение смешиваемости. Функция потерь называется η-смешиваемой, если существует функция подстановки $\Sigma(g_t)$ такая, что

$$\lambda(\omega, \Sigma(g_t)) \leqslant g_t(\omega)$$

для всех $\omega \in \Omega$, где g_t определена по (5.18).

5.8. Метод защитных предсказаний и агрегирующий алгоритм

.

В этом разделе излагается метод "защитных" предсказаний (defensive forecasts), предложенный Вовком и Такемурой (см. работы, [64], [66], [69], [70], [20]). Рассматривается протокол игры на предсказания с участниками *Природа* и *Предсказатель*, представленный на рис. 5.1.

Обозначим $(p, \omega)^n = p_1, \omega_1, \ldots, p_n, \omega_n$, где $\omega_i \in \{0, 1\}$, $p_i \in [0, 1]$ при $1 \leqslant i \leqslant n$. Полагаем $(p, \omega)^0 = \lambda$. Последовательность функций $M_n((p, \omega)^n)$, $n = 0, 1, \ldots$, принимающих неотрицательные вещественные значения, называется теоретико-игровым супермартингалом с переменными вероятностями, если выполнены условия

- $M_0(\lambda) \leqslant 1$;
- для всех n, p_i, ω_i при $1 \leqslant i \leqslant n-1$ и всех p.

$$pM_n((p, \omega)^{n-1}, p, 1) + (1-p)M_n((p, \omega)^{n-1}, p, 0) \leqslant \\ \leqslant M_{n-1}((p, \omega)^{n-1}). \qquad (5.51)$$

Супермартингал M_n называется непрерывным по прогнозу, если для любого n функция $M_n((p, \omega)^n) = M_n((p, \omega)^{n-1}, p_n, \omega_n)$ непрерывна по переменной p_n при любых фиксированных значениях остальных параметров.

Теорема 5.1. *Для любого непрерывного по прогнозу супермартингала M_n можно так выбирать прогнозы $p_1, p_2, \ldots$, что выполнено $M_{n-1}((p,\omega)^{n-1}) \geqslant M_n((p,\omega)^n)$ для всех n, как бы Природа не выбирала исходы $\omega_1, \omega_2 \ldots$.*

Доказательство. Определим

$$f_n(\omega, p) = M_n((p,\omega)^{n-1}, p, \omega) - M_{n-1}((p,\omega)^{n-1}),$$

где $p \in [0,1]$, $\omega \in \{0,1\}$.

Для любого n функция $f_n(\omega, p)$ непрерывна по p. Нетрудно проверить, для нее также выполнено неравенство $pf_n(1,p) + (1-p)f_n(0,p) \leqslant 0$ для всех p. В частности, $f_n(0,0) \leqslant 0$ и $f_n(1,1) \leqslant 0$.

Докажем, что существует p^* такое, что $f_n(\omega, p^*) \leqslant 0$ для всех $\omega \in \{0,1\}$. Если $f_n(1,0) \leqslant 0$, то полагаем $p^* = 0$. Если $f_n(0,1) \leqslant 0$, то полагаем $p^* = 1$. Остается рассмотреть случай $f_n(1,0) > 0$ и $f_n(0,1) > 0$. В этом случае, разность $f_n(p) = f_n(1,p) - f_n(0,p)$ положительна при $p = 0$ отрицательна при $p = 1$. По теореме о промежуточном значении $f_n(p^*) = 0$ для некоторого $p^* \in (0,1)$. Отсюда следует, что $f_n(\omega, p^*) \leqslant 0$ для $\omega = 0$ и $\omega = 1$. $\triangle$.

Частным случаем этой теоремы является теорема 3.5 из раздела 3.4.

Приведем конкретные примеры теоретико-игровых супермартингалов.

Рассмотрим игру на предсказания с η-смешиваемой функцией потерь $\lambda(\omega, p)$ и с протоколом, представленным на рис. 5.1. Пусть $\{1, \ldots, N\}$ – множество экспертов, ξ_t^i – прогноз эксперта i, p_t – прогноз *Предсказателя* на шаге t, $\omega_t \in \{0,1\}$ – исход, представленной *Природой* на шаге t.

Введем обозначения $r_{i,t}(\omega, p) = \lambda(\omega, p) - \lambda(\omega_t, \xi_t^i)$ – регрет алгоритма предсказания относительно эксперта i на шаге t при прогнозе предсказателя p и $\omega \in \{0,1\}$, $R_{i,n} = \sum\limits_{t=1}^{n} r_{i,t}(\omega_t, p_t)$ – соответствующий кумулятивный регрет за первые n шагов, где $p_1, \ldots, p_n$ – прогнозы предсказателя. Определим $M_n^i((\omega,p)^n) = \exp(\eta R_{i,n})$.

Приведем примеры функций потерь, для которых эта функция является теоретико-игровым супермартингалом, непрерывным относительно прогноза p_n.

Свойство (5.51) для $M_n^i((\omega,p)^n)$ можно записать в эквивалентном виде

$$pe^{(\eta r_{i,n}(1,p))} + (1-p)e^{(\eta r_{i,n}(0,p))} \leqslant 1. \qquad (5.52)$$

Логарифмическая функция потерь.

Пусть $\lambda(\omega,p) = -\ln p(\omega)$, где $p(1) = p$ и $p(0) = 1-p$, $0 < p < 1$. Эта функция является η-смешиваемой при $0 < \eta \leqslant 1$.

Проверим свойство (5.52):

$$pe^{\eta(-\ln p + \ln \xi_n^i)} + (1-p)e^{\eta(-\ln(1-p) + \ln(1-\xi_t^i))} \leqslant 1,$$

где $0 < \eta \leqslant 1$. Это неравенство перепишем в виде

$$p^{1-\eta}(\xi_n^i)^\eta + (1-p)^{1-\eta}(1-\xi_n^i)^\eta \leqslant 1.$$

Легко видеть, что это неравенство является следствием обобщенного неравенства между арифметическим и геометрическим средними: $a^\alpha b^{1-\alpha} \leqslant \alpha a + (1-\alpha)b$, где $a, b \geqslant 0$ и $0 \leqslant \alpha \leqslant 1$. Последнее неравенство следует из вогнутости логарифма.

Квадратичная функция потерь.

Пусть $\lambda(\omega,p) = (\omega - p)^2$, где $\omega \in \{0,1\}$ и $p \in [0,1]$. Эта функция является η-смешиваемой при $0 < \eta \leqslant 2$.

Проверим свойство (5.52):

$$pe^{\eta((p-1)^2 - (1-\xi_n^i)^2)} + (1-p)e^{\eta(p^2 - (\xi_t^i)^2)} \leqslant 1,$$

Обозначим $\xi_n^i = p + x$ и перепишем это неравенство в виде

$$pe^{2\eta(1-p)x} + (1-p)e^{-2\eta px} \leqslant e^{\eta x^2}.$$

Последнее неравенство является частным случаем неравенства Хефдинга $\ln Ee^{sX} \leqslant sE(X) + \frac{s^2(b-a)^2}{8}$, где X – случайная величина, для которой $a \leqslant X \leqslant b$ и $s \in \mathcal{R}$. Действительно, возьмем в качестве X случайную величину, которая равна 1 с вероятностью p и 0 с вероятностью $1-p$. Для нее при $a = 0$ и $b = 1$ будет $pe^{s(1-p)} + (1-p)e^{-sp} \leqslant e^{s^2/8}$. Положим $s = 2\eta x$ и получим $pe^{2\eta(1-p)x} + (1-p)e^{-2\eta px} \leqslant e^{\eta^2 x^2/2} \leqslant e^{\eta x^2}$, так как $\eta \leqslant 2$.

Ясно, что для логарифмической и квадратичной функций потерь супермартингал $M_n^i((\omega,p)^n)$ будет непрерывным по прогнозу p_n.

Сформулируем полученные результаты в виде теоремы.

Теорема 5.2. *Для логарифмической и квадратичной функций потерь функция* $M_n^i((\omega,p)^n) = \exp(\eta R_{i,n})$ *является непрерывным по прогнозу теоретико-игровым супермартингалом для любого* i.

Альтернативный алгоритм для вычисления прогнозов агрегирующего алгоритма AA.

Приведем теперь алгоритм для вычисления предсказаний агрегирующего алгоритма AA на основе метода защитных предсказаний.

Предварительно найдем некоторую величину, которая является инвариантом агрегирующего алгоритма – она не возрастает в процессе игры на предсказания как бы *Природа* не выбирала свои исходы. На основе этой величины будет построен теоретико-игровой супермартингал. Алгоритм, приведенный в теореме 5.1, будет давать предсказания, которые обеспечат регрет (5.15).

Напомним некоторые элементы анализа агрегирующего алгоритма AA из начальных разделов главы 5. Задано начальное распределение $P_0(i)$ на множестве всех экспертов $\{1,\dots,N\}$. Пусть $\eta > 0$ и задана η-смешиваемая функция потерь $\lambda(\omega,p)$.

Допустим, что на шаге T прогнозы $\gamma_1,\dots,\gamma_{T-1}$ уже определены, $\xi_1^i,\dots,\xi_T^i$ – прогнозы экспертов. Из результатов раздела 5 следует, что оценка (5.15) будет иметь место, если для любого T прогноз γ_T удовлетворяет на шаге T условию (5.13). Перепишем условие (5.13) в виде

$$\sum_{j=1}^N P_{T-1}(j) \geqslant \sum_{j=1}^N P_{T-1}(j) e^{-\eta\lambda(\omega,\xi_T^j)} e^{\eta\lambda(\omega,\gamma_T)} \tag{5.53}$$

при $\omega \in \{0,1\}$. Напомним, что $P_{T-1}(j) = P_0(j)\prod_{t=1}^{T-1} e^{-\eta\lambda(\omega_t,\xi_t^j)}$. Используем это равенство, а также умножим обе части неравенства (5.53) на $\prod_{t=1}^{T-1} e^{\eta\lambda(\omega_t,\gamma_t)}$ и получим после некоторых преобразований

$$\sum_{j=1}^N P_0(j) Q_{T-1}(j) \geqslant \sum_{j=1}^N P_0(j) Q_{T-1}(j) e^{\eta(\lambda(\omega,\gamma_T)-\lambda(\omega,\xi_T^j))}, \tag{5.54}$$

где

$$Q_{T-1}(j) = \exp\left(\eta\left(\sum_{t=1}^{T-1}(\lambda(\omega_t,\gamma_t) - \lambda(\omega_t,\xi_t^j)\right)\right). \qquad (5.55)$$

Также полагаем $Q_0(j) = 1$ для всех j.

Таким образом, условие (5.13) эквивалентно тому, что существует γ_T такое, что для всех ω выполнено

$$\sum_{j=1}^{N} P_0(j)\tilde{Q}_T(j) \leqslant \sum_{j=1}^{N} P_0(j)Q_{T-1}(j), \qquad (5.56)$$

где $\tilde{Q}_T(j)$ есть $Q_T(j)$, в котором ω_T заменено на ω.

Таким образом выбор прогнозов γ_t согласно алгоритму AA гарантирует, что величина $\sum\limits_{j=1}^{N} P_0(j)Q_T(j)$ не возрастает как бы *Природа* не выбирала свои исходы ω_t. Верно и обратное, если прогнозы выбираются так, что эта сумма не возрастает с ростом T, то условие (5.13) для этих прогнозов выполнено и оценка (5.15) регрета алгоритма AA имеет место.

Вспоминая определение кумулятивного регрета $R_{i,T}$ заметим, что $Q_T(i) = \exp(\eta R_{i,T})$. По теореме 5.2 для любого i функция

$$M_n^i((\omega,p)^n) = \exp(\eta R_{i,n})$$

является непрерывным по прогнозу теоретико-игровым супермартингалом для логарифмической и квадратичной функций потерь.

Легко проверить, что выпуклая комбинация супермартингалов также является супермартингалом. В частности, функция

$$M_n((\omega,p)^n) = \sum_{i=1}^{N} M_n^i((\omega,p)^n)P_0(i) =$$
$$= \sum_{i=1}^{N} \exp(\eta R_{i,n})P_0(i) = \sum_{i=1}^{N} P_0(i)Q_n(i)$$

также является супермартингалом.

Теорема 5.1 дает нам алгоритм для вычисления предсказаний γ_t, при которых величина супермартингала $M_t((\omega,p)^t)$ не возрастает и остается меньшим или равным 1. Иными словами, существует алгоритм вычисления прогнозов так, что для всех T

$$\sum_{i=1}^{N} P_0(i)Q_T(i) \leqslant \sum_{i=1}^{N} P_0(i)Q_{T-1}(i). \tag{5.57}$$

В частности, $\sum\limits_{i=1}^{N} P_0(i)Q_T(i) \leqslant 1$ для всех T. Из определения (5.55) следует, что

$$Q_T(i) = \exp\left(\eta(L_T - L_T^i)\right),$$

где $L_T = \sum\limits_{t=1}^{T} \lambda(\omega_t, \gamma_t)$ и $L_T^i = \sum\limits_{t=1}^{T} \lambda(\omega_t, \xi_t^i)$. Из неравенства

$$\sum_{i=1}^{N} P_0(i) \exp\left(\eta(L_T - L_T^i)\right) \leqslant 1$$

следует, что каждое слагаемое не превосходит 1, т.е.

$$P_0(i) \exp\left(\eta(L_T - L_T^i)\right) \leqslant 1.$$

Отсюда следует, что для любого i

$$L_T \leqslant L_T^i + \frac{1}{\eta} \log \frac{1}{P_0(i)}$$

для всех T как бы *Природа* не выбирала исходы. В случае равномерного начального распределения на множестве экспертов $P_0(i) = 1/N$ для всех i получаем

$$L_T \leqslant L_T^i + \frac{\ln N}{\eta}.$$

Дальнейшее обсуждение связи между свойством смешиваемости функции потерь и теоретико-игровыми супермартингалами см. в работах [20], [18]. Другой алгоритм подобного типа приведен в статье [26].

5.9. Универсальный портфель

Рассмотрим следующую игру Ковера [22]. Имеется N финансовых инструментов, например акций. Время разделено на интервалы $t = 1, 2, \ldots$ Поведение рынка в момент t характеризуется вектором изменений цен акций от момента $t-1$ к моменту t (вектором доходностей)

$$\bar{\omega}_t = (\omega_{1,t}, \ \ldots, \ \omega_{N,t}),$$

где

$$\omega_{i,t} = \frac{S_{i,t}}{S_{i,t-1}},$$

$S_{i,t}$ – цена акции i при закрытии интервала в момент t. По определению $\omega_{i,t} \in [0, \infty)$, причем считаем, что не все $\omega_{i,t}$ равны нулю.

Инвестиции в данные N финансовых инструментов характеризуются *портфелем* – вектором $\bar{\gamma}_t \in [0,1]^N$, где

$$\bar{\gamma}_t = (\gamma_{1,t}, \ \ldots, \ \gamma_{N,t})$$

и $\gamma_{1,t} + \cdots + \gamma_{N,t} = 1$. Величины $\gamma_{i,t}$ определяют пропорции, по которым текущая сумма денег вкладывается в финансовые инструменты в момент $t-1$. Тогда инвестиции, вложенные согласно портфелю $\bar{\gamma}_t$, увеличиваются в

$$(\bar{\gamma}_t \cdot \bar{\omega}_t) = \sum_{i=1}^{N} \gamma_{i,t}\omega_{i,t}$$

раз к моменту t.

Приведем пример когда использование портфеля финансовых инструментов дает больший доход чем вложения в каждый из этих инструментов. [1]

Пусть вектор доходностей двух финансовых инструментов изменяется со временем как

$$\bar{\omega}_t = (1, \frac{1}{2}), (1, 2), (1, \frac{1}{2}), (1, 2), (1, \frac{1}{2}), (1, 2), \ldots$$

[1] Стратегия “Buy and Hold” предусматривает покупку некоторого количества единиц финансового инструмента в начале периода и продажу их в конце периода.

Легко видеть, что стратегия “Buy and Hold” не дает выигрыша ни на одном из этих финансовых инструментов по отдельности, тогда как выигрыш портфельной стратегии $\bar{\gamma}_t = \bar{\gamma} = (\frac{1}{2}, \frac{1}{2})$ за $2T$ шагов игры дает выигрыш на единицу вложения

$$K_T = 1 \cdot \left(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2\right) \cdots = \left(\frac{9}{8}\right)^T \to \infty$$

при $T \to \infty$.

Пусть Γ – множество всех портфелей. По определению это множество является N-мерным симплексом.

Определим функцию потерь в виде

$$\lambda(\bar{\omega}, \bar{\gamma}) = -\ln(\bar{\gamma} \cdot \bar{\omega}). \tag{5.58}$$

Рассмотрим постоянных экспертов – каждый эксперт всегда будет давать в качестве прогноза один и тот же портфель: $\bar{\gamma} \in \Gamma$. Таким образом, каждый портфель из Γ рассматривается в качестве эксперта.

Применим агрегирующий алгоритм AA к этому множеству экспертов и к этой функции потерь.

Доход от инвестиций от одной вложенной денежной единицы при постоянном портфеле $\bar{\gamma} \in \Gamma$ за первые T шагов (моментов времени) равен

$$K(\bar{\gamma}) = \prod_{t=1}^{T} (\bar{\gamma} \cdot \bar{\omega}_t).$$

Тогда кумулятивные потери эксперта $\bar{\gamma}$ за тот же период равны

$$L_T(\bar{\gamma}) = -\sum_{t=1}^{T} \ln(\bar{\gamma} \cdot \bar{\omega}_t).$$

Допустим, что задано априорное распределение $P_0(d\bar{\gamma})$ на симплексе Γ всех портфелей.

Согласно (5.20) потери обобщенного алгоритма APA равны

$$g_T(\omega) = \log_\beta \int_\Gamma \frac{\beta^{\lambda(\bar{\omega},\bar{\gamma}) + L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\bar{\gamma})} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}), \tag{5.59}$$

где $\beta = e^{-\eta}$, $0 < \eta \leqslant 1$.

Теорема 5.3. *Игра (функция потерь* (5.58)) *Ковера является смешиваемой при* $0 < \eta \leqslant 1$. *Функция подстановки задается выражением*

$$\Sigma(g_T) = \int_\Gamma \bar{\gamma} P^*_{T-1}(d\bar{\gamma}) =$$
$$= \int_\Theta \bar{\gamma} \frac{\beta^{L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\gamma)} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}). \tag{5.60}$$

Доказательство. Нам надо доказать, что для всех $\bar{\omega}$

$$\lambda\left(\bar{\omega}, \int_\Gamma \bar{\gamma} dP(\bar{\gamma})\right) \leqslant \log_\beta \int_\Gamma \beta^{\lambda(\bar{\omega},\bar{\gamma})} dP(\bar{\gamma}).$$

Это неравенство эквивалентно неравенству

$$f\left(\int_\Gamma \bar{\gamma} dP(\bar{\gamma})\right) \geqslant \int_\Gamma f(\bar{\gamma}) dP(\bar{\gamma}), \tag{5.61}$$

где $f(\bar{\gamma}) = \beta^{\lambda(\bar{\omega},\bar{\gamma})} = (\bar{\gamma} \cdot \bar{\omega})^\eta$. Неравенство (5.61) следует из вогнутости функции $f(\bar{\gamma})$ при $0 < \eta \leqslant 1$. При $\eta = 1$ это неравенство переходит в равенство

$$\left(\bar{\omega} \cdot \int_\Gamma \bar{\gamma} dP(\bar{\gamma})\right) = \int_\Gamma (\bar{\omega} \cdot \bar{\gamma}) dP(\bar{\gamma}).$$

$\triangle$

Прогноз портфеля (5.60) можно записать полностью, используя представление (5.21) (из леммы 5.2) суммарных потерь обобщенного алгоритма за T шагов:

$$L_T(APA) = \log_\beta \int_\Gamma \beta^{L_T(\bar{\gamma})} P_0(d\bar{\gamma}).$$

Полагаем $\eta = 1$. Так как при нашей функции потерь суммарные потери одного эксперта $\bar{\gamma}$ за T шагов равны

$$L_T(\bar{\gamma}) = -\ln \prod_{t=1}^{T} (\bar{\gamma} \cdot \bar{\omega}_t),$$

получаем выражение для оптимального портфеля – прогноза нашего алгоритма:

$$\bar{\gamma}_T = \frac{\int_\Gamma \bar{\gamma} \prod_{t=1}^{T-1} (\bar{\gamma} \cdot \bar{\omega}_t) P_0(d\bar{\gamma})}{\int_\Gamma \prod_{t=1}^{T-1} (\bar{\gamma} \cdot \bar{\omega}_t) P_0(d\bar{\gamma})}.$$

Удобно рассмотреть в качестве априорного распределения на симплексе всех постоянных портфелей распределение Дирихле с параметрами $(1/2, \ \ldots, \ 1/2)$:

$$P_0(d\bar{\gamma}) = \frac{\Gamma(N/2)}{[\Gamma(1/2)]^N} \prod_{j=1}^{N} \gamma_j^{-1/2} d\bar{\gamma},$$

где

$$\Gamma(a) = \int\limits_0^\infty x^{a-1} e^{-x} dx.$$

Заметим, что $\Gamma(N+1) = N!$.

Приведем верхнюю оценку потерь агрегирующего алгоритма в игре Ковера. Постановка задачи построения универсального портфеля при отсутствии стохастических предположений о природе вектора доходностей финансовых инструментов, алгоритм вычисления оптимального портфеля и оценка его капитала были впервые предложены Ковером [21], [22], [23]. Позже Вовк [58] показал, что алгоритм Ковера и его оценка могут быть получены как результат применения агрегирующего алгоритма AA. Игра Ковера и соответствующие утверждения представлены также в монографии [23].

Теорема 5.4. *Для любого постоянного портфеля $\bar{\gamma}$ выполнено неравенство*

$$L_T(AA) \leqslant L_T(\bar{\gamma}) + \frac{N-1}{2} \ln T + c \tag{5.62}$$

для всех T, где c – положительная константа.

Доказательство. Пусть $\eta = 1$. Тогда $\beta^{\lambda(\bar{\omega}, \bar{\gamma})} = (\bar{\omega} \cdot \bar{\gamma})$.

Кумулятивные потери агрегирующего алгоритма за T шагов удовлетворяют неравенству

$$L_T(AA) \leqslant \sum_{t=1}^{T} g_t(\bar{\omega}_t) = -\ln \int_{\Gamma} e^{-L_T(\bar{\gamma})} P_0(d\bar{\gamma}) =$$

$$= -\ln \int_{\Gamma} \prod_{t=1}^{T} (\bar{\omega}_t \cdot \bar{\gamma}) P_0(d\bar{\gamma}).$$

Для того, чтобы получить оценку (5.62), нам надо доказать неравенство

$$\int_{\Gamma} \prod_{t=1}^{T} (\bar{\omega}_t \cdot \bar{\gamma}) P_0(d\bar{\gamma}) \geqslant \epsilon T^{-\frac{N-1}{2}} \sup_{\bar{\gamma} \in \Gamma} \prod_{t=1}^{T} (\bar{\omega}_t \cdot \bar{\gamma}),$$

где ϵ – некоторая константа. Перепишем это неравенство в виде

$$\frac{\sup\limits_{\bar{\gamma} \in \Gamma} \prod_{t=1}^{T} (\bar{\omega}_t \cdot \bar{\gamma})}{\int_{\Gamma} \prod_{t=1}^{T} (\bar{\omega}_t \cdot \bar{\gamma}) P_0(d\bar{\gamma})} \leqslant \epsilon T^{\frac{N-1}{2}}. \tag{5.63}$$

Приведем две вспомогательные леммы.

Лемма 5.4. *Для любых неотрицательных вещественных чисел* $\alpha_1, \ldots, \alpha_T$ *и* $\beta_1, \ldots, \beta_T$ *выполнено* [2]

$$\frac{\sum\limits_{t=1}^{T} \alpha_t}{\sum\limits_{t=1}^{T} \beta_t} \leqslant \max_{1 \leqslant t \leqslant T} \frac{\alpha_t}{\beta_t}. \tag{5.64}$$

Доказательство. Пусть $J = \operatorname{argmax}_{1 \leqslant t \leqslant T} \frac{\alpha_t}{\beta_t}$. Если $\alpha_J = 0$, то неравенство (5.64) очевидно. Если $\alpha_J > 0$ и $\beta_J = 0$, то правая часть неравенства (5.64) равна бесконечности и неравенство вновь верно. Далее предполагаем, что $\alpha_J > 0$ и $\beta_J > 0$.

[2] Полагаем $\frac{0}{0} = 0$.

Для любого t из $\alpha_t/\beta_t \leqslant \alpha_J/\beta_J$ следует $\alpha_t/\alpha_J \leqslant \beta_t/\beta_J$. Поэтому

$$\frac{\sum\limits_{t=1}^{T}\alpha_t}{\sum\limits_{t=1}^{T}\beta_t} = \frac{\alpha_J\left(1+\sum\limits_{t\neq J}\frac{\alpha_t}{\alpha_J}\right)}{\beta_J\left(1+\sum\limits_{t\neq J}\frac{\beta_t}{\beta_J}\right)} \leqslant \frac{\alpha_J}{\beta_J}.$$

Лемма доказана. $\triangle$

Будем рассматривать произвольные последовательности чисел $j^T = j_1 \dots j_T$, где $j_1 \dots j_T \in \{1, \dots, N\}$.

Лемма 5.5. *Пусть* $\bar{\gamma}^* = \operatorname{argmin}_{\bar{\gamma}\in\Gamma} L_T(\gamma)$. *Для любого распределения вероятностей* P_0 *на симплексе* Γ *выполнено* [3]

$$\frac{\prod_{t=1}^{T}(\bar{\gamma}^*\cdot\bar{\omega}_t)}{\int_\Gamma \prod_{t=1}^{T}(\bar{\gamma}\cdot\bar{\omega}_t)P_0(d\bar{\gamma})} \leqslant \max_{j^T=j_1\dots j_T} \frac{\prod_{t=1}^{T}\gamma_{j_t}}{\int_\Gamma \prod_{t=1}^{T}\gamma_{j_t}P_0(d\bar{\gamma})}. \qquad (5.65)$$

Доказательство. Пусть $\bar{\gamma}^* = (\gamma_1^*, \dots, \gamma_N^*)$ и $\bar{\omega}_t = (\omega_{t,1}, \dots, \omega_{t,N})$.

Записываем произведение сумм в виде суммы всевозможных произведений составленных путем выбора по одному элементу из каждой суммы

$$\prod_{t=1}^{T}(\bar{\gamma}^*\cdot\bar{\omega}_t) = \prod_{t=1}^{T}\sum_{j=1}^{N}\gamma_j^*\omega_{t,j} = \sum_{j^T=j_1\dots j_T}\prod_{t=1}^{T}\gamma_{j_t}^*\omega_{t,j_t}. \qquad (5.66)$$

Аналогичным образом получаем представление

$$\int_\Gamma \prod_{t=1}^{T}(\bar{\omega}_t\cdot\bar{\gamma})P_0(d\bar{\gamma}) = \sum_{j^T=j_1\dots j_T}\int_\Gamma \prod_{t=1}^{T}\gamma_{j_t}\omega_{t,j_t})P_0(d\bar{\gamma}). \qquad (5.67)$$

Делим (5.66) на (5.67) и оцениваем частное по лемме 5.4, где $\alpha_{j^T} = \prod_{t=1}^{T}\gamma_{j_t}^*\omega_{t,j_t}$ и $\beta_{j^T} = \int_\Gamma \prod_{t=1}^{T}\gamma_{j_t}\omega_{t,j_t})P_0(d\bar{\gamma})$. При этом рассматриваем эти величины только для тех наборов индексов j^T,

[3] Левая часть неравенства не зависит от $\bar{\omega}$.

для которых $\prod_{t=1}^{T} \omega_{t,j_t} > 0$.

$$\frac{\prod_{t=1}^{T}(\bar{\gamma}^* \cdot \bar{\omega}_t)}{\int_\Gamma \prod_{t=1}^{T}(\bar{\omega}_t \cdot \bar{\gamma}) P_0(d\bar{\gamma})} =$$

$$= \frac{\sum\limits_{j^T=j_1\ldots j_T} \prod_{t=1}^{T} \gamma^*_{j_t}\omega_{t,j_t}).}{\sum\limits_{j^T=j_1\ldots j_T} \int_\Gamma \prod_{t=1}^{T} \gamma^*_{j_t}\omega_{t,j_t}) P_0(d\bar{\gamma})} \leqslant$$

$$\leqslant \max_{j^T=jj_1\ldots j_T} \frac{\prod_{t=1}^{T} \gamma_{j_t}\omega_{t,j_t}}{\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t}\omega_{t,j_t} P_0(d\bar{\gamma})} =$$

$$= \max_{j^T=j_1\ldots j_T} \frac{\prod_{t=1}^{T} \gamma_{j_t}}{\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t} P_0(d\bar{\gamma})}.$$

Последнее равенство было получено сокращением числителя и знаменателя на множитель ω_{t,j_t}.

Лемма доказана. $\triangle$

Для завершения доказательства теоремы нам необходимо оценить сверху правую часть неравенства (5.65).

Мы приведем вывод для более слабой оценки. когда в качестве распределения P_0 используется равномерное $(1,\ldots,1)$ – распределение Дирихле на симплексе Γ.

Полная оценка приведена в работе [22] и в монографии [23].

Оценим сверху величину

$$\frac{\max\limits_{j^T=j_1\ldots j_T} \prod_{t=1}^{T} \gamma_{j_t}}{\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t} P_0(d\bar{\gamma})}.$$

Пусть $n_i(j^T) = |\{t \leqslant T : j_t = i\}|$ при $i \in \{1,\ldots,N\}$, $\nu_i(j^T) = n_i(j^T)/T$ – соответствующие частоты встречаемости индексов в последовательности j^T, $H(\nu_1(j^T),\ldots,\nu_N(j^T)) = -\sum\limits_{i=1}^{N} \nu_i(j^T)\ln\nu_i(j^T)$ – частотная энтропия.

Методом Лагранжа максимизируем величину $\prod_{t=1}^{T} \gamma_{j_t}$ и получаем неравенство

$$\prod_{t=1}^{T} \gamma_{j_t} \leqslant 2^{-TH(\nu_1(j^T),\ldots,\nu_N(j^T))} \tag{5.68}$$

для любых $\gamma_{j_1}, \ldots, \gamma_{j_T}$, таких что $\sum\limits_{i=1}^{N} \gamma_{j_i} = 1$, причем равенство достигается при $\gamma_{j_i} = \nu_i(j^T)$, $i = 1, \ldots, N$.

Из определения интеграл от $\gamma_{j_1} \ldots, \gamma_{j_T}$ по равномерному распределению P_0 на Γ равен

$$\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t} P_0(d\bar{\gamma}) = \frac{1}{\binom{T+N-1}{N-1} \mathcal{N}(\nu_1(j^T), \ldots, \nu_N(j^T))}, \tag{5.69}$$

где $\mathcal{N}(\nu_1(j^T), \ldots, \nu_N(j^T))$ – число наборов $i_1, \ldots, i_T$, в которых частоты встречаемости элементов $\{1, \ldots, N\}$ равны $\nu_1(j^T), \ldots, \nu_N(j^T)$. Такой набор частот называется типом последовательности j^T. Здесь $\binom{T+N-1}{N-1}$ – число типов.[4] Тогда из (5.69) получаем

$$\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t} P_0(d\bar{\gamma}) \geqslant \frac{1}{\binom{T+N-1}{N-1} 2^{TH(\nu_1(j^T), \ldots, \nu_N(j^T))}}. \tag{5.70}$$

Из оценок (5.68) и (5.69) получаем оценку

$$\frac{\max\limits_{j^T = j_1 \ldots j_T} \prod_{t=1}^{T} \gamma_{j_t}}{\int_\Gamma \prod_{t=1}^{T} \gamma_{j_t} P_0(d\bar{\gamma})} \leqslant \binom{T+N-1}{N-1} \leqslant (T+1)^{N-1}.$$

Таким образом, мы получили оценку (5.63), в которой величина $\epsilon T^{\frac{N-1}{2}}$ заменена на ϵT^{N-1}.

На этом доказательство теоремы 5.4 завершается. $\triangle$

Так как доход при следовании стратегии инвестирования, предлагаемой алгоритмом AA, равен

$$K_T(AA) = e^{-L_T(AA)},$$

неравенство (5.62) можно переписать в виде

$$K_T(AA) \geqslant e^{-c} T^{-\frac{N-1}{2}} K_T(\bar{\gamma}),$$

[4] Число наборов в одном типе $(\nu_1, \ldots, \nu_N)$ равно $\mathcal{N}(\nu_1, \ldots, \nu_N) = \binom{T}{n_1, \ldots, n_N}$, где $\nu_i = \frac{n_i}{T}$, $1 \leqslant i \leqslant N$, а число всех типов равно числу решений уравнения $n_1 + \cdots + n_N = N$ в неотрицательных целых числах, а именно, $\binom{T+N-1}{N-1}$. Кроме этого, хорошо известно, что $\binom{T}{n_1, \ldots, n_N} \leqslant 2^{TH(\nu_1, \ldots, \nu_N)}$.

где c – константа, $K_T(\bar{\gamma})$ – доход, полученный при использовании произвольного постоянного портфеля $\bar{\gamma}$.

Данная оценка эффективна в том случае, когда доход наилучшего алгоритма, использующего постоянный портфель акций, растет экспоненциально с ростом числа временных шагов T игры. В этом случае, благодаря полиномиальности оценки (5.71) по T, доход универсального алгоритма также растет экспоненциально по T. К числу недостатков алгоритма Ковера является большая вычислительная сложность, которая растет экспоненциально в зависимости от числа акций.

5.10. Многомерная онлайн регрессия

5.10.1. Многомерная регрессия с помощью агрегирующего алгоритма

В этом разделе мы рассмотрим применение агрегирующего алгоритма для решения задачи регрессии. В отличие от обычной многомерной регрессии, которая использует обучающую выборку для определения своих параметров, AA-алгоритм обучается в режиме онлайн. Соответствующий алгоритм предложен р работе [61].

Рассмотрим многомерную линейную регрессию. *Природа* выдает значения (x_t, y_t), где $x_t \in \mathcal{R}^n$ и $y_t \in \mathcal{R}$ при $t = 1, 2, \ \ldots$

Задача регрессии заключается в вычислении на каждом шаге $t > 1$ прогноза величины y_t по ранее полученным значениям $(x_1, y_1), \ \ldots, \ (x_{t-1}, y_{t-1})$ и значению аргумента x_t.

В этом разделе мы не подчеркиваем векторы чертой сверху.

В соответствии с теорией предсказаний с использованием экспертных стратегий введем в эту игру экспертов $\theta \in \mathcal{R}^n$. Эксперт θ дает на шаге t предсказание – значение линейной функции: $\xi_t^\theta = (\theta \cdot x_t)$.

Задачей регрессии в режиме онлайн является построение на каждом шаге t прогноза величины y_t, используя прогнозы таких линейных экспертов.

Общая схема регрессии регулируется протоколом, приведенным на рис. 5.2.

FOR $t = 1, 2, \ldots$
Природа анонсирует $x_t \in \mathcal{R}^n$.
Эксперты анонсируют прогнозы $\xi_t^\theta = (\theta \cdot x_t)$, $\theta \in \mathcal{R}^n$.
Предсказатель представляет предсказание $\gamma_t \in \mathcal{R}$.
Природа анонсирует $y_t \in [-Y, Y]$.
ENDFOR

Рис. 5.2: Общая схема регрессии

Рассматривается игра с полной информацией между игроками: *Эксперт* θ, *Предсказатель* и *Природа*.

Для вычисления потерь используется квадратичная функция потерь. На шаге t *Эксперт* θ вычисляет свои потери $(y_t - (\theta \cdot x_t))^2$. *Предсказатель* вычисляет свои потери $(y_t - \gamma_t)^2$.

Применим к этой игре алгоритм AA с параметром обучения $\eta = 1/2Y$. [5]

Введем априорное распределение

$$P_0(d\theta) = (a\eta/\pi)^{n/2} e^{-a\eta\|\theta\|^2} d\theta, \tag{5.71}$$

где a – некоторый параметр (аналогичный параметру, который используется в гребневой регрессии), а константы выбраны из условия нормализации. Здесь используется евклидова норма вектора $\theta = (\theta_1, \ldots, \theta_n)$, заданная формулой $\|\theta\| = \sqrt{\theta_1^2 + \cdots + \theta_n^2}$. Напомним также, что мы отождествляем скалярное произведение $(\theta \cdot x)$ и одноэлементную матрицу $x'\theta$, где x' – вектор-строка, θ – вектор-столбец.

Тогда потери произвольного эксперта $\theta \in \mathcal{R}^n$ на шаге t равны:

$$\lambda(y_t, x_t'\theta) = (y_t - x_t'\theta)^2 = \theta'(x_t x_t')\theta - 2(y_t x_t')\theta + y_t^2. \tag{5.72}$$

Напомним, что $x_t' = (x_{1,t}, \ldots, x_{n,t})$, $\theta' = (\theta_1, \ldots, \theta_n)$, а x_t, θ – эти же векторы, записанные в виде столбцов. Здесь мы использовали равенство $x_t'\theta x_t'\theta = \theta'(x_t x_t')\theta$, которое можно проверить

[5] Можно доказать, что при таком значении параметра функция подстановки существует и имеет вид (5.50).

по-координатными преобразованиями:

$$x_t'\theta x_t'\theta = \left(\sum_{i=1}^{n} x_{t,i}\theta_i\right)\left(\sum_{j=1}^{n} x_{t,j}\theta_j\right) =$$

$$= \sum_{i,j=1}^{n} \theta_i x_{t,i} x_{t,j}\theta_j = \theta'(x_t x_t')\theta.$$

Потери этого эксперта $\theta \in \mathcal{R}^n$ за первые T шагов равны

$$L_T(\theta) = \sum_{t=1}^{T}(y_t - x_t'\theta)^2 =$$

$$= \theta'\left(\sum_{t=1}^{T} x_t x_t'\right)\theta - 2\left(\sum_{t=1}^{T} y_t x_t'\right)\theta + \sum_{t=1}^{T} y_t^2. \tag{5.73}$$

Согласно (5.8) и (5.22) имеем

$$P_{t-1}(d\theta) = \beta^{L_{t-1}(\theta)} P_0(d\theta).$$

Поэтому произвольное псевдопредсказание на шаге t имеет вид

$$g_t(y) = \log_\beta \int \beta^{\lambda(y, x_t'\theta)} P_{t-1}^*(d\theta) =$$

$$= \log_\beta \int \beta^{\lambda(y, x_t'\theta)} \frac{P_{t-1}(d\theta)}{P_{t-1}(\Theta)} =$$

$$= \log_\beta \int \beta^{\lambda(y, x_t'\theta) + L_{t-1}(\theta)} \frac{1}{P_{t-1}(\Theta)} P_0(d\theta). \tag{5.74}$$

Отсюда, учитывая представления (5.71) для априорного распределения и (5.72) для функции потерь, получим

$$g_T(-Y) = \log_\beta \int \beta^{\lambda(-Y, x_T'\theta) + L_{T-1}(\theta)} \frac{1}{P_{T-1}(\Theta)} P_0(d\theta) = \tag{5.75}$$

$$= \int_{\mathcal{R}^n} e^{-\eta\theta'(aI + \sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} \frac{d\theta}{P_{T-1}(\Theta)}.$$

Аналогичное представление имеет место для $g_T(Y)$.

В случае квадратичной функции потерь и множества предсказаний $[-Y, Y]$ можно показать, что функция подстановки существует и имеет вид (5.50) (см. [61]).

Тогда, используя формулу (5.75) для $g_T(-Y)$ и аналогичную формулу для $g_T(Y)$, получаем

$$\gamma_T = \frac{1}{4Y}(g_T(-Y) - g_T(Y)) = \frac{1}{4Y} \times$$

$$\times \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_T')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_t')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_T')\theta} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta e^{-\eta F\left(aI+\sum\limits_{t=1}^{T} x_t x_t', -2\sum\limits_{t=1}^{T-1} y_t x_t', 2Y x_T'\right)} =$$

$$= \frac{1}{4Y} F\left(aI + \sum_{t=1}^{T} x_t x_t', -2\sum_{t=1}^{T-1} y_t x_t', 2Y x_T'\right) =$$

$$= \left(\sum_{t=1}^{T-1} y_t x_t'\right)\left(aI + \sum_{t=1}^{T} x_t x_t'\right)^{-1} \cdot x_T. \qquad (5.76)$$

Здесь мы сразу сократили общий множитель $\frac{1}{P_{T-1}(\Theta)}$ в числителе и знаменателе 2-й строки. При переходе от 2-й строки к 3-й множитель $e^{-\eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)}$ в числителе и знаменателе был вынесен из под интеграла и сокращен.

При переходе от 3-й строки к 4-й мы используем следующую ниже лемму 5.6, из которой следует, что интеграл в числителе 3-й строки равен

$$\frac{\pi^{n/2}}{\sqrt{\det A}} e^{-\eta \inf\limits_{\theta\in\mathcal{R}^n}(\theta' A\theta + c'\theta + x'\theta)},$$

а интеграл в знаменателе 3-й строки равен

$$\frac{\pi^{n/2}}{\sqrt{\det A}} e^{-\eta \inf\limits_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta - x'\theta)},$$

где

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$c = -2 \sum_{t=1}^{T-1} y_t x_t',$$
$$x = 2Y x_T'.$$

В 4-й строке мы использовали обозначение

$$F(A, c, x) = \inf_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta + x'\theta) - \\ - \inf_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta - x'\theta), \tag{5.77}$$

а при переходе от 5-й строки к 6-й – следующую ниже лемму 5.7, согласно которой $F(A, c, x) = -c'A^{-1}x$.

Приведем формулировки и доказательства лемм 5.6 и 5.7.

Лемма 5.6. *Пусть $Q(\theta) = \theta' A\theta + c'\theta + d$, где $\theta, c \in \mathcal{R}^n$, $d \in \mathcal{R}$ и A – симметричная положительно определенная матрица типа $(n \times n)$. Тогда*

$$\int\limits_{\mathcal{R}^n} e^{-Q(\theta)} d\theta = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A}}, \tag{5.78}$$

где $Q_0 = \min_{\theta \in \mathcal{R}^n} Q(\theta)$.

Доказательство. Пусть минимум квадратичной формы

$$Q(\theta) = \theta' A\theta + c'\theta + d$$

достигается при $\theta = \theta_0$. Полагаем $\xi = \theta - \theta_0$ и $\tilde{Q}(\xi) = Q(\xi + \theta_0)$. Легко видеть, что квадратичная часть формы $\tilde{Q}$ есть $\xi' A\xi$. Так

как минимум новой формы $\tilde{Q}$ достигается при $\theta = \bar{0}$, где $\bar{0} = (0, \ldots, 0)$, эта форма не может иметь линейной части. Действительно, в достаточно малой окрестности $\bar{0}$ линейная часть доминировала бы над квадратичной частью и тогда бы на $\bar{0}$ форма $\tilde{Q}$ не принимала бы минимальное значение.

Так как минимум $\tilde{Q}(\xi)$ равен Q_0, можно заключить, что константа формы есть Q_0. Таким образом, $\tilde{Q}(\xi) = \xi' A\xi + Q_0$.

Остается доказать, что

$$\int\limits_{\mathcal{R}^n} e^{-\xi' A\xi} d\xi = \pi^{n/2}/\sqrt{\det A}.$$

Это следует из теоремы 3 (раздела 2.7) монографии [1]. $\triangle$

Лемма 5.7 показывает, что $F(A, c, x) = -c' A^{-1} x$.

Лемма 5.7. *Пусть A – симметричная положительно определенная матрица типа $(n \times n)$, $b, x \in \mathcal{R}^n$. Тогда*

$$F(A, c, x) = \min_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta + x'\theta) -$$
$$- \min_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta - x'\theta) = -c' A^{-1} x. \qquad (5.79)$$

Доказательство. Для нахождения первого минимума приравниваем частные производные квадратичной формы $\theta' A\theta + c'\theta + x'\theta$ по θ_i к нулю. Здесь $\theta = (\theta_1, \ldots, \theta_n)$. Получаем систему уравнений $2A\theta + c' + x' = \bar{0}$. Отсюда легко видеть, что этот минимум достигается при значении вектора $\theta_1 = -\frac{1}{2} A^{-1}(c + x)$. Аналогично, минимум второй части достигается при $\theta_1 = -\frac{1}{2} A^{-1}(c - x)$. Утверждение леммы получается подстановкой этих значений в разность минимумов. $\triangle$

$A = aI$; $b' = \bar{0}$.
FOR $t = 1, 2, \ \ldots$
Алгоритм получает $x_t \in \mathcal{R}^n$.
Вычисляем $A = A + x_t x_t'$.
Выдаем предсказание $\gamma_t = b' A^{-1} x_t$.
Алгоритм получает $y_t \in [-Y, Y]$.
Вычисляем $b' = b' + y_t x_t'$.
ENDFOR

Рис. 5.3: Алгоритм регрессии AAR

Таким образом, согласно выражению (5.76) для γ_T, на шаге T

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$b = \sum_{t=1}^{T-1} y_t x_t',$$
$$\gamma_T = b' A^{-1} x_T =$$
$$= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T.$$

Эти формулы показывают, что мы можем записать теперь алгоритм AAR для регрессии следующим образом – см. рис. 5.3.

Сравнение потерь алгоритма AAR с потерями наилучшего эксперта дает следующая теорема.

Теорема 5.5. *Для произвольного T*

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|^2) + Y^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) \leqslant$$
$$\leqslant \inf_\theta (L_T(\theta) + a\|\theta\|^2) + Y^2 \sum_{i=1}^{n} \ln \left(1 + \frac{1}{a} \sum_{t=1}^{T} x_{t,i}^2 \right).$$

Если к тому же $|x_{t,i}| \leqslant X$ *для всех* t *и* i, *то*

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|^2) + nY^2 \ln\left(\frac{TX^2}{a} + 1\right).$$

Доказательство. Пусть $\eta = \frac{1}{2Y^2}$. По лемме 5.2 потери алгоритма APA выражаются в виде формулы (5.21). В нашем случае эта формула записывается в виде

$$L_T(APA) = \log_\beta \int_{\mathcal{R}^n} e^{-\eta L_T(\theta)} P_0(d\theta) =$$

$$= \log_\beta \int_{\mathcal{R}^n} (a\eta/\pi)^{n/2} e^{-\eta\theta'(aI + \sum\limits_{t=1}^T x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^T y_t x_t')\theta - \eta \sum\limits_{t=1}^T y_t^2} d\theta. \quad (5.80)$$

Экспонента в (5.80) имеет вид $e^{-\eta F(\theta)}$, где

$$F(\theta) = \theta' \left(aI + \sum_{t=1}^T x_t x_t'\right)\theta - 2\left(\sum_{t=1}^T y_t x_t'\right)\theta + \sum_{t=1}^T y_t^2.$$

Пусть минимум $F(\theta)$ достигается при $\theta = \theta_0$.

Тогда по лемме 5.6 выражение (5.80), представляющее потери APA, равно

$$L_T(APA) =$$

$$= \log_\beta \left(((a\eta/\pi)^{n/2}) \frac{\pi^{n/2} e^{-\eta F(\theta_0)}}{\sqrt{\det\left(a\eta I + \eta \sum\limits_{t=1}^T x_t x_t')\right)}} \right) =$$

$$= F(\theta_0) - \frac{1}{2}\log_\beta \det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right) =$$

$$= F(\theta_0) + \frac{1}{2\eta}\ln\det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right) =$$

$$= F(\theta_0) + Y^2 \ln\det\left(I + \frac{1}{a}\sum_{t=1}^T x_t x_t'\right).$$

По определению (5.73) кумулятивных потерь эксперта θ

$$F(\theta_0) = \theta_0' \left(aI + \sum_{t=1}^{T} x_t x_t' \right) - 2 \left(\sum_{t=1}^{T} y_t x_t' \right) \theta_0 + \sum_{t=1}^{T} y_t^2 =$$
$$= a\|\theta_0\|^2 + \sum_{t=1}^{T} (y_t - x_t'\theta_0)^2 =$$
$$= a\|\theta_0\|^2 + L_T(\theta_0).$$

Так как $L_T(AAR) \leqslant L_T(APA)$, отсюда получаем утверждение теоремы. $\triangle$

5.10.2. Переход к ядерной многомерной регрессии

Для перехода к ядерной регрессии необходимо перевести все полученные алгоритмы и оценки ошибок регрессии в форму, при которой все они зависят только от скалярных произведений векторов исходных данных. После этого мы предположим, что гиперплоскость регрессии была проведена в пространстве признаков, при этом скалярные произведения будут свернуты в виде значений ядра.

Вспомним формулы для вычисления основных параметров алгоритма:

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$b = \sum_{t=1}^{T-1} y_t x_t',$$
$$\gamma_T = b' A^{-1} x_T =$$
$$= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T. \tag{5.81}$$

Пусть $K(x, x')$ – некоторое ядро, где $x, x' \in \mathcal{R}^n$, и мы получаем на вход выборку $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots)$

Вводем обозначения:

$K_T = (K(x_i, x_j))_{i,j=1}^T$ – матрица значений ядра;

$k_T = (k(x_i, x_T))_{i=1}^T$ – последний столбец матрицы K_T;

Y_T – вектор-столбец исходов;

$(Y_{T-1}, 0) = (y_1, \ldots, y_{T-1}, 0)$ – неполный вектор-столбец исходов, дополненный нулем.

Запишем онлайн алгоритм линейной регрессии (5.81) в виде удобном для перевода в ядерную форму.

Введем матрицу типа $T \times n$

$$X_T = \begin{pmatrix} x_1' \\ x_2' \\ \ldots \\ x_T' \end{pmatrix} = \begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1T} \\ x_{21}, x_{22}, \ldots, x_{2T} \\ \ldots \\ x_{n1}, x_{n2}, \ldots, x_{nT} \end{pmatrix},$$

у которой строки – векторы-строки $x_1', \ldots, x_T'$. Тогда легко проверить, что

$$\sum_{t=1}^{T} x_t x_t' = X_T' X_T,$$

а также

$$\sum_{t=1}^{T-1} y_t x_t = (Y_{T-1}, 0)' X_T.$$

Имеет место следующая

Лемма 5.8. *Для любой $n \times m$ матрицы B и любой $m \times n$ матрицы C таких, что матрицы $aI_n + CB$ и $aI_m + BC$ обратимы, имеет место равенство*

$$B(aI_n + CB)^{-1} = (aI_m + BC)^{-1} B, \tag{5.82}$$

где a – любое вещественное число и I_n – единичная матрица размера n.

Доказательство. Равенство (5.82) эквивалентно равенству

$$(aI_n + BC)B = B(aI_m + CB),$$

которое очевидно ввиду дистрибутивности умножения матриц. $\triangle$

Используя лемму представим значение предсказания линейной регрессии

$$\begin{aligned}
\gamma_T = b' A^{-1} x_T = \\
= \left(\sum_{t=1}^{T-1} y_t x_t'\right) \left(aI + \sum_{t=1}^{T} x_t x_t'\right)^{-1} \cdot x_T = \\
= (Y_{T-1}, 0)' \, X_T \left(aI + X_T' X_T\right)^{-1} x_T = \\
(Y_{T-1}, 0)' \left(aI + X_T X_T'\right)^{-1} X_T x_T = \\
= (Y_{T-1}, 0)' \left(aI + \tilde{K}_T\right)^{-1} \tilde{k}_T, \qquad (5.83)
\end{aligned}$$

где $\tilde{K}_T = X_T X_T'$ и $\tilde{k}_T = X_T x_T$. Заметим также, что

$$\begin{aligned}
X_T X_T' = (x_t \cdot x_{t'})_{t,t'=1}^{T}, \\
\tilde{k}_T = (x_t \cdot x_T)_{t=1}^{T},
\end{aligned}$$

т.е. элементы матрицы и вектора представляют собой скалярные произведения векторов $x_1, \ldots, \ x_T$.

Адаптивный алгоритм ядерной версии получается из алгоритма линейной онлайн регрессии заменой скалярных произведений из матрицы $\tilde{K}_T = X_T X_T'$ и вектора $\tilde{k}_T = X_T x_T$ на значения ядра $K_T = (K(x_i, x_j))_{i,j=1}^{T}$ и $k_T = (k(x_i, x_T))_{i=1}^{T}$. Получим выражение для прогноза ядерной версии адаптивной многомерной регрессии

$$\gamma_T = (Y_{T-1}, 0)' \, (aI + K_T)^{-1} \, k_T.$$

Оценка ошибки обучения для ядерной многомерной регрессии имеет вид

Теорема 5.6. *Для всех T*

$$L_T(AAR) \leqslant \inf_{\theta} (L_T(\theta) + a\|\theta\|^2) + Y^2 \ln det \left(\frac{\tilde{K}_T}{a} + I\right)$$

для всех T.

5.10.3. Ядерная форма гребневой регрессии

Пусть алгоритм предсказания $\mathcal{A}$ использует входные векторы $x_1, x_2, \dots x_T$ и результат его работы зависит только от их скалярных произведений. Также задано ядро $K(x, y)$, где $x, y \in \mathcal{R}^n$. Заменим всюду скалярные произведения $(x_i \cdot x_j)$ на $K(x_i, x_j)$.

В частности, формула для предсказания методом гребневой регрессии (см. раздел 2.8.2)

$$\gamma_{T+1} = w'x = ((aI + X_T'X_T)^{-1}X_T'Y_T)'x_{T+1}$$

может быть преобразована с помощью леммы 5.8 к виду

$$\begin{aligned}\gamma_{T+1} = w'x = ((aI + X_T'X_T)^{-1}X_T'Y_T)'x_{T+1} = \\ = Y_T'X_T(aI + X_T'X_T)^{-1}x_{T+1} = \\ = Y_T'(aI + X_TX_T')^{-1}X_Tx_{T+1} = \\ = Y_T'(aI + X_TX_T')^{-1}k_T,\end{aligned}$$

где $k_T = (x_{T+1} \cdot x_t)_{t=1}^T$.

Предсказание для ядерной гребневой регрессии имеет вид

$$\gamma_{T+1} = Y_T'(aI + K_T)^{-1}k_T,$$

где $K_T = (K(x_i, x_j))_{i=1}^T$ и $k_T = K(x_{T+1}, x_t)_{t=1}^T$.

Заметим, что в этих же обозначениях формула (5.83) для предсказания адаптивной ядерной регрессии имеет несколько отличный вид:

$$\gamma_{T+1} = (Y_T, 0)' \left(aI + K_{T+1}\right)^{-1} k_{T+1},$$

где $K_{T+1} = (K(x_i, x_j))_{i=1}^{T+1}$ и $k_{T+1} = K(x_{T+1}, x_t)_{t=1}^{T+1}$.

5.11. Задачи и упражнения

1. Нарисовать графики предсказаний и области суперпредсказаний, а также их образы в экспоненциальном пространстве для квадратичной, логарифмической, абсолютной и простой функций потерь для различных $\eta > 0$. Привести примеры η, при которых

для логарифмической и квадратичной функций потерь соответствующие области в экспоненциальном пространстве будут выпуклыми (а также невыпуклыми).

2. Функция потерь $\lambda(\omega, \gamma)$ называется η-экспонециально вогнутой, если функция $f(\gamma) = e^{-\eta\lambda(\omega,\gamma)}$ является вогнутой для всех $\omega \in \Omega$. Называем функцию $\lambda(\omega, \gamma)$ экспонециально вогнутой, если она η-экспонециально вогнута для некоторого $\eta > 0$.

Доказать, что всякая экспоненциально вогнутая функция $\lambda(\omega, \gamma)$ является выпуклой по γ для любого ω. Привести пример, когда обратное утверждение неверно (*Указание.* Изучить абсолютную функцию потерь $\lambda(\omega, \gamma) = |\omega - \gamma|$).

3. Доказать, что если некоторая функция потерь является η-экспоненциально вогнутой для некоторого $\eta > 0$, то она является η'-экспоненциально вогнутой для любого $0 < \eta' \leqslant \eta$.

4. Объясните в чем отличие понятий смешиваемости и экспоненциальной вогнутости. Доказать, что класс смешиваемых функций потерь строго шире чем класс экспоненциально вогнутых функций (*Указание.* Для некоторых $\eta > 0$ квадратичная функция потерь может быть $\eta > 0$-смешиваемой, но не экспоненциально вогнутой. Изучить свойства функции $f(\gamma) = e^{-\eta(\omega-\gamma)^2}$ для различных $\eta > 0$).

5. Доказать, что функция потерь $\lambda(\omega, p) = (\omega - p)^2$, где $\omega \in \{0, 1\}$ и $p \in [0, 1]$ является η-смешиваемой при $0 < \eta \leqslant 2$.

6. Доказать, что абсолютная функция потерь $\lambda(\omega, \gamma) = |\omega - \gamma|$ не является смешиваемой. Вычислить константу $c(\beta)$ (и $a(\beta)$) из оценки (5.27) для случая $\omega \in \{0, 1\}$ и $\gamma \in [0, 1]$.

(*Ответ*:

$$c(\beta) = \frac{\frac{1}{2} ln \frac{1}{\beta}}{\ln\left(\frac{2}{1+\beta}\right)} = \frac{\frac{1}{2}\eta}{\ln\left(\frac{2}{1+e^{-\eta}}\right)},$$

где $\beta = e^{-\eta}$.

Можно получить это значение путем минимизации показателя степени c в неравенстве:

$$e^{-\eta\lambda(\omega,\gamma)} \geqslant \left(\sum_{i=1}^{N} e^{-\eta\lambda(\omega,\xi_t^i)} P_{t-1}^*(i)\right)^c$$

в наихудшем случае – для произвольных ω, γ, ξ_t^i и P_{t-1}^*. Удобно использовать геометрическое представление – кривую:

$$\left(e^{-\eta\gamma}, e^{-\eta(1-\gamma)}\right)$$

при $0 \leqslant \gamma \leqslant 1$. Найти c для случая, когда расстояние от точки на кривой до самой длинной хорды максимально по всем ω, γ и ξ_i).

7. Проверить выпуклость кривой (5.37).

8. Нарисовать график кривой $c(\beta)$, заданной равенством (5.38).

9. Вывести неравенство (5.39). Исследовать зависимость множителя в оценке (5.39) от значения β.

Перестроить эту оценку в оценку с единичным множителем и регретом порядка $O(\sqrt{T \ln N})$ аналогично тому, как это было сделано в алгоритме распределения потерь в режиме онлайн из раздела 4.2.

10. Построить модификацию Fixed-Share для агрегирующего алгоритмом Вовка. Получить оценку соответствующего регрета.

5.12. Лабораторные работы

Использовать данные из следующих вебсайтов для решения задач регрессии: $http://www.csie.ntu.edu.tw$; База данных UCI Machine Learning Repository находится по адресу $http://archive.ics.uci.edu$. Она содержит большое число наборов данных для классификации и регрессии.

Лабораторная работа 1

Построить простую линейную, гребневую регрессии, а также регрессию с помощью стандартного программного обеспечения SVM, данных по формулам разделов 2.8 и 2.8.2, 2.9.1. Дать сравнительный анализ точности регрессии для всех использованных методов.

Лабораторная работа 2

Провести также эксперименты с ядерными версиями этих методов. Дать сравнительный анализ точности регрессии для всех использованных методов.

Лабораторная работа 3

Провести линейную регрессию в режиме онлайн с помощью агрегирующего алгоритма из раздела 5.10. Провести сравнительный анализ точности регрессии с другими методами.

Лабораторная работа 4

Реализовать следующую стратегию построения портфеля акций на основе данных об исторических ценах акций из сайта www.finam.ru (см. [53]).

Имеется N акций. Пусть $\omega_{i,t}$ – доходность акции i, $S_{i,t}$ – количество денег, вложенных в акцию i, в момент времени t. Рассматривается N стратегий типа "Buy and Hold". Алгоритм инвестирования заключается в перераспределении инвестиций между этими стратегиями: для произвольной акции i на шаге t оставляем сумму $(1-\alpha)S_{i,t}$, вложенными в эту акцию, а сумму $\alpha S_{i,t}$ распределяем равномерно между остальными $N-1$ акциями. Здесь $0 \leqslant \alpha \leqslant 1$ – параметр алгоритма. В этом случае капитал, вложенный в акцию i, изменяется по формуле

$$S_{i,t+1} = \left((1-\alpha)S_{i,t} + \frac{\alpha}{N-1} \sum_{j \neq i} S_{j,t}, \right) \omega_{i,t+1}.$$

при $1 \leqslant i \leqslant N$.

Написать точный алгоритм и реализовать его в виде компьютерной программы. Провести численные эксперименты.

Сформулировать этот алгоритм в форме Fixed-Share с весами $w_{i,t} = S_{i,t} / \sum\limits_{j=1}^{N} S_{j,t}$. Получить оценку регрета.

Глава 6

Выпуклая онлайн оптимизация

6.1. Алгоритмы следования за лидером FTL и FTRL

Задано выпуклое подмножество $S \subseteq \mathcal{R}^k$. Рассматривается следующая повторяющаяся игра, в которой предсказатель принимает решение $\mathbf{w}_t \in S$, после чего природа выдает функцию $f_t : S \to \mathcal{R}$. В результате предсказатель несет потери $f_t(\mathbf{w}_t)$.

Кумулятивные потери предсказателя за первые T шагов равны $H_T = \sum\limits_{t=1}^{T} f_t(\mathbf{w}_t)$. Кумулятивные потери фиксированного решения $\mathbf{u} \in S$ равны $L_T(\mathbf{u}) = \sum\limits_{t=1}^{T} f_t(\mathbf{u})$. Регрет предсказателя относительно заданного решения $\mathbf{u} \in S$ определяется как

$$R_T(\mathbf{u}) = H_T - L_T(\mathbf{u}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}).$$

Цель предсказателя заключается в том, чтобы минимизировать регрет какие бы функции $f_1, \ldots, f_T$ природа не выбирала. Точнее, предсказатель должен выбирать элементы $\mathbf{w}_t \in S$ так, чтобы величина $\sup_{f_1,\ldots,f_T} \min_{\mathbf{u}\in S} R_T(\mathbf{u})$ была бы минимальной.

Онлайн оптимизация
Полагаем $\mathbf{w}_1 \in S$ – любая точка.
FOR $t = 1,\ 2,\ \ldots$
Предсказатель принимает решение $\mathbf{w}_t \in S$.
Природа выдает функцию $f_t : S \to \mathcal{R}$.
Предсказатель несет потери $f_t(\mathbf{w}_t)$.
ENDFOR

Рис. 6.1: Онлайн оптимизация

Протокол игры онлайн оптимизации представлен на рис. 6.1.

Примером такой игры является алгоритм Hedge, где на каждом шаге t предсказатель выбирает элемент симплекса $\mathbf{w}_t = (w_{1,t}, \ldots, w_{k,t})$, а природа представляет функцию вида $f_t(\mathbf{w}) = (\mathbf{w} \cdot l_t)$, где $l_t = (l_{1,t}, \ldots, l_{k,t})$ – вектор потерь экспертов. Потери предсказателя на шаге t равны $(\mathbf{w}_t \cdot l_t) = \sum_{i=1}^{k} w_{i,t} l_{i,t}$.

Алгоритм FTL. В алгоритме следования за лидером FTL в протоколе (рис. 6.1) на каждом шаге t в качестве предсказания выбирается наилучшее решение на основании прошлой информации

$$\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}), \tag{6.1}$$

при этом, $\mathbf{w}_1$ – произвольный элемент S.[1]

Основное свойство алгоритма FTL представлено в следующем утверждении.

Предложение 6.1. *Для любого* $\mathbf{u} \in S$

$$R_T^{FTL}(\mathbf{u}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}). \tag{6.2}$$

[1]Предполагаем, что минимум в (6.1) достигается в некоторой точке $\mathbf{w} \in S$. В частности, этот минимум существует, если для каждого t функция f_t выпуклая. Если точка минимума не единственная, то выбираем одну из таких точек согласно некоторому правилу.

Доказательство. Вычтем $\sum\limits_{t=1}^{T} f_t(\mathbf{w}_t)$ из обоих частей неравенства (6.2) и получим, что надо доказать, что

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leqslant \sum_{t=1}^{T} f_t(\mathbf{u}) \tag{6.3}$$

для любого $\mathbf{u} \in S$. Будем доказывать это утверждение с помощью матемтической индукции по T. При $T = 1$ утверждение следует из определения $\mathbf{w}_2$. Допустим, что это утверждение верно для $T-1$:

$$\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1}) \leqslant \sum_{t=1}^{T-1} f_t(\mathbf{u})$$

для любого $\mathbf{u} \in S$. Тогда

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leqslant \sum_{t=1}^{T-1} f_t(\mathbf{u}) + f_{T+1}(\mathbf{w}_{T+1})$$

для любого $\mathbf{u} \in S$. В частности, при $\mathbf{u} = \mathbf{w}_{T+1}$ будет

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leqslant \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1}) = \min_{\mathbf{u}\in S} \sum_{t=1}^{T} f_t(\mathbf{u}).$$

Отсюда следует (6.3), а из него следует (6.2). $\triangle$

Приведем простейшее приложение предложения 6.1.

Пример 1. Онлайн квадратичная оптимизация. Пусть на шаге t *Природа* выбирает функцию $f_t(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}_t\|^2$, где $\mathbf{w}, \mathbf{z}_t \in \mathcal{R}$ и $\|\cdot\| = \|\cdot\|_2$ эвклидова норма в $\mathcal{R}^k$, а *Предсказатель* выбирает

$$\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^{t-1} \|\mathbf{w} - \mathbf{z}_i\|^2. \tag{6.4}$$

Находим минимум по $\mathbf{w}$ – приравниваем к нулю производную правой части (6.4). получаем $\sum\limits_{i=1}^{t-1} \mathbf{w} - \mathbf{z}_i = \mathbf{0}$. Отсюда $\mathbf{w}_t = \frac{1}{t-1}\sum_{i=1}^{t-1} \mathbf{z}_i$ при $t \geqslant 2$ и $\mathbf{w}_1 = \mathbf{0}$. Тогда

$$\mathbf{w}_{t+1} = \frac{1}{t}(\mathbf{z}_t + (t-1)\mathbf{w}_t) = \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \frac{1}{t}\mathbf{z}_t.$$

Запишем это равенство в виде $\mathbf{w}_{t+1} - \mathbf{z}_t = \left(1 - \frac{1}{t}\right)(\mathbf{w}_t - \mathbf{z}_t)$. Тогда

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) = \frac{1}{2}\|\mathbf{w}_t - \mathbf{z}_t\|^2 - \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{z}_{t+1}\|^2 =$$
$$\frac{1}{2}\left(1 - \left(1 - \frac{1}{t}\right)^2\right)\|\mathbf{w}_t - \mathbf{z}_t\|^2 \leqslant \frac{1}{t}\|\mathbf{w}_t - \mathbf{z}_t\|^2.$$

Обозначим $B_T = \max_{1 \leqslant t \leqslant T} \|\mathbf{z}_t\|$. Тогда из определения $\|\mathbf{w}_t\| \leqslant B_T$ и $\|\mathbf{w}_t - \mathbf{z}_t\| \leqslant 2B_T$ при $1 \leqslant t \leqslant T$. По предложению 6.1 получаем оценку регрета алгоритма FTL

$$R_T^{FTL} \leqslant \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) =$$
$$\sum_{t=1}^{T} \frac{1}{2}\|\mathbf{w}_t - \mathbf{z}_t\|^2 - \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{z}_{t+1}\|^2 \leqslant$$
$$(2B_T)^2 \sum_{t=1}^{T} \frac{1}{t} \leqslant 4B_T^2(\ln T + 1).$$

Таким образом, мы получили логарифмическую оценку регрета алгоритма FTL.

В некоторых случаях алгоритм FTL может иметь линейный регрет.

Пример 2. Линейный регрет.

Пусть $S = [-1, 1]$ и $f_t(w) = wz_t$, где $w, z_t \in S$ и

$$z_t = \begin{cases} -\frac{1}{2} \text{ если } t = 1, \\ 1 \text{ если } t \text{ четное}, \\ -1 \text{ если } t > 1 \text{ и нечетное}. \end{cases}$$

Тогда так как $w_t = \operatorname{argmin}_{-1 \leqslant w \leqslant 1} w \sum_{i=1}^{t-1} z_i$, будет $w_t = 0, 1, -1, 1, -1, \ldots$ и $H_t = 0, 1, 2, 3, 4, \ldots$. Поэтому $R_T^{FTL}(0) = H_T - L_T(0) = T - 1$.

Алгоритм следования за регуляризованным лидером FTRL.

Недостатки FTL, вроде приведенных в примере 2, преодолеваются при переходе к онлайн оптимизации с регуляризацией.

Допустим, что задана некоторая функция $R(\mathbf{w})$ – регуляризатор. В алгоритме следования за регуляризованным лидером FTRL в протоколе (рис. 6.1) на каждом шаге t в качестве предсказания выбирается наилучшее решение

$$\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}\in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w}), \tag{6.5}$$

при этом, $\mathbf{w}_1 = \operatorname{argmin}_{\mathbf{w}\in S} R(\mathbf{w})$, а в качестве $\mathbf{w}_0$ может быть выбран произвольный элемент множества S.[2]

Предложение 6.2. *Для любого* $\mathbf{u} \in S$ *имеет место верхняя оценка регрета алгоритма FTRL относительно решения* $\mathbf{u}$

$$R_T^{FTRL}(\mathbf{u}) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant$$
$$R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}). \tag{6.6}$$

Доказательство. Обозначим $f_0 = R$ и применим алгоритм FTL при $t = 0, 1, \ldots$. Тогда по предложению 6.2

$$R_T^{FTL}(\mathbf{u}) = \sum_{t=0}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant \sum_{t=0}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) =$$
$$f_0(\mathbf{w}_0) - f_0(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) =$$
$$R(\mathbf{w}_0) - R(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}). \tag{6.7}$$

Так как

$$R_T^{FTRL}(\mathbf{u}) = f_0(\mathbf{w}_0) - f_0(\mathbf{u}) + R_T^{FTL}(\mathbf{u}) =$$
$$R(\mathbf{w}_0) - R(\mathbf{u}) + R_T^{FTL}(\mathbf{u}), \tag{6.8}$$

[2] См. ссылку к (6.1). Здесь также минимум достигается в некоторой точке $\mathbf{w} \in S$, если для каждого t функция f_t выпуклая, а в качестве регуляризатора $R(\mathbf{w})$ также используется выпуклая функция.

сокращая слагаемое $R(\mathbf{w}_0)$ в (6.7) и (6.8), получим

$$R_T^{FTRL}(\mathbf{u}) \leqslant \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) + R(\mathbf{u}) - R(\mathbf{w}_1).$$

Предложение доказано. $\triangle$

Пример 3. Оценка регрета для линейной оптимизации с регуляризацией. Пусть $f_t(\mathbf{w}) = (\mathbf{w} \cdot \mathbf{z}_t)$, где $\mathbf{w}, \mathbf{z}_t \in \mathcal{R}^k$, и $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|^2$, где $\eta > 0$. В этом случае при $t > 1$, решая задачу минимизации (6.5), получим $\sum_{i=1}^{t-1} \mathbf{z}_i + \frac{1}{\eta}\mathbf{w}_t = \mathbf{0}$, откуда $\mathbf{w}_t = -\eta \sum_{i=1}^{t-1} \mathbf{z}_i$. Можно также переписать это в виде $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta z_t$. Мы получили частный случай алгоритма онлайн градиентного спуска для линейных функций и $S = \mathcal{R}^k$.

Оценка регрета дается в следующем утверждении.

Предложение 6.3. *В случае линейной оптимизации с регуляризацией, для произвольного* $\mathbf{u}$ *регрет алгоритма FTRL относительно решения* $\mathbf{u}$ *имеет верхнюю оценку*

$$R_T^{FTRL}(\mathbf{u}) \leqslant \frac{1}{2\eta}\|\mathbf{u}\|^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|^2.$$

Доказательство. По предложению 6.2

$$R_T^{FTRL}(\mathbf{u}) \leqslant R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) =$$
$$\frac{1}{2\eta}\|\mathbf{u}\|^2 + \sum_{t=1}^{T}((\mathbf{w}_t - \mathbf{w}_{t+1}) \cdot \mathbf{z}_t) =$$
$$\frac{1}{2\eta}\|\mathbf{u}\|^2 + \eta \sum_{t=1}^{T} \|\mathbf{z}_t\|^2,$$

так как $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{z}_t$ – результат решения задачи минимизации в алгоритме FTRL. $\triangle$

Частный случай предложени 6.3 представлен в виде следующего следствия.

Следствие 6.1. *Пусть* $\mathbf{u} \in \{\mathbf{u} : \|\mathbf{u}\| \leqslant B\}$ *и* $\|\mathbf{z}_t\| \leqslant D$ *для всех* t. *Тогда при* $\eta = \frac{B}{2D\sqrt{2T}}$ *будет* $R_T^{FTRL}(\mathbf{u}) \leqslant BD\sqrt{2T}$ *для любого* T.

Линериализация выпуклых функций. Задано выпуклое множество $S \subseteq \mathcal{R}^k$ Будем использовать характеристическое свойство выпуклых функций: функция $f : S \to \mathcal{R}$ является выпуклой тогда и только тогда, когда для каждого $\mathbf{w} \in S$ существует $\mathbf{z}$ такое, что для каждого $\mathbf{u} \in S$ выполнено $f(\mathbf{u}) \geqslant f(\mathbf{w}) + ((\mathbf{u} - \mathbf{w}) \cdot \mathbf{z})$. Такое $\mathbf{z}$ называется субградиентом в точке $\mathbf{w}$. Субградиент может быть не единственным, множество всех субградиентов в точке $\mathbf{w}$ обозначается $\partial f(\mathbf{w})$.

Если функция дифференцируема, то субградиент единственный и равен вектору производной $\nabla f(\mathbf{w}) = \frac{df(\mathbf{w})}{d\mathbf{w}}$ функции от многих переменных $f(\mathbf{w})$. Мы часто будем и в общем случае обозначать произвольный градиент $\mathbf{z} = \nabla f(\mathbf{w})$.

Функция $f(\mathbf{w})$ удовлетворяет условию Липшица с константой L относительно нормы $\|\cdot\|$, если $|f(\mathbf{w}) - f(\mathbf{u})| \leqslant L\|\mathbf{w} - \mathbf{u}\|$ для всех $\mathbf{w}, \mathbf{u} \in S$.[3]

Следующее утверждение устанавливает эквивалентость свойства функции удовлетворять условию Липшица и ограниченностью двойственной нормы ее субградиента. Двойственная норма определяется как $\|\mathbf{z}\|_* = \max_{\mathbf{v}:\|\mathbf{v}\|=1}(\mathbf{v} \cdot \mathbf{z})$. Так как $\|\mathbf{z}\|_* = \max_{\mathbf{v}}(\frac{\mathbf{v}}{\|\mathbf{v}\|} \cdot \mathbf{z})$, выполнено $\|\mathbf{z}\|_* \geqslant \frac{1}{\|\mathbf{v}\|}(\mathbf{v} \cdot \mathbf{z})$ для любого $\mathbf{v}$. Отсюда $(\mathbf{v} \cdot \mathbf{z}) \leqslant \|\mathbf{z}\|_*\|\mathbf{v}\|$ для любого $\mathbf{v}$.

Легко видеть, что норма, двойственная к эвклидовой норме, также эвклидова.

Предложение 6.4. *Выпуклая функция* $f : S \to \mathcal{R}$ *удовлетворяет условию Липшица с константой* L *относительно нормы* $\|\cdot\|$ *тогда и только тогда, когда для любого* $\mathbf{w} \in S$ *двойственная норма любого ее субградиента* $\mathbf{z} \in \partial f(\mathbf{w})$ *ограничена числом* L.

Доказательство. Пусть f удовлетворяет условию Липшица с константой L и $\mathbf{w} \in S$ любое. Пусть $\mathbf{v}_* = \operatorname{argmax}_{\mathbf{v}:\|\mathbf{v}\|=1}(\mathbf{v} \cdot \mathbf{z})$ и $\mathbf{u} = \mathbf{w} + \mathbf{v}_*$. Тогда $\|\mathbf{u} - \mathbf{w}\| = 1$ и $((\mathbf{u} - \mathbf{w}) \cdot \mathbf{z}) = \|\mathbf{z}\|_*$. По определению

[3] Это определение и следующее ниже предложение 6.4 имеют смысл для произвольной нормы.

субградиента $f(\mathbf{u}) - f(\mathbf{w}) \geqslant (\mathbf{z} \cdot (\mathbf{u} - \mathbf{w})) = \|\mathbf{z}\|_*$. Добавляя условие Липшица, получим $L = L\|\mathbf{u} - \mathbf{w}\| \geqslant f(\mathbf{u}) - f(\mathbf{w}) \geqslant \|\mathbf{z}\|_*$.

Пусть теперь $\mathbf{z} \in \partial f(\mathbf{w})$ и $\|\mathbf{z}\|_* \leqslant L$. Из определения субградиента $f(\mathbf{w}) - f(\mathbf{u}) \leqslant (\mathbf{z} \cdot (\mathbf{w} - \mathbf{u})) \leqslant \|\mathbf{z}\|_* \|\mathbf{w} - \mathbf{u}\| \leqslant L\|\mathbf{w} - \mathbf{u}\|$, т.е функция f удовлетворяет условию Липшица. $\triangle$

Теорема 6.1. *Рассмотрим алгоритм FTRL с регуляризатором* $R(\mathbf{u}) = \frac{1}{2\eta}\|\mathbf{u}\|^2$. *Пусть функции* $f_1, f_2, \ldots$ *выпуклые и удовлетворяют условию Липшица с одной и той же константой* D. *Тогда для любого* $\mathbf{u} \in S$

$$R_T^{FTRL}(\mathbf{u}) \leqslant \frac{1}{2\eta}\|u\|^2 + \eta D^2 T$$

для всех T.

Если к тому же $\mathbf{u} \in S = \{\mathbf{v} : \|\mathbf{v}\| \leqslant B\}$ *и горизонт прогнозирования* T *заранее задан, то при* $\eta = \frac{B}{2D\sqrt{2T}}$ *будет иметь место оценка регрета* $R_T^{FTRL}(\mathbf{u}) \leqslant BD\sqrt{2T}$.

Доказательство. Утверждение следует из лемм 6.1 и 6.2.

Лемма 6.1. *Пусть* $f_1, f_2, \ldots$ *– последовательность выпуклых функций из протокола (рис.6.1), регуляризатор* $R(\mathbf{w})$ *также выпуклая функция. Пусть также* $\mathbf{w}_1, \mathbf{w}_2, \ldots$ *– последовательность решений предсказателя. Тогда*

$$R_T^{FTRL}(\mathbf{u}) \leqslant R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T}((\mathbf{w}_t - \mathbf{w}_{t+1}) \cdot \nabla f_t(\mathbf{w}_t)).$$

Доказательство. Для любого t и $\mathbf{u} \in S$ выполнено

$$f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant ((\mathbf{w}_t - \mathbf{u}) \cdot \nabla f_t(\mathbf{w}_t)),$$

где $\nabla f_t(\mathbf{w}_t) \in \partial f(\mathbf{w}_t)$ – любой субградиент функции f в точке $\mathbf{w}_t$.

Отсюда по предложению 6.3 для любого $\mathbf{u} \in S$

$$R_T^{FTRL}(\mathbf{u}) \leqslant R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leqslant$$

$$R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leqslant$$

$$R(\mathbf{u}) - R(\mathbf{w}_1) + \sum_{t=1}^{T} ((\mathbf{w}_t - \mathbf{w}_{t+1}) \cdot \nabla f_t(\mathbf{w}_t)).$$

$\triangle$

Лемма 6.2. *Для любого t выполнено $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leqslant \eta D$.*

Доказательство. Прямым вычислением получаем

$$\sum_{i=1}^{t-1} (\nabla f_i(\mathbf{w}_i) \cdot \mathbf{w}) + \frac{1}{2\eta}\|w\|^2 =$$

$$\frac{1}{2\eta}\left(\|w\|^2 + 2\eta \sum_{i=1}^{t-1} (\nabla f_i(\mathbf{w}_i) \cdot \mathbf{w})\right) =$$

$$\frac{1}{2\eta}\left\|\mathbf{w} - \big(-\eta \sum_{i=1}^{t-1} \nabla f_i(\mathbf{w}_i)\big)\right\|^2 - \frac{1}{2}\eta \left\|\sum_{i=1}^{t-1} \nabla f_i(\mathbf{w}_i)\right\|^2 \tag{6.9}$$

Так как вычитаемое (6.9) не зависит от $\mathbf{w}$, правило (6.5) можно заменить на

$$\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} \left\|\mathbf{w} - \big(-\eta \sum_{i=1}^{t-1} \nabla f_i(\mathbf{w}_i)\big)\right\| =$$

$$P_S\big(-\eta \sum_{i=1}^{t-1} \nabla f_i(\mathbf{w}_i)\big).$$

Отсюда

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\| = \left\|P_S\big(-\eta \sum_{i=1}^{t} \nabla f_i(\mathbf{w}_i)\big) - P_S\big(-\eta \sum_{i=1}^{t-1} \nabla f_i(\mathbf{w}_i)\big)\right\| \leqslant$$

$$\|-\eta \nabla f_t(\mathbf{w}_t)\| \leqslant \eta D,$$

где $P_S(\mathbf{z}) = \min_{\mathbf{u}\in S} \|\mathbf{u} - \mathbf{z}\|$. Мы использовали свойство проекции $\|P_S(\mathbf{y}) - P_S(\mathbf{z})\| \leqslant \|\mathbf{y} - \mathbf{z}\|$ для любых $\mathbf{y}$ и $\mathbf{z}$, а также равномерную ограниченность градиента. $\triangle$

Утверждение теоремы 6.1 является прямым следствием этих двух лемм. $\triangle$

Детальное изложение этих вопросов можно найти в [48].

6.2. Онлайн градиентный спуск

Задачу предсказанию с использованием экспертных стратегий можно также рассматривать в терминах теории онлайн оптимизации [74].

Рассматриваем эвклидову норму $\|\cdot\| = \|\cdot\|_2$ на $\mathcal{R}^n$. Заданы выпуклое замкнутое множество $S \subseteq \mathcal{R}^n$ и класс $\mathcal{F}$ выпуклых вещественных функций определенных на этом множестве.

Пусть $d(x, y) = \|x - y\|$ – эвклидово расстояние между точками (векторами) $x, y \in \mathcal{R}^n$ и $(x \cdot y)$ – соответствующее скалярное произведение.[4]

Обозначим $P(y) = \operatorname{argmin}_{x\in S} d(x, y)$ – эвклидова проекция произвольной точки $y \in \mathcal{R}^n$ на выпуклое замкнутое подмножество S.

Далее будут использоваться следующие допущения.

- Множество S ограниченное: существует константа C_S такая, что $d(x, y) \leqslant C_S$ для всех $x, y \in S$.
- Для каждой функции $c \in \mathcal{F}$ и $x \in S$ существует субградиент $g(x)$, для которого выполнено $c(y) - c(x) \geqslant g(x)(y - x)$ для всех $y \in S$. Эту функцию $g(x)$ также обозначаем $\nabla c(x)$.
- Существует константа $C_{\mathcal{F}}$ такая, что $\|\nabla c(x)\| \leqslant C_{\mathcal{F}}$ для всех $c \in \mathcal{F}$ и всех $x \in S$.
- Существуют алгоритмы для вычисления значений $\nabla c(x)$ и $P(y) = \operatorname{argmin}_{x\in S} d(x, y)$.

Протокол алгоритма онлайн градиентного спуска представлен на рис. 6.2.

[4] В этом и следующем разделах мы не подчеркиваем векторы сверху.

Алгоритм онлайн градиентного спуска
Полагаем $x_1 \in S$ – любая точка, $L_0 = 0$, $\eta_t = 1/\sqrt{t}$.
FOR $t = 1,\ 2,\ \ldots$
Предсказатель выдает $x_t \in S$
Природа выдает $c_t \in \mathcal{F}$
Кумулятивные потери $L_t = L_{t-1} + c_t(x_t)$
Подготавливаем прогноз для следующего шага
$x_{t+1} = P(x_t - \eta_t \nabla c_t(x_t))$
ENDFOR

Рис. 6.2: Алгоритм онлайн градиентного спуска

Регрет алгоритма за первые T шагов определяется

$$R_T = \sum_{t=1}^{T} c_t(x_t) - \min_{x \in S} \sum_{t=1}^{T} c_t(x).$$

Пример: $S \subset \mathcal{R}^n$ – стандартный n-мерный симплекс в $\mathcal{R}^n$, $x_t = (x_{1,t}, \ldots, x_{n,t})$, где $x_{i,t} \geqslant 0$ и $\sum\limits_{i=1}^{n} x_{t,i} = 1$ для всех t, $c_t(x) = (l_t \cdot x)$, где вектор $l_t \in [0,1]^n$ задает функцию c_t и $x \in S$ – ее аргумент.

Теорема 6.2. *Алгоритм онлайн градиентного спуска имеет верхнюю оценку регрета*

$$R_T \leqslant \frac{1}{2} C_S^2 \sqrt{T} + (\sqrt{T} - \frac{1}{2}) C_{\mathcal{F}}^2.$$

Доказательство. Сначала убедимся в том, что достаточно провести оценку регрета для линейных функций вида $c_t(x) = (g_t(x_t) \cdot x)$. Действительно, для произвольной выпуклой функции $c_t \in \mathcal{F}$ будет $c_t(x) - c_t(x_t) \geqslant \nabla c_t(x_t)(x - x_t)$ для всех x.[5] Пусть $x^* = \operatorname{argmin}_{x \in S} \sum\limits_{t=1}^{T} c_t(x)$. Для этой точки выполнено $c_t(x^*) - c_t(x_t) \geqslant \nabla c_t(x_t)(x^* - x_t)$. Отсюда

$$c_t(x_t) - c_t(x^*) \leqslant -\nabla c_t(x_t)(x^* - x_t) = (\nabla c_t(x_t) \cdot x_t) - (\nabla c_t(x_t) \cdot x^*).$$

[5]Здесь и далее произведения типа $\nabla c_t(x_t)(x - x_t)$ понимаются как скалярные.

Тогда

$$R_T = \sum_{t=1}^{T} c_t(x_t) - c_t(x^*) \leqslant \sum_{t=1}^{T} (\nabla c_t(x_t) \cdot x_t) - \sum_{t=1}^{T} (\nabla c_t(x_t) \cdot x^*).$$

Далее оценим регрет алгоритма для функций вида $c_t(x) = (g_t(x_t) \cdot x)$.

Обозначим $y_{t+1} = x_t - \eta_t g_t$, где $g_t = \nabla c_t(x_t)$, $x_{t+1} = P(y_{t+1})$.

Имеем

$$y_{t+1} - x^* = x_t - x^* - \eta_t g_t.$$

Возводим обе части этого равенства в квадрат (скалярное произведение вектора на себя):

$$(y_{t+1} - x^*)^2 = (x_t - x^*)^2 - 2\eta_t (x_t - x^*) g_t + \eta_t^2 \|g_t\|^2. \qquad (6.10)$$

Из свойства проекции на выпуклое множество имеем $\|x - y\| \geqslant \|P(y) - x\|$ для всех $x \in S$ и $y \in \mathcal{R}^n$. Отсюда $(y_{t+1} - x^*)^2 \geqslant (x_{t+1} - x^*)^2$. Отсюда и из (6.10) получаем

$$(x_{t+1} - x^*)^2 \leqslant (x_t - x^*)^2 - 2\eta_t (x_t - x^*) g_t + \eta_t^2 C_{\mathcal{F}}^2.$$

или

$$(x_t - x^*) g_t \leqslant \frac{1}{2\eta_t} ((x_t - x^*)^2 - (x_{t+1} - x^*)^2) + \frac{\eta_t}{2} C_{\mathcal{F}}^2. \qquad (6.11)$$

Суммируем (6.11) по t и получаем, с учетом (6.10),

$$\begin{aligned} R_T &\leqslant \sum_{t=1}^{T} (x_t - x^*) g_t \leqslant \\ &\leqslant \sum_{t=1}^{T} \frac{1}{2\eta_t} ((x_t - x^*)^2 - (x_{t+1} - x^*)^2) + \frac{\eta_t}{2} C_{\mathcal{F}}^2 = \\ &= \frac{1}{2\eta_1} (x_1 - x^*)^2 + \sum_{t=2}^{T} \frac{1}{2\eta_{t-1}} (x_t - x^*)^2 + \\ &+ \frac{1}{2} \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (x_t - x^*)^2 - \sum_{t=1}^{T} \frac{1}{2\eta_t} (x_{t+1} - x^*)^2 + \\ &+ \frac{C_{\mathcal{F}}^2}{2} \sum_{t=1}^{T} \eta_t. \qquad (6.12) \end{aligned}$$

Второе и четвертое слагаемые суммы (6.12) упрощаются

$$\sum_{t=2}^{T} \frac{1}{2\eta_{t-1}}(x_t - x^*)^2 - \sum_{t=1}^{T} \frac{1}{2\eta_t}(x_{t+1} - x^*)^2 = \\ = -\frac{1}{2\eta_T}(x_{T+1} - x^*)^2 \leqslant 0. \qquad (6.13)$$

Продолжаем (6.12) с учетом (6.13)

$$R_T \leqslant \frac{1}{2\eta_1}(x_1 - x^*)^2 + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(x_t - x^*)^2 + \frac{C_{\mathcal{F}}^2}{2}\sum_{t=1}^{T}\eta_t \leqslant \\ \leqslant C_S^2\left(\frac{1}{2\eta_1} + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\right) + \frac{C_{\mathcal{F}}^2}{2}\sum_{t=1}^{T}\eta_t \leqslant \\ \leqslant C_S^2\frac{1}{2\eta_T} + \frac{C_{\mathcal{F}}^2}{2}\sum_{t=1}^{T}\eta_t \leqslant \frac{C_S^2}{2}\sqrt{T} + \frac{C_{\mathcal{F}}^2}{2}(2\sqrt{T} - 1).$$

Теорема доказана. $\triangle$

Логарифмическая оценка регрета. Для широкого класса функций можно получить логарифмическую оценку регрета. Функция $c(x)$ называется сильно выпуклой, если

$$c(\mathbf{x}) - c(\mathbf{y}) \geqslant (\mathbf{x} - \mathbf{y}) \cdot \nabla c(\mathbf{y}) + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

для любых $\mathbf{x}, \mathbf{y} \in S$, где $\sigma > 0$ – модуль сильной выпуклости. Подробнее о сильно выпуклых функциях см. в разделе 6.3.

В доказательстве теоремы 6.2 имеем при $g_t = \nabla c_t(\mathbf{x}_t)$

$$c_t(\mathbf{x}^*) - c(\mathbf{x}_t) \geqslant (\mathbf{x}^* - \mathbf{x}_t) \cdot g_t + \frac{\sigma}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2.$$

Отсюда следует, что можно модифицировать неравенство (6.12) в следующем виде. Полагаем при этом $\eta_t = \frac{1}{\sigma t}$.

$$R_T \leqslant \sum_{t=1}^{T}(x_t - x^*)g_t - \frac{\sigma}{2}(\mathbf{x}_t - \mathbf{x}^*)^2 \leqslant$$

$$\leqslant \sum_{t=1}^{T} \frac{1}{2\eta_t}\left((x_t - x^*)^2 - (x_{t+1} - x^*)^2 - \eta_t\sigma(\mathbf{x}_t - \mathbf{x}^*)^2\right) +$$

$$\frac{\eta_t}{2}C_{\mathcal{F}}^2 = \frac{1}{2}\left(\frac{1}{\eta_1} - \sigma\right)(x_1 - x^*)^2 + \sum_{t=2}^{T}\frac{1}{2\eta_{t-1}}(x_t - x^*)^2 +$$

$$\frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma\right)(x_t - x^*)^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}(x_{t+1} - x^*)^2 +$$

$$\frac{C_{\mathcal{F}}^2}{2}\sum_{t=1}^{T}\eta_t \leqslant \frac{C_{\mathcal{F}}^2}{2}\sum_{t=1}^{T}\eta_t. \quad (6.14)$$

Сумма второго и четвертого слагаемых как и в (6.13) не превосходит 0. Первое слагаемое равно 0, так как $\eta_1 = 1/\sigma$. Третье слагаемое также равно 0, так как $\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \sigma = 0$. С учетом этого, продолжаем (6.14)

$$R_T \leqslant \frac{C_{\mathcal{F}}^2}{2\sigma}\sum_{t=1}^{T}\frac{1}{t} \leqslant \frac{C_{\mathcal{F}}^2}{2\sigma}(\ln T + 1).$$

Таким образом, имеет место следующее утверждение

Теорема 6.3. *Пусть в алгоритме онлайн градиентного спуска (рис. 6.2) функции c_t являются сильно выпуклыми с модулем сильной выпуклости σ и выполнены все условия теоремы 6.2. Тогда имеет место логарифмическая оценка регрета*

$$R_T \leqslant \frac{C_{\mathcal{F}}^2}{2\sigma}(\ln T + 1).$$

6.3. Онлайн зеркальный спуск и связанные с ним вопросы

Приведем алгоритм онлайн оптимизации общего вида основанного на методе зеркального спуска, частными случаями которого

являются онлайн градиентный спуск и алгоритм экспоненциального смешивания.

Приведем необходимые определения. Пусть $(x \cdot y)$ – скалярное произведение на эвклидовом пространстве $\mathcal{R}^n$. Пусть $\|x\|$ – некоторая норма на $\mathcal{R}^n$. Далее будем рассматривать следующие виды норм на $\mathcal{R}^n$: 1-норма $\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|$, 2-норма (эвклидова норма) $\|x\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$ и ∞-норма $\|x\|_\infty = \max\limits_{1 \leqslant i \leqslant n} |x_i|$.

Пусть Ω – непустое подмножество $\mathcal{R}^n$ такое, что множество его внутренних точек Ω^c является выпуклым. Рассмотрим функцию R, определенную на Ω и непрерывно дифференцируемую на Ω^c. Обозначим $\nabla R(b) = (\frac{\partial R}{\partial x_1}(b), \ldots, \frac{\partial R}{\partial x_n}(b))$ – градиент функции R.[6]

Расхождение Брэгмана.

Функция $R : \Omega \to \mathcal{R}$ называется *функцией Лежандра*, если

- R – строго выпуклая и непрерывно дифференцируемая на Ω^c;
- если какая либо последовательность точек $\{x_n\}$ из Ω сходится к некоторой точке на границе этого множества, то $\|\nabla R(x_n)\| \to \infty$ при $n \to \infty$.

По произвольной функции Лежандра R определяется *расхождение Брэгмана* как разность между приращением функции и слагаемыми первого порядка в разложении Тейлора

$$D_R(x, y) = R(x) - R(y) - (\nabla R(y) \cdot (x - y)), \tag{6.15}$$

где $x \in \Omega$ и $y \in \Omega^c$. Легко видеть, что $D_R(x, y) > 0$ при $x \neq y$, $D_R(x, x) = 0$; кроме этого, можно привести примеры, когда расхождение Брэгмана не симметричная функция.

Функция R называется *сильно выпуклой* относительно нормы $\|\cdot\|$, если для всех $x \in \Omega$ и $y \in \Omega^c$ выполнено

$$R(x) - R(y) \geqslant (\nabla R(y) \cdot (x - y)) + \frac{\sigma}{2}\|x - y\|^2,$$

[6] В более общем случае можно рассматривать $\nabla R(b)$ как субградиент.

где σ – положительная константа (модуль сильной выпуклости).

Для строго выпуклой функции $D_R(x,y) > 0$, а для сильно выпуклой функции с модулем сильной выпуклости σ выполнено $D_R(x,y) \geqslant \frac{\sigma}{2}\|x-y\|^2$ для всех x и y.

Пример 1.

Пусть $R(x) = \frac{1}{2}\|x\|_2^2$ на $\Omega = \mathcal{R}^n$. Тогда $\nabla R(y) = y$ и $D_R(x,y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - (y\cdot(x-y)) = \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2 - (x\cdot y) = \frac{1}{2}\|x-y\|_2^2$. Здесь модуль сильной выпуклости относительно эвклидовой нормы равен 1.

Пример 2.

Функция $R(x) = \sum\limits_{i=1}^{n} x_i \ln x_i - x_i$ рассматривается на $\mathcal{R}_+^n$ (множество всех n-мерных векторов с неотрицательными координатами). Здесь определим по непрерывности $0\ln 0 = 0$. Для нее $(\nabla R(y))_i = \ln y_i$ при $y_i > 0$, $D_R(x,y) = \sum\limits_{i=1}^{n} x_i \ln x_i - \sum\limits_{i=1}^{n} y_i \ln y_i - \sum\limits_{i=1}^{n} \ln y_i (x_i - y_i) = \sum\limits_{i=1}^{n} x_i \ln \frac{x_i}{y_i}$ – расхождение Кульбака–Лейблера между x и y. Модуль сильной выпуклости функции R относительно 1-нормы равен 1 (см. задачу из раздела 6.4).

Лемма 6.3. $(\nabla R(x) - \nabla R(y))\cdot(z-y) = D_R(z,y) + D_R(y,x) - D_R(z,x)$.

Доказательство. Представим все расхождения в удобном виде

$$\begin{aligned} D_R(z,y) &= R(z) - R(y) - (\nabla R(y)\cdot(z-y)) \\ D_R(y,x) &= R(y) - R(x) - (\nabla R(x)\cdot(y-x)) \\ -D_R(z,x) &= -R(z) + R(x) + (\nabla R(x)\cdot(z-x)) \end{aligned}$$

Складываем эти равенства и получаем

$$D_R(z,y) + D_R(y,x) - D_R(z,x) = (\nabla R(x) - \nabla R(y))\cdot(z-y).$$

$\triangle$

В следующей лемме представлено обобщенное неравенство Пифагора для расхождения Брэгмана (см. [43]).

Лемма 6.4. *Пусть $S \subseteq \Omega$ – замкнутое выпуклое множество такое, что $S \cap \Omega^c \neq \emptyset$, $x_0 \in \Omega^c$ и $x^* = \operatorname{argmin}_{x \in S} D_R(x, x_0)$ – проекция Брэгмана точки x_0 на S.*[7] *Тогда $x^* \in S \cap \Omega^c$ и для каждого $y \in S \cap \Omega^c$*

$$D_R(y, x_0) \geqslant D_R(x^*, x_0) + D_R(y, x^*).$$

Доказательство. Определим $G(x) = D_R(x, x_0) - D_R(x, x^*)$. Из определения расхождения Брэгмана и сокращения одинаковых членов получаем

$$G(x) = -R(x_0) - (x - x_0) \cdot \nabla R(x_0) + R(x^*) + (x - x^*) \cdot \nabla R(x^*).$$

Функция G является линейной. Пусть $x_\alpha = \alpha y + (1-\alpha)x^*$ – точка на прямой, соединяющей y и x^*. Из выпуклости S, $x_\alpha \in S$ для всех α. Из линейности $G(x_\alpha) = \alpha G(y) + (1-\alpha)G(x^*)$. Тогда

$$D_R(x_\alpha, x_0) - D_R(x_\alpha, x^*) = \alpha(D_R(y, x_0) - D_R(y, x^*)) + (1-\alpha) D_R(x^*, x_0).$$

Отсюда при $\alpha > 0$ будет

$$\begin{aligned} D_R(y, x_0) - D_R(y, x^*) - D_R(x^*, x_0) = \\ = \frac{D_R(x_\alpha, x_0) - D_R(x_\alpha, x^*) - D_R(x^*, x_0)}{\alpha} \geqslant \\ \geqslant \frac{-D_R(x_\alpha, x^*)}{\alpha}. \end{aligned} \tag{6.16}$$

Здесь мы использовали неравенство $D_R(x_\alpha, x_0) \geqslant D_R(x^*, x_0)$.

Определим $D(x) = D_R(x, x^*)$. По второму свойству функции Лежандра, так как $x_0 \in \Omega^c$ и x^* – точка минимума, она не может находиться на границе множества Ω, т.е. $x^* \in \Omega^c$. Поэтому по первому свойству функции Лежандра производная функции D по направлению $y - x^*$ в точке x^*

$$D'_{y-x^*}(x^*) = \lim_{\alpha \to 0+} \frac{D(x_\alpha)}{\alpha} = \lim_{\alpha \to 0+} \frac{D(x^* + \alpha(y - x^*)) - D(x^*)}{\alpha}$$

существует и конечна.

Имеем $\nabla D(x^*) = \nabla_x D_R(x^*, x^*) = \nabla R(x^*) - \nabla R(x^*) = 0$. Отсюда $D'_{y-x^*}(x^*) = (y - x^*) \cdot \nabla D(x^*) = 0$. Так как в (6.16)

[7] Нетрудно доказать, что такая точка x^* существует и единственная.

$\lim\limits_{\alpha\to 0+} \frac{-D_R(x_\alpha,x^*)}{\alpha} = 0$, а левая часть (6.16) не зависит от α, будет $D_R(y,x_0) - D_R(y,x^*) - D_R(x^*,x_0) \geqslant 0$, что и требовалось доказать. $\triangle$

Выпуклое сопряжение.

Пусть R – функция Лежандра, определенная на множестве Ω. Определим операцию сопряжения

$$R^*(p) = \sup_{x\in\Omega}((p\cdot x) - R(x)). \tag{6.17}$$

Так как множество Ω^c выпуклое, а функция R строго выпуклая, точка x, где достигается максимум в (6.17), единственная, по второму свойству функции Лежандра $x \in \Omega^c$, при этом $p = \nabla R(x)$. Нетрудно проверить, что R^* – также функция Лежандра, определенная на множестве $\Omega^* = \nabla R(\Omega^c)$. Называем это множество двойственным пространством.

Можно расмотреть функцию сопряженную к R^*

$$(R^*)^*(x) = R^{**}(x) = \sup_{p\in\Omega^*}((p\cdot x) - R^*(p)). \tag{6.18}$$

Нетрудно доказать, что $R^{**}(x) = R(x)$ (см.задачу из раздела 6.4).

Лемма 6.5. *При $x \in \Omega$ и $p \in \Omega^*$ будет $p = \nabla R(x)$ тогда и только тогда, когда $x = \nabla R^*(p)$.*

Доказательство. Если для некоторых p и x выполнено $R^*(p) = (p\cdot x) - R(x)$, то x есть точка максимума в левой части (6.17) и поэтому $p = \nabla R(x)$. Обратное также верно – если для p и x выполнено $p = \nabla R(x)$, то x является точкой максимума в (6.17) и выполнено $R^*(p) = (p\cdot x) - R(x)$.

Таким образом, $R^*(p) = (p\cdot x) - R(x)$ тогда и только тогда, когда $p = \nabla R(x)$.

Аналогично, в (6.18), $R^{**}(x) + R^*(p) = (x\cdot p)$ выполнено тогда и только тогда, когда $x = \nabla R^*(p)$.

Так как $R = R^{**}$, равенство $R^{**}(x) + R^*(p) = (x\cdot p)$ совпадает с равенством $R(x) + R^*(p) = (x\cdot p)$.

Следовательно, $p = \nabla R(x)$ тогда и только тогда, когда $x = \nabla R^*(p)$. $\triangle$

Алгоритм зеркального спуска.
Полагаем $\hat{y}_1 \in S \cap \Omega^c$ – начальная точка, $L_0 = 0$, $\eta > 0$.
FOR $t = 1, \ldots, T$
Предсказатель выдает $\hat{y}_t$ (вычислено на предыдущем шаге).
Природа выдает $x_t \in F$.
Кумулятивные потери $L_t = L_{t-1} + (\hat{y}_t \cdot x_t)$.
Подготавливаем прогноз для следующего шага
$\tilde{y}_{t+1} = \nabla R^*(\nabla R(\hat{y}_t) - \eta x_t)$ (эта операция определена при условии $\nabla R(y) - \eta x \in \Omega^*$ при всех $y \in \Omega^c$ и $x \in F$).
$\hat{y}_{t+1} = \mathrm{argmin}_{y \in S} D_R(y, \tilde{y}_{t+1})$.
ENDFOR

Рис. 6.3: Алгоритм онлайн зеркального спуска

Из леммы 6.5 прямыми преобразованиями легко получить следующее равенство.

Следствие 6.2. $D_R(x, y) = D_{R^*}(\nabla R(x), \nabla R(y))$.

Лемма 6.5 и следствие 6.2 показывают механизм использования функции Лежандра: отображение $\nabla R(x)$ отображает пространство Ω^c в двойственное пространство Ω^* (образ $\nabla R(\Omega^c)$), при этом $\nabla R^*(x)$ – обратное отображение и сохраняется величина расхождения Брэгмана между любыми двумя точками и между их образами.

Алгоритм онлайн зеркального спуска.

Переходим к изложению метода онлайн зеркального спуска. Задана норма $\|x\|$ на $\mathcal{R}^n$. Пусть $\Omega \subseteq \mathcal{R}^n$ и задана сильно выпуклая с модулем сильной выпуклости σ относительно нормы $\|\cdot\|$ функция Лежандра R, определенная на Ω ($\Omega^* = \nabla R(\Omega^c)$ – двойственное пространство). Заданы также замкнутое выпуклое множество $S \subseteq \Omega$ такое, что $S \cap \Omega^c \neq \emptyset$, а также замкнутое выпуклое множество $F \subseteq \mathcal{R}^n$ (см. примеры 1а и 2а ниже).

Рассматривается повторяющаяся игра. На каждом раунде t предсказатель выдает решение $\hat{y}_t \in S$ и получает ответ $x_t \in F$. Потери от этого решения исчисляются в виде $(\hat{y}_t \cdot x_t)$. Протокол алгоритма онлайн зеркального спуска представлен на рис. 6.3.

Регрет алгоритма за первые T шагов определяется

$$R_T = \sum_{t=1}^{T}(\hat{y}_t \cdot x_t) - \min_{y \in S} \sum_{t=1}^{T}(y \cdot x_t).$$

Теорема 6.4. *Пусть $\eta > 0$ и $Y = \sup_{y,y' \in S}(R(y) - R(y'))$, $X = \sup_{x \in F} \|x\|_*$.*[8] *Тогда алгоритм онлайн зеркального спуска имеет верхнюю оценку регрета*

$$R_T \leqslant \frac{Y}{\eta} + \frac{\eta}{2\sigma} T X^2.$$

При $\eta = \frac{1}{X}\sqrt{\frac{2\sigma Y}{T}}$ эта оценка принимает вид

$$R_T \leqslant \sqrt{\frac{2YX^2}{\sigma} T}.$$

Эквивалентная формулировка: для любого T и $\eta_T = \frac{1}{X}\sqrt{\frac{2\sigma Y}{T}}$ выполнено

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t \cdot x_t) - \frac{1}{T}\min_{y \in S}\sum_{t=1}^{T}(y \cdot x_t) \leqslant \sqrt{\frac{2YX^2}{\sigma T}},$$

как бы Природа не выбирала свои ходы.

Доказательство. Пусть $y^* = \operatorname{argmin}_{y \in S} \sum\limits_{t=1}^{T}(y \cdot x_t)$. По определению

$$\eta R_T = \eta \sum_{t=1}^{T}((\hat{y}_t - y^*) \cdot x_t) =$$

$$= \eta \sum_{t=1}^{T}((\hat{y}_t - \tilde{y}_{t+1}) \cdot x_t) + \eta \sum_{t=1}^{T}((\tilde{y}_{t+1} - y^*) \cdot x_t). \tag{6.19}$$

Проведем оценку слагаемого первой суммы из правой части (6.19).

$$\eta((\hat{y}_t - \tilde{y}_{t+1}) \cdot x_t) \leqslant \eta\|\hat{y}_t - \tilde{y}_{t+1}\| \cdot \|x_t\|_* =$$
$$= (\sqrt{\sigma}\|\hat{y}_t - \tilde{y}_{t+1}\|)(\frac{\eta}{\sqrt{\sigma}}\|x_t\|_*). \tag{6.20}$$

[8] Двойственная норма определяется $\|x\|_* = \sup_{\|u\| \leqslant 1}(x \cdot u)$.

Здесь мы использовали неравенство $(y \cdot x) \leqslant \|y\| \cdot \|x\|_*$ для всех x и y, где $\|x\|_*$ – двойственная норма.[9]

Из неравенства $ab \leqslant \frac{a^2}{2} + \frac{b^2}{2}$ и сильной выпуклости функции R получим оценку для произвольного слагаемого (6.20) первой суммы из правой части (6.19)

$$\eta((\hat{y}_t - \tilde{y}_{t+1}) \cdot x_t) \leqslant \frac{\sigma}{2}\|\hat{y}_t - \tilde{y}_{t+1}\|^2 + \frac{\eta^2}{2\sigma}\|x_t\|_*^2 \leqslant \\ \leqslant D_R(\tilde{y}_{t+1}, \hat{y}_t) + \frac{\eta^2}{2\sigma}X^2. \qquad (6.21)$$

Получим теперь верхнюю оценку слагаемого второй суммы из правой части (6.19). По определению $\tilde{y}_{t+1} = \nabla R^*(\nabla R(\hat{y}_t) - \eta x_t)$. По лемме 6.5 $\nabla R(\tilde{y}_{t+1}) = \nabla R(\hat{y}_t) - \eta x_t$ или $\eta x_t = \nabla R(\hat{y}_t) - \nabla R(\tilde{y}_{t+1})$. Отсюда используя лемму 6.3 получим

$$\eta((\tilde{y}_{t+1} - y^*) \cdot x_t) = (\tilde{y}_{t+1} - y^*) \cdot (\nabla R(\hat{y}_t) - \nabla R(\tilde{y}_{t+1})) = \\ = D_R(y^*, \hat{y}_t) - D_R(y^*, \tilde{y}_{t+1}) - D_R(\tilde{y}_{t+1}, \hat{y}_t) \leqslant \\ \leqslant D_R(y^*, \hat{y}_t) - D_R(y^*, \hat{y}_{t+1}) - D_R(\tilde{y}_{t+1}, \hat{y}_t). \quad (6.22)$$

Здесь мы использовали неравенство $D_R(y^*, \tilde{y}_{t+1}) \geqslant D_R(y^*, \hat{y}_{t+1})$, которое следует из леммы 6.4.

Суммируя (6.22) по t получим

$$\eta \sum_{t=1}^{T}((\tilde{y}_{t+1} - y^*) \cdot x_t) \leqslant \\ \leqslant D_R(y^*, \hat{y}_1) - D_R(y^*, \hat{y}_{T+1}) - \sum_{t=1}^{T} D_R(\tilde{y}_{t+1}, \hat{y}_t). \qquad (6.23)$$

[9]Из определения двойственной нормы $\|x\|_* = \sup_{\|u\| \leqslant 1}(u \cdot x) = \sup_u(\frac{u}{\|u\|} \cdot x) \geqslant \frac{1}{\|y\|}(y \cdot x)$. Отсюда получаем необходимое неравенство.

Сложим неравенства (6.21) и (6.23)

$$\eta R_T = \sum_{t=1}^{T} (\hat{y}_t - y^*) \cdot x_t \leqslant \sum_{t=1}^{T} D_R(\tilde{y}_{t+1}, \hat{y}_t) + \frac{\eta^2}{2\sigma} T X^2 +$$

$$+ D_R(y^*, \hat{y}_1) - D_R(y^*, \hat{y}_{T+1}) - \sum_{t=1}^{T} D_R(\tilde{y}_{t+1}, \hat{y}_t) \leqslant$$

$$\leqslant D_R(y^*, \hat{y}_1) + \frac{\eta^2}{2\sigma} T X^2 \leqslant$$

$$\leqslant R(y^*) - R(\hat{y}_1) + \frac{\eta^2}{2\sigma} T X^2 \leqslant$$

$$\leqslant Y + \frac{\eta^2}{2\sigma} T X^2 \qquad (6.24)$$

Из (6.24) получаем оценку регрета $R_T \leqslant \frac{Y}{\eta} + \frac{\eta}{2\sigma} T X^2$. Оптимизируем по η и получим $\eta = \frac{1}{X}\sqrt{\frac{2\sigma Y}{T}}$ и $R_T \leqslant \sqrt{\frac{2YX^2}{\sigma} T}$. Теорема доказана. $\triangle$

Вернемся к примерам 1 и 2.

Пример 1а.

Пусть $S = F = \mathcal{R}^n$, $R(y) = \frac{1}{2}\|y\|_2^2$, где $y = (y_1, \ldots, y_n)$. Тогда $\nabla R(y) = y$ и $\nabla R^*(p) = p$. Тогда алгоритм зеркального спуска имеет вид

$$\hat{y}_{t+1} = \nabla R^*(\nabla R(y_t) - \eta x_t) = \hat{y}_t - \eta x_t,$$
$$\tilde{y}_{t+1} = \hat{y}_{t=1}. \qquad (6.25)$$

и представляет собой алгоритм градиентного спуска с постоянным параметром $\eta > 0$.

Пример 2а.

Пусть $S = \{x \in \mathcal{R}^n : x_i \geqslant 0, \sum\limits_{i=1}^{n} x_i = 1\}$ есть n-мерный симплекс в $\mathcal{R}^n$, $F = \mathcal{R}^n$, $R(y) = \sum\limits_{i=1}^{n} y_i \ln y_i - y_i$. Тогда $\nabla R(y) = (\ln y_1, \ldots \ln y_n)$. Обозначим $\nabla R(y) = \ln y$. Если $z = \nabla R(y)$, то $y = \nabla R^*(z)$. В координатах $z_i = \ln y_i$ и $y_i = e^{z_i}$, т.е., $y =$

$\nabla R^*(z) = e^z = (e^{z_1}, \ldots, e^{z_n})$. Отсюда получаем выражение аналогичное (6.25) в координатах

$$\tilde{y}_{i,t+1} = (\nabla R^*(\nabla R(\hat{y}_t) - \eta x_t))_i = e^{\ln \hat{y}_{i,t} - \eta x_{i,t}} = \hat{y}_{i,t} e^{-\eta x_{i,t}}.$$

Проекция Брэгмана на симплекс – это нормализация в 1-норме (см. задачу из раздела (6.4) ниже) $\hat{y}_{t+1} = \frac{\tilde{y}_{t+1}}{\|\tilde{y}_{t+1}\|_1}$ или в координатах

$$\hat{y}_{i,t+1} = \frac{\hat{y}_{i,t} e^{-\eta x_{i,t}}}{\sum\limits_{j=1}^{n} \hat{y}_{j,t} e^{-\eta x_{j,t}}}.$$

Получаем аналог алгоритма экспоненциального смешивания Hedge.

6.4. Задачи и упражнения

1. Доказать, что модуль сильной выпуклости функции R из примера 2 относительно 1-нормы равен 1 (см. [6]). (*Указание.* Предварительно доказать неравенство Пинскера: $\sum\limits_{i=1}^{n} x_i \ln \frac{x_i}{y_i} \geqslant \frac{1}{2}\|x - y\|_1^2$, где $x = (x_1, \ldots, x_n)$ и $y = (y_1, \ldots, y_n)$.)

2. Доказать, что $R^{**} = R$.

3. Доказать равенство $D_R(x, y) = D_{R^*}(\nabla R(x), \nabla R(y))$ (Следствие 6.2).

4. Доказать, что проекция Брэгмана точки $y \in \Omega$ на симплекс S в примере 2а равна нормализации y в 1-норме $\hat{y} = \frac{y}{\|y\|_1} = \frac{y}{\sum_{i=1}^{n} |y_i|}$, в частности, $\hat{y}_{i,t+1} = \frac{\tilde{y}_{i,t+1} e^{-\eta x_{i,t}}}{\sum\limits_{j=1}^{n} \tilde{y}_{j,t+1} e^{-\eta x_{j,t}}}$.

5. Алгоритм следования за регуляризованным лидером получается из алгоритма, представленного на рис. 6.3, если прогноз для следующего шага определить

$$\tilde{y}_{t+1} = \operatorname{argmin}_{y \in \Omega} \left((y \cdot x_t) + \eta^{-1} D_R(y, \hat{y}_t) \right)$$
$$\hat{y}_{t+1} = \operatorname{argmin}_{y \in S} D_R(y, \tilde{y}_{t+1})$$

Доказать эквивалентность алгоритма следования за регуляризованным лидером и алгоритма онлайн зеркального спуска в том смысле, что они вычисляют одинаковые прогнозы $\hat{y}_t$.

6. Проверить, что функция $R(x) = -2\sum\limits_{i=1}^{n}\sqrt{x_i}$, определенная на $\mathcal{R}^n_+$, есть функция Лежандра. Вычислить для нее $R^*(p)$, $\nabla R(x)$, $D_R(x,y)$, $D_{R^*}(x,y)$.

Часть III

Игры и предсказания

Глава 7

Элементы теории игр

В данной главе мы рассмотрим классические вопросы теории игр – игры двух лиц с нулевой суммой. Мы докажем минимаксную теорему Дж. фон Неймана, а также рассмотрим методы решения таких игр. Будут также рассмотрены произвольные игры между конечным числом игроков. Для таких игр будут определены понятия равновесия Нэша и коррелированного равновесия Аумана.

Понятия и результаты этой главы будут использоваться в далее в главах 8 и 9.

7.1. Антагонистические игры двух игроков

Пусть X и Y – множества произвольной природы. Рассмотрим антагонистическую игру двух лиц. Первый игрок выбирает *стратегию* (ход) $x \in X$; одновременно с ним второй игрок выбирает стратегию $y \in Y$. В *нормальной форме* игры каждый игрок выбирает стратегию независимо от выбора другого игрока. Задана функция $f(x, y)$ выигрыша первого игрока, которая одновременно является функцией проигрыша второго игрока. Функция $f(x, y)$ определена на декартовом произведении $X \times Y$.

В случае $f(x, y) < 0$ выигрыш первого игрока отрицательный, т.е. является его проигрышем.

Цель первого игрока – максимизация своего выигрыша, цель второго игрока заключается в минимизации своего проигрыша.

Если первый игрок выбрал стратегию x, то его выигрыш будет не меньше чем $\inf_{y\in Y} f(x,y)$ независимо от выбора второго игрока. Эта величина называется гарантированным результатом для первого игрока. Наилучший гарантированный результат для первого игрока:

$$\underline{v} = \sup_{x\in X} \inf_{y\in Y} f(x,y)$$

называется *нижним значением игры.*

Стратегия x^0 первого игрока называется *максиминной*, если

$$\inf_{y\in Y} f(x^0,y) = \underline{v}.$$

С точки зрения второго игрока, выбор стратегии y гарантирует ему максимальный проигрыш: $\sup_{x\in X} f(x,y)$ – его *гарантированный результат.* Наилучший гарантированный результат второго игрока – величина

$$\overline{v} = \inf_{y\in Y} \sup_{x\in X} f(x,y)$$

называется верхним значением игры.

Стратегия y^0 второго игрока называется минимаксной, если

$$\sup_{x\in X} f(x,y^0) = \overline{v}.$$

Лемма 7.1. *В любой антагонистической игре $\underline{v} \leqslant \overline{v}$, т.е.*

$$\sup_{x\in X} \inf_{y\in Y} f(x,y) \leqslant \inf_{y\in Y} \sup_{x\in X} f(x,y).$$

Доказательство. Имеем для любых $x \in X$ и $y \in Y$

$$\inf_{y\in Y} f(x,y) \leqslant f(x,y) \leqslant \sup_{x\in X} f(x,y).$$

Отсюда

$$\inf_{y\in Y} f(x,y) \leqslant \sup_{x\in X} f(x,y).$$

Левая часть последнего неравенства зависит от x, а правая – нет. Поэтому

$$\sup_{x\in X} \inf_{y\in Y} f(x,y) \leqslant \sup_{x\in X} f(x,y)$$

для всех y, следовательно,

$$\underline{v} = \sup_{x\in X} \inf_{y\in Y} f(x,y) \leqslant \inf_{y\in Y} \sup_{x\in X} f(x,y) = \overline{v}.$$

Лемма доказана. $\triangle$

Точка $(x^0, y^0) \in X \times Y$ называется *седловой точкой* функции f, если

$$f(x, y^0) \leqslant f(x^0, y^0) \leqslant f(x^0, y) \tag{7.1}$$

для всех $x \in X$ и $y \in Y$.

Условие (7.1) эквивалентно условию

$$\max_{x\in X} f(x, y^0) = f(x^0, y^0) = \min_{y\in Y} f(x^0, y). \tag{7.2}$$

Заметим, что когда мы пишем min вместо inf или max вместо sup, то имеем ввиду, что эти экстремальные значения достигаются в некоторой точке.

Говорят, что антагонистическая игра *имеет решение*, если функция $f(x,y)$ имеет седловую точку (x^0, y^0). Число $v = f(x^0, y^0)$ называется *значением*, или *ценой* игры, x^0, y^0 – *оптимальные стратегии* игроков, (x^0, y^0, v) – *решение игры.* Эти названия оправдываются следующей теоремой.

Теорема 7.1. 1) *Для того чтобы функция $f(x,y)$ имела седловую точку, необходимо и достаточно, чтобы было выполнено условие*

$$\max_{x\in X} \inf_{y\in Y} f(x,y) = \min_{y\in Y} \sup_{x\in X} f(x,y). \tag{7.3}$$

2) *Пусть выполнено* (7.3). *Тогда пара (x^0, y^0) тогда и только тогда является седловой точкой, когда x^0 – максиминная, а y^0 – минимаксная стратегии игроков.*

Доказательство. Доказательство необходимости 1) и 2). Пусть (x^0, y^0) – седловая точка функции $f(x,y)$. Тогда

$$\overline{v} \leqslant \sup_{x\in X} f(x, y^0) = f(x^0, y^0) = v = \inf_{y\in Y} f(x^0, y) \leqslant \underline{v}. \tag{7.4}$$

Отсюда $\overline{v} \leqslant \underline{v}$. По лемме 7.1 имеем равенство $\overline{v} = \underline{v}$. Тогда в (7.4) имеют место равенства, и поэтому x^0 – максиминная, а y^0 – минимаксная стратегии.

Доказательство достаточности. Допустим, что (7.3) выполнено. Возьмем x^0 – максиминную, y^0 – минимаксную стратегии. Покажем, что пара (x_0, y_0) является седловой точкой. Действительно,

$$f(x^0, y^0) \geqslant \inf_{y \in Y} f(x^0, y) = \underline{v} = \overline{v} = \sup_{x \in X} f(x, y^0) \geqslant f(x^0, y^0).$$

Отсюда следует, что во всех этих неравенствах можно поставить знаки равенства. Таким образом, (x^0, y^0) – седловая точка. $\triangle$

Игра в орлянку, при которой первый игрок загадывает число 0 или 1, а второй отгадывает, с матрицей выплат

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

не имеет седловой точки. Для нее наилучший гарантированный результат для первого игрока равен: $\underline{v} = \max_i \min_j a_{i,j} = -1$, а наилучший гарантированный результат для второго игрока (т.е. его проигрыш) равен: $\overline{v} = \min_j \max_i a_{i,j} = 1$. Эта игра не имеет решения.

7.2. Достаточное условие существования седловой точки

Докажем достаточное условие существования седловой точки, следствием к которому является минимаксная теорема.

Предварительно напомним, что подмножество $Z \subseteq \mathcal{R}^n$ евклидового пространства $\mathcal{R}^n$ называется *выпуклым*, если для любых точек $z, z' \in Z$ и любого числа $0 \leqslant p \leqslant 1$ точка $pz + (1-p)z' \in Z$.

Функция $h(z)$, определенная на выпуклом множестве Z, называется выпуклой, если для любых $z, z' \in Z$ и любого числа $0 \leqslant p \leqslant 1$ выполнено неравенство

$$h(pz + (1-p)z') \leqslant ph(z) + (1-p)h(z'). \tag{7.5}$$

Функция $h(z)$ называется вогнутой, если выполнено неравенство (7.5), где знак $\leqslant$ заменен на $\geqslant$.

Теорема 7.2. *Пусть X, Y – выпуклые подмножества $\mathcal{R}^n$ и $\mathcal{R}^m$ соответственно (где n и m – произвольные натуральные числа), Y – компакт, функция $f(x,y)$ определена на $X \times Y$, принимает вещественные значения и ограничена по абсолютной величине, функция $f(x,\cdot)$ – выпуклая и непрерывная (по y) для каждого значения $x \in X$, $f(\cdot,y)$ – вогнутая для каждого значения $y \in Y$. Тогда*

$$\sup_{x\in X}\inf_{y\in Y} f(x,y) = \inf_{y\in Y}\sup_{x\in X} f(x,y).$$

Доказательство. По лемме 7.1 надо доказать, что

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X}\inf_{y\in Y} f(x,y).$$

Допустим для простоты, что $f(x,y) \in [0,1]$.

Фиксируем достаточно малое $\epsilon > 0$ и достаточное большое натуральное число n. Из компактности Y следует, что существует ϵ-сеть в Y, т.е. конечное множество точек $\{y^1, \ldots, y^N\}$ такое, что каждая точка $y \in Y$ находится в ϵ-окрестности одной из точек y^i.

Определим последовательность точек $y_1, y_2, \ldots, y_n \in Y$ и последовательность точек $x_1, x_2, \ldots, x_n \in X$ рекурсивно. Пусть x_0 – любая точка X. Определим при $t = 1, \ldots, n$:

$$y_t = \frac{\sum\limits_{i=1}^{N} y^i e^{-\eta\sum_{s=0}^{t-1} f(x_s,y^i)}}{\sum\limits_{j=1}^{N} e^{-\eta\sum_{s=0}^{t-1} f(x_s,y^j)}}, \tag{7.6}$$

где $\eta = \sqrt{(8\ln N)/n}$ и x_t выбирается так, чтобы было

$$f(x_t,y_t) \geqslant \sup_{x\in X} f(x,y_t) - \frac{1}{n}.$$

Так как функция f является выпуклой по второму аргументу, мы можем применить теорему 4.9 с функцией потерь $\lambda(x,y) = f(x,y)$.

Применим теорему 4.9. Для этого нормируем функцию $f(x,y)$ так, чтобы ее значения лежали в интервале $[0,1]$ (прибавим и умножим на подходящие константы).

В алгоритме экспоненциального взвешивания (7.6) величины y^i – прогнозы экспертов, $i=1,\dots,N$, x_t – исходы, $t=1,\dots,n$, y_t – прогноз *Предсказателя*. По (4.54) получим

$$\sum_{t=1}^{n} f(x_t,y_t) \leqslant \min_{i=1,\dots,N} \sum_{t=1}^{n} f(x_t,y^i) + \sqrt{\frac{1}{2} n \ln N}.$$

Делим это неравенство на n :

$$\frac{1}{n}\sum_{t=1}^{n} f(x_t,y_t) \leqslant \min_{i=1,\dots,N} \frac{1}{n}\sum_{t=1}^{n} f(x_t,y^i) + \sqrt{\frac{\ln N}{2n}}. \tag{7.7}$$

Пользуемся выпуклостью функции f по второму аргументу и вогнутостью по первому, а также используя (7.7), получаем

$$\begin{aligned}
&\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \\
&\leqslant \sup_{x\in X} f\left(x, \frac{1}{n}\sum_{t=1}^{n} y_t\right) \leqslant \\
&\leqslant \sup_{x\in X} \frac{1}{n}\sum_{t=1}^{n} f(x,y_t) \leqslant \\
&\leqslant \frac{1}{n}\sum_{t=1}^{n} \sup_{x\in X} f(x,y_t) \leqslant \\
&\leqslant \frac{1}{n}\sum_{t=1}^{n} f(x_t,y_t) + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ \dots,\ N} \frac{1}{n}\sum_{t=1}^{n} f(x_t,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ \dots,\ N} f\left(\frac{1}{n}\sum_{t=1}^{n} x_t, y^i\right) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \sup_{x\in X} \min_{i=1,\ \dots,\ N} f(x,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}.
\end{aligned} \tag{7.8}$$

Переход от 1-й строки ко 2-й происходит по определению; переход от 2-й к 3-й – по выпуклости $f(x,\cdot)$; переход от 3-й к 4-й происходит, так как супремум суммы не превосходит суммы супремумов; переход от 4-й к 5-й происходит по определению x_t; переход от 5-й к 6-й происходит по (7.7); переход от 6-й к 7-й происходит по вогнутости функции $f(\cdot,y)$; переход от 7-й к 8-й происходит по определению супремума.

Таким образом, мы доказали, что для всех n

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X}\min_{i=1,\ \ldots,\ N} f(x,y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}.$$

Устремляем n к бесконечности и получаем

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X}\min_{i=1,\ldots,N} f(x,y^i).$$

Устремляем $\epsilon \to 0$ и получаем

$$\inf_{y\in Y}\sup_{x\in X} f(x,y) \leqslant \sup_{x\in X}\inf_{y\in Y} f(x,y).$$

Теорема доказана. $\triangle$

Доказательство теоремы 7.2 содержит метод вычисления цены игры, так как из 1-й, 5-й и 8-й строк неравенства (7.8) следует, что величина $\frac{1}{n}\sum\limits_{t=1}^{n} f(x_t,y_t)$ является как угодно близким приближением к цене игры при достаточно малом ϵ и достаточно большом n.

7.3. Смешанные расширения матричных игр

7.3.1. Минимаксная теорема

Пусть теперь $X = \{1,\ \ldots,\ N\}$ и $Y = \{1,\ \ldots,\ M\}$. Соответствующая игра называется матричной. Функция выигрыша $f(i,j) = a_{i,j}$ может быть представлена в виде матрицы. Первый игрок выбирает номер строки, второй игрок – номер столбца, элемент $a_{i,j}$, находящийся на их пересечении, определяет выигрыш первого игрока и проигрыш второго.

Смешанной стратегией игрока называется распределение вероятностей на множестве его ходов. Смешанное расширение матричной игры $(X, Y, f(x, y))$ определяется как игра $(\mathcal{X}, \mathcal{Y}, f(\bar{p}, \bar{q}))$, где $\mathcal{X}$ – множество смешанных стратегий первого игрока, $\mathcal{Y}$ – множество смешанных стратегий второго игрока, $f(\bar{p}, \bar{q})$ – среднее значение выигрыша относительно меры $\bar{p} \times \bar{q}$:

$$\mathcal{X} = \{\bar{p} = (p_1, \ \ldots, \ p_N) : \sum_{i=1}^{N} p_i = 1, \ p_i \geqslant 0\};$$

$$\mathcal{Y} = \{\bar{q} = (q_1, \ \ldots, \ q_M) : \sum_{i=1}^{M} q_i = 1, \ q_i \geqslant 0\};$$

$$f(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} f(i, j) p_i q_j.$$

Имеет место минимаксная теорема Дж. фон Неймана.

Теорема 7.3. *Всякая матричная игра имеет решение в смешанных стратегиях:*

$$\max_{\bar{p} \in \mathcal{X}} \min_{\bar{q} \in \mathcal{Y}} f(\bar{p}, \bar{q}) = \min_{\bar{q} \in \mathcal{Y}} \max_{\bar{p} \in \mathcal{X}} f(\bar{p}, \bar{q}).$$

Доказательство. Достаточно доказать, что функция $f(\bar{p}, \bar{q})$ имеет седловую точку. Применим теорему 7.2. Множества $\mathcal{X}$ и $\mathcal{Y}$ – симплексы в евклидовых пространствах, поэтому являются выпуклыми. Функция $f(\bar{p}, \bar{q})$ – билинейная и поэтому непрерывна по обоим аргументам, вогнута и выпукла по ним. $\triangle$

Замечание. Можно также рассмотреть последовательный вариант игры двух игроков: сначала первый игрок выбирает элемент $\bar{p} \in \mathcal{X}$, потом второй игрок выбирает $\bar{q} \in \mathcal{Y}$; при этом второй игрок знает выбор первого игрока. В этом случае первый игрок по-прежнему предполагает, что второй игрок своим выбором будет пытаться минимизировать его выигрыш. Поэтому его оптимальная стратегия состоит в том, чтобы добиться того, чтобы достигался $\max_{\bar{p} \in \mathcal{X}} \min_{\bar{q} \in \mathcal{Y}} f(\bar{p}, \bar{q})$.

При другой последовательности действий сначала второй игрок выбирает $\bar{q} \in \mathcal{Y}$, а затем первый игрок, зная его выбор, выбирает $\bar{p} \in \mathcal{X}$. Здесь второй игрок зная, что первый игрок в ответ на его ход будет максимизировать его проигрыш, выберет свой ход $\bar{q} \in \mathcal{Y}$ так, чтобы достигался $\min_{\bar{q}\in\mathcal{Y}} \max_{\bar{p}\in\mathcal{X}} f(\bar{p}, \bar{q})$.

D этом случае по-прежнему действует минимаксная теорема 7.3:

$$\max_{\bar{p}\in\mathcal{X}} \min_{\bar{q}\in\mathcal{Y}} f(\bar{p}, \bar{q}) = \min_{\bar{q}\in\mathcal{Y}} \max_{\bar{p}\in\mathcal{X}} f(\bar{p}, \bar{q}).$$

7.3.2. Чистые стратегии

Рассмотрим матричную игру на $X = \{1, \ldots, N\}, Y = \{1, \ldots, M\}$ с функцией выигрыша $f(i, j) = a_{i,j}$. Приведем три простых утверждения, которые более детально описывают структуру оптимального решения в терминах чистых стратегий.

Обозначим $1_i = (0, \ldots, 1, \ldots, 0)$ – *чистую стратегию*, которая представляет собой распределение вероятностей на множестве X, сосредоточенное на $i \in X$ (вектор длины N, у которого i-я координата равна 1, остальные координаты равны 0). Аналогичным образом рассматриваются чистые стратегии на Y. Заметим, что $f(1_i, 1_j) = f(i, j) = a_{i,j}$.

Теорема 7.4. *Для того чтобы пара смешанных стратегий $(\bar{p}^*, \bar{q}^*)$ была решением (седловой точкой) смешанного расширения матричной игры $(\mathcal{X}, \mathcal{Y}, f)$, необходимо и достаточно, чтобы выполнялось неравенство*

$$f(1_i, \bar{q}^*) \leqslant f(\bar{p}^*, \bar{q}^*) \leqslant f(\bar{p}^*, 1_j) \tag{7.9}$$

для всех $i \in X$ и $j \in Y$.

Доказательство. Необходимость следует из теоремы 7.1. Для доказательства достаточности заметим, что каждая смешанная стратегия $\bar{p} = (p_1, \ldots, p_N)$ матричной игры является линейной комбинацией чистых стратегий $\bar{p} = \sum\limits_{i=1}^{N} p_i 1_i$, аналогичным образом

представляется смешанная стратегия $\bar{q} = \sum\limits_{j=1}^{M} q_j 1_j$. Поэтому можно рассмотреть дважды линейную комбинацию неравенства (7.9). Получим

$$f(\bar{p}, \bar{q}^*) = \sum_{i=1}^{N} p_i f(1_i, \bar{q}^*) \leqslant \sum_{i=1}^{N} p_i f(\bar{p}^*, \bar{q}^*) = f(\bar{p}^*, \bar{q}^*),$$
$$f(\bar{p}^*, \bar{q}^*) = \sum_{j=1}^{M} q_j f(\bar{p}^*, \bar{q}^*) \leqslant \sum_{j=1}^{M} q_j f(\bar{p}^*, 1_j) = f(\bar{p}^*, \bar{q})$$

для всех $\bar{p}$ и $\bar{q}$. Отсюда получаем условие седловой точки:

$$f(\bar{p}, \bar{q}^*) \leqslant f(\bar{p}^*, \bar{q}^*) \leqslant f(\bar{p}^*, \bar{q})$$

для любых $\bar{p}$ и $\bar{q}$. $\triangle$

Теорема 7.5. *Для смешанного расширения произвольной матричной игры справедливы соотношения*

$$\min_{\bar{q}} f(\bar{p}, \bar{q}) = \min_{j} f(\bar{p}, 1_j),$$
$$\max_{\bar{p}} f(\bar{p}, \bar{q}) = \max_{i} f(1_i, \bar{q}).$$

Доказательство. Очевидно, что

$$\min_{\bar{q}} f(\bar{p}, \bar{q}) \leqslant \min_{j} f(\bar{p}, 1_j),$$
$$\max_{\bar{p}} f(\bar{p}, \bar{q}) \geqslant \max_{i} f(1_i, \bar{q}).$$

Противоположное неравенство следует из неравенства

$$f(\bar{p},\bar{q}) = \sum_{i=1}^{N}\sum_{j=1}^{M} a_{i,j}p_iq_j =$$
$$= \sum_{j=1}^{M}\left(\sum_{i=1}^{N} a_{i,j}p_i\right)q_j \geqslant$$
$$\geqslant \left(\min_j \sum_{i=1}^{N} p_i a_{i,j}\right)\left(\sum_{j=1}^{M} q_j\right) =$$
$$= \min_j \sum_{i=1}^{N} p_i a_{i,j} = \min_j f(\bar{p}, 1_j),$$

которое имеет место для любого $\bar{q}$. Это неравенство означает, что минимум взвешенной линейной комбинации достигается, когда весь вес сосредоточен на наименьшем элементе. Следовательно,

$$\min_{\bar{q}} f(\bar{p},\bar{q}) \geqslant \min_j f(\bar{p}, 1_j).$$

Второе неравенство доказывается аналогичным образом. $\triangle$

Из этой теоремы получаем

Следствие 7.1. *В смешанном расширении произвольной матричной игры выполнено равенство*

$$v = \max_{\bar{p}} \min_j f(\bar{p}, 1_j) = \min_{\bar{q}} \max_i f(1_i, \bar{q}),$$

где v – значение игры.

Найдем решение игры в орлянку в смешанных стратегиях. Матрица выплат этой игры типа (2×2) имеет вид

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Смешанные стратегии этой игры: $\bar{p} = (p, 1-p)$ и $\bar{q} = (q, 1-q)$, а среднее значение выигрыша имеет вид

$$f(\bar{p}, \bar{q}) = \sum_{i,j=1}^{2} a_{i,j} p_i q_j =$$
$$= q(-p + 1 - p) + (1-q)(p - (1-p)) =$$
$$= -4\left(p - \frac{1}{2}\right)\left(q - \frac{1}{2}\right).$$

Среднее значение выигрыша:

$$f(\bar{p}, \bar{q}) = -4\left(p - \frac{1}{2}\right)\left(q - \frac{1}{2}\right) \tag{7.10}$$

представляет собой уравнение однополостного гиперболоида.

Пусть

$$v(p) = \min_j f(\bar{p}, 1_j) = \min\{1 - 2p, 2p - 1\}.$$

По следствию 7.1 решение игры достигается в точке p^*, на которой $v(p)$ достигает своего максимума – это $p^* = \frac{1}{2}$.

Аналогичные рассуждения показывают, что $q^* = \frac{1}{2}$. Значение игры: $v^* = f(\bar{p}^*, \bar{q}^*) = 0$. Точка (p^*, q^*) является седловой точкой однополостного гиперболоида (7.10).

7.3.3. Решение матричной игры типа $(2 \times M)$

Для нахождения решений в смешанных расширениях матричных игр $(2 \times M)$ (или $(N \times 2)$) можно использовать геометрическое представление стратегий. Согласно следствию 7.1 значение такой игры равно

$$v = \max_p \min_{1 \leqslant j \leqslant M} (a_{1,j} p + a_{2,j}(1-p)).$$

Здесь первый игрок выбирает смешанную стратегию – распределение вероятностей $\bar{p} = (p, 1-p)$ на строках матрицы, а второй игрок выбирает чистую стратегию – столбец j матрицы.

Значит, для нахождения значения игры и ее решения для первого игрока надо просто найти значение $p = p^*$, при котором функция

$$v(p) = \min_{1 \leqslant j \leqslant M} (a_{1,j}p + a_{2,j}(1-p))$$

достигает своего максимального значения на отрезке $[0, 1]$. Это значение $v(p^*)$ и будет значением игры.

Для нахождения решения строим все M прямых вида

$$L_j(p) = a_{1,j}p + a_{2,j}(1-p),$$

где $j = 1, \ \ldots, \ M$.

Для каждого $p \in [0, 1]$ проводим вертикальную прямую до пересечения с прямой с наименьшим значением ординаты. Точки пересечения образуют ломаную линию $y = v(p)$ – *нижнюю огибающую* для всех этих прямых. Верхняя точка нижней огибающей определяет оптимальную стратегию первого игрока (ее абсцисса p^*) и значение игры (ордината точки $v(p^*)$).

Задача. Найти решение смешанного расширения матричной игры:

$$\begin{pmatrix} 7 & 3 & 3 & 1 & -1 & 0 \\ -1 & -1 & 1 & 0 & 5 & 3 \end{pmatrix}.$$

Строим все прямые вида $L_j(p) = a_{1,j}p + a_{2,j}(1-p)$ при $j = 1, \ \ldots, \ 6$:

$$\begin{aligned} L_1(p) &= 7p - (1-p), \\ L_2(p) &= 3p - (1-p), \\ L_3(p) &= 3p + (1-p), \\ L_4(p) &= p, \\ L_5(p) &= -p + 5(1-p), \\ L_6(p) &= 3(1-p). \end{aligned}$$

Строим нижнюю огибающую. Точка p^* является точкой пересечения прямой 4 и прямой 5, т.е. решаем уравнение $p = -p + 5(1-p)$. Получаем: $p^* = 5/7$, $v(p^*) = 5/7$.

Для нахождения оптимальной стратегии второго игрока используем следующую теорему.

Теорема 7.6. *Пусть* $(\bar{p}^*, \bar{q}^*)$ *– решение матричной игры в смешанных стратегиях,* v^* *– значение игры. Тогда*

- *из* $p_i^* > 0$ *следует* $f(1_i, \bar{q}^*) = v^*$,
- *из* $q_j^* > 0$ *следует* $f(\bar{p}^*, 1_j) = v^*$.

Доказательство. Докажем первое утверждение. По определению $f(1_i, \bar{q}^*) \leqslant v^*$, $i = 1, \ \ldots, \ N$.

Допустим, что существует i_0 такое, что $p_{i_0}^* > 0$ и одновременно $f(1_{i_0}, \bar{q}^*) < v^*$. Рассмотрим линейную комбинацию неравенств $f(1_i, \bar{q}^*) \leqslant v^*$ с коэффициентами p_i^*, $i = 1, \ \ldots, \ N$, и, так как одно из складываемых неравенств является строгим, получим

$$v^* = f(\bar{p}^*, \bar{q}^*) = \sum_{i=1}^{N} f(1_i, \bar{q}^*) p_i^* < v^* = f(\bar{p}^*, \bar{q}^*).$$

Это противоречие доказывает первое утверждение.

Второе утверждение доказываем аналогично. $\triangle$

Следствие 7.2. *Пусть* $(\bar{p}^*, \bar{q}^*)$ *– решение матричной игры в смешанных стратегиях,* v^* *– значение игры. Тогда*

- *из* $f(1_i, \bar{q}^*) < v^*$ *следует* $p_i^* = 0$,
- *из* $f(\bar{p}^*, 1_j) > v^*$ *следует* $q_j^* = 0$.

Условие $f(\bar{p}^*, 1_j) = pa_{1,j} + (1-p)a_{2,j} > v^*$ означает, что соответствующая прямая в точке p^* проходит выше точки пересечения (двух) прямых, на которых достигается значение игры.

Завершим решение задачи – найдем оптимальную стратегию второго игрока.

Для 1-й, 2-й, 3-й, 6-й чистых стратегий второго игрока (т.е. соответствующих прямых) выполняется неравенство

$$f(\bar{p}^*, 1_j) = L_j(p^*) > v^*$$

при $j = 1, 2, 3, 6$.

По следствию 7.2 для оптимальной стратегии

$$\bar{q}^* = (q_1^*, q_2^*, q_3^*, q_4^*, q_5^*, q_6^*)$$

будет $q_1^* = 0$, $q_2^* = 0$, $q_3^* = 0$, $q_6^* = 0$, $q_4^* = q$, $q_5^* = 1 - q$.

Пусть теперь первый игрок выбирает чистую стратегию на строках – одну из строк $i = 1, 2$. Второй игрок выбирает смешанную стратегию $\bar{q}^* = (0, 0, 0, q_4^*, 1 - q_4^*, 0)$ на столбцах. Тогда

$$v^* = \min_q \max_{1 \leqslant i \leqslant 2} (a_{i,4} q + a_{i,5}(1 - q)) = \max_{1 \leqslant i \leqslant 2} (a_{i,4} q_4^* + a_{i,5}(1 - q_4^*)).$$

Для $j = 4, 5$ получаем: $q_4^* - (1 - q_4^*) = 5/7$ и $5(1 - q_4^*) = 5/7$. Отсюда $q_4^* = 6/7$, $q_5^* = 1/7$.

Полное решение игры имеет вид

$$\bar{p}^* = \left(\frac{5}{7}, \frac{2}{7}\right),$$
$$\bar{q}^* = (0, 0, 0, \frac{6}{7}, \frac{1}{7}, 0),$$
$$v^* = \frac{5}{7}.$$

7.3.4. Решение игры типа $(N \times M)$

В общем случае рассматривается матричная игра с матрицей $A = (a_{i,j})$, где $i = 1, \ldots, N$, $j = 1, \ldots, M$. Без ограничения общности можно считать, что все элементы матрицы A строго положительны; поэтому значение v игры в смешанных стратегиях также строго положительно. [1]

По следствию 7.1 в смешанном расширении произвольной матричной игры выполнено равенство

$$v = \max_p \min_j f(\bar{p}, 1_j) = \min_{\bar{q}} \max_i f(1_i, \bar{q}), \tag{7.11}$$

где v – значение игры. Поэтому существует смешанная стратегия $\bar{p} = (p_1, \ldots, p_N)$ первого игрока, такая, что $f(\bar{p}, 1_j) \geqslant v$ для

[1] Для того чтобы этого добиться, достаточно прибавить некоторую достаточно большую положительную константу к каждому элементу платежной матрицы игры.

любой чистой стратегии 1_j второго игрока. Иными словами, выполняются условия

$$\sum_{i=1}^{N} a_{i,j} p_i \geqslant v \text{ при } j = 1, \ \ldots, \ M,$$

$$\sum_{i=1}^{N} p_i = 1,$$

$$p_i \geqslant 0 \text{ при } i = 1, \ \ldots, \ N.$$

Введем обозначения $x_i = p_i/v$, $i = 1, \ \ldots, \ M$. Тогда эти условия превращаются в соотношения

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \text{ при } j = 1, \ \ldots, \ M,$$

$$\sum_{i=1}^{N} x_i = 1/v,$$

$$x_i \geqslant 0 \text{ при } i = 1, \ \ldots, \ N.$$

Задача поиска решения в матричной игре сводится к задаче линейного программирования: найти $x_1, \ \ldots, \ x_N$ такие, что

$$\sum_{i=1}^{N} x_i \to \min$$

при условиях

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \text{ при } j = 1, \ \ldots, \ M,$$

$$x_i \geqslant 0 \text{ при } i = 1, \ \ldots, \ N.$$

По (7.11) существует смешанная стратегия $\bar{q} = (q_1, \ \ldots, \ q_N)$ первого игрока, такая, что $f(1_i, \bar{q}) \leqslant v$ для любой чистой страте-

гии 1_i первого игрока. Иными словами, выполняются условия

$$\sum_{j=1}^{M} a_{i,j} q_j \leqslant v \text{ при } i = 1, \ \ldots, \ N,$$

$$\sum_{j=1}^{M} q_j = 1,$$

$$q_j \geqslant 0 \text{ при } j = 1, \ \ldots, \ M.$$

Введем обозначения: $x'_j = q_j/v$, $j = 1, \ \ldots, \ M$. Тогда эти условия превращаются в соотношения

$$\sum_{j=1}^{M} a_{i,j} x'_j \leqslant 1 \text{ при } i = 1, \ \ldots, \ N,$$

$$\sum_{j=1}^{M} x'_j = 1/v,$$

$$x'_j \geqslant 0 \text{ при } j = 1, \ \ldots, \ M.$$

Задача поиска решения в матричной игре сводится к задаче линейного программирования: найти $x'_1, \ \ldots, \ x'_M$ такие, что

$$\sum_{j=1}^{M} x'_j \to \max$$

при условиях

$$\sum_{j=1}^{M} a_{i,j} x'_j \leqslant 1 \text{ при } i = 1, \ \ldots, \ N,$$

$$x'_j \geqslant 0 \text{ при } j = 1, \ \ldots, \ M.$$

Это – задача линейного программирования, двойственная к прямой задаче для переменных x_i, $i = 1, \ \ldots, \ N$.

7.3.5. Конечная игра между K игроками

В общем случае конечная игра между K игроками в *нормальной форме* определяется следующим образом. Игрок $k \in \{1, \ldots, K\}$ имеет N_k возможных стратегий (ходов или чистых стратегий). Пусть $\bar{i} = (i_1, \ldots, i_K)$ – некоторый набор стратегий всех K игроков, где $i_j \in \{1, \ldots, N_j\}$, $j = 1, \ldots, K$.

Тогда выигрыш k-го игрока обозначается $f^k(\bar{i}) = f^k(i_1, \ldots, i_K)$ (в другой постановке его потери равны – $f^k(\bar{i})$).

Смешанная стратегия k-го игрока – это распределение вероятностей $\bar{p}^k = (p_1^k, \ldots, p_{N_k}^k)$ на множестве всех его стратегий $\{1, \ldots, N_k\}$. Здесь p_j^k – вероятность выбора игроком стратегии $j \in \{1, \ldots, N_k\}$.

Пусть I^k – случайная величина, принимающая значение $i \in \{1, \ldots, N_k\}$ с вероятностью p_i^k.

Пусть $\bar{I} = (I^1, \ldots, I^K)$ – векторная случайная величина, представляющая набор стратегий всех игроков. Значениями такой случайной величины являются векторы $\bar{i} = (i_1, \ldots, i_K)$, где $i_j \in \{1, \ldots, N_j\}$, $j = 1, \ldots, K$.

Обычно предполагается, что случайные величины $I^1, \ldots, I^K$ независимы. На множестве векторов $\bar{I}$ рассматривается вероятностная мера $\pi = \bar{p}^1 \times \cdots \times \bar{p}^K$, которая определяет вероятность элементарного исхода $\bar{i} = (i_1, \ldots, i_K)$, равную произведению вероятностей исходов:

$$\pi(\bar{i}) = \pi(\bar{I} = \bar{i}) = p_{i_1}^1 \cdot \ \ldots \ \cdot p_{i_K}^K.$$

Математическое ожидание выигрыша k-го игрока равно

$$E_\pi(f^k(\bar{I})) = \sum_{\bar{i}} \pi(\bar{i}) f^k(\bar{i}) =$$

$$= \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} p_{i_1}^1 \cdot \ \ldots \ \cdot p_{i_K}^K f^k(i_1, \ldots, i_K).$$

Равновесие Нэша. Набор смешанных стратегий всех K игроков

$$\pi = (\bar{p}^1, \ldots, \bar{p}^k, \ldots, \bar{p}^K)$$

называется *равновесием Нэша*, если для любого $k = 1, \ldots, K$ и любой смешанной стратегии $\bar{p}'^k$ будет

$$E_\pi(f^k) \geqslant E_{\pi'}(f^k),$$

где стратегия

$$\pi' = (\bar{p}^1, \ \ldots, \ \bar{p}'^k, \ \ldots, \ \bar{p}^K)$$

получена из стратегии π заменой вероятностного распределения k-го игрока $\bar{p}^k$ на другое распределение $\bar{p}'^k$.

Можно сказать, что если π – равновесие Нэша, то никакому игроку не выгодно изменять свою стратегию, если другие игроки не меняют свои стратегии.

Минимаксная теорема Дж. фон Неймана является частным случаем утверждения о существовании равновесия Нэша для случая игры с нулевой суммой для двух игроков. В этом случае функции выигрышей игроков равны $f^1(i,j) = f(i,j)$ и $f^2(i,j) = -f(i,j)$, где $f(i,j)$ – функция выигрыша в игре двух лиц с нулевой суммой.

В частности, седловая точка $(\bar{p}^0, \bar{q}^0)$ в игре двух лиц в смешанных стратегиях с нулевой суммой является равновесием Нэша, так как для любых смешанных стратегий $\bar{p}$ и $\bar{q}$ выполнено

$$f(\bar{p}, \bar{q}^0) \leqslant f(\bar{p}^0, \bar{q}^0) \leqslant f(\bar{p}^0, \bar{q}),$$

где $f(\bar{p}, \bar{q})$ – математическое ожидание выигрыша первого игрока, а $-f(\bar{p}, \bar{q})$ – математическое ожидание выигрыша второго игрока.

В случае игры двух лиц с нулевой суммой множество всех равновесий Нэша описано в следующем утверждении.

Предложение 7.1. *Пара $(\bar{p}^*, \bar{q}^*)$ является точкой равновесия Нэша в игре двух лиц с нулевой суммой тогда и только тогда, когда*

$$\bar{q}^* \in \{\bar{q} : \min_{\bar{p}} f(\bar{p}, \bar{q}) \to \max\},$$
$$\bar{p}^* \in \{\bar{p} : \max_{\bar{q}} f(\bar{p}, \bar{q}) \to \min\}.$$

Для любой такой пары $(\bar{p}^*, \bar{q}^*)$ выполнено $f(\bar{p}^*, \bar{q}^*) = v$, где v – цена игры.

Доказательство предложения 7.1 предоставляется читателю в виде задачи.

В общем случае конечной игры K игроков имеет место следующая основная теорема.

Теорема 7.7. *Каждая конечная игра имеет по крайней мере одно равновесие Нэша.*

Доказательство этой теоремы основано на использовании теоремы Брауэра о неподвижной точке.

Приведем примеры игр и равновесий Нэша. Рассмотрим игры двух игроков, каждый из которых имеет две стратегии.

Пример 1. Первая игра – ранее рассмотренная игра в орлянку, в которой первый игрок загадывает число 0 или 1, а второй отгадывает, с матрицей выплат

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

Эта игра с нулевой суммой имеет решение в смешанных стратегиях: для первого и второго игрока соответствующие смешанные стратегии имеют вид $\vec{p}^* = (\frac{1}{2}, \frac{1}{2})$ и $\vec{q}^* = (\frac{1}{2}, \frac{1}{2})$. Это решение и является единственным равновесием Нэша в этой игре.

Мы перепишем матрицу выплат этой игры в виде таблицы более общего вида:

Ход	0	1
0	(-1,1)	(1,-1)
1	(1,-1)	(-1,1)

Пример 2. Два игрока решают идти на концерт слушать Баха или идти на концерт слушать Пендерецкого. [2] Один предпочитает слушать Баха, а другой Пендерецкого. При этом, оба они предпочитают идти вместе на один концерт, чем каждому на свой концерт. Таблица предпочтений имеет вид:

[2]Музыка композитора Кшиштофа Пендерецкого содержит эксперименты в области атональной музыки и не всегда рассчитана на широкого слушателя.

Ход	Бах	Пендерецкий
Бах	(2,1)	(0,0)
Пендерецкий	(0,0)	(1,2)

Имеется два равновесия Нэша в чистых стратегиях в этой игре (Б,Б) и (П,П).

Пример 3. Два игрока живут в соседних комнатах. Каждый может слушать громкую или тихую музыку. Каждый игрок предпочитает слушать громкую музыку, а также, чтобы его сосед слушал тихую музыку. Таблица предпочтений степени громкости имеет вид:

Ход	Тихо	Громко
Тихо	(3,3)	(1,4)
Громко	(4,1)	(2,2)

В этой игре имеется только одно равновесие Нэша. Это чистая стратегия (Г,Г) (доказательство в виде задачи).[3]

Коррелированное равновесие. Обобщением равновесия Нэша является коррелированное равновесие Аумана. Распределение вероятностей P на множестве

$$\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$$

всех возможных наборов $\bar{i} = (i_1,\ \ldots,\ i_K)$, составленных из всевозможных стратегий всех K игроков, называется *коррелированным равновесием*, если для всех $k = 1,\ \ldots,\ K$ и для любой функции $h : \{1,\ \ldots,\ N_k\} \to \{1,\ \ldots,\ N_k\}$ будет

$$E_P(f^k(\bar{i})) \geqslant E_P(f^k(h(i_k), \bar{i}_{-k})), \tag{7.12}$$

[3]Эта игра также называется “дилемма заключенного”. Двое изолированных друг от друга заключенных получают наименьшие сроки заключения (выигрыш – 3), если оба не признаются в совершенном преступлении – (Т,Т). В то же время, каждый из них будет освобожден (выигрыш – 4), если признается сам и будет свидетельствовать против другого – (Т,Г), (Г,Т). Другой заключенный, если он при этом не признался, получит значительно больший срок (выигрыш – 1). Ни один из заключенных не знает точно, что сделает другой. Для каждого безопаснее всего признаться – (Г,Г) и получить выигрыши по 2 каждый.

где вектор $\bar{i} = (i_1, \ \ldots, \ i_K)$ распределен в соответствии с вероятностным распределением P, а также

$$\bar{i}_{-k} = (i_1, \ \ldots, \ i_{k-1}, i_{k+1}, \ \ldots, \ i_K)$$
$$(h(i_k), \bar{i}_{-k}) = (i_1, \ \ldots, \ i_{k-1}, h(i_k), i_{k+1}, \ \ldots, \ i_K).$$

В отличие от равновесия Нэша величины i_k более не считаются независимыми, а вероятностная мера P не является произведением мер – смешанных стратегий игроков.

Следующая лемма дает эквивалентное описание коррелированного равновесия в геометрических терминах.

Лемма 7.2. *Распределение вероятностей P на множестве*

$$\prod_{k=1}^{K} \{1, \ \ldots, \ N_k\}$$

последовательностей стратегий типа $\bar{i} = (i_1, \ \ldots, \ i_K)$ является коррелированным равновесием тогда и только тогда, когда для каждого игрока $k \in \{1, \ \ldots, \ K\}$ и любых стратегий $j, j' \in \{1, \ \ldots, \ N_k\}$ выполнено

$$\sum_{\bar{i}: i_k = j} P(\bar{i})(f^k(\bar{i}) - f^k(j', \bar{i}_{-k})) \geqslant 0, \tag{7.13}$$

где $(j', \bar{i}_{-k}) = (i_1, \ \ldots, \ i_{k-1}, j', i_k, \ \ldots, \ i_K)$.

Условие (7.13) можно записать также в виде:

$$E(f^k(\bar{i}) | i_k = j) \geqslant E(f^k(j', \bar{i}_{-k}) | i_k = j),$$

где E – условное математическое ожидание относительно распределения P.

Доказательство. Условие (7.12) коррелированного равновесия эквивалентно совокупности условий:

$$\sum_{\bar{i}} P(\bar{i})(f^k(\bar{i}) - f^k(h(i_k), \bar{i}_{-k})) \geqslant 0, \tag{7.14}$$

где $k \in \{1, \ldots, K\}$ и h – произвольная функция типа

$$h : \{1, \ldots, N_k\} \to \{1, \ldots, N_k\}.$$

Для произвольных $j, j' \in \{1, \ldots, N_k\}$ рассмотрим функцию h, такую, что $h(j) = j'$ и $h(i_k) = i_k$ для всех $i_k \neq j$. Тогда в сумме (7.14) останутся только слагаемые, соответствующие наборам $\bar{i}$, в которых $i_k = j$, а в остальных слагаемых соответствующие разности сократятся. Таким образом, сумма (7.14) превратится в сумму (7.13).

В обратную сторону, утверждение тривиально. $\triangle$

Каждое условие типа (7.13) задает замкнутую полуплоскость, поэтому множество всех коррелированных равновесий представляет собой замкнутый выпуклый многогранник в пространстве всех мер на множестве $\prod_{k=1}^{K}\{1, \ldots, N_k\}$.

Пусть P – некоторое распределение вероятностей на множестве $\prod_{k=1}^{K}\{1, \ldots, N_k\}$ и $j \in A_k$ для некоторого $1 \leqslant k \leqslant K$. Обозначим посредством $P_{-i}(\cdot|i_k = j)$ соответствующее условное распределение на множестве $\prod_{s=1, s\neq k}^{K}\{1, \ldots, N_s\}$ произвольных наборов $\bar{i}_{-k}$ из этого множества при известном $i_k = j$. Введем также обозначение

$$f^k(j', \bar{P}_{-k}(\cdot|i_k = j))) = E_{\bar{P}_{-k}(\cdot|i_k=j))}(f^k(j', \bar{i}_{-k})).$$

Будем также писать более компактно:

$$f^k(j', \bar{P}_{-k}) = E_{\bar{P}_{-k}}(f^k(j', \bar{i}_{-k})),$$

имея ввиду, что $\bar{P}_{-k}$ есть распределение на $\bar{i}_{-k}$, порожденное распределением P при условии $i_k = j$.

Теперь можно записать условие (7.13) коррелированного равновесия в эквивалентной форме:

Следствие 7.3. *Распределение вероятностей P на множестве $\prod_{k=1}^{K}\{1, \ldots, N_k\}$ последовательностей стратегий типа $\bar{i} = (i_1, \ldots, i_K)$ является коррелированным равновесием тогда и только тогда, когда для каждого игрока $k \in \{1, \ldots, K\}$ и любых*

стратегий $j, j' \in \{1, \ \ldots, \ N_k\}$ *выполнено* [4]

$$f^k(j, \bar{P}_{-k}(\cdot|i_k = j)) = \max_{j' \in A_i} f^k(j', \bar{P}_{-k}(\cdot|i_k = j)). \tag{7.15}$$

Существование равновесия Нэша в любой конечной игре означает, что коррелированное равновесие существует. Множество всех коррелированных равновесий более обширное и имеет более простое описание, чем множество всех равновесий Нэша.

7.4. Задачи и упражнения

1. Доказать, что в смешанном расширении произвольной матричной игры произвольная максиминная (минимаксная) стратегия одного игрока достигается при чистой стратегии другого игрока:

$$\begin{aligned} \min_{\bar{q}} f(\bar{p}^*, \bar{q}) &= \min_j f(\bar{p}^*, 1_j), \\ \max_{\bar{p}} f(\bar{p}, \bar{q}^*) &= \max_i f(1_i, \bar{q}^*), \end{aligned} \tag{7.16}$$

где $(\bar{p}^*, \bar{q}^*)$ – решение игры (седловая точка).

2. Доказать предложение 7.1.

3. Доказать, что в игре из Примера 2 (раздел (7.3.5)) имеется также равновесие Нэша в смешанных стратегиях: первый игрок выбирает Б с вероятностью $\frac{2}{3}$ и П с вероятностью $\frac{1}{3}$; второй игрок выбирает Б с вероятностью $\frac{1}{3}$ и П – с вероятностью $\frac{2}{3}$.

Имеются ли другие равновесия Нэша в этой игре?

4. Доказать, что в игре из Примера 3 (раздел (7.3.5)) имеется только одно равновесие Нэша. Это чистая стратегия (Г,Г).

5. Показать, что произвольная выпуклая комбинация равновесий Нэша является коррелированным равновесием.

[4] Это условие будет использоваться в дальнейшем в качестве определения коррелированного равновесия.

Глава 8

Теоретико-игровая интерпретация теории вероятностей

В этой главе мы рассмотрим новый теоретико-игровой подход к теории вероятностей, предложенный Вовком и Шейфером [49].

В рамках этого подхода формулируются игры, в которых, при определенных условиях, выполнены различные законы теории вероятностей. Примеры таких законов – закон больших чисел, закон повторного логарифма, центральная предельная теорема и т.д.

Игровая интерпретация теории вероятностей из книги [49] будет продемонстрирована в разделе 8.1 на примере закона больших чисел.

В рамках этого подхода также наиболее естественным образом формулируется задача построения универсальных предсказаний, рассмотренная в главе 3. Игры на универсальные предсказания будут рассмотрены в разделе 8.3.

8.1. Теоретико-игровой закон больших чисел

Игровая интерпретация теории вероятностей основана на идех и понятиях из финансов. В игровой постановке Вовка и Шейфе-

ра [49] для каждого закона теории вероятностей (например, усиленного закона больших чисел или закона повторного логарифма) формулируется некоторая повторяющаяся игра с полной информацией, в которой на каждом раунде (шаге) игры один участник – *Предсказатель*, выдает среднее значение будущего исхода, а после этого, другой участник – *Природа*, выдает новый исход. [1] Третий участник игры – *Скептик*, определяет цель игры. Зная прогноз, *Скептик* делает ставку на его отклонение от будущего исхода и выигрывает или проигрывает некоторую величину. Перед началом игры *Скептик* располагает некоторым начальным капиталом и в течении игры он не может брать в долг – его стратегия должна быть *безопасной*. Игра устроена таким образом, что если закон теории вероятностей нарушается для последовательности прогнозов и исходов, то *Скептик* может наращивать свой выигрыш до бесконечности даже находясь в рамках указанных ограничений. Это эквивалентно тому, что на тех последовательностях прогнозов и исходов, для которых закон теории вероятностей выполнен капитал *Скептика* всегда останется ограниченным, если он использует безопасную стратегию.

Рассмотрим бесконечно повторяющуюся *ограниченную* игру на предсказания между тремя игроками: *Предсказатель*, *Скептик* и *Природа*.

Действия игроков регулируются следующим протоколом – см. рис. 8.1

Данную игру можно рассматривать как финансовую. В ней на каждом шаге n *Скептик* покупает M_n единиц некоторого финансового инструмента по p_n за каждую единицу. В конце шага объявляется новая цена ω_n и *Скептик* продает их по новой цене, а его капитал увеличивается или уменьшается на соответствующую величину. Заметим, что может быть $M_n < 0$. В этом случае *Скептик* продает $|M_n|$ единиц финансового инструмента в начале шага и покупает их в конце шага.

Скептик выигрывает в этой игре, если (1) $\mathcal{K}_n \geqslant 0$ для всех n и (2) либо $\sup_n \mathcal{K}_n = \infty$, либо выполнен закон больших чисел (8.1)

[1] В случае бинарных исходов 0 и 1 среднее значение равно вероятности единицы.

Пусть $\mathcal{K}_0 = 1$.
FOR $n = 1, 2, \ \ldots$
Предсказатель предъявляет прогноз $p_n \in [0, 1]$.
Скептик предъявляет число $M_n \in \mathcal{R}$.
Природа предъявляет исход $\omega_n \in [0, 1]$.
Скептик обновляет свой *выигрыш*: $\mathcal{K}_n = \mathcal{K}_{n-1} + M_n(\omega_n - p_n)$.
ENDFOR

Рис. 8.1: Протокол бесконечно продолжающейся игры с предсказаниями

(см. ниже); в противном случае, выигрывают *Природа* и *Предсказатель*.

В общем случае не предполагается что имеется закон для выбора ходов участников игры. Если такой закон имеется, называем его стратегией. Например, будет определена стратегия *Скептика*: на каждом шаге значение M_n будет определяться с помощью последовательности функций от всевозможных предшествующих ходов участников игры:

$$M_n = M_n(p_1, \omega_1, p_2, \omega_2, \ldots \ , p_{n-1}, \omega_{n-1}, p_n).$$

Теоретико-игровой закон больших чисел формулируется в виде следующей теоремы.

Теорема 8.1. *Существует стратегия Скептика, при которой он выигрывает независимо от того какие бы ходы не выбирали Предсказатель и Природа. Иными словами, существует стратегия Скептика такая, что $\mathcal{K}_n \geqslant 0$ для всех n и выполнено одно из двух условий:*[2]

- $\sup_n \mathcal{K}_n = \infty$
- *или*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0. \tag{8.1}$$

[2]Одна из задач из раздела 8.5 утверждает, что это условие можно заменить на более сильное условие $\lim\limits_{n\to\infty} \mathcal{K}_n = \infty$.

Доказательство. Допустим, что закон больших чисел (8.1) не выполнен. Это означает, что для некоторого $\epsilon > 0$ будет выполнено

$$\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) > 2\epsilon \tag{8.2}$$

для бесконечно многих n или же для некоторого $\epsilon > 0$ будет выполнено

$$\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) < -2\epsilon \tag{8.3}$$

для бесконечно многих n. Мы предполагаем, что $\epsilon < \frac{1}{2}$.

Рассмотрим первый случай. Так как $|\omega_i - p_i| \leqslant 1$, отсюда

$$\epsilon\sum_{i=1}^{n}(\omega_i - p_i) - \epsilon^2\sum_{i=1}^{n}(\omega_i - p_i)^2 > \epsilon^2 n$$

для бесконечно многих n. Используем неравенство $t - t^2 \leqslant \ln(1+t)$, которое выполнено при всех $t \geqslant -\frac{1}{2}$, и получаем

$$\sum_{i=1}^{n}\ln(1 + \epsilon(\omega_i - p_i)) > \epsilon^2 n$$

для бесконечно многих n.

Пусть в игре *Скептик* выбирает на каждом шаге n

$$M_n = \epsilon\mathcal{K}_{n-1},$$

где $\mathcal{K}_{n-1}$ – его текущий выигрыш. Легко видеть, что выигрыш *Скептика* на шаге n равен

$$\mathcal{K}_n = \prod_{i=1}^{n}(1 + \epsilon(\omega_i - p_i)), \tag{8.4}$$

а для его логарифма выполнено

$$\ln\mathcal{K}_n = \sum_{i=1}^{n}\ln(1 + \epsilon(\omega_i - p_i)) > \epsilon^2 n \tag{8.5}$$

для бесконечно многих n. Отсюда $\sup_n \mathcal{K}_n = \infty$.

Заметим также, что из определения (8.4), $\mathcal{K}_n \geqslant 0$ для всех n как бы не выбирались значения ω_i и p_i в процессе игры.

Аналогичным образом, в случае когда выполнено (8.3) для бесконечно многих n, можно построить стратегию *Скептика*

$$M_n = -\epsilon \mathcal{K}_{n-1},$$

где $\mathcal{K}_{n-1}$ – его текущий выигрыш в соответствующей игре.

Недостаточность этого рассуждения заключается в том, что *Скептик* не имеет информации о том, какое из условий (8.2) или (8.3) выполнено для бесконечно многих n, а также для какого $\epsilon > 0$ оно выполнено.

Для того, чтобы обойти эту трудность, усложним стратегию *Скептика* так, чтобы она учитывала оба случая и все возможные значения $\epsilon > 0$. Полагаем $\epsilon_k = 2^{-k}$ при $k = 1, 2, \ldots$. Определим $\mathcal{K}_0^{1,k} = 1$ и $\mathcal{K}_0^{2,k} = 1$ для всех k.

Рассмотрим последовательность стратегий и соответствующих вспомогательных игр

$$\begin{aligned} M_n^{1,k} &= \epsilon_k \mathcal{K}_{n-1}^{1,k}, \\ M_n^{2,k} &= -\epsilon_k \mathcal{K}_{n-1}^{2,k}, \\ M_n^{+} &= \sum_{k=1}^{\infty} 2^{-k} M_n^{1,k}, \\ M_n^{-} &= \sum_{k=1}^{\infty} 2^{-k} M_n^{2,k}, \\ M_n &= \frac{1}{2}(M_n^{+} + M_n^{-}). \end{aligned}$$

Объединим вспомогательные игры и стратегии в одну игру и одну

смешанную стратегию M_n с одним выигрышем $\mathcal{K}_n$:

$$\mathcal{K}_n^+ = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{1,k},$$
$$\mathcal{K}_n^- = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{2,k},$$
$$\mathcal{K}_n = \frac{1}{2}(\mathcal{K}_n^+ + \mathcal{K}_n^-).$$

Все эти ряды сходятся, так как для любого фиксированного n будет $\mathcal{K}_n^{1,k} \leqslant 2^n$ для всех k. Отсюда будет выполнено $|M_n^{2,k}| \leqslant 2^{n-1}$ для всех n.

Заметим, что каждый из выигрышей удовлетворяет условиям $\mathcal{K}_n^{1,k} \geqslant 0$ и $\mathcal{K}_n^{2,k} \geqslant 0$ для всех n и k.

Если закон больших чисел (8.1) не выполнен, то условие (8.2) или условие (8.3) будет выполнено при некотором $\epsilon = \epsilon_k$ для бесконечно многих n. Из условия (8.5), в котором $\mathcal{K}_n = \mathcal{K}_n^{s,k}$ для $s = 0$ или $s = 1$, следует, что $\ln \mathcal{K}_n^{s,k} > \epsilon_k^2 n$ для бесконечно многих n. Отсюда следует, что $\sup_n \mathcal{K}_n = \infty$.

Теорема доказана. $\triangle$

Теоретико-игровую форму закона больших чисел получим с помощью эквивалентной записи утверждения теоремы 8.1.

Следствие 8.1. *Можно построить такую безопасную стратегию Скептика, что для любой реализации ограниченной игры с предсказаниями выполнена следующая импликация:*

$$\sup_n \mathcal{K}_n < \infty \Rightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0,$$

где $\mathcal{K}_n$ – капитал Скептика на шаге n.

Другими словами, закон больших чисел выполнен для тех траекторий игры, на которых нельзя неограниченно увеличивать свой капитал, используя безопасную стратегию.

8.2. Теоретико-игровая вероятность

В теоретико-игровом подходе к теории вероятности Вовка и Шейфера [49] исходным является понятие игры с предсказаниями. Понятие вероятности события является производным и определяется с помощью понятия игры. Правила игры регулируются протоколом.

Предварительно сделаем неформальные пояснения. Рассмотрим какую-либо игру. В этой игре мы выделяем игроков *Скептика* и *Внешнее окружение*, которое также делает свои ходы (в ответ на ходы произведенные *Скептиком*). *Скептик* делает свои ходы, используя известные ему ходы *Внешнего окружения* и изменяя при этом свой капитал $\mathcal{K}$. *Скептик* имеет в начале игры начальный капитал (возможно нулевой) и может брать в долг без процентов.

Каждой реализации игры соответствует траектория ξ *Внешнего окружения*. Для каждой такой траектории определен капитал *Скептика* $\mathcal{K}(\xi)$.

Можно формально определить траектории игры подобного типа и эволюцию капитала *Скептика*. На каждом раунде (шаге) игры *Внешнее окружение* и *Скептик* делают свои ходы: *Внешнее окружение* предъявляет элемент w произвольной природы, *Скептик* предъявляет элемент m.

Траектория *Внешнего окружения* состоит из последовательности его ходов $w_1, w_2, \ldots$. В общем случае можно считать. что каждый ход w *Внешнего окружения* лежит в некотором множестве W, которое может зависеть от начального фрагмента траектории игры, известного до хода w, а очередной ход m *Скептика* является элементом некоторого множества S также зависящего от предшествующей части траектории.

Под стратегией *Скептика* понимается функция $\mathcal{M}$ от начальных фрагментов траекторий. *Скептик* вычисляет свой очередной ход в виде $m = \mathcal{M}(\xi)$, где ξ – начальная часть траектории, известная к моменту его хода.

Стратегия игры определяет капитал *Скептика* $\mathcal{K}^{\mathcal{M}}(\xi)$, который зависит от известной части траектории ξ игры. Капитал *Скептика* $\mathcal{K}^{\mathcal{M}}$ изменяется в процессе игры с помощью *функции*

выигрыша λ на $W \times S$. После очередного хода m *Скептика* и очередного хода w *Внешнего окружения* и происходит изменение капитала *Скептика*:

$$\mathcal{K}^{\mathcal{M}}(\xi w) = \mathcal{K}^{\mathcal{M}}(\xi) + \lambda(w, m),$$

где $\mathcal{K}^{\mathcal{M}}(\emptyset) = \mathcal{K}_0^{\mathcal{M}}$ – его начальный капитал.

Мы будем предполагать, что все ходы *Скептика* образуют линейное пространство: $\alpha_1 m + \alpha_2 m' \in S$ для любых $m, m' \in S$ и вещественных чисел α_1, α_2, а функция λ является линейной по второму аргументу:

$$\lambda(w, \alpha_1 m + \alpha_2 m') = \alpha_1 \lambda(w, m) + \alpha_2 \lambda(w, m')$$

для любых m, m' и вещественных чисел α_1, α_2. В частности, для любых двух стратегий $\mathcal{M}_1$ и $\mathcal{M}_2$ и вещественных чисел α_1 и α_2 линейная комбинация $\mathcal{M} = \alpha_1 \mathcal{M}_1 + \alpha_2 \mathcal{M}_2$ также является статегией и для капиталов соответствующим этим стратегиям выполнено:

$$\mathcal{K}^{\mathcal{M}}(\xi) = \alpha_1 \mathcal{K}^{\mathcal{M}_1}(\xi) + \alpha_2 \mathcal{K}^{\mathcal{M}_2}(\xi)$$

для всех траекторий ξ.

Пусть задано некоторое множество траекторий Ω и $x = x(\xi)$ – функция от траекторий $\xi \in \Omega$. Такая функция будет называться *переменной*, по аналогии со случайной переменной в теории вероятностей.

В финансовой интерпретации, переменная x – это обязательство (контракт) выплатить $x(\xi)$ единиц (денег) по какой бы траектории ξ не развивалась игра.

Купить обязательство x за α означает, что покупатель платит продавцу величину α в начале игры, а продавец обязан возвратить покупателю величину $x(\xi)$ в конце игры, какая бы траектория ξ игры не имела место. Рассмотрим вопрос о том, по какой минимальной цене α продавец может продать (а покупатель купить) переменную $x(\xi)$.

Покупатель может заплатить продавцу величину α меньшую чем возможная выплата $x(\xi)$ в конце игры. В этом случае, продавец должен компенсировать эту разницу путем игры, т.е., использовать полученную сумму α как начальный капитал для игры с

некоторой стратегией $\mathcal{M}$ и получить в конце игры капитал достаточный (или даже больший) для выплаты $x(\xi)$ по обязательству: $\mathcal{K}^{\mathcal{M}}(\xi) + \alpha \geqslant x(\xi)$ для любой траектории ξ игры.

Далее, для произвольной переменной y, выражение $\mathcal{K}^{\mathcal{M}} \geqslant y$ будет означать, что $\mathcal{K}^{\mathcal{M}}(\xi) \geqslant y(\xi)$ для всех траекторий ξ игры.

В данном случае роль продавца играет *Скептик*, который должен обеспечить выполнение обязательства x путем хеджирования в рассматриваемой игре.

Верхней ценой переменной x называется наименьшая цена α, по которой *Скептик* может продать переменную x так, чтобы у него существовала бы стратегия $\mathcal{M}$ игры, при использовании которой он мог бы выполнить обязательство выплатить $x(\xi)$ при любой траектории игры:

$$\overline{E}x = \inf\{\alpha : \exists\mathcal{M}(\mathcal{K}^{\mathcal{M}} \geqslant x - \alpha)\}$$

Верхняя цена определяется по траектории ξ, на которой величина выплаты $x(\xi)$ максимальная.

Рассмотрим теперь вопрос о том, за какую максимальную цену α покупатель может купить (а продавец продать) переменную $x(\xi)$. Продавец получает от покупателя α в начале игры и обязуется выплатить $x(\xi)$ в конце игры. Роль покупателя играет *Скептик*, который должен путем хеджирования компенсировать разность $\alpha - x(\xi)$.

Нижней ценой переменной x называется наибольшая цена α, при которой существует такая стратегия $\mathcal{N}$ *Скептика*, что выполнено условие $\mathcal{K}^{\mathcal{N}} \geqslant \alpha - x$:

$$\underline{E}x = \sup\{\alpha : \exists\mathcal{N}(\mathcal{K}^{\mathcal{N}} \geqslant \alpha - x)\}$$

Нижняя цена определяется по траектории ξ, на которой величина выплаты $x(\xi)$ минимальная.

Купить переменную x за α эквивалентно тому, чтобы продать переменную $-x$ за $-\alpha$. Формально это можно записать в виде

$$\begin{aligned}&\underline{E}(-x) = \sup\{\alpha : \exists\mathcal{N}(\mathcal{K}^{\mathcal{N}} \geqslant \alpha + x)\} = \\ &= -\inf\{\alpha : \exists\mathcal{M}(\mathcal{K}^{\mathcal{M}} \geqslant x - \alpha)\} = -\overline{E}x\end{aligned}$$

Протокол называется *когерентным*, если для каждой стратегии $\mathcal{M}$ существует такая траектория ξ, что $\mathcal{K}^{\mathcal{M}}(\xi) \leqslant \mathcal{K}_0^{\mathcal{M}}$, где $\mathcal{K}_0^{\mathcal{M}}$ – начальный капитал *Скептика*. Это означает, что для любой стратегии *Скептика* существует такая траектория игры, на которой он не может выиграть ничего более чем его начальный капитал.

Предложение 8.1. *Если протокол когерентный, то* $\underline{E}x \leqslant \overline{E}x$ *и* $\underline{E}a = \overline{E}a = a$*, где* a *– переменная, для которой* $a(\xi) = a$ *для всех* ξ*.*

Доказательство. Если $\underline{E}x > \overline{E}x$, то существуют константы $\alpha_1 < \alpha_2$ такие, что

$$\overline{E}x < \alpha_1 < \alpha_2 < \underline{E}x,$$

и стратегии $\mathcal{M}_1$ и $\mathcal{M}_2$ такие, что $\mathcal{K}^{\mathcal{M}_1} \geqslant x - \alpha_1$ и $\mathcal{K}^{\mathcal{M}_2} \geqslant \alpha_2 - x$. Тогда для стратегии $\mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2$ выполнено

$$\mathcal{K}^{\mathcal{M}} = \mathcal{K}^{\mathcal{M}_1} + \mathcal{K}^{\mathcal{M}_2} \geqslant \alpha_2 - \alpha_1 > 0,$$

что противоречит предположению о когерентности протокола игры.

Доказательство второго утверждения предоставляется читателю в виде задачи. $\triangle$

Если $\underline{E}x = \overline{E}x$, это общее значение называется ценой переменной и и обозначается Ex.

Заметим, что в финансовой интерпретации верхняя цена определяется интересами продавца переменной, а нижняя цена определяется интересами покупателя переменной.

В случае, когда верхняя и нижняя цены совпадают: $\underline{E}x = \overline{E}x$ Можно также ввести нижнюю и верхние вариации переменной x: $\underline{V}x = \underline{E}(x - Ex)^2$ и $\overline{V}x = \overline{E}(x - Ex)^2$, где Ex – цена переменной x.

Верхняя и нижняя вероятности события $S \subseteq \Omega$ вводятся для каждого теоретико-игрового протокола. Рассмотрим индикаторную функцию события S:

$$1_S(\xi) = \begin{cases} 1, \text{ если } \xi \in S, \\ 0 \text{ в противном случае.} \end{cases}$$

Эта функция также является переменной, определенной на траекториях игры. Поэтому она имеет верхнюю и нижнюю цены.

Верхней вероятностью события S называется величина

$$\overline{P}(S) = \overline{E}(1_S).$$

Из определения $\overline{P}(S) \leqslant 1$ для любого события S и

$$\overline{P}(S) = \inf\{\alpha : \mathcal{K}_0 = 0 \text{ и } \exists\mathcal{M}\forall\xi(\mathcal{K}^{\mathcal{M}}(\xi) + \alpha \geqslant 1 \text{ если } \xi \in S, \\ \mathcal{K}^{\mathcal{M}}(\xi) + \alpha \geqslant 0 \text{ если } \xi \notin S)\}.$$

Нижней вероятностью события S называется величина

$$\underline{P}(S) = \underline{E}(1_S).$$

По определению,

$$\underline{P}(S) = \sup\{\alpha : \mathcal{K}_0 = 0 \text{ и } \exists\mathcal{N}\forall\xi(\mathcal{K}^{\mathcal{N}}(\xi) \geqslant \alpha - 1 \text{ если } \xi \in S, \\ \mathcal{K}^{\mathcal{N}}(\xi) \geqslant \alpha \text{ если } \xi \notin S)\}.$$

Если протокол когерентный, то

$$0 \leqslant \underline{P}(S) \leqslant \overline{P}(S) \leqslant 1$$

и

$$\underline{P}(S) = 1 - \overline{P}(\Omega \setminus S).$$

В качестве примера применения понятия теоретико-игровой вероятности рассмотрим конкретные траектории и соответствующий вариант теоремы Бернулли.

На рис. 8.2 приведен протокол игры с полной информацией – протокол игры Бернулли.

Обозначим $S_N = \sum\limits_{n=1}^{N} x_n$. По определению *Скептик* выигрывает в этой игре, если $\mathcal{K}_n \geqslant 0$ при $n = 1, \ldots, N$ и $\mathcal{K}_n \geqslant 1$ либо $\left|\frac{S_N}{N}\right| < \epsilon$.

Заметим, что этот протокол является когерентным, так как в ответ на любой ход *Скептика* M_n *Природа* может предъявить исход

$$x_n = \begin{cases} 1, \text{ если } M_n < 0, \\ -1 \text{ в противном случае.} \end{cases}$$

В том случае $\mathcal{K}_n \leqslant \mathcal{K}_0$ для всех n.

Участники игры: *Скептик* и *Природа*.[3]
Заданы числа $0 < \epsilon \leqslant 1$ и $\alpha > 0$ – начальный капитал *Скептика*: $\mathcal{K}_0 = \alpha$.
FOR $n = 1, 2, \ \ldots$
Скептик предъявляет число $M_n \in \mathcal{R}$.
Природа предъявляет исход $x_n \in [-1, 1]$.
Скептик обновляет свой капитал: $\mathcal{K}_n = \mathcal{K}_{n-1} + M_n x_n$.
ENDFOR

Рис. 8.2: Протокол игры Бернулли

Теорема 8.2. *Скептик имеет выигрышную стратегию при* $N \geqslant \frac{1}{\alpha\epsilon^2}$. *Кроме этого,*

$$\overline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \frac{1}{N\epsilon^2}.$$

Из когерентности протокола следует, что аналогичное неравенство выполнено для нижней вероятности.

Доказательство. Предварительно заметим, что

$$\begin{aligned} S_n^2 &= S_{n-1}^2 + 2x_n S_{n-1} + x_n^2 = \\ &= S_{n-1}^2 + 2x_n S_{n-1} + 1, \end{aligned}$$

где $S_0 = 0$.

Пусть *Скептик* выбирает свою стратегию в виде: $M_n = \frac{2\alpha S_{n-1}}{N}$ на каждом шаге n игры. Тогда

$$\begin{aligned} \mathcal{K}_N - \mathcal{K}_0 &= \frac{\alpha}{N}\sum_{n=1}^{N} 2S_{n-1}x_n = \\ &= \frac{\alpha}{N}(S_N^2 - N) = \alpha\left(\frac{S_N^2}{N} - 1\right). \end{aligned}$$

Следовательно, $\mathcal{K}_N = \alpha\frac{S_N^2}{N}$. Отсюда

$$\left|\frac{S_N}{N}\right| \leqslant \sqrt{\frac{\mathcal{K}_N}{\alpha N}}. \tag{8.6}$$

По определению, *Скептик* выигрывает, если $\mathcal{K}_N \geqslant 1$. Если $\mathcal{K}_N < 1$ и $N > \frac{1}{\alpha\epsilon^2}$, то по (8.6)

$$\left|\frac{S_N}{N}\right| < \epsilon.$$

Таким образом, *Скептик* опять выигрывает.

Сумма S_N зависит от траектории игры: $S_N = S_N(\xi)$. Оценим верхнюю вероятность события, состоящего в том, что среднее значение указанной суммы отклоняется от нуля более чем на некоторое число ϵ:

$$E = \left\{\xi : \left|\frac{S_N(\xi)}{N}\right| > \epsilon\right\},$$

состоящее из всех траекторий игры для которых выполнено указанное неравенство. Его верхняя вероятность равна

$$\overline{P}(E) = \inf\{\alpha : \mathcal{K}_0 = \alpha \text{ и } \exists\mathcal{M}\forall\xi(\mathcal{K}_N^{\mathcal{M}}(\xi) \geqslant 1, \text{ если } \left|\frac{S_N(\xi)}{N}\right| > \epsilon,$$
$$\mathcal{K}_N^{\mathcal{M}}(\xi) \geqslant 0, \text{ если } \left|\frac{S_N(\xi)}{N}\right| \leqslant \epsilon)\}.$$

Как мы только что доказали, такая стратегия $\mathcal{M}$ существует при $\alpha = \frac{1}{N\epsilon^2}$. Следовательно,

$$\underline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \overline{P}\left\{\left|\frac{S_N}{N}\right| > \epsilon\right\} \leqslant \frac{1}{N\epsilon^2}.$$

Теорема доказана. $\triangle$

8.3. Игры на универсальные предсказания

В этом разделе мы покажем, что при некоторой модификации игры из раздела 8.1 *Скептик* используя безопасную стратегию может “вынудить” *Предсказателя* выдавать прогнозы, которые калибруются на произвольной бесконечной последовательности исходов, выдаваемых *Природой*.

Рассмотрим некоторую бесконечно повторяющуюся игру между тремя игроками: *Предсказатель*, *Скептик* и *Природа*.

Пусть $\mathcal{K}_0 = 1$.

FOR $n = 1, 2, \ \ldots$
Скептик предъявляет функцию $S_n : [0,1] \to \mathcal{R}$.
Предсказатель предъявляет прогноз $p_n \in [0,1]$.
Природа предъявляет исход $\omega_n \in \{0,1\}$.
Скептик обновляет свой *выигрыш*: $\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(\omega_n - p_n)$.
ENDFOR

Победители в бесконечной детерминированной игре:
Предсказатель выигрывает в этой игре, если выигрыш *Скептика* $\mathcal{K}_n$ остается ограниченным; в противном случае выигрывают *Скептик* и *Природа*.

Рис. 8.3: Протокол игры с детерминированными предсказаниями

Действия игроков регулируются протоколом, приведенным на рис. 8.3.

Основное свойство игры представлено в следующей теореме.

Теорема 8.3. *Скептик и Природа имеют выигрышную стратегию в детерминированной игре на предсказание.*

Доказательство. Действительно, *Скептик* может определить

$$S_n(p) = \begin{cases} 1, \text{ если } p < 0.5, \\ -1 \text{ в противном случае.} \end{cases}$$

Природа может определять на каждом шаге n игры

$$\omega_n = \begin{cases} 1, \text{ если } p_n < 0.5, \\ 0 \text{ в противном случае.} \end{cases}$$

Тогда в такой игре на каждом шаге $n > 0$, если $\omega_n = 0$, то $p_n \geqslant \frac{1}{2}$ и поэтому $\omega_n - p_n \leqslant -\frac{1}{2}$ и $S_n(p_n) = -1$; если же $\omega_n = 1$, то $p_n < \frac{1}{2}$ и поэтому $\omega_n - p_n \geqslant \frac{1}{2}$ и $S_n(p_n) = 1$. Отсюда следует, что

$$\mathcal{K}_n \geqslant \mathcal{K}_{n-1} + \frac{1}{2}$$

для всех n, и выигрыш *Скептика* неограничен. $\triangle$

Пусть $\mathcal{K}_0 = 1$ и $\mathcal{F}_0 = 1$.
FOR $n = 1, 2, \ \ldots$
Скептик предъявляет функцию $S_n : [0,1] \to \mathcal{R}$.
Предсказатель предъявляет распределение вероятностей на множестве всех предсказаний: $P_n \in \mathcal{P}[0,1]$.
Природа предъявляет исход $\omega_n \in \{0,1\}$.
Предсказатель предъявляет тест случайности $f_n : [0,1] \to \mathcal{R}$, удовлетворяющий условию корректности относительно меры P_n, а именно, $\int f_n(p) P_n(dp) \leqslant 0$.
Генератор случайных чисел предъявляет число $p_n \sim P_n$.
Скептик обновляет свой выигрыш: $\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(\omega_n - p_n)$.
Предсказатель обновляет свой выигрыш: $\mathcal{F}_n = \mathcal{F}_{n-1} + f_n(p_n)$.
ENDFOR

Рис. 8.4: Протокол игры с вероятностными предсказаниями

В этой игре "враждебная" *Природа* использует прогноз *Предсказателя* для формирования своего исхода.

Оказывается, что в рандомизированном варианте этой игры выигрывает *Предсказатель*. В рандомизированном варианте игры *Природа* не будет знать точного прогноза *Предсказателя*, ей известно только распределение вероятностей, согласно которому генерируется этот прогноз.

Рассмотрим бесконечно повторяющуюся игру между четырьмя игроками: *Предсказатель*, *Скептик*, *Природа*, *Генератор случайных чисел*, множество исходов – $\{0,1\}$, $\mathcal{P}\{0,1\}$ – множество всех вероятностных мер на $\{0,1\}$. Каждая мера $Q \in \mathcal{P}\{0,1\}$ задается двумя числами $(q, 1-q)$, где $q = Q\{1\}$ – вероятность 1, $\mathcal{P}[0,1]$ – множество всех вероятностных мер на $[0,1]$.

Игра регулируется протоколом, представленным на рис. 8.4.

Протокол определяет порядок действий игроков и доступную им информацию. Каждый игрок при выборе своей стратегии может использовать всю информацию, которая появилась до его хода – исходы, прогнозы, стратегии других игроков.

Ограничения для Скептика: Скептик должен выбирать S_n так, что его выигрыш $\mathcal{K}_n \geqslant 0$ для всех n независимо от ходов

других игроков.

Ограничения для Предсказателя: Предсказатель должен выбирать P_n и f_n так, чтобы его выигрыш $\mathcal{F}_n \geqslant 0$ для всех n независимо от ходов других игроков. [4]

Победители в рандомизированной игре на предсказания:

Предполагаем, что стратегии игроков таковы, что данные ограничения выполнены. Если игрок хотя бы один раз нарушает ограничение, то он уже не может быть победителем в игре.

Предсказатель выигрывает в этой игре, если (i) его выигрыш $\mathcal{F}_n$ неограничен или если (ii) выигрыш *Скептика* $\mathcal{K}_n$ остается ограниченным; в остальных случаях выигрывают другие игроки.

В следующей теореме мы докажем, что *Предсказатель* имеет выигрышную стратегию в этой игре.

Теорема 8.4. *Предсказатель имеет выигрышную стратегию в вероятностной игре на предсказания.*

Доказательство. На каждом шаге n нашей игры рассмотрим вспомогательную игру с нулевой суммой между *Природой* и *Предсказателем*, которая заключается в следующем.

Предсказатель выбирает число $p_n \in [0,1]$, *Природа* выбирает число $\omega_n \in \{0,1\}$. Потери *Предсказателя* (выигрыш *Природы*) равны:

$$F(\omega_n, p_n) = S(p_n)(\omega_n - p_n).$$

Для любой смешанной стратегии *Природы* $Q_n \in \mathcal{P}\{0,1\}$, *Предсказатель* предъявляет чистую стратегию $p_n = Q\{1\}$. Чистая стратегия p_n соответствует смешанной стратегии $P_n(p_n) = 1$, $P_n(r) = 0$ при $r \in [0,1] \setminus \{p_n\}$.

[4] Выигрыш $\mathcal{F}_n$ соответствует понятию ограниченного снизу супермартингала в теории вероятностей, а $f_n(p)$ соответствует супермартингал-разности $\mathcal{F}_n - \mathcal{F}_{n-1}$. Условия игры требуют, чтобы $\mathcal{F}_0 = 1$ и в процессе игры $\mathcal{F}_n \geqslant 0$ для всех n. Условие $\int f_n(p) P_n(dp) \leqslant 0$ для всех n влечет $\int \mathcal{F}_n P_n(dp) \leqslant \mathcal{F}_{n-1}$ для всех n. Эти свойства составляют определение супермартингала в теории вероятностей. В нашем случае, эти свойства должны выполняться только на траектории прогнозов $p_1, p_2, \ldots$, которые генерируются в процессе игры.

Тогда математическое ожидание выигрыша *Природы* относительно смешанной стратегии Q и чистой стратегии p_n равно

$$\begin{aligned} F(Q_n, P_n) = Q\{0\}F(0, p_n) + Q\{1\}F(1, p_n) = \\ = Q\{0\}S(p_n)(-p_n) + Q\{1\}S(p_n)(1 - p_n) = \\ = (1 - Q\{1\})S(p_n)(-Q\{1\}) + Q\{1\}S(p_n)(1 - Q\{1\}) = 0. \end{aligned}$$

Таким образом, $\forall\ Q\ \exists\ P\ F(Q, P) \leqslant 0$ или

$$\sup_Q \inf_P F(Q, P) \leqslant 0. \tag{8.7}$$

Для того чтобы применить минимаксную теорему, надо превратить эту игру в матричную.

Мы уже имеем две строки, которые соответствуют предсказаниям *Природы* $\omega_n \in \{0, 1\}$.

Рассмотрим некоторое приближение к вспомогательной игре, в котором множество столбцов, соответствующих ходам *Предсказателя*, конечное. Для произвольного $\Delta > 0$ выберем в множестве $[0, 1]$ конечную ϵ-сеть N_ϵ, состоящую из рациональных точек, такую, что каждая точка этого отрезка находится на расстоянии не более чем ϵ от одной из точек этого множества, а также, чтобы нижнее значение игры не превосходило $\Delta/2$, если *Предсказатель* выбирает $p_n \in N_\epsilon$.

Такую ϵ-сеть можно выбрать, так как $|S_n(p)| \leqslant \mathcal{K}_{n-1} \leqslant 2^{n-1}$ ограничено по p (зависит только от n). [5] Тогда неравенство (8.7) будет преобразовано в неравенство

$$\sup_Q \inf_P F(Q, P) \leqslant \Delta/2.$$

Согласно минимаксной теореме,

$$\inf_P \sup_Q F(Q, P) = \sup_Q \inf_P F(Q, P) \leqslant \Delta/2.$$

Поэтому *Предсказатель* имеет смешанную стратегию $P \in \mathcal{P}[0, 1]$, сосредоточенную на множестве N_ϵ, такую, что

$$\sup_Q F(Q, P) \leqslant \Delta,$$

[5] Скептик должен выбирать $S_n(p)$ так, что $\mathcal{K}_n \geqslant 0$ для всех n независимо от действий других игроков.

откуда следует, что

$$\int S_n(p)(\omega_n - p)dP(p) \leqslant \Delta \tag{8.8}$$

для обоих значений $\omega_n = 0$ и $\omega_n = 1$.

Пусть E_Δ – подмножество множества $\mathcal{P}[0,1]$ всех вероятностных мер на единичном отрезке, состоящее из мер P, удовлетворяющих условию (8.8) для $\omega_n = 0$ и $\omega_n = 1$ одновременно.

На множестве мер $\mathcal{P}[0,1]$ можно рассмотреть топологию слабой сходимости. Из теории меры известно, что $\mathcal{P}[0,1]$ компактно в этой топологии. Кроме того, E_Δ замкнуто в этой топологии.

Выберем последовательность монотонно убывающих к 0 рациональных чисел Δ_i, $i = 1, 2, \ldots$ Пересечение бесконечной последовательности замкнутых вложенных друг в друга подмножеств компакта непусто, поэтому $\cap E_{\Delta_i} \neq \emptyset$.

Тогда существует вероятностная мера $P_n \in \cap E_{\Delta_i} \subseteq \mathcal{P}[0,1]$, такая, что

$$\int S_n(p)(\omega_n - p)P_n(dp) \leqslant 0 \tag{8.9}$$

для $\omega_n = 0$ и $\omega_n = 1$ одновременно.

Вернемся теперь к нашей основной игре. Стратегия *Предсказателя* будет заключаться в выборе на шаге n вероятностного распределения P_n, которое было определено в вспомогательной игре. Его вторым ходом будет выбор теста f_n для проверки случайного числа

$$f_n(p) = S_n(p)(\omega_n - p).$$

Тогда $\mathcal{F}_n = \mathcal{K}_n$ для всех n.

Среднее значение f_n по мере P_n не превосходит 0 по (8.9), т.е. f_n – корректный относительно меры P_n тест.

Из $\mathcal{F}_n = \mathcal{K}_n$ получаем, что всегда будет выполнено одно из двух: либо выигрыш *Скептика* ограничен, либо выигрыш *Предсказателя* неограничен. В обоих случаях *Предсказатель* выигрывает. $\triangle$

Будем говорить, что *Генератор случайных чисел* выдает правильные случайные числа, если $\sup_n \mathcal{F}_n < \infty$.

8.4. Рандомизированные калибруемые предсказания

В этом разделе мы покажем, что *Скептик*, выбирая специальным образом свою безопасную стратегию $S_n(p)$, может добиться того, чтобы *Предсказатель* выбирал свои прогнозы так, чтобы они были хорошо калибруемыми на произвольной последовательности исходов, как бы *Природа* их не выбирала.

Предварительно рассмотрим простой случай – *Предсказатель* будет выбирать свои прогнозы так, чтобы выполнялся некоторый теоретико-игровой вариант закона больших чисел. Идея конструкции та же, что и в разделе 8.1.

Пусть ϵ – произвольное положительное число такое, что выполнено $0 < \epsilon < \frac{1}{2}$. Полагаем $\mathcal{K}_0^1 = 1$. В определенной выше рандомизированной игре на предсказание полагаем

$$S_n^1(p) = \epsilon \mathcal{K}_{n-1}^1,$$

т.е. стратегия *Скептика* не зависит от прогноза *Предсказателя* на шаге n, а зависит от выигрыша *Скептика*, полученного им на шагах $< n$.

В этом случае выигрыш *Скептика* на шаге n равен

$$\mathcal{K}_n^1 = \prod_{i=1}^{n} (1 + \epsilon(\omega_i - p_i)), \tag{8.10}$$

где $\omega_1, \ldots, \omega_n$ – последовательность исходов, предложенных *Природой*, а $p_1, \ldots, p_n$ – последовательность прогнозов, предложенных *Предсказателем* на шагах $1, \ldots, n$.

Так как $|\omega_i - p_i| \leqslant 1$ для всех i, $\mathcal{K}_n^1 \geqslant 0$ для всех n независимо от действий других игроков, т.е. основное требование к к стратегии *Скептика* выполнено.

По теореме 8.4 *Предсказатель* имеет выигрышную стратегию в вероятностной игре на предсказания. Это означает, что если *Генератор случайных чисел* выдает правильные случайные числа, т.е. $\sup_n \mathcal{F}_n < \infty$, то как бы *Природа* не выдавала свои исходы $\omega_1, \ldots, \omega_n$, *Предсказатель* обладает методом прогнозирования,

при котором для его прогнозов $p_1, \ldots, p_n$ выигрыш *Скептика* $\mathcal{K}_n^1$ ограничен, скажем некоторым числом $C > 0$:

$$\prod_{i=1}^{n}(1+\epsilon(\omega_i - p_i)) \leqslant C$$

для всех n. Это неравенство можно переписать в виде

$$\begin{aligned}
\sum_{i=1}^{n} \ln(1+\epsilon(\omega_i - p_i)) &\leqslant \ln C, \\
\epsilon \sum_{i=1}^{n}(\omega_i - p_i) - \epsilon^2 \sum_{i=1}^{n}(\omega_i - p_i)^2 &\leqslant \ln C, \\
\epsilon \sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \ln C + \epsilon^2 n, \\
\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \frac{\ln C}{\epsilon n} + \epsilon \qquad (8.11)
\end{aligned}$$

для всех n. Здесь мы использовали неравенство $\ln(1+t) \geqslant t - t^2$ при $t \geqslant -0.5$.

Отсюда следует, что

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \leqslant \epsilon. \qquad (8.12)$$

Аналогичным образом, полагая $\mathcal{K}_0^2 = 1$ и выбирая стратегию

$$S_n^2(p) = -\epsilon \mathcal{K}_{n-1}^2,$$

Скептик может добиться того, чтобы *Предсказатель* выдавал свои прогнозы так, чтобы было выполнено неравенство

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \geqslant -\epsilon. \qquad (8.13)$$

Обе эти стратегии можно соединить в одну стратегию, которая обеспечивает одновременное выполнение обоих неравенств (8.12)

и (8.13). В этом случае стратегии $S_n^1(p)$ и $S_n^2(p)$ и соответствующие капиталы $\mathcal{K}_n^1(p)$, $\mathcal{K}_n^2(p)$ рассматриваются *Скептиком* как вспомогательные в его расчетах.

Для игры *Скептик* выбирает стратегию

$$S_n(p) = \frac{1}{2}(S_n^1(p) + S_n^2(p)).$$

Тогда его выигрыш на шаге n будет равен

$$\mathcal{K}_n = \frac{1}{2}(\mathcal{K}_n^1 + \mathcal{K}_n^2).$$

Заметим, что каждый из выигрышей будет удовлетворять условиям $\mathcal{K}_n^1 \geqslant 0$ и $\mathcal{K}_n^2 \geqslant 0$ для всех n. На первом шаге $S_1(p) = 0$, так как $S_1^1(p) = -S_1^2(p)$, затем значения $S_n^1(p)$ и $S_n^2(p)$ разойдутся, так как они определяются на основании своих выигрышей $\mathcal{K}_n^1(p)$ и $\mathcal{K}_n^2(p)$.

Пусть *Генератор случайных чисел* выдает правильные случайные числа, т.е. $\sup_n \mathcal{F}_n < \infty$.

Из ограниченности суммарного выигрыша $\mathcal{K}_n$ будет следовать ограниченность каждого из выигрышей $\mathcal{K}_n^1$ и $\mathcal{K}_n^2$. Как было доказано выше, такая ограниченность выигрышей влечет одновременное выполнение предельных неравенств (8.12) и (8.13).

Следующий шаг заключается в том, чтобы построить стратегию *Скептика*, которая обеспечивает одновременное выполнение неравенств (8.12) и (8.13) для всех $\epsilon > 0$.

Для этого введем последовательность $\epsilon_k = 2^{-k}$ для всех k. Определим $\mathcal{K}_0^{1,k} = 1$ и $\mathcal{K}_0^{2,k} = 1$ для всех k. Рассмотрим последовательность стратегий

$$\begin{aligned}
S_n^{1,k}(p) &= \epsilon_k \mathcal{K}_{n-1}^{1,k}, \\
S_n^{2,k}(p) &= -\epsilon_k \mathcal{K}_{n-1}^{2,k}, \\
S_n^{+}(p) &= \sum_{k=1}^{\infty} 2^{-k} S_n^{1,k}(p), \\
S_n^{-}(p) &= \sum_{k=1}^{\infty} 2^{-k} S_n^{2,k}(p), \\
S_n(p) &= \frac{1}{2}(S_n^{+}(p) + S_n^{-}(p)).
\end{aligned}$$

Соответствующие выигрыши связаны условиями

$$\mathcal{K}_n^+ = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{1,k},$$
$$\mathcal{K}_n^- = \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{2,k},$$
$$\mathcal{K}_n = \frac{1}{2}(\mathcal{K}_n^+ + \mathcal{K}_n^-).$$

Все эти ряды сходятся, так как для любого фиксированного n будет $\mathcal{K}_n^{1,k} \leqslant 2^n$ для всех k по формуле (8.10). Отсюда и из определения $|S_n^{2,k}(p)| \leqslant 2^{n-1}$ для всех n.

Заметим, что каждый из выигрышей удовлетворяет условиям $\mathcal{K}_n^{1,k} \geqslant 0$ и $\mathcal{K}_n^{2,k} \geqslant 0$ для всех n и k. Поэтому из равномерной ограниченности суммарного выигрыша $\mathcal{K}_n$ следует ограниченность каждого из выигрышей $\mathcal{K}_n^{1,k}$ и $\mathcal{K}_n^{2,k}$. Как было доказано выше, ограниченность каждого из этих выигрышей влечет одновременное выполнение предельных неравенств (8.12) и (8.13) уже теперь для всех ϵ_k, $k = 1, 2, \ldots$

Отсюда получаем, что смешанная стратегия *Скептика* вынуждает *Предсказателя*, чтобы выиграть в игре согласно теореме 8.4, выдавать такие прогнозы, что выполнено следующее условие калибруемости:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0. \tag{8.14}$$

Определение калибруемости (8.14) обладает очевидным недостатком. Например, хорошо калибруемыми прогнозами для последовательности $\omega_1, \omega_2, \ldots = 0, 1, 0, 1, 0, 1, 0, 1, \ldots$ является последовательность $p_1, p_2, \ldots = \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots$ Однако, если рассматривать только члены последовательности исходов, имеющие четные (или нечетные) индексы, подобные прогнозы уже не будут хорошо калибруемыми на соответствующей подпоследовательности. Поэтому необходимо ввести в рассмотрение дополнительные правила выбора подпоследовательностей.

Пусть в процессе игры *Природа* выдает последовательность исходов $\omega_1,\ \omega_2\ldots,$ *Предсказатель* выдает прогнозы $p_1, p_2,\ \ldots$ *Правилом выбора* называется функция

$$F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{n-1}, \omega_{n-1}, p_n),$$

определенная на последовательностях типа

$$p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{n-1}, \omega_{n-1}, p_n,$$

где p_n – прогноз *Предсказателя* на шаге n, $n = 1, 2,\ \ldots,$ и принимающая только два значения: 0 и 1.

Последовательность прогнозов $p_1, p_2,\ \ldots$ называется *хорошо калибруемой* на последовательности исходов $\omega_1, \omega_2,\ \ldots$ относительно правила выбора $F(p_1, \omega_1, p_2, \omega_2,\ \ldots\ , p_{n-1}, \omega_{n-1}, p_n)$, если выполнено

$$\sup_n \sum_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{i-1}, \omega_{i-1}, p_i) < \infty$$

или

$$\lim_{n\to\infty} \frac{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{i-1}, \omega_{i-1}, p_i)(\omega_i - p_i)}{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{i-1}, \omega_{i-1}, p_i)} = 0. \quad (8.15)$$

Основной результат теории универсального прогнозирования утверждает

Теорема 8.5. *Для любой счетной последовательности правил выбора F_k, $k = 1, 2,\ \ldots,$ существует такая стратегия Предсказателя (алгоритм вычисления прогнозов P_n по предыстории $p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{n-1}, \omega_{n-1}$), что при любой стратегии Природы, выдающей исходы ω_n на основании известных значений*

$$p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{n-1}, \omega_{n-1}, p_n,$$

последовательность прогнозов $p_1, p_2,\ \ldots,$ выдаваемая генератором случайных чисел, будет хорошо калибруемой относительно любого правила выбора F_k, при условии $sup_n \mathcal{F}_n < \infty$ (т.е. когда генератор случайных чисел выдает правильные случайные числа).

Доказательство. Доказательство теоремы представляет собой следующий шаг усложнения рассматриваемой выше конструкции.

В конструкции стратегий $S_n^{1,k}(p)$ и $S_n^{1,k}(p)$ число ϵ_k заменим на пару $\epsilon_k F_s$, где $k, s = 1, 2, \ \ldots$

Рассмотрим бесконечные последовательности вспомогательных стратегий *Скептика*:

$$S_n^{1,k,s}(p) = \epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p)\mathcal{K}_{n-1}^{1,k,s},$$
$$S_n^{2,k,s}(p) = -\epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{n-1}, \omega_{n-1}, p)\mathcal{K}_{n-1}^{2,k,s}.$$

Введем какую-нибудь эффективную нумерацию всех пар натуральных чисел (k, s). Пусть для такой пары с номером i будет $p(i) =$ и $q(i) = s$. Такую нумерацию и функции $p(i)$ и $q(i)$ легко опрелелить конкретным образом. Определим

$$S_n^+(p) = \sum_{j=1}^{\infty} 2^{-j} S_n^{1,p(j),q(j)}(p),$$
$$S_n^-(p) = \sum_{j=1}^{\infty} 2^{-j} S_n^{2,p(j),q(j)}(p),$$
$$S_n(p) = \frac{1}{2}(S_n^+(p) + S_n^-(p)).$$

Далее доказательство проходит аналогично случаю смешивания стратегий *Скептика* с множителями ϵ_k.

Заметим, что суммирование в модернизированном варианте (8.11) должно производиться только по тем i, для которых

$$F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i) = 1.$$

В модернизированном варианте (8.11) и в (8.14) для того, чтобы получить (8.15), надо n в знаменателе заменить на

$$\sum_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i).$$

8.5. Задачи и упражнения

1. Доказать, что в теореме 8.1 можно заменить условие $\sup \mathcal{K}_n = \infty$ на условие $\lim\limits_{n\to\infty} \mathcal{K}_n = \infty$ (Указание: для этого надо вместо одной стратегии $M_n = \epsilon\mathcal{K}_{n-1}$ рассмотреть бесконечно много стратегий вида

$$M_n^C = \begin{cases} M_n, \text{ если } \mathcal{K}_{n-1} \leqslant 2^C, \\ 0, \text{ в противном случае,} \end{cases}$$

где C – произвольное положительное целое число. После этого, рассмотрим смесь этих стратегий $\tilde{M}_n = \sum\limits_{C=1}^{\infty} 2^{-C} M_n^C$. Соответствующий капитал Скептика обозначаем $\mathcal{K}_n^C$. Необходимо показать, что на произвольном шаге n капитал Скептика, который придерживается стратегии $\tilde{M}_n$, равен $\tilde{\mathcal{K}}_n = \sum\limits_{C=1}^{\infty} 2^{-C} \mathcal{K}_n^C$. Отсюда уже легко получить, что $\sup \mathcal{K}_n = \infty$ тогда и только тогда, когда $\lim\limits_{n\to\infty} \tilde{\mathcal{K}}_n = \infty$).

2. Доказать, что в произвольной игре для верхней цены любых переменных x, x_1, x_2 выполнены неравенства:

a) $\overline{E}x \leqslant \sup\limits_{\xi\in\Omega} x(\xi)$.

b) $\underline{E}a = \overline{E}a = a$, где a – константа.

c) $\overline{E}(x_1 + x_2) \leqslant \overline{E}x_1 + \overline{E}x_2$.

d) $\overline{E}(x + \alpha) = \overline{E}x + \alpha$, где α константа.

e) $\overline{E}(\alpha x) = \alpha\overline{E}x$ при $\alpha > 0$.

f) если $x_1 \leqslant x_2$, то $\overline{E}x_1 \leqslant \overline{E}x_2$.

3. Сформулировать и доказать аналогичные неравенства для нижней цены $\underline{E}x$ переменной x.

4. Доказать, что в любой игре $0 \leqslant \overline{P}(S) \leqslant 1$ и $\underline{P}(S) \leqslant 1$.

5. Доказать, что в игре с когерентным протоколом для верхней и нижней вероятностей любых событий E, E_1 и E_2 выполнены неравенства:

a) $0 \leqslant \underline{P}(E) \leqslant \overline{P}(E) \leqslant 1$.

b) $\underline{P}(\Omega) = \overline{P}(\Omega) = 1$.

c) $\overline{P}(E) = 1 - \underline{P}(\Omega \setminus E)$.

d) $\overline{P}(E_1 \cup E_2) \leqslant \overline{P}(E_1) + \overline{P}(E_2)$.

e) $\underline{P}(E_1 \cap E_2) \geqslant \underline{P}(E_1) + \underline{P}(E_2) - 1$.

f) если $E_1 \subseteq E_2$, то $\overline{P}(E_1) \leqslant \overline{P}(E_2)$.

6. Пусть простая игра на предсказания из раздела 8.2 (протокол Бернулли) с исходами из множества $\{-1, 1\}$ продолжается N раундов. Доказать, что верхняя и нижняя вероятность любой траектории ξ_0 длины N равны 2^{-N} и, соответственно, верхняя и нижняя вероятность любого конечного множества S равны $2^{-N}|S|$ (Указание: Пусть траектория ξ_0 фиксирована и может использоваться стратегией *Скептика*. Для оценки сверху верхней вероятности заданной траектории ξ_0 рассмотрим полагаем $\mathcal{K}_0 = 2^{-N}$ и определим стратегию *Скептика*: $M_1 = 2^{-N}$ и $M_t = \mathcal{K}_{t-1}$ при $t \geqslant 2$ вдоль этой траектории. Определим $M_t = 0$ после того как на шаге $t-1$ траектория ξ игры разошлась с заданной траекторией ξ_0. Тогда капитал *Скептика* будет удваиваться на каждом шаге, пока траектория игры ξ совпадает с заданной ξ_0. Капитал станет равным нулю, как только траектория ξ игры разойдется с заданной траекторией ξ_0. Поэтому $\mathcal{K}_N(\xi_0) = 1$ и $\mathcal{K}_N(\xi) = 0$ при $\xi \neq \xi_0$.

Для оценки снизу нижней вероятности полагаем $\alpha = 2^{-N}$ и рассмотрим стратегию $M_1 = -\alpha$ и $M_t = \mathcal{K}_{t-1}$ при $t \geqslant 2$ вдоль заданной траектории. Определим $M_t = 0$ после того как траектория игры разошлась с заданной траекторией ξ. При этом $\mathcal{K}_0 = 0$. Пока траектория игры совпадает с заданной траекторией ξ долг *Скептика* в конце шага s равен $-\sum_{i=0}^{s-1} 2^i\alpha = -(2^s - 1)\alpha$. Если траектория игры впервые разошлась с траекторией ξ на шаге s, то *Скептик* выигрывает $\alpha 2^s$ и после отдачи долга у него остается α. Если траектория игры совпадает с траекторией ξ, долг *Скептика* в конце игры равен $-(2^N - 1)\alpha$. Поэтому по окончании игры капитал *Скептика* равен α).

Привести примеры событий, для которых можно точно вычислить верхние и нижние вероятности.

7. Естественно рассматривать протоколы, когерентные в более широком смысле, чем рассматривалось ранее. Рассматриваем все траектории $\xi^N = \xi_1, \ldots, \xi_N$ длины N.

Протокол называется *когерентным*, если для каждой стратегии $\mathcal{M}$ и любой неполной траектории $\xi^n = \xi_1, \ldots, \xi_n$ длины $n < N$ существует такое ее продолжение $\xi_{n+1}^N = \xi_{n+1}, \ldots, \xi_N$, что $\mathcal{K}^{\mathcal{M}}(\xi^N) \leqslant \mathcal{K}^{\mathcal{M}}(\xi^n)$, где $\xi^N = \xi^n\xi_{n+1}^N$. Это значит, что для лю-

бой точки траектории ξ существует ее продолжение, при котором стратегия $\mathcal{M}$ не может выиграть больше чем она имела в этой точке.

Доказать, что при таком определении когерентного протокола верхняя вероятность события S равна

$$\overline{P}(S) = \inf\{\alpha : \exists\mathcal{M}\forall\xi(\forall n(\mathcal{K}^{\mathcal{M}}(\xi^n) + \alpha \geqslant 0) \text{ и } \\ \mathcal{K}^{\mathcal{M}}(\xi) + \alpha \geqslant 1 \text{ если } \xi \in S)\}.$$

Первое условие означает, что теперь при определении верхней вероятности события мы можем потребовать, чтобы капитал *Скептика* являлся неотрицательным не только в конце игры, но и на всех ее шагах.

8. Завершить доказательство теорем 8.1 и 8.5.

Глава 9

Повторяющиеся игры

В главе 7 рассматривались однократные реализации игр и вычислялись их характеристики. Вычисление точек равновесия в таких играх является вычислительно трудоемкими процедурами. В частности, как было показано, для такого вычисления необходимо решать задачу линейного программирования. В этой главе мы покажем, что используя калибруемые предсказания можно приближать точки равновесия Нэша или точки коррелированого равновесия в неограниченно повторяющихся играх с помощью частотных распределений.

В разделе 9.1 мы рассмотрим асимптотические характеристики бесконечно повторяющихся игр с нулевой суммой и покажем, что построенные ранее алгоритмы, машинного обучения приближают точки равновесия Нэша.

В разделе 9.2 мы докажем теорему Блекуэлла о достижимости, которая является обобщением минимаксной теоремы на случай векторно значных функций выигрыша. В разделе 9.3 мы применим эту теорему для построения калибруемых предсказаний для случая произвольного конечного числа исходов.

Далее, в разделе 9.4 будет показано, что если в некоторой неограниченно повторяющейся игре все игроки используют предсказания, которые калибруются на последовательностях стратегий, выбранных в игре этими оппонентами, и выбирают “оптимальный ответ” на эти предсказания, то совместное частотное

распределение стратегий игроков сходится к множеству коррелированных равновесий игры.

В разделе 9.6 мы приводим метод построения генеративно-состязательной сети (GAN) с помощью бесконечно повторяющейся игры с нулевой суммой и алгоритма онлайн градиентного спуска.

9.1. Бесконечно повторяющиеся игры двух игроков с нулевой суммой

В этом разделе мы рассмотрим игры, повторяющиеся во времени.

Допустим, что на каждом шаге $t = 1, 2, \ldots$ первый игрок выбирает ход $I_t \in \{1, \ldots, N\}$ в соответствии с распределением вероятностей $\bar{p}_t = (p_{1,t}, \ldots, p_{N,t})$ (смешанной стратегией), а второй игрок выбирает ход $J_t \in \{1, \ldots, M\}$ в соответствии с распределением вероятностей $\bar{q}_t = (q_{1,t}, \ldots, q_{N,t})$. Смешанные стратегии игроков $\bar{p}_t$ и $\bar{q}_t$ могут зависеть от предшествующих ходов игроков и их результатов. Выигрыш первого игрока на шаге t равен $f(I_t, J_t)$, а выигрыш второго игрока равен $-f(I_t, J_t)$.

Будем сравнивать кумулятивный выигрыш каждого игрока за n шагов с кумулятивным выигрышем его наилучшей константной стратегии:

$$\max_{i=1,\ldots,N} \sum_{t=1}^{n} f(i, J_t) - \sum_{t=1}^{n} f(I_t, J_t)$$

для первого игрока и

$$\sum_{t=1}^{n} f(I_t, J_t) - \min_{j=1,\ldots,M} \sum_{t=1}^{n} f(I_t, j)$$

для второго игрока.

Мы применим теорию предсказаний с использованием экспертных стратегий для вычисления приближений к равновесию в таких играх. При анализе действий первого игрока множество его ходов $\{1, \ldots, N\}$ будет рассматриваться как множество вспомогательных экспертов. Каждый эксперт i выдает на всех шагах одно и то же предсказание равное $i \in \{1, \ldots, N\}$. Первый игрок

рассматривается как *Предсказатель*, который выдает на каждом шаге t предсказание I_t. При этом ходы $J_t \in \{1, \ldots, M\}$ второго игрока интерпретируются как исходы *Природы*.

Аналогично, при анализе действий второго игрока множество его ходов $\{1, \ldots, M\}$ также рассматривается как множество вспомогательных экспертов. Каждый эксперт j выдает на всех шагах одно и то же предсказание равное $j \in \{1, \ldots, M\}$. Второй игрок рассматривается как *Предсказатель*, который выдает на каждом шаге t предсказание J_t. При этом ходы $I_t \in \{1, \ldots, N\}$ первого игрока интерпретируются как исходы *Природы*.

Разъясним теперь, какие функции потерь используются при этом анализе. Потери первого игрока равны $\lambda^1(J_t, I_t) = -f(I_t, J_t)$, где J_t – исход *Природы*, а I_t – прогноз первого игрока на шаге t. Потери второго игрока равны $\lambda^2(I_t, J_t) = f(I_t, J_t)$, где где I_t – исход *Природы*, а J_t – прогноз второго игрока на шаге t.

Первый (или второй) игрок может выбирать свои ходы (смешанные стратегии) согласно некоторому правилу или алгоритму, который на каждом шаге t выдает распределение вероятностей $\bar{p}_t$ (или $\bar{q}_t$). Алгоритм подобного рода будет называться *онлайн стратегией* первого (или второго) игрока в *бесконечно повторяющейся игре*.

Допустим, что оба игрока выбирают свой ходы в соответствии с онлайн стратегиями, состоятельными по Ханнану (см. определение (4.65)). Например, можно использовать алгоритм экспоненциального взвешивания из раздела 4.6.2. Согласно этому алгоритму первый игрок выбирает свою смешанную стратегию $\bar{p}_t = (p_{1,t}, \ldots, p_{N,t})$ по формуле:

$$p_{i,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \lambda^1(J_s, i)\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \lambda^1(J_s, k)\right)},$$

где $i = 1, \ldots, N$, η_t – переменный параметр обучения. При этом ход J_s второго игрока рассматривается как исход *Природы*.

По следствию 4.8 первый игрок является состоятельным по Ханнану, т.е., при соответствующем выборе параметров η_t имеет

место

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\lambda^1(J_t,I_t)-\min_{i=1,\dots,\ N}\frac{1}{n}\sum_{t=1}^{n}\lambda^1(J_t,i)\right)\leqslant 0 \qquad (9.1)$$

с вероятностью 1, где распределение вероятностей на траекториях первого игрока определяется по последовательности распределений $\bar{p}_t$, $t=1,2,\dots$.

Второй игрок может также может применять аналогичную стратегию. В этом случаю он также будет состоятельным по Ханнану, т.е. с вероятностью 1 имеет место

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\lambda^2(I_t,J_t)-\min_{j=1,\dots,\ M}\frac{1}{n}\sum_{t=1}^{n}\lambda^2(I_t,j)\right)\leqslant 0. \qquad (9.2)$$

В терминах выигрышей (9.1) имеет вид: с вероятностью 1 выполнено

$$\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}f(I_t,J_t)-\max_{i=1,\dots,\ N}\frac{1}{n}\sum_{t=1}^{n}f(i,J_t)\right)\geqslant 0, \qquad (9.3)$$

где траектория $I_1,I_2,\dots$ распределена по мере $\bar{p}_1\times\bar{p}_2\times\dots$ – произведению смешанных стратегий первого игрока.

Соотношение (9.2) имеет вид: с вероятностью 1 имеет место

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}f(I_t,J_t)-\min_{j=1,\dots,\ M}\frac{1}{n}\sum_{t=1}^{n}f(I_t,j)\right)\leqslant 0, \qquad (9.4)$$

где траектория $J_1,J_2,\dots$ распределена по мере $\bar{q}_1\times\bar{q}_2\times\dots$ – произведению смешанных стратегий второго игрока.

Следующая теорема утверждает, что если первый игрок выбирает свои ходы согласно состоятельной по Ханнану онлайн стратегии для бесконечно повторяющейся игры, то независимо от того как второй игрок выбирает свои ходы, средний выигрыш первого игрока не может быть намного меньше чем цена игры. Аналогичное утверждение верно для второго игрока – если второй игрок выбирает свои ходы согласно состоятельной по Ханнану онлайн стратегии, то независимо от того как выбирает свои ходы первый игрок, средний проигрыш второго игрока не может быть намного больше чем цена игры.

Теорема 9.1. *Допустим, что в игре двух лиц с нулевой суммой первый игрок выбирает свои ходы согласно онлайн стратегии, состоятельной по Ханнану. Тогда независимо от того как второй игрок выбирает свои ходы*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} f(I_i, J_t) \geqslant v, \tag{9.5}$$

почти всюду, где v – цена игры.

Если каждый игрок придерживается онлайн стратегии, состоятельной по Ханнану, то, с вероятностью 1, имеет место равенство

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} f(I_i, J_t) = v. \tag{9.6}$$

Доказательство. Цена игры представляется в виде

$$v = \max_{\bar{p}} \min_{\bar{q}} f(\bar{p}, \bar{q}) = \min_{\bar{q}} \max_{\bar{p}} f(\bar{p}, \bar{q}).$$

Кроме того,

$$f(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j f(i, j),$$

$$f(\bar{p}, j) = \sum_{i=1}^{N} p_i f(i, j),$$

$$f(i, \bar{q}) = \sum_{j=1}^{M} q_j f(i, j).$$

Согласно соотношению (9.3), для доказательства первого утверждения (9.5) теоремы достаточно показать, что для произвольной последовательности $J_1, J_2, \ldots$ ходов второго игрока

$$\max_{i=1,\ldots,\ N} \frac{1}{n} \sum_{t=1}^{n} f(i, J_t) \geqslant v \tag{9.7}$$

для всех n. Для доказательства заметим, что

$$\max_{i=1,\dots,\ N} \frac{1}{n}\sum_{t=1}^{n} f(i, J_t) = \max_{\bar{p}} \frac{1}{n}\sum_{t=1}^{n} f(\bar{p}, J_t),$$

так как $\sum\limits_{t=1}^{n} f(\bar{p}, J_t)$ линейно по $\bar{p}$, а максимум линейной функции, определенной на симплексе вероятностных распределений на $\{1, \dots,\ N\}$, достигается в одной из его вершин.

Пусть

$$\hat{q}_{j,n} = \frac{1}{n}\sum_{t=1}^{n} 1_{\{J_t = j\}}$$

– частота шагов, на которых второй игрок выбирает ход j. Пусть также $\hat{q}_n = (\hat{q}_{1,n}, \dots,\ \hat{q}_{M,n})$. Тогда

$$\max_{\bar{p}} \frac{1}{n}\sum_{t=1}^{n} f(\bar{p}, J_t) = \max_{\bar{p}} \sum_{j=1}^{M} \hat{q}_{j,n} f(\bar{p}, j) =$$
$$= \max_{\bar{p}} f(\bar{p}, \hat{q}_n) \geqslant \min_{\bar{q}} \max_{\bar{p}} f(\bar{p}, \bar{q}) = v$$

для произвольной последовательности $J_1, J_2, \dots$ ходов второго игрока

Для доказательства второго утверждения (9.6) теоремы воспользуемся условием (9.4) состоятельности по Ханнану и докажем, что

$$\min_{j=1,\dots,\ M} \frac{1}{n}\sum_{t=1}^{n} f(I_t, j) \leqslant v = \max_{\bar{p}} \min_{\bar{q}} f(\bar{p}, \bar{q})$$

для произвольной последовательности $I_1, I_2, \dots$ ходов первого игрока. Доказательство этого утверждения аналогично доказательству неравенства (9.7).

Отсюда получим

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} f(I_i, J_t) \leqslant v, \tag{9.8}$$

почти всюду, где v – цена игры.

Объединяя (9.8) и (9.5) получим (9.6). Теорема доказана. $\triangle$

9.2. Теорема Блекуэлла о достижимости

Теорема 9.1 из предыдущего раздела утверждает, что первый игрок при достаточно большом числе шагов, придерживаясь онлайн стратегии состоятельной по Ханнану, может сделать среднее значение своего выигрыша асимптотически не меньше цены игры, какой бы стратегии не придерживался второй игрок.

В этом разделе мы рассмотрим обобщение этого утверждения на случай векторно–значной функции выигрыша и произвольного замкнутого выпуклого множества S вместо цены игры. Будет доказана знаменитая теорема Блекуэлла о достижимости (Blackwell approachability theorem). Эта теорема представляет необходимые и достаточные условия, при которых существует рандомизированная онлайн стратегия первого игрока для бесконечно повторяющейся игры, придерживаясь которой он с вероятностью 1 при неограиченном продолжении игры может как угодно близко приблизить среднее значение вектора своего выигрыша к заданному множеству S, независимо от того, как бы не выбирал свои ходы второй игрок.

В 1956 г. Блекуэлл [15] предложил обобщение минимаксной теоремы на случай векторнозначной функции выигрыша. Позже было замечено, что эта теорема может быть использована для построения калибруемых предсказаний.

По-прежнему рассматривается игра двух лиц. Только теперь функция выигрыша $\bar{f}(i,j)$ принимает значения в d-мерном пространстве $\mathcal{R}^d$. Напомним, что стратегии первого игрока принадлежат конечному множеству $\mathcal{I} = \{1, \ldots, N\}$, а стратегии второго игрока принадлежат конечному множеству $\mathcal{J} = \{1, \ldots, M\}$. Смешанные стратегии игроков – это распределения вероятностей на множествах $\mathcal{I}$ и $\mathcal{J}$. Их множества обозначаются $\mathcal{P}(\mathcal{I})$ и $\mathcal{P}(\mathcal{J})$ соответственно. При $\bar{p} = (p_1, \ldots, p_N) \in \mathcal{P}(\mathcal{I})$ и $\bar{q} = (q_1, \ldots, q_M) \in$

$\mathcal{P}(\mathcal{J})$

$$\bar{f}(\bar{p}, j) = \sum_{i=1}^{N} p_i \bar{f}(i, j),$$

$$\bar{f}(i, \bar{q}) = \sum_{j=1}^{M} q_j \bar{f}(i, j),$$

$$\bar{f}(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j \bar{f}(i, j).$$

Как обычно, рассматриваем евклидово расстояние

$$\|\bar{x} - \bar{y}\| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

между любыми двумя векторами $\bar{x}, \bar{y} \in \mathcal{R}^d$. Если $S \subseteq \mathcal{R}^d$ и $\bar{x} \in \mathcal{R}^d$, то расстояние от точки $\bar{x}$ до множества S определяется как

$$\operatorname{dist}(\bar{x}, S) = \inf_{\bar{y} \in S} \|\bar{x} - \bar{y}\|.$$

Для замкнутого множества S пусть $d_S(\bar{x})$ обозначает какой-нибудь элемент $\bar{y} \in S$, для которого расстояние $\operatorname{dist}(\bar{x}, \bar{y})$ минимальное. Если к тому же S – выпуклое, то такой элемент единственный.

Множество $S \subseteq \mathcal{R}^d$ называется *достижимым* (approachable), если существует такая рандомизированная онлайн стратегия первого игрока $\bar{p}_1, \bar{p}_2, \ldots$, что для любой последовательности ходов $J_1, J_2, \ldots$ второго игрока для P-почти всех последовательностей $I_1, I_2, \ldots$ ходов первого игрока выполнено

$$\lim_{T \to \infty} \operatorname{dist}\left(\frac{1}{T} \sum_{t=1}^{T} \bar{f}(I_t, J_t), S \right) = 0,$$

где P – общее распределение вероятностей на траекториях $I_1, I_2, \ldots$ ходов первого игрока, которое определяется распределениями $\bar{p}_1, \bar{p}_2, \ldots$.

Следующая теорема дает достаточное условие достижимости замкнутого выпуклого подмножества из $\mathcal{R}^d$.

Предполагаем, что рассматриваемые далее множества S и значения $\bar{f}(i, j)$ находятся в единичном шаре пространства $\mathcal{R}^d$.

Теорема 9.2. *Задано замкнутое выпуклое подмножество $S \subseteq \mathcal{R}^d$. Для каждого вектора $\bar{x} \notin S$ рассмотрим гиперплоскость $\Pi_{\bar{x}}$, проходящую через точку $d_S(\bar{x})$ и ортогональную прямой, соединяющей точки $\bar{x}$ и $d_S(\bar{x})$.*

Допустим, что для каждого вектора $\bar{x} \notin S$ существует распределение $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что все точки $\bar{f}(\bar{p}, 1), \ldots, \bar{f}(\bar{p}, M)$ и точка $\bar{x}$ лежат по разные стороны гиперплоскости $\Pi_{\bar{x}}$. Тогда множество S достижимо.

Доказательство. Пусть $I_1, I_2, \ldots$ и $J_1, J_2, \ldots$ – какие-либо последовательности ходов первого и второго игроков. Обозначим вектор среднего значения выигрышей за первые t шагов

$$\bar{m}_t = \frac{1}{t} \sum_{i=1}^{t} \bar{f}(I_i, J_i).$$

Пусть на шагах игры меньших t игроки уже произвели ходы $I_1, \ldots, I_{t-1}$ и $J_1, \ldots, J_{t-1}$. Предположим, что $\bar{m}_{t-1} \notin S$.

Уравнение гиперплоскости $\Pi_{\bar{m}_{t-1}}$, проходящей через точку $d_S(\bar{m}_{t-1})$ и ортогональную прямой, соединяющей точки $\bar{m}_{t-1}$ и $d_S(\bar{m}_{t-1})$, имеет вид

$$(\bar{w}_{t-1} \cdot \bar{x}) - b_{t-1} = 0,$$

где

$$\bar{w}_{t-1} = \frac{\bar{m}_{t-1} - d_S(\bar{m}_{t-1})}{\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|}$$

и

$$b_{t-1} = (\bar{w}_{t-1} \cdot d_S(\bar{m}_{t-1})).$$

Предполагаем, что $\bar{m}_0 = \bar{0}$.

Заметим, что точка $\bar{m}_{t-1}$ находится выше гиперплоскости (так как является концом направляющего вектора этой гиперплоскости).

По условию теоремы для $\bar{x} = \bar{m}_{t-1}$ существует смешанная стратегия $\bar{p}_t$ первого игрока, для которой все точки

$$\bar{f}(\bar{p}_t, 1), \ldots, \bar{f}(\bar{p}_t, M)$$

находятся ниже данной гиперплоскости:

$$(\bar{w}_{t-1} \cdot \bar{f}(\bar{p}_t, j)) - b_{t-1} \leqslant 0$$

для всех $j = 1, \ldots, M$. Это условие можно также записать в виде

$$\max_{1 \leqslant j \leqslant M} (\bar{w}_{t-1} \cdot (\bar{f}(\bar{p}_t, j) - d_S(\bar{m}_{t-1}))) \leqslant 0. \tag{9.9}$$

Смешанная стратегия $\bar{p}_t$ определяется путем решения задачи линейного программирования (9.9).

Проверим, что точка $\bar{m}_t$ "приближается" к множеству S. Из определения

$$\operatorname{dist}(\bar{m}_t, S) = \|\bar{m}_t - d_S(\bar{m}_t)\| \leqslant \|\bar{m}_t - d_S(\bar{m}_{t-1})\|. \tag{9.10}$$

Нетрудно проверить, что

$$\bar{m}_t = \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t}\bar{f}(I_t, J_t). \tag{9.11}$$

Возведем в квадрат неравенство (9.10) и продолжим выкладки с использованием (9.11) :

$$\begin{aligned}
\operatorname{dist}(\bar{m}_t, S)^2 \leqslant \left\| \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t}\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1}) \right\|^2 = \\
= \left\| \frac{t-1}{t}(\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) + \frac{1}{t}(\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1})) \right\|^2 = \\
= \left(\frac{t-1}{t}\right)^2 \|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 + \\
+2\frac{t-1}{t^2}((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1}))) + \\
+\frac{1}{t^2}\|\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1})\|^2.
\end{aligned} \tag{9.12}$$

Так как множество S и все значения $\bar{f}(i, j)$ находятся в единичном шаре пространства $\mathcal{R}^d$, выполнено неравенство

$$\|\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1})\| \leqslant 2.$$

Используя это неравенство преобразуем неравенства (9.11) и (9.12) в неравенство

$$t^2\|\bar{m}_t - d_S(\bar{m}_t)\|^2 - (t-1)^2\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 \leqslant$$
$$\leqslant 4 + 2(t-1)((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1}))). \quad (9.13)$$

Обозначим

$$K_{t-1} = \frac{t-1}{T}\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|.$$

Выполнено $0 \leqslant K_{t-1} \leqslant 2$ при $t \leqslant T$. Суммируем по $t = 1, \dots, T$ левую и правую части неравенства (9.13) и разделим их на T^2 :

$$\|\bar{m}_T - d_S(\bar{m}_T)\|^2 \leqslant$$
$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (\bar{f}(I_t, J_t) - d_S(\bar{m}_{t-1}))) \leqslant$$
$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (\bar{f}(I_t, J_t) - \bar{f}(\bar{p}_t, J_t)). \quad (9.14)$$

Для получения последнего неравенства мы использовали неравенство (9.9).

Второе слагаемое последнего члена (9.14) представляет собой мартингал-разность. [1] Поэтому оно по следствию 10.6 к неравенству Хефдинга–Азумы почти всюду стремится к 0 при $T \to \infty$. Отсюда

$$\mathrm{dist}(\bar{m}_T, S) = \|\bar{m}_T - d_S(\bar{m}_T)\| \to 0 \text{ при } T \to \infty$$

с вероятностью 1. Теорема доказана. $\triangle$

В следующей теореме дается необходимое и достаточное условие, при котором произвольное замкнутое выпуклое множество достижимо первым игроком.

Теорема 9.3. *Замкнутое выпуклое подмножество $S \subseteq \mathcal{R}^d$ достижимо первым игроком тогда и только тогда, когда для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что $\bar{f}(\bar{p}, \bar{q}) \in S$.*

[1] Так как $E_{\bar{p}_t}(\bar{f}(I_t, J_t)) = \bar{f}(\bar{p}_t, J_t)$, где E – символ математического ожидания. Также K_{t-1} – предсказуемая случайная величина.

Доказательство. Допустим, что для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что $\bar{f}(\bar{p}, \bar{q}) \in S$. Пусть также $\bar{x}_0 \notin S$ и $d_S(\bar{x}_0)$ – ближайшая к $\bar{x}_0$ точка из S. Рассмотрим вспомогательную матричную игру с функцией выигрыша $a(i,j) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{f}(i,j))$. По минимаксной теореме

$$\max_{\bar{p}} \min_{j} a(\bar{p}, j) = \min_{\bar{q}} \max_{i} a(i, \bar{q}). \tag{9.15}$$

По условию теоремы для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p}$ такое, что $\bar{f}(\bar{p}, \bar{q}) \in S$. Отсюда и из (9.15) получаем

$$\begin{aligned} &\max_{\bar{p}} \min_{j}((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{f}(\bar{p}, j)) = \\ &= \min_{\bar{q}} \max_{\bar{p}}((\bar{d}_S(x_0) - \bar{x}_0) \cdot \bar{f}(\bar{p}, \bar{q})) \geqslant \\ &\geqslant \min_{\bar{s} \in S}((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{s}) = \\ &= ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)). \end{aligned} \tag{9.16}$$

Последнее равенство из (9.16) следует из определения вектора $d_S(\bar{x}_0)$ и выпуклости множества S.

Рассмотрим гиперплоскость

$$L(\bar{x}) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{x}) - ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)) = 0,$$

проходящую через точку $d_S(x_0)$ и ортогональную вектору $d_S(\bar{x}_0) - \bar{x}_0$. Легко проверить, что

$$((d_S(x_0) - \bar{x}_0) \cdot \bar{x}_0) < ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)).$$

Отсюда получаем $L(\bar{x}_0) < 0$, т.е., точка $\bar{x}_0$ лежит ниже гиперплоскости $L(\bar{x}) = 0$.

Из неравенства между первым и последним членом цепочки равенств и неравенств (9.16) следует, что существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что для любого $j = 1, \ldots, M$

$$((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{f}(\bar{p}, j)) \geqslant ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)).$$

Иными словами, $L(\bar{f}(\bar{p}, j)) \geqslant 0$ для любого $j = 1, \ldots, M$. Таким образом, гиперплоскость $L(\bar{x}) = 0$ разделяет эти точки и точку $\bar{x}_0$. Следовательно, множество S достижимо по теореме 9.2.

Для доказательства обратного утверждения, допустим, что существует $\bar{q}_0 \in \mathcal{P}(\mathcal{J})$ такое, что $\bar{f}(\bar{p}, \bar{q}_0) \notin S$ для всех $\bar{p} \in \mathcal{P}(\mathcal{I})$.

Применим теорему 9.2 для игры с транспонированной функцией выигрыша $\bar{f}'(i,j) = \bar{f}(j,i)$ и выпуклого замкнутого множества $T(\bar{q}_0) = \{\bar{f}(\bar{p}, \bar{q}_0) : \bar{p} \in \mathcal{P}(\mathcal{I})\}$.

По определению все значения $\bar{f}'(\bar{q}_0, 0), \ldots, \ \bar{f}'(\bar{q}_0, N) \in T(\bar{q}_0)$. Из выпуклости множества $T(\bar{q}_0)$ следует, что для любого $\bar{x} \notin T(\bar{q}_0)$ точки $\bar{x}$ и $\bar{f}'(\bar{q}_0, 0), \ldots, \ \bar{f}'(\bar{q}_0, N)$ находятся по разные стороны от гиперплоскости $\Pi_{\bar{x}}$, проходящей через точку $d_{T(q_0)}(\bar{x})$ и ортогональную прямой, соединяющей точки $\bar{x}$ и $d_{T(q_0)}(\bar{x})$.

Тогда по теореме 9.2 множество $T(\bar{q}_0)$ достижимо для игры с транспонированной функцией выигрыша и с постоянной стратегией $\bar{q}_0$.

Мы допустили, что $T(\bar{q}_0) \cap S = \emptyset$. По условию теоремы и по определению множества S и $T(q_0)$ замкнутые. Нетрудно показать, что в этом случае множество S не достижимо для исходной игры (см. задачу из раздела 9.5). Теорема доказана. $\triangle$

В качестве первого применения теоремы 9.2 построим онлайн стратегию предсказания состоятельную по Ханнану.

Пусть заданы множества ходов $\mathcal{I} = \{1, \ldots, \ N\}$ – первого игрока и $\mathcal{J} = \{1, \ldots, \ M\}$ – второго игрока. $\mathcal{P}(\mathcal{I})$ и $\mathcal{P}(\mathcal{J})$ – множества их смешанных стратегий. Рассматривается игра с функцией потерь $l(i,j)$, где $0 \leqslant l(i,j) \leqslant 1$ для всех i, j.

Наша цель определить на каждом шаге t смешанную стратегию $\bar{p}_t$ первого игрока такую, что для любой последовательности ходов $J_1, J_2, \ldots$ второго игрока с вероятностью 1 было выполнено

$$\limsup_{T \to \infty} \left(\frac{1}{T} \sum_{t=1}^{T} l(I_t, J_t) - \min_{1 \leqslant i \leqslant N} \frac{1}{T} \sum_{t=1}^{T} l(i, J_t) \right) \leqslant 0, \qquad (9.17)$$

где последовательность ходов $I_1, I_2, \ldots$ распределена по мере P порожденной распределениями $\bar{p}_1, \bar{p}_2, \ldots$.

Для того, чтобы применить теорему 9.2 рассмотрим выпуклое замкнутое множество

$$S = \{(u_1, \ldots, \ u_N) : u_i \leqslant 0, i = 1, \ldots, \ N\},$$

а также векторнозначную платежную функцию

$$\bar{f}(i,j) = \begin{pmatrix} l(i,j) - l(1,j) \\ \dots \\ l(i,j) - l(k,j) \\ \dots \\ l(i,j) - l(N,j) \end{pmatrix}.$$

Заметим, что значения $\bar{f}(i,j)$ лежат в N-мерном шаре радиуса $\sqrt{N}$ с центром в начале координат. Умножая эту функцию на константу $1/\sqrt{N}$ можно добиться, чтобы значения $\bar{f}(i,j)$ лежали в единичном шаре.

Рассмотрим произвольный вектор $\bar{x}_0 \notin S$. Достаточно рассмотреть случай когда $d_S(\bar{x}_0) = \bar{0}$ и уравнение гиперплоскости $\Pi_{\bar{x}_0}$ имеет вид $(\bar{w} \cdot \bar{x}) = 0$, где все компоненты w_i нормального вектора $\bar{w}$ гиперплоскости неотрицательные.

Для доказательства существования стратегии такой, что выполнено (9.17), достаточно доказать, что существует смешанная стратегия $\bar{p} \in \mathcal{P}(\mathcal{I})$ такая, что все векторы $\bar{f}(\bar{p},1), \dots,\ \bar{f}(\bar{p},M)$ лежат ниже гиперплоскости $(\bar{w} \cdot \bar{x}) = 0$, т.е. выполнено

$$\sum_{k=1}^{N} w_k (l(\bar{p},j) - l(k,j)) \leqslant 0$$

для всех $j = 1, \dots M$. Легко проверить, что это условие выполнено при

$$\bar{p} = \frac{\bar{w}}{\sum\limits_{i=1}^{N} w_i}.$$

По теореме 9.2 существует последовательность смешанных стратегии $\bar{p}_1, \dots,\ \bar{p}_t, \dots$ первого игрока такая, что условие (9.17) выполнено с вероятностью 1.

9.3. Калибруемые предсказания

В этом разделе мы приведем метод построения калибруемых предсказаний в случае произвольного конечного множества исходов на

основе теоремы 9.3. Этот метод был предложен в работе Маннора и Штольца [44].

В разделе 3.2 рассматривались задача универсального предсказания среднего значения p_i будущего исхода ω_i и соответствующее понятие калибруемости. В этом разделе будет рассматриваться задача универсального предсказания вероятностного распределения будущих исходов. В случае бинарного множества исходов $\{0,1\}$ обе эти задачи эквивалентны, так как вероятность единицы p_i равна среднему значению будущего исхода $\omega_i \in \{0,1\}$.

Будем предполагать, что исходы принадлежат конечному множеству $A = \{a_1, \ldots, a_m\}$. Обозначим $\mathcal{P}(A)$ – множество всех распределений вероятностей на множестве A. Каждое такое распределение (смешанная стратегия) есть вектор $\bar{p} = (p_1, \ldots, p_m)$ неотрицательных вещественных чисел сумма которых равна 1. На векторах–распределениях будет рассматриваться норма $\|\bar{p}\|_\infty = \max_{1 \leqslant i \leqslant m} |p_i|$. Можно также рассматривать эвклидову норму $\|\bar{p}\|_2$. Известно, что эти нормы эквивалентны в пространстве $\mathcal{R}^m$. В дальнейшем мы будем использовать обозначение $\|\bar{p}\|$ имея ввиду любую из этих норм.

Пусть $\bar{\delta}[a_i] = (0, \ldots, 1, \ldots, 0)$ обозначает вероятностное распределение, сосредоточенное на элементе a_i множества A. В этом векторе i-я координата равна 1, остальные координаты – нулевые.

Рассмотрим игру с полной информацией между *Предсказателем* и *Природой*. На каждом шаге t *Предсказатель* выдает вероятностное распределение $\bar{p}_t \in \mathcal{P}(A)$, после чего *Природа* выдает исход $a_t \in A$.

В терминах теории игр, $\bar{p}_t$ – смешанная стратегия *Предсказателя*, $\bar{\delta}[a_t]$ – чистая стратегия *Природы*.

Для выбора стратегий $\bar{p}_1, \bar{p}_2, \ldots$ *Предсказатель* будет применять рандомизированную стратегию, точнее, *Предсказатель* будет выдавать на каждом шаге t случайный вектор $\bar{p}_t \in \mathcal{P}(A)$, распределенный согласно некоторому вероятностному распределению $\bar{P}_t \in \mathcal{P}(\mathcal{P}(A))$.

Каждый игрок может использовать всю информацию, известную до его действия. На стратегию *Природы* не накладывается никаких ограничений.

По теореме Ионеско-Тульчи [4] вероятностные меры P_t порождают общее распределение P на траекториях $\bar{p}_1, \bar{p}_2, \ldots$.

Пусть задано число $\epsilon > 0$. Цель *Предсказателя* выдавать рандомизированные прогнозы $\bar{p}_t$ распределенные по мере P так, чтобы для любого $\bar{p} \in \mathcal{P}(A)$ для произвольной стратегии $a_1, a_2, \ldots$ *Природы* почти всюду было выполнено условие ϵ-калибруемости:

$$\limsup_{T\to\infty} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon} (\bar{\delta}[a_t] - \bar{p}_t) \right\| \leqslant \epsilon, \tag{9.18}$$

где векторы $\bar{p}_1, \bar{p}_2, \ldots$ распределены по мере P и

$$I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon} = \begin{cases} 1, \text{ если } \|\bar{p}_t - \bar{p}\| \leqslant \epsilon, \\ 0, \text{ в противном случае.} \end{cases}$$

Предсказатель будет выбирать свои прогнозы $\bar{p}_t$ из некоторого фиксированного конечного множества стратегий $\mathcal{P}_\epsilon$. Для задания этого множества построим какую-нибудь ϵ-сеть $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$ в множестве всех векторов $\mathcal{P}(A)$. Таким образом, для любого вектора $\bar{p} \in \mathcal{P}(A)$ найдется элемент ϵ-сети $\bar{s}_i \in \mathcal{P}_\epsilon$ такой, что $\|\bar{p} - \bar{s}_i\| < \epsilon$.

Мы будем строить вероятностные распределения $P_t \in \mathcal{P}(\mathcal{P}(A))$, сконцентрированные на конечном множестве $\mathcal{P}_\epsilon$. Для простоты мы отождествляем конечное множество $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$ и множество индексов его элементов $\mathcal{I} = \{1, 2, \ldots, N\}$, а также будем рассматривать на каждом шаге t вероятностные распределение P_t на $\mathcal{I}$.

Пусть P – общее распределение на траекториях $i_1, i_2, \ldots$ номеров элементов множества $\mathcal{P}_\epsilon$ порожденное распределениями P_t. Тогда условие (9.18) очевидным образом следует из следующего условия: почти всюду выполнено условие

$$\limsup_{T\to\infty} \sum_{k=1}^{N} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\{i_t = k\}} (\bar{\delta}[a_t] - \bar{s}_k) \right\| \leqslant \epsilon. \tag{9.19}$$

Существование ϵ-калибруемой стратегии в общей форме утверждается в следующей теореме.

Теорема 9.4. *Для произвольного $\epsilon > 0$ можно построить последовательность вероятностных распределений P_t, $t = 1, 2, \ldots$, такую, что P-почти всюду выполнено условие ϵ-калибруемости (9.18), где P – вероятностное распределение на траекториях $\bar{p}_1, \bar{p}_2, \ldots$ порожденное последовательностью P_t.*[2]

Доказательство. Мы применим теорему 9.3, в которой первый игрок – это *Предсказатель* с множеством стратегий [3] $\mathcal{I} = \{1, 2, \ldots, N\}$, а второй игрок – *Природа* с множеством стратегий $\mathcal{J} = A$. Функция выигрыша принимает в качестве значений векторы размерности $N|A|$:

$$\bar{f}(k, a) = \begin{pmatrix} \bar{0} \\ \ldots \\ \bar{0} \\ \bar{\delta}[a] - \bar{s}_k \\ \bar{0} \\ \ldots \\ \bar{0} \end{pmatrix}.$$

где $k \in \mathcal{I}$ и $a \in \mathcal{J}$, $\bar{0}$ – m-мерный нулевой вектор, $m = |A|$, а также $\bar{\delta}[a] - \bar{p}_k$ – разность двух векторов – столбцов размерности m, которая занимает k-ю компоненту сложного вектора.

Определим теперь выпуклое множество в пространстве $\mathcal{R}^{mN}$. Мы записываем векторы пространства $\mathcal{R}^{mN}$ как сложные векторы размерности N с вектор-компонентами в $\mathcal{R}^m : \bar{X} = (\bar{x}_1, \ldots, \bar{x}_N)$, где $\bar{x}_i \in \mathcal{R}^m$. Определим замкнутое выпуклое множество

$$C = \left\{ \bar{X} : \sum_{k=1}^{N} \|\bar{x}_k\| \leqslant \epsilon \right\}.$$

По теореме 9.3 замкнутое выпуклое подмножество C достижимо тогда и только тогда, когда для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что $\bar{f}(\bar{p}, \bar{q}) \in C$.

[2] Условие (9.18) эквивалентно условию (9.19).

[3] Мы отождествляем $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$ и множество индексов $\mathcal{I} = \{1, 2, \ldots, N\}$.

Условие теоремы 9.3 о достижимости выполнено для множества C, так как для любой смешанной стратегии $\bar{q} \in \mathcal{P}(\mathcal{J}) = \mathcal{P}(A)$ второго игрока, найдется точка $\bar{s}_k$ из $\mathcal{P}_\epsilon$ такая, что $\|\bar{q} - \bar{s}_k\| \leqslant \epsilon$, т.е. $\bar{f}(k, \bar{q}) \in C$. В этом случае мы берем в теореме 9.3 в качестве $\bar{p}$ распределение $\bar{\delta}[k]$ на $\mathcal{I} = \{1, \dots N\}$, сосредоточенное на числе k, где $1 \leqslant k \leqslant N$. В этом случае $\bar{f}(\bar{p}, \bar{q}) \in C$.

По теореме 9.2 можно построить рандомизированную стратегию *Предсказателя* $P_t \in \mathcal{P}(\mathcal{I})$ такую, что как бы Природа не выбирала последовательность $a_1, a_2, \dots$ последовательность векторнозначных выигрышей

$$\frac{1}{T}\sum_{t=1}^{T} \bar{f}(i_t, a_t) = \begin{pmatrix} \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=1\}}(\bar{\delta}[a_t] - \bar{s}_1) \\ \dots \\ \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=N\}}(\bar{\delta}[a_t] - \bar{s}_N) \end{pmatrix}.$$

почти всюду сходится к множеству C, где траектория $i_1, i_2, \dots$ распределена по мере P порожденной последовательностью рандомизированных стратегий P_t.

Таким образом, условие калибруемости (9.19) выполнено почти всюду. Теорема доказана $\triangle$

Последовательность предсказаний называется (хорошо) калибруемой на последовательности исходов, если она является ϵ-калибруемой для любого $\epsilon > 0$.

Предсказания, которые выбираются из конечного множества $\mathcal{P}_\epsilon = \{\bar{s}_1, \ \dots, \ \bar{s}_{N_\epsilon}\}$ и удовлетворяют условию (9.19), называются ϵ-калибруемыми предсказаниями.

Можно усилить теорему 9.4 и добиться калибруемости предсказаний.

Теорема 9.5. *Можно построить рандомизированную стратегию Предсказателя P_t такую, что для любого $\bar{p} \in \mathcal{P}(A)$ и любого $\epsilon > 0$ условие калибруемости*

$$\lim_{T\to\infty} \left\| \frac{1}{T}\sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon}(\bar{\delta}[a_t] - \bar{p}_t) \right\| = 0 \tag{9.20}$$

выполнено почти всюду.

Пусть ϵ_i – строго убывающая последовательность положительных вещественных чисел такая, что $\epsilon_i \to 0$ при $i \to \infty$. Для построения необходимой последовательности предсказаний надо разделить шаги конструкции на достаточно большие по размеру интервалы – "эпохи", в каждой из которых в качестве предсказания выбираются элементы соответствующей ϵ_i-сети в множестве $\mathcal{P}(\mathcal{P}(A))$. Эти предсказания ϵ_i-калибруются на верхней границе i-й эпохе и ϵ_{i-1}-калибруются на всей i-й эпохе. Эта идея была реализована при доказательстве теоремы 3.7 в разделе 3.5.1.

С помощью этой же конструкции можно усилить (9.19) до утверждения:

Теорема 9.6. *Можно построить такую рандомизированную стратегию, что*

$$\lim_{T\to\infty} \sum_{\bar{p}\in\mathcal{P}(A)} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\{\bar{p}_t=\bar{p}\}} (\bar{\delta}[a_t] - \bar{p}) \right\| = 0 \tag{9.21}$$

выполнено почти всюду.

Заметим, что во внешней сумме из (9.21) только конечное число слагаемых отлично от нуля: суммирование происходит только по $\bar{p} \in \{\bar{p}_t : 1 \leqslant t \leqslant T\}$.

Мы не будем останавливаться на деталях соответствующей конструкции.

9.4. Калибруемые предсказания и коррелированное равновесие

В этом разделе мы покажем, что если в некоторой неограниченно повторяющейся игре все игроки используют предсказания будущих ходов оппонентов, которые калибруются на последовательностях стратегий, выбранных в игре этими оппонентами, и выбирают "оптимальный ответ" на эти предсказания, то совместное частотное распределение стратегий игроков сходится к множеству коррелированных равновесий игры.

Каждое распределение вероятностей на конечном множестве A мощности N есть N-мерный вектор: обозначим такое распределение $\bar{p}$. Для любого $a \in A$ посредством $p(a)$ обозначаем a-ю координату $\bar{p}$. Другими словами, $p(a)$ – вероятность приписываемая элементу $a \in A$.

В качестве нормы на таких распределениях $\bar{p}$ можно рассматривать одну их эквивалентных норм $\|\bar{p}\|$ на $\mathcal{R}^N$ (эвклидову или максимум) и соответствующее расстояние $d(\bar{p}, \bar{q}) = \|\bar{p} - \bar{q}\|$. Расстояние от элемента $\bar{p} \in \mathcal{R}^N$ до множества $S \subseteq \mathcal{R}^N$ определяется:

$$d(\bar{p}, S) = \inf_{\bar{q} \in S} d(\bar{p}, \bar{q}).$$

Бесконечная последовательность $\bar{p}_1, \bar{p}_2, \ldots$ сходится к множеству S, если

$$\lim_{t \to \infty} d(\bar{p}_t, S) = 0.$$

Рассмотрим произвольную игру k игроков, заданную в нормальной форме. Для каждого игрока i задано конечное множество его ходов (стратегий): $A_i = \{1, \ldots, N_i\}$, $i = 1, \ldots, k$. Кроме этого, для каждого игрока i задана функция его выигрыша $f^i(i_1, \ldots, i_k)$, где $i_s \in A_s$, $s = 1, \ldots k$, – ходы всех игроков.

Смешанная стратегия игрока s – это вероятностное распределение на множестве его ходов A_s. Мы также будем рассматривать смешанные стратегии групп игроков $s_1, \ldots, s_l$ – совместные вероятностные распределения на множествах их ходов $A_{s_1} \times \ldots A_{s_l}$.

Обозначим $A = \prod_{j=1}^{k} A_j$ и $A_{-i} = \prod_{j \neq i} A_j$. Пусть $\bar{p}^t_{-i}$ – произвольное распределение вероятностей на множестве ходов всех игроков, кроме i, – их совместная смешанная стратегия. Здесь нижний индекс подчеркивает, что $\bar{p}_{-i} \in \mathcal{P}(A_{-i})$.

Будем также использовать обозначения:

$$f^i(a, \bar{p}_{-i}) = E_{\bar{p}_{-i}}(f^i(a, \cdot)) = \sum_{\bar{a}_{-i} \in A_{-i}} f^i(a, \bar{a}_{-i}) p_{-i}(\bar{a}_{-i}),$$

$$\bar{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_k),$$

$$(a, \bar{a}_{-i}) = (a_1, \ldots, a_{i-1}, a, a_{i+1}, \ldots, a_k),$$

где $a \in A_i$, $\bar{a}_{-i} \in A_{-i}$, E – символ математического ожидания относительно меры $\bar{p}_{-i}$.

FOR $t = 1, 2, \ldots$
Для каждого $i = 1, \ldots, k$, игрок i выдает предсказание набора будущих ходов своих оппонентов $j \neq i$ – распределение вероятностей $\bar{p}^t_{-i}$ (смешанная совместная стратегия этих игроков) и выбирает свою стратегию $a^t_i \in A_i$, при которой его выигрыш максимален, при условии, что его оппоненты будут придерживаются совместной смешанной стратегии $\bar{p}^t_{-i}$:

$$a^t_i \in \mathrm{argmax}_{a \in A_i} f^i(a, \bar{p}^t_{-i}). \tag{9.22}$$

ENDFOR

Рис. 9.1: Протокол игры

Пусть теперь игроки повторяют свою игру согласно протоколу, представленному на рис. 9.1. Мы называем любую стратегию a игрока i, на которой достигается максимум функции $f^i(a, \bar{p}^t_{-i})$, *оптимальным ответом* на предсказание $\bar{p}^t_{-i}$ ходов остальных игроков. Если имеется несколько таких стратегий, то выбирается одна из них – a^t_i, согласно какому-либо заранее фиксированному правилу. Называем эту стратегию *выбранным оптимальным ответом.*

Пусть $\bar{a}^t = (a^t_1, \ldots, a^t_k)$ – набор ходов всех игроков на шаге t. Обозначим

$$\bar{p}_T = \frac{1}{T} \sum_{t=1}^{T} \bar{\delta}[\bar{a}^t] \tag{9.23}$$

- эмпирическое частотное распределение наборов стратегий, выбранных всеми игроками за T шагов игры. Здесь $\bar{\delta}[\bar{a}]$ есть вектор размерности $\prod_{i=1}^k |A_i|$, в котором одна координата, соответствующая набору $\bar{a}$, равна 1, а все остальные его координаты равны 0.

Координатами вектора $\bar{p}_T$ являются частоты встречаемости каждого набора стратегий $\bar{a} = (a_1, \ldots, a_k)$ в последовательности наборов $\bar{a}^t = (a^t_1, \ldots, a^t_k)$, выбранных игроками на шагах $t = 1, \ldots, T$ игры. Размерность вектора $\bar{p}_T$, как и вектора $\bar{\delta}[\bar{a}^t]$, равна общему числу наборов $(a^t_1, \ldots, a^t_k)$, а именно, $\prod_{i=1}^k |A_i|$.

Каждому набору стратегий $\bar{a} = (a_1, \dots, a_k)$ соответствует число

$$p_T(\bar{a}) = \frac{1}{T}|\{t : 1 \leqslant t \leqslant T, \bar{a}^t = \bar{a}\}| \tag{9.24}$$

– значение частотного распределения на наборе $\bar{a}$ (соответствующая координата вектора p_T).

Следующая теорема показывает, что если каждый игрок использует предсказания ходов остальных игроков, которые калибруются на наборах стратегий оппонентов в смысле (9.21), и выбирает оптимальный ответ (9.22) на эти предсказания, то совместное частотное распределение стратегий игроков сходится к множеству $\mathcal{C}$ коррелированных равновесий игры.

Теорема 9.7. *Если для каждого i предсказания $\bar{p}^1_{-i}, \bar{p}^2_{-i}, \dots$ калибруются на последовательности $\bar{a}^1_{-i}, \bar{a}^2_{-i}, \dots$ стратегий оппонентов i, то последовательность частотных эмпирических распределений $\bar{p}_T$, определенных по (9.23), сходится к множеству $\mathcal{C}$ коррелированных равновесий.*

Доказательство. Для доказательства теоремы нам надо показать, что

$$d(\bar{p}_T, \mathcal{C}) \to 0$$

при $T \to \infty$, где $\mathcal{C}$ – множество всех коррелированных равновесий игры. Мы также покажем, что $\mathcal{C} \neq \emptyset$.

Симплекс всех распределений вероятностей на многограннике $A = \prod_{i=1}^k A_i$ (векторов размерности $|A|$) является компактным множеством. Поэтому последовательность распределений $\{\bar{p}_T : T = 1, 2, \dots\}$, определенных по (9.23), содержит бесконечную сходящуюся подпоследовательность $\bar{p}_{T_j}$. Пусть $\bar{p}^*$ – предельная точка этой подпоследовательности. Мы докажем, что $\bar{p}^*$ является коррелированным равновесием.

Фиксируем i и ход $a \in A_i$ игрока i такие, что

$$p^*(a) = \sum_{\bar{a}: a_i = a} p^*(\bar{a}) > 0,$$

где $\bar{a} = (a_1, \dots\ , a_k)$, $a_s \in A_s$, $s = 1, \dots\ , k$. [4]

Обозначим $f = f^i$ и определим два подмножества (зависящие от i и a) $B, \tilde{B} \subseteq \mathcal{P}(A_{-i})$:

$$B = \{\bar{q}_{-i} : f(a, \bar{q}_{-i}) = \max_{a' \in A_i} f(a', \bar{q}_{-i})\}$$

– множество всех смешанных стратегий оппонентов игрока i, для которых его чистая стратегия a является оптимальным ответом. Легко видеть, что B – замкнутое выпуклое множество. Определим также

$$\tilde{B} = \left\{\bar{q}_{-i} : \exists t \left(\bar{q}_{-i} = \bar{q}^t_{-i})\& f(a, \bar{q}_{-i}) = \max_{a' \in A_i} f(a', \bar{q}_{-i})\right)\right\}$$

– множество всех смешанных стратегий, выбранных оппонентами игрока i на тех шагах $t = 1, 2, \dots$ игры, на которых он выбрал ход a в качестве оптимального ответа. Из определения $\tilde{B} \subseteq B$.

Из определения следует, что множество $\tilde{B}$ является не более чем счетным, так как на каждом шаге к нему добавляется не более одного элемента.

Рассмотрим условную вероятность произвольного вектора ходов $\bar{a}_{-i}$ всех игроков, кроме i, при известном $a_i = a$ (где a было выбрано выше) относительно предельного распределения $\bar{p}^*$:

$$p^*(\bar{a}_{-i}|a_i = a) = p^*((a, \bar{a}_{-i})|a_i = a) = \frac{p^*(a, \bar{a}_{-i})}{p^*(a_i = a)}. \qquad (9.25)$$

По следствию 7.3 распределение вероятностей p на множестве $\prod_{k=1}^{K}\{1, \dots, N_k\}$ последовательностей стратегий $\bar{a} = (a_1, \dots, a_K)$ является коррелированным равновесием тогда и только тогда, когда для каждого игрока $i \in \{1, \dots, K\}$ и любой стратегии $a \in A_i$ выполнено

$$f^i(a, \bar{p}(\cdot|a_i = a)) = \max_{a' \in A_i} f^i(a', \bar{p}(\cdot|a_i = a)).$$

[4] Если $p^*(a) = 0$, то то ход a можно не учитывать при подсчете частотного распределения; это эквивалентно случаю, когда i-й игрок не использует a. В этом случае a можно игнорировать.

Отсюда следует, что вероятностное распределение $\bar{p}^*$ является коррелированным равновесием тогда и только тогда, когда условное распределение $\bar{p}^*(\cdot|a_i = a) \in B$ для всех i и $a \in A_i$.

Вернемся к фиксированным числу i и элементу $a \in A_i$. Мы докажем, что $\bar{p}^*(\cdot|a_i = a) \in B$ рассматривая приближение к нему с помощью соответствующего частотного распределения.

Пусть

$$N_T(a) = |\{t : 1 \leqslant t \leqslant T,\ a_i^t = a\}|$$

– число шагов $\leqslant T$, на которых игрок i выбирает стратегию a,

$$N_T(\bar{p}_{-i}) = |\{t : 1 \leqslant t \leqslant T,\ \bar{p}_{-i}^t = \bar{p}_{-i}\}|$$

– число шагов $\leqslant T$, на которых оппоненты игрока i выбирают набор смешанных стратегий $\bar{p}_{-i} \in \mathcal{P}(A_{-i})$.

Определим соответствующее распределению $\bar{p}_T$ условное частотное распределение $\bar{p}_T(\cdot|a_i = a)$ стратегий, выбранных всеми игроками кроме i, а именно, рассмотрим условную вероятность $\bar{a}_{-i}$ при известном $a_i = a$ относительно распределения p_T :

$$p_T(\bar{a}_{-i}|a_i = a) = \frac{p_T(a, \bar{a}_{-i})}{p_T(a_i = a)}. \tag{9.26}$$

Согласно (9.25) $p_{T_j}(a_{-i}|a_i = a) \to p^*(a_{-i}|a_i = a)$ при $j \to \infty$.

По определению множества $\tilde{B}$ элемент $a \in A_i$ появляется в наборе стратегий $\bar{a}^t$ в качестве i-й координаты только при $\bar{p}_{-i}^t \in \tilde{B}$. Отсюда следует, что частота встречаемости любого набора $(a, \bar{a}_{-i})$ в последовательности

$$\{\bar{a}^t :\ 1 \leqslant t \leqslant T\}$$

равна частоте встречаемости набора $\bar{a}_{-i}$ в последовательности

$$\{\bar{a}_{-i}^t : \bar{p}_{-i}^t \in \tilde{B},\ 1 \leqslant t \leqslant T\}.$$

Поэтому по (9.24) получаем:

$$p_T(a, \bar{a}_{-i}) = p_T(\bar{a}) = \frac{1}{T}|\{t : 1 \leqslant t \leqslant T,\ \bar{p}_{-i}^t \in \tilde{B},\ \bar{a}_{-i}^t = \bar{a}_{-i}\}|.$$

По определению $p_T(a_i = a) = \frac{N_T(a)}{T}$. Отсюда получаем выражения для условного частотного распределения, образованного последовательностью $\bar{a}^t$, где $a_i^t = a$, $t = 1, \ldots, T$:

$$\bar{p}_T(\cdot|a_i = a) = \frac{1}{N_T(a)} \sum_{1 \leqslant t \leqslant T,\ \bar{p}_{-i}^t \in \tilde{B}} \bar{\delta}[\bar{a}_{-i}^t] =$$

$$= \left(\frac{T}{N_T(a)}\right) \frac{1}{T} \sum_{\bar{p}_{-i} \in \tilde{B}} \sum_{1 \leqslant t \leqslant T} I_{\{\bar{p}_{-i}^t = \bar{p}_{-i}\}} (\bar{\delta}[\bar{a}_{-i}^t] - \bar{p}_{-i}) + \qquad (9.27)$$

$$+ \sum_{\bar{p}_{-i} \in \tilde{B}} \left(\frac{N_T(\bar{p}_{-i})}{N_T(a)}\right) \bar{p}_{-i}. \qquad (9.28)$$

Так как $p^*(a_i = a) > 0$ и $\bar{p}^*$ есть предел распределений $\bar{p}_{T_j}$ при $j \to \infty$, поэтому множитель $\left(\frac{T_j}{N_{T_j}(a)}\right)$ ограничен сверху.

По определению

$$N_T(a) = \sum_{\bar{p}_{-i} \in \tilde{B}} N_T(\bar{p}_{-i}).$$

Таким образом, по (9.21) сумма (9.27) стремится к нулю и поэтому распределение $\bar{p}_{T_j}(\cdot|a_i = a)$ сходится при $T_j \to \infty$ к множеству выпуклых комбинаций элементов из множества $\tilde{B}$, которое в свою очередь является подмножеством выпуклого замкнутого множества B. Отсюда

$$d(\bar{p}_{T_j}(\cdot|a_i = a), B) \to 0$$

при $T_j \to \infty$.

Из сходимости $\bar{p}_{T_j} \to \bar{p}^*$ при $j \to \infty$ следует, что для произвольного вектора $\bar{a}_{-i}$ будет $p_{T_j}(\bar{a}_{-i}|a_i = a) \to p^*(\bar{a}_{-i}|a_i = a)$ при $T_j \to \infty$. Отсюда и из замкнутости множества B следует, что $\bar{p}^*(\cdot|a_i = a) \in B$ для всех i и $a \in A_i$. Таким образом, мы доказали, что вероятностное распределение $\bar{p}^*$ является коррелированным равновесием. Отсюда следует утверждение теоремы. $\triangle$

Из теорем 9.4 и 9.7 получаем следующее следствие.

Следствие 9.1. *Можно построить рандомизированный алгоритм, который для любого* $1 \leqslant i \leqslant N$ *вычисляет последовательность предсказаний* $\bar{p}^1_{-i}, \bar{p}^2_{-i}, \ldots$ *ходов оппонентов игрока* i*, так что выполнено следующее:*

- *Пусть каждый игрок выбирает выбирает в качестве своего хода наилучший ответ на предсказание этого алгоритма.*
- *Тогда эмпирические частоты* $\bar{p}_T$ *ходов всех игроков сходятся к множеству* $\mathcal{C}$ *коррелированных равновесий игры при* $T \to \infty$ *с вероятностью 1.*

9.5. Задачи и упражнения

1. Доказать неравенство (9.8).

2. Доказать, что минимаксная теорема является следствием теоремы Блекуэлла о достижимости.

3. Докажите, что если замкнутое множество S достижимо в игре с матрицей $f(i,j)$, то любое замкнутое подмножество его дополнения не достижимо в игре с матрицей $f'(i,j) = f(j,i)$.

9.6. Генеративно-состязательные сети (GANs)

В этом разделе мы рассмотрим повторяющуюся игру с использованием выпуклой оптимизации для построения генеративно–состязательной нейронной сети (Generative Adversarial Network (GAN)). Такая нейронная сеть представляет собой комбинацию двух нейронных сетей, одна из которых – генератор G, генерирует образцы, а другая – дискриминатор h, должна отличать правильные ("подлинные") образцы (из некоторой базы данных)[5] от искусственно сгенерированных генератором G. Так как сети G и h имеют противоположные цели – генерировать образцы (неотличимые от подлинных) и отбраковать искусственные образцы, между ними возникает антагонистическая игра. Генеративно-состязательную сети были впервые предложены в [32]. Излагаемые ниже результаты опубликованы в [33].

[5]Во многих примерах образцы это изображения (фотографии).

Приводим математическое описание этой игры. Задано вероятностное распределение p_{data} на множестве всех образцов. Образцы представлены многомерными векторами.

Пусть X и Y – выпуклые подмножества $\mathcal{R}^k$ и $\mathcal{R}^m$, $G_{\mathbf{x}}$ – функция генератор, где $\mathbf{x} \in X$ – вектор его параметров. Соответствующее моделируемое распределение $p_{\mathbf{x}}$ на на множестве всех образцов порождается генератором из равномерного (или гауссовского) распределения λ: $p_{\mathbf{x}}(A) = \lambda(G_{\mathbf{x}}^{-1}(A)$. Пусть $h_{\mathbf{y}}$ – дискриминатор – функция, определенная на множестве образцов и принимающая значения в интервале $[0, 1]$, $\mathbf{y} \in Y$ – вектор ее параметров. Генератор и дискриминатор реализуются в виде нейронных сетей определенной архитектуры.

Значение $h_{\mathbf{y}}(\mathbf{u})$ можно понимать как вероятность того, что образец $\mathbf{u}$ сгенерирован распределением p_{data}. Мы обучаем нейронную сеть $h_{\mathbf{y}}$ так, чтобы максимизировать вероятность правильной классификации образцов. Одновременно, мы обучаем нейронную сеть $G_{\mathbf{x}}$ так, чтобы минимизировать среднее значение $\log(1 - h_{\mathbf{y}}(G_{\mathbf{x}}(\mathbf{z})))$ при $\mathbf{z} \sim \lambda$. Между генератором и дискриминатором возникает антагонистическая игра с некоторой платежной функцией $M : \mathcal{R}^k \times \mathcal{R}^m \to \mathcal{R}$.

В [32] рассматривается платежная функция

$$M(\mathbf{x}, \mathbf{y}) = E_{\mathbf{u} \sim p_{data}}[\log h_{\mathbf{y}}(\mathbf{u})] + E_{\mathbf{z} \sim \lambda}[\log(1 - h_{\mathbf{y}}(G_{\mathbf{x}}(\mathbf{z}))]. \quad (9.29)$$

Для построения оптимальных сетей решается задача поиска седловой точки функции $M(\mathbf{x}, \mathbf{y})$ в смешанных стратегиях. Это означает, что будем искать распределения вероятностей P^1 и P^2 такие, что

$$E_{(\mathbf{x},\mathbf{y}) \sim P^1 \times P^2}[M(\mathbf{x}, \mathbf{y})] \geqslant \max_{\mathbf{y} \in Y} E_{\mathbf{x} \sim P^1}[M(\mathbf{x}, \mathbf{y})]. \quad (9.30)$$

$$E_{(\mathbf{x},\mathbf{y}) \sim P^1 \times P_T^2}[M(\mathbf{x}, \mathbf{y})] \leqslant \min_{\mathbf{x} \in X} E_{\mathbf{y} \sim P^2}[M(\mathbf{x}, \mathbf{y})]. \quad (9.31)$$

Далее будет указан алгоритм для приближенного решения этой задачи: неравенства (9.30) и (9.31) будут выполнены с некоторой (как угодно большой) точностью, а распределения вероятностей P^1 и P^2 будут сконцентрированы на конечных подмножествах X и Y.

Соответствующие нейронные сети могут быть построены если функция $M(\mathbf{x}, \mathbf{y})$ является вогнутой по $\mathbf{y}$. Предполагаем также, что $|M(\mathbf{x}, \mathbf{y})| \leqslant C$ для всех $\mathbf{x}$ и $\mathbf{y}$.

Покажем, что условие вогнутости по $\mathbf{y}$ для функции $M(\mathbf{x}, \mathbf{y})$, определенной согласно (9.29), выполнено, если дискриминатор $h_{\mathbf{y}}(\mathbf{u})$ реализован в виде нейронной сети с одним внутренним слоем и пороговой функцией $h_{\mathbf{y}}(\mathbf{u}) = \frac{1}{1+\exp(-(\mathbf{y}\cdot\mathbf{u}))}$.

Легко проверить, что функция $\log h_{\mathbf{y}}(\mathbf{u}) = \log \frac{1}{1+\exp(-(\mathbf{y}\cdot\mathbf{u}))}$ является вогнутой по $\mathbf{y}$ для каждого фиксированного $\mathbf{u}$.[6] Отсюда $E_{\mathbf{u}\sim p_{data}}[\log h_{\mathbf{y}}(\mathbf{u})]$ также вогнутая по $\mathbf{y}$. Аналогичным образом доказывается, что функция $E_{\mathbf{z}\sim\lambda}[\log(1 - h_{\mathbf{y}}(G_{\mathbf{x}}(\mathbf{z}))]$ является вогнутой по $\mathbf{y}$. Для каждого $\mathbf{x}$ платежная функция $M(\mathbf{x}, \mathbf{y})$ является вогнутой по $\mathbf{y}$ как сумма двух вогнутых функций.

Для поиска седловой точки антагонистической игры между генератором и дискриминатором мы используем технику онлайн выпуклой оптимизации. Будем рассуждать в терминах выпуклой по $\mathbf{y}$ функции потерь $f(\mathbf{x}, \mathbf{y}) = -M(\mathbf{x}, \mathbf{y})$ и будем решать задачу поиска распределений P^1 и P^2 таких, что (с некоторым приближением) выполнены неравенства

$$E_{(\mathbf{x},\mathbf{y})\sim P^1\times P^2}[f(\mathbf{x}, \mathbf{y})] \leqslant \min_{\mathbf{y}\in Y} E_{\mathbf{x}\sim P^1}[f(\mathbf{x}, \mathbf{y})]. \tag{9.32}$$

$$E_{(\mathbf{x},\mathbf{y})\sim P^1\times P^2_T}[f(\mathbf{x}, \mathbf{y})] \geqslant \max_{\mathbf{x}\in X} E_{\mathbf{y}\sim P^2}[f(\mathbf{x}, \mathbf{y})]. \tag{9.33}$$

Сформулируем основные допущения. Предполагаем, что существуют константы C, B и D такие, что

- $|f(\mathbf{x}, \mathbf{y})| \leqslant C$ для всех $\mathbf{x}$ и $\mathbf{y}$,
- $\|y\| \leqslant B$ для всех $\mathbf{y} \in Y$,
- для любого $\mathbf{x}$ функция $g(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$ выпуклая по $\mathbf{y}$
- для произвольного $\mathbf{x} \in X$ функция $M(\mathbf{x}, \mathbf{y})$ (а значит и $f(\mathbf{x}, \mathbf{y})$) удовлетворяет условию Липшица по $\mathbf{y}$ с константой D. Тогда, согласно предложению 6.4, для любого $\mathbf{x} \in X$

[6] Легко проверить, что одномерная функция $h(x) = \log \frac{1}{1+\exp(-x)}$ является вогнутой, а многомерная функция $h_{\mathbf{y}}(\mathbf{u})$ является композицией вогнутой и линейной функций.

субградиент функции $g(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$ равномерно ограничен: $\|\nabla g(\mathbf{y})\| \leqslant D$ для всех $\mathbf{y} \in Y$.

- полагаем $\eta = \frac{B}{2D\sqrt{2T}}$.

Свойство выпуклости функций позволяет при оценке регрета алгоритм FTL заменить эти функции на их линеаризации. Пусть задана последовательность выпуклых функций $g_1(\mathbf{y}), \ldots, g_T(\mathbf{y})$ и последовательность аргументов $\mathbf{y}_1, \ldots, \mathbf{y}_T \in Y$. Пусть

$$\mathbf{y}^* = \operatorname{argmin}_{\mathbf{y} \in Y} \sum_{t=1}^{T} g_t(\mathbf{y}).$$

Так как для любой точки $\mathbf{y}$ выполнено неравенство

$$g_t(\mathbf{y}^*) - g_t(\mathbf{y}) \geqslant \nabla g_t(\mathbf{y})(\mathbf{y}^* - \mathbf{y}),$$

для каждого t будет

$$g_t(\mathbf{y}_t) - g_t(\mathbf{y}^*) \leqslant (\nabla g_t(\mathbf{y}_t) \cdot \mathbf{y}_t) - (\nabla g_t(\mathbf{y}_t) \cdot \mathbf{y}^*).$$

Тогда регрет относительно решения $\mathbf{y}^*$ для последовательности функций g_t, $t = 1, \ldots, T$ не превосходит регрета относительно их линериализаций:

$$R_T = \sum_{t=1}^{T} g_t(\mathbf{y}_t) - g_t(\mathbf{y}^*) \leqslant$$
$$\sum_{t=1}^{T} (\nabla g_t(\mathbf{y}_t) \cdot \mathbf{y}_t) - (\nabla g_t(\mathbf{y}_t) \cdot \mathbf{y}^*).$$

Мы воспользуемся этим свойством в следующем алгоритме, приведенном на рис. 9.2. Этот алгоритм построен по схеме рассуждений при доказательстве минимаксной теоремы 7.2, при этом не предполагается, что множества ходом игроков конечные.

Будем последовательно по очереди применять алгоритмы FTL и алгоритм онлайн градиентного спуска, приведенные в разделах 6.1 и 6.2.

В следующих двух леммах 9.1 и 9.2 приводятся верхние оценки регрета алгоритмов онлайн градиентного спуска (рис.9.35) b FTL.

Алгоритм поиска седловой точки

FOR $t = 1, \ldots, T$

- Обозначаем $f_t(\mathbf{x}) = f(\mathbf{x}, \mathbf{y}_t)$ и $g_t(\mathbf{y}) = f(\mathbf{x}_t, \mathbf{y})$. Функции $f_t(\mathbf{x})$ и $g_t(\mathbf{y})$ при $1 \leqslant t \leqslant t-1$ определены на прошлых шагах.

- Вычисляем согласно алгоритму FTL:

$$\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x} \in X} \sum_{i+1}^{t-1} f_i(\mathbf{x}). \tag{9.34}$$

  Здесь мы предполагаем, что максимум в (9.34 достигается в некоторой точке $\mathbf{x} \in X$. Это так, если предположить, что X – компактное подмножество $\mathcal{R}^k$. Ввиду отсутствия свойства вогнутости $f(\mathbf{x}, \mathbf{y})$ по $\mathbf{x}$, поиск такого максимума может оказаться нетривиальной задачей. Можно также искать приближенный максимум как в теореме 7.2, например, с точностью до $\frac{1}{T}$. В этом случае предложение 6.1 будет верно с точностью до константы.

- Вычисляем согласно алгоритму онлайн градиентного спуска (рис. 6.2) из раздела 6.2:

$$\mathbf{y}_t = P_Y(\mathbf{y}_{t-1} - \eta \nabla g_i(\mathbf{y}_{t-1})). \tag{9.35}$$

ENDFOR

Рис. 9.2: Алгоритм поиска седловой точки

Лемма 9.1. *Регрет алгоритма онлайн градиентного спуска имеет оценку*

$$R_T^{FTRL} = \sum_{t=1}^{T} f(\mathbf{x}_t, \mathbf{y}_t) - \min_{\mathbf{y}\in Y} \sum_{t=1}^{T} f(\mathbf{x}_t, \mathbf{y}) \leqslant R_T^2, \tag{9.36}$$

где $R_T^2 = \frac{1}{2}\left(\frac{(2B)^2}{2\eta} + D^2\eta T\right) = \frac{33}{8} BD\sqrt{2T}$.

Эта утверждение является частным случаем теоремы 6.2.

Лемма 9.2. *Регрет алгоритма FTL имеет оценку*

$$\max_{\mathbf{x}\in X} \sum_{t=1}^{T} f(\mathbf{x}, \mathbf{y}_t) - \sum_{t=1}^{T} f(\mathbf{x}_t, \mathbf{y}_t) \leqslant R_T^1,$$

где $R_T^1 = \eta D^2 T + C = BD\sqrt{\frac{1}{8}T} + C = O(\sqrt{T})$.

Доказательство. По лемме 6.2 (которая применено к функции $-f_t$) регрет алгоритма FTL удовлетворяет неравенству

$$\max_{\mathbf{x}\in X} \sum_{t=1}^{T} f(\mathbf{x}, \mathbf{y}_t) - \sum_{t=1}^{T} f(\mathbf{x}_t, \mathbf{y}_t) \leqslant \tag{9.37}$$

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}). \tag{9.38}$$

Продолжим это неравенство

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}) =$$

$$\sum_{t=1}^{T-1} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1} + f_{t+1}(\mathbf{x}_{t+1}) - f_{t+1}(\mathbf{x}_{t+1}) +$$

$$f_T(\mathbf{x}_T) - f_T(\mathbf{x}_{T+1}) =$$

$$\sum_{t=1}^{T-1} f_{t+1}(\mathbf{x}_{t+1}) - f_t(\mathbf{x}_{t+1}) + \sum_{t=1}^{T-1} f_t(\mathbf{x}_t) - f_{t+1}(\mathbf{x}_{t+1}) +$$

$$f_T(\mathbf{x}_T) - f_T(\mathbf{x}_{T+1}) =$$

$$\sum_{t=1}^{T-1} f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - f(\mathbf{x}_{t+1}, \mathbf{y}_t) + f_1(\mathbf{x}_1) - f_T(\mathbf{x}_{T+1}) \leqslant \tag{9.39}$$

$$\sum_{t=1}^{T-1} D\|y_{t+1} - y_t\| + f_1(\mathbf{x}_1) - f_T(\mathbf{x}_{T+1}) \leqslant \tag{9.40}$$

$$\eta D^2 T + 2C = BD\sqrt{\frac{1}{8}T} + 2C = O(\sqrt{T}). \tag{9.41}$$

Здесь при переходе от (9.39) к (9.40) мы использовали условие Липшица для функции $f(\mathbf{x}_t, \mathbf{y})$ по $\mathbf{y}$. При переходе от (9.40) к (9.41) мы использовали лемму 6.2. $\triangle$

Теорема 9.8. *Пусть P_1 – равномерное распределение на $\{\mathbf{x}_1, \dots \mathbf{x}_T\}$, P_2 – равномерное распределение на $\{\mathbf{y}_1, \dots \mathbf{y}_T\}$. Тогда для любого T смешанная стратегия $P_T^1 \times P_T^2$ является приближенной седловой точкой функции $M(x, y)$:*

$$E_{(\mathbf{x},\mathbf{y})\sim P_T^1 \times P_T^2}[M(\mathbf{x}, \mathbf{y})] \geqslant \max_{\mathbf{y}\in Y} E_{\mathbf{x}\sim P_T^1}[M(\mathbf{x}, \mathbf{y})] - \frac{R_T^1 + R_T^2}{T}.$$

$$E_{(\mathbf{x},\mathbf{y})\sim P_T^1 \times P_T^2}[f(\mathbf{x}, \mathbf{y})] \leqslant \min_{\mathbf{x}\in X} E_{\mathbf{y}\sim P_T^2}[M(\mathbf{x}, \mathbf{y})] + \frac{R_T^1 + R_T^2}{T}, . \tag{9.42}$$

где R_T^1 и R_T^2 такие, как в леммах 9.1 и 9.2, причем $R_T^1 + R_T^2 = O(\sqrt{T})$.

Доказательство. По леммам 9.1 и 9.2 имеем

$$\max_{\mathbf{x}\in X}\sum_{t=1}^{T} f(\mathbf{x},\mathbf{y}_t) - \sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}_t) \leqslant R_T^1, \tag{9.43}$$

$$\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}_t) - \min_{\mathbf{y}\in Y}\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}) \leqslant R_T^2. \tag{9.44}$$

Имеет место неравенство между максимумом и математическим ожиданием

$$\max_{\mathbf{x}\in X} f(\mathbf{x},\mathbf{y}_t) \geqslant E_{\mathbf{x}\sim P_T^1}\left[\sum_{t=1}^{T} f(\mathbf{x},\mathbf{y}_t)\right].$$

Вместе с (9.43) это влечет

$$E_{\mathbf{x}\sim P_T^1}\left[\sum_{t=1}^{T} f(\mathbf{x},\mathbf{y}_t)\right] - \sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}_t) \leqslant R_T^1. \tag{9.45}$$

Складываем (9.45) с (9.44), делим на T и получаем

$$E_{\mathbf{x}\sim P_T^1}\left[\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x},\mathbf{y}_t)\right] - \min_{\mathbf{y}\in Y}\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t,\mathbf{y}) \leqslant \frac{R_T^1+R_T^2}{T}. \tag{9.46}$$

Из определения неравенство (9.46) можно переписать в виде

$$E_{(\mathbf{x},\mathbf{y})\sim P_T^1\times P_T^2}[f(\mathbf{x},\mathbf{y})] \leqslant \min_{\mathbf{y}\in Y} E_{\mathbf{x}\sim P_T^1}[f(\mathbf{x},\mathbf{y})] + \frac{R_T^1+R_T^2}{T}. \tag{9.47}$$

Аналогичным образом получим неравенство

$$E_{(\mathbf{x},\mathbf{y})\sim P_T^1\times P_T^2}[f(\mathbf{x},\mathbf{y})] \geqslant \max_{\mathbf{x}\in X} E_{\mathbf{y}\sim P_T^2}[f(\mathbf{x},\mathbf{y})] - \frac{R_T^1+R_T^2}{T}. \tag{9.48}$$

Неравенства (9.47) и (9.47) означают, что для произвольного T смешанная стратегия $P_T^1\times P_T^2$ со степенью точности $\frac{R_T^1+R_T^2}{T}$ является седловой точкой $f(\mathbf{x},\mathbf{y}) = -M(\mathbf{x},\mathbf{y})$ функции $f(\mathbf{x},\mathbf{y})$.

Производим подстановку $f(\mathbf{x},\mathbf{y}) = -M(\mathbf{x},\mathbf{y})$ в (9.47) и (9.48) и получаем (9.42). $\triangle$

Часть IV

Вспомогательные утверждения

Глава 10

Некоторые замечательные неравенства

Приведем несколько замечательных неравенств, которые неоднократно используются в доказательствах теорем. Основным таким неравенством будет неравенство Хефдинга.

Лемма 10.1. *Пусть X – случайная величина и $a \leqslant X \leqslant b$. Тогда для произвольного $s \in \mathcal{R}$*

$$\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8}. \tag{10.1}$$

Доказательство. Так как

$$\ln E(e^{sX}) = sE(X) + \ln E(e^{s(X-E(X))}),$$

достаточно доказать, что для любой случайной величины X с $E(X) = 0$, $a \leqslant X \leqslant b$, будет

$$E(e^{sX}) \leqslant e^{s^2(b-a)^2/8}.$$

Из выпуклости экспоненты имеем

$$e^{sx} \leqslant \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

при $a \leqslant x \leqslant b$.

Обозначим $p = -\frac{a}{b-a}$. Так как $E(X) = 0$, то применяя математическое ожидание к обеим частям этого неравенства, получим при $x = X$:

$$E(e^{sX}) \leqslant -\frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} =$$
$$= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} = e^{\varphi(u)},$$

где $u = s(b-a)$, $\varphi(u) = -pu + \ln(1 - p + pe^u)$.

Производная $\varphi(u)$ по u равна

$$\varphi'(u) = -p + \frac{p}{p + (1-p)e^{-u}}.$$

Имеем $\varphi(0) = \varphi'(0) = 0$. Кроме того,

$$\varphi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leqslant \frac{1}{4}.$$

Действительно, обозначим $q = (1-p)e^{-u}$. Тогда надо доказать неравенство $\frac{pq}{(p+q)^2} \leqslant \frac{1}{4}$, которое следует из $(p-q)^2 \geqslant 0$.

По формуле Тейлора для некоторого $\theta \in [0, u]$ получаем

$$\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{u^2}{2}\varphi''(\theta) \leqslant \frac{u^2}{8} = \frac{s^2(b-a)^2}{8},$$

так как $u = s(b-a)$. Лемма доказана. $\triangle$

Рассмотрим несколько следствий, разъясняюших значение этого неравенства.

Следствие 10.1. *Пусть X – случайная величина такая, что $P\{a \leqslant X \leqslant b\} = 1$, $c > 0$. Тогда*

$$P\{|X - E(X)| > c\} \leqslant 2e^{-\frac{2c^2}{(b-a)^2}}. \tag{10.2}$$

Доказательство. Предварительно напомним неравенство Маркова. Пусть X – случайная величина, $X \geqslant 0$. Из

$$E(X) = \int XdP \geqslant \int_{\{X>c\}} XdP \geqslant cP\{X > c\}$$

следует, что $P\{X > c\} \leqslant E(X)/c$.

Используя это неравенство и неравенство (10.1), получим

$$P\{X - E(X) > c\} = P\{e^{s(X-E(X))} > e^{cs}\} \leqslant e^{-cs+\frac{s^2(b-a)^2}{8}}$$

для всех $s > 0$. Находим минимум правой части по s. Он достигается при $s = 4c/(b-a)^2$. Отсюда получаем

$$P\{X - E(X) > c\} \leqslant e^{-\frac{2c^2}{(b-a)^2}}.$$

Аналогично получаем

$$P\{X - E(X) < -c\} \leqslant e^{-\frac{2c^2}{(b-a)^2}}.$$

Окончательно получаем

$$P\{|X - E(X)| > c\} \leqslant 2e^{-\frac{2c^2}{(b-a)^2}}.$$

$\triangle$

Более известным является следующее следствие из этой леммы – неравенство Чернова. [1]

Следствие 10.2. *Пусть $X_1, X_2, \ldots$ – последовательность независимых случайных величин таких, что при всех $i = 1, 2, \ldots$ выполнено $P\{a_i \leqslant X \leqslant b_i\} = 1$. Тогда для любого $\epsilon > 0$:*

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i > \epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right), \qquad (10.3)$$

а также

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i < -\epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

[1] Для удобства иногда используем обозначение $\exp(x) = e^x$.

Доказательство. Доказательство аналогично доказательству следствия 10.1. Из неравенства Маркова и неравенства (10.1) получаем

$$P\left\{\sum_{i=1}^{n}(X_i - E(X_i)) > \epsilon\right\} \leqslant$$

$$\leqslant \frac{E\left(\exp(s\sum\limits_{i=1}^{n}(X_i - E(X_i)))\right)}{\exp(s\epsilon)} =$$

$$= \frac{\prod\limits_{i=1}^{n} E(\exp(s(X_i - E(X_i))))}{\exp(\epsilon s)} \leqslant$$

$$\leqslant \frac{\prod\limits_{i=1}^{n} \exp\left(\frac{s^2(b_i-a_i)^2}{8}\right)}{\exp(\epsilon s)} \leqslant$$

$$\leqslant \exp\left(-\epsilon s + \frac{s^2\sum\limits_{i=1}^{n}(b_i - a_i)^2}{8}\right) \leqslant$$

$$\leqslant \exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

При преходе от второй строки к третьей мы использовали независимость случайных величин $X_1, X_2, \ldots$.

При переходе от предпоследней строки к последней строке мы использовали минимизацию по s. Второе неравенство получается аналогичным образом. $\triangle$

Из этого следствия можно получить оценку скорости сходимости для закона больших чисел.

Следствие 10.3. *Пусть $X_1, X_2, \ldots$ – последовательность независимых случайных величин таких, что при всех $i = 1, 2, \ldots$*

выполнено $P\{a_i \leqslant X \leqslant b_i\} = 1$. *Тогда для любого* $\epsilon > 0$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-\frac{2n^2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

Если $a_i = 0$, $b_i = 1$ *для всех* i, *то*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2n\epsilon^2\right). \tag{10.4}$$

Последовательность случайных величин $V_1, V_2, \ldots$ называется мартингал-разностью относительно последовательности случайных величин $X_1, X_2, \ldots$, если для любого $i > 1$ величина V_i есть функция от $X_1, \ldots, X_i$ и

$$E(V_{i+1}|X_1, \ldots, X_i) = 0$$

с вероятностью 1. Следующее неравенство называется неравенством Хефдинга–Адзумы.

Лемма 10.2. *Пусть* $V_1, V_2, \ldots$ *– мартингал-разность относительно последовательности случайных величин* $X_1, X_2, \ldots$, *кроме этого,* $V_i \in [A_i, A_i + c_i]$ *для некоторой случайной величины* A_i, *измеримой относительно* $X_1, \ldots, X_{i-1}$, *и некоторой последовательности положительных констант* c_i. *Если* $S_n = \sum\limits_{i=1}^{n} V_i$, *то для любого* $s > 0$

$$E(e^{sS_n}) \leqslant e^{(s^2/8)\sum\limits_{i=1}^{n} c_i^2}.$$

Доказательство. Имеем

$$\begin{aligned} E(e^{sS_n}) = E(e^{sS_{n-1}}E(e^{sV_n}|X_1, \ldots, X_{n-1})) \leqslant \\ \leqslant E(e^{sS_{n-1}}e^{s^2c_n^2/8}) = \\ = e^{s^2c_n^2/8}E(e^{sS_{n-1}}). \end{aligned} \tag{10.5}$$

Здесь при переходе от первой строки ко второй была использована лемма 10.1.

Результат леммы получается путем итерации неравенства (10.5).
$\triangle$

Следующее следствие доказывается аналогично следствию 10.1.

Следствие 10.4. *Пусть $V_1, V_2, \ldots$ – мартингал-разность относительно последовательности случайных величин $X_1, X_2, \ldots$, кроме этого, $V_i \in [A_i, A_i + c_i]$ для некоторой случайной величины A_i, измеримой относительно $X_1, \ldots, X_{i-1}$, и некоторой последовательности положительных констант c_i. Если $S_n = \sum\limits_{i=1}^{n} V_i$, то для любого $n > 0$*

$$P\{|S_n| > c\} \leqslant 2\exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right).$$

Доказательство. Используем неравенство Маркова

$$P\{X > c\} \leqslant E(X)/c$$

и неравенство (10.1). Получим для произвольного n :

$$P\{S_n > c\} = P\{e^{sS_n} > e^{cs}\} \leqslant \exp\left(-cs + \frac{s^2 \sum\limits_{i=1}^{n} c_i^2}{8}\right)$$

для всех s. Находим минимум правой части по s. Он достигается при $s = 4c/\sum\limits_{i=1}^{n} c_i^2$. Отсюда получаем

$$P\{S_n > c\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right). \tag{10.6}$$

Аналогично получаем

$$P\{S_n < -c\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right).$$

Окончательно получаем

$$P\{|S_n| > c\} \leqslant 2\exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right).$$

Следствие 10.5. *В условиях следствия 10.4, где к тому же $c_i = 1$ для всех i, получаем*

$$P\left\{\frac{1}{n}|S_n| > c\right\} \leqslant 2e^{-2nc^2}. \tag{10.7}$$

Так как ряд экспонент в правой части неравенства (10.7) сходится, по лемме Бореля–Кантелли получим

Следствие 10.6. *В условиях следствия 10.4, где к тому же выполнено $B_1 < c_i < B_2$ для всех i, для некоторых положительных констант B_1, B_2 получаем усиленный мартингальный закон больших чисел:*

$$P\left\{\lim_{n\to\infty}\frac{S_n}{n} = 0\right\} = 1. \tag{10.8}$$

Неравенство Мак-Диармида. Следующее утверждение использовалось при оценках среднего Радемахера (см. также работы [45] и [51]).

Предложение 10.1. *Пусть $f : \mathcal{X}^n \to \mathcal{R}$ – функция, удовлетворяющая условию*

$$\begin{gathered}|f(x_1,\ \ldots,\ x_{i-1}, x_i, x_{i+1},\ \ldots,\ x_n) - \\ -f(x_1,\ \ldots,\ x_{i-1}, x_i',\ x_{i+1},\ \ldots,\ x_n)| \leqslant c_i\end{gathered} \tag{10.9}$$

для любого i и для всех $x_1,\ \ldots,\ x_i, x_i',\ \ldots,\ x_n \in \mathcal{X}$, где $c_1,\ \ldots,\ c_n$ – некоторые константы.

Пусть также $X_1,\ \ldots,\ X_n$ – независимые одинаково распределенные (согласно вероятностному распределению P) случайные величины, принимающие значения в множестве $\mathcal{X}$.

Тогда имеет место неравенство:

$$P^n\{f(X_1,\ \ldots,\ X_n) - E(f(X_1,\ \ldots,\ X_n)) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{n} c_i^2}\right), \qquad (10.10)$$

где E – символ математического ожидания по распределению P^n на выборках длины n.

Так как условие утверждения выполнено при замене f на $-f$, выполнено также неравенство:

$$P^n\{E(f(X_1,\ \ldots,\ X_n)) - f(X_1,\ \ldots,\ X_n) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{n} c_i^2}\right). \qquad (10.11)$$

Доказательство. Пусть функция $f(x_1, \ldots, x_n)$ определена на множестве $\mathcal{X}^n$ и удовлетворяет условию (10.9) с константами $c_1, \ldots, c_n$. Пусть также $X_1,\ \ldots,\ X_n$ – независимые одинаково распределенные (согласно вероятностному распределению P) случайные величины, принимающие значения в множестве $\mathcal{X}$. Определим мартингал-разности

$$V_i = E_{X_{i+1}\sim P,\ldots,X_n\sim P}(f|X_1, \ldots, X_i) - E_{X_i\sim P,\ldots,X_n\sim P}(f|X_1, \ldots, X_{i-1})$$

при $1 \leqslant i \leqslant n$. Действительно, из определения,

$$E_{X_i\sim P,\ldots,X_n\sim P}(V_i|X_1, \ldots, X_{i-1}) = 0$$

для всех i. Кроме этого, также легко видеть, что

$$\sum_{i=1}^{n} V_i = \sum_{i=1}^{n} f(X_1, \ldots, X_n) - E_{P^n}(f).$$

Границы изменения каждой величины V_i есть предсказуемые случайные величины

$$L_i = \inf_x E_{X_{i+1}\sim P,\ldots,X_n\sim P}(f|X_1,\ldots,X_{i-1},x) - \\ -E_{X_i\sim P,\ldots,X_n\sim P}(f|X_1,\ldots,X_{i-1}),$$
$$M_i = \sup_x E_{X_{i+1}\sim P,\ldots,X_n\sim P}(f|X_1,\ldots,X_{i-1},x) - \\ -E_{X_i\sim P,\ldots,X_n\sim P}(f|X_1,\ldots,X_{i-1}).$$

По свойству (10.9) будет $M_i - L_i \leqslant c_i$ при $1 \leqslant i \leqslant n$. Тогда по неравенству Хефдинга–Адзумы

$$P^n\{f(X_1,\ldots,X_n) - E(f) > t\} \leqslant e^{-\frac{2t^2}{\sum\limits_{i=1}^n c_i^2}}.$$

Предложение доказано. $\triangle$

Литература

[1] Э. Беккенбах, Р. Беллман. Неравенства. – М.: Мир, 1965. – 276 с.

[2] В.Н. Вапник, А.Я. Червоненкис. Теория распознавания образов (статистические проблемы обучения). – М.: Наука, 1974 – 416 с.

[3] В.В. Вьюгин. Элементы математической теории машинного обучения: учеб. пособие. – М.: МФТИ: ИППИ РАН, 2010. – 231с.

[4] А.Н. Ширяев. Вероятность. – М.: МЦНМО, 2007. – 968 с.

[5] Ю. М. Штарьков. Универсальное последовательное кодирование отдельных сообщений // Пробл. передачи информ. 23(3) 1987. 3–17.

[6] А.Б. Юдицкий, А.В. Назин, А.Б. Цыбаков, Н. Ваятис. Рекуррентное агрегирование оценок методом зеркального спуска с усреднением // Проблемы передачи информации. 41(4) 2005. 78–96.

[7] Е.В. Шикин, Г.Е. Шикина. Исследование операций: учебное пособие. – М.: ТК Велби, изд. Проспект, 2006. – 280 с.

[8] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability // J. ACM V. 1997. 44(4). P. 615–631.

[9] M. Anthony M, P.L. Bartlett. Neural network learning: Theoretical foundations, Cambridge: Cambridge University Press, 1999.

[10] N. Aronszajn. Theory of reproducing kernels // Transactions of the American Mathematical Society. 1950. V. 68. P. 337–404.

[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem // SIAM J. Comput. 32(1) P.48–77. 2002.

[12] P. Bartlett, S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results // Journal of Machine Learning Research. 2002. V.3. P. 463-482.

[13] P. Bartlett, O. Bousquet, S. Mendelson. Local Rademacher Complexities. // The Annals of Statistics. 2005, V. 33, No. 4, 1497–1537.

[14] O. Bousquet, M. Warmuth. Tracking a small set of experts by mixing past posteriors // Journal of Machine Learning Research. 2002, V.3, 363–396

[15] D. Blackwell. An analog of the minimax theorem for vector payoffs // Pacific Journal of Mathematics. 1956. V. 6. P. 1–8.

[16] Bousquet Oliver, Stephane Boucheron, and Gabor Lugosi. Introduction to statistical learning theory. Advanced Lectures on Machine Learning. 2004. P. 169–207.

[17] S. Bubeck, S., Cesa-Bianchi, N. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems // Foundations and Trends in Machine Learning V.5 N 1 (2012) 1–122.

[18] A. Chernov, V. Vovk. Prediction with Expert EvaluatorsЎ Advice. In: R. Gavald‘a, G. Lugosi, T. Zeugmann, S. Zilles (eds.) ALT 2009 Proceed- ings, LNCS(LNAI) vol. 5809, pp. 8Ц22. Springer, 2009. (See also, Technical Report arXiv:0902.4127v1 [cs.LG].)

[19] A. Chernov, F. Zhdanov. Prediction with expert advice under discounted loss. Algorithmic Learning Theory Volume 6331 of the series Lecture Notes in Computer Science pp 255–269 (See also, Technical report, arXiv:1005.1918v1 [cs.LG], 2010).

[20] A. Chernov, Y. Kalnishkan, F. Zhdanov, Vladimir Vovk // Supermartingales in Prediction with Expert Advice arXiv:1003.2218v1 [cs.LG] 10 Mar 2010.

[21] Cover, T.: Universal portfolios. Mathematical Finance. V.1 (1991) 1-Ţ29

[22] T. Cover, E. Ordentlich. Universal portfolio with side information // IEEE Transaction on Information Theory – 1996. – V. 42. – P. 348–363.

[23] Cover Thomas M., Thomas Joy A. Elements of Information Theory. 2nd ed. John Wiley and Sons, Inc., 2006. 748 p.

[24] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines. – Cambridge UK: Cambridge University Press, 2000.

[25] A.P. Dawid. Calibration-based empirical probability [with discussion] // Ann. Statist. – 1985. – V. 13. – P. 1251–1285.

[26] Kamalika Chaudhuri, Yoav Freund, and Daniel Hsu. A Parameter-free Hedging Algorithm. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 2009 ,22, P. 297–305.

[27] M. Herbster and M. Warmuth. Tracking the best expert. Machine Learning, 32(2) ,151–178, 1998.

[28] D.P. Foster, R. Vohra. Asymptotic calibration // Biometrika. – 1998. – V. 85. – P. 379–390.

[29] Y. Freund, R.E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // Journal of Computer and System Sciences – 1997. – V. 55. – P. 119–139.

[30] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, Contributions to the Theory of Games 3, pages 97-139, Princeton University Press, 1957.

[31] M. Hutter, J. Poland. Adaptive online prediction by following the perturbed leader // Journal of Machine Learning Research, 6:639–660, 2005.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, 2672Ц2680, 2014.

[33] Paulina Grnarova, Kfir Y. Levy, Aurelien Lucchi, Thomas Hofmann, Andreas Krause. An Online Learning Approach to GenerativeAdversarial Networks. arXiv:1706.03269v1 [cs.LG] 10 Jun 2017.

[34] S.N. Kakade, D.P. Foster. Deterministic calibration and Nash equilibrium // Lecture Notes in Computer Science – Berlin: Springer, 2004. – V. 3120. – P. 33–48.

[35] Sham Kakade, Ambuj Tewari. Topics in Artificial Intelligence (Learning Theory) - Spring 2008. Lecture Notes. *http : //ttic.uchicago.edu/ tewari/LT$_S$P2008.html*

[36] A. Kalai, S. Vempala. Efficient algorithms for online decisions. In Bernhard Scholkopf, Manfred K. Warmuth, editors, *Proceedings of the 16th Annual Conference on Learning Theory COLT 2003, Lecture Notes in Computer Science 2777*, pages 506–521, Springer-Verlag, Berlin, 2003. Extended version in Journal of Computer and System Sciences, 71:291–307, 2005.

[37] Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. In Proceedings of the 18th Annual Conference on Learning Theory (COLT), P.188Ц203, 2005.

[38] Yuri Kalnishkan. Kernel Methods (Introduction), 2008 (Unpublished manuscript). $http : //onlineprediction.net/?n = Main.KernelMethods\#toc8$

[39] G.S. Kimeldorf, G. Wahba. Some results on Tchebycheffian spline functions // J. Math. Anal. Appl. – 1971 –V. 33 – 82-Ц95.

[40] J.Kivinen, M.K.Warmuth. Averaging expert prediction. In Paul Fisher and Hans Ulrich Simon, editors, Computational Learning Theory: 4th European Conference (EuroColt '99), 1999, 153–167. Springer

[41] M. Ledoux, M. Talagrand. Probability in Banach Spaces: Isoperimetry and Processes. Springer, New York. 1991.

[42] N. Littlestone, M. Warmuth. The weighted majority algorithm // Information and Computation – 1994 – V. 108 – P. 212–261.

[43] G. Lugosi, N. Cesa-Bianchi. Prediction, Learning and Games. – New York: Cambridge University Press, 2006.

[44] S. Mannor, G. Stoltz. A Geometric Proof of Calibration // arXiv:0912.3604v2. 2009.

[45] C. McDiarmid. On the method of bounded differences. London Mathematical Society Lecture Notes Series. Surveys in Combinatorics. Cambridge University Press. V. 141. P. 148-Ц188. 1989.

[46] H. Robbins. Some aspects of the sequential design of experiment //. Bull. Amer. Math. Soc., 58(5) P. 527-Ц535, 1952.

[47] S. de Rooij, T. van Erven, D. Grunwald, M. Koolen. Follow the Leader If You Can, Hedge If You Must // Journal of Machine Learning Research. 2014, 15, 1281–1316.

[48] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization // Machine Learning 4 (2) P.107–194, 2011.

[49] G. Shafer, V. Vovk. Probability and Finance. It's Only a Game! – New York: Wiley. 2001.

[50] J. Shawe-Taylor, N. Cristianini. Margin distribution bounds on generalization // In Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'99. P.263–273. 1999.

[51] J. Shawe-Taylor, N. Cristianini. Kernel Methods for Pattern Analysis. – Cambridge UK: Cambridge University Press, 2004.

[52] B. Scholkopf, A. Smola. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

[53] Y. Singer. Switching portfolios. Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (UAI'98), 1998. P.488–495.

[54] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research, 2, 67-ЦI93, 2001

[55] L.G. Valiant. A theory of the learnable, Communications of the ACM V. 27(11). P. 1134-1142. 1984.

[56] V.N. Vapnik. Statistical Learning Theory. – New York: Wiley, 1998.

[57] V. Vovk. Aggregating strategies // Proceedings of the 3rd Annual Workshop on Computational Learning Theory (M. Fulk and J. Case, editors,) – San Mateo, CA: Morgan Kaufmann, 1990. – P. 371–383.

[58] V. Vovk. A game of prediction with expert advice // Journal of Computer and System Sciences – 1998 – V. 56. – No. 2. P. 153–173.

[59] V. Vovk. Derandomizing stochastic prediction strategies. Machine Learning, 35(3):247ЦI282, 1999.

[60] V. Vovk, C. Watkins. Universal portfolio selection // Proceedings of the 11th Annual Conference on Computational Learning Theory – New York: ACM Press, 1998. – P. 12–23.

[61] V. Vovk. Competitive on-line statistics // International Statistical Review – 2001 – V. 69. – P. 213–248.

[62] V. Vovk, A. Gammerman, G. Shafer. Algorithmic Learning in a Random World. Springer, New York, 2005.

[63] V. Vovk, G. Shafer. Good randomized sequential probability forecasting is always possible // J. Royal Stat. Soc. B. – 2005 – V. 67 – P. 747–763.

[64] V. Vovk, A. Takemura, G. Shafer. Defensive forecasting // Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (ed. by R. G. Cowell and Z. Ghahramani) – Cambridge UK: Society for Artificial Intelligence and Statistics, 2005. – P. 365–372.

[65] V. Vovk. On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract). TAMS Lecture Notes in Computer Science – Berlin: Springer, 3959, 2006, 452–463

[66] V. Vovk. Defensive Prediction with Expert Advice. In: S. Jain, H. Simon, E. Tomita (eds.) ALT 2005 Proceedings, LNCS(LNAI) vol. 3734, pp. 444–458. Springer, 2005. (See also: Competitive On-line Learning with a Convex Loss Function. Technical Report arXiv:cs/0506041v3 [cs.LG], arXiv.org e-Print archive, September 2005.)

[67] V. Vovk. On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract) // Lecture Notes in Computer Science – Berlin: Springer, 2006. – V. 3959. – P. 452–463.

[68] V. Vovk. Predictions as statements and decisions // Theoretical Computer Science – 2008. – V. 405. – No. 3. – P. 285–296.

[69] V. Vovk. Defensive Forecasting for Optimal Prediction with Expert Advice. Technical Report arXiv:0708.1503 [cs.LG], arXiv.org e-Print archive, August 2007.

[70] V. Vovk. Continuous and Randomized Defensive Forecasting: Unified View. Technical Report arXiv:0708.2353v2 [cs.LG], arXiv.org e-Print archive, August 2007.

[71] V. Vovk. Some comments on “A parameter-free hedging algorithm” by Chaudhuri, Freund, and Hsu (unpublished manuscript)

[72] V. V’yugin, V. Trunov. Universal algorithmic trading. Journal of Investment Strategies. V.2 (1), Winter 2012/13 P. 63–88.

[73] V. Vyugin. Universal Algorithm for Trading in Stock Market Based on the Method of Calibration // Lecture Notes in Artificial Intelligence (LNAI). 8139, P.53-67 (2013).

[74] Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the Twentieth International Conference on Machine Learning (ICML), 2003.