МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное агентство по образованию
Московский физико-технический институт
(государственный университет)
Учреждение Российской академии наук
Институт проблем передачи информации им. А.А. Харкевича
РАН

В.В.Вьюгин

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ МАШИННОГО ОБУЧЕНИЯ

Допущено
Учебно-методическим объединением
высших учебных заведений Российской Федерации
по образованию в области прикладных математики и физики
в качестве учебного пособия для студентов

по направлению «Прикладные математика и физика»

МОСКВА
МФТИ
2012

УДК 005.519.8(075.8)
ББК 65.290-2в6я73

Рецензенты:
д.ф-м.н. проф. А.В.Бернштейн,
д.ф-м.н. проф. Н.К.Верещагин

Вьюгин В.В. «Элементы математической теории машинного обучения» (учебное пособие). М.: 2010. - 341 с.

Предназначено для первоначального знакомства с математическими основами современной теории машинного обучения (Machine Learning) и теории игр на предсказания. Цель данного пособия – дать краткий обзор основных математических методов и алгоритмов, наиболее широко обсуждаемых в мировой научной литературе последних лет. В первой части излагаются основы статистической теории машинного обучения, рассматриваются задачи классификации и регрессии с опорными векторами, теория обобщения Вапника–Червоненкиса и алгоритмы построения разделяющих гиперплоскостей. Во второй части рассматриваются задачи адаптивного прогнозирования в режиме онлайн в теоретико-игровой и сравнительной постановке: игры с рандомизированными предсказаниями, предсказания с использованием экспертных стратегий (Prediction with Expert Advice).

Для студентов и аспирантов математических и прикладных математических специальностей, а также для специалистов в области исследования операций, прогнозирования и теории игр.

Библ. 22.

Оглавление

Введение

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

Многие задачи, возникающие в реальной жизни, 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 поступающих к нам данных. 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)*.

Отметим два типа методов машинного обучения. При первом методе 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 методе обучение никогда не прекращается, как говорится, оно происходит в режиме *онлайн*, т.е. предсказания и обучение происходят постоянно в процессе поступления данных.

Методы машинного обучения первого типа будут рассмотрены в главах 1 и 2, которые посвящены *статистической теории машинного обучения*, методы второго типа будут изучаться в главе 3 в теории *хорошо калибруемых* предсказаний (Calibration) и в главах 4 и 7, в которых представлена теория последовательных предсказаний с использованием *предсказаний экспертов* (Prediction with Expert Advice).

В главе 5 излагается алгоритм усиления слабых классификаторов – бустинг (Boosting). Приводится алгоритм AdaBoost, решающий эту задачу.

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.

После того как схема предсказания определена, нам необходимо оценить ее возможности, т.е. качество ее предсказаний.

Предварительно напомним, как оцениваются модели предсказания в классической статистической теории. В классической статистической теории последовательного предсказания мы предполагаем, что последовательность 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. 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. Оценки ошибок классификации или регрессии являются равномерными по всем таким вероятностным распределениям. Вероятность ошибочной классификации или регрессии называется ошибкой обобщения.

Для оценки качества схемы классификации или регрессии служит 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 результате функция классификации согласована с обучающей выборкой. Важнейшим параметром такой оценки является «сложность» – *размерность* класса функций классификации. 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.

Глава 2 посвящена построению алгоритмов классификации и регрессии. В основном, это алгоритмы, использующие метод опорных векторов.

Теория последовательного предсказания (глава 3) идет несколько дальше. Она вообще не использует гипотез о стохастических механизмах, генерирующих данные. Наблюдаемые исходы могут генерироваться совершенно неизвестным нам механизмом, который может быть как детерминистским так и стохастическим, 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 качество предсказаний. В отсутствие вероятностной модели функция риска в виде математического ожидания не может быть определена. Ее заменяют конкретные тесты, оценивающие рассогласованность между предсказаниями и соответствующими исходами. Один из видов таких тестов – серия тестов на калибруемость. Цель алгоритма – выдавать такие предсказания,

которые выдерживают все тесты на калибруемость.

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 – экспертов. 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. Наш алгоритм предсказания может использовать эти прогнозы, а также кумулятивные потери экспертов. Качество нашего предсказателя оценивается в наихудшем случае, а именно в виде разности между кумулятивными потерями предсказателя и кумулятивными потерями экспертов. Ошибка алгоритма предсказателя (регрет – regret) определяется как минимальное значение такой разности.

Будет рассмотрен метод распределения потерь в режиме онлайн, применимый в наиболее общей ситуации. Основной метод, использованный в главе 4, – это метод экспоненциального смешивания экспертных прогнозов.

В главе 5 метод распределения потерь в режиме онлайн будет применен для усиления слабых алгоритмов классификации. Слабый алгоритм классификации делает лишь незначительно меньшее число ошибок, чем простое случайное угадывание. Алгоритм AdaBoost, приводимый в этой главе, усиливает слабый алгоритм классификации до алгоритма, который с некоторого момента в процессе обучения начинает делать как угодно малое число ошибок.

В главе 6 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 3. Будет также построен соответствую-

щий алгоритм многомерной регрессии в режиме онлайн, основанный на применении агрегирующего алгоритма.

Предсказания в режиме онлайн тесно связаны с теорией игр. Теория игр рассматривается 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 [24]. 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.

В рамках этого подхода также наиболее естественным образом формулируется задача построения универсальных предсказаний, рассмотренная в главе 3. Рассматриваются бесконечно повторяющиеся игры с несколькими игроками. Выяснилось, что наиболее простым и естественным образом задача универсального предсказания формулируется в рамках теории игр. Процесс предсказания может рассматриваться как повторяющаяся игра между *Предсказателем* и *Природой*, генерирующей исходы; могут существовать также другие участники игры. Правила игры регулируются протоколом игры. Основные участники игры вычисляют свой выигрыш. Выигрывает тот участник, выигрыш которого неограниченно возрастает в процессе игры, либо его стратегия не позволяет другим участникам игры неограниченно наращивать свой выигрыш.

Специальный участник игры задает цель игры. Присоединяясь к *Природе*, он может вынуждать *Предсказателя* выдавать прогнозы, удовлетворяющие критерию, который он задал. Например, этот участник может вынуждать *Предсказателя* выдавать такие прогнозы, которые образуют распределения вероятностей, удовлетворяющие всем тестам на калибруемость прогнозов на по-

следовательности исходов, выдаваемой *Природой*. Универсальная стратегия *Предсказателя* будет строиться с использованием минимаксной теоремы.

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.

Данное учебное пособие представляет собой расширенный вариант курса лекций, прочитанных автором в 2008–2012 годах в Московском физико-техническом институте (МФТИ) в рамках специализации «Математические и информационные технологии».

Автор благодарен студентам МФТИ П.Д. Ерофееву, И.А. Жарову, А.А. Крещуку, А.Д. Шишкину за сделанные ими замечания.

Автор также благодарен Владимиру Вовку и Юрию Калнишкану за ценные замечания и советы по поводу изложения материала данного учебного пособия.

При составлении этого учебного пособия были использованы монографии: В.Н. Вапник и А.Я. Червоненкис [2], Vladimir Vapnik [28], Nello Cristianini, John Shawe-Taylor [10], Gabor Lugosi, Nicolo Cesa-Bianchi [21], Glenn Shafer, Vladimir Vovk [24], а также учебное пособие Е.В. Шикин, Г.Е. Шикина [4].

Глава 1

Элементы теории классификации

1.1. Задача классификации

1.1.1. Постановка задачи 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.

Мы рассмотрим два основных класса задач теории машинного обучения: *задачи классификации* и *задачи регрессии*.

В этом разделе будет рассматриваться задача классификации. Пусть задано множество объектов $\mathcal{X}$ и множество D классов этих объектов. В дальнейшем $\mathcal{X} \subseteq \mathcal{R}^n$, где $\mathcal{R}$ – множество всех действительных чисел, а D – 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 называть векторами (точками) и 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

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

умножения на вещественное число

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

где $\bar{x} = (x_1,\ \dots,\ x_n)'$ и $\bar{y} = (y_1,\ \dots,\ y_n)'$. [1]

На векторах из $\mathcal{R}^n$ также определено их скалярное произведение $(\bar{x} \cdot \bar{y}) = x_1y_1 + \dots + x_ny_n$. Норма (длина) вектора $\bar{x}$ определяется как

$$\|\bar{x}\| = \sqrt{(\bar{x} \cdot \bar{x})} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

При решении задачи классификации мы исходим из *обучающей выборки* $S = ((\bar{x}_1, y_1),\ \dots,\ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{X}$ – вектор евклидового пространства $\mathcal{R}^n$ большой 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 $S = ((\bar{x}_1, y_1),\ \dots,\ (\bar{x}_l, y_l))$, при этом элементы y_i обычно являются вещественными числами, т.е. $D = \mathcal{R}$. Задача регрессии будет рассмотрена в разделах 2.8, 2.9, 2.9.2, а также в разделе 6.9.

Мы предполагаем, что выборка $S = ((\bar{x}_1, y_1),\ \dots,\ (\bar{x}_l, y_l))$ генерируется (порождается) некоторым источником. Основное предположение об источнике, порождающем выборку S, заключается

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

в том, что на парах $(\bar{x}, y)$, т.е. на пространстве $\mathcal{X} \times D$ задано распределение вероятностей P, а пары $(\bar{x}_i, y_i)$, образующие выборку S, одинаково и независимо распределены.

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

Строго говоря, пары $(\bar{x}, y)$ являются реализациями случайной величины $(\bar{X}, Y)$, которая имеет распределение вероятностей P. Плотность распределения P будет обозначаться так же как $P(\bar{x}, y)$.

Правило или *функция классификации* – это функция типа $h : \mathcal{X} \to D$, которая разбивает элементы $\bar{x}_i \in \mathcal{X}$ 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: $S^+ = ((\bar{x}_i, y_i) : y_i = 1)$ – положительные примеры (или первый класс) и $S^- = ((\bar{x}_i, y_i) : y_i = -1)$ – отрицательные примеры (или второй класс).

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

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

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

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

Здесь $h(X)$ – функция от случайной величины X, также является случайной величиной, поэтому можно рассматривать вероятность события $\{h(\bar{X}) \neq Y\}$.

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.

1.1.2. Байесовский классификатор

Предварительно рассмотрим один простейший метод классификации. Легко построить оптимальный классификатор, если распределение вероятностей P, генерирующее пары $(\bar{x}_i, y_i)$, известно.

Рассмотрим пары случайных переменных $(\bar{X}, Y)$, принимающих значения в множестве типа $\mathcal{X} \times \{-1, 1\}$. Предполагаем, что на этим парам соответствует распределение вероятностей P и соответствующая плотность вероятности $P(\bar{x}, y)$. Предполагаем, что существуют условные плотности $P(\bar{x}|Y = 1)$ – плотность распределения векторов первого класса, а также $P(\bar{x}|Y = -1)$ – плотность распределения векторов второго класса. Величины $P\{Y = 1\}$ и $P\{Y = -1\}$ определяют вероятности появления векторов первого и второго классов соответственно. Все эти вероятности и плотности вероятностей легко вычисляются по плотности вероятности $P(\bar{x}, y)$. Например, $P\{Y = 1\} = \int\limits_{\mathcal{X}} P(\bar{x}, 1)d\bar{x}$, а $P(\bar{x}|Y = 1) = P(\bar{x}, 1)/P\{Y = 1\}$. [2]

Используя эти вероятности, можно по формуле Байеса определить апостериорные вероятности принадлежности вектора $\bar{x}$ к первому и второму классу

$$P\{Y = 1|\bar{X} = \bar{x}\} = cP(\bar{x}|Y = 1)P\{Y = 1\},$$
$$P\{Y = -1|\bar{X} = \bar{x}\} = cP(\bar{x}|Y = -1)P\{Y = -1\},$$

где

$$c = \frac{1}{P(\bar{x}|Y = 1)P\{Y = 1\} + P(\bar{x}|Y = -1)P\{Y = -1\}},$$

Рассмотрим условную вероятность того, что вектор $\bar{x}$ принадлежит первому классу

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

[2]Здесь и далее мы предполагаем, что $P\{Y = 1\} > 0$ и $P\{Y = -1\} > 0$, а также $P(\bar{x}|Y = 1) > 0$ и $P(\bar{x}|Y = -1) > 0$.

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

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

Байесовский классификатор определяется как

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

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* $g : \mathcal{X} \to \{-1, 1\}$

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

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

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

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

Аналогичное неравенство выполнено для классификатора $h(\bar{x})$.

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

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

по определению байесовского классификатора h.

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

Байесовский классифкатор 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

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

Пусть некоторый классификатор $g_l \in \mathcal{D}$ построен некоторым алгоритмом $\mathcal{A}$ по случайной выборке $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ 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 $h(\bar{x})$ вероятностное распределение P, генерирующее пары $(\bar{x}, y)$. 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 $(\bar{X}, Y)$.

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

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

Байесовский классификатор служит для оценки качества других алгоритмов классификации.

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 в пространстве $\mathcal{R}^n$.

В некоторых персептронах рассматриваются преобразованные признаки, которые принимают всего два значения (бинарные признаки), например, рассматривается случай $\mathcal{Y} = \{-1, 1\}$.

Для простоты считаем, что персептрон различает два понятия или класса объектов. Персептрон относит вектор $\bar{x}$ к первому

классу, если

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

в противном случае, персептрон относит вектор $\bar{x}$ ко 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

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

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

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

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

для всех i.

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

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

для всех i.

Обозначим

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

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

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

Переходим теперь к описанию алгоритма Розенблатта построения разделяющей гиперплоскости.

Пусть задана произвольная бесконечная обучающая выборка

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

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

$$(\Lambda^* \cdot \tilde{y}_i) > 0 \tag{1.6}$$

для всех i. Считаем, что $|\Lambda^*| = 1$. Для бесконечной выборки мы усилим условие строгого разделения (1.6): предполагаем, что существует порог разделения – число $\rho_0 > 0$ такое, что

$$(\Lambda^* \cdot \tilde{y}_i) > \rho_0 \tag{1.7}$$

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

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

Алгоритм Розенблатта построения разделяющей гиперплоскости

Обучение персептрона заключается в изменении координат вектора весов Λ на каждом шаге алгоритма.

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 $\tilde{y}_1, \tilde{y}_2,\ \ldots$.

Полагаем $\Lambda_0 = (0,\ \ldots,\ 0)$.

FOR $t\ =\ 1,\ 2,\ \ldots$

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

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*

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

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

$$\left[\frac{D^2}{\rho_0^2}\right]$$

раз. Это значит, что неравенство $\Lambda_t \neq \Lambda_{t-1}$ *выполнено для не более чем* $\left[\frac{D^2}{\rho_0^2}\right]$ *различных* t. [3] *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

$$\|\Lambda_t\|^2 \leqslant kD^2. \tag{1.8}$$

По условию разделимости (1.7) существует единичный вектор Λ^* такой, что

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

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

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

[3] $[r]$ – целая часть вещественного числа r.

По неравенству Коши–Буняковского

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

Поэтому имеет место неравенство

$$\|\Lambda_t\| \geqslant k\rho_0. \tag{1.9}$$

Объединяем неравенства (1.8) и (1.9), получаем

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

Таким образом, число исправлений не превосходит

$$k \leqslant \left[\frac{D^2}{\rho_0^2}\right].$$

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 $\left[\frac{D^2}{\rho_0^2}\right]$, 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

$$k \leqslant \left[\frac{n}{\rho_0^2}\right],$$

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.

Возникает естественный вопрос, всегда ли существует такое отображение $\bar{y} = \bar{\phi}(\bar{x})$, при котором образы любых двух непересекающихся в исходном пространстве $\mathcal{X}$ множеств были бы разделены в спрямляющем пространстве $\mathcal{Y}$ гиперплоскостью.

Было показано, что если исходное пространство $\mathcal{X}$ бинарное, то такое отображение существует, более того, оно может быть осуществлено с помощью линейных функций (см. [2]). Недостатком этого результата является то, что размерность спрямляющего пространства оказывается очень большой. Для большинства пар множеств отношение $\frac{D^2}{\rho_0^2}$ слишком велико и поэтому обучение персептрона может потребовать значительно большого времени.

По этой причине, обычно отображение $\bar{y} = \bar{\phi}(\bar{x})$ выбирается на основании знания конкретной предметной области.

Многослойная нейронная сеть

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

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

Функция σ называется функцией активации; на место аргумента в ней подставлено значение персептрона. Примеры функций активации:

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

где

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

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-ой 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*. Задачей теории обобщения является оценить вероятность ошибки классификации на произвольной тестовой выборке.

Теория обобщения Вапника–Червоненкиса позволяет вычислить вероятность ошибки классификации или регрессии (относительно, возможно неизвестного нам, распределения вероятностей, генерирующего данные) 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. Это позволяет контролировать зависимость параметров обучения и вероятности ошибки в будущем.

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 $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ – обучающая выборка. Здесь $\bar{x}_i \in \mathcal{X}$ и $y_i \in \{-1, 1\}$ при $1 \leqslant i \leqslant l$. В дальнейшем $\mathcal{X} \subseteq \mathcal{R}^n$ – n-мерное евклидово векторное пространство.

В данном разделе при вероятностном анализе мы предполагаем, что выборка S – это случайная величина, состоящая из случайных величин $(\bar{x}_i, y_i)$, $i = 1,\ \ldots\ l$. Для удобства (в отличие от раздела 1.1.2) мы обозначаем случайные величины $(\bar{x}_i, y_i)$ маленькими буквами.

Пусть задано правило (или функция) $h : \mathcal{X} \to \{-1, 1\}$. Риск

функционал (или ошибка классификации) определяется как

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

Эта величина равна вероятности неправильной классификации.

Гипотеза классификации h согласована с выборкой

$$S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l)),$$

если $h(\bar{x}_i) = y_i$ для всех $1 \leqslant i \leqslant l$. Обозначим

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

– относительное число ошибок классификации h на выборке S. Здесь $|A|$ – число элементов множества A. Тогда гипотеза классификации h согласована с выборкой S, если $\mathrm{err}_S(h) = 0$.

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

$$\begin{aligned} P^l\{S : \mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon\} = \\ = \prod_{i=1}^{l} P\{h(\bar{x}_i) = y_i\} = \\ = \prod_{i=1}^{l}(1 - P\{h(\bar{x}_i) \neq y_i\}) = \\ = (1 - \mathrm{err}_P(h))^l \leqslant e^{-l\epsilon}. \end{aligned} \tag{1.10}$$

Здесь мы использовали независимость ошибок на элементах выборки.

Пусть H – некоторый класс гипотез классификации. Если класс H конечный, то из (1.10) получаем оценку

$$P^l\{S : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant |H|e^{-l\epsilon}. \tag{1.11}$$

Интерпретация (1.11) заключается в следующем.

Пусть задан критический уровень $\delta > 0$ принятия ошибочной гипотезы классификации $h \in H$, согласованный с обучающей выборкой S. Тогда по (1.11) мы можем утверждать, что с вероятностью $\geqslant 1 - \delta$ гипотеза классификации $h_S \in H$, построенная

по случайной обучающей выборке S и согласованная с ней, будет иметь ошибку классификации $\mathrm{err}_P(h) \leqslant \epsilon = \frac{1}{l} \ln \frac{|H|}{\delta}$.

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

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

$$B_H(l) = \max_{(\bar{x}_1, \bar{x}_2, \ \ldots, \ \bar{x}_l)} |\{(h(\bar{x}_1), h(\bar{x}_2), \ \ldots, \ h(\bar{x}_l)) : h \in H\}|.$$

Свойства этой функции будут изучаться далее.

Имеет место теорема – аналог соотношения (1.11) 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 $1_A(\bar{x}) = 1$, если $\bar{x} \in A$, и $1_A(\bar{x}) = 0$, если $\bar{x} \notin A$. Аналогично $1_{h(\bar{x}) \neq y}(\bar{x}, y)$ есть случайная величина, равная 1, если $h(\bar{x}) \neq y$ и равная 0, в противном случае. Тогда

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

где E – математическое ожидание по мере P. По определению

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

– 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*

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \epsilon)\} \leqslant$$
$$\leqslant\ 2P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \frac{1}{2}\epsilon)\}. \qquad (1.12)$$

Доказательство. Легко видеть, что неравенство (1.12) эквивалентно неравенству

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

Докажем (1.13). Для каждой выборки S из множества левой части неравенства (1.13) 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 [4]

$$1_{\mathrm{err}_S(h_S)\ =\ 0 \& \mathrm{err}_P(h_S)\ >\ \epsilon} 1_{\mathrm{err}_P(h_S)\ -\ \mathrm{err}_{S'}(h_S)\ \leqslant\ \frac{1}{2}\epsilon} \leqslant$$
$$\leqslant 1_{\mathrm{err}_S(h_S)\ =\ 0 \& \mathrm{err}_{S'}(h_S) > \frac{1}{2}\epsilon}. \qquad (1.14)$$

Возьмем математическое ожидание по второй выборке S' от обеих частей неравенства (1.14). Получим неравенство для случайных величин, зависящих от первой выборки S :

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

[4] Здесь $1_{\mathrm{err}_S(h_S)=0 \& \mathrm{err}_P(h_S)>\epsilon}(S) = 0$, если S не лежит в множестве из левой части неравенства (1.13). Также $1_{\mathrm{err}_S(h_S)=0 \& \mathrm{err}_{S'}(h_S)>\frac{1}{2}\epsilon}(SS') = 0$, если SS' не лежит в множестве из правой части неравенства (1.13).

Используя свойства биномиального распределения получаем

$$P^l\{S' : \mathrm{err}_P(h_S) - \mathrm{err}_{S'}(h_S) \leqslant \frac{1}{2}\epsilon\} =$$
$$= P^l\{S' : \mathrm{err}_{S'}(h_S) \geqslant \mathrm{err}_P(h_S) - \frac{1}{2}\epsilon\} =$$
$$= \sum_{\{k:k/l \geqslant p-\epsilon/2\}} \binom{l}{k} p^k (1-p)^{n-k} > \frac{1}{2} \tag{1.16}$$

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

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

Делаем замену переменных в этой сумме:

$$\sum_{\{k:k<lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} =$$
$$= \sum_{\{k:l-k>l(1-p)+1\}} \binom{l}{k} p^k (1-p)^{n-k} =$$
$$= \sum_{\{k:k>lp+1\}} \binom{l}{k} p^k (1-p)^{n-k}. \tag{1.17}$$

Сумма первой и третьей сумм из (1.17) меньше 1. Поэтому каждая из них меньше $1/2$.

Подставляя неравенство (1.16) в (1.15), получим

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

Возьмем среднее по S и получим

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

Отсюда получаем (1.13). 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. Определим функцию η, которая по произвольной выборке $SS' = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_{2l}, y_{2l}))$ длины $2l$ выдает ее состав Υ, т.е. множество пар ее составляющих вместе с кратностями

$$\eta(SS') = \Upsilon = \{((\bar{x}_1, y_1), k_1), \ \ldots, \ ((\bar{x}_L, y_L), k_L)\},$$

где k_i – число вхождений пары $(\bar{x}_i, y_i)$ в выборку SS', $i = 1, \ \ldots, \ L$, L – число различных пар $(\bar{x}_i, y_i)$ в выборке SS'; по определению $k_1 + \cdots + k_L = 2l$.

В отличие от выборки ее состав – неупорядоченное множество. Мера P^{2l} на выборках длины $2l$ индуцирует меру $\hat{P}$ на их составах:

$$\hat{P}(\Xi) = P^{2l}\{SS' : \eta(SS') \in \Xi\},$$

где Ξ – множество, состоящее из составов типа Υ.

Фиксируем некоторый состав Υ для выборок длины $2l$. Предварительно также фиксируем некоторую функцию классификации h.

Для каждой двойной выборки $SS' = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_{2l}, y_{2l}))$ определим бинарную последовательность $\epsilon_1, \ \ldots, \ \epsilon_{2l}$ ошибок классификации, где

$$\epsilon_i = \begin{cases} 1, \text{ если } h(\bar{x}_i) \neq y_i, \\ -1, \text{ если } h(\bar{x}_i) = y_i, \end{cases}$$

где $i = 1, \ldots, 2l$.

Поскольку ошибки классификации описываются бернуллиевским распределением с вероятностью ошибки $P\{h(\bar{x}) \neq y\}$, любые два набора ошибок $\epsilon_1, \ldots, \epsilon_{2l}$ и $\epsilon'_1, \ldots, \epsilon'_{2l}$ на двух выборках с одним и тем же составом Υ равновероятны. [5]

Поэтому вероятность того, что для некоторого фиксированного состава Υ на некоторой двойной выборке SS', имеющей состав Υ (т.е. такой, что $\eta(SS') = \Upsilon$), функция h делает m ошибок и все они сосредоточены на второй половине этой выборки, оценивается сверху при $m \geqslant \epsilon l$:

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

Пусть теперь функция классификации h принимает любое значение из класса H. Число всех функций, которые получаются ограничением области определения функций из H на множество всех объектов $\{\bar{x}_1, \ldots, \bar{x}_{2l}\}$ из выборок SS' данного состава $\eta(SS') = \Upsilon$, не превосходит числа элементов множества $\{(h(\bar{x}_1), h(\bar{x}_2), \ldots, h(\bar{x}_{2l}) : h \in H\}$, состоящего из бинарных последовательностей длины $2l$.

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

$$B_H(l) = \max_{(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_l)} |\{(h(\bar{x}_1), h(\bar{x}_2), \ldots, h(\bar{x}_l)) : h \in H\}|.$$

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

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

[5] Эти вероятности определяются биномиальным распределением и равны $\binom{2l}{k} p^k (1-p)^{2l-k}$, где $p = P\{h(\bar{x}) \neq y\}$ и k – число единиц (число ошибок) в выборке.

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

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

Левая часть этого неравенства представляет собой случайную величину (функцию от состава Υ). Правая часть неравенства не зависит от состава Υ.

Интегрируя это неравенство по мере $\hat{P}$ на составах Υ, получим безусловное неравенство

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

Лемма 1.3 доказана. $\triangle$

Теорема 1.2 непосредственно следует из лемм 1.2 и 1.3.

Из теоремы 1.2 следует, что всякая гипотеза классификации h, имеющая ошибку $\mathrm{err}_P(h) > \epsilon$, с вероятностью $\geqslant 1 - 2B_H(2l)e^{-\epsilon l/4}$ 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 следующее следствие

Следствие 1.1. *Допустим, что класс H функций классификации имеет конечную VC-размерность d.* [6]

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

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

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

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

при $l \geqslant d$.

Все эти результаты можно усилить на случай обучения с ошибками. 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* $l > 2/\epsilon$ *имеет место неравенство*

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

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*

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

Доказательство этой леммы аналогично доказательству леммы 1.3. При этом удобно использовать неравенство Хефдинга (см. следствие 4.5 из раздела 4.7 ниже).

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

Теорема 1.3. *Имеет место оценка*

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

при $l > 2/\epsilon$.

Отсюда получаем следующее следствие, связывающее вероятность ошибки обобщения и среднее число ошибок на обучающей выборке.

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

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

при $l \geqslant d$.

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 не зависящие от размерности пространства (см. теоремы 1.9, 2.10, следствие 2.2 далее).

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 H – произвольный класс функций классификации, функция $h \in H$. Бинарный набор $(h(\bar{x}_1),\ \ldots,\ h(\bar{x}_l))$, состоящий из элементов множества $\{-1, 1\}$, определяет разделение множества $\{\bar{x}_1,\ \ldots,\ \bar{x}_l\}$ на два подмножества $\{\bar{x}_i : h(\bar{x}_i) = 1\}$ – положительные примеры и $\{\bar{x}_i : h(\bar{x}_i) = -1\}$ – отрицательные примеры.

Множество $\{\bar{x}_1,\ \ldots,\ \bar{x}_l\}$ полностью разделено функциями из H (shattered by H), если

$$\{(h(\bar{x}_1),\ \ldots,\ h(\bar{x}_l)) : h \in H\} = \{-1, 1\}^l.$$

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

$$B_H(l) = \max_{(\bar{x}_1, \bar{x}_2,\ \ldots,\ \bar{x}_l)} |\{(h(\bar{x}_1), h(\bar{x}_2),\ \ldots,\ h(\bar{x}_l) : h \in H\}|.$$

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

Основная теорема теории *VC*-размерности (лемма Сауэра) [7]:

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

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*

$$B_H(l) \leqslant \sum_{i=0}^{d} \binom{l}{i} \leqslant \left(\frac{el}{d}\right)^d \tag{1.21}$$

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. *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 (1.21) 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 $< l+d$, 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

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

Тогда нам надо доказать, что для любого класса функций H с VC-размерностью $\leqslant d$ будет $B_H(l) \leqslant h(l, d)$ 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 с VC-размерностью d и пусть $X_1 = \{x_1, x_2, \ldots,\ x_l\}$ – произвольное множество объектов мощьности l, $X_2 = \{x_2, \ldots,\ x_l\}$ – оно же, но без первого элемента.

Рассмотрим ограничения $H_1 = H_{|X_1}$ функций из класса H на множество X_1 и $H_2 = H_{|X_2}$ – ограничения функций из класса H на множества X_2.

Пусть также класс функций H_3 состоит функций f из класса H_2, для каждой из которых найдется функция $f' \in H$ такая, что $f'(x_1) = -f(x_1)$.

Заметим, 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 класс функций полностью разделяет некоторое множество объектов мощности d, то сущствуют 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 этом элементе). В это случае, класс H_1 также полностью разделяет это расширенное множество, а его размерность равна $d+1$. Противоречие.

По предположению индукции

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

Поэтому

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

Так множество X произвольное, получаем

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

Неравенство (1.21) 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:

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

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

$$= \left(\frac{l}{n}\right)^d \left(1 + \frac{d}{l}\right)^l < \left(\frac{l}{d}\right)^d e^d = \left(\frac{el}{d}\right)^d. \qquad (1.22)$$

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 $h(\bar{x}) = \text{sign}(L(\bar{x}))$, где $L(\bar{x})$ – линейная функция. Здесь $\text{sign}(r) = 1$, если $r > 0$, $\text{sign}(r) = -1$, в противном случае. Иногда классификатором будем называть соответствующую функцию $L(\bar{x})$.

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 $\text{sign}(L(\bar{x})) = \text{sign}(\bar{a} \cdot \bar{x})$ 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 классифи-*

кации над $\mathcal{R}^n$ выполнено

$$G_{\mathcal{L}}(l) = \ln H_{\mathcal{L}}(l) = \ln\left(2\sum_{i=0}^{n-1}\binom{l-1}{i}\right) <$$
$$< (n-1)(\ln(l-1) - \ln(n-1) + 1) + \ln 2 \qquad (1.23)$$

при $l > n$.

Доказательство. Предварительно 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 $h(\bar{x}) = \operatorname{sign}(L(\bar{x}))$, где $L(\bar{x}) = a_1x_1 + \cdots + a_nx_n$, 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 $z_{i,j} = (\bar{a}_j \cdot \bar{u}_i)$, $i = 1, \ldots, n+1$, $j = 1, \ldots, 2^{n+1}$,

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

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,

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

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

$$L'(\bar{x}) = a_1 x_1 + \cdots + a_n x_n + b x_{n+1}.$$

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 разделе-

ние 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

$$\Phi(n, l) = \Phi(n, l-1) + \Phi(n-1, l-1) \tag{1.24}$$

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

Доказать в виде задачи, что рекуррентное соотношение (1.24) 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 (1.23) (а также из неравенства (1.21)) 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 (1.22). Доказательство теоремы закончено. $\triangle$

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

Пусть $\mathcal{F}$ – некоторый класс иидикаторных 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 (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, также по лемме Сауэра, выполнены сле-

дующие неравенства:

$$B_{\mathcal{F}}(m) \leqslant \prod_{i=1}^{l} B_{\mathcal{F}^i}(m) \leqslant$$

$$\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} B_{\mathcal{F}^{i,j}}(m) \leqslant$$

$$\leqslant \prod_{i=1}^{l} \prod_{j=1}^{n_i} \left(\frac{le}{n_{i-1}+1} \right)^{d_{i-1}+1} =$$

$$= \prod_{i=1}^{l} \left(\frac{me}{n_{i-1}+1} \right)^{n_i(n_{i-1}+1)} \leqslant (me)^N,$$

где

$$N = \sum_{i=1}^{l} d_i(d_{i-1}+1)$$

– 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

векторы, 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 векторы 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 $\mathcal{R}^n$, $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ – выборка длины l, $\epsilon > 0$.

Каждой функции $f \in \mathcal{F}$ сопоставим индикаторную функцию классификации

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

Границей ошибки при классификации примера $(\bar{x}_i, y_i)$ с помощью вещественной функции f называется величина $\gamma_i = y_i f(\bar{x}_i)$. Заметим, что $\gamma_i > 0$ означает, что классификация с помощью f является правильной. 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.

Назовем *ϵ-покрытием* множества функций $\mathcal{F}$ относительно множества $X = \{\bar{x}_1,\ \ldots,\ \bar{x}_l\}$ любое конечное множество функций $\mathcal{B}$ такое, что для любого $f \in \mathcal{F}$ существует $g \in \mathcal{B}$ такая, что $|f(\bar{x}_i) - g(\bar{x}_i)| < \epsilon$ при $i = 1,\ \ldots,\ l$.

Пусть $\mathcal{N}(\epsilon, \mathcal{F}, X)$ – размер покрытия $\mathcal{B}$, имеющего наименьшее значение величины $|\mathcal{B}|$ (минимальное покрытие).

Максимум этих величин по всем множествам X мощности l

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

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

По определению $\mathrm{err}_S(f)$ – доля ошибок классификации с помощью функции f на выборке $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$. Эта величина равна доле векторов $\bar{x}_i$ таких, что $y_i f(\bar{x}_i) \leqslant 0$.

Пусть P – распределение вероятностей на $\mathcal{R} \times \{-1, 1\}$ генерирующее элементы выборки S. Рассмотрим вероятность ошибочной классификации примера $(\bar{x}, y)$ с помощью функции f :

$$\mathrm{err}_P(f) = P\{yf(\bar{x}) \leqslant 0\}.$$

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 (1.12) 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 1.4.

Вторая лемма аналогична лемме 1.5

Лемма 1.7. *При* $l > 2/\epsilon$

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

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

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

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 $m_S(g) > \gamma/2$, а на второе половине выборки либо делает ошибки, либо имеет порог разделения $\leqslant \gamma/2$ в доле $\mathrm{err}_{\hat{S}}(\gamma/2, g) > \frac{\epsilon}{2}$ примеров. Оценка такая же как и в лемме 1.5, а именно, (1.20).

В результате получаем оценку

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

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

Из теоремы 1.6 получаем

Следствие 1.3. *Заданы класс $\mathcal{F}$ вещественных функций и числа $\gamma > 0$, $\delta > 0$. Тогда для любого распределения вероятностей P на $\mathcal{R}^n \times \{-1, 1\}$ с вероятностью $1 - \delta$ на случайной выборке S длины l любая функция $f \in \mathcal{F}$, которая классифицирует S с границей ошибки $m_S(f) > \gamma$, имеет верхнюю границу вероятности ошибочной классификации*

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

при всех l.

Каждый класс функций $\mathcal{F}$ порождает так называемую пороговую размерность или fat-размерность (fat-shattered dimension). Пусть $\gamma > 0$. Множество $X = \{\bar{x}_1, \ldots, \bar{x}_l\}$ называется *γ-разделимым*, если существуют вещественные числа $r_1, \ldots, r_l$ такие, что для любого подмножества $E \subseteq X$ существует функция $f_E \in \mathcal{F}$ такая, что $f_E(\bar{x}_i) \geqslant r_i + \gamma$, если $\bar{x}_i \in E$, и $f_E(\bar{x}_i) < r_i - \gamma$, если $\bar{x}_i \notin E$.

Множество X называется *γ-разделимым на одном уровне*, если $r_i = r$ для всех i.

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 $\mathcal{F}$ зависит от параметра $\gamma > 0$. 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 $\mathcal{R}^n \to [a, b]$, где $a < b$. Выберем $0 < \gamma < 1$ и обозначим $d = \mathrm{fat}_{\gamma/4}(\mathcal{F})$. Тогда*

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

Теорема 1.7 вместе со следствием 1.3 влечет следующее 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 верхнюю границу вероятности ошибочной классификации*

$$\mathrm{err}_P(f) \leqslant \frac{4}{l} \left(d \ln \frac{8el}{d\gamma} \ln \frac{32l}{\gamma^2} + \ln \frac{4}{\delta} \right)$$

при $l \geqslant d$, 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 – шар радиуса R в n-мерном евклидовом пространстве: $X = \{\bar{x} : |\bar{x}| \leqslant R\}$, 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 $\mathcal{F}$. Тогда для произвольного подмножества $\hat{Y} \subseteq Y$ найдется весовой вектор $\bar{w}$, $\|\bar{w}\| \leqslant 1$, такой что $(\bar{w} \cdot \bar{x}_i) > \gamma$ для $\bar{x}_i \in \hat{Y}$ и $(\bar{w} \cdot \bar{x}_i) \leqslant -\gamma$ для $\bar{x}_i \notin \hat{Y}$. Имеем

$$\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \notin \hat{Y}} (\bar{w} \cdot \bar{x}_i) \geqslant \gamma l. \tag{1.25}$$

Из неравенства Коши–Буняковского для евклидовой нормы получаем

$$\begin{aligned} &\sum_{x_i \in \hat{Y}} (\bar{w} \cdot \bar{x}_i) - \sum_{x_i \notin \hat{Y}} (\bar{w} \cdot \bar{x}_i) = \\ &= \left(\bar{w} \cdot \left(\sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i\right)\right) \leqslant \\ &\leqslant \left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\|^2 \cdot \|\bar{w}\|^2. \end{aligned} \tag{1.26}$$

Из (1.25), (1.26) и из неравенства $|\bar{w} \leqslant 1$ получаем нижнюю оценку

$$\left\| \sum_{x_i \in \hat{Y}} \bar{x}_i - \sum_{x_i \notin \hat{Y}} \bar{x}_i \right\| \geqslant \gamma l. \tag{1.27}$$

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 (1.27) относительно случайного разбиения множества Y определяемого вектором $\bar{\xi}$. Имеем

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

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

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

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

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

Вместе с неравенством (1.27) это влечет $R\sqrt{l} \geqslant \gamma l$. Отсюда получаем $l \leqslant (R/\gamma)^2$. Это означает, что $\mathrm{fat}_\gamma(\mathcal{F}) \leqslant (R/\gamma)^2$. $\triangle$

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

Теорема 1.9. *Рассмотрим задачу классификации с помощью линейных функций $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b \in \mathcal{L}$, где $\bar{x} \in \mathcal{R}^n$, $\|\bar{w}\| = 1$. Задано число $\gamma > 0$.*

Для произвольного распределения вероятностей P, сконцентрированного в шаре радиуса R с центром в начале координат 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

$$\operatorname{err}_P(f) = P\{yf(\bar{x}) \leqslant 0\} \leqslant$$
$$\leqslant \frac{4}{l}\left(\frac{64R^2}{\gamma^2}\ln\frac{el\gamma}{8R^2}\ln\frac{32l}{\gamma^2} + \ln\frac{4}{\delta}\right) \tag{1.28}$$

при $l > \frac{64R^2}{\gamma^2}$.

Оценки теорем 1.8 и 1.9 послужат основой для получения не зависящих от размерности пространства оценок вероятности ошибки обобщения для машин на опорных векторах в теореме 2.4 из раздела 2.6.1 и в теоремах 2.9 и 2.10 из раздела 2.9.1.

1.3.2. Покрытия и упаковки

При составлении данного раздела были использованы лекции Какаде и Тевари [17].

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

$$\mathcal{N}_d(\alpha, A) = \min\{|B| : B \text{ является } \alpha\text{-покрытием } A\}. \tag{1.29}$$

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

$$\mathcal{M}_d(\alpha, A) = \max\{|B| : B \subseteq A \text{ является } \alpha\text{-отделимым}\}. \tag{1.30}$$

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 большем α. 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. 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 $\mathcal{F} \subseteq B^{\mathcal{X}}$, 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 $l(f,g) > 2$. Иными словами существует $x \in \mathcal{X}$ такое, что $|f(x) - g(x)| > 2$. 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

$$f^{\alpha}(x) = \left\lfloor \frac{f(x)}{\alpha} \right\rfloor$$

для всех x, а 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 $\mathcal{F}^{\alpha}$.

Число покрытия $\mathcal{N}_d(\alpha, A)$ и число упаковки $\mathcal{M}_d(\alpha, A)$ были определены согласно (1.29) и (1.30).

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:

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

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

Лемма 1.9. *Пусть $\mathcal{F} \subseteq B^{\mathcal{X}}$ и $\alpha > 0$. Тогда*

$$Sdim(\mathcal{F}^{\alpha}) \leqslant fat_{\alpha/2}(\mathcal{F}), \tag{1.31}$$

$$\mathcal{M}(\alpha, \mathcal{F}, X) \leqslant \mathcal{M}(2, \mathcal{F}^{\alpha/2}, X) \tag{1.32}$$

Доказательство леммы предлагается в качестве задачи.

Следующая лемма представляет техническую часть основного результата этого раздела.

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

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

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

Доказательство. Допустим, что $b \geqslant 2$ и определим функцию $t(h, n)$. Значение 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. Пусть S_h – состоит из всех таких F. Каждый такой класс функций F может строго разделять некоторые множества $X \subseteq \mathcal{X}$ относительно некоторых последовательностей s. Пусть k_F – число всех таких возможных пар (X, s) строго разделимых классом F. Полагаем $t(h, n) = \min_{F \in \mathcal{S}_h} k_F$.

Эквивалентным образом, функцию $t(h, n)$ можно задаеть формальным условием:

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

Когда мы говорим, что F строго разделяет пару (X, s), мы имеем ввиду, что F строго разделяет пару X относительно последовательности s.

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

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

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 $< b^i$ 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

$$\sum_{i=0}^{d} \binom{n}{i} b^i = y$$

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

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

$$t\left(2(n(b+1)^2)^{\lceil \log y \rceil}\right) \geqslant y, \tag{1.33}$$

где $y = \sum\limits_{i=0}^{d} \binom{n}{i} b^i$.

Для того, чтобы доказать неравенство (1.33) предварительно докажем следующее утверждение.

Лемма 1.12.

$$t(2, n) \geqslant 1 \text{ } \textit{при } n \geqslant 1, \tag{1.34}$$

$$t(2mn(b+1)^2, n) \geqslant 2t(2m, n-1) \text{ } \textit{при } n \geqslant 2, m \geqslant 1. \tag{1.35}$$

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 (1.34) выполнено.

Для доказательства (1.35) 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 (1.35) автоматически выполнено. Разделим произвольным образом функции из F на пары $\{f, g\}$. Всего таких пар не менее чем $mn(b+1)^2$. Пусть P обозначает это множество пар. Для произвольной пары $\{f, g\} \in P$ 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

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

Общее число таких множеств не превосходит

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

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

Число все пар равно $mn(b+1)^2$. Отсюда должны существовать x^*, i^* и j^*, $j^* > i^*+1$ 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 размера $\geqslant t(2m, n-1)$, 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 $E \subseteq X$ и $f(x) \geqslant s_i + 1$ при $x \in E$, $f(x) \leqslant s_i - 1$ при $x \notin E$ для некоторого набора $s = (s_1, \ldots, s_{n-1})$. Аналогично, пусть $g(x) \geqslant s'_i + 1$ при $x \in E$, $f(x) \leqslant s'_i - 1$ при $x \notin E$ 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 (1.35) и лемма 1.12 доказаны. $\triangle$

Переходим теперь к доказательству леммы 1.10. Применяя неравенства (1.34) и (1.35) рекурсивным образом, получим

$$t(2(n(b+1)^2)^r, n) \geqslant 2^r t(2, n-r) \geqslant 2^r \tag{1.36}$$

при $n > r \geqslant 1$.

Если $\lceil \log y \rceil < n$, то полагаем $r = \lceil \log y \rceil$ в (1.36) 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, Сеза-Бьянки и Хаусслера [5].

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 (1.32) леммы 1.9, а также лемму 1.10, получим следующую цепочку неравенств

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

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

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

Из неравенства (1.31) леммы 1.9 получаем $d' \leqslant fat_{\alpha/4}(\mathcal{F}) = d$. Отсюда

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

В частности, $\log y \leqslant (ben/d)$.

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 мы введем еще одно определение емкости класса функций – среднее по Радемахеру. Это понятие позволяет получить верхнюю оценку ошибки обобщения.

При составлении данного раздела были использованы лекции Какаде и Тевари [17].

Пусть $z^l = (z_1, \ldots, z_l)$ – некоторая выборка. Пусть элементы выборки принадлежат некоторому множеству $\mathcal{X}$, на котором задано вероятностное распределение P. Мы предположим, что эле-

менты выборки независимо и одинаково распределены согласно распределению P.

Задан класс $\mathcal{F}$ равномерно ограниченных функций, определенных на множестве $\mathcal{X}$: $a \leqslant f(x) \leqslant b$ для всех $x \in \mathcal{X}$ и всех $f \in \mathcal{F}$, где $a < b$.

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 усовное 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 распределению 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 равна среднему значению выборочного среднего по Радемахеру относительно распределения P^l.

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

Пусть элементы выборки $z^l = (z_1, \ldots, z_l)$ независимо друг от друга генерируются с помощью вероятностного распределения P. По определению выборочное среднее функции f на выборке z^l

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

равно

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

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 $\bar{F}$.

Теорема 1.11. *Для произвольной функции $f \in \mathcal{F}$ имеет место неравенство*

$$E_{z^l \sim P^l}(\sup_{f\in\mathcal{F}}(E_P(f) - \tilde{E}(f(z^l)))) \leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.37}$$

Доказательство. Пусть $\tilde{z}^l = (\tilde{z}_1, \ \ldots, \ \tilde{z}_l)$ – случайные величины, распределенные также 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:

$$E_{z^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(E_P(f(z))-\frac{1}{l}\sum_{i=1}^{l}f(z_i)\right)\right)=$$

$$=E_{z^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}E_{\tilde{z}_i\sim P}(f(\tilde{z}_i))-f(z_i)\right)\right)\leqslant$$

$$\leqslant E_{z^l\sim P^l}\left(E_{\tilde{z}^l\sim P^l}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i))-f(z_i)\right)\right)\right)=$$

$$=E_{z^l\tilde{z}^l\sim P^{2l}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}(f(\tilde{z}_i))-f(z_i)\right)\right)=$$

$$=E_{z^l\tilde{z}^l\sim P^{2l}}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(\tilde{z}_i))-f(z_i)\right)\right)\leqslant$$

$$=E_{\tilde{z}^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(\tilde{z}_i)\right)\right)+$$

$$+E_{z^l\sim P^l}E_{\sigma\sim B_{1/2}}\left(\sup_{f\in\mathcal{F}}\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right)\right)=$$

$$=2\mathcal{R}_l(\mathcal{F}).\quad(1.38)$$

Переход от 2-й строки к 3-й происходит по свойству:

$$\sup_{f\in\mathcal{F}}\int f(\tilde{z})dP\leqslant\int\sup_{f\in\mathcal{F}}f(\tilde{z})dP,$$

которое в свою очередь следует из свойства: супремум суммы не превосходит суммы супремумов. Появление в 5-й строке σ_i не изменило супремум, так как иатематическое 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 (1.37) доказано. $\triangle$

Приведем два следствия из теоремы 1.11.

Во-первых, неравенство (1.37) можно обратить:

Следствие 1.5. *Для произвольной функции $f \in \mathcal{F}$ имеет место неравенство*

$$E_{P^l}(\sup_{f\in\mathcal{F}}(\tilde{E}(f(z^l))) - E_P(f))) \leqslant 2\mathcal{R}_l(\mathcal{F}). \tag{1.39}$$

Неравенство ()1.39 прямо следует из неравенства (1.37) 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*

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

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

$$P^l\{f(\tilde{z}_1, \ \ldots, \ \tilde{z}_l) - E_{P^l}(f(\tilde{z}_1, \ \ldots, \ \tilde{z}_l)) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right), \tag{1.40}$$

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

Доказательство этой леммы можно найти в работе [23] и в монографии [26].

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

$$P^l\{E_{P^l}(f(z_1,\ \ldots,\ z_l)) - f(z_1,\ \ldots,\ z_l) \geqslant t\} \leqslant \\ \leqslant \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{l} c_i^2}\right). \qquad (1.41)$$

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 $1-\delta$ для произвольной функции $f \in \mathcal{F}$ выполнено*

$$E_P(f(z)) \leqslant \hat{E}(f(z)) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2l}} \leqslant \\ \leqslant E_P(f(z)) \leqslant \hat{E}(f(z)) + 2\tilde{\mathcal{R}}_l(\mathcal{F}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}. \qquad (1.42)$$

Доказательство. Для заданной f имеет место очевидное неравенство

$$E_P(f(z)) \leqslant \hat{E}(f(z)) + \sup_{h\in\mathcal{F}}(E_P(h) - \tilde{E}(h)). \qquad (1.43)$$

Применим неравенство (1.40) леммы 1.13 ко второму члену (1.43).

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 (1.40) 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 (1.40) следует, что с вероятностью $1 - \delta/2$ выполнено

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}(h)) \leqslant E_{P^l}(\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}(h))) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.44)$$

Неравенство (1.37) утверждает, что

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

Отсюда и из (1.44) получаем

$$\sup_{h \in \mathcal{F}}(E_P(h) - \tilde{E}(h)) \leqslant 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.45)$$

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

$$E_P(f) \leqslant \tilde{E}(f)) + 2\mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}} \quad (1.46)$$

для любой функции $f \in \mathcal{F}$. Таким образом выполнено первое неравенство (1.42) следствия.

Аналогичным образом, с помощью неравенства (1.41) леммы 1.13 получаем, что с вероятностью $1 - \delta/2$ выполнено

$$\mathcal{R}_l(\mathcal{F}) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}. \quad (1.47)$$

Из неравенств (1.46) и (1.47) получаем, с вероятностью $1 - \delta$ выполнено второе неравенство (1.42) 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:

$$\tilde{\mathcal{R}}_l(\phi(\mathcal{F})) \leqslant L\tilde{\mathcal{R}}_l(\mathcal{F}), \tag{1.48}$$

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant L\mathcal{R}_l(\mathcal{F}). \tag{1.49}$$

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 случайныех величин со значениями из множества $\{+1, -1\}$, и пусть σ – соответствующее распределение на наборах этих величин.

Преобразования ниже 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 (1.48) и (1.49).

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

$$\mathcal{R}_l(\phi(\mathcal{F})) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi(f(z_i))\right). \tag{1.50}$$

Для простоты рассуждений предполагаем, что $L = 1$. [9] Нам необходимо доказать неравенство

$$\mathcal{R}_l(\phi(\mathcal{F})) \leqslant \mathcal{R}_l(\mathcal{F}) = E\left(\frac{1}{l}\sum_{i=1}^{l}\sigma_i f(z_i)\right). \tag{1.51}$$

Мы осуществим переход от (1.50) к (1.51) 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

[9] Можно заменить функцию ϕ на ϕ/L.

тождественную функцию: $\phi_i' = I$. При этом будет выполнена следующая цепочка неравенств:

$$
\begin{aligned}
&E(\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i(f(z_i))) = \\
&\frac{1}{2l}E(\sup_{f\in\mathcal{F}}(\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i))) + \\
&+\sup_{f\in\mathcal{F}}(-\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)))) = \\
&= \frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(\phi(f(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) - \\
&-\phi(f'(z_1)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) \leqslant \\
&\leqslant \frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(|f(z_1) - f'(z_1)| + \\
&+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) = \\
&= \frac{1}{2l}E(\sup_{f,f'\in\mathcal{F}}(f(z_1) - f'(z_1) + \\
&+\sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i)))) = \\
&= \frac{1}{2l}E(\sup_{f\in\mathcal{F}}(f(z_1) + \sum_{i=2}^{l}\sigma_i\phi_i(f(z_i)) + \\
&\sup_{f'\in\mathcal{F}}(-f'(z_1) + \sum_{i=2}^{l}\sigma_i\phi_i(f'(z_i))) = \\
&= E(\sup_{f\in\mathcal{F}}\frac{1}{l}\sum_{i=1}^{l}\sigma_i\phi_i'(f(z_i))), \qquad (1.52)
\end{aligned}
$$

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

тождественных функций чем набор $\phi_1, \ldots, \phi_l$.

В цепочке (1.52) 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 супремум достигается при неотрицательном значении разности $\phi(f(z_1)) - \phi(f'(z_1))$, поэтому можно заменить ее 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 (1.52) мы получим выражение

$$E(\sup_{f\in\mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma_i \phi_i'(f(z_i))), \tag{1.53}$$

в котором все ϕ_i' являются тождественными функциями, т.е. сумма (1.53) равна $\mathcal{R}_l(\mathcal{F})$.

Первая строка цепочки (1.52) 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 (1.48) и (1.49) 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*

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

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 логарифма. 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 строки этого неравенства и получаем неравенство

$$E\left(\sup_{a \in A} \sum_{i=1}^{l} \sigma_i a_i\right) \leqslant \frac{\ln |A|}{\lambda} + \frac{\lambda r^2}{2}. \tag{1.54}$$

Легко проверяется, что правая часть неравенства (1.54) достигает своего минимума при $\lambda = \sqrt{2 \ln |A| / r^2}$. Подставляем это значение λ в правую часть неравенства (1.54) 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*

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

для всех l.

Доказательство. Пусть $E = E_{P^l}$ и бинарная строка $a = (a_1, \ldots, a_l)$ представляет значения $(f(z_1), \ldots, f(z_l))$. Имеет ме-

сто следующая цепочка неравенств:

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

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*

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \inf_\alpha\left(\sqrt{\frac{2\ln\mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha\right). \tag{1.55}$$

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

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

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

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

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

$$+E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i(f(x_i)-g(x_i))\right). \quad (1.56)$$

Для среднего из последней строки (1.56) выполнено неравенство

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

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

$$\leqslant E_\sigma\left(\sup_{g\in B}\sup_{f\in B_\alpha(g)}\frac{1}{l}\sum_{i=1}^{l}\sigma_i|f(x_i)-g(x_i)|\right) \leqslant \alpha. \quad (1.57)$$

По лемме 1.14 получаем

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

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

$$\leqslant \sqrt{\frac{2\ln|B|}{l}} = \sqrt{\frac{2\ln\mathcal{N}(\alpha,\mathcal{F},x^l)}{l}}. \quad (1.58)$$

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

Соединяем вместе неравенства (1.57) и (1.58) и получаем неравенство

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \left(\sqrt{\frac{2 \ln \mathcal{N}(\alpha, \mathcal{F}, x^l)}{l}} + \alpha \right). \tag{1.59}$$

Так как неравенство (1.59) выполнено для любого $\alpha > 0$, оно выполнено и для нижней грани по $\alpha > 0$. Отсюда получаем неравенство (1.55). Теорема доказана. $\triangle$

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

Следствие 1.7.

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

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. Найти значения функции роста $B_H(3)$, $B_H(4)$, $B_H(5)$, $\ldots$, где

a) H – класс всех однородных классификаторов;

b) H – класс всех линейных классификаторов;

c) H – класс всех классификаторов, порожденных многочленами 2-го порядка, 3-го порядка и т.д.

4. Привести примеры классов функций – классификаторов, для которых VC-размерность равна ∞. Рассмотреть класс функций $\mathcal{F} = \{\text{sign}(\sin(tx)) : t \in \mathcal{R}\}$.

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

6. Доказать, что рекуррентное соотношение (1.24) имеет решение:

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

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 $(\bar{w} \cdot \bar{x}_i) - c = 0$, 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 $(\bar{w} \cdot \bar{x}) - c = 0$ (см. (2.1)).

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 частями 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 ближайшей к ней (сверху и снизу) точек выборки максимальна среди всех разделяющих S гиперплоскостей.

Лемма 2.2. *Оптимальная гиперплоскость – единственная гиперплоскость такая, что сумма расстояний от ближайшей 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 $|\bar{w}| \leqslant 1$ 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 $y_i = 1$:

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

Случай $y_i = -1$ разбирается аналогичным образом.

Отсюда получаем противоречие, так как вектор $\bar{w}_1^*$ имеет меньшую норму, чем норма вектора $\bar{w}_0^*$. 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) части 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. 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 ограничениям (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 полагаем $W(\bar{\alpha}) = L(\bar{w}, b, \bar{\alpha})$. С учетом (2.13) получим

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{x}_i \cdot \bar{x}_j). \tag{2.14}$$

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 $\bar{w}_0$

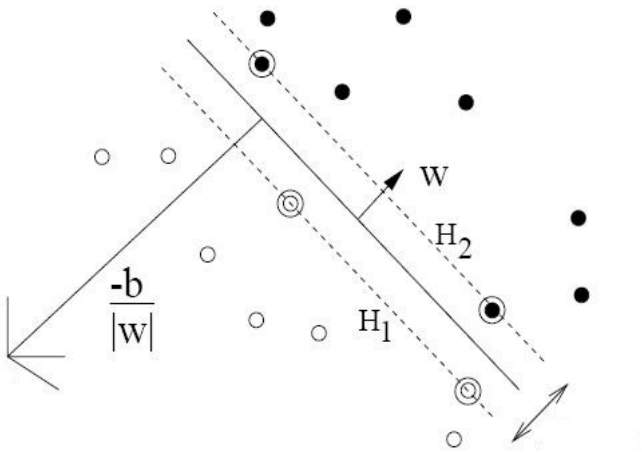

Рис. 1.1. Опорные векторы расположены на граничных гиперплоскостях H_1 и H_2

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, неопорные 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 индексов 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 $> \epsilon$, ограничена $(1-\epsilon)^{l-d} \leqslant \exp(-\epsilon(l-d))$. 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 $X \times \{-1, 1\}$, с P^l-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*

$$\operatorname{err}_P(h_{\hat{S}}) \leqslant \frac{1}{l-d}\left(d \ln\left(\frac{el}{d}\right) + \ln\left(\frac{l}{\delta}\right)\right).$$

Из этой теоремы следует, что при $d > 2$ и при достаточно больших l

$$\operatorname{err}_P(h_{\hat{S}}) \leqslant \frac{d \ln l}{l-d},$$

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 2-й степени от n переменных. Тогда можно рассмотреть следующую конструкцию.

Вводим новые переменные в пространстве признаков

$$\begin{aligned} z_1 = x_1,\ \ldots,\ z_n = x_n, \\ z_{n+1} = x_1^2,\ \ldots,\ z_{2n} = x_n^2, \\ z_{2n+1} = x_1 x_2,\ \ldots,\ z_N = x_n x_{n-1}. \end{aligned}$$

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 $\leqslant n$.

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 $\mathcal{R}^N$:

$$\begin{aligned}
&\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = \\
&= \sum_{j=1}^{N} \left(\sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \right) \phi_j(\bar{x}) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i \sum_{j=1}^{N} \phi_j(\bar{x}) \phi_j(\bar{x}_i) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}_i)) + b = \\
&= \sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}, \bar{x}_i) + b = 0, \qquad (2.23)
\end{aligned}$$

где

$$K(\bar{x}_i, \bar{x}) = (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x})). \qquad (2.24)$$

Таким образом, все рассуждения для «линейных» SVM-машин (оптимальных гиперплоскостей) в пространстве $\mathcal{R}^n$ годятся и для «нелинейных машин» 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:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) =$$

$$= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\bar{x}_i, \bar{x}_j) \tag{2.26}$$

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 $\mathcal{R}^{\frac{n(n+1)}{2}}$:

$$\bar{\phi}(\bar{x}) = \bar{\phi}(x_1, \ldots, x_n) =$$
$$= (1, x_1^2, \ \ldots, \ x_n^2, \sqrt{2}x_1x_2, \ \ldots, \ \sqrt{2}x_1x_2, \ \ldots, \ \sqrt{2}x_{n-1}x_n).$$

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

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i (\bar{x}_i \cdot \bar{x})^2 + b. \tag{2.28}$$

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 ядра 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, определеные 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 $\Xi^* = \cup_{n=0}^{\infty}\Xi^n$.

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:

$$K_n(s,t) = \sum_{u \in \Xi^n} (\phi_u^n(s) \cdot \phi_u^n(t)) =$$
$$= \sum_{u \in \Xi^n} \sum_{\bar{i}: u = s[\bar{i}]} \lambda^{l(\bar{i})} \sum_{\bar{j}: u = s[\bar{j}]} \lambda^{l(\bar{j})} =$$
$$= \sum_{u \in \Xi^n} \sum_{\bar{i}: u = s[\bar{i}]} \sum_{\bar{j}: u = s[\bar{j}]} \lambda^{l(\bar{i}) + l(\bar{j})}.$$

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. [10]).

Ядра такого типа используются при классификации текстов.

Более подробно о ядрах см. монографию Шолькопфа и Смолы [27].

2.5.1. 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: $K(\bar{x}, \bar{y}) = K(\bar{y}, \bar{x})$ для всех $\bar{x}$ и $\bar{y}$. Кроме этого, выполнено еще одно важное свойство: для любой последовательности элементов $\bar{x}_1, \ldots, \bar{x}_n$ и любой последовательности вещественных чисел $\alpha_1, \ldots, \alpha_n$ выполнено

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x}_j)) =$$
$$= \left(\sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \cdot \sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \right) = \left\| \sum_{i=1}^{n} \alpha_i \bar{\phi}(\bar{x}_i) \right\|^2 \geqslant 0. \qquad (2.29)$$

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 $x_1 , \ldots , x_n$ и любого набора вещественных чисел $\alpha_1 , \ldots , \alpha_n$ выполнено

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

Согласно (2.29) функция $K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}))$ является положительно определенной.

Матрица $(K(x_i, x_j)_{i,j=1}^{n}$ называется матрицей Грама.

Гильбертово пространство порожденное воспроизводящим ядром

По каждой симметричной положительно определенной функции $K(x, y)$ определим некоторое каноническое функциональное гильбертово пространство $\mathcal{F}$. Определим отображение $\Phi : X \to \mathcal{R}^X$ из множества X в множество всех функций из X в $\mathcal{R}$:

$$\Phi(x) = K(\cdot, x) = K_x.$$

По определению $\phi(x)$ это функция, для которой: $K_x(y) = K(x, y)$ для всех y. Определим пространство 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 и g. Она также является положительно определенной.

Предварительно заметим, что

$$(f \cdot f) = \sum_{i=1}^{n} \sum_{i,j=1}^{n'} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0.$$

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 $(K_x \cdot f) = f(x)$ и, в частности, $(K_x \cdot K_y) = K(x, y)$ для всех x и y. Из-за этих свойств симметричная положительно определенная функция $K(x, y)$ называется *воспроизводящим ядром.*

Из этих свойств и из свойства (2.30) также следует, что

$$|f(x)|^2 = |(f \cdot K_x)|^2 \leqslant K(x, x)(f \cdot f).$$

В частности, из $(f \cdot f) = 0$ следует, что $f(x) = 0$ для всех x.

Функция $\|f\| = \sqrt{(f \cdot f)}$ является нормой, так как она определена по скалярному произведению. Рассмотрим пополнение множества всех линейных комбинаций (2.30) относительно этой нормы до полного метрического пространства $\mathcal{F}$. Полученное гильбертово пространство называется *гильбертовым пространством порожденным воспроизводящим ядром* (Reproducing Kernel Hilbert Space – RKHS).

Другой вариант определения RKHS заключается в следующем. RKHS – это 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 $K(x, y) = (K_x \cdot K_y)$.

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.

Заметим без доказательства, что хотя ядро, определенное по некоторому отображению ϕ с помощью равенства (2.27), является симметричным и положительно определенным, восстановленное по нему отображение в каноническое гильбертово пространство $\mathcal{F}$ может не совпадать с исходным отображением ϕ. С другой стороны, каждому симметричному положительно определенному ядру соотсетствует единственное каноническое гильбертово пространство RKHS [6].

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 Киммельдорфом и Вахбой [19]. См. также [27].

Теорема 2.3. *Пусть X – некоторое множество объектов и $S = ((x_1, y_1)\ , \ldots,\ (x_l, y_j))$ – 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 единственное каноническое гильбертово пространство RKHS с нормой $\|\cdot\|$.*

Заданы также функция потерь $c : (X^2 \times \mathcal{R})^l \to \mathcal{R} \cup \{\infty\}$ и некоторая строго монотонно возрастающая функция Ω на множестве всех неотрицательных вещественных чисел.

Тогда любая функция $f \in \mathcal{F}$, минимизирующая регуляризованный риск функционал

$$c((x_1, y_1, f(x_1))\ , \ldots,\ (x_l, y_j, f(x_l))) + \Omega(\|f\|) \tag{2.32}$$

может быть представлена в виде

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x), \tag{2.33}$$

для некоторых чисел $\alpha_1 \ , \ldots , \ \alpha_l$.

Пример такого риск функционала для задачи регрессии в пространстве признаков $f \in \mathcal{F}$:

$$c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_j, f(x_l))) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda \|f\|^2,$$

где $\lambda > 0$.

Доказательство. Напомним, что $K_{x_i} = K(x_i, \cdot)$ – 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 $c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_j, f(x_l)))$ регуляризованного функционала (2.32) не зависит от f_*.

Так как f_* ортогонально элементу $\sum\limits_{i=1}^{l} \alpha_i K_{x_i}$ и функция Ω является 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 (2.32) должно быть $f_* = 0$. Отсюда решение задачи минимизации функционала (2.32) должно иметь вид (2.33):

$$f(x) = \sum_{i=1}^{l} \alpha_i K_{x_i}.$$

Теорема доказана. $\triangle$

Теорема 2.3 показывает, что для решения задачи (2.32) 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:

$$c((x_1, y_1, f(x_1)) \ , \ldots , \ (x_l, y_j, f(x_l))) =$$

$$= \frac{1}{\lambda} \sum_{i=1}^{l} \max\{0, 1 - y_i f(x_i)) + \|f\|^2,$$

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 (2.32), имеет вид

$$f = \sum_{i=1}^{l} \alpha_i K_{x_i}.$$

2.6. Случай неразделимой выборки

Предварительно получим 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

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

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

$$\xi_i = \max\{0, \gamma - y_i f(\bar{x}_i)\}.$$

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.

$\xi_i = 0$ 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 наличие шума в исходных данных.

В целом норма вектора ошибок $\bar{\xi}$ отражает размер ошибок классификации, а такжe 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-Тэйлору и Кристианини [25] представлен в следующей ниже теореме.

Теорема 2.4. *Пусть $\mathcal{L}$ – класс всех линейных функций вида $L(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$ с единичным весовым вектором $\|\bar{w}\| = 1$. Пусть $\gamma > 0$. Тогда существует константа c такая, что для произвольного распределения вероятностей P на $\mathcal{X} \times \{-1, 1\}$ с носителем внутри шара радиуса R и с центром в начале координат с P^l-вероятностью $1 - \delta$ произвольная функция $f \in \mathcal{L}$ имеет на*

случайной выборке S длины l вероятность ошибки [1] :

$$\mathrm{err}_P(f) = P\{yf(\bar{x}) < 0\} \leqslant \frac{c}{l}\left(\frac{R^2 + \|\bar{\xi}\|^2}{\gamma^2}\ln^2 l + \ln\frac{1}{\delta}\right), \quad (2.34)$$

где $\bar{\xi}$ – вектор переменных мягкого отступа относительно гиперплоскости L и порога $\gamma > 0$, а $l \geqslant \frac{64(R+\|\xi\|)^2}{\gamma^2}$.

Доказательство. Рассмотрим линейный классификатор $f(\bar{x}) = (\bar{w}\cdot\bar{x}) + b$, где $\|\bar{w}\| = 1$. Из определения переменной мягкого отступа $\bar{\xi} = (\xi_1, \ldots, \xi_l)$, определенной для этого классификатора и выборки $S = (\bar{x}_1, \ldots, \bar{x}_l)$, будет

$$y_i f(\bar{x}_i) + \xi_i \geqslant \gamma \quad (2.35)$$

при $i = 1, \ldots, l$.

Пусть $\nu > 0$ – параметр, значение которого мы оптимизируем позже. Заменим векторы обучающей выборки $\bar{x}_1, \ldots, \bar{x}_l$ размерности n на вспомогательные векторы $\bar{x}'_1, \ldots, \bar{x}'_l$ размерности $n+l$, которые определяются следующим образом:

$$\bar{x}'_i = (x_{i,1}, \ldots, x_{i,n}, 0, \ldots, \nu, \ldots, 0),$$

при $i = 1, \ldots, l$, где $(n+i)$-я координата вектора $\bar{x}'_i$ равна ν, а остальные дополнительные координаты равны 0. Полученную выборку обозначим $S' = ((\bar{x}'_1, y_1) \ldots, (\bar{x}'_l, y_l))$.

Гиперплоскость $f(\bar{x}) = (\bar{w}\cdot\bar{x}) + b$ заменяем на гиперплоскость

$$f'(\bar{x}') = (\bar{w}'\cdot\bar{x}') + b, \text{ где}$$
$$\bar{w}' = (w_1, \ldots, w_n, \frac{1}{\nu}y_1\xi_1, \ldots, \frac{1}{\nu}y_l\xi_l), \quad (2.36)$$

а $\bar{x}'$ – произвольный вектор размерности $n + l$.

Из условия (2.35) следует, что новая выборка S' оказвается разделенной новым классификатором (2.36) с порогом γ :

$$y_i((\bar{w}'\cdot\bar{x}'_i) + b) = y_i((\bar{w}\cdot\bar{x}_i) + b) + (y_i)^2\xi_i \geqslant \gamma \quad (2.37)$$

[1] Имеется в виду шар в пространстве $\mathcal{R}^n$, которому принадлежат классифицируемые элементы $\bar{x}_1, \ldots, \bar{x}_l$ выборки S.

при $i = 1,\ \ldots,\ l$.

Для того, чтобы применить к новой выборке и новому классификатору теорему 1.9 из раздела 1.3, необходимо нормировать направляющий вектор гиперплоскости (2.36). Имеет место равенство

$$\|\bar{w}'\|^2 = \|\bar{w}\| + \frac{1}{\nu^2}\|\bar{\xi}\|^2 = 1 + \frac{1}{\nu^2}\|\bar{\xi}\|^2.$$

Кроме того, все векторы $\bar{x}_i'$ содержатся в шаре радиуса R', где $R'^2 = R^2 + \nu^2$.

После нормировки условие (2.37) превращается в условие

$$y_i\left(\left(\frac{\bar{w}'}{\|\bar{w}'\|}\cdot\bar{x}_i'\right) + b\right) \geqslant \gamma' = \frac{\gamma}{\|\bar{w}'\|}.$$

при $i = 1,\ \ldots,\ l$. Отсюда следует, что главный множитель из оценки следствия 1.4 имеет вид

$$\frac{R'^2}{\gamma'^2} = \frac{(R^2+\nu^2)(1+\frac{1}{\nu^2}\|\xi\|^2)}{\gamma^2}.$$

Преобразуем

$$(R^2+\nu^2)(1+\frac{1}{\nu^2}\|\xi\|^2) = R^2 + \|\xi\|^2 + \nu^2 + \frac{1}{\nu^2}R^2\|\xi\|^2.$$

Минимум этого выражения достигается при $\nu^2 = R\|\xi\|$, а само выражение приобретает вид

$$R^2 + 2R\|\xi\| + \|\xi\|^2 = (R+\|\xi\|)^2 \leqslant 2(R^2+\|\xi\|^2).$$

Применяя теорему 1.9 из раздела 1.3, получаем оценку (2.34) при $l \geqslant \frac{64(R+\|\xi\|)^2}{\gamma^2}$. Теорема доказана. $\triangle$

2.6.2. Оптимизационные задачи для классификации с ошибками

Случай квадратичной нормы

В случае, когда выборка $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ неразделима, рассматривается задача оптимизации с переменными мягкого отступа ξ_i, $i = 1,\ \ldots,\ l$.

Найдем векторы $\bar{w}$, $\bar{\xi}$ и число b, так чтобы

$$(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i^2 \to \min, \tag{2.38}$$

$$y_i((\bar{w} \cdot \bar{x}_i) + b) \geqslant 1 - \xi_i, \tag{2.39}$$

$$\xi_i \geqslant 0 \tag{2.40}$$

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 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 (2.38) – (2.40) имеет вид

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 - \\ - \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i), \tag{2.41}$$

где $\alpha_i \geqslant 0$ – множители Лагранжа.

Соответствующая двойственная задача формулируется путем дифференцирования лагранжиана

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0}, \tag{2.42}$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{\xi}} = C\bar{\xi} - \bar{\alpha} = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0,$$

а также подстановкой этих соотношений в (2.41) :

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) + \\ + \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}) - \frac{1}{C}(\bar{\alpha} \cdot \bar{\alpha}) = \\ = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) - \frac{1}{2C}(\bar{\alpha} \cdot \bar{\alpha}). \qquad (2.43)$$

Таким образом, мы должны максимизировать по $\bar{\alpha}$ величину

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j ((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C}\delta_{ij}) \qquad (2.44)$$

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 (2.42) 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 $y_i((\bar{w} \cdot \bar{x}_i) + b) \leqslant 1$, при этом $\alpha_i \geqslant 0$ и $\xi_i \geqslant 0$. Это те векторы, которые лежат на граничных гиперплоскостях или же неправильно ими классифицируются, в этом случае $y_i((\bar{w} \cdot \bar{x}_i) + b) < 1$ и $\xi_i > 0$.

Сформулируем задачу оптимизации для пространства признаков, заданного некоторым ядром $K(\bar{x}_i, \bar{x}_j)$.

Теорема 2.5. *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:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (K(\bar{x}_i, \bar{x}_j) + \frac{1}{C}\delta_{ij}) \to \max \tag{2.45}$$

$$\text{при условиях } \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$\alpha_i \geqslant 0 \ i = 1, \ \ldots, \ l. \tag{2.46}$$

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 $\operatorname{sign}(f(\bar{x}))$ *разделяет элементы выборки так же, как соответствующая гиперплоскость, полученная в результате решения задачи оптимизации* (2.38) – (2.40) *в пространстве признаков, определенном ядром* $K(\bar{x}, \bar{z})$, *где переменные отступа определяются для границы 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 $\alpha_i = C\xi_i$, также условия Каруша–Куна–Таккера:

$$\alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

при $i = 1, \ \ldots, \ l$.

Величина $\rho(\bar{w}) = \frac{1}{|\bar{w}|}$, определяющая расстояние между гиперплоскостями $(\bar{w} \cdot \bar{x}_i) + b) = \pm 1$ (в ядерном случае) определяется следующим образом:

$$
\begin{aligned}
(\bar{w} \cdot \bar{w}) &= \sum_{i,j=1}^{l} y_i y_j \alpha_i^* \alpha_j^* K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} y_j \alpha_j^* \sum_{i \in sv} y_i \alpha_i^* K(\bar{x}_i, \bar{x}_j) = \\
&= \sum_{j \in sv} \alpha_j^* (1 - \xi_j^* - y_j b^*) = \\
&= \sum_{j \in sv} \alpha_j^* - \sum_{j \in sv} \alpha_j^* \xi_j^* = \\
&= \sum_{j \in sv} \alpha_j^* - \frac{1}{C} (\bar{\alpha}^* \cdot \bar{\alpha}^*).
\end{aligned}
$$

В (2.45) можно заменить ядро $K(\bar{x}_i, \bar{x}_j)$ на

$$K'(\bar{x}_i, \bar{x}_j) = K(\bar{x}_i, \bar{x}_j) + \frac{1}{C} \delta_{\bar{x}}(\bar{y})$$

и далее использовать методы построения оптимальной гиперплоскости, приведенные выше.

Верхняя оценка (2.34) вероятности ошибки классификации при обобщении не зависит от размерности пространства, что позволяет применять ядра $K(\bar{x}, \bar{z})$, порождающие пространства признаков 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 2.4. Оценка вероятности ошибки (2.34) имеет место для пороговых линейных функций с единичным весовым вектором $\bar{w}$. Для того чтобы ее применить к задаче (2.38) – (2.40), поделим обе части неравенства (2.39) на $\|\bar{w}\|$, где $\bar{w}$ – оптимальное решение задачи. Тогда в (2.34) надо взять в качестве ξ_i величину $\xi_i / \|\bar{w}\|$, а $\gamma = 1/\|\bar{w}\|$. Получим новую версию вероятности ошибки

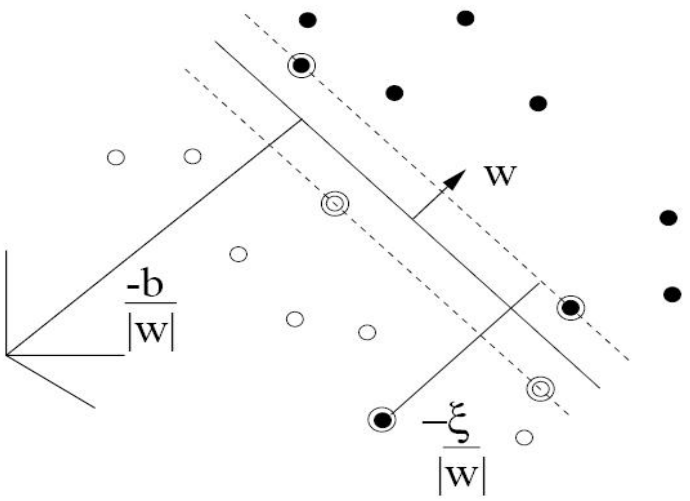

Рис. 1.2. Опорные векторы расположены на граничных гиперплоскостях или же неправильно ими классифицируются

(2.34) :

$$\begin{aligned} \mathrm{err}_P(f) = P\{yf(\bar{x}) < 0\} \leqslant \\ \leqslant \frac{c}{l}((\|\bar{w}\|^2 R^2 + \|\bar{\xi}\|^2)\ln^2 l + \ln\frac{1}{\delta}). \end{aligned} \tag{2.47}$$

Неравенство (2.47) показывает, что для минимизации верхней оценки вероятности ошибки обобщения нам действительно необходимо минимизировать величину (2.38).

Случай линейной нормы

На практике также часто рассматривается аналогичная задача оптимизации, в которой вместо квадратичной нормы вектора переменных мягкого отступа $\bar{\xi}$ используется линейная норма. В этом случае возникает следующая задача оптимизации.

Находим векторы $\bar{w}$, $\bar{\xi}$ и число b, так чтобы

$$(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i \to \min, \tag{2.48}$$

$$y_i((\bar{w} \cdot \bar{x}_i) + b_0) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0 \tag{2.49}$$

при $i = 1, \ \ldots, \ l$. Константа C определяет баланс между двумя частями функционала.

Соответствующий лагранжиан имеет вид

$$L(\bar{w}, b, \xi, \alpha, \bar{r}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i -$$

$$- \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b_0) - 1 + \xi_i) - \sum_{i=1}^{l} r_i \xi_i,$$

где $\alpha_i \geqslant 0$, $r_i \geqslant 0$ при $i = 1, \ \ldots, \ l$.

Соответственная двойственная задача получается путем приравнивания к нулю производных:

$$\frac{\partial L(\bar{w}, b, \xi, \alpha, \bar{r})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \xi, \alpha, \bar{r})}{\partial \xi_i} = C - \alpha_i - r_i = 0,$$

$$\frac{\partial L(\bar{w}, b, \xi, \alpha, \bar{r})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0.$$

Подставляем решения этих уравнений в прямую задачу и получаем двойственное представление задачи в виде задачи максимизации функционала:

$$L(\bar{w}, b, \xi, \alpha, \bar{r}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j),$$

который почти совпадает с функционалом для случая квадратичной нормы переменных мягкого отступа.

Единственное 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

$$\alpha_i(y_i((\bar{x}_i \cdot \bar{x}_j) + b) - 1 + \xi_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$\xi_i(\alpha_i - C) = 0,\ i = 1,\ \ldots,\ l.$$

Согласно этим условиям переменная мягкого отступа ξ_i отлична от нуля только при $\alpha_i = C$.

Опорные векторы – это те векторы $\bar{x}_i$, где $\alpha_i > 0$ (в таком случае $\alpha_i = C$). Для них выполнено условие $y_i((\bar{x}_i \cdot \bar{x}_j) + b) \leqslant 1$ и $\xi_i \geqslant 0$. Это те векторы, которые лежат на границах гиперплоскостей или неправильно ими классифицируются. Легко видеть, что расстояние от такого вектора до соответствующей разделяющей гиперплоскости равно $-\frac{\xi_i}{\|\bar{w}\|}$ (см. рис. 1.2).

Для произвольного ядра получаем следующее утверждение.

Теорема 2.6. *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:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) \to \max \qquad (2.50)$$

$$\textit{при условиях } \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$C \geqslant \alpha_i \geqslant 0\ i = 1,\ \ldots,\ l. \qquad (2.51)$$

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* $C > \alpha_i^* > 0$.

Тогда функция классификации $\mathrm{sign}(f(\bar{x}))$ *разделяет элементы выборки так же, как соответствующая гиперплоскость, полученная в результате решения задачи оптимизации* (2.48) – (2.49) *в пространстве признаков, определенном ядром* $K(\bar{x}, \bar{z})$*, где переменные мягкого отступа определены для границы ошибки:*

$$\gamma = \left(\sum_{i,j \in sv} y_i y_j \alpha_i^* \alpha_j^* K(\bar{x}_i, \bar{x}_j) \right)^{-1/2}.$$

Таким образом, задача оптимизации (2.48) – (2.49) эквивалентна задаче оптимизации (2.38) – (2.40) 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 линейной в пространстве признаков, заданном некоторым ядром. [2] Напомним основные понятия теории пороговой классификации. Задан класс $\mathcal{F}$ функций типа $\mathcal{X} \to \mathcal{R}$, с помощью которых производится классификация. В наших приложениях область определения $\mathcal{X}$ функций из $\mathcal{F}$ является подмножеством $\mathcal{R}^n$. По каждой функции $f \in \mathcal{F}$ определим индикаторную функцию классификации

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

В наших приложениях $\bar{x}$ – вектор из $\mathcal{R}^n$.

Задана обучающая выборка $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, где $\bar{x}_i \in \mathcal{R}^n$ и $y_i \in \{-1, 1\}$.

[2] Материал данного раздела использует результаты из монографии Шо-Тэйлора и Кристианини [26].

Задано отображение $\bar{\phi} : \mathcal{R}^n \to \mathcal{R}^N$ заданное ядром $K(\bar{x}, \bar{y})$. Пусть $\mathcal{F}$ – класс всех функций линейных в пространстве признаков $\mathcal{R}^N$ с ограниченным весовым вектором. Каждая функция из класса $\mathcal{F}$ имеет вид $f(\bar{x}) = (\bar{w} \cdot \bar{\phi}(\bar{x})) + b$, где $\bar{w}, \bar{\phi}(\bar{x}) \in \mathcal{R}^N$ и $\|w\| \leqslant 1$. Для простоты полагаем $b = 0$.

Пусть $\mathbf{K} = (K(\bar{x}_i, \bar{x}_j))_{i,j=1}^n$ – матрица значений ядра на элементах для выборки S; $tr(\mathbf{K}) = \sum_{i=1}^l K(\bar{x}_i, \bar{x}_i)$ – след матрицы $\mathbf{K}$.

Приведем оценку выборочного среднего Радемахера для класса $\mathcal{F}$ относительно обучающей выборки S.

Теорема 2.7. *Выборочное среднее Радемахера класса $\mathcal{F}$ относительно обучающей выборки $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ удовлетворяет неравенству*

$$\tilde{\mathcal{R}}_l(\mathcal{F}) \leqslant \frac{1}{l}\sqrt{tr(\mathbf{K})}. \tag{2.52}$$

Доказательство. Имеет место следующая цепочка равенств и неравенств:

$$\begin{aligned}
\tilde{\mathcal{R}}_l(\mathcal{F}) = E_\sigma \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{l} \sum_{i=1}^l \sigma_i f(\bar{x}_i) \right| \right) = \\
= E_\sigma \left(\sup_{\|\bar{w}\| \leqslant 1} \left| \left(\bar{w} \cdot \frac{1}{l} \sum_{i=1}^l \sigma_i \bar{\phi}(\bar{x}_i) \right) \right| \right) \leqslant \\
\leqslant \frac{1}{l} E_\sigma \left(\left\| \sum_{i=1}^l \sigma_i \bar{\phi}(\bar{x}_i) \right\| \right) = \\
= \frac{1}{l} E_\sigma \left(\left(\sum_{i=1}^l \sigma_i \bar{\phi}(\bar{x}_i) \cdot \sum_{i=1}^l \sigma_i \bar{\phi}(\bar{x}_i) \right)^{1/2} \right) \leqslant \\
\leqslant \frac{1}{l} \left(E_\sigma \left(\sum_{i,j=1} \sigma_i \sigma_j K(\bar{x}_i, \bar{x}_j) \right) \right)^{1/2} = \\
= \frac{1}{l} \left(\sum_{i=1}^l K(\bar{x}_i, \bar{x}_i) \right)^{1/2}.
\end{aligned}$$

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 $E(\sigma_i) = 0$ и $E(\sigma_i\sigma_j) = E(\sigma_i)E(\sigma_j) = 0$ при $i \neq j$. Теорема доказана. $\triangle$

Напомним, что число $\gamma_i = y_i f(\bar{x}_i)$ называется границей 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.

Задана выборка $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$. Пусть задано число $\gamma > 0$. Переменная мягкого отступа примера $(\bar{x}_i, y_i)$ для пороговой функции f и границы ошибки γ определяется как

$$\xi_i = (\gamma - y_i f(\bar{x}_i))_+. \tag{2.53}$$

Здесь $(x)_+ = \max\{0, x\}$. Напомним, 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 $S = ((\bar{x}_1, y_1), \ldots, (\bar{x}_l, y_l))$.

Определим вспомогательную функцию $f(\bar{x}, y) = -yf(\bar{x})$. Пусть

$$\chi(x) = \begin{cases} 1, \text{ если } x > 0, \\ 0 \text{ в противном случае.} \end{cases}$$

Предполагаем, что элементы выборки S генерируются независимо друг от друга с помощью вероятностного распределения P. Легко проверить, что

$$P^l\{(\bar{x}, y) : y \neq \operatorname{sign}(f(\bar{x}))\} = E_P(\chi(-yf(\bar{x}))).$$

Пусть $\mathbf{K} = (K(\bar{x}_i, \bar{x}_j))_{i,j=1}^n$ – матрица значений ядра на элементах для выборки S.

В следующей теореме дается верхняя оценка ошибки обобщения классификатора линейного в пространстве признаков, заданного ядром K.

Теорема 2.8. *Для произвольного $\delta > 0$ с вероятностью $1 - \delta$ выполнено*

$$P^l\{y \neq \operatorname{sign}(f(\bar{x}))\} \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i + \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}. \quad (2.54)$$

Заметим, что правая часть неравенства (2.54) является случайной величиной. Матрица $\mathbf{K}$ построена по случайной выборке S. Переменные мягкого отступа ξ_i также зависят от элементов выборки $\bar{x}_i$ и поэтому также являются случайными величинами.

Доказательство. Напомним, что $\gamma > 0$ – порог функции классификации. Определим вспомогательную функцию $g : \mathcal{R} \to [0, 1]$:

$$g(r) = \begin{cases} 1, \text{ если } r > 0, \\ 1 + r/\gamma, \text{ если } -\gamma \leqslant r \leqslant 0, \\ 0 \text{ в противном случае.} \end{cases}$$

Из определения этой функции следует, что $g(r) \geqslant \chi(r)$. Отсюда и по следствию 1.6 с вероятностью $1 - \delta$ выполнено

$$E_P(\chi(f(\bar{x}, y))) \leqslant E_P(g(f(\bar{x}, y))) \leqslant$$
$$\leqslant \tilde{E}_P(g(f(\bar{x}, y))) + 2\tilde{\mathcal{R}}_l(g \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \quad (2.55)$$

По определению переменной мягкого отступа (2.53) выполнено

$$g(-y_i f(\bar{x}_i)) \leqslant \xi_i/\gamma$$

при $1 \leqslant i \leqslant l$.

По теореме 1.12, где $L = 1/\gamma$, имеем

$$\tilde{\mathcal{R}}_l(g \circ \mathcal{F}) \leqslant \tilde{\mathcal{R}}_l(\mathcal{F})/\gamma.$$

Из неравенства (2.55) и неравенства (2.52) теоремы 2.7 следует, что с вероятностью $1 - \delta$,

$$E_P(\chi(f(\bar{x}, y))) \leqslant \frac{1}{l\gamma}\sum_{i=1}^{l}\xi_i + \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \quad (2.56)$$

Теорема доказана. $\triangle$

В частности, если функция $f(\bar{x})$ разделяет выборку S без ошибок, то имеет место оценка:

Следствие 2.1. *Допустим, что функция $f(\bar{x})$ разделяет выборку S без ошибок, а также выполнены все приведенные выше предположения. Тогда для произвольного $\delta > 0$ с вероятностью $1 - \delta$ выполнено*

$$P^l\{y \neq \operatorname{sign}(f(\bar{x}))\} \leqslant \frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2l}}.$$

Оценка (2.54) по порядку уступает аналогичной оценке, полученной с помощью понятия пороговой размерности. Пусть $\|\bar{x}_i\| \leqslant R$ для всех $1 \leqslant i \leqslant l$. Для малых своих значений величина

$$\frac{2}{l\gamma}\sqrt{tr(\mathbf{K})} \leqslant \frac{2}{l\gamma}\sqrt{lR^2} = 2\sqrt{\frac{R^2}{l\gamma^2}}$$

по порядку больше главного члена оценки (1.28) теоремы 1.9 или главного члена оценки (2.34) теоремы 2.4, имеющих порядок $O\left(\frac{R^2}{l\gamma^2}\right)$. Оценки (1.28) и (2.34) были получена в рамках теории пороговой размерности классов функций.

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 $i = 1,\ \ldots,\ l$.

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 рассматриваться 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 $\bar{x}'\bar{z}$:

$$\bar{x}'\bar{z} = (x_1, \ \ldots, \ x_n) \begin{pmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{pmatrix} = (x_1 z_1 + \ldots x_n z_n)$$

типа (1×1) с единственным элементом, равным этому скалярному произведению.

Согласно методу наименьших квадратов, минимизируем *квадратичную функцию потерь*:

$$L(\bar{w}, b) = \sum_{i=1}^{l} (y_i - (\bar{w} \cdot \bar{x}_i) - b)^2. \tag{2.57}$$

Обозначим $\tilde{w}$ – расширенный вектор-столбец весовых коэффици-

ентов и свободного члена

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix}.$$

Аналогичным образом, обозначим $\tilde{x}$ – 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:

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}). \tag{2.58}$$

Рассмотрим матрицу типа $(l \times (n+1))$, строками которой являются расширенные векторы-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 (2.57) можно записать в матричном виде как квадрат нормы вектора-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:

$$L(\tilde{w}) = \|\tilde{X}\tilde{w} - \bar{y}\|^2 \to \min. \tag{2.59}$$

Геометрически это может интерпретироваться так же, как поиск проекции наименьшей длины вектора $\bar{y}$ на подпространство (гиперплоскость), порожденное векторами – 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, что избавиться от свободного члена в уравнении регрессии, мы рассматриваем задачу регрессии с весовой переменной $\tilde{w}$ – расширенный вектор-столбец весовых коэффициентов и свободного члена

$$\tilde{w} = \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \\ b \end{pmatrix},$$

а также $\tilde{x}$ – расширенный вектор-столбец 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:

$$L(\tilde{w}) = \lambda(\tilde{w} \cdot \tilde{w}) + \sum_{i=1}^{l}(y_i - (\tilde{w} \cdot \tilde{x}_i))^2 = \\ = \lambda\|\tilde{w}\|^2 + \|\tilde{X}\tilde{w} - \bar{y}\|^2. \qquad (2.60)$$

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 2.9.3.

2.9. Регрессия с опорными векторами

2.9.1. Ошибка обобщения при регрессии

Задача регрессии заключается в нахождении функции $f(\bar{x})$, которая аппроксимирует отображение из некоторого подмножества $\mathcal{X} \subseteq \mathcal{R}^n$ в множество действительных чисел $\mathcal{R}$. Данная функция определяется по обучающей выборке. Соответствующая теория обобщения дает вероятность ошибки аппроксимации на тестовой выборке.

Пусть $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$ – обучающая выборка, $\bar{x}_i \in \mathcal{X}$ и $y_i \in \mathcal{R}$.

Напомним, что разности $|f(\bar{x}_i) - y_i|$ называются *остатками*. *Функция потерь* сводит эти остатки в одну величину, на основе которой вычисляется качество аппроксимации. Пример функции потерь – квадратичная функция потерь $\sum\limits_{i=1}^{l}(f(\bar{x}_i) - y_i)^2$.

Задача оценки ошибки обобщения для регрессии с помощью функции $f(\bar{x})$ сводится к задаче оценки ошибки обобщения неко-

торой искусственной задачи классификации. В этом случае можно применить результаты раздела 1.3.

Пусть задано число $\theta > 0$ – точность регрессии. Тогда функция $g(\bar{x}, y) = \theta - |f(\bar{x}) - y|$ определяет способ решения некоторой задачи классификации пар $(\bar{x}, y)$: условие $g(\bar{x}, y) > 0$ означает, что пара $(\bar{x}, y)$ классифицируется функцией g как положительный пример; в противном случае она классифицируется как отрицательный пример.

Для получения оценки вероятности ошибки обобщения, не зависящей от размерности пространства, вводим число γ – нижнюю границу ошибки соответствующей задачи классификации, $0 < \gamma < \theta$. По определению условие $g(\bar{x}, y) \geqslant \gamma$ эквивалентно условию $|f(\bar{x}) - y| \leqslant \theta - \gamma$.

Теперь теорема 1.9 может быть переформулирована для случая многомерной регрессии. Соответствующая теорема 2.9 имеет место для случая, когда при обучении требуется точное попадание данных выборки в слой размера $\theta - \gamma$ вокруг гиперплоскости (см. [10], теорема 4.26).

Теорема 2.9. *Рассмотрим задачу регрессии с помощью линейных функций* $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b \in \mathcal{L}$, *где* $\bar{x} \in \mathcal{R}^n$, $|\bar{w}| = 1$.

Пусть $0 < \gamma < \theta \in \mathcal{R}$ *и* P *– распределение вероятностей, сконцентрированное в шаре радиуса* R *с центром в начале координат генерирующее выборку* $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$.

Тогда с вероятностью $1-\delta$ *произвольная гипотеза* $f \in \mathcal{L}$, *для которой* $|f(\bar{x}_i) - y_i| \leqslant \theta - \gamma$ *для всех* $i = 1, \ \ldots, \ l$, *имеет верхнюю оценку вероятности ошибки:*

$$\begin{gathered} \operatorname{err}_P(f) = P\{|f(\bar{x}) - y| \geqslant \theta\} \leqslant \\ \leqslant \frac{4}{l}\left(\frac{64R^2}{\gamma^2}\ln\frac{el\gamma}{8R^2}\ln\frac{32l}{\gamma^2} + \ln\frac{4}{\delta}\right) \end{gathered}$$

при $l > \frac{64R^2}{\gamma^2}$.

Другая оценка вероятности ошибки дается для случая переменных мягкого отступа:

$$\xi_i = \max\{0, |y_i - f(\bar{x}_i)| - (\theta - \gamma)\},$$

где θ – заданная точность приближения, γ – потери на границе.

Пусть $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ – вектор переменных мягкого отступа для обучающей выборки $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$.

Роль вектора переменных мягкого отступа надо понимать следующим образом.

Если в результате обучения $\xi_i > 0$, *то*

$$\xi_i = |y_i - f(\bar{x}_i)| - (\theta - \gamma),$$

т.е. эта величина показывает, насколько в результате обучения i*-й остаток вышел из слоя размера* $\epsilon = \theta - \gamma$, *проходящего вокруг значений функции регрессии.* [3] *Величина* ξ_i *тем больше, чем больше ошибка регрессии на примере* $(\bar{x}_i, y_i)$.

Если $|y_i - f(\bar{x}_i)| \leqslant \epsilon$, *то* $\xi_i = 0$, *и мы считаем, что ошибки регрессии на примере нет; значение исходного данного* y_i *лежит в слое размера* ϵ, *расположенном вокруг значений функции регрессии на обучающей выборке.*

Норма $\|\bar{\xi}\| = \sqrt{\sum_{i=1}^{l} \xi_i^2}$ вектора переменных мягкого отступа отражает величину всех остатков регрессии и поэтому должна минимизироваться при обучении.

Заметим, что из $\xi_i > \gamma$ следует, что ошибка на $(\bar{x}_i, y_i)$ больше чем θ.

В монографии [10] приведена теорема (теорема 4.28), которая в отличие от теоремы 2.9 не требует точного попадания данных выборки в слой размера $\epsilon = \theta - \gamma$ вокруг гиперплоскости, а использует вектор переменных мягкого отступа $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$. Здесь возможно произвольное отклонение данных при обучении от гиперплоскости регрессии. При этом большие отклонения ухудшают оценку вероятности ошибки регрессии.

Верхняя оценка этой теоремы служит основой для выбора критерия оптимизации при выборе разделяющей гиперплоскости.

Теорема 2.10. *Рассмотрим задачу регрессии с помощью линейных функций* $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b \in \mathcal{L}$.

[3] Заметим, что размер слоя ϵ исчисляется по вертикали.

Пусть $0 < \gamma \leqslant \theta \in \mathcal{R}$. *Тогда существует такая константа* c, *что для произвольного распределения вероятностей* P, *сконцентрированного в шаре радиуса* R *с центром в начале координат, выполнено следующее: с* P^l*-вероятностью* $1 - \delta$ *на выборке* $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ *произвольная гипотеза* $f \in \mathcal{L}$ *с вектором переменных мягкого отступа* $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ *имеет оценку вероятности ошибки обобщения:*

$$\begin{gathered} \operatorname{err}_P(f) = P\{|f(\bar{x}) - y| \geqslant \theta\} \leqslant \\ \leqslant \frac{c}{l}\left(\frac{\|\bar{w}\|^2 R^2 + \|\bar{\xi}\|^2}{\gamma^2} \ln^2 l + \ln\frac{1}{\delta}\right), \end{gathered} \tag{2.61}$$

где $\xi_i = \max\{0, |y_i - f(\bar{x}_i)| - \epsilon\}$, $i = 1, \ldots, \ l$, *где* $\epsilon = \theta - \gamma$.

Можно положить $\gamma = \theta$. Тогда $\xi_i = |y_i - f(\bar{x}_i)|$, $i = 1, \ \ldots, \ l$. В этом случае

$$\|\bar{\xi}\|^2 = \sum_{i=1}^{l} (y_i - f(\bar{x}_i))^2.$$

Упрощенный вариант теоремы 2.10 выглядит следующим образом.

Следствие 2.2. *Рассмотрим задачу регрессии с помощью линейных функций* $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b \in \mathcal{L}$, *где* $\bar{x} \in \mathcal{X} \subseteq \mathcal{R}^n$.

Пусть $\theta > 0$. *Тогда существует такая константа* c, *что для произвольного распределения вероятностей* P, *сконцентрированного в шаре радиуса* R *с центром в начале координат, выполнено следующее: с* P^l*-вероятностью* $1 - \delta$ *на случайной выборке* $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$ *произвольная гипотеза* $f \in \mathcal{L}$ *имеет оценку вероятности ошибки обобщения:*

$$\begin{gathered} \operatorname{err}_P(f) = P\{|f(\bar{x}) - y| \geqslant \theta\} \leqslant \\ \leqslant \frac{c}{l}\left(\frac{\|\bar{w}\|^2 R^2 + \sum\limits_{i=1}^{l} (y_i - f(\bar{x}_i))^2}{\theta^2} \ln^2 l + \ln\frac{1}{\delta}\right). \end{gathered} \tag{2.62}$$

Оценка (2.62) может быть использована для гребневой регрессии, где $\epsilon = 0$.

Отметим, что все эти оценки ошибки обобщения имеют место для произвольного неизвестного нам распределения вероятностей P, генерирующего элементы выборки.

2.9.2. 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

$$L^{\epsilon}(\bar{x}, y, f) = |y - f(\bar{x})|_{\epsilon} = \max\{0, |y - f(\bar{x})| - \epsilon\}, \qquad (2.63)$$

где f – произвольная функция типа $\mathcal{R}^n \to \mathcal{R}$.

Пусть $\bar{\xi} = (\xi_1, \ \ldots, \ \xi_l)$ – вектор переменных мягкого отступа, где $\xi_i = L^{\epsilon}(\bar{x}_i, y_i, f)$, $i = 1, \ \ldots, \ l$, где $\epsilon = \theta - \gamma$.

Аналогично, ϵ-нечувствительная квадратичная функция потерь определяется

$$L_2^{\epsilon}(\bar{x}, y, f) = (|y - f(\bar{x})|_{\epsilon})^2. \qquad (2.64)$$

Задача минимизации в случае квадратичной функции потерь

Оценка (2.61) теоремы 2.10 показывает, что для уменьшения верхней оценки вероятности ошибки обобщения метода регрессии нам надо минимизировать величину

$$R^2\|\bar{w}\|^2 + \sum_{i=1}^{l} L_2^{\epsilon}(\bar{x}_i, y_i, f),$$

где $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$.

Мы будем минимизировать функцию потерь сразу при всех возможных значениях параметра γ и при всех возможных значениях параметра $\theta = \epsilon + \gamma$ 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 (2.64) при фиксированных значениях параметров C и ϵ формулируется следующим образом:

$$\begin{aligned} \|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) &\to \min \\ \text{при условиях } ((\bar{w}\cdot\bar{x}_i)+b) - y_i &\leqslant \epsilon + \xi_i,\ i = 1,\ \ldots,\ l, \\ y_i - ((\bar{w}\cdot\bar{x}_i)+b) &\leqslant \epsilon + \hat{\xi}_i,\ i = 1,\ \ldots,\ l, \\ \xi_i, \hat{\xi}_i &\geqslant 0,\ i = 1,\ \ldots,\ l. \end{aligned} \quad (2.65)$$

На практике параметр C подбирается путем процедуры типа перебора для данной обучающей выборки.

Лагранжиан прямой задачи имеет вид

$$\begin{aligned} L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = |\bar{w}|^2 + C\sum_{i=1}^{l}(\xi_i^2 + \hat{\xi}_i^2) + \\ + \sum_{i=1}^{l}\alpha_i((\bar{w}\cdot\bar{x}_i) + b - y_i - \epsilon - \xi_i) + \\ + \sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w}\cdot\bar{x}_i) - b - \epsilon - \hat{\xi}_i). \end{aligned}$$

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 частные производные к

нулю

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \bar{w}} = \bar{0},$$
$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial b} = 0,$$
$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \bar{\xi}} = \bar{0},$$
$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha})}{\partial \hat{\xi}} = \bar{0}.$$

Из первого равенства получаем выражение для весового вектора

$$\bar{w} = \frac{1}{2} \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) \bar{x}_i. \tag{2.66}$$

Заметим, что для любого допустимого решения задачи (2.65) 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:

$$\sum_{i=1}^{l} y_i (\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l} (\hat{\alpha}_i + \alpha_i) -$$
$$-\frac{1}{2} \sum_{i,j=1}^{l} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C}\delta_{ij}) \to \max$$
$$\text{при условиях } \sum_{i=1}^{l} (\hat{\alpha}_i - \alpha_i) = 0,$$
$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0,\ i = 1,\ \ldots,\ l, \tag{2.67}$$

где $\delta_{ij} = 1$ тогда и только тогда, когда $i = j$.

Условия Каруша–Куна–Таккера следующие:

$$\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \epsilon - \xi_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \epsilon - \hat{\xi}_i) = 0,\ i = 1,\ \ldots,\ l,$$
$$\xi_i \hat{\xi}_i = 0,\ \alpha_i \hat{\alpha}_i = 0,\ i = 1,\ \ldots,\ l. \tag{2.68}$$

Если ввести обозначение $\beta_i = \hat{\alpha}_i - \bar{\alpha}_i$, $i = 1,\ \ldots,\ l$, и использовать соотношения $\alpha_i \hat{\alpha}_i = 0$, то двойственная задача (2.67) формулируется аналогично двойственной задаче для случая классификации:

$$\sum_{i=1}^{l} y_i \beta_i - \epsilon \sum_{i=1}^{l} |\beta_i| -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} \beta_i \beta_j ((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C} \delta_{ij}) \to \max$$

$$\text{при условиях} \sum_{i=1}^{l} \beta_i = 0,\ i = 1,\ \ldots,\ l. \tag{2.69}$$

Из условий Каруша–Куна–Таккера (2.68) 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 (2.66) соответствующие слагаемые отсутствуют («опорных векторов» становится меньше), и решение задачи нахождения максимума в двойственной задаче упрощается. Кроме этого, уменьшается норма вектора переменных мягкого отступа в верхней оценке вероятности ошибки обобщения (2.61) из теоремы 2.10. Заметим, что опорными являются векторы $\bar{x}_i$, для которых $(\bar{w} \cdot \bar{x}_i) + b \leqslant y_i - \epsilon$ или $(\bar{w} \cdot \bar{x}_i) + b \geqslant y_i + \epsilon$.

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 с помощью SVM

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. Для удобства изложения обозначим β_i посредством α_i, т.е. α_i имеет теперь другой смысл, чем ранее.

Теорема 2.11. *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 $\bar{\alpha}^*$ *является решением квадратичной оптимизационной задачи:*

$$W(\bar{\alpha}) = \sum_{i=1}^{l} y_i \alpha_i - \epsilon \sum_{i=1}^{l} |\alpha_i| -$$

$$-\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j (K(\bar{x}_i, \bar{x}_j) + \frac{1}{C} \delta_{ij}) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l} \alpha_i = 0, \ i = 1, \ \ldots, \ l. \tag{2.70}$$

Пусть также $f(\bar{x}) = \sum\limits_{i=1}^{l} \alpha_i K(\bar{x}_i, \bar{x}) + b^*$, *где* b^* *выбирается так, чтобы* $f(\bar{x}_i) - y_i = -\epsilon - \alpha_i / C$ *для произвольного* i *с* $\alpha_i > 0$.

Тогда функция $f(\bar{x})$ *эквивалентна гиперплоскости в пространстве признаков, определяемом ядром* $K(\bar{x}_i, \bar{x})$, *которая решает задачу оптимизации* (2.65).

Задача минимизации в случае линейной функции потерь

Аналогичным образом рассматривается задача регрессии при линейной функции потерь (2.63).

Оценка, аналогичная оценке (2.61) теоремы 2.10 (в данном пособии эта оценка и соответствующая теорема не приводятся, см. теорему 4.30 из [10]), показывает, что для уменьшения верхней оценки вероятности ошибки обобщения метода регрессии нам надо минимизировать величину

$$\frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^{l} L^{\epsilon}(\bar{x}_i, y_i, f),$$

где $f(\bar{x}) = (\bar{w} \cdot \bar{x}) + b$, C – параметр, контролирующий баланс между сложностью регрессионной гипотезы и суммой величин линейных остатков для этой гипотезы.

Прямая задача минимизации в случае линейной функции потерь (2.63) при фиксированных значениях параметров C и ϵ формулируется следующим образом:

$$\frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) \to \min$$

$$\text{при условиях } ((\bar{w}\cdot\bar{x}_i) + b) - y_i \leqslant \epsilon + \xi_i,\ i = 1,\ \ldots,\ l,$$

$$y_i - ((\bar{w}\cdot\bar{x}_i) + b) \leqslant \epsilon + \hat{\xi}_i,\ i = 1,\ \ldots,\ l,$$

$$\xi_i, \hat{\xi}_i \geqslant 0,\ i = 1,\ \ldots,\ l. \quad (2.71)$$

На практике параметр C подбирается путем процедуры типа перебора для данной обучающей выборки.

Лагранжиан прямой задачи имеет вид

$$L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = \|\bar{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i) +$$

$$+\sum_{i=1}^{l}\alpha_i((\bar{w}\cdot\bar{x}_i) + b - y_i - \epsilon - \xi_i) +$$

$$+\sum_{i=1}^{l}\hat{\alpha}_i(y_i - (\bar{w}\cdot\bar{x}_i) - b - \epsilon - \hat{\xi}_i).$$

Соответствующая прямой задаче (2.71) двойственная задача формулируется следующим образом:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon\sum_{i=1}^{l}(\hat{\alpha}_i + \alpha_i) -$$

$$-\frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(\bar{x}_i\cdot\bar{x}_j) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0, \quad (2.72)$$

$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C,\ i = 1,\ \ldots,\ l, \quad (2.73)$$

а ее решение такое, как указана далее.

Условия Каруша–Куна–Таккера имеют вид:

$$\alpha_i((\bar{w}\cdot\bar{x}_i)+b-y_i-\epsilon-\xi_i)=0,\ i=1,\ \ldots,\ l,$$
$$\hat{\alpha}_i(y_i-(\bar{w}\cdot\bar{x}_i)-b-\epsilon-\hat{\xi}_i)=0,\ i=1,\ \ldots,\ l,$$
$$(\alpha_i-C)\xi_i=0,\ (\hat{\alpha}_i-C)\hat{\xi}_i=0, \tag{2.74}$$
$$\xi_i\hat{\xi}_i=0,\ \alpha_i\hat{\alpha}_i=0,\ i=1,\ \ldots,\ l. \tag{2.75}$$

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 $\hat{\alpha}_i=C$ (для линейной нормы).

$0<\alpha_i<C$ может быть только для векторов со значениями y_i 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

$$(\bar{w}\cdot\bar{x}_i)+b+\epsilon<y_i, \qquad \xi_i>0,$$
и
$$(\bar{w}\cdot\bar{x}_i)+b-\epsilon<y_i, \qquad \hat{\xi}_i>0.$$

Весовой вектор – линейная комбинация опорных векторов

$$\bar{w}=\frac{1}{2}\sum_{i=1}^{l}(\hat{\alpha}_i-\alpha_i)\bar{x}_i.$$

Выполнено $\alpha_i\hat{\alpha}_i=0$.

Функция линейной регрессии имеет вид:

$$f(\bar{x})=\sum_{i=1}^{l}\alpha_i(\bar{x}_i\cdot\bar{x})+b^*.$$

Для пространства с ядром двойственная задача имеет вид:

$$\sum_{i=1}^{l} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon \sum_{i=1}^{l}(\hat{\alpha}_i + \alpha_i) -$$

$$-\frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)K(\bar{x}_i, \bar{x}_j) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l}(\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geqslant 0, \alpha_i \geqslant 0,$$

$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C,\ i = 1,\ \ldots,\ l.$$

Функция регрессии для ядерной версии имеет вид:

$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i K(\bar{x}_i, \bar{x}) + b^*.$$

2.9.3. Гребневая регрессия в двойственной форме

Гребневая регрессия может быть представлена как частный случай регрессии с опорными векторами при ϵ-нечувствительной квадратичной функции потерь (2.64), 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.

Оценка ошибки обобщения для гребневой регрессии дается неравенством (2.62), приведенном в следствии 2.2.

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

$$L(\tilde{w}, \bar{\xi}, \bar{\alpha}) = \lambda|\tilde{w}|^2 + \sum_{i=1}^{l} \xi_i^2 + \sum_{i=1}^{l} \alpha_i (y_i - (\tilde{w} \cdot \tilde{x}_i) - \xi_i). \qquad (2.76)$$

Приравниваем к нулю частные производные лагранжиана (2.76) 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 (2.76), получим задачу в двойственной форме

$$W(\bar{\alpha}) = \sum_{i=1}^{l} y_i \alpha_i - \frac{1}{4\lambda} \sum_{i,j=1}^{l} \alpha_i \alpha_j (\tilde{x}_i \cdot \tilde{x}_j) - \\ - \frac{1}{4} \sum_{i=1}^{l} \alpha_i^2 \to \max. \qquad (2.77)$$

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 (2.77) 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:

$$\bar{\alpha} = 2\lambda (K + \lambda I)^{-1} \bar{y}. \qquad (2.78)$$

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 (2.78) имеем

$$\bar{\alpha}' = 2\lambda \bar{y}' (K + \lambda I)^{-1}.$$

Отсюда функция регрессии имеет вид

$$f(\tilde{x}) = (\tilde{w} \cdot \tilde{x}) = \bar{y}'(K + \lambda I)^{-1}\bar{k}. \tag{2.79}$$

Отметим один недостаток данной постановки. Поскольку в данной постановке $\epsilon = \theta - \gamma = 0$, число параметров α_i равно размеру выборки l, поэтому 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 (2.79), 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

$$f(\tilde{x}) = \bar{y}'(\lambda I + K)^{-1}\bar{z}, \tag{2.80}$$

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 (2.80) совсем не обязательно знать конкретный вид отображения в пространство признаков, достаточно знать ядро $K(\tilde{x}, \tilde{z})$.

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 (Krieging)*. 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

$$\inf_{\bar{w}} f(\bar{w}) \text{ при условиях}$$

$$g_i(\bar{w}) \leqslant 0,\ i = 1,\ \ldots,\ m, \tag{2.81}$$

$$h_i(\bar{w}) = 0,\ i = 1,\ \ldots,\ m. \tag{2.82}$$

Последние два условия можно записать в векторном виде $\bar{g}(\bar{w}) \leqslant \bar{0}$ и $\bar{h}(\bar{w}) = \bar{0}$. Пусть

$$\mathcal{R} = \{\bar{w} \in \mathcal{R}^n : \bar{g}(\bar{w}) \leqslant \bar{0}, \bar{h}(\bar{w}) = 0\}$$

– *область допустимости решений.*

Решение задачи оптимизации – это такой вектор $\bar{w}^*$, что $\bar{w}^* \in \mathcal{R}$ 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 (2.82) 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

$$\max_{(\bar{\alpha}, \bar{\beta})} \Theta(\bar{\alpha}, \bar{\beta}) \text{ при условиях} \\ \alpha_i \geqslant 0,\ i = 1,\ \ldots,\ m. \tag{2.83}$$

Ниже приводится слабая теорема двойственности.

Теорема 2.12. *Пусть вектор $\bar{w}$ удовлетворяет условиям* (2.81) *и* (2.82) *прямой задачи оптимизации (в частности, он может быть решением прямой задачи), а $(\bar{\alpha}, \bar{\beta})$ – решение двойственной задачи* (2.83)*. Тогда $f(\bar{w}) \geqslant \Theta(\bar{\alpha}, \bar{\beta})$.*

Доказательство. Имеем

$$\Theta(\bar{\alpha}, \bar{\beta}) = \inf_{\bar{u}} L(\bar{u}, \bar{\alpha}, \bar{\beta}) \leqslant L(\bar{w}, \bar{\alpha}, \bar{\beta}) = \\ = f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}) \leqslant f(\bar{w}). \tag{2.84}$$

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 2.3. *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 2.4. *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 (2.84) 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 2.13. *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.*

Теорема Куна–Таккера – основная теорема выпуклой нелинейной оптимизации.

Теорема 2.14. *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*

$$\begin{aligned} \inf f(\bar{w}),\ \bar{w} \in \Omega,\ &\textit{при условиях} \\ \bar{g}(\bar{w}) &\leqslant \bar{0}, \\ \bar{h}(\bar{w}) &= \bar{0} \end{aligned}$$

тогда и только тогда, когда существует пара $(\bar{\alpha}^*, \bar{\beta}^*)$ *такая, что*

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{w}} = \bar{0},$$

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{\beta}} = \bar{0}, \tag{2.85}$$

$$\alpha_i^* g_i(\bar{w}^*) = 0,\ i = 1,\ \ldots,\ m, \tag{2.86}$$

$$g_i(\bar{w}^*) \leqslant 0,\ i = 1,\ \ldots,\ m,$$

$$\alpha_i^* \geqslant 0,\ i = 1,\ \ldots,\ m.$$

Условия достижения максимума по $\bar{\beta}$ линейной по $\bar{\alpha}$ и $\bar{\beta}$ функции $L(\bar{w}^*, \bar{\alpha}, \bar{\beta})$ задаются условиями (2.85); 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 (2.86), 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 (2.86) 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

$$\frac{1}{2}\bar{w}' Q \bar{w} - \bar{k}\bar{w} \to \min$$

$$\text{при условии } X\bar{w} \leqslant \bar{c}, \tag{2.87}$$

где Q – $n \times n$-положительно определенная матрица, $\bar{k}$ – n-вектор, $\bar{c}$ – m-вектор, $\bar{w}$ – n-вектор неизвестных, X – (m, n)-матрица.

Допускаем, что условия определяют непустое множество векторов. Тогда задача может быть переписана в виде: найти максимум

$$\max_{\bar{\alpha}} \left(\min_{\bar{w}} \left(\frac{1}{2}\bar{w}' Q \bar{w} - \bar{k}\bar{w} + \bar{\alpha}'(X\bar{w} - \bar{c}) \right) \right) \tag{2.88}$$

$$\text{при условии } \bar{\alpha} \geqslant \bar{0}.$$

Минимум по $\bar{w}$ в (2.88) достигается при

$$\bar{w} = Q^{-1}(\bar{k} - X'\bar{\alpha}).$$

Подставляем это выражение в (2.87), получим двойственную задачу

$$-\frac{1}{2}\bar{\alpha}'P\bar{\alpha} - \bar{\alpha}'\bar{d} - \frac{1}{2}\bar{k}'Q\bar{k} \to \max$$
$$\text{при условии } \bar{\alpha} \geqslant \bar{0}, \tag{2.89}$$

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 некомформности – была введена Вовком и Гаммерманом [33]. Мера некомформности 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 некомформности 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, определенным ядром $K(\bar{x}, \bar{x}')$. После этого, строится разделяющая гиперплоскость в пространстве признаков, Согласно (2.42)

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 2.5 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 построния SVM по выборке S, при этом будут вычислены коэффициенты Лагранжа α_i. Возьмем в качестве *меры некомформности* 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 $y_i((\bar{w} \cdot \bar{\phi}(\bar{x}_i)) + b) > 1$. Такие векторы $\bar{x}_i$ правильно классифицируются и лежат с внешней стороны относительно граничных гиперплоскостей. Опорными являются те векторы $\bar{x}_i$, для которых выполнено $y_i((\bar{w} \cdot \bar{\phi}(\bar{x}_i)) + b) \leqslant 1$, при этом $\alpha_i \geqslant 0$ и $\xi_i = \alpha_i / C$. Это те векторы, образы которых лежат на граничных гиперплоскостях или же неправильно ими классифицируются, в этом случае $y_i((\bar{w} \cdot \bar{\phi}(\bar{x}_i)) + b) < 1$. В случае линейной нормы добавляется условие $\alpha_i \leqslant C$, где C – соответственная константа из задачи оптимизации. Таким образом:

- примеры с $\alpha_i = 0$ правильно классифицируются и поэтому имеют высшую степень согласованности с выборкой (по которой построена гиперповерхность);
- примеры с положительными значениями α_i либо лежат на граничных гиперплоскостях, либо неправильно классифицируются и поэтому степень согласованности с выборкой тем хуже, чем больше значение α_i.

Введенная мера некомформности применяется к примеру $(\bar{x}_i, y_i)$. Определяется *p-mecm (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 некомформности среди примеров выборки.

На основе введенного понятия p-теста можно построить 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 (подровнее см. в [33]).

Меры некомформности строятся исходя из специфики моделей данных. В монографии [33] 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 заключатся 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 $\bar{x} \in \mathcal{R}^n$, $y \in D$, где 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.

$(\bar{x}, y)$ – некоторый пример. Определим меру некомформности 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 эвклидово 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 некомформности 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 ($k = 3, 4, \ldots$), а также соответствующие функции классификации вида (2.25).

3. Для любой положительно определенной функции $K(x, y)$ выполнено неравенство типа Коши–Буняковского:

$$K(x_1, x_2) \leqslant \sqrt{K(x_1, x_1)K(x_2, x_2)} \text{ для всех } x_1, x_2 \in X. \quad (2.90)$$

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

(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.

Заметим, что функция $K(x, y)$ в общем случае не является билинейной.

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 $K(x, y) = (K_x \cdot K_y)$.

Доказать, что функция $K(x, y) = (K_x \cdot K_y)$ является симметричной и положительно определенной.

6. Пусть $K_1(x, y), K_2(x, y), \ldots,$ – положительно определенные ядра на X. Доказать, что следующие их комбинации также являются 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 $K = PP'$);

(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 в оптимизационной задаче (2.70) значение b^* не зависит от i.

8. Показать, что соответствующая прямой задаче (2.71) двойственная задача формулируется в виде (2.73). Обосновать соотношения (2.75) для этой задачи.

9. Доказать, что матрица Грама $K_{i,j} = (\tilde{x}_i \cdot \tilde{x}_j)$ обратима тогда и только тогда, когда векторы $\tilde{x}_1, \ \ldots, \ \tilde{x}_l$ линейно независимы.

10. (i) Найти максимум объема параллелепипеда при заданной площади поверхности.

(ii) Найти максимум энтропии $H(p_1, \ \ldots, \ p_n) = -\sum_{i=1}^{n} p_i \ln p_i$ при условиях $\sum p_i = 1$, $\sum c_i p_i = e$.

11. Провести все необходимые выкладки для получения решения (2.89) квадратичной задачи.

12. Доказать, что для класса $\mathcal{F}$ всех линейных (однородных) функций множество является γ-разделимым тогда и только тогда, когда оно γ-разделимо (может быть для другого γ) на одном уровне, причем $r = 0$.

13. Вывести соотношения (2.69) для двойственной задачи регрессии.

2.13. Лабораторные работы по теме SVM

В этом разделе предлагаются стандартные лабораторные работы для решения задачи классификации с помощью SVM.

Выполнение работы включает следующие процедуры:

- Загрузить исходные данные из 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 $K(\bar{x}, \bar{y}) = tanh(\gamma(\bar{x} \cdot \bar{y}) + r)$.

  Рекомендуется первоначально выбрать гауссово ядро $K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x}-\bar{y}\|^2}{\sigma^2}}$. 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/cjlin/libsvm/

www.support-vector.net

По адресу

http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf

можно найти инструкции по практическому применению программ SVM и подготовке исходных данных. Там же приведены примеры.

На сайте

http://archive.ics.uci.edu/ml/

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/roweis/data.html

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 следующих сайтов. Выбрать набор данных, провести обучение на SVM и тестирование на тестовой выборке.

Библиотека LIBSVM для машин с опорными векторами находится по адресу

http://www.csie.ntu.edu.tw/cjlin/libsvm/

База данных для машинного обучения находится по адресу

http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/

Примеры данных и задач:

1) Провести обучение и классификацию видов лейкемии по медицинским данным из следующего сайта:

http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/

2) Откалиброванные данные для классификации вин (по 13 признакам) имеются по адресу

http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/

multiclass/wine.scale

База данных UCI Machine Learning Repository находится на сайте

http://archive.ics.uci.edu/ml/datasets.html

Она содержит 190 наборов данных для классификации и регрессии.

Лабораторная работа 3

Провести обучение и классификацию на предыдущих данных с помощью персептрона и алгоритма Розенблатта. Провести сравнение времени работы.

Глава 3

Универсальные предсказания

3.1. Универсальное прогнозирование в режиме онлайн

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 событий 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.

Исторически первой процедурой универсального прогнозирования является правило Лапласа. Эта процедура использует гипотезу 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. Особенность данной задачи

заключается в том, что мы не знаем истинного значения p и хотим построить процедуру прогнозирования, которая бы годилась для всех p таких, что $0 \leqslant p \leqslant 1$.

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 :

$$P(\omega^n) = \int_0^1 p^{n_1}(1-p)^{n_2} dp.$$

Значение этого интеграла легко вычислить.

Лемма 3.1.

$$\int_0^1 p^{n_1}(1-p)^{n_2} dp = \frac{1}{(n+1)\binom{n}{n_1}}.$$

Доказательство. Проверим это равенство обратной индукцией по n_1. При $n_1 = n$ имеем $\int_0^1 p^n dp = \frac{1}{(n+1)}$.

Предположим, что

$$\int_0^1 p^{n_1+1}(1-p)^{n_2-1} dp = \frac{1}{(n+1)\binom{n}{n_1+1}}.$$

Интегрируя по частям, получим

$$\int_0^1 p^{n_1}(1-p)^{n_2}dp = \frac{n-n_1}{n_1+1}\int_0^1 p^{n_1+1}(1-p)^{n_2-1}dp = \\ = \frac{n-n_1}{n_1+1}\frac{1}{(n+1)\binom{n}{n_1+1}} = \frac{1}{(n+1)\binom{n}{n_1}}.$$

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 такой функции потерь – логарифмическая функция потерь:

$$L_p(\omega^n) = -\ln p(\omega^n) = -\ln(p^{n_1}(1-p)^{n_2}).$$

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 данным распределением вероятностей.

Для правила Лапласа

$$L(\omega^n) = \ln P(\omega^n) = -\ln\int_0^1 p^{n_1}(1-p)^{n_2}dp.$$

Тогда для произвольной последовательности ω^n выполнено

$$L(\omega^n) - \inf_{0\leqslant p\leqslant 1} L_p(\omega^n) = \ln \frac{\sup\limits_{0\leqslant p\leqslant 1} p^{n_1}(1-p)^{n_2}}{\int\limits_0^1 p^{n_1}(1-p)^{n_2}dp} =$$
$$= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{(n+1)\binom{n}{n_1}}} \leqslant \ln(n+1).$$

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 истинных вероятностей источника, порождающего исходы ω_i.

Легко проверить, что

$$\sup_{0\leqslant p\leqslant 1} p^{n_1}(1-p)^{n_2} = \left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}.$$

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 1 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 3.5 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 регрет асимптотически в два раза меньше, чем в соответствующей оценке для метода Лапласа.

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

$$\frac{\sum_{i=1\&p_i\approx p^*}^{n} \omega_i}{\sum_{i=1\&p_i\approx p^*}^{n} 1_{p_i\approx p^*}} \approx p^*, \tag{3.1}$$

если знаменатель отношения (3.1) стремится к бесконечности при $n \to \infty$. Здесь мы использовали символ $\approx$ приближенного равенства, потому что на практике число p^* можно задавать только с некоторой точностью. Условие $p_i \approx p^*$ требует дальнейшего уточнения.

Уточним схему действий *Предсказателя* и *Природы* в виде следующего протокола игры с участием этих двух игроков.
FOR $n = 1, 2, \ldots$
Предсказатель анонсирует прогноз $p_n \in [0,1]$.
Природа анонсирует исход $\omega_n \in \{0,1\}$.
ENDFOR

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 [11]. Рассмотрим произвольные подынтервалы $I = [a,b]$, $(a,b]$, $[a,b)$, (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.

$$\frac{\sum_{i=1}^{n} I(p_i)(\omega_i - p_i)}{\sum_{i=1}^{n} I(p_i)} \longrightarrow 0, \tag{3.2}$$

если знаменатель отношения (3.2) 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.

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 (3.2) нарушается.

Данная последовательность $\omega = \omega_1\omega_2\ldots$ является простейшим примером «адаптивно враждебной» стратегии *Природы*. При генерации очередного исхода ω_i согласно приведенному выше протоколу *Природа* уже знает наш прогноз p_i и использует это знание для формирования очередного исхода.

Подобные трудности предсказания оказались преодолимыми с помощью понятия рандомизированной предсказательной системы. Пусть $\mathcal{P}[0,1]$ – множество всех вероятностных мер на отрезке $[0,1]$. Под *рандомизированной* предсказательной системой понимается функция $f : \Xi \to \mathcal{P}[0,1]$, значениями которой являются распределения вероятностей на отрезке $[0,1]$. Мы также обозначаем $\Pr_x(\cdot) = f(x)$, где x – конечная последовательность исходов.

Обозначаем $\omega^{i-1} = \omega_1 \ldots \omega_{i-1}$. Для каждой бесконечной последовательности $\omega = \omega_1\omega_2\ldots$ (которая в данном случае является параметром) условные вероятностные распределения $\Pr_{\omega^{i-1}}(\cdot)$ порождают распределение вероятностей $\Pr = \prod_{i=1}^{\infty} \Pr_{\omega^{i-1}}$ на множестве всех бесконечных последовательностей прогнозов $p_1, p_2, \ldots,$

где $p_i \in [0,1]$, $i = 1, 2, \ldots$ Данное распределение вероятностей Pr является бесконечным произведением мер $\Pr_{\omega^{i-1}}$, $i = 1, 2, \ldots$

В этом случае при фиксированной последовательности ω можно рассматривать вероятность Pr события (3.2).

Заметим, что такая мера Pr существует и в гораздо более общем случае, а именно, когда последовательность $\omega = \omega_1\omega_2\ldots$ зависит от последовательности прогнозов $p_1, p_2, \ldots$, точнее последовательность $\omega^{i-1} = \omega_1\ldots\omega_{i-1}$ является измеримой функцией от последовательности $p_1, \ldots, p_{i-1}$ для всех i. В этом случае по теореме Ионеско–Тульчи [3] о продолжении меры существует вероятностная мера Pr на множестве $[0,1]^\infty$ всех бесконечных траекторий прогнозов $p_1, p_2, \ldots$ такая, что условные вероятности, соответствующие этой мере, для всех n удовлетворяют условиям

$$\Pr\{p_n \in A | p_1, \ldots, p_{n-1}\} = \Pr_{\omega^{n-1}}(A)$$

для любого борелевского подмножества A единичного интервала.

Фостер и Воора [12], а также Какаде и Фостер [16], определили для произвольного параметра $\Delta > 0$ рандомизированную предсказательную систему f такую, что для произвольной бесконечной последовательности $\omega = \omega_1\omega_2\ldots$ с Pr-вероятностью 1 :

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \right| \leqslant \Delta,$$

где траектории прогнозов $\tilde{p}_1, \tilde{p}_2, \ldots$ распределены по вероятностной мере Pr, а $I(p)$ – характеристическая функция произвольного подынтервала $[0,1]$.

Более точные оценки, приведенные в этих и последующих работах, показывают, что это условие можно заменить, например, на условие

$$\limsup_{n\to\infty} \left| \frac{1}{\sqrt{n\alpha(n)}} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \right| \leqslant \Delta,$$

где $\alpha(n)$ – произвольная неограниченная неубывающая функция.

В этом случае (3.2) также выполнено для алгоритма Какаде и Фостера, если

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n\alpha(n)}} \sum_{i=1}^{n} I(\tilde{p}_i) = \infty.$$

3.3. Алгоритм вычисления калибруемых прогнозов

Приведем некоторый модернизированный вариант рандомизированного алгоритма Какаде и Фостера. Пусть $\omega_1\omega_1\ldots$ – произвольная последовательность элементов $\{0,1\}$ или $[0,1]$, поступающая в режиме онлайн. Построим алгоритм для вычисления случайной величины, выдающей прогноз $p_n \in [0,1]$ 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 разобъем 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

$$(\mu_n(v))^2 = (\mu_{n-1}(v))^2 + 2w_v(p_n)\mu_{n-1}(v)(\omega_n - p_n) + \\ +(w_v(p_n))^2(\omega_n - p_n)^2. \qquad (3.3)$$

Суммируем (3.3) по v :

$$\sum_{v\in V}(\mu_n(v))^2 = \sum_{v\in V}(\mu_{n-1}(v))^2 + \\ +2(\omega_n - p_n)\sum_{v\in V} w_v(p_n)\mu_{n-1}(v) + \\ + \sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2. \qquad (3.4)$$

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

$$K(p, p_i) = (\bar{w}(p) \cdot \bar{w}(p_i)) \tag{3.5}$$

– скалярное произведение соответствующих векторов (ядро). По определению $K(p, p_i)$ – непрерывная функция.

Второй член правой части равенства (3.4) при подходящем значении p_n всегда можно сделать меньшим или равным нулю. Действительно, в качестве p_n берем корень $p_n = p$ уравнения

$$\sum_{v \in V} w_v(p)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i) = 0, \tag{3.6}$$

если он существует. В противном случае если левая часть уравнения (3.6) (которая является непрерывной по p функцией) больше нуля для всех значений p_n, то полагаем $p_n = 1$, если она меньше нуля, то полагаем $p_n = 0$. Определенное таким образом значение p_n выдаем в качестве детерминированного прогноза.

Третий член (3.4) 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 (3.4), если последовательно выбирать прогнозы p_i согласно указанному правилу, получим

$$\sum_{v \in V} (\mu_n(v))^2 \leqslant \sum_{i=1}^{n} \sum_{v \in V} (w_v(p_i))^2 (\omega_i - p_i))^2 \leqslant n. \tag{3.7}$$

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

$$E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) = \sum_{v \in V} w_v(p_i) I(v)(\omega_i - v). \tag{3.8}$$

Согласно усиленному мартингальному закону больших чисел (см. следствие 4.9 ниже) с Pr-вероятностью 1 :

$$\left| \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) - \frac{1}{n} \sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) \right| \to 0 \tag{3.9}$$

при $n \to \infty$.

По определению детерминированного прогноза p_i и функции $w_v(p)$

$$\left| \sum_{v \in V} w_v(p_i) I(v)(\omega_i - v) - \sum_{v \in V} w_v(p_i) I(v)(\omega_i - p_i) \right| < \Delta \tag{3.10}$$

для каждого i.

Применяем неравенство Коши–Буняковского к векторам $\{\mu_n(v) : v \in V\}$ и $\{I(v) : v \in V\}$, учитываем (3.10), и получаем

$$\begin{aligned} &\left| \sum_{i=1}^{n} \sum_{v \in V} w_v(p_i) I(v)(\omega_i - p_i) \right| = \\ &= \left| \sum_{v \in V} I(v) \sum_{i=1}^{n} w_v(p_i)(\omega_i - p_i) \right| \leqslant \\ &\leqslant \sqrt{\sum_{v \in V} (\mu_n(v))^2} \sqrt{\sum_{v \in V} I(v)} \leqslant \\ &\leqslant \sqrt{Kn}, \end{aligned} \tag{3.11}$$

где $K = 1/\Delta$ - число подынтервалов разбиения.

Используя (3.10) и (3.11), получаем верхнюю оценку для абсо-

лютной величины суммы математических ожиданий (3.8) :

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))\right| =$$

$$= \left|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v)\right| \leqslant \tag{3.12}$$

$$\leqslant \Delta n + \sqrt{n/\Delta} \tag{3.13}$$

для всех n.

Из (3.12) и (3.9) получаем, что с Pr-вероятностью 1 :

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leqslant \Delta. \tag{3.14}$$

Сформулируем результаты этого раздела в виде следующей теоремы.

Теорема 3.1. *Для каждого $\Delta > 0$ можно построить рандомизированную предсказательную систему f такую, что для произвольной бесконечной последовательности $\omega = \omega_1\omega_2\ldots$ с вероятностью 1 выполнено:*

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leqslant \Delta,$$

где бесконечные траектории прогнозов $\tilde{p}_1, \tilde{p}_2, \ldots$ распределены по мере Pr*, $I(p)$ – характеристическая функция произвольного подынтервала $[0,1]$. Такая функция называется правилом выбора, определенным прогнозом.*

Если в процессе конструкции в определенные моменты времени n_s, $s = 1, 2, \ldots$, изменять $\Delta = \Delta_s$ так что $\Delta_s \to 0$ при $s \to \infty$, можно достичь асимптотически точного результата:

Теорема 3.2. *Можно построить рандомизированную предсказательную систему f такую, что для произвольной бесконечной последовательности $\omega = \omega_1\omega_2\ldots$ с вероятностью 1 выполнено:*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) = 0,$$

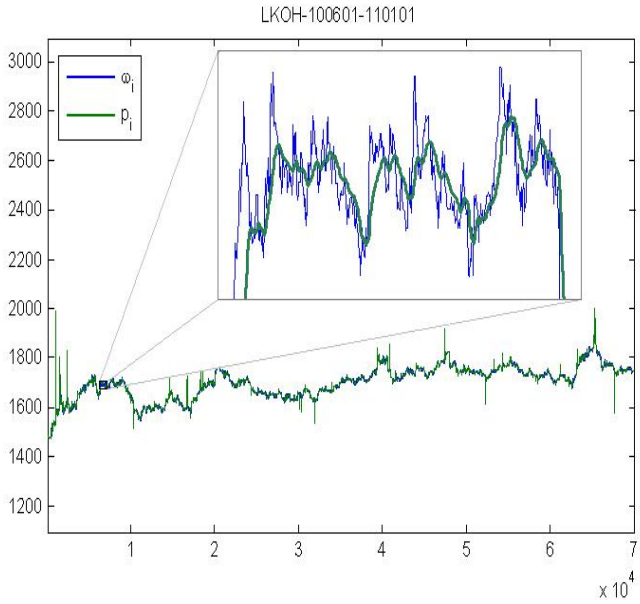

Рис. 2.1. Пример последовательности данных $\omega_1, \omega_2, \ldots$ и последовательности калибруемых прогнозов $p_1, p_2, \ldots$

где $I(p)$ *– характеристическая функция произвольного подынтервала* $[0, 1]$.

Детали конструкции оставляем читателю.

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 единич-

ного интервала;

- универсальное прогнозирование, при котором прогноз является детерминированным, однако в качестве правил выбора разрешается использовать только гладкие приближения к характеристическим функциям подынтервалов единичного интервала.

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. [16]).

Второй метод прогнозирования будет рассмотрен в этом разделе.

Метод построения алгоритмов универсального прогнозирования, предложенный в работах Фостера и Вооры [12], а также Какаде и Фостера [16], был обобщен В. Вовком на случай произвольных ядер в работах [35] и [37]. В этом разделе мы представим основную идею этого обобщения.

Сформулируем задачу прогнозирования в виде некоторой игры между игроками: *Природа*, *Предсказатель* и *Скептик*.

В этой игре прогнозы будут детерминированными, подобно прогнозам p_i, которые вычисляются в виде корней уравнений типа (3.6) 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* $\mathcal{K}_0 = 1$.

Игра регулируется следующим протоколом.
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

Следующая теорема показывает, что *Предсказатель* имеет стратегию, при которой выигрыш *Скептика* не возрастает.

Теорема 3.3. *Предсказатель имеет стратегию, 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. 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

$$S_n(p) = 0, \tag{3.15}$$

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 $S_n(p)$, 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 $K((p,\bar{x}),(p',\bar{x}'))$ - вещественную гладкую функцию на $([0,1] \times X)^2$. Пример ядра – гауссово ядро:

$$K((p,\bar{x}),(p',\bar{x}')) = \\ = \exp\left(-\frac{(p-p_i)^2}{\sigma_1^2} - \frac{\|\bar{x}-\bar{x}'\|^2}{\sigma_2^2}\right), \tag{3.16}$$

где σ_1, σ_2 – параметры ядра.

Рассмотрим следующую стратегию *Скептика*, которая будет вынуждать *Предсказателя* делать на каждом шаге n «хорошо калибруемые» прогнозы:

$$S_n(p) = \sum_{i=1}^{n-1} K((p, \bar{x}_n), (p_i, \bar{x}_i))(y_i - p_i).$$

Пусть *Предсказатель* использует стратегию, описанную в теореме 3.3. Тогда выигрыш *Скептика* за N шагов игры удовлетворяет соотношениям

$$\begin{aligned}
\mathcal{K}_N - \mathcal{K}_0 &= \sum_{n=1}^{N} S_n(p_n)(y_n - p_n) = \\
&= \sum_{n=1}^{N} \sum_{i=1}^{n-1} K((p_n, \bar{x}_n), (p_i, \bar{x}_i))(y_i - p_i)(y_n - p_n) = \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} K((p_n, \bar{x}_n), (p_i, \bar{x}_i))(y_i - p_i)(y_n - p_n) - \\
&\qquad - \frac{1}{2} \sum_{n=1}^{N} K((p_n, \bar{x}_n), (p_n, \bar{x}_n))(y_n - p_n)^2. \qquad (3.17)
\end{aligned}$$

По теореме Мерсера (см. раздел 2.5) существует 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 «$\cdot$» – скалярное произведение в пространстве $\mathcal{H}$ (далее $\|\cdot\|$ – соответствующая норма).

Величина $c_{\mathcal{H}} = \sup_a \|\Phi(a)\|$ называется константой вложения (embedding constant). Мы рассматриваем гильбертовы пространства H, для которых эта величина конечна: $c_{\mathcal{H}} < \infty$.

Перепишем (3.17) в виде

$$\mathcal{K}_N - \mathcal{K}_0 = \frac{1}{2}\left|\left|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right|\right|^2 - \\ -\frac{1}{2}\sum_{n=1}^{N} ||\bar{\Phi}((p_n, \bar{x}_n)(y_n - p_n)||^2. \tag{3.18}$$

По предположению

$$c_{\mathcal{H}} = \sup_{p,\bar{x}} ||\bar{\Phi}(p, \bar{x})|| < \infty.$$

По теореме 3.3 неравенство $\mathcal{K}_N - \mathcal{K}_0 \leqslant 0$ выполнено для всех n. Тогда из (3.18) следует

$$\frac{1}{2}\left|\left|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right|\right|^2 \leqslant \frac{1}{2}NC^2. \tag{3.19}$$

Неравенство (3.19) перепишем в виде

$$\left|\left|\sum_{n=1}^{N} \bar{\Phi}(p_n, \bar{x}_n)(y_n - p_n)\right|\right| \leqslant \sqrt{N}C. \tag{3.20}$$

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 какоенибудь 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

$$K((p^*, \bar{x}^*), (p, \bar{x})) = I_{(p^*, \bar{x}^*)}(p, \bar{x}). \tag{3.21}$$

Примером такого семейства $I_{p^*}(p)$ является семейство гауссовых ядер типа (3.16).

Для прогнозов p_i будет выполнено

Следствие 3.1.

$$\left|\frac{1}{N}\sum_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)\right| \leqslant \frac{C^2}{\sqrt{N}} \tag{3.22}$$

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 (3.20) 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 (3.22). $\triangle$

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 (3.22) можно переписать в виде

$$\left|\frac{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)(y_n-p_n)}{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)}\right| \leqslant \frac{C^2\sqrt{N}}{\sum\limits_{n=1}^{N} I_{(p^*,\bar{x}^*)}(p_n,\bar{x}_n)}. \tag{3.23}$$

Оценка (3.23) 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. Задачи и упражнения

1. Доказать, что при использовании смешивания по методу Кричевского и Трофимова условная вероятность 1 после n бинарных наблюдений $\omega^n = \omega_1, \ \ldots, \ \omega_n$ равна

$$P(1|\omega^n) = \frac{n_1 + 1/2}{n+1}.$$

2. Доказать, что также имеет место оценка:

$$\int_0^1 \frac{p^{n_1}(1-p)^{n_2}}{\pi\sqrt{p(1-p)}} dp \geqslant \frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}.$$

3. 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 $\omega_1\omega_2\ldots$.

3.6. Лабораторные работы

Алгоритм, описанный в разделе 3.3, может быть легко реализован в виде компьютерной программы. При этом для вычисления корня уравнения (3.6) лучше всего использовать гладкое приближение к ядру (3.5) – гауссово ядро $K(p,p') = e^{-\gamma(p-p')^2}$, для некоторого $\gamma > 0$. [1] Можно также использовать ядро вида $K(p,p') = \cos(\gamma(p-p')^2)$. Корень уравнения (3.6) или (3.15) можно искать методом деления отрезка пополам.

Различные временные ряды можно загружать с сайта FINAM:

http://old.finam.ru/analysis/export/default.asp

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.

[1] В данном случае сигналы отсутствуют. Для использования сигналов можно использовать ядро (3.16).

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

Элементы сравнительной теории машинного обучения

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 на точное предсказание будущего исхода. Имеется два возможных исхода

$\{0,1\}$. Имеются N экспертов (стратегий), которые на каждом шаге 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, например это эксперт 1, точно предсказывает будущий исход: $p_t^1 = \omega_t$ 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.*

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 [20].

Приведем протокол игры на предсказания с экспертами. Участники игры: *Эксперт* i, $i = 1, \ldots, N$, *Статистик*, *Природа*. Каждому участнику игры в момент его действия доступна информация о всех действиях других игроков в моменты, предшествующие данному. Говорим, что это игра с полной информацией.

Пусть ϵ – параметр, $0 < \epsilon < 1$.

Алгоритм $WMA(\epsilon)$
Полагаем $w_1^i = 1$ при $i = 1, \ldots, N$.

FOR $t = 1, 2, \ldots, T$
Эксперт i выдает прогноз $p_t^i \in \{0, 1\}$, $i = 1, \ldots, N$
Статистик 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 $\omega_t \in \{0, 1\}$
Статистик 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

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 *Статистика*, 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 ми-

нимум 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

$$\frac{W_T}{W_0} = \prod_{t=0}^{T-1} \frac{W_{t+1}}{W_t} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M, \tag{4.2}$$

где $M = L_T$ – общее число ошибок алгоритма $WMA(\epsilon)$ на первых T шагах.

Заметим, что $W_0 = \sum\limits_{i=1}^{N} w_0^i = N$. Из (4.1) и (4.2) следует

$$\frac{(1-\epsilon)^m}{N} < \frac{W_T}{W_0} \leqslant \left(1 - \frac{\epsilon}{2}\right)^M.$$

Вычисляем натуральный логарифм от обоих частей этого нера-

венства, проводим следующие переходы:

$$\begin{aligned} m\ln(1-\epsilon) - \ln N &< M\ln\left(1-\frac{\epsilon}{2}\right) \\ m\ln(1-\epsilon) - \ln N &< -\frac{\epsilon}{2}M \\ m\ln\left(\frac{1}{1-\epsilon}\right) + \ln N &> \frac{\epsilon}{2}M \\ m\left(\frac{2}{\epsilon}\right)\ln\left(\frac{1}{1-\epsilon}\right) + \left(\frac{2}{\epsilon}\right)\ln N &> M \\ \left(\frac{2}{1-\epsilon}\right)m + \left(\frac{2}{\epsilon}\right)\ln N &> M, \end{aligned} \tag{4.3}$$

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.

Теорема 4.1 является частным случаем теоремы 4.2.

Исторически, по-видимому, это первый алгоритм такого рода. Он был предложен Литлстоуном и Вармутом в 1989 году и назывался Weighted Majority Algorithm [20]. Несколько позже, в 1990 году, В.Г. Вовк предложил более общий агрегирующий алгоритм (Aggregating Algorithm) и понятие перемешиваемости, которые работают для игр более общего типа [29].

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 [13].

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 стратегию, 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

$$L_T = \sum_{t=1}^{T} (\bar{p}_t \cdot \bar{l}_t).$$

Цель *Распределителя* заключается в выборе такой стратегии распределения $\bar{p}_t$, $t = 1, 2, \ldots, T$, чтобы минимизировать величину

$$R_T = L_T - \min_i L_T^i.$$

Для решения этой задачи рассмотрим алгоритм $Hedge(\beta)$ из работы [13]. 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 $\sum_{i=1}^{N} w_1^i = 1$.

Алгоритм $Hedge(\beta)$
FOR $t = 1, 2, \ldots, T$
Распределитель вычисляет распределение экспертных стратегий:

$$\bar{p}_t = \frac{\bar{w}_t}{\sum_{i=1}^{N} w_t^i}. \tag{4.4}$$

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

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 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

$$\begin{aligned}
\sum_{i=1}^{N} w_{T+1}^i \leqslant \\
\leqslant \prod_{t=1}^{T}(1-(1-\beta)(\bar{p}_t \cdot \bar{l}_t)) \leqslant \\
\leqslant \exp\left(-(1-\beta)\sum_{t=1}^{T}(\bar{p}_t \cdot \bar{l}_t)\right).
\end{aligned}$$

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

$$w_{T+1}^i = w_1^i \prod_{t=1}^{T} \beta^{l_i} = w_1^i \beta^{L_T^i}. \tag{4.9}$$

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 «регрет» $\frac{\ln N}{1-\beta}$.

В работе [30] 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*

$$g(x) = \frac{1}{1+\sqrt{\frac{2}{x}}}.$$

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 преобразований приводит к нужному результату:

$$\begin{aligned}
L\frac{-\ln\beta}{1-\beta} + \frac{1}{1-\beta}R &\leqslant L\frac{1+\beta}{2\beta} + \frac{1}{1-\beta}R = \\
&= \frac{1}{2}L\left(1+\frac{1}{\beta}\right) + \frac{1}{1-\beta}R = \\
&= L + \frac{1}{2}L\sqrt{\frac{2\tilde{R}}{\tilde{L}}} + \frac{1}{1-\frac{1}{1+\sqrt{\frac{2\tilde{R}}{\tilde{L}}}}}R \leqslant \\
&\leqslant L + \sqrt{\frac{1}{2}\tilde{L}\tilde{R}} + R + R\sqrt{\frac{\tilde{L}}{2\tilde{R}}} \leqslant \\
&\leqslant L + \sqrt{2\tilde{L}\tilde{R}} + R.
\end{aligned}$$

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 4.4.

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. 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 [14], а также статью Калаи и Вемпала [18] и монографию Сеза-Бианки и Лугоши [21].

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 – 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 пред-

полагаем, 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 *Статистик* наблюдает кумулятивные потери экспертов $s_{1:t-1}^i = s_1^i + \cdots + s_{t-1}^i$ за прошлые шаги $< t$, $i = 1, \ldots N$. *Статистик* принимает решение следовать за одним из этих экспертов, скажем за экспертом i. В конце шага *Статистик* 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 *Статистика*, значительно превышающим потери каждого из экспертов.

Пусть потери каждого эксперта на шагах $t = 0, 1, \ldots, 6$ есть $s_{0,1,2,3,4,5,6}^1 = (\frac{1}{2}, 0, 1, 0, 1, 0, 1)$ and $s_{0.1,2,3,4,5,6}^2 = (0, 1, 0, 1, 0, 1, 0)$. 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.

Метод следования за возмущенным лидером был впервые предложен Ханнаном [14] в 1957г. Калаи и Вемпала [18] заново переоткрыли этот метод и опубликовали более простое доказательство. Они предложили название «Follow the Perturbed Leader – FPL» Дальнейшее изучение этого алгоритма проводилось Хуттером и Поландом [15], которые обобщили его на счетный класс экспертов.

Алгоритм FPL выдает в качестве предсказания номер эксперта i, для которого эяляется минимальной величина

$$s_{1:t-1}^i - \frac{1}{\epsilon}\xi^i,$$

где ϵ это *параметр обучения*, и ξ^i, $i = 1, \ldots N$, $t = 1, 2, \ldots$, есть последовательность независимых одинаково распределенных слу-

FPL алгоритм.

FOR $t = 1, \dots T$
Статистик выбирает эксперта, имеющего наименьшие возмущенные кумулятивные потери на шагах $< t$:

$$I_t = \text{argmin}_{i=1,2,\dots N} \left\{ s^i_{1:t-1} - \frac{1}{\epsilon_t} \xi^i \right\}.$$

Эксперт i несет потери s^i_t for $i = 1, \dots, N$.
Статистик несет потери $s_t = s^{I_t}_t$.
ENDFOR

Рис. 4.1: Псевдокод FPL алгоритма

чайных внличин распределенных согласно экспоненциальному закону с плотностью $p(x) = \exp\{-x\}$, $x \geqslant 0$.

Заметим, что можно выбрать эти случайные величины перед процессом обучения алгоритма.

Мы будем использовать свойства экспоненциального распределения: $P\{\xi > a\} = e^{-a}$ и $P\{\xi > a + b\} = e^b P\{\xi > a\}$ для всех неотрицательных значений a и b. Эти и другие свойства экспоненциального распределения предлагаются в виде задач в разделе 5.3.

На шаге t игры каждые из N экспертов несет потери $s^i_t \in [0, 1]$, $i = 1, \dots N$; 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 $s^i_0 = v_0 = 0$ для всех i и $\epsilon_0 = \infty$.

Псевдокод FPL алгоритма представлен на рис. 4.1.

Пусть $s_{1:T} = \sum_{t=1}^{T} s^{I_t}_t$ – кумулятивные потери алгоритма FPL на шагах $\leqslant T$.

В следующей ниже теореме дается верхняя оценка среднего значения кумулятивных потерь алгоритма FPL.

Теорема 4.5. *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:*

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + 2\sqrt{2T\ln N} \tag{4.13}$$

Доказательство. Анализ оптимальности алгоритма FPL основан на сравнении его потерь с потерями вспомогательного алгоритма IFPL (Infeasible FPL) (see рис. 4.2).

Алгоритм IFPL делает свои предсказания на основе использования величин $s^i_{1:t}$, $i = 1, \dots N$, которые еще неизвестны *Статистику* 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

$$l_t = E(s_t^{I_t}) \text{ и } r_t = E(s_t^{J_t}),$$
$$l_{1:T} = \sum_{t=1}^{T} l_t \text{ и } r_{1:T} = \sum_{t=1}^{T} r_t,$$

соответственно, где $s_t^{I_t}$ – потери алгоритма FPL на шаге t и $s_t^{J_t}$ – потери алгоритма IFPL на шаге t, символ E обозначает математическое ожидание.

Напомним, что $I_t = \operatorname{argmin}_i\{s^i_{1:t-1} - \frac{1}{\epsilon_t}\xi^i\}$ и $J_t = \operatorname{argmin}_i\{s^i_{1:t} - \frac{1}{\epsilon_t}\xi^i\}$.

Лемма 4.3. *Средние кумулятивные потери алгоритммов* FPL *и* IFPL *удовлетворяют неравенству*

$$l_{1:T} \leqslant r_{1:T} + \sum_{t=1}^{T} \epsilon_t \tag{4.14}$$

для всех T.

IFPL алгоритм.

FOR $t = 1, \dots T$
Статистик выбирает эксперта, имеющего наименьшие возмущенные кумулятивные потери на шагах $\leqslant t$:

$$J_t = \text{argmin}_{i=1,2,\dots N} \left\{ s^i_{1:t} - \frac{1}{\epsilon_t} \xi^i \right\}.$$

Эксперт i несет потери s^i_t for $i = 1, \dots, N$.
Статистик несет потери $s^{J_t}_t$.
ENDFOR

Рис. 4.2: Псевдокод IFPL алгоритма

Доказательство. Пусть $c_1, \dots c_N$ – поизвольные 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:

$$\begin{aligned}
P\{I_t = j | \xi^i = c_i, i \neq j\} = \\
= P\{s^j_{1:t-1} - \frac{1}{\epsilon_t}\xi^j \leqslant m_j | \xi^i = c_i, i \neq j\} = \\
= P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} - m_j) | \xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} - m_j + 1) | \xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t-1} + s^i_t - m_j) | \xi^i = c_i, i \neq j\} \leqslant \\
\leqslant e^{\epsilon_t} P\{\xi^j \geqslant \epsilon_t(s^j_{1:t} - m'_j) | \xi^i = c_i, i \neq j\} = \\
= e^{\epsilon_t} P\{s^j_{1:t} - \frac{1}{\epsilon_t}\xi^j \leqslant m'_j | \xi^i = c_i, i \neq j\} = \\
= e^{\epsilon_t} P\{J_t = j | \xi^i = c_i, i \neq j\}. \qquad (4.15)
\end{aligned}$$

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:

$$P\{I_t = j\} \leqslant e^{\epsilon_t} P\{J_t = j\}. \qquad (4.16)$$

для всех $t = 1, 2, \ldots$ и $j = 1, \ldots N$.

Суммируем (4.16) по $t = 1, \ldots, T$ и получим неравенство

$$l_t = E(s^{I_t}_t) = \sum_{j=1}^{T} s^j_t P\{I_t = j\} \leqslant e^{\epsilon_t} \sum_{j=1}^{T} s^j_t P\{J_t = j\} = e^{\epsilon_t} r_t.$$

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 4.4. *Математическое ожидание кумулятивных потерь алгоритма* IFPL *ограничено сверху*

$$r_{1:T} \leqslant \min_i s^i_{1:T} + \frac{\ln N}{\epsilon_T} \tag{4.17}$$

for all T.

Доказательство. Введем в этом доказательстве $\mathbf{s}_t = (s^1_t, \dots s^N_t)$ – вектор одношаговых потерь экспертов и $\mathbf{s}_{1:t} = (s^1_{1:t}, \dots s^N_{1:t})$ – вектор кумулятивных потерь экспертов. Пусть также $\xi = (\xi^1, \dots \xi^N)$ – вектор координатами которого являются экспоненциально распределенные случайные величины.

Рассмотрим вспомогательные векторы:

$$\tilde{\mathbf{s}}_t = \mathbf{s}_t - \xi\left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}}\right) \tag{4.18}$$

$$\tilde{\mathbf{s}}_{1:t} = \mathbf{s}_{1:t} - \frac{1}{\epsilon_t}\xi \tag{4.19}$$

при $t = 1, 2, \dots$.

Для произвольного вектора $\mathbf{s} = (s^1, \dots, s^N)$ и единичного вектора $\mathbf{d} = (0, \dots, 1, \dots, 0)$ обозначим

$$M(\mathbf{s}) = \operatorname{argmin}_{\mathbf{d}\in D}\{\mathbf{d}\cdot\mathbf{s}\},$$

где $D = \{(0, \dots, 1), \dots, (1, \dots, 0)\}$ – 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

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \tilde{\mathbf{s}}_t \leqslant M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T}. \tag{4.20}$$

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 (4.20) для шага T используя предположение индукции для шага $T-1$:

$$\begin{aligned} M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} = M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_T \geqslant \\ \geqslant M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_{1:T-1} + M(\tilde{\mathbf{s}}_{1:T-1}) \cdot \tilde{\mathbf{s}}_T \geqslant \\ \geqslant \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \tilde{\mathbf{s}}_t. \end{aligned}$$

Вспоминая определение (4.18) вектора $\tilde{\mathbf{s}}_t$, мы можем переписать (4.20) следующим образом:

$$\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \mathbf{s}_t \leqslant M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} + \sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{1:t}) \cdot \xi \left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}} \right) \tag{4.21}$$

Аналогично, используя определение (4.19) вектора $\tilde{\mathbf{s}}_{1:t}$ и то что критерий выбора координаты вновь был изменен, получаем неравенство

$$\begin{aligned} M(\tilde{\mathbf{s}}_{1:T}) \cdot \tilde{\mathbf{s}}_{1:T} \leqslant M(\mathbf{s}_{1:T}) \cdot \left(\mathbf{s}_{1:T} - \frac{\xi}{\epsilon_T} \right) = \\ = \min_{\mathbf{d} \in D} \{ \mathbf{d} \cdot \mathbf{s}_{1:T} \} - \frac{M(\mathbf{s}_{1:T}) \cdot \xi}{\epsilon_T}. \end{aligned} \tag{4.22}$$

По определению $(M(\mathbf{s}_{1:T}) \cdot \xi) = \xi^k$ для некоторог k.

Так как $E(\xi) = 1$ для экспоненциально распределенной случайной величины ξ, математическое ожидание вычитаемого члена в (4.22) равно

$$E\left(\frac{M(\mathbf{s}_{1:T}) \cdot \xi}{\epsilon_T}\right) = \frac{1}{\epsilon_T} E(\xi^k) = \frac{1}{\epsilon_T}. \tag{4.23}$$

Второй член (4.21) удовлетвояет

$$\sum_{t=1}^{T} (M(\tilde{\mathbf{s}}_{\mathbf{1}:\mathbf{t}}) \cdot \xi) \left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}}\right) \leqslant$$

$$\leqslant \sum_{t=1}^{T} \max_{1 \leqslant i \leqslant N} \xi^i \left(\frac{1}{\epsilon_t} - \frac{1}{\epsilon_{t-1}}\right) = \frac{1}{\epsilon_T} \max_{1 \leqslant i \leqslant N} \xi^i. \tag{4.24}$$

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:

$$0 \leqslant E(\max_{1 \leqslant i \leqslant N} \xi^i) \leqslant 1 + \ln N. \tag{4.25}$$

Действительно, для экспоненциально распределенных случйных величин ξ^i, $i = 1, \ldots N$, выполнено

$$P\{\max_i \xi^i \geqslant a\} = P\{\exists i (\xi^i \geqslant a)\} \leqslant$$

$$\leqslant \sum_{i=1}^{N} P\{\xi^i \geqslant a\} = N \exp\{-a\}. \tag{4.26}$$

Для произвольной неотрицательной случайной величины η выполнено

$$E(\eta) = \int_0^\infty P\{\eta \geqslant y\} dy. \tag{4.27}$$

Доказательство этого соотношение предоставляется читателю в виде задачи из раздела 5.3. Тогда по (4.26) 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 (4.25) математическое ожидание (4.24) ограничено сверху числом $\frac{1}{\epsilon_T}(1 + \ln N)$.

Комбинируя оценки (4.21)–(4.24) и (4.23), получим

$$r_{1:T} = E\left(\sum_{t=1}^{T} M(\tilde{\mathbf{s}}_{\mathbf{1:t}}) \cdot \mathbf{s_t}\right) \leqslant$$

$$\leqslant \min_i s^i_{1:T} + \frac{\ln N}{\epsilon_T}. \tag{4.28}$$

Лемма доказана. $\triangle$.

Завершим доказательство теоремы.

Неравенство (4.14) леммы 4.3 и неравенство (4.17) леммы 4.4 влекут неравенство

$$E(s_{1:T}) \leqslant \min_i s^i_{1:T} + a \sum_{t=1}^{T} \frac{1}{\sqrt{t}} + \frac{1}{a} \ln N\sqrt{T} \leqslant$$

$$\leqslant \min_i s^i_{1:T} + 2a\sqrt{T} + \frac{1}{a} \ln N\sqrt{T}. \tag{4.29}$$

for all T. Минимизируем (4.29) по a, получим оптимальное значение $a = \sqrt{2 \ln N}$. Таким образом, мы получили оценку (4.13) 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 неравенство 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 $\xi_1^t, \ldots, \xi_N^t$, $t = 1, 2, \ldots$, так, что все эти случайные величины рассматриваемые вместе независимы.

В алгоритме FPL на рис. 4.1 на шаге t мы будем возмущать каждого эксперта с помощью серии случайных величин $\xi_t^1, \ldots, \xi_t^N$. *Статистик* выбирает эксперта, имеющего наименьшие возмущенные кумулятивные потери на шагах $< t$:

$$I_t = \operatorname{argmin}_{i=1,2,\ldots N} \left\{ s_{1:t-1}^i - \frac{1}{\epsilon_t} \xi_t^i \right\}.$$

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 4.3 остается тем же, доказательство леммы 4.4 изменяется незначительно, надо только в неравенствах (4.21), (4.22) и (4.24) сразу рассмотреть математическое ожидание от обоих их частей и использовать то, что $E(\xi_t^i) = 1$ для всех i и t.

Следствие 4.1. *Для произвольного $\delta > 0$ с вероятностью $1 - \delta$ выполнено неравенство*

$$s_{1:T} \leqslant \min_i s_{1:T}^i + 2\sqrt{2T \ln N} + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}. \tag{4.30}$$

Алгоритм FPL *является асимптотически состоятельным:*

$$\limsup_{T \to \infty} \frac{1}{T} (s_{1:T} - \min_{i=1,\ldots N} s_{1:T}^i) \leqslant 0 \tag{4.31}$$

с вероятностью 1.

Доказательство. Для доказательства первого утверждения мы используем вариант неравенство Чернова (4.60) из следствия 4.5:

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 $\epsilon > 0$

$$P\left\{\sum_{i=1}^{T} X_i - E\sum_{i=1}^{T} X_i > \epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{T}\right). \tag{4.32}$$

Полагаем $\delta = \exp\left(-\frac{2\epsilon^2}{T}\right)$. Отсюда $\epsilon = \sqrt{\frac{T}{2}\ln\frac{2}{\delta}}$. При $X_t = s_t$ из неравенства (4.32) 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 (4.13) теоремы 4.5 получаем неравенство (4.30).

Для доказательства утверждения (4.31) мы применим другой вариант (4.61) неравенства Чернова

$$P\left\{\left|\frac{1}{T}\sum_{i=1}^{T}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2T\epsilon^2\right). \tag{4.33}$$

Здесь полагаем $X_t = s_t$. Так как 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 (4.13) теоремы 4.5 получаем неравенство (4.31) для верхнего предела.

Следствие 4.1 верно и в более общем случае «адаптивно враждебных» экспертов, потери которых на каждом шаге t могут зависеть от значений случайных величин $s_{t'}$ при $t' < t$. В этом случае случайные величины $X - t = s_t$ не будут независимыми, но величины $X_t - E(X_t) = s_t - E(s_t)$ образуют мартингал-разности, и мы можем применить соответствующее неравенство Хефдинга–Азумы (4.65) и усиленный мартингальный закон больших чисел (4.63).

4.4. Алгоритм экспоненциального взвешивания экспертных решений

Напомним, что $\mathcal{R}$ – это множество всех вещественных (действительных) чисел. Пусть Ω – множество исходов, Γ – множество решений или предсказаний (прогнозов), Θ – множество параметров (экспертных стратегий, экспертов). Предполагаем, что Θ – конечное множество, $\Gamma \subseteq \mathcal{R}^n$. В этой главе Ω – произвольное множество объектов любой природы.

Оценка принятого решения (или предсказания) $\gamma \in \Gamma$ при исходе $\omega \in \Omega$ производится с помощью функции потерь $\lambda(\omega, \gamma)$, принимающей неотрицательные действительные значения. Далее мы будем предполагать, что значения функции потерь лежат в отрезке $[0, 1]$.

Рассматривается игра с полной информацией между тремя игроками: *Статистик* (Learner), *Множество экспертов* (Decision Pool), *Природа* (Nature). Игра происходит в соответствии со следующим протоколом.
FOR $t = 1, 2, \ldots$
Эксперты $\theta \in \Theta$ делают предсказания: $\xi_t^\theta \in \Gamma$.
Статистик принимает свое решение: $\gamma_t \in \Gamma$.
Природа анонсирует исход: $\omega_t \in \Omega$.
Эксперты $\theta \in \Theta$ вычисляют свои суммарные потери на шаге t игры:

$$L_t(\theta) = L_{t-1}(\theta) + \lambda(\omega_t, \xi_t^\theta).$$

Статистик вычисляет свои суммарные потери на шаге t игры:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

Здесь $L_0(\theta) = L_0 = 0$ для всех θ.
ENDFOR

Протокол определяет порядок действий (ходы) игроков. Каждый игрок может при определении своего действия использовать всю информацию, которая известна к началу его хода.

Целью *Статистика* является выбор такой последовательности прогнозов $\gamma_1, \gamma_2, \ldots$, чтобы для каждого t его суммарные потери L_t были бы с некоторой степенью точности не больше чем

суммарные потери наиболее эффективного эксперта, т.е. не больше чем $\inf\limits_{\theta} L_t(\theta)$.

Природа может быть *враждебной* по отношению к *Статистику*: выдаваемые ею исходы ω_t могут зависеть от прогнозов γ_t, так как *Природа* выдает исход ω_t тогда, когда прогноз γ_t уже выдан *Статистиком*.

Количественной оценкой метода прогнозирования является кумулятивная ошибка, или кумулятивный *регрет* (cumulative *regret*) относительно эксперта θ :

$$R_{\theta,T} = \sum_{t=1}^{T} (\lambda(\omega_t, \gamma_t) - \lambda(\omega_t, \xi_t^{\theta})) = L_T - L_T(\theta). \tag{4.34}$$

Цель метода предсказания заключается в том, чтобы

$$\limsup_{T\to\infty} \frac{1}{T}(L_T - \inf_{\theta} L_T(\theta)) \leqslant 0. \tag{4.35}$$

Заметим, что здесь не исключается тот случай, когда *Статистик* может предсказывать даже лучше, чем эксперт, имеющий наименьшие потери.

Прогнозы будут элементами n-мерного евклидового пространства $\mathcal{R}^n$. Таким образом, их можно складывать и умножать на вещественные числа.

Подножество Z 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 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

$$h(pz' + (1-p)z'') \leqslant ph(z') + (1-p)h(z''). \tag{4.36}$$

Заданы множества исходов Ω и множество прогнозов Γ. Задана некоторая функция потерь $\lambda(\omega, \gamma)$.

Пусть множество экспертов конечно: $\Theta = \{1, \ldots, N\}$. В этом разделе предполагаем, что множество прогнозов Γ – выпуклое

подмножество $\mathcal{R}^n$, а функция потерь $\lambda(\omega, \gamma)$ является выпуклой по прогнозу γ.

Простейший алгоритм взвешивания экспертных прогнозов вычисляет прогноз *Статистика* по формуле

$$\gamma_t = \frac{\sum\limits_{i=1}^{N} w_{i,t-1}\xi_t^i}{\sum\limits_{j=1}^{N} w_{j,t-1}} = \sum_{i=1}^{N} w_{i,t-1}^* \xi_t^i, \tag{4.37}$$

где $\xi_t^i \in \mathcal{R}^n$ – прогноз i-го эксперта на шаге t, $w_{i,t-1}$, $i = 1, \ldots, N$, – веса, приписанные экспертам на шаге t,

$$w_{i,t-1}^* = \frac{w_{i,t-1}}{\sum\limits_{j=1}^{N} w_{j,t-1}} \tag{4.38}$$

– нормированные веса. Так как Γ – выпуклое множество, $\gamma_t \in \Gamma$ для всех t.

В алгоритме экспоненциального взвешивания в качестве весов экспертов берут величины

$$w_{i,t-1} = e^{-\eta L_{t-1}(i)}, \tag{4.39}$$

$i = 1, \ldots, N$, где $L_{t-1}(i)$ – суммарные потери i-го эксперта на шагах от 1 до $t-1$, $\eta > 0$ – некоторый параметр – *параметр обучения*.

В этом случае прогноз *Статистика* вычисляется по формуле

$$\gamma_t = \frac{\sum\limits_{i=1}^{N} \xi_t^i e^{-\eta L_{t-1}(i)}}{\sum\limits_{j=1}^{N} e^{-\eta L_{t-1}(j)}} = \sum_{i=1}^{N} w_{i,t-1}^* \xi_t^i, \tag{4.40}$$

где

$$w_{i,t-1}^* = \frac{e^{-\eta L_{t-1}(i)}}{\sum\limits_{j=1}^{N} e^{-\eta L_{t-1}(j)}} \tag{4.41}$$

– вес эксперта i, $i = 1, \ldots, N$.

Оптимальные свойства алгоритма экспоненциального взвешивания изучаются в следующей теореме.

Теорема 4.6. *Допустим, что функция потерь $\lambda(\omega, \gamma)$ является выпуклой по второму аргументу и принимает значения в $[0, 1]$. Тогда для любых $\eta > 0$, T и $\omega_1, \ldots, \omega_T \in \Omega$ кумулятивная ошибка алгоритма экспоненциального взвешивания удовлетворяет неравенству*

$$L_T - \min_{i=1,\ldots,N} L_T(i) \leqslant \frac{\ln N}{\eta} + \frac{T\eta}{8}. \tag{4.42}$$

При $\eta = \sqrt{8 \ln N / T}$ верхняя оценка имеет вид: $\sqrt{\frac{1}{2} T \ln N}$.

Доказательство. Определим вспомогательные величины

$$W_t = \sum_{i=1}^{N} w_{i,t} = \sum_{i=1}^{N} e^{-\eta L_t(i)}, \tag{4.43}$$

$W_0 = N$.

В доказательстве будет использоваться *неравенство Хеффдинга*, которое сформулировано и доказано в разделе 4.7 в виде леммы 4.5. Эта лемма утверждает следующее: пусть X – случайная величина и $a \leqslant X \leqslant b$. Тогда для произвольного $s \in \mathcal{R}$

$$\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8},$$

где E – математическое ожидание.

Доказательство теоремы будет основано на сравнении нижней и верхней оценок величины $\ln \frac{W_T}{W_0}$.

Нижняя оценка получается следующим образом. Заметим, что так как $w_{i,0} = 1$ для всех $i = 1, \ldots, N$,

$$\begin{aligned} \ln \frac{W_T}{W_0} &= \ln \left(\sum_{i=1}^{N} e^{-\eta L_T(i)} \right) - \ln N \geqslant \\ &\geqslant \ln \left(\max_{i=1,\ \ldots,\ N} e^{-\eta L_T(i)} \right) - \ln N = \\ &= -\eta \min_{i=1,\ \ldots,\ N} L_T(i) - \ln N. \end{aligned} \tag{4.44}$$

Верхняя оценка величины $\ln \frac{W_T}{W_0}$ получается с помощью следующих выкладок. Имеем для произвольного t

$$\ln \frac{W_t}{W_{t-1}} = \ln \frac{\sum\limits_{i=1}^{N} e^{-\eta\lambda(\omega_t,\xi_t^i)} e^{-\eta L_{t-1}(i)}}{\sum\limits_{i=1}^{N} e^{-\eta L_{t-1}(i)}} =$$

$$= \ln \frac{\sum\limits_{i=1}^{N} w_{i,t-1} e^{-\eta\lambda(\omega_t,\xi_t^i)}}{\sum\limits_{j=1}^{N} w_{j,t-1}} = E(e^{-\eta\lambda(\omega_t,\xi_t^i)}), \tag{4.45}$$

где математическое ожидание рассматривается относительно распределения вероятностей:

$$w_{i,t-1}^* = \frac{w_{i,t-1}}{\sum\limits_{j=1}^{N} w_{j,t-1}},$$

$i = 1, \ \ldots, \ N$.

Применим неравенство Хефдинга (4.58), в котором $a = 0$, $b = 1$, случайная величина X принимает значение $\lambda(\omega_t, \xi_t^i)$ с вероятностью $w_{i,t-1}^*$. Используем также выпуклость функции потерь $\lambda(\omega, \gamma)$ по второму аргументу. В результате получаем следующие неравенства:

$$\ln \frac{W_t}{W_{t-1}} \leqslant -\eta \frac{\sum\limits_{i=1}^{N} w_{i,t-1}\lambda(\omega_t, \xi_t^i)}{\sum\limits_{j=1}^{N} w_{j,t-1}} + \frac{\eta^2}{8} \leqslant$$

$$\leqslant -\eta\lambda\left(\omega_t, \frac{\sum\limits_{i=1}^{N} w_{i,t-1}\xi_t^i}{\sum\limits_{j=1}^{N} w_{j,t-1}}\right) + \frac{\eta^2}{8} =$$

$$= -\eta\lambda(\omega_t, \gamma_t) + \frac{\eta^2}{8}, \tag{4.46}$$

где γ_t – прогноз по алгоритму экспоненциального смешивания (4.40).

Отсюда, суммируя (4.46) по $t = 1, \ldots, T$, получим

$$\ln \frac{W_T}{W_0} = \sum_{i=1}^{T} \ln \frac{W_t}{W_{t-1}} \leqslant -\eta L_T + \frac{\eta^2}{8} T. \tag{4.47}$$

Используя нижнюю оценку (4.44) и верхнюю оценку (4.47), получим

$$L_T \leqslant \min_{i=1,\ldots,\ N} L_T(i) + \frac{\ln N}{\eta} + \frac{\eta}{8} T. \tag{4.48}$$

Теорема доказана. $\triangle$

Напомним, что при $\eta = \sqrt{8 \ln N / T}$ верхняя оценка имеет вид: $\sqrt{\frac{1}{2} T \ln N}$. Очевидный недостаток при выборе параметра η заключается в том, что для его выбора надо фиксировать величину T – горизонт, до которого делается предсказание.

Значительно более лучшая оценка, основанная на использовании переменного параметра обучения, приведена в следующем разделе.

4.5. Алгоритм экспоненциального взвешивания с переменным параметром обучения

Рассмотрим технически более сложную конструкцию алгоритма экспоненциального взвешивания с переменным параметром обучения, предложенную Алексеем Черновым [8].

Далее, $L_T(i)$ – суммарные потери i-го эксперта за первые T шагов, $\widehat{L}_T$ – суммарные потери *Статистика*.

Множество экспертов конечно: $\Theta = \{1, \ldots, N\}$, множество прогнозов Γ – выпуклое подмножество $\mathcal{R}^n$, а функция потерь $\lambda(\omega, \gamma)$ является выпуклой по прогнозу γ.

Модифицируем алгоритм экспоненциального взвешивания – в качестве весов экспертов берем величины

$$w_{i,t-1} = e^{-\eta_t L_{t-1}(i)},$$

$i = 1, \ldots, N$, где $L_{t-1}(i)$ – суммарные потери i-го эксперта на шагах от 1 до $t-1$, $\eta_t > 0$ – переменный параметр обучения.

В этом случае можно добиться равномерной по шагам верхней оценки ошибки предсказания.

Теорема 4.7. *Для любой последовательности положительных вещественных чисел $\eta_1 \geqslant \eta_2 \geqslant \ldots$, для любого $n \geqslant 1$ и для любых $\omega_1, \ldots, \omega_n \in \Omega$, ошибка (регрет) алгоритма экспоненциального взвешивания с переменным параметром обучения η_t удовлетворяет неравенству*

$$\widehat{L}_T - \min_{i=1,\ldots,N} L_T(i) \leqslant \frac{\ln N}{\eta_T} + \frac{1}{8}\sum_{t=1}^{T} \eta_t \,. \tag{4.49}$$

В частности, для $\eta_t = \sqrt{\frac{4\ln N}{t}}$, $t = 1, \ldots, n$, выполнено

$$\widehat{L}_T - \min_{i=1,\ldots,N} L_T(i) \leqslant \sqrt{T \ln N}.$$

Доказательство. На шаге t *Статистик* вычисляет свое предсказание по формуле $\widehat{p}_t = \sum_{i=1}^{N} \xi_t^i w_{i,t-1}/W_{t-1}$, где $w_{i,t-1} = e^{-\eta_t L_{t-1}(i)}$ и $W_{t-1} = \sum_{j=1}^{N} w_{j,t-1}$. Из выпуклости функции λ по второму аргументу получаем

$$\lambda(\omega_t, \widehat{p}_t) \leqslant \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}}\, \lambda(\omega_t, \xi_t^i) \,.$$

Применяем неравенство Хефдинга и получаем

$$e^{-\eta_t \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}}\, \lambda(\omega_t, \xi_t^i)} \geqslant \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} e^{-\eta_t \lambda(\omega_t, \xi_t^i) - \eta_t^2/8}$$

Перепишем это неравенство в виде

$$e^{-\eta_t \lambda(\omega_t, \widehat{p}_t)} \geqslant \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} e^{-\eta_t \lambda(\omega_t, \xi_t^i) - \eta_t^2/8} \,. \tag{4.50}$$

Введем вспомогательные величины

$$s_{i,t-1} = e^{-\eta_{t-1}L_{t-1}(i)+\eta_{t-1}\left(\widehat{L}_{t-1}-\frac{1}{8}\sum_{k=1}^{t-1}\eta_k\right)}$$

и заметим, что

$$\frac{w_{i,t-1}}{W_{t-1}} = \frac{\frac{1}{N}\left(s_{i,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}}}{\sum_{j=1}^{N}\frac{1}{N}\left(s_{j,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}}}\,. \tag{4.51}$$

Докажем, что $\sum_{j=1}^{N}\frac{1}{N}s_{j,t} \leqslant 1$ математической индукцией по t. При $t=0$ это утверждение выполнено, так как $s_{i,0}=1$ для всех i. Допустим, что $\sum_{j=1}^{N}\frac{1}{N}s_{j,t-1} \leqslant 1$. Тогда

$$\sum_{j=1}^{N}\frac{1}{N}\left(s_{j,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}} \leqslant \left(\sum_{j=1}^{N}\frac{1}{N}s_{j,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}} \leqslant 1\,, \tag{4.52}$$

так как функция $x \mapsto x^{\alpha}$ вогнутая и монотонная по $x \geqslant 0$ и $\alpha \in [0,1]$ и так как $0 \leqslant \eta_t \leqslant \eta_{t-1}$. Используя (4.52) в качестве границы знаменателя из правой части (4.51), получим $w_{i,t-1}/W_{t-1} \geqslant \left(s_{i,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}}/N$. Комбинируя это с (4.50), получим

$$e^{-\eta_t\lambda(\widehat{p}_t,y_t)} \geqslant \sum_{i=1}^{N}\frac{1}{N}\left(s_{i,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}}e^{-\eta_t\lambda(\omega_t,\xi_t^i)-\eta_t^2/8}\,.$$

Заметим, что

$$s_{i,t} = \left(s_{i,t-1}\right)^{\frac{\eta_t}{\eta_{t-1}}}e^{-\eta_t\lambda(\omega_t,\xi_t^i)+\eta_t\lambda(\omega_t,\widehat{p}_t)-\eta_t^2/8}.$$

Отсюда получим $\sum_{j=1}^{N}\frac{1}{N}s_{j,t} \leqslant 1$.

Так как для произвольного i выполнено

$$\frac{1}{N}s_{i,n} \leqslant \sum_{j=1}^{N}\frac{1}{N}s_{j,n} \leqslant 1,$$

получаем

$$-\eta_n L_n(i) + \eta_n\left(\widehat{L}_n - \frac{1}{8}\sum_{k=1}^{n}\eta_k\right) \leqslant \ln N,$$

отсюда следует (4.49). □

4.6. 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.

Напомним протокол детерминированной игры на предсказания с использованием экспертных прогнозов.
Пусть $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ldots, N$.
FOR $t = 1, 2, \ldots$
Эксперт i выбирает прогноз: $\xi_t^i \in \Gamma$, $i = 1, \ldots, N$.
Статистик выбирает прогноз: $\gamma_t \in \Gamma$.
Природа выбирает исход: $\omega_t \in \Omega$.
Эксперт i вычисляет свои суммарные потери на шаге t игры:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i),$$

где $i = 1, \ldots, N$.
Статистик вычисляет свои суммарные потери на шаге t игры:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

Каждый игрок может при определении своего действия использовать всю информацию, которая известна к началу его хода.

Потери *Статистика* на шагах $t = 1, \ldots, T$ равны

$$L_T = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t).$$

Потери эксперта i на шагах $t = 1, \ldots, T$ равны

$$L_T(i) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i).$$

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 детерминированной стратегии *Статистика* $\gamma_1, \gamma_2, \ldots$ существует такая последовательность исходов $\omega_1, \omega_2, \ldots$, что потери *Статистика* максимальны, т.е. $L_T = T$ для всех T. Действительно, *Природа* может определить для всех $t = 1, 2, \ldots$:

$$\omega_t = \begin{cases} 2, \text{ если } \gamma_t \leqslant 1, \\ 1 \text{ в противном случае.} \end{cases}$$

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 *Статистик* 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 произвольной 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 стратегии *Статистика* «адаптивно враждебная» *Природа* может предоставить последовательность $\omega_1, \omega_2, \ldots$ такую, что

$$L_T - \min_{i=1,2} L_t(i) \geqslant T/2$$

для всех T.

Приведенный пример показывает, что для некоторых невыпуклых функциях потерь *Природа* может выдавать последова-

тельность исходов так, что при любых детерминированных действиях *Статистика* его потери за любой период игры T имеют ошибку предсказания $\geqslant T/2$.

Этот же пример подходит для $\Omega = \Gamma = \{1, 2\}$.

Данную проблему *Статистик* 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 теперь на каждом шаге t игры *Статистик* выдает прогноз в виде смешанной стратегии – распределения вероятностей $\bar{p}_t = \{p_{1,t},\ \ldots,\ p_{N,t}\}$ на множестве экспертов $\{1,\ \ldots,\ N\}$.

Мы введем еще одного игрока – *Генератора случайных чисел*, который будет генерировать элементы множества $\{1,\ \ldots,\ N\}$ согласно заданному ему распределению вероятностей.

Протокол рандомизированной игры имеет следующий вид.
Пусть $L_0 = 0$, $L_0(i) = 0$, $i = 1,\ \ldots,\ N$.
FOR $t = 1, 2,\ \ldots$
Эксперт i выбирает прогноз: $\xi_t^i \in \Gamma$, $i = 1,\ \ldots,\ N$.
Статистик выдает распределение вероятностей: $p_{1,t},\ \ldots,\ p_{N,t}$ на множестве экспертов $\{1,\ \ldots,\ N\}$.
Природа выбирает исход: $\omega_t \in \Omega$.
Генератор случайных чисел выбирает эксперта: $i_t \in \{1,\ \ldots,\ N\}$ с вероятностью $p_{i,t}$.
Эксперт i вычисляет свои суммарные потери на шаге t игры:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i),$$

где $i = 1,\ \ldots,\ N$.
Статистик вычисляет свои суммарные потери на шаге t игры:

$$L_t = L_{t-1} + \lambda(\omega_t, \xi_t^{i_t}).$$

ENDFOR

Генератор случайных чисел относительно произвольного распределения вероятностей можно реализовать на основе генерато-

ра равномерно распределенных случайных чисел следующим образом.

Вводим случайные переменные I_t, так что $I_t = i$ тогда и только тогда, когда

$$U_t \in \left[\sum_{j=1}^{i-1} p_{j,t}, \sum_{j=1}^{i} p_{j,t} \right],$$

где величины $U_1, U_2, \ldots$ – независимые равномерно распределенные в единичном отрезке случайные величины. Из определения следует, что $P\{I_t = i\} = p_{i,t}$ для всех t.

В такой игре потери *Статистика* $\lambda(\omega_t, \xi_t^{I_t})$ являются случайной величиной. В этом случае качество *Статистика* измеряется также случайной величиной – случайной ошибкой предсказания:

$$L_T - \min_{i=1,\ldots,N} L_T(i) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^{I_t}) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \quad (4.53)$$

Рассмотрим постановку, при которой целью *Статистика* является минимизация математического ожидания ошибки (4.53) :

$$\begin{aligned} E(L_T - \min_{i=1,\ldots,N} L_T(i)) = \\ = E(L_T) - \min_{i=1,\ldots,N} L_T(i) = \\ = \sum_{t=1}^{T} E(\lambda(\omega_t, \xi_t^{I_t})) - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i) = \\ = \sum_{t=1}^{T} \sum_{i=1}^{N} \lambda(\omega_t, \xi_t^i) p_{i,t} - \min_{i=1,\ldots,N} \sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i). \end{aligned} \quad (4.54)$$

Будем вычислять распределение вероятностей на множестве экспертов с помощью соотношений (4.4) из раздела 4.2. На шаге t определим

$$p_{i,t} = \frac{\beta^{\sum_{s=1}^{t-1} l_s^i}}{\sum_{j=1}^{N} \beta^{\sum_{s=1}^{t-1} l_s^j}}, \quad (4.55)$$

где $l_s^i = \lambda(\omega_s, \xi_s^i)$ при $i = 1, \ldots, N$, $0 < \beta < 1$.

Алгоритм вычисления вероятностных стратегий (4.55) будет называться *вероятностным алгоритмом экспоненциального взвешивания*.

Из леммы 4.2 следует

Теорема 4.8. *Пусть L_T – случайные кумулятивные потери алгоритма $Hedge(\beta)$ за T шагов при $\beta = g(T/\ln N)$, где*

$$g(x) = \frac{1}{1 + \sqrt{\frac{2}{x}}}.$$

Тогда имеет место оценка математического ожидания случайных потерь вероятностного алгоритма экспоненциального взвешивания

$$E(L_T) \leqslant \min_i L_T(i) + \sqrt{2T \ln N} + \ln N. \tag{4.56}$$

Недостатком этой оценки является то, что параметр β зависит от горизонта T. См. также комментарий в конце раздела 4.4.

Заметим, что для любого t вектор вероятностей $\{p_{1,t}, \ldots, p_{N,t}\}$ на множестве экспертов зависит от последовательности предшествующих исходов $\omega_1, \ldots, \omega_{t-1}$, выдаваемых *Природой*, а последовательность $\omega_1, \ldots, \omega_t$, в свою очередь может зависеть от последовательности распределений $\{p_{1,s}, \ldots, p_{N,s}\}$, $s = 1, \ldots t$, выдаваемых *Статистиком*.

По теореме Ионеско–Тульчи (см. [3]) существует распределение вероятностей $\mathcal{P}$, определенное на бесконечных траекториях выбираемых экспертов $i_1, i_2, \ldots$, где $i_t \in \{1, \ldots N\}$, при всех $t = 1, 2, \ldots$, порожденное всеми распределениями $\{p_{1,t}, \ldots, p_{N,t}\}$, $t = 1, 2, \ldots$

Используя следствие 4.7 из неравенства Хефдинга–Азумы (лемма 4.6 ниже), можно получить следующее следствие из неравенства 4.56.

Следствие 4.2. *Пусть $0 < \delta < 1$. Тогда ошибка (регрет) вероятностного алгоритма экспоненциального взвешивания с веро-*

ятностью $1-\delta$ *удовлетворяет неравенству*

$$\sum_{t=1}^{T}\lambda(\omega_t,\xi_t^{I_t})-\min_{i=1,\ldots,\ N}\sum_{t=1}^{T}\lambda(\omega_t,\xi_t^i)\leqslant$$
$$\leqslant\sqrt{2T\ln N}+\ln N+\sqrt{\frac{1}{2}T\ln\frac{1}{\delta}}.$$

Доказательство. По определению случайные величины

$$X_t=\lambda(\omega_t,\xi_t^{I_t})-E(\lambda(\omega_t,\xi_t^{I_t}))=$$
$$=\lambda(\omega_t,\xi_t^{I_t})-\sum_{i=1}^{N}\lambda(\omega_t,\xi_t^i)p_{i,t}$$

представляют собой последовательность ограниченных мартингал-разностей. Поэтому для их сумм

$$S_T=\sum_{t=1}^{T}X_t$$

по следствию 4.7 выполнено неравенство

$$P\{S_T>c\}\leqslant e^{-\frac{2c^2}{T}},$$

где c – произвольное положительное число (см. неравенство (4.63)). 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 (4.54) и (4.56). $\triangle$

Пусть в игре на предсказания с использованием экспертов $i=1,\ldots,\ N$ некоторый предсказатель выдает прогнозы $\xi_1,\xi_2,\ldots,$ а i-й эксперт выдает прогнозы $\xi_1^i,\xi_2^i,\ldots.$

Предсказатель называется состоятельным по Ханнану, если с $\mathcal{P}$-вероятностью 1

$$\limsup_{T\to\infty}\frac{1}{T}\left(\sum_{t=1}^{T}\lambda(\omega_t,\xi_t)-\min_{i=1,\ldots,\ N}\sum_{t=1}^{T}\lambda(\omega_t,\xi_t^i)\right)\leqslant 0. \qquad (4.57)$$

Следствие 4.2 влечет следующее следствие.

Следствие 4.3. *Вероятностный алгоритм экспоненциального взвешивания является состоятельным по Ханнану.*

При этом прогнозы предсказателя определяются как $\xi_t = \xi_t^{I_t}$ для всех t.

Поясним, как соотносится пример, приведенный в начале этого раздела, со следствием 4.3. В примере *Статистик* 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 $\mathcal{P}$-вероятность выбрать эту траекторию, а также любую другую, на которой нарушается условие (4.57), равна 0.

Сравнение с теоремой 4.2 показывает, что радомизированный алгоритм, примененный к простой функции потерь, имеет примерно в два раза меньшую оценку ошибки, чем детерминированный алгоритм взвешенного большинства WMA.

4.7. Некоторые замечательные неравенства

Приведем несколько замечательных неравенств, которые неоднократно используются в доказательствах теорем. Основным таким неравенством будет неравенство Хефдинга.

Лемма 4.5. *Пусть X – случайная величина и $a \leqslant X \leqslant b$. Тогда для произвольного $s \in \mathcal{R}$*

$$\ln E(e^{sX}) \leqslant sE(X) + \frac{s^2(b-a)^2}{8}. \tag{4.58}$$

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, разъясняющих значение этого неравенства.

Следствие 4.4. *Пусть X – случайная величина такая, что $P\{a \leqslant X \leqslant b\} = 1$. Тогда*

$$P\{|X - E(X)| > c\} \leqslant 2e^{-\frac{2c^2}{(b-a)^2}}. \tag{4.59}$$

Доказательство. Предварительно напомним неравенство Маркова. Пусть X – случайная величина, $X \geqslant 0$. Из

$$E(X) = \int X dP \geqslant \int_{\{X>c\}} X dP \geqslant cP\{X > c\}$$

следует, что $P\{X > c\} \leqslant E(X)/c$.

Используя это неравенство и неравенство (4.58), 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. 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 4.5. *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 $\epsilon > 0$:*

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i > \epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right), \qquad (4.60)$$

[1] Для удобства иногда используем обозначение $\exp(x) = e^x$.

а также

$$P\left\{\sum_{i=1}^{n} X_i - E\sum_{i=1}^{n} X_i < -\epsilon\right\} \leqslant \exp\left(-\frac{2\epsilon^2}{\sum\limits_{i=1}^{n}(b_i-a_i)^2}\right).$$

Доказательство. Доказательство аналогично доказательству следствия 4.4. Из неравенства Маркова и неравенства (4.58) 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 4.6. *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

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| > \epsilon\right\} \leqslant 2\exp\left(-2n\epsilon^2\right). \tag{4.61}$$

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–Азумы.

Лемма 4.6. *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 $X_1, \ldots, X_i$, и некоторой последовательности положительных констант c_i. Если $S_k = \sum\limits_{i=1}^{k} V_i$, то для любого $s > 0$*

$$E(e^{sS_n}) \leqslant e^{(s^2/8)\sum\limits_{i=1}^{k} c_i^2}.$$

Доказательство. Имеем

$$\begin{aligned} E(e^{sS_n}) = E(e^{sS_{n-1}}E(e^{sV_n}|X_1, \ldots, X_{n-1})) \leqslant \\ \leqslant E(e^{sS_{n-1}}e^{s^2c_n^2/8}) = \\ = e^{s^2c_n^2/8}E(e^{sS_{n-1}}). \end{aligned} \tag{4.62}$$

Здесь при переходе от первой строки ко второй была использована лемма 4.5.

Результат леммы получается путем итерации неравенства (4.62). $\triangle$

Следующее следствие доказывается аналогично следствию 4.4.

Следствие 4.7. *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* $X_1, \ldots, X_i$, *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 (4.58). 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

$$P\{S_n > c\} \leqslant \exp\left(-\frac{2c^2}{\sum\limits_{i=1}^{n} c_i^2}\right). \tag{4.63}$$

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 4.8. *В условиях следствия 4.7, где к тому же $c_i = 1$ для всех i, получаем*

$$P\left\{\frac{1}{n}|S_n| > c\right\} \leqslant 2e^{-2nc^2}. \tag{4.64}$$

Так как ряд экспонент в правой части неравенства (4.64) сходится, по лемме Бореля–Кантелли получим

Следствие 4.9. *В условиях следствия 4.7, 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:*

$$P\left\{\lim_{n\to\infty}\frac{S_n}{n} = 0\right\} = 1.$$

4.8. Задачи и упражнения

1. Построить вариант алгоритма большинства для случая года ммеется эксперт, про которого извество, что он делает не более k ошибок. Получить оценку числа его ошибок.

2. Допустим, что потери экспертов равны 0 или 1. Пусть s – эксперт, имеющий минимальные кумулятивные потери L_T^s, L_T – потери предсказателя в алгоритме Hedge. Доказать, что $L_T \geqslant L_T^s$.

3. Рассмотрим протокол игры на предсказания с использованием экспертов, в котором *Природа* выдает последовательность $0^T(01)^T1^T$. Имеется три эксперта, каждй из которых выдает постоянное предсказание: *Эксперт 1* всегда предсказывает $\xi_t^1 = 0$ для всех $t = 1, \ldots, 4T$, *Эксперт 2* предсказывает $\xi_t^1 = 1$ ля всех $t = 1, \ldots, 4T$, *Эксперт 3* предсказывает $\xi_t^1 = 1/2$ 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.

4. 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.

5. Доказать, что для любой неотрицательной случайной величины η с плотностью распределения $p(t)$ выполнено соотношение:

$$E(\eta) = \int_0^\infty P\{\eta \geqslant y\} dy.$$

Замечание. 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 $E(\eta) = \int_0^\infty tp(t)dt$.

6. Провести доказательство леммы 4.4 для того случая, когда на каждом шаге t в алгоритмах FPL и IFPL для рандомизации используется вся серия случайных величин $\xi_t^1, \ldots, \xi_t^N, t = 1, 2, \ldots$.

Глава 5

Усиление простых классификаторов – бустинг

В этой главе рассматривается метод усиления простых классификаторов, который называется *бустинг* (Boosting). Этот метод основан на комбинировании примитивных «слабых» классификаторов в один «сильный». Под «силой» классификатора в данном случае подразумевается эффективность (качество) решения задачи классификации, которое обычно измеряется средним числом ошибок классификации на обучающей выборке.

Будет изучаться алгоритм AdaBoost (от английских слов «адаптивность» и «усиление»), предложенный Фройндом и Шапире [13]. Этот алгоритм был успешно использован во многих областях, в частности для задачи 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.

5.1. 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, $S = ((\bar{x}_1, y_1),\ \ldots,\ (\bar{x}_l, y_l))$, где $\bar{x}_i \in X$ и $y_i \in Y$. Мы предполагаем, что $Y = \{0, 1\}$, $X = \mathcal{R}^n$ – подмножество n-мерного арифметического векторного пространства. Мы также предполагаем, что для всех i пары $(\bar{x}_i, y_i)$ 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

$$\epsilon = D\{i : h(\bar{x}_i) \neq y_i\} = \sum_{i:h(\bar{x}_i)\neq y_i} D(i).$$

В частности, при распределении $D(i) = 1/l$ ошибка обучения равна доле числа неправильных классификаций объектов:

$$\epsilon = |\{i : h(\bar{x}_i) \neq y_i\}|/l.$$

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 $(\bar{x}_i, y_i)$ 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. Будет приведен алгоритм AdaBoost (предложенный Фройндом и Шапире [13]), который 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).

Алгоритм AdaBoost

Вход алгоритма: выборка $S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l))$, 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 :

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(\bar{x}_i) - y_i|.$$

4) Полагаем $\beta_t = \epsilon_t/(1 - \epsilon_t)$.

5) Определим адаптированные веса при $i = 1, \ldots, l$:

$$w_{t+1}^i = w_t^i \beta_t^{1-|h_t(\bar{x}_i)-y_i|}.$$

ENDFOR

Результат работы алгоритма: выдать гипотезу – индикаторную функцию:

$$h(\bar{x}) = \begin{cases} 1, \text{ если } f(\bar{x}) \geqslant \frac{1}{2}, \\ 0 \text{ в противном случае,} \end{cases}$$

где пороговая функция f определяется в виде линейной комбинации гипотез алгоритма WeakLearn

$$f(\bar{x}) = \sum_{t=1}^{T} q_t h_t(\bar{x}),$$

с весами

$$q_t = \frac{\ln(1/\beta_t)}{\sum\limits_{t=1}^{T} \ln(1/\beta_t)},$$

при $t = 1, \ldots, T$.

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 $l_t^i = 1 - |h_t(\bar{x}_i) - y_i|$, такие потери равны нулю, если гипотеза h_t неправильно классифицирует объект x_i, и они максимальны (единица), если классификация – правильная. Соответственно, вес неправильной классификации 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 5.1. *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*

$$\epsilon = D\{h(\bar{x}_i) \neq y_i\} = \sum_{h(\bar{x}_i)\neq y_i} D(i)$$

результирующей гипотезы h, *выданной алгоритмом* AdaBoost *после* T *шагов работы, ограничена*

$$\epsilon \leqslant 2^T \prod_{t=1}^{T} \sqrt{\epsilon_t(1-\epsilon_t)}. \tag{5.1}$$

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:

$$\sum_{i=1}^{l} w_{t+1}^i = \sum_{i=1}^{l} w_t^i \beta_t^{1-|h_t(\bar{x}_i)-y_i|} \leqslant$$

$$\leqslant \sum_{i=1}^{l} w_t^i (1-(1-\beta_t)(1-|h_t(\bar{x}_i)-y_i|)) =$$

$$= \left(\sum_{i=1}^{l} w_t^i\right)(1-(1-\beta_t)(1-\epsilon_t)). \tag{5.2}$$

Используя (5.2) T раз, получим

$$\sum_{i=1}^{l} w_{T+1}^i \leqslant \prod_{t=1}^{T}(1-(1-\beta_t)(1-\epsilon_t)). \tag{5.3}$$

Здесь было использовано определение ошибки обучения ϵ_t алгоритма WeakLearn на шаге t :

$$\epsilon_t = \sum_{i=1}^{l} p_t^i |h_t(\bar{x}_i) - y_i| = \sum_{i=1}^{l} \left(\frac{w_t^i}{\sum_{j=1}^{l} w_t^j} \right) |h_t(\bar{x}_i) - y_i|.$$

Лемма 5.1. *Результирующий классификатор h делает ошибку на объекте $\bar{x}_i$ тогда и только тогда, когда*

$$\prod_{t=1}^{T} \beta_t^{-|h_t(\bar{x}_i)-y_i|} \geqslant \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{5.4}$$

Доказательство. Действительно, это утверждение прямо следует из определения классификатора h_f в случае, когда $y_i = 0$, так как в таком случае $\beta_t^{-|h_t(\bar{x}_i)-y_i|} = \beta_t^{-h_t(\bar{x}_i)}$ для всех t. По определению равенство $h(\bar{x}_i) = 1$ может быть тогда и только тогда, когда

$$\sum_{t=1}^{T} \ln(1/\beta_t) h_t(\bar{x}_i) \geqslant \frac{1}{2} \sum_{t=1}^{T} \ln(1/\beta_t). \tag{5.5}$$

Неравенство (5.5) эквивалентно неравенству (5.4).

Пусть теперь $y_i = 1$. Тогда $h_t(\bar{x}_i) \leqslant y_i$ для всех t. Поэтому $\beta_t^{-|h_t(\bar{x}_i)-y_i|} = \beta_t^{-(1-h_t(\bar{x}_i))}$ для всех t. В этом случае для всех $1 \leqslant t \leqslant T$

$$\beta_t^{-|h_t(\bar{x}_i)-y_i|} = \beta_t^{-1+h_t(\bar{x}_i)}. \tag{5.6}$$

Равенство $h(x_i) = 0$ по определению возможно только при

$$\prod_{t=1}^{T} \beta_t^{-h_t(\bar{x}_i)} < \left(\prod_{t=1}^{T} \beta_t \right)^{-1/2}. \tag{5.7}$$

Неравенство (5.7) эквивалентно неравенству

$$\prod_{t=1}^{T} \beta_t^{h_t(\bar{x}_i)} > \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}. \tag{5.8}$$

Неравенство (5.8) с учетом равенства (5.6) эквивалентно неравенству (5.4) леммы. Лемма доказана. $\triangle$

Возвращаясь к доказательству теоремы, заметим, что по определению

$$w_{T+1}^i = D(i) \prod_{t=1}^{T} \beta_t^{1-|h_t(\bar{x}_i)-y_i|}. \tag{5.9}$$

По лемме 5.1 из (5.4) и (5.9) получаем

$$\begin{aligned} \sum_{i=1}^{l} w_{T+1}^i \geqslant \sum_{i:h(\bar{x}_i)\neq y_i} w_{T+1}^i \geqslant \\ \geqslant \left(\sum_{i:h(\bar{x}_i)\neq y_i} D(i) \right) \left(\prod_{t=1}^{T} \beta_t \right)^{1/2} = \\ = \epsilon \left(\prod_{t=1}^{T} \beta_t \right)^{1/2}, \end{aligned} \tag{5.10}$$

где ϵ – ошибка обучения результирующего классификатора h относительно распределения D.

Комбинируя (5.3) и (5.10), получим

$$\epsilon \leqslant \prod_{t=1}^{T} \frac{1-(1-\beta_t)(1-\epsilon_t)}{\sqrt{\beta_t}}. \tag{5.11}$$

Так как элементы произведения (5.11) 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 (5.11) и получаем (5.1). Теорема доказана. $\triangle$

Следствие 5.1. *Ошибка обучения результирующего классификатора h удовлетворяет неравенству*

$$\epsilon \leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right), \tag{5.12}$$

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* (5.12) *упрощается до*

$$\epsilon \leqslant \exp(-2T\gamma^2). \tag{5.13}$$

Доказательство. Действительно, в оценке (5.1) теоремы 5.1 при $\epsilon_t = \frac{1}{2} - \gamma_t$ будет

$$2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1-4\gamma_t^2}.$$

Отсюда

$$\epsilon \leqslant \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2} =$$

$$= \exp\left(\sum_{t=1}^{T}\frac{1}{2}\ln(1-4\gamma_t^2)\right) \leqslant$$

$$\leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right). \tag{5.14}$$

Неравенство (5.12) доказано.

Для доказательства (5.13) заметим, что неравенство (5.12) при $\gamma_t = \gamma$ превращается в

$$\epsilon \leqslant (1-4\gamma^2)^{T/2} = \exp((T/2)\ln(1-4\gamma^2)) \leqslant \exp(-2T\gamma^2).$$

Оценка (5.13) представляет собой обычную экспоненциально убывающую оценку ошибки обучения типа неравенства Хефдинга.

Неравенство (5.13) позволяет оценить число итераций алгоритма AdaBoost, необходимых для достижения ошибки обучения $\leqslant \epsilon$ результирующего классификатора h :

$$T \geqslant \frac{1}{2\gamma^2}\ln\frac{1}{\epsilon}.$$

5.2. Лабораторные работы

Лабораторная работа 1

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 сайта

http : //www.cs.toronto.edu/roweis/data.html

По этому адресу имеются данные из базы USPS в формате MATLAB, содержащие цифровые образы различных написаний рукописных цифр.

5.3. Problems

1. Construct a variant of the weighted majority algorithm for the case when an expert exists in the pool which makes no more than k mistakes. Compute a performance bound for this algorithm.

2. Assume that one-step losses of all experts are 0 and 1. Let s be the best expert which has minimal cumulative loss L_T^s at the first T steps, and let L_T be a cumulative loss of the algorithm $Hedge$. Prove that $L_T \geqslant L_T^s$.

3. Consider the protocol of the game of prediction with expert advice, where *Nature* outputs a sequence $0^T(01)^T1^T$. There are three constant experts. *Expert 1* ouputs $\xi_t^1 = 0$ for all $t = 1, \ldots, 4T$, *Expert 2* ouputs $\xi_t^1 = 1$ for all $t = 1, \ldots, 4T$, *Expert 3* ouputs $\xi_t^1 = 1/2$ for all $t = 1, \ldots, 4T$. The loss function is $\lambda(\omega, \gamma) = |\omega - \gamma|$.

Compute at all time points $t = 1, \ldots, 4T$:

(i) the weights of experts;

(ii) the loss *Allocator* and prediction of the exponentially weighted forecaster.

4. Check the simplest properties of the exponential distribution with density $p(x) = e^{-x}$: $P\{\xi > a\} = e^{-a}$ and $P\{\xi > a + b\} = e^b P\{\xi > a\}$ for all nonnegative a and b.

5. Prove that for any non-negative random variable η with a density $p(t)$ the following equality is valid:

$$E(\eta) = \int_0^\infty P\{\eta \geqslant y\} dy.$$

Note. Use $p(y) = F'(y)$, where $F(y) = \int_0^y p(t)dt = 1 - P\{\eta \geqslant y\}$. After that, apply the integration by parts of $E(\eta) = \int_0^\infty tp(t)dt$.

Глава 6

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 $O(\ln N)$.

6.1. Экспоненциально вогнутые 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*

$$\Pi_\lambda = \{(x,y) : \exists\, p\ (\lambda(0,p) = x, \lambda(1,p) = y)\} \tag{6.1}$$

и *множество суперпредсказаний*

$$\Sigma_\lambda = \{(x,y) : \exists\, p\ (\lambda(0,p) \leqslant x, \lambda(1,p) \leqslant y)\}. \tag{6.2}$$

Из первого свойства функции потерь и компактности $[0,1]$ следует, что множество суперпредсказаний замкнуто. Для рассматриваемых ниже функций потерь множество предсказаний (6.1) является границей множества суперпредсказаний (6.2).

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*

$$E_\eta(x,y) = (e^{-\eta x}, e^{-\eta y}) \tag{6.3}$$

для всех $x, y \in [0,+\infty)$.

При этом гомоморфизме множество предсказаний (6.1) переходит в множество

$$E_\eta(\Pi_\lambda) = \{(e^{-\eta\lambda(0,p)}, e^{-\eta\lambda(1,p)}) : p \in \Gamma\},$$

а множество суперпредсказаний (6.2) переходит в множество

$$E_\eta(\Sigma_\lambda) = \{(x,y) : \exists\, p\ (0 \leqslant x \leqslant e^{-\eta\lambda(0,p)}, 0 \leqslant y \leqslant e^{-\eta\lambda(1,p)})\}. \tag{6.4}$$

Функция потерь $\lambda(\omega,\gamma)$ называется *η-смешиваемой*, если множество $E_\eta(\Sigma_\lambda)$ является выпуклым. Функция потерь называется

смешиваемой (или экспоненциально вогнутой), если она является η-смешиваемой для некоторого $\eta > 0$.

Ясно, что для всякой смешиваемой функции потерь множество суперпредсказаний является выпуклым. Далее мы увидим, что не всякая функция потерь с выпуклым множеством суперпредсказаний является смешиваемой. Таким образом, смешиваемость – более сильное требование, чем просто выпуклость множества суперпредсказаний.

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

$$\lambda(\omega, \gamma) = -\frac{1}{\eta} \ln(\omega\gamma + (1 - \omega)(1 - \gamma)), \tag{6.5}$$

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 $\Gamma = [0, 1]$.

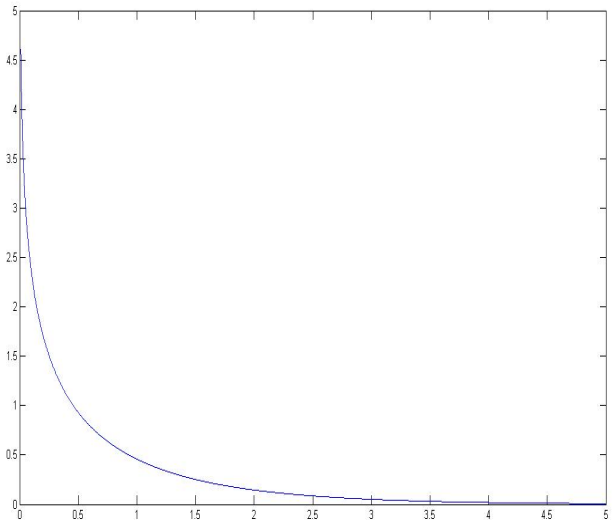

Рис. 6.1. Множество предсказаний и суперпредсказаний логарифмической функции потерь

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.

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 свой-

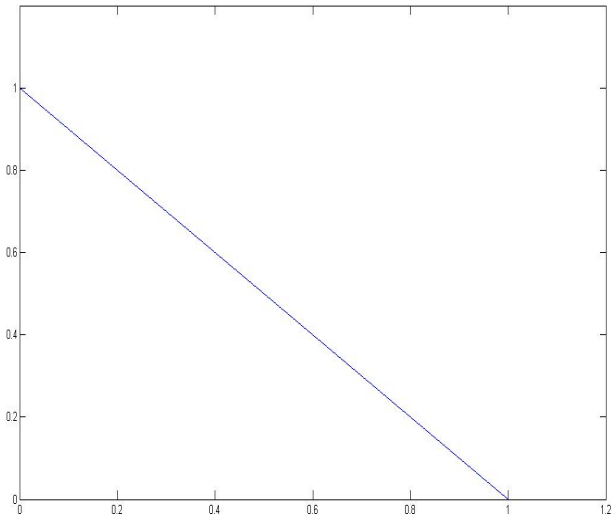

Рис. 6.2. Образы множества предсказаний и суперпредсказаний логарифмической функции потерь в экспоненциальном пространстве

ству

$$\lambda(\omega,\gamma) = \begin{cases} 0, \text{ если } \omega = \gamma, \\ 1 \text{ в противном случае.} \end{cases}$$

Обсудим геометрические свойства смешиваемых функций потерь. Здесь обобщенная логарифмическая функция потерь играет особую роль.

Легко видеть, что множество предсказаний (6.1) обобщенной логарифмической функции потерь (6.5) есть кривая:

$$\{(x,y) : e^{-\eta x} + e^{-\eta y} = 1\}. \tag{6.6}$$

Мы будем рассматривать параллельные сдвиги кривой (6.6) в плоскости суперпредсказаний, т.е. кривые вида

$$\{(x,y) : e^{-\eta(x-\alpha)} + e^{-\eta(y-\beta)} = 1\}, \tag{6.7}$$

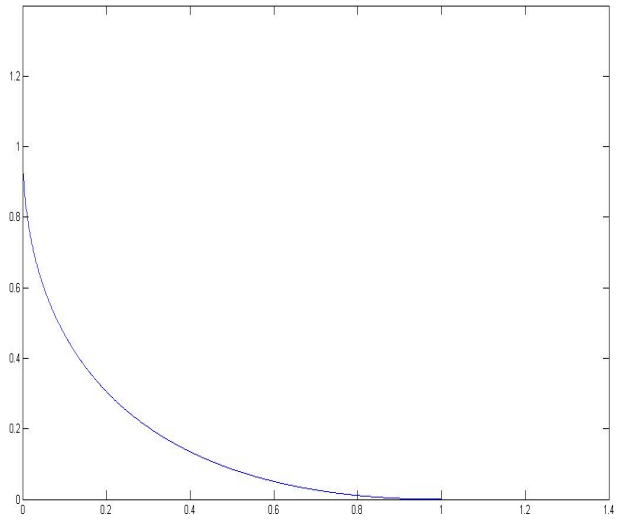

Рис. 6.3. Множество предсказаний и суперпредсказаний квадратичной функции потерь

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 $y_1 \geqslant y_2$.

Множество $A \subseteq \mathcal{R}^2$ находится северо-восточнее некоторого параллельного сдвига кривой (6.6), если каждая его точка находится северо-восточнее некоторой точки, лежащей на этом сдвиге (6.7).

Заметим, что прообразами всех прямых вида $ax + by = c$, где $a > 0$ и $b > 0$, рассматриваемых в экспоненциальном пространстве, при гомоморфизме (6.3) 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

$$ae^{-\eta x} + be^{-\eta y} = c,$$

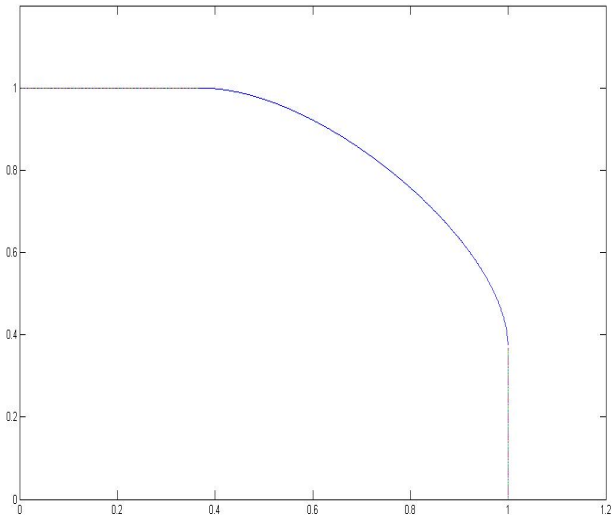

Рис. 6.4. Образы множества предсказаний и суперпредсказаний квадратичной функции потерь в экспоненциальном пространстве

т.е. параллельный сдвиг кривой $e^{-\eta x} + e^{-\eta y} = 1$ на вектор

$$\left(-\frac{1}{\eta} \ln \frac{a}{c}, -\frac{1}{\eta} \ln \frac{b}{c} \right).$$

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 6.1. *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 [29]. 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 [20]. Агрегирующий алгоритм Вовка имеет ошибку предсказания, которая зависит только от числа экспертов и не зависит от длины последовательности.

6.2. Конечное множество экспертов

В разделах 4.4 и 4.6 алгоритмы предсказания имели ошибку порядка $O(\sqrt{T \ln N})$, где T – длина периода, а N – число экспертов. Алгоритмы и результаты этих разделов относились к функциям потерь произвольного вида (в разделе 4.4 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 ошибку предсказания, не зависящую от длины периода T, эта ошибка имеет вид $O(\ln N)$, где N – число экспертов.

В дальнейшем мы построим стратегию предсказателя, для которой

$$L_T \leqslant c(\eta) \inf_\theta L_T(\theta) + a(\eta) \ln N$$

для всех T, где в общем случае может быть $c(\eta) > 1$ для любого значения параметра обучения алгоритма $\eta \in (0, \infty)$.

В случае смешиваемых функций потерь будет $c(\eta) = 1$ для некоторых значений параметра обучения η.

Предварительно рассмотрим схему алгоритма в случае множества исходов $\Omega = \{0, 1\}$ и конечного множества экспертов $\Theta = \{1, 2, \ldots, N\}$. Прогнозы могут принимать любые действительные значения $\Gamma = \mathcal{R}$. 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 на предсказания с использованием экспертных прогнозов.
Пусть $L_0 = 0$, $L_0(i) = 0$, $i = 1, \ldots, N$.
FOR $t = 1, 2, \ldots$
Эксперт i выбирает прогноз $\xi_t^i \in \Gamma$, $i = 1, \ldots, N$.
Статистик выбирает прогноз $\gamma_t \in \Gamma$.
Природа выбирает исход $\omega_t \in \Omega$.
Эксперт i вычисляет свои суммарные потери на шаге t игры:

$$L_t(i) = L_{t-1}(i) + \lambda(\omega_t, \xi_t^i).$$

Статистик вычисляет свои суммарные потери на шаге t игры:

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t).$$

ENDFOR

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 $t = 1, 2, \ldots$ *Статистик* перестраивает веса экспертов $i = 1, \ldots, N$ согласно формуле

$$P_t(i) = \beta^{\lambda(\omega_t, \xi_t^i)} P_{t-1}(i). \tag{6.8}$$

Таким образом, вес эксперта, имеющего большие потери, уменьшается.

Веса экспертов (6.8) нормируем:

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)}, \tag{6.9}$$

чтобы сумма нормированных весов стала равной 1.

Введем вспомогательную функцию, которая называется «псевдопредсказанием» :

$$g_t(\omega) = \log_\beta \sum_{j=1}^{N} \beta^{\lambda(\omega, \xi_t^i)} P_{t-1}^*(i). \tag{6.10}$$

Алгоритм, выдающий псевдопредсказания, вычисленные по формуле (6.10), обозначаем APA (Aggregating Pseudo Algorithm). Обозначим суммарные потери алгоритма APA за T шагов на последовательности исходов $\omega_1, \ldots, \omega_T$:

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{6.11}$$

Следующая лемма представляет суммарные потери алгоритма APA в более простом и ясном виде.

Лемма 6.1. *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 (6.8) следует, что

$$P_T(i) = \beta^{\sum_{t=1}^{T} \lambda(\omega_t, \xi_t^i)} P_0(i) = \beta^{L_T(i)} P_0(i).$$

Из определения имеют место следующие равенства

$$\begin{aligned}
\log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) - \log_\beta \sum_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i) = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_T(i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i) + \lambda(\omega_T, \xi_T^i)} P_0(i)}{\sum\limits_{i=1}^{N} \beta^{L_{T-1}(i)} P_0(i)} = \\
= \log_\beta \frac{\sum\limits_{i=1}^{N} \beta^{\lambda(\omega_T, \xi_T^i)} P_{T-1}(i)}{\sum\limits_{i=1}^{N} P_{T-1}(i)} = \\
= \log_\beta \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P^*_{T-1}(j) = g_T(\omega_T). \qquad (6.12)
\end{aligned}$$

Последнее равенство следует из определения (6.10). Поскольку (6.12) 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.

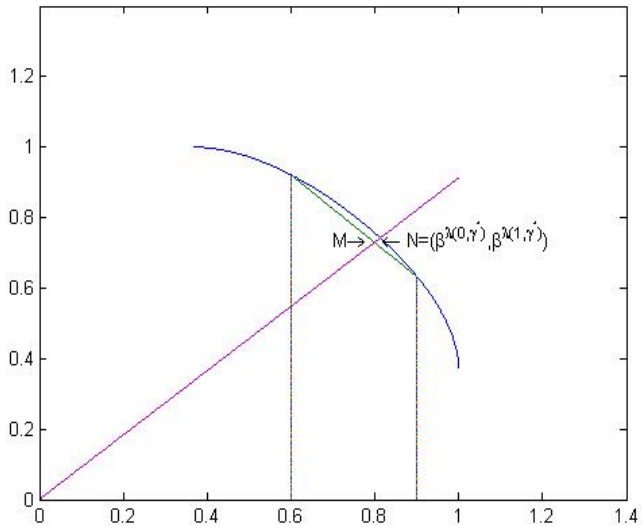

Рис. 6.5. Пример определения предсказания γ^*. Прямая, проходящая через точку M, отмечает точку $N = (\beta^{\lambda(0,\gamma^*)}, \beta^{\lambda(1,\gamma^*)})$ на кривой, по которой вычисляется предсказание γ^*

Мы покажем, что функция подстановки существует, если функция потерь $\lambda(\omega, \gamma_t)$ является смешиваемой.

Предложение 6.2. *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

$$\beta^{\lambda(\omega_T, \gamma^*)} \geqslant \sum_{j=1}^{N} \beta^{\lambda(\omega_T, \xi_T^j)} P_{T-1}^*(j) \tag{6.13}$$

для всех $\omega_T \in \{0, 1\}$. Неравенство (6.13) 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.

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 (6.13). 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 6.1 следует, что будет иметь место неравенство

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant$$

$$\leqslant L_T(APA) = \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i). \tag{6.14}$$

Припишем каждому эксперту одинаковый вес $P_0(i) = 1/N$. Тогда

из (6.14) следует, что для произвольного $i \in \Theta$, для всех T

$$L_T(AA) \leqslant \log_\beta \left(\frac{1}{N} \sum_{i=1}^{N} \beta^{L_T(i)} \right) \leqslant$$
$$\leqslant \log_\beta \left(\frac{1}{N} \beta^{L_T(i)} \right) = L_T(i) + \frac{\ln N}{\eta}. \qquad (6.15)$$

Оценка (6.15) 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 (ошибку предсказания), 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.

6.3. 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.

На шаге $t = 1, 2, \ldots$ *Статистик* перестраивает веса экспертов в соответствии с формулой

$$P_t(d\theta) = \beta^{\lambda(\omega_t, \xi_t^\theta)} P_{t-1}(d\theta). \qquad (6.16)$$

Таким образом, вес эксперта, имеющего большие потери, уменьшается. По определению задание весов (6.16) 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 (6.16) нормируем:

$$P_t^*(d\theta) = \frac{P_t(d\theta)}{P_t(\Theta)}. \tag{6.17}$$

Нормированные веса представляют собой вероятностную меру; для нее $P_t^*(\Theta) = 1$.

Аналогичным образом введем «псевдопредсказание»

$$g_t(\omega) = \log_\beta \int_\Theta \beta^{\lambda(\omega,\xi_t^\theta)} P_{t-1}^*(d\theta). \tag{6.18}$$

Алгоритм, выдающий псевдопредсказания, также обозначается APA, а суммарные потери алгоритма APA за T шагов равны

$$L_T(APA) = \sum_{t=1}^{T} g_t(\omega_t). \tag{6.19}$$

Из (6.16) 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 (6.18) переписывается в виде

$$g_T(\omega) = \log_\beta \int_\Theta \frac{\beta^{\lambda(\omega,\xi_T^\theta)+L_{T-1}(\theta)}}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} P_0(d\theta). \tag{6.20}$$

Имеет место аналог леммы 6.1.

Лемма 6.2. *Суммарные потери обобщенного алгоритма за T шагов могут быть представлены в виде*

$$L_T(APA) = \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \tag{6.21}$$

Доказательство. Доказательство леммы аналогично доказательству леммы 6.1. Из (6.8) следует, что

$$P_t(d\theta) = \beta^{\sum_{t=1}^{T} \lambda(\omega_t, \xi_t^\theta)} P_0(d\theta) = \beta^{L_T(\theta)} P_0(d\theta). \tag{6.22}$$

Из определения имеют место следующие равенства:

$$\begin{aligned}
\log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta) - \log_\beta \int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta) = \\
= \log_\beta \frac{\int_\Theta \beta^{L_T(\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} = \\
= \log_\beta \frac{\int_\Theta \beta^{L_{T-1}(\theta) + \lambda(\omega_T, \xi_T^\theta)} P_0(d\theta)}{\int_\Theta \beta^{L_{T-1}(\theta)} P_0(d\theta)} = \\
= \log_\beta \frac{\int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}(d\theta)}{\int_\Theta P_{T-1}(d\theta)} = \\
= \log_\beta \int_\Theta \beta^{\lambda(\omega_T, \xi_T^\theta)} P_{T-1}^*(d\theta) = g_T(\omega_T).
\end{aligned} \tag{6.23}$$

Последнее равенство следует из определения (6.18).

Поскольку (6.23) 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 (6.18). Полагаем $\Sigma(g_t) = \gamma^*$.

В этом случае по лемме 6.2 будет иметь место неравенство

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) \leqslant \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta). \qquad (6.24)$$

6.4. 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) $c(\eta)$:

$$c(\eta) = \inf \left\{ c : \forall\ P\ \exists\ \delta \in \Gamma\ \forall\ \omega \left(\lambda(\omega, \delta) \leqslant c \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma) \right) \right\}.$$

При некоторых естественных предположениях на исходные множества функция $c(\eta)$ является непрерывной и невозрастающей.

В этом случае функция подстановки определяется как функция, удовлетворяющая

$$\forall\ \omega : \lambda(\omega, \Sigma_\eta(g)) \leqslant c(\eta) g(\omega) \qquad (6.25)$$

для любого псевдопредсказания

$$g(\omega) = \log_\beta \int_\Gamma \beta^{\lambda(\omega,\gamma)} P(d\gamma)$$

и вероятностного распределения P на Γ.

Можно определить *минимаксную* функцию подстановки.

$$\Sigma_\eta(g) \in \arg \min_{\gamma \in \Gamma} \sup_{\omega \in \Omega} \frac{\lambda(\omega, \gamma)}{g(\omega)}. \qquad (6.26)$$

По определению любая минимаксная функция подстановки $\Sigma_\eta(g)$, удовлетворяющая (6.26), удовлетворяет и неравенству (6.25).

Заметим, что могут существовать другие – не минимаксные, функции подстановки такие, что выполнено условие (6.25). Часто их проще вычислить.

В общем случае, для произвольной функции потерь, не обязательно смешиваемой, вместо (6.24) имеем

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma_\eta(g_t)) \leqslant c(\eta) \log_\beta \int_\Theta \beta^{L_T(\theta)} P_0(d\theta).$$

Все аналогичные неравенства будут верными, если ввести в них множитель $c(\eta)$.

В случае конечного числа экспертов неравенство (6.15) переходит в неравенство

$$L_T(AA) \leqslant c(\eta) \log_\beta \left(\frac{1}{N} \sum_{i=1}^{N} \beta^{L_T(i)} \right) \leqslant$$

$$\leqslant c(\eta) \log_\beta \left(\frac{1}{N} \beta^{L_T(i)} \right) = c(\eta) L_T(k) + c(\eta) \frac{\ln N}{\eta}$$

для всех T и всех $k = 1, \ \ldots, \ N$.

6.5. 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 *Статистик* 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:

$$Q_i(\omega | \omega_1, \ \ldots, \ \omega_{t-1}) = \xi_t^i(\omega) \in \Gamma. \tag{6.27}$$

Такое распределение можно интерпретировать как субъективное условное распределение эксперта i на t-м шаге. Величина (6.27) 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

$$Q_i(\omega_1, \ldots, \omega_t) = \xi_1^i(\omega_1)\xi_2^i(\omega_2) \cdot \ \ldots \ \cdot \xi_t^i(\omega_t). \tag{6.28}$$

Веса экспертов перестраиваются согласно (6.8). В данном случае вес i-го эксперта переопределяется на шаге t :

$$\begin{aligned} P_t(i) &= \beta^{\lambda(\omega_t, \xi_t^i)} P_{t-1}(i) = \\ &= \xi_1^i(\omega_1)\xi_2^i(\omega_2) \cdot \ \ldots \ \cdot \xi_t^i(\omega_t) P_0(i) = \\ &= Q_i(\omega_1, \ \ldots, \ \omega_t) P_0(i). \end{aligned} \tag{6.29}$$

Веса (6.29) экспертов нормируются как

$$P_t^*(i) = \frac{P_t(i)}{\sum\limits_{j=1}^{N} P_t(j)} = \frac{Q_i(\omega_1, \ \ldots, \ \omega_t) P_0(i)}{\sum\limits_{j=1}^{N} Q_j(\omega_1, \ \ldots, \ \omega_t) P_0(j)}. \tag{6.30}$$

Вероятность $P_t^*(i)$ представляет собой апостериорную вероятность эксперта i после наблюдения исходов $\omega_1, \ \ldots, \ \omega_t$.

Так как $\beta^{\lambda(\omega_t, \xi_t^i)} = \xi_t^i(\omega_t)$, псевдопредсказание (6.10) превращается в логарифм байесовской смеси распределений, предлагаемых на шаге t экспертами,

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \xi_t^i(\omega) P_{t-1}^*(i). \tag{6.31}$$

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 *Статистика*, 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 АА –

$$\gamma_1(\omega) = \sum_{i=1}^{N} \xi_1^i(\omega) P_0(i),$$

представляет собой байесовскую смесь вероятностных распределений экспертов относительно априорного распределения P_0 на множестве всех экспертов.

После того как появился первый исход ω_1, *Статистик* 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 AA

$$\gamma_2(\omega) = \sum_{i=1}^{N} \xi_2^i(\omega) P_1^*(i)$$

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, *Статистик* 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 AA представляет собой последовательное применение байесовского правила в режиме онлайн.

Потери i-го эксперта за T шагов равны

$$\begin{aligned} L_T(i) &= \sum_{i=1}^{T} \lambda(\omega_t, \xi_t^i) = \\ &= -\ln(\xi_1^i(\omega_1) \cdot \ \ldots \ \cdot \xi_T^i(\omega_T)) = \\ &= -\ln Q_i(\omega_1, \ldots, \ \omega_T). \end{aligned} \tag{6.32}$$

Это равенство использует определение субъективной вероятности (6.27), которую *Статистик* приписывает экспертам на шаге t.

Потери *Статистика*, использующего алгоритм AA, за T шагов равны

$$L_T(AA) = \sum_{t=1}^{T} \lambda(\omega_t, \Sigma(g_t)) =$$
$$= \log_\beta \sum_{i=1}^{N} \beta^{L_T(i)} P_0(i) =$$
$$= \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \ldots, \omega_T) P_0(i). \tag{6.33}$$

Таким образом, потери *Статистика* 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 (6.15) превращается в неравенство

$$L_T(AA) = \log_\beta \sum_{i=1}^{N} Q_i(\omega_1, \ldots, \omega_T) P_0(i) \leqslant$$
$$\leqslant -\ln Q_k(\omega_1, \ldots, \omega_T) - \ln P_0(k) \tag{6.34}$$

для всех T и $k = 1, \ldots, N$.

6.6. Простая игра на предсказания

Напомним, что простая игра на предсказание 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

$$g(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega,\xi_t^i)} P_{t-1}^*(i) \tag{6.35}$$

представляется в виде точки $(g(0), g(1))$ на координатной плоскости $\mathcal{R}^2$. Эта точка имеет вид

$$(\log_\beta(\beta p + (1-p)),\ \log_\beta(p + \beta(1-p))), \tag{6.36}$$

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 (6.36) 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 (6.36) получаем

$$\frac{1}{c(\beta)} = \log_\beta \left(\frac{1+\beta}{2}\right),$$

или

$$c(\beta) = \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}}. \tag{6.37}$$

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 (6.36), 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 (6.25), 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 (6.36). 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

$$L_T(AA) \leqslant \left(\frac{\ln\frac{1}{\beta}}{\ln\frac{2}{1+\beta}}\right) L_T(\theta) - \ln\left(\frac{1+\beta}{2}\right) \ln P_0(\theta). \qquad (6.38)$$

6.7. Игра с квадратичной функцией потерь

Изучим игру с квадратичной функцией потерь в простейшем случае, когда пространство исходов двухэлементное: $\Omega = \{-1, 1\}$,

пространство прогнозов это по-прежнему все действительные числа $\Gamma = \mathcal{R}$. Функция потерь – квадрат разности между исходом и прогнозом $\lambda(\omega, \gamma) = (\omega - \gamma)^2$.

Мы расматриваем случай $\Omega = \{-1, 1\}$, поскольку доказательства в этом случае проще. Все приведенные ниже утверждения также верны и для случая $\Omega = [-1, 1]$.

Лемма 6.3. *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

$$\frac{d^2y}{d^2x} = \frac{d\gamma}{dx}\frac{x'(\gamma)y''(\gamma) - x''(\gamma)y'(\gamma)}{(x'(\gamma))^2}. \tag{6.39}$$

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:

$$\begin{aligned} x'(\gamma) &= -2\eta(1+\gamma)e^{-\eta(1+\gamma)^2}, \\ x''(\gamma) &= 2\eta(-1+2\eta(1+\gamma)^2)e^{-\eta(1+\gamma)^2}, \\ y'(\gamma) &= 2\eta(1-\gamma)e^{-\eta(1-\gamma)^2}, \\ y''(\gamma) &= 2\eta(-1+2\eta(1-\gamma)^2)e^{-\eta(1-\gamma)^2}. \end{aligned} \tag{6.40}$$

Тогда условие η-смешиваемости требует, чтобы для всех значений $\gamma \in [-1,1]$

$$\begin{aligned} -(1+\gamma)(-1+2\eta(1-\gamma)^2) - \\ -(1-\gamma)(-1+2\eta(1+\gamma)^2) \geqslant 0, \\ \eta(1-\gamma^2) \leqslant \frac{1}{2}, \\ \eta \leqslant \frac{1}{2}. \end{aligned} \tag{6.41}$$

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

$$g_t(\omega) = \log_\beta \sum_{i=1}^{N} \beta^{\lambda(\omega,\xi_t^i)} P_{t-1}^*(i) \tag{6.42}$$

задается точкой

$$\begin{aligned} (e^{-\frac{1}{2}g(-1)}, e^{-\frac{1}{2}g(1)}) = \\ = \left(\sum_{i=1}^{N} \beta^{\lambda(-1,\xi_t^i)} P_{t-1}^*(i), \sum_{i=1}^{N} \beta^{\lambda(1,\xi_t^i)} P_{t-1}^*(i) \right), \end{aligned} \tag{6.43}$$

которая расположена под вогнутой кривой

$$\left(\beta^{\lambda(-1,\gamma)}, \beta^{\lambda(1,\gamma)} \right), \tag{6.44}$$

при $\gamma \in [-1.1]$.

Проведем прямую, проходящую через начало координат и точку (6.43). Коэффициент наклона этой прямой равен

$$k = \frac{\beta^{g(1)}}{\beta^{g_t(-1)}} = e^{\frac{1}{2}g_t(-1) - \frac{1}{2}g_t(1)}. \tag{6.45}$$

Точка пересечения $\left(\beta^{\lambda(-1,\gamma^*)}, \beta^{\lambda(1,\gamma^*)}\right)$ этой прямой и кривой (6.44) имеет абсциссу и ординату по величине не меньше, чем абсцисса и ордината точки (6.43) :

$$\begin{aligned} \beta^{\lambda(-1,\gamma^*)} &\geqslant \beta^{g_t(-1)}, \\ \beta^{\lambda(1,\gamma^*)} &\geqslant \beta^{g_t(1)}. \end{aligned} \tag{6.46}$$

Эквивалентная запись (6.46) имеет вид

$$\begin{aligned} \lambda(-1,\gamma^*) &\leqslant g_t(-1), \\ \lambda(1,\gamma^*) &\leqslant g_t(1). \end{aligned} \tag{6.47}$$

Вычислим предсказание γ^*. Значение γ^* находим из уравнения

$$\frac{\beta^{g_t(1)}}{\beta^{g_t(-1)}} = \beta^{g_t(1)-g_t(-1)} = \frac{\beta^{\lambda(1,\gamma^*)}}{\beta^{\lambda(-1,\gamma^*)}} = \beta^{\lambda(1,\gamma^*)-\lambda(-1,\gamma^*)}. \tag{6.48}$$

Остается найти корень уравнения

$$\lambda(1,\gamma^*) - \lambda(-1,\gamma^*) = (1-\gamma^*)^2 - (-1-\gamma^*)^2 = g_t(1) - g_t(-1),$$

который равен

$$\gamma^* = \frac{1}{4}(g_t(-1) - g_t(1)). \tag{6.49}$$

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. [32]).

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

$$\lambda(\omega, \Sigma(g_t)) \leqslant g_y(\omega)$$

для всех $\omega \in \Omega$, где g_t определена по (6.18).

6.8. Универсальный портфель

Рассмотрим следующую игру Ковера [9]. Имеется N финансовых инструментов, например акций. Время разделено на моменты $t = 1, 2, \ldots$ Поведение рынка в момент t характеризуется вектором изменений цен акций от момента $t-1$ к моменту t :

$$\bar{\omega}_t = (\omega_{1,t}, \ \ldots, \ \omega_{N,t}),$$

где $\omega_{i,t} = c_{i,t}/c_{i,t-1}$, $c_{i,t}$ – цена акции i при закрытии рынка 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. Тогда инвестиции, вложенные согласно портфелю $\bar{\gamma}$, увеличиваются в

$$(\bar{\gamma} \cdot \bar{\omega}) = \sum_{i=1}^{N} \gamma_i \omega_i$$

раз.

Определим функцию потерь в виде

$$\lambda(\bar{\omega}, \bar{\gamma}) = -\ln(\bar{\gamma} \cdot \bar{\omega}). \tag{6.50}$$

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 [31].

Допустим, что задано априорное распределение $P_0(d\bar{\gamma})$ на симплексе Γ всех портфелей.

Согласно (6.20) потери обобщенного алгоритма APA равны

$$g_T(\omega) = \log_\beta \int_\Gamma \frac{\beta^{\lambda(\bar{\omega},\bar{\gamma})+L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\bar{\gamma})} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}), \tag{6.51}$$

где $\beta = e^{-\eta}$, $0 < \eta \leqslant 1$.

Теорема 6.1. *Игра* (*функция потерь* (6.50)) *Ковера является смешиваемой при* $0 < \eta \leqslant 1$. *Функция подстановки задается выражением*

$$\Sigma(g_T) = \bar{\gamma}^* = \int_\Gamma \bar{\gamma} P_{T-1}(d\bar{\gamma}) =$$
$$= \int_\Theta \bar{\gamma} \frac{\beta^{L_{T-1}(\bar{\gamma})}}{\int_\Gamma \beta^{L_{T-1}(\gamma)} P_0(d\bar{\gamma})} P_0(d\bar{\gamma}). \tag{6.52}$$

Доказательство. Нам надо доказать, что для всех $\bar{\omega}$

$$\lambda\left(\bar{\omega}, \int_\Gamma \bar{\gamma} P(d\bar{\gamma})\right) \leqslant \log_\beta \int_\Gamma \beta^{\lambda(\bar{\omega},\bar{\gamma})} P(d\bar{\gamma}).$$

Это неравенство эквивалентно неравенству

$$f\left(\int_\Gamma \bar{\gamma} P(d\bar{\gamma})\right) \geqslant \int_\Gamma f(\bar{\gamma}) P(d\bar{\gamma}), \tag{6.53}$$

где $f(\bar{\gamma}) = \beta^{\lambda(\bar{\omega},\bar{\gamma})} = (\bar{\gamma} \cdot \bar{\omega})^\eta$. Неравенство (6.53) следует из вогнутости функции $f(\bar{\gamma})$ при $0 < \eta \leqslant 1$. $\triangle$

Прогноз портфеля (6.52) можно записать полностью, используя представление (6.21) (из леммы 6.2) 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:

$$\bar\gamma_T = \frac{\int_\Gamma \bar\gamma \prod_{t=1}^{T-1} (\bar\gamma \cdot \bar\omega_t) P_0(d\bar\gamma)}{\int_\Gamma \prod_{t=1}^{T-1} (\bar\gamma \cdot \bar\omega_t) P_0(d\bar\gamma)}.$$

В случае произвольного пространства экспертов Γ удобно рассмотреть в качестве априорного распределения распределение Дирихле с параметрами $(1/2,\ \ldots,\ 1/2)$ на симплексе Γ :

$$P_0(d\bar\gamma) = \frac{\Gamma(N/2)}{[\Gamma(1/2)]^N} \prod_{j=1}^{N} \gamma_j^{-1/2} d\bar\gamma,$$

где

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Заметим, что $\Gamma(N+1) = N!$.

Мы приведем без доказательства основной результат статьи [30], который дает оценку оптимальности алгоритма AA.

Теорема 6.2. *Пусть множество экспертов* Γ *– произвольное,* $\eta = 1$*. Тогда суммарные потери алгоритма* AA *удовлетворяют неравенству*

$$L_T(AA) \leqslant \inf_{\bar\gamma} L_T(\bar\gamma) + \frac{N-1}{2} \ln T + c \tag{6.54}$$

для всех T*, где* c *– положительная константа.*

Так как доход при следовании стратегии инвестирования, предлагаемой алгоритмом AA, равен

$$K_T(AA) = e^{-L_T(AA)},$$

неравенство (6.54) можно переписать в виде

$$K_T(AA) \geqslant T^{-\frac{N-1}{2}} K_T(\bar{\gamma}),$$

где $K_T(\bar{\gamma})$ – доход, полученный при использовании произвольного постоянного портфеля $\bar{\gamma}$.

6.9. Многомерная онлайн регрессия

6.9.1. 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.

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. Имеются эксперты – линейные функции $f(x) = (\theta \cdot x)$, где $\theta, x \in \mathcal{R}^n$. Значения этих функций при $x = x_t$ интерпретируются как прогнозы экспертов θ на шаге t. В этом разделе мы не подчеркиваем векторы чертой сверху.

Задачей регрессии в режиме онлайн является построение на каждом шаге t прогноза величины y_t, используя прогнозы таких линейных экспертов.

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: $\xi_t^\theta = (\theta \cdot x_t)$.

Общая схема регресии регулируется следующем протоколом. Рассматривается игра с полной информацией между игроками: *Эксперт* θ, *Статистик* и *Природа*.

FOR $t = 1, 2, \ldots$
Природа анонсирует $x_t \in \mathcal{R}^n$.
Эксперты анонсируют прогнозы $\xi_t^\theta = (\theta \cdot x_t)$, $\theta \in \mathcal{R}^n$.
Статистик представляет предсказание $\gamma_t \in \mathcal{R}$.
Природа анонсирует $y_t \in [-Y, Y]$.
ENDFOR

Для исчисления потерь используется квадратичная функция потерь. На шаге t *Эксперт* θ вычисляет свои потери $(y_t - (\theta \cdot x_t))^2$. *Статистик* вычисляет свои потери $(y_t - \gamma_t)^2$.

Применим к этой игре алгоритм AA с параметром обучения $\eta = 1/2Y^2$. [1]

Введем априорное распределение

$$P_0(d\theta) = (a\eta/\pi)^{n/2} e^{-a\eta\|\theta\|^2} d\theta, \tag{6.55}$$

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:

$$\lambda(y_t, x_t'\theta) = (y_t - x_t'\theta)^2 = \theta'(x_t x_t')\theta - 2(y_t x_t')\theta + y_t^2. \tag{6.56}$$

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

[1] Можно доказать, что при таком значении параметра функция подстановки существует и имеет вид (6.49).

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

$$L_T(\theta) = \sum_{t=1}^T (y_t - x_t'\theta)^2 =$$
$$= \theta'\left(\sum_{t=1}^T x_t x_t'\right)\theta - 2\left(\sum_{t=1}^T y_t x_t'\right)\theta + \sum_{t=1}^T y_t^2. \tag{6.57}$$

Согласно (6.8) и (6.22) имеем

$$P_{t-1}(d\theta) = \beta^{L_{t-1}(\theta)} P_0(d\theta).$$

Поэтому произвольное псевдопредсказание на шаге t имеет вид

$$g_t(y) = \log_\beta \int \beta^{\lambda(y, x_t'\theta)} P_{t-1}^*(d\theta) =$$
$$= \log_\beta \int \beta^{\lambda(y, x_t'\theta)} \frac{P_{t-1}(d\theta)}{P_{t-1}(\Theta)} =$$
$$= \log_\beta \int \beta^{\lambda(y, x_t'\theta) + L_{t-1}(\theta)} \frac{1}{P_{t-1}(\Theta)} P_0(d\theta). \tag{6.58}$$

Отсюда, учитывая представления (6.55) для априорного распределения и (6.56) для функции потерь, получим

$$g_T(-Y) = \log_\beta \int \beta^{\lambda(-Y, x_T'\theta) + L_{T-1}(\theta)} \frac{1}{P_{T-1}(\Theta)} P_0(d\theta) = \tag{6.59}$$

$$= \int_{\mathcal{R}^n} e^{-\eta\theta'(aI + \sum_{t=1}^T x_t x_t')\theta + 2\eta(\sum_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum_{t=1}^{T-1} y_t^2 + Y^2)} \frac{d\theta}{P_{T-1}(\Theta)}.$$

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 (6.49) (см. [32]).

Тогда, используя формулу (6.59) для $g_T(-Y)$ и аналогичную формулу для $g_T(Y)$, получаем

$$\gamma_T = \frac{1}{4Y}(g_T(-Y) - g_T(Y)) = \frac{1}{4Y} \times$$

$$\times \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_t')\theta - \eta(\sum\limits_{t=1}^{T-1} y_t^2 + Y^2)} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta \frac{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' - Y x_T')\theta} d\theta}{\int_{\mathcal{R}^n} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T-1} y_t x_t' + Y x_T')\theta} d\theta} =$$

$$= \frac{1}{4Y} \log_\beta e^{-\eta F\left(aI+\sum\limits_{t=1}^{T} x_t x_t', -2\sum\limits_{t=1}^{T-1} y_t x_t', 2Y x_T'\right)} =$$

$$= \frac{1}{4Y} F\left(aI + \sum_{t=1}^{T} x_t x_t', -2\sum_{t=1}^{T-1} y_t x_t', 2Y x_T'\right) =$$

$$= \left(\sum_{t=1}^{T-1} y_t x_t'\right)\left(aI + \sum_{t=1}^{T} x_t x_t'\right)^{-1} \cdot x_T. \tag{6.60}$$

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 6.4, 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

$$\begin{aligned} A &= aI + \sum_{t=1}^{T} x_t x_t', \\ c &= -2\sum_{t=1}^{T-1} y_t x_t', \\ x &= 2Y x_T'. \end{aligned}$$

В 4-й строке мы использовали обозначение

$$\begin{aligned} F(A,c,x) = &\inf_{\theta\in\mathcal{R}^n}(\theta' A\theta + c'\theta + x'\theta) - \\ &- \inf_{\theta\in\mathcal{R}^n}(\theta' A\theta + c'\theta - x'\theta), \end{aligned} \tag{6.61}$$

а при переходе от 5-й строки к 6-й – следующую ниже лемму 6.5, согласно которой $F(A,c,x) = -c'A^{-1}x$.

Приведем формулировки и доказательства лемм 6.4 и 6.5.

Лемма 6.4. *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*

$$\int\limits_{\mathcal{R}^n} e^{-Q(\theta)} d\theta = e^{-Q_0} \frac{\pi^{n/2}}{\sqrt{\det A}}, \tag{6.62}$$

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 6.5 показывает, что $F(A, c, x) = -c' A^{-1} x$.

Лемма 6.5. *Пусть A – симметричная положительно определенная матрица типа $(n \times n)$, $b, x \in \mathcal{R}^n$. Тогда*

$$F(A, c, x) = \min_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta + x'\theta) -$$
$$- \min_{\theta \in \mathcal{R}^n} (\theta' A\theta + c'\theta - x'\theta) = -c' A^{-1} x. \qquad (6.63)$$

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. $\triangle$

Таким образом, согласно выражению (6.60) 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:

$A = aI$; $b' = \bar{0}$.
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

Сравнение потерь алгоритма AAR с потерями наилучшего эксперта дает следующая теорема.

Теорема 6.3. *Для произвольного T*

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + Y^2 \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) \leqslant$$

$$\leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + Y^2 \sum_{t=1}^{n} \ln \left(1 + \frac{1}{a} \sum_{t=1}^{T} x_{t,i}^2 \right).$$

Если к тому же $\|x_t\| \leqslant X$ для всех t, то

$$L_T(AAR) \leqslant \inf_\theta (L_T(\theta) + a\|\theta\|_2^2) + nY^2 \ln \left(\frac{TX^2}{a} + 1 \right).$$

Доказательство. Пусть $\eta = \frac{1}{2Y}$. По лемме 6.2 потери алгоритма APA выражаются в виде формулы (6.21). В нашем случае эта формула записывается в виде

$$L_T(APA) = \log_\beta \int_{\mathcal{R}^n} e^{-\eta L_T(\theta)} P_0(d\theta) =$$

$$= \log_\beta \int_{\mathcal{R}^n} (a\eta/\pi)^{n/2} e^{-\eta\theta'(aI+\sum\limits_{t=1}^{T} x_t x_t')\theta + 2\eta(\sum\limits_{t=1}^{T} y_t x_t')\theta - \eta \sum\limits_{t=1}^{T} y_t^2} d\theta. \quad (6.64)$$

Экспонента в (6.64) 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 6.4 выражение (6.64), представляющее потери APA, равно

$$L_T(APA) =$$

$$= \log_\beta \left(((a\eta/\pi)^{n/2}) \frac{\pi^{n/2} e^{-\eta F(\theta_0)}}{\sqrt{\det \left(a\eta I + \eta \sum\limits_{t=1}^{T} x_t x_t' \right)}} \right) =$$

$$= F(\theta_0) - \frac{1}{2} \log_\beta \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) =$$

$$= F(\theta_0) + \frac{1}{2\eta} \ln \det \left(I + \frac{1}{a} \sum_{t=1}^{T} x_t x_t' \right) =$$

$$= F(\theta_0) + \frac{1}{2\eta} Y^2 \ln \det \left(I + \sum_{t=1}^{T} x_t x_t' \right).$$

По определению (6.57) 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. $\triangle$

6.9.2. 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:

$$A = aI + \sum_{t=1}^{T} x_t x_t',$$
$$b = \sum_{t=1}^{T-1} y_t x_t',$$
$$\gamma_T = b' A^{-1} x_T =$$
$$= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T. \qquad (6.65)$$

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 (6.65) 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 6.6. *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*

$$B(aI_n + CB)^{-1} = (aI_m + BC)^{-1} B, \tag{6.66}$$

где a – любое вещественное число и I_n – единичная матрица размера n. [2]

[2] Далее индекс n опускаем.

Доказательство. Равенство (6.66) эквивалентно равенству

$$(aI_n + BC)B = B(aI_m + CB),$$

которое очевидно ввиду дистрибутивности умножения матриц. $\triangle$

Используя лемму представим значение предсказания линейной регрессии

$$\begin{aligned}
\gamma_T = b' A^{-1} x_T = \\
= \left(\sum_{t=1}^{T-1} y_t x_t' \right) \left(aI + \sum_{t=1}^{T} x_t x_t' \right)^{-1} \cdot x_T = \\
= (Y_{T-1}, 0)' X_T \left(aI + X_T' X_T \right)^{-1} x_T = \\
(Y_{T-1}, 0)' \left(aI + X_T X_T' \right)^{-1} X_T x_T = \\
= (Y_{T-1}, 0)' \left(aI + \tilde{K}_T \right)^{-1} \tilde{k}_T, \qquad (6.67)
\end{aligned}$$

где $\tilde{K}_T = X_T X_T'$ и $\tilde{k}_T = X_T x_T$. Заметим также, что

$$X_T X_T' = (x_t \cdot x_{t'})_{t,t'=1}^{T},$$
$$\tilde{k}_T = (x_t \cdot x_T)_{t=1}^{T},$$

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 регресии 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 ядерной многомерной регрессии

$$\gamma_T = (Y_{T-1}, 0)' \left(aI + K_T \right)^{-1} k_T.$$

Оценка ошибки предсказания для ядерной многомерной регрессии имеет вид

Теорема 6.4. *Если* $|y_t| \leqslant X$ *и* $||x_t||_2 \leqslant X$ *для всех* t, *то*

$$L_T(AAR) \leqslant \inf_{\theta} (L_T(\theta) + a|\theta|_2^2) + nY^2 \ln \left(\frac{TX^2}{a} + 1 \right)$$

для всех T.

6.9.3. Двойственная форма задачи регрессии

Дадим теперь определение двойственной формы произвольного алгоритма предсказания.

Пусть алгоритм предсказания $\mathcal{A}$ использует входные векторы $x_1, x_2, \dots x_T$ только в виде их скалярных произведений.

Пусть задано ядро $K(x, y)$, где $x, y \in \mathcal{R}^n$.

Формула для предсказания методом гребневой регрессии

$$\gamma_{T+1} = w'x = ((aI + X_T'X_T)^{-1}X_T'Y_T)' \cdot x_{T+1}$$

может быть преобразована с помощью леммы 6.6 к виду

$$\begin{aligned}\gamma_{T+1} = w'x = ((aI + X_T'X_T)^{-1}X_T'Y_T)' \cdot x_{T+1} = \\ = Y_T'X_T(aI + X_T'X_T)^{-1} \cdot x_{T+1} = \\ = Y_T'(aI + X_TX_T')^{-1}X_T \cdot x_{T+1} = \\ = Y_T'(aI + X_TX_T')^{-1}X_T \cdot k_{T+1}.\end{aligned}$$

Заметим, что в этих же обозначениях формула (6.67) для адаптивной ядерной регрессии имеет несколько отличный вид:

$$\gamma_{T+1} = (Y_T, 0)'\,(aI + K_{T+1})^{-1} \cdot k_{T+1}.$$

6.10. Задачи и упражнения

1. Нарисовать графики предсказаний и области суперпредсказаний, а также их образы в экспоненциальном пространстве для квадратичной, логарифмической, асолютной 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. Доказать, что абсолютная функция потерь не является смешиваемой.

3. Проверить выпуклость кривой (6.36).

4. Нарисовать график кривой $c(\beta)$, заданной равенством (6.37).

5. Вывести неравенство (6.38). Найти оптимальное значение β, для которого регрет в (6.38) принимает минимальное значение.

6. Использовать лемму 4.2 из раздела 4.2 для получения оценки с регретом $O(\sqrt{T})$.

6.11. Лабораторные работы

Использовать данные из следующих сайтов для решения задач регрессии.

$$http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/.$$

База данных UCI Machine Learning Repository находится на сайте

$$http://archive.ics.uci.edu/ml/datasets.html$$

Она содержит 185 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, 2.9.2. 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 6.9. Провести сравнительный анализ точности регрессии с другими методами.

Глава 7

Элементы теории игр

В данной главе мы сначала 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. Далее, в разделе 9, мы применим минимаксную теорему для решения бесконечно повторяющихся игр на предсказания. В частности, будет рассмотрен новый теоретико-игровой подход к теории вероятностей, предложенный Вовком и Шейфером [24]. В рамках этого подхода наиболее естественным образом формулируется и решается задача построения универсальных предсказаний, рассмотренная в главе 3.

7.1. Антагонистические игры двух игроков

Пусть X и Y – множества произвольной природы. Рссмотрим антагонистическую игру двух лиц. Первый игрок выбирает *стратегию* $x \in X$; 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 4.6 с функцией потерь $\lambda(x,y) = f(x,y)$.

В алгоритме экспоненциального взвешивания (7.6) величины y^i – прогнозы экспертов, $i = 1, \ldots, N$, x_t – исходы, $t = 1, \ldots, n$, y_t – прогноз *Статистика*. По (4.42) получим

$$\sum_{t=1}^{n} f(x_t, y_t) \leqslant \min_{i=1,\ldots,N} \sum_{t=1}^{n} f(x_t, y^i) + \sqrt{\frac{1}{2} n \ln N}.$$

Делим это неравенство на n :

$$\frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) \leqslant \min_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} f(x_t, y^i) + \sqrt{\frac{\ln N}{2n}}. \tag{7.7}$$

Пользуемся выпуклостью функции f по второму аргументу и вогнутостью по первому, а также используя (7.7), получаем

$$\begin{aligned}
&\inf_{y \in Y} \sup_{x \in X} f(x, y) \leqslant \\
&\leqslant \sup_{x \in X} f\left(x, \frac{1}{n} \sum_{t=1}^{n} y_t\right) \leqslant \\
&\leqslant \sup_{x \in X} \frac{1}{n} \sum_{t=1}^{n} f(x, y_t) \leqslant \\
&\leqslant \frac{1}{n} \sum_{t=1}^{n} \sup_{x \in X} f(x, y_t) \leqslant \\
&\leqslant \frac{1}{n} \sum_{t=1}^{n} f(x_t, y_t) + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ ...,\ N} \frac{1}{n} \sum_{t=1}^{n} f(x_t, y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \min_{i=1,\ ...,\ N} f\left(\frac{1}{n} \sum_{t=1}^{n} x_t, y^i\right) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \leqslant \\
&\leqslant \sup_{x \in X} \min_{i=1,\ ...,\ N} f(x, y^i) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n}.
\end{aligned} \tag{7.8}$$

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 $X = \{1,\ \ldots,\ N\}$, $Y = \{1,\ \ldots,\ M\}$. 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 $(\mathcal{X}, \mathcal{Y}, \bar{f}(\bar{p}, \bar{q}))$,

где $\mathcal{X}$ – множество смешанных стратегий первого игрока, $\mathcal{Y}$ – множество смешанных стратегий второго игрока, $\bar{f}(\bar{p},\bar{q})$ – среднее значение выигрыша относительно меры $p \times q$:

$$\mathcal{X} = \{\bar{p} = (p_1,\ \ldots,\ p_N) : \sum_{i=1}^{N} p_i = 1,\ p_i \geqslant 0\};$$

$$\mathcal{Y} = \{\bar{q} = (q_1,\ \ldots,\ q_M) : \sum_{i=1}^{M} q_i = 1,\ q_i \geqslant 0\};$$

$$\bar{f}(\bar{p},\bar{q})) = \sum_{i=1}^{N}\sum_{j=1}^{M} f(i,j)p_i q_j.$$

Имеет место минимаксная теорема Дж. фон Неймана.

Теорема 7.3. *Всякая матричная игра имеет решение в смешанных стратегиях:*

$$\max_{\bar{p}\in\mathcal{X}} \min_{\bar{q}\in\mathcal{Y}} \bar{f}(\bar{p},\bar{q}) = \min_{\bar{q}\in\mathcal{Y}} \max_{\bar{p}\in\mathcal{X}} \bar{f}(\bar{p},\bar{q}).$$

Доказательство. Достаточно доказать, что функция $\bar{f}(\bar{p},\bar{q}))$ имеет седловую точку. Применим теорему 7.2. Множества $\mathcal{X}$ и $\mathcal{Y}$ – симплексы в евклидовых пространствах, поэтому являются выпуклыми. Функция $\bar{f}(\bar{p},\bar{q})$ – 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 $p \in \mathcal{X}$, 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 $\max_{\bar{p}\in\mathcal{X}} \min_{\bar{q}\in\mathcal{Y}} \bar{f}(\bar{p},\bar{q})$.

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 $\min_{\bar{q}\in\mathcal{Y}} \max_{\bar{p}\in\mathcal{X}} \bar{f}(\bar{p}, \bar{q})$.

Нетрудно убедиться, что в этом случае по-прежнему верна минимаксная теорема 7.3, т.е.

$$\max_{\bar{p}\in\mathcal{X}} \min_{\bar{q}\in\mathcal{Y}} \bar{f}(\bar{p}, \bar{q}) = \min_{\bar{q}\in\mathcal{Y}} \max_{\bar{p}\in\mathcal{X}} \bar{f}(\bar{p}, \bar{q}).$$

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 $\bar{f}(1_i, 1_j) = f(i, j) = a_{i,j}$.

Теорема 7.4. *Для того чтобы пара смешанных стратегий $(\bar{p}^*, \bar{q}^*)$ была решением (седловой точкой) смешанного расширения матричной игры $(\mathcal{X}, \mathcal{Y}, \bar{f}(\bar{p}^*, \bar{q}^*))$, необходимо и достаточно, чтобы выполнялось неравенство*

$$\bar{f}(1_i, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, \bar{q}^*) \leqslant \bar{f}(\bar{p}^*, 1_j) \tag{7.9}$$

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 $\bar{p} = \sum_{i=1}^{N} p_i 1_i$, аналогичным образом представляется смешанная стратегия $\bar{q} = \sum_{j=1}^{M} q_j 1_j$. Поэтому можно рассмотреть дважды линейную комбинацию неравенства (7.9).

Получим

$$\bar{f}(\bar{p},\bar{q}^*)=\sum_{i=1}^{N}p_i\bar{f}(1_i,\bar{q}^*)\leqslant\sum_{i=1}^{N}p_i\bar{f}(\bar{p}^*,\bar{q}^*)=\bar{f}(\bar{p}^*,\bar{q}^*),$$
$$\bar{f}(\bar{p}^*,\bar{q}^*)=\sum_{j=1}^{M}q_j\bar{f}(\bar{p}^*,\bar{q}^*)\leqslant\sum_{j=1}^{M}q_j\bar{f}(\bar{p}^*,1_j)=\bar{f}(\bar{p}^*,\bar{q})$$

для всех $\bar{p}$ и $\bar{q}$. Отсюда получаем условие седловой точки:

$$\bar{f}(\bar{p},\bar{q}^*)\leqslant\bar{f}(\bar{p}^*,\bar{q}^*)\leqslant\bar{f}(\bar{p}^*,\bar{q})$$

для любых $\bar{p}$ и $\bar{q}$. $\triangle$

Теорема 7.5. *Для смешанного расширения произвольной матричной игры справедливы соотношения*

$$\min_{\bar{q}}\bar{f}(\bar{p},\bar{q})=\min_{j}\bar{f}(\bar{p},1_j),$$
$$\max_{\bar{p}}\bar{f}(\bar{p},\bar{q})=\max_{i}\bar{f}(1_i,\bar{q}).$$

Доказательство. Очевидно, что

$$\min_{\bar{q}}\bar{f}(\bar{p},\bar{q})\leqslant\min_{j}\bar{f}(\bar{p},1_j),$$
$$\max_{\bar{p}}\bar{f}(\bar{p},\bar{q})\leqslant\max_{i}\bar{f}(1_i,\bar{q}).$$

Противоположное неравенство следует из неравенства

$$\bar{f}(\bar{p},\bar{q})=\sum_{i=1}^{N}\sum_{j=1}^{M}a_{i,j}p_iq_j=$$
$$=\sum_{j=1}^{M}\left(\sum_{i=1}^{N}a_{i,j}p_i\right)q_j\geqslant$$
$$\geqslant\left(\min_{j}\sum_{i=1}^{N}p_ia_{i,j}\right)\left(\sum_{j=1}^{M}q_j\right)=$$
$$=\min_{j}\sum_{i=1}^{N}p_ia_{i,j}=\min_{j}\bar{f}(\bar{p},1_j),$$

которое имеет место для любого $\bar{q}$. Это неравенство означает, что минимум взвешенной линейной комбинации достигается, когда весь вес сосредоточен на наименьшем элементе. Следовательно,

$$\min_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) \geqslant \min_{j} \bar{f}(\bar{p}, 1_j).$$

Второе неравенство доказывается аналогичным образом. $\triangle$

Из этой теоремы получаем

Следствие 7.1. *В смешанном расширении произвольной матричной игры выполнено равенство*

$$v = \max_{\bar{p}} \min_{j} \bar{f}(\bar{p}, 1_j) = \min_{\bar{q}} \max_{i} \bar{f}(1_i, \bar{q}),$$

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

$$\begin{aligned} \bar{f}(\bar{p}, \bar{q}) = \sum_{i,j=1}^{2} a_{i,j} p_i p_j = \\ = q(-p + 1 - p) + (1-q)(p - (1-p)) = \\ = 4\left(p - \frac{1}{2}\right)\left(q - \frac{1}{2}\right). \end{aligned}$$

Среднее выигрыша:

$$\bar{f}(\bar{p}, \bar{q}) = 4\left(p - \frac{1}{2}\right)\left(q - \frac{1}{2}\right) \tag{7.10}$$

представляет собой уравнение однополостного гиперболоида.

Пусть

$$v(p) = \min_{j} \bar{f}(\bar{p}, 1_j) = \min\{1 - 2p, 2p - 1\}.$$

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: $v^* = \bar{f}(\bar{p}^*, \bar{q}^*) = 0$. 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 p^* 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 $\bar{f}(1_i, \bar{q}^*) = v^*$,*
- *из $q_j^* > 0$ следует $\bar{f}(\bar{p}^*, 1_j) = v^*$.*

Доказательство. Докажем первое утверждение. По определению $\bar{f}(1_i, \bar{q}^*) \leqslant v^*$, $i = 1, \ldots, N$.

Допустим, что существует i_0 такое, что $p_{i_0}^* > 0$ и одновременно $\bar{f}(1_{i_0}, \bar{q}^*) < v^*$. Рассмотрим линейную комбинацию неравенств $\bar{f}(1_i, \bar{q}^*) \leqslant v^*$ с коэффициентами p_i^*, $i = 1, \ldots, N$, и, так как одно из складываемых неравенств является строгим, получим

$$v^* = \bar{f}(\bar{p}^*, \bar{q}^*) = \sum_{i=1}^{N} \bar{f}(1_i, \bar{q}^*) p_i^* < v^* = \bar{f}(\bar{p}^*, \bar{q}^*).$$

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* $\bar{f}(1_i, \bar{q}^*) < v^*$ *следует* $p_i^* = 0$,
- *из* $\bar{f}(\bar{p}^*, 1_j) > v^*$ *следует* $q_j^* = 0$.

Условие $\bar{f}(\bar{p}^*, 1_j) = pa_{1,j} + (1-p)a_{2,j} > v^*$ 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

$$\bar{f}(\bar{p}^*, 1_j) = L_j(p^*) > v^*$$

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

$$v = \max_{p} \min_{j} \bar{f}(\bar{p}, 1_j) = \min_{\bar{q}} \max_{i} \bar{f}(1_i, \bar{q}), \tag{7.11}$$

где v – значение игры. Поэтому существует смешанная стратегия $\bar{p} = (p_1, \ldots, p_N)$ первого игрока, такая, что $\bar{f}(\bar{p}, 1_j) \geqslant v$ для 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

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \text{ при } j = 1, \ldots, M,$$

$$\sum_{i=1}^{N} x_i = 1/v,$$

$$x_i \geqslant 0 \text{ при } i = 1, \ldots, N.$$

[1] Для того чтобы этого добиться, достаточно прибавить некоторую достаточно большую положительную константу к каждому элементу платежной матрицы игры.

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

$$\sum_{i=1}^{N} a_{i,j} x_i \geqslant 1 \text{ при } j = 1,\ \ldots,\ M,$$

$$x_i \geqslant 0 \text{ при } i = 1,\ \ldots,\ N.$$

Для второго игрока по (7.11) существует смешанная стратегия $\bar{q} = (q_1,\ \ldots,\ q_N)$ первого игрока, такая, что $\bar{f}(1_i, \bar{q}) \leqslant v$ для любой чистой стратегии 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 $\bar{p}^k = (p^k_1,\ \ldots,\ p^k_{N_k})$ на множестве всех его стратегий $\{1,\ \ldots,\ N_k\}$. Здесь p^k_j – 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^k_i.

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:

$$\pi(\bar{i}) = \pi(\bar{I} = \bar{i}) = p^1_{i_1} \cdot\ \ldots\ \cdot p^K_{i_K}.$$

Математическое ожидание выигрыша k-го игрока равно

$$E_\pi(f^k(\bar{I})) = \sum_{\bar{i}} \pi(\bar{i}) f^k(\bar{i}) =$$

$$= \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} p^1_{i_1} \cdot \ \ldots \ \cdot p^K_{i_K} f^k(i_1, \ \ldots, \ i_K).$$

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 существаниии 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

$$\bar{f}(\bar{p}, \bar{q}^0) \leqslant \bar{f}(\bar{p}^0, \bar{q}^0) \leqslant \bar{f}(\bar{p}^0, \bar{q}),$$

где $\bar{f}(\bar{p}, \bar{q})$ – математическое ожидание выигрыша первого игрока, а $-\bar{f}(\bar{p}, \bar{q})$ – 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 лий с нулевой суммой тогда и только тогда, когда*

$$\bar{q}^* \in \{\bar{q} : \min_{\bar{p}} \bar{f}(\bar{p}, \bar{q}) \to \max\},$$
$$\bar{p}^* \in \{\bar{p} : \max_{\bar{q}} \bar{f}(\bar{p}, \bar{q}) \to \min\}.$$

Для любой такой пары $(\bar{p}^*, \bar{q}^*)$ *выполнено* $\bar{f}(\bar{p}^*, \bar{q}^*) = v$, *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 не имеет седловой точки, но имеет решение в смешанных стратегиях: для первого и второго игрока их смешанные стратегии имеют вид $\bar{p}^* = (\frac{1}{2}, \frac{1}{2})$ и $\bar{q}^* = (\frac{1}{2}, \frac{1}{2})$. 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. 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:

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 (T, T) (доказательство в виде задачи).

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

$$E_P(f^k(\bar{i})) \geqslant E_P(f^k(\bar{i}_{-k}, h(i_k))), \tag{7.12}$$

где вектор $\bar{i} = (i_1,\ \ldots,\ i_K)$ распределен в соответствии с вероятностным распределением P, а также

$$\bar{i}_{-k} = (i_1,\ \ldots,\ i_{k-1}, i_{k+1},\ \ldots,\ i_K)$$
$$(\bar{i}_{-k}, h(i_k)) = (i_1,\ \ldots,\ i_{k-1}, h(i_k), i_{k+1},\ \ldots,\ i_K).$$

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

$$\sum_{\bar{i}: i_k = j} P(\bar{i})(f^k(\bar{i}) - f^k(\bar{i}_{-k}, j') \geqslant 0, \tag{7.13}$$

где $(\bar{i}_{-k}, j') = (i_1,\ \ldots,\ i_{k-1}, j', i_k,\ \ldots,\ i_K)$.

Условие (7.13) можно записать также в виде:

$$E(f^k(\bar{i})|i_k = j) \geqslant E(f^k(\bar{i}_{-k}, j')|i_k = j), \tag{7.14}$$

где E – условное математическое ожидание относительно распределения P. [2]

Доказательство. Условие (7.12) коррелированного равновесия эквивалентно совокупности условий:

$$\sum_{\bar{i}} P(\bar{i})(f^k(\bar{i}) - f^k(\bar{i}_{-k}, h(i_k))) \geqslant 0, \tag{7.15}$$

[2] Часто именно это условие удобно принять в качестве определения коррелированного равновесия.

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 (7.15) останутся только слагаемые, соответствующие наборам $\bar{i}$, в которых $i_k = j$, а в остальных слагаемых соответствующие разности сократятся. Таким образом, сумма (7.15) превратится в сумму (7.13).

В обратную сторону, утверждение тривиально. $\triangle$

Пусть P – некоторое распределение вероятностей на множестве $\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$ и $a \in A_k$ для некоторого $1 \leqslant k \leqslant K$. Обозначим посредством $P_{-i}(\cdot|i_k = a)$ соответствующее условное распределение на множестве $\prod_{s=1,s\neq k}^{K}\{1,\ \ldots,\ N_s\}$ произвольного набора $\bar{i}_{-k}$ из этого множества при известном $i_k = a$. Введем также обозначение

$$f^k(j, \bar{P}_{-k}(\cdot|i_k = a))) = E_{\bar{P}_{-k}(\cdot|i_k=a))}(f^k(j, \bar{i}_{-k}))$$

– математическое ожидание функции выигрыша, в которой $i_k = a$, относительно этого условного распределения. 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 $i_k = a$.

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 стратегии $j, j' \in \{1,\ \ldots,\ N_k\}$ выполнено*

$$f^k(j, \bar{P}_{-k}(\cdot|i_k = j)) = \max_{j' \in A_i} f^k(j', \bar{P}_{-k}(\cdot|i_k = j)). \qquad (7.16)$$

Каждое условие типа (7.13) задает замкнутую полуплоскость, поэтому множество всех коррелированных равновесий представляет собой замкнутый выпуклый многогранник в пространстве всех мер на множестве $\prod_{k=1}^{K}\{1,\ \ldots,\ N_k\}$.

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:

$$\begin{aligned} \min_{\bar{q}} \bar{f}(\bar{p}^*, \bar{q}) &= \min_{j} \bar{f}(\bar{p}^*, 1_j), \\ \max_{\bar{p}} \bar{f}(\bar{p}, \bar{q}^*) &= \max_{i} \bar{f}(1_i, \bar{q}^*), \end{aligned} \tag{7.17}$$

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 (T, T).

5. Показать, что произвольная выпуклая комбинация равновесий Нэша является коррелированным равновесием.

Глава 8

Теоретико-игровая интерпретация теории вероятностей

В этой главе мы рассмотрим новый теоретико-игровой подход к теории вероятностей, предложенный Вовком и Шейфером [24].

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 [24] 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 8.2.

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 [24] 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 ограничен безопасной стратегией.

Рассмотрим бесконечно повторяющуюся *ограниченную* игру на предсказания между тремя игроками: *Предсказатель*, *Скептик* и *Природа*.

Действия игроков регулируются следующим протоколом:

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

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 капитал *Скептика* увеличивается или уменьшается на соответствующую величину. Заметим, что может

[1] В случае бинарных исходов 0 и 1 среднее значение равно вероятности 1.

быть $M_n < 0$. В этом случае *Скептик* продает некоторое количество единиц финансового инструмента.

Скептик выигрывает в этой игре, если $\mathcal{K}_n \geqslant 0$ для всех n и $\sup \mathcal{K}_n = \infty$ (какие бы ходы не предпринимали *Предсказатель* и *Природа*); в противном случае выигрывают *Природа* и *Предсказатель*.

Под реализацией игры называется последовательность ходов всех игроков: $p_1, M_1, \omega_1, p_2, M_2, \omega_2, \ldots$. При этом не предполагается, имеется закон для выбора ходов участников. Если такой закон имеется, называем его стратегией. Например, будет определена стратегия *Скептика*: на каждом шаге значение M_n будет определяться с помощью последовательности функций

$$M_n = M_n(p_1, M_1, \omega_1, p_2, M_2, \omega_2, \ldots\ , p_{n-1}, M_{n-1}, \omega_{n-1}, p_n).$$

Теоретико-игровой закон больших чисел формулируется в виде следующей теоремы.

Теорема 8.1. *Можно построить такую безопасную стратегию Скептика, что для любой реализации приведенной выше игры выполнено следующее: если не выполнен усиленный закон больших чисел*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0, \tag{8.1}$$

то Скептик выигрывает в ограниченной игре на предсказание, более того, он может так выбирать свои ходы M_n, что $\mathcal{K}_n \geqslant 0$ для всех n и $\limsup\limits_{n\to\infty} \frac{\ln \mathcal{K}_n}{n} > 0$.

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.

Рассмотрим первый случай. Так как $|\omega_i - p_i| \leqslant 1$, отсюда

$$\epsilon \sum_{i=1}^{n} (\omega_i - p_i) - \epsilon^2 \sum_{i=1}^{n} (\omega_i - p_i)^2 > \epsilon^2 n.$$

Используем неравенство $t - t^2 \leqslant \ln(1+t)$, которое выполнено при всех $t \geqslant 1/2$, 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 его логарифм равен

$$\ln \mathcal{K}_n = \sum_{i=1}^{n} \ln(1 + \epsilon(\omega_i - p_i)) > \epsilon^2 n,$$

Отсюда получаем, что

$$\limsup_{n \to \infty} \frac{\ln \mathcal{K}_n}{n} \epsilon^2 > 0. \tag{8.5}$$

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.

Недостаток 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 $\mathcal{K}_n$:

$$\begin{aligned} \mathcal{K}_n^{+} &= \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{1,k}, \\ \mathcal{K}_n^{-} &= \sum_{k=1}^{\infty} 2^{-k} \mathcal{K}_n^{2,k}, \\ \mathcal{K}_n &= \frac{1}{2}(\mathcal{K}_n^{+} + \mathcal{K}_n^{-}). \end{aligned}$$

Все эти ряды сходятся, так как для любого фиксированного n будет $\mathcal{K}_n^{1,k} \leqslant 2^n$ для всех k. Отсюда и из определения $|S_n^{2,k}| \leqslant 2^{n-1}$ 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) нарушается при некотором $\epsilon = \epsilon_k$ для бесконечно многих n. Из условия (8.5), в котором $\mathcal{K}_n = \mathcal{K}_n^{s,k}$, следует, что

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n^{s,k}}{n} > 0$$

для $s = 0$ или 1. Отсюда следует, что

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n}{n} > 0.$$

Теорема доказана. $\triangle$

Теоретико-игровую форму закона больших чисел получим путем обращения (и некоторого ослабления) утверждения теоремы 8.1.

Следствие 8.1. *Можно построить такую безопасную стратегию Скептика, что для любой реализации ограниченой игры на предсказания 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. 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*.

Действия игроков регулируются следующим протоколом:

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 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*.

Теорема 8.2. *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. $\triangle$

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 мер на $\{0,1\}$. [2]

Игра регулируется следующим протоколом.

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 $p_n \in [0,1]$.

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

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. [3]

---

[2] Каждая мера $Q \in \mathcal{P}\{0,1\}$ задается двумя числами $(q, 1-q)$, где $q = Q\{1\}$ – вероятность 1.

[3] Выигрыш $\mathcal{F}_n$ соответствует понятию ограниченного снизу супермартин-

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 8.3. *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 $r \in [0,1] \setminus \{p_n\}$.

Тогда математическое ожидание выигрыша *Природы* относительно смешанной стратегии Q и чистой стратегии p_n равно

$$\begin{aligned} F(Q_n, P_n) &= Q\{0\}F(0,p_n) + Q\{1\}F(1,p_n) = \\ &= Q\{0\}S(p_n)(-p_n) + Q\{1\}S(p_n)(1-p_n) = \\ &= (1 - Q\{1\})S(p_n)(-Q\{1\}) + Q\{1\}S(p_n)(1 - Q\{1\}) = 0. \end{aligned}$$

гала 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.

Таким образом, $\forall\ Q\ \exists\ P\ F(Q,P) \leqslant 0$ или

$$\max_Q \min_P F(Q,P) \leqslant 0. \tag{8.6}$$

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 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). [4] Тогда неравенство (8.6) будет преобразовано в неравенство

$$\max_Q \min_P F(Q,P) \leqslant \Delta/2.$$

Согласно минимаксной теореме,

$$\min_P \max_Q F(Q,P) = \max_Q \min_P F(Q,P) \leqslant \Delta/2.$$

Поэтому *Предсказатель* имеет смешанную стратегию $P \in \mathcal{P}[0,1]$, сосредоточенную на множестве N_ϵ, такую, что

$$\max_Q F(Q,P) \leqslant \Delta,$$

откуда следует, что

$$\int S_n(p)(\omega_n - p)P(dp) \leqslant \Delta \tag{8.7}$$

[4] Скептик должен выбирать $S_n(p)$ так, что $\mathcal{K}_n \geqslant 0$ для всех n независимо от действий других игроков.

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 (8.7) 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

$$\int S_n(p)(\omega_n - p)P_n(dp) \leqslant 0 \tag{8.8}$$

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 (8.8), 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 $\sup_n \mathcal{F}_n < \infty$.

8.3. 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 $0 < \epsilon < 1$. 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

$$\mathcal{K}_n^1 = \prod_{i=1}^{n} (1 + \epsilon(\omega_i - p_i)), \tag{8.9}$$

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 8.3 *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

$$\begin{aligned}
\sum_{i=1}^{n}\ln(1+\epsilon(\omega_i - p_i)) &\leqslant \ln C,\\
\epsilon\sum_{i=1}^{n}(\omega_i - p_i) - \epsilon^2\sum_{i=1}^{n}(\omega_i - p_i)^2 &\leqslant \ln C,\\
\epsilon\sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \ln C + \epsilon^2 n,\\
\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) &\leqslant \frac{\ln C}{\epsilon n} + \epsilon \qquad (8.10)
\end{aligned}$$

для всех n. Здесь мы использовали неравенство $\ln(1+t) \geqslant t - t^2$ при $|t| \leqslant 0.5$.

Отсюда следует, что

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \leqslant \epsilon. \qquad (8.11)$$

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

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\omega_i - p_i) \geqslant -\epsilon. \qquad (8.12)$$

Обе эти стратегии можно соединить в одну стратегию, которая обеспечивает одновременное выполнение обоих неравенств (8.11)

и (8.12). 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 (8.11) и (8.12).

Следующий шаг заключается в том, чтобы построить стратегию *Скептика*, которая обеспечивает одновременное выполнение неравенств (8.11) и (8.12) 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 (8.9). 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 (8.11) и (8.12) уже теперь для всех ϵ_k, $k = 1, 2, \ldots$

Отсюда получаем, что смешанная стратегия *Скептика* вынуждает *Предсказателя*, чтобы выиграть в игре согласно теореме 8.3, выдавать такие прогнозы, что выполнено следующее условие калибруемости:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\omega_i - p_i) = 0. \tag{8.13}$$

Определение калибруемости (8.13) 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

$$\lim_{n\to\infty} \frac{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{i-1}, \omega_{i-1}, p_i)(\omega_i - p_i)}{\sum\limits_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2,\ \ldots,\ p_{i-1}, \omega_{i-1}, p_i)} = 0. \quad (8.14)$$

Основной результат теории универсального прогнозирования утверждает

Теорема 8.4. *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*:

$$S_n^{1,k,s}(p) = \epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p)\mathcal{K}_{n-1}^{1,k,s},$$
$$S_n^{2,k,s}(p) = -\epsilon_k F_s(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p)\mathcal{K}_{n-1}^{2,k,s}.$$

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 (8.10) 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 (8.10) и в (8.13) для того, чтобы получить (8.14), надо n в знаменателе заменить на

$$\sum_{i=1}^{n} F(p_1, \omega_1, p_2, \omega_2, \ \ldots, \ p_{i-1}, \omega_{i-1}, p_i).$$

8.4. Задачи и упражнения

1. Провести все выкладки для стратегии $S_n^2(p)$ и получить неравенство (8.12).

2. Завершить доказательство теоремы 8.4.

Глава 9

Повторяющиеся игры

В предыдущих разделах 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 корелированого равновесия в неограниченно повторяющихся мграх 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.

9.1. Бесконечно повторяющиеся игры двух игроков с нулевой суммой

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 $\bar{f}(\bar{p}_t, \bar{q}_t)$, а выигрыш второго игрока равен $-\bar{f}(\bar{p}_t, \bar{q}_t)$.

Будем сравнивать кумулятивный выигрыш каждого игрока за n шагов с его наилучшей константной стратегией:

$$\max_{i=1,\ldots,N} \sum_{t=1}^{n} f(i, J_t) - \sum_{t=1}^{n} f(I_t, J_t)$$

– для первого игрока и

$$\sum_{i=1}^{M} f(I_t, J_t) - \min_{j=1,\ldots,N} \sum_{t=1}^{n} f(I_t, j)$$

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 эксперта 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 $I_t \in \{1, \ldots, N\}$ второго эксперта 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.

Допустим, что оба игрока выбирают свой ходы в соответствии со смешанными стратегиями, состоятельными по Ханнану (см. определение (4.57)). Например, можно использовать алгоритм экспоненциального взвешивания из разделов 4.2 и 4.6. Согласно этому алгоритму первый игрок выбирает свою смешанную стратегию по формуле

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \lambda^1(i, J_s)\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \lambda^1(k, J_s)\right)},$$

где $i = 1, \ldots, N$, η – параметр обучения. При этом ход J_s второго игрока рассматривается как исход *Природы*.

По следствию 4.3 первый игрок является состоятельным по Ханнану, т.е. с вероятностью 1 имеет место

$$\limsup_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{n} \lambda^1(J_t, I_t) - \min_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} \lambda^1(J_t, i) \right) \leqslant 0 \qquad (9.1)$$

при соответствующем выборе параметра η.

Второй игрок может также может применять аналогичную стратегию. В этом случаю он также будет состоятельным по Ханнану, т.е. с вероятностью 1 имеет место

$$\limsup_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{n} \lambda^2(I_t, J_t) - \min_{j=1,\ldots,M} \frac{1}{n} \sum_{t=1}^{n} \lambda^2(I_t, j) \right) \leqslant 0. \qquad (9.2)$$

В терминах выигрышей (9.1) имеет вид: с вероятностью 1 вы-

полнено

$$\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} f(I_t,J_t) - \max_{i=1,\dots,\ N}\frac{1}{n}\sum_{t=1}^{n} f(i,J_t)\right)\geqslant 0, \tag{9.3}$$

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

$$\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} f(I_t,J_t) - \min_{j=1,\dots,\ M}\frac{1}{n}\sum_{t=1}^{n} f(I_t,j)\right)\leqslant 0, \tag{9.4}$$

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 смешанной стратегии, то независимо от того какой стратегии придерживается второй игрок, 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 смешанной стратегии, то независимо от того какой стратегии придерживается первый игрок, средний выигрыш 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 смешанной стратегии, состоятельной по Ханнану. Тогда*

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} f(I_i,J_t)\geqslant v, \tag{9.5}$$

почти всюду, где v – цена игры.

Если каждый игрок придерживается 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

$$v = \max_{\bar p} \min_{\bar q} \bar f(\bar p, \bar q) = \min_{\bar q} \max_{\bar p} \bar f(\bar p, \bar q).$$

Кроме того,

$$\bar f(\bar p, \bar q) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j f(p_i, q_j),$$

$$\bar f(\bar p, j) = \sum_{i=1}^{N} p_i f(p_i, j),$$

$$\bar f(i, \bar q) = \sum_{j=1}^{M} q_j f(i, q_j).$$

Согласно соотношению (9.3), для доказательства первого утверждения (9.5) теоремы достаточно показать, что

$$\max_{i=1,\dots,\ N} \frac{1}{n} \sum_{t=1}^{n} f(i, J_t) \geqslant v \tag{9.7}$$

для всех n. Для доказательства заметим, что

$$\max_{i=1,\dots,\ N} \frac{1}{n} \sum_{t=1}^{n} f(i, J_t) = \max_{\bar p} \frac{1}{n} \sum_{t=1}^{n} \bar f(\bar p, J_t),$$

так как $\sum\limits_{t=1}^{n} \bar f(\bar p, J_t)$ 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 $\hat q_n = (\hat q_{1,n}, \dots,\ q_{M,n})$. Тогда

$$\max_{\bar p} \frac{1}{n} \sum_{t=1}^{n} \bar f(\bar p, J_t) = \max_{\bar p} \sum_{j=1}^{M} \hat q_{j,n} \bar f(\bar p, j) =$$

$$= \max_{\bar p} \bar f(\bar p, \hat q_n) \geqslant \min_{\bar q} \max_{\bar p} \bar f(\bar p, \bar q) = v.$$

Для доказательства второго утверждения (9.6) теоремы воспользуемся условием (9.4) состоятельности по Ханнану и докажем, что

$$\min_{j=1,\ldots,\ M} \frac{1}{n} \sum_{t=1}^{n} f(I_t, j) \leqslant v = \max_{\bar{p}} \min_{\bar{q}} \bar{f}(\bar{p}, \bar{q}).$$

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 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 стратегия первого игрока, придерживаясь которой он с вероятностью 1 при неограниченном продолжении игры может как угодно близко приблизить среднее значение вектора своего выигрыша к заданному множеству S, независимо от того, какой бы стратегии не придерживался второй игрок.

В 1956 г. Блекуэлл [7] 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 $f(i,j)$ 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 $\bar{p} \in \mathcal{P}(\mathcal{I})$ и $\bar{q} \in \mathcal{P}(\mathcal{J})$ определяется

$$f(\bar{p}, j) = \sum_{i=1}^{N} f(p_i, j),$$
$$f(i, \bar{q}) = \sum_{j=1}^{M} f(i, q_j),$$
$$f(\bar{p}, \bar{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} f(i,j).$$

Для смешанных стратегий $\bar{p} = (p_1, \ldots, p_N) \in \mathcal{P}(\mathcal{I})$ и $\bar{q} = (q_1, \ldots, q_M) \in \mathcal{P}(\mathcal{J})$ получаем линейную комбинацию векторов $f(i,j)$:

$$\bar{f}(\bar{p}, \bar{q}) = \sum_{i,j} f(i,j) p_i q_j.$$

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 $d_S(\bar{y})$ обозначает какой-нибудь элемент $\bar{y} \in S$, для которого расстояние $\operatorname{dist}(\bar{x}, S)$ минимальное. Если к тому же S – выпуклое, то такой элемент единственный.

Множество $S \subseteq \mathcal{R}^d$ называется *достижимым (approachable)*, если существует такая последовательность $\bar{p}_1, \bar{p}_2, \ldots$ рандомизированных стратегий первого игрока, 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

$$\lim_{T\to\infty} \inf_{c\in S} \left\| c - \frac{1}{T}\sum_{t=1}^{T} f(I_t, J_t) \right\| = 0,$$

где $P = \prod \bar{p}_t$ – общее распределение вероятностей на траекториях $I_1, I_2, \ldots$ ходов первого игрока, которое определяется своими условными распределениями $\bar{p}_1, \bar{p}_2, \ldots$.

Следующая теорема дает достаточное условие достижимостых замкнутого выпуклого подмножества из $\mathcal{R}^d$.

Предполагаем, что рассматриваемые далее множества S и значения $f(i,j)$ находятся в единичном шаре пространства $\mathcal{R}^d$.

Теорема 9.2. *Задано замкнутое 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 $f(\bar{p}, 1), \ldots, f(\bar{p}, M)$ 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 стратегии первого и второго игроков. Обозначим вектор среднего значения выигрышей за первые t шагов

$$\bar{m}_t = \frac{1}{t}\sum_{i=1}^{t} f(I_i, J_i).$$

Пусть на шагах $< t$ игры игроки уже произвели ходы $I_1, \ldots, I_{t-1}$ и $J_1, \ldots, J_{t-1}$. Уравнение гиперплоскости $\Pi_{\bar{x}}$, проходящей через точку $\bar{x} = d_S(\bar{m}_{t-1})$ 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

$$b_{t-1} = (\bar{m}_{t-1} \cdot d_S(\bar{m}_{t-1})).$$

Предполагаем, что $\bar{m}_0 = \bar{0}$.

Заметим, что точка $d_S(\bar{m}_{t-1})$ находится выше гиперплоскости (так как является концом направляющего вектора этой гиперплоскости).

По условию теоремы для $\bar{x} = d_S(\bar{m}_{t-1})$ существует смешанная стратегия $\bar{p}_t$ первого игрока, для которой все точки

$$f(\bar{p}_t, 1), \ldots, \ f(\bar{p}_t, M)$$

находятся ниже данной гипервлоскости:

$$(\bar{w}_{t-1} \cdot f(\bar{p}_t, j)) - b_{t-1} \leqslant 0$$

для всех $j = 1, \ldots, \ M$. Это условие можно также записать в виде

$$\max_{1 \leqslant j \leqslant M} (\bar{w}_{t-1} \cdot (f(\bar{p}_t, j) - d_S(\bar{m}_{t-1}))) \leqslant 0. \tag{9.9}$$

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

$$d(\bar{m}_t, S) = \|\bar{m}_t - d_S(\bar{m}_t)\| \leqslant \|\bar{m}_t - d_S(\bar{m}_{t-1})\|. \tag{9.10}$$

Нетрудно проверить, что

$$\bar{m}_t = \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t} f(I_t, J_t). \tag{9.11}$$

Возведем в квадрат неравенство (9.10) и продолжим выкладки с

использованием (9.11) :

$$d(\bar{m}_t, S)^2 \leqslant \left\| \frac{t-1}{t}\bar{m}_{t-1} + \frac{1}{t}f(I_t, J_t) - d_S(\bar{m}_{t-1}) \right\|^2 =$$

$$= \left\| \frac{t-1}{t}(\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) + \frac{1}{t}(f(I_t, J_t) - d_S(\bar{m}_{t-1})) \right\|^2 =$$

$$= \left(\frac{t-1}{t}\right)^2 \|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 +$$

$$+2\frac{t-1}{t^2}((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))) +$$

$$+\frac{1}{t^2}\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2. \qquad (9.12)$$

Так как множество S и все значения $f(i,j)$ находятся в единичном шаре пространства $\mathcal{R}^d$, выполнено неравенство

$$\|f(I_t, J_t) - d_S(\bar{m}_{t-1})\| \leqslant 2.$$

Используя это неравенство преобразуем неравенства (9.11) и (9.12) в неравенство

$$t^2\|\bar{m}_t - d_S(\bar{m}_t)\|^2 - (t-1)^2\|\bar{m}_{t-1} - d_S(\bar{m}_{t-1})\|^2 \leqslant$$
$$\leqslant 4 + 2(t-1)((\bar{m}_{t-1} - d_S(\bar{m}_{t-1})) \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))). \quad (9.13)$$

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 $t \leqslant T$. Просуммируем при $t = 1, \ldots, T$ левую и правую части неравенства (9.13) и разделим их на T^2 :

$$\|\bar{m}_T - d_S(\bar{m}_T)\|^2 \leqslant$$

$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (f(I_t, J_t) - d_S(\bar{m}_{t-1}))) \leqslant$$

$$\leqslant \frac{4}{T} + \frac{2}{T}\sum_{t=1}^{T} K_{t-1}(\bar{w}_{t-1} \cdot (f(I_t, J_t) - f(\bar{p}_t, J_t)). \qquad (9.14)$$

Для получения последнего неравенства мы использовали неравенство (9.10).

Второе слагаемое последнего члена (9.14) представляет собой мартингал-разность. Поэтому оно по следствию 4.9 к неравенству Хефдинга–Азумы почти всюду стремится к 0 при $T \to \infty$. Отсюда

$$d(\bar{m}_T, S) = \|\bar{m}_T - d_S(\bar{m}_T)\| \to 0 \text{ при } T \to \infty$$

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 тогда и только тогда, когда для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что $f(\bar{p}, \bar{q}) \in S$.*

Доказательство. Допустим, что для каждого $\bar{q} \in \mathcal{P}(\mathcal{J})$ существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что $f(\bar{p}, \bar{q}) \in S$. 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 $a(i,j) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(i,j))$. 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 i такое, что $f(i, \bar{q}) \in S$. Отсюда и из (9.15) получаем

$$\begin{aligned} &\max_{\bar{p}} \min_{j}((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(\bar{p}, j)) = \\ &= \min_{\bar{q}} \max_{i}((\bar{d_S}(x_0) - \bar{x}_0) \cdot f(i, \bar{q})) \geqslant \\ &\qquad \geqslant \min_{\bar{s} \in S}((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{s}) = \\ &\qquad = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)). \end{aligned} \tag{9.16}$$

Рассмотрим гиперплоскость

$$L(\bar{x}) = ((d_S(\bar{x}_0) - \bar{x}_0) \cdot \bar{x}) - ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(x_0)) = 0$$

Легко проверить, что

$$((\bar{y} - \bar{x}_0) \cdot x_0) < ((\bar{y} - \bar{x}_0) \cdot y)$$

при $\bar{y} = d_S(\bar{x}_0)$. Отсюда $L(\bar{x}_0) < 0$. Из (9.16) следует, что существует $\bar{p} \in \mathcal{P}(\mathcal{I})$ такое, что для любого $j = 1, \ldots, M$

$$((d_S(\bar{x}_0) - \bar{x}_0) \cdot f(\bar{p}, j)) \geqslant ((d_S(\bar{x}_0) - \bar{x}_0) \cdot d_S(\bar{x}_0)).$$

Иными словами, $L(f(\bar{p}, j)) \geqslant 0$ для любого $j = 1, \ldots, M$, т.е. выполнено условие теоремы 9.2. Следовательно множество S достижимо.

Для доказательства обратного утверждения, допустим, что существует $\bar{q}_0 \in \mathcal{P}(\mathcal{J})$ такое что $f(\bar{p}, \bar{q}_0) \notin S$ для всех $\bar{p} \in \mathcal{P}(\mathcal{I})$.

Применим теорему 9.2 для игры с транспонированной матрицей (функцией) выигрыша $f'(i, j) = f(j, i)$ и выпуклого замкнутого множества $T(\bar{q}_0) = \{f(\bar{p}, \bar{q}_0) : \bar{p} \in \mathcal{P}(\mathcal{I})\}$.

Существует $\bar{q}_0$ такое, что все значения $f'(\bar{q}_0, 0), \ldots, f'(\bar{q}_0, M) \in T(\bar{q}_0)$. Из выпуклости множества $T(\bar{q}_0)$ следует, что для любого $\bar{x} \notin T(\bar{q}_0)$ точки $\bar{x}$ и $f'(\bar{q}_0, 0), \ldots, f'(\bar{q}_0, M)$ находятся по разные стороны от гиперплоскости $\Pi_{\bar{x}}$. Тогда по теореме 9.2 множество $T(\bar{q}_0)$ достижимо (для второго игрока первоначальной игры с постоянной стратегией $\bar{q}_0$).

Мы допустили, что $T(\bar{q}_0) \cap S = \emptyset$. Значит множество S не достижимо для первого игрока.

Теорема доказана. $\triangle$

В качестве первого применения теоремы 9.2 построим 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 $\prod_t \bar{p}_t$.

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 потерь

$$\bar{f}(i,j) = \begin{pmatrix} l(i,j) - l(1,j) \\ \cdots \\ l(i,j) - l(k,j) \\ \cdots \\ l(i,j) - l(N,j) \end{pmatrix}.$$

Заметим, что значения $f(i,j)$ лежат в N-мерном шаре радиуса $\sqrt{N}$ с центром в начале координат. Умножая функцию потерь на константу $1/\sqrt{N}$ можно добиться, чтобы значения $f(i,j)$ лежали в единичном шаре.

Рассмотрим произвольный вектор $\bar{x}_0 \notin S$. Достаточно рассмотреть случай когда $d_S(x_0) = \bar{0}$ и уравнение гиперлоскости $\Pi_{\bar{x}_0}$ 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 $f(\bar{p}, 1), \ldots,\ f(\bar{p}, M)$ 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 $j = 1, \ldots M$. 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 $\bar{p}_1, \ldots,\ \bar{p}_t, \ldots$ 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 [22].

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 $\|\bar{p}\|_1 = \max_{1 \leqslant i \leqslant m} |p_i|$. Можно также рассматривать широко известную эвклидову норму $\|\bar{p}\|_2$ в $\mathcal{R}^m$. 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-й элемент равен 1, остальные элементы – 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 $\bar{P}_t \in \mathcal{P}(\mathcal{P}(A))$. По теореме Ионеску-Тульчи [3] вероятност-

ные меры P_t можно рассматривать как условные распределения относительно некоторого общего распределения $P = \prod P_t$ на траекториях $\bar{p}_1, \bar{p}_2, \ldots$.

Каждый игрок может использовать всю информацию, известную до его действия. На стратегию *Природы* не накладывается никаких ограничений.

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* P-почти всюду было выполнено условие ϵ-калибруемости:

$$\limsup_{T\to\infty} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon} (\bar{p}_t - \bar{\delta}[a_t]) \right\| \leqslant \epsilon, \tag{9.18}$$

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 $\mathcal{P}_\epsilon = \{\bar{s}_1, \ldots, \bar{s}_N\}$. Для задания этого множества построим какую-нибудь ϵ-сеть 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 $\mathcal{I}$.

Общее распределение на тректориях $i_1, i_2, \ldots$ номеров определяется как $P = \prod P_t$. Тогда условие (9.18) очевидным образом следует из следующего условия: P-почти всюду выполнено условие

$$\limsup_{T\to\infty} \sum_{k=1}^{N} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\{i_t = k\}} (\bar{s}_k - \bar{\delta}[a_t]) \right\| \leqslant \epsilon, \tag{9.19}$$

где траектория $i_1, i_2, \ldots$ распределена по мере P.

Существование ϵ-калибруемой стратегии в общей форме утверждается в следующей теореме.

Теорема 9.4. *Для произвольного $\epsilon > 0$ можно построить вероятностное распределение P такое, что P-почти всюду выполнено условие ϵ-калибруемости (9.18), где векторы $\bar{p}_1, \bar{p}_2, \ldots$ распределены согласно P.* [1]

Доказательство. Мы применим теорему 9.3, в которой первый игрок – это *Предсказатель* с множеством стратегий [2] $\mathcal{I} = \{1, 2, \ldots, N\}$, а второй игрок – *Природа* с множеством стратегий $\mathcal{J} = A$. Функция выигрыша принимает в качестве значений векторы размерности $N|A|$:

$$f(k, a) = \begin{pmatrix} \bar{0} \\ \ldots \\ \bar{0} \\ \bar{s}_k - \bar{\delta}[a] \\ \bar{0} \\ \ldots \\ \bar{0} \end{pmatrix}.$$

где $k \in \mathcal{I}$ и $a \in \mathcal{J}$, $\bar{0}$ – m-мерный нулевой вектор, $m = |A|$, а также $\bar{s}_k - \bar{\delta}[a]$ – 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 $f(\bar{p}, \bar{q}) \in C$.

[1] Условие (9.18) эквивалентно условию (9.19).

[2] 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 $\|\bar{s}_i - \bar{q}\| \leqslant \epsilon$, т.е. $f(k, \bar{q}) \in C$. В этом случае мы берем в теореме 9.2 в качестве $\bar{p}$ распределение $\bar{\delta}[k]$ на $\mathcal{I} = \{1, \dots N\}$, сосредоточенное на числе k, где $1 \leqslant k \leqslant N$.

По теореме 9.2 можно построить рандомизированную стратегию *Предсказателя* $P = \prod P_t$, где $P_t \in \mathcal{P}(\mathcal{I})$ такую, что как бы Природа не выбирала последовательность $a_1, a_2, \dots$ последовательность векторно-значных выигрышей

$$\frac{1}{T}\sum_{t=1}^{T} f(i_t, a_t) = \begin{pmatrix} \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=1\}}(\bar{s}_1 - \bar{\delta}[a_t]) \\ \dots \\ \frac{1}{T}\sum\limits_{t=1}^{T} I_{\{i_t=N\}}(\bar{s}_N - \bar{\delta}[a_t]) \end{pmatrix}.$$

P-почти всюду сходится к множеству C, где траектория $i_1, i_2, \dots$ распределена по мере P.

Таким образом, условие калибруемости (9.19) выполнено почти всюду. Теорема доказана $\triangle$

Последовательность предсказаний называется 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 FV-предсказаниями.

Можно усилить теорему 9.4 и добиться калибруемости предсказаний.

Теорема 9.5. *Можно построить рандомизированную стратегию Предсказателя P такую, что для любого $\bar{p} \in \mathcal{P}(A)$ условие калибруемости*

$$\lim_{T\to\infty} \left\| \frac{1}{T}\sum_{t=1}^{T} I_{\|\bar{p}_t - \bar{p}\| \leqslant \epsilon}(\bar{p}_t - \bar{\delta}[a_t]) \right\| = 0 \tag{9.20}$$

выполнено P-почти, где последовательность векторов $\bar{p}_1, \bar{p}_2, \dots$ распределена по мере P.

Пусть ϵ_i – последовательность рациональных 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 надо строить предсказания, которые ϵ_i-калибруются на правой границе i-й эпохе и ϵ_{i-1}-калибруются на всей i-й эпохе. Мы не будем останавливаться на деталях этой 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 распределене вероятностей на конечном множестве мощности N есть N-мерный вектор: обозначим такое распределение $\bar{p}$. Поэтому в качестве нормы на таких распределениях можно рассматривать одну их эквивалентных норм $\|\bar{p}\|$ на $\mathcal{R}^N$ (эвклидову или максимум) и соответствующее расстояние $\mathrm{dist}(\bar{p}, \bar{q}) = \|\bar{p} - \bar{q}\|$. Расстояние от элемента $\bar{p} \in \mathcal{R}^N$ до множества $S \subseteq \mathcal{R}^N$ определяется:

$$\mathrm{dist}(\bar{p}, S) = \inf_{\bar{q} \in S} \mathrm{dist}(\bar{p}, \bar{q}).$$

Бесконечная последовательность $\bar{p}_1, \bar{p}_2, \ldots$ сходится к множеству S, если

$$\lim_{t \to \infty} \mathrm{dist}(\bar{p}_t, S) = 0.$$

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:

$$f^i(a, \bar{p}_{-i}) = E_{\bar{p}_{-i}}(f^i(a, \cdot)) = \sum_{\bar{a}_{-i} \in A_{-i}} f^i(a, \bar{a}_{-i}) \bar{p}_{-i}(\bar{a}_{-i}),$$

$$\bar{a}_{-i} = (a_1, \ldots, \ a_{i-1}, a_{i+1}, \ldots, \ a_k),$$

$$(a, \bar{a}_{-i}) = (a_1, \ldots, \ a_{i-1}, a, a_{i+1}, \ldots, \ a_k),$$

где $a \in A_i$, $\bar{a}_{-i} \in A_{-i}$, E – символ математического ожидания относительно меры $\bar{p}_{-i}$.

Пусть теперь игроки повторяют свою игру на шагах $t = 1, 2, \ldots$ согласно следующему протоколу.

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 $\bar{p}^t_{-i}$:

$$a^t_i \in \operatorname{argmax}_{a \in A_i} f^i(a, \bar{p}^t_{-i}). \tag{9.21}$$

ENDFOR

Мы называем любую стратегию a игрока i, на которой достигается максимум функции $\bar{f}^i(a, \bar{p}^t_{-i})$, *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 $\bar{a}^t = (a_1^t, \ldots, a_k^t)$ – набор ходов всех игроков на шаге t. Обозначим

$$\bar{p}_T = \frac{1}{T} \sum_{t=1}^{T} \bar{\delta}[\bar{a}^t] \tag{9.22}$$

- 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 $\bar{a}^t = (a_1^t, \ldots, a_k^t)$, 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 $(a_1^t, \ldots, a_k^t)$: $\prod_{i=1}^{k} |A_i|$.

Каждому набору стратегий $\bar{a} = (a_1, \ldots, a_k)$ соответствует число

$$\bar{p}_T(\bar{a}) = \frac{1}{T} |\{t : 1 \leqslant t \leqslant T, \bar{a}^t = \bar{a}\}| \tag{9.23}$$

– значение частотного распределения на наборе $\bar{a}$ (соответствующая координата вектора p_T).

Напомним, что последовательность предсказаний i-го игрока $\bar{p}_{-i}^t$ калибруется на последовательности исходов всех игроков, кроме i-го игрока, $\bar{a}_{-i}^t$, $t = 1, 2, \ldots$, если для произвольного вероятностного распределения $\bar{p}_{-i} \in A_{-i}$

$$\lim_{T \to \infty} \left\| \frac{1}{T} \sum_{t=1}^{T} I_{\|\bar{p}_{-i}^t - \bar{p}_{-i}\| \leqslant \epsilon} (\bar{p}_{-i}^t - \bar{\delta}[a_{-i}^t]) \right\| = 0, \tag{9.24}$$

где $\epsilon \geqslant 0$ – произвольное и I – индикаторная функция соответствующего условия.

Следующая теорема показывает, что если каждый игрок использует предсказания ходов остальных игроков, которые калибруются на наборах стратегий оппонентов, и выбирает оптимальный ответ (9.21) на эти предсказания, то совместное частотное распределение стратегий игроков сходится к множеству $\mathcal{C}$ коррелированных равновесий игры.

Теорема 9.6. *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 (9.22), сходится к множеству $\mathcal{C}$ коррелированных равновесий.*

Доказательство. Для доказательства теоремы, надо показать, что

$$\operatorname{dist}(\bar{p}_T, \mathcal{C}) \to 0$$

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 (9.22), 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

$$\bar{p}^*(a) = \sum_{\bar{a}: a_i = a} \bar{p}^*(\bar{a}) > 0,$$

где $\bar{a} = (a_1, \dots, a_k)$, $a_s \in A_s$, $s = 1, \dots, k$. [3]

Пишем $f = f^i$, и определим два подмножества (зависящие от i и a) $B, \tilde{B} \subseteq \mathcal{P}(A_{-i})$:

$$B = \{\bar{q}_{-i} : \bar{f}(a, \bar{q}_{-i}) = \max_{a' \in A_i} \bar{f}(a', \bar{q}_{-i})\}$$

– 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

$$\tilde{B} = \left\{ \bar{q}_{-i} : \exists t \left(\bar{q}_{-i} = \bar{q}^t_{-i} \right) \& \bar{f}(a, \bar{q}_{-i}) = \max_{a' \in A_i} \bar{f}(a', \bar{q}_{-i}) \right\}$$

[3] Если $\bar{p}^*(a) = 0$, 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.

– множество всех смешанных стратегий, выбранных оппонентами игрока i на тех шагах $t = 1, 2, \ldots$ 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 $\bar{p}^*$:

$$\bar{p}^*(a_{-i}|a_i = a) = \bar{p}^*((a, \bar{a}_{-i})|a_i = a) = \frac{\bar{p}^*(a, \bar{a}_{-i})}{\bar{p}^*(a_i = a)}. \qquad (9.25)$$

По следствию 7.3 распределение вероятностей p на множестве $\prod_{k=1}^{K}\{1, \ldots, N_k\}$ последовательностей стратегий $\bar{a} = (a_1, \ldots, a_K)$ является коррелированным равновесием тогда и только тогда, когда для каждого игрока $i \in \{1, \ldots, K\}$ и любых стратегий $a, a' \in \{1, \ldots, N_i\}$ выполнено

$$f^i(a, \bar{p}(\cdot|a_i = a)) = \max_{a' \in A_i} f^i(a', \bar{p}(\cdot|a_i = a)).$$

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 $a \in A_i$.

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 :

$$\bar{p}_T(a_{-i}|a_i = a) = \frac{\bar{p}_T(a, \bar{a}_{-i})}{\bar{p}_T(a_i = a)}. \tag{9.26}$$

Согласно (9.25) $\bar{p}_{T_j}(a_{-i}|a_i = a) \to p^*(a_{-i}|a_i = a)$ 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 $\bar{p}^t_{-i} \in \tilde{B}$. 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

$$\{\bar{a}^t_{-i} : \bar{p}^t_{-i} \in \tilde{B}, \ 1 \leqslant t \leqslant T\}.$$

Поэтому по (9.23) получаем:

$$\bar{p}_T(a, \bar{a}_{-i}) = \bar{p}_T(\bar{a}) = \frac{1}{T}|\{t : 1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}, \ \bar{a}^t_{-i} = \bar{a}_{-i}\}|.$$

По определению $\bar{p}_T(a_i = a) = \frac{N_T(a)}{T}$. Отсюда получаем выражения для условного частотного распределения, образованного последовательностью $\bar{a}^t$, где $a^t_i = a$, $t = 1, \ldots, \ T$:

$$\bar{p}_T(\cdot|a_i = a) = \frac{1}{N_T(a)} \sum_{1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}} \bar{\delta}[\bar{a}^t_{-i}] =$$

$$= \left(\frac{T}{N_T(a)}\right) \frac{1}{T} \sum_{1 \leqslant t \leqslant T, \ \bar{p}^t_{-i} \in \tilde{B}} (\bar{\delta}[\bar{a}^t_{-i}] - \bar{p}^t_{-i}) + \tag{9.27}$$

$$+ \sum_{1 \leqslant t \leqslant T, \ \bar{p}_{-i} \in \tilde{B}} \left(\frac{N_T(\bar{p}_{-i})}{N_T(a)}\right) \bar{p}_{-i}. \tag{9.28}$$

Пусть $\epsilon > 0$ – произвольное. Мы будем считать, что предсказания выбираются из конечного множества $\mathcal{P}_\epsilon = \{\bar{s}_1, \ \ldots, \ \bar{s}_{N_\epsilon}\}$ и удовлетворяют условию (9.19). Такие предсказания были названы ϵ-калибруемыми FV-предсказаниями. В этом случае множество $\tilde{B}$

конечно. Поэтому из условия (9.19) следует, что

$$\limsup_{T\to\infty}\left\|\frac{1}{T}\sum_{1\leqslant t\leqslant T,\ \bar{p}^t_{-i}\in\tilde{B}}(\bar{\delta}[\bar{a}^t_{-i}]-\bar{p}^t_{-i})\right\|\leqslant\epsilon. \tag{9.29}$$

Более того, по теореме 9.5 можно постепенно уменьшать число ϵ и строить предсказания таким образом, что среднее (9.29) будет стремиться к нулю при $T\to\infty$.

Таким образом, сумма (9.27) стремится к нулю при $T\to\infty$.

Так как $\bar{p}^*(a_i=a)>0$ 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

$$N_T(a)=\sum_{\bar{p}_{-i}\in\tilde{B}_T}N_T(\bar{p}_{-i}).$$

Таким образом, по (9.28) 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

$$\operatorname{dist}(\bar{p}_{T_j}(\cdot|a_i=a),B)\to 0$$

при $T_j\to\infty$.

Из сходимости $\bar{p}_{T_j}\to\bar{p}^*$ следует, что для произвольного вектора $\bar{a}_{-i}$ будет $\bar{p}_{T_j}(\bar{a}_{-i}|a_i=a)\to\bar{p}^*(a_{-i}|a_i=a)$ 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 $a\in A_i$, и, тем самым, вероятностное распределение $\bar{p}^*$ является коррелированным равновесием. Отсюда следует утверждение теоремы. $\triangle$

Литература

[1] Беккенбах Э., Беллман Р. Неравенства. – М.: Мир, 1965. – 276 с.

[2] Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). – М.: Наука, 1974 – 416 с.

[3] Ширяев А.Н. Вероятность. – М.: МЦНМО, 2007. – 968 с.

[4] Шикин Е.В., Шикина Г.Е. Исследование операций: учебное пособие. – М.: ТК Велби, изд. Проспект, 2006. – 280 с.

[5] Alon N., Ben-David S., Cesa-Bianchi N., Haussler D. // Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM V. 1997. 44(4). P. 615-631.

[6] Aronszajn N. Theory of reproducing kernels // Transactions of the American Mathematical Society. 1950. V. 68. P. 337Ц404.

[7] Blackwell D. An analog of the minimax theorem for vector payoffs // Pacific Journal of Mathematics. 1956. V. 6. P. 1–8.

[8] A. Chernov, F. Zhdanov. Prediction with expert advice under discounted loss. Technical report, arXiv:1005.1918v1 [cs.LG], 2010.

[9] Cover T., Ordentlich E. Universal portfolio with side information // IEEE Transaction on Information Theory – 1996. – V. 42. – P. 348–363.

[10] Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines. – Cambridge UK: Cambridge University Press, 2000.

[11] Dawid A.P. Calibration-based empirical probability [with discussion] // Ann. Statist. – 1985. – V. 13. – P. 1251–1285.

[12] Foster D.P., Vohra R. Asymptotic calibration // Biometrika. – 1998. – V. 85. – P. 379–390.

[13] Freund Y., Schapire R.E. 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.

[14] 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.

[15] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader // Journal of Machine Learning Research, 6:639–660, 2005.

[16] Kakade, S.M., Foster, D.P. Deterministic calibration and Nash equilibrium // Lecture Notes in Computer Science – Berlin: Springer, 2004. – V. 3120. – P. 33–48.

[17] Kakade S., Tewari A. Learning Theory Lectures 1-17. CMSC 35900 (Spring 2008):
http://ttic.uchicago.edu/ tewari/lectures/lecture11.pdf

[18] A. Kalai and 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.

[19] Kimeldorf G. S. and Wahba G. Some results on Tchebycheffian spline functions // J. Math. Anal. Appl. – 1971 –V. 33 – 82-Ц95.

[20] Littlestone N., Warmuth M. The weighted majority algorithm // Information and Computation – 1994 – V. 108 – P. 212–261.

[21] Lugosi G., Cesa-Bianchi N. Prediction, Learning and Games. – New York: Cambridge University Press, 2006.

[22] Mannor S., Stoltz G. A Geometric Proof of Calibration // arXiv:0912.3604v2. 2009.

[23] McDiarmid C. On the method of bounded differences. London Mathematical Society Lecture Notes Series. Surveys in Combinatorics. Cambridge University Press. V. 141. pp. 148Ц188. 1989.

[24] Shafer G., Vovk V. Probability and Finance. It's Only a Game! – New York: Wiley. 2001.

[25] Shawe-Taylor J., Cristianini N. Margin distribution bounds on generalization // In Proceedings of the European Conference on Computational Learning Theory, EuroCOLT'99. P.263–273. 1999.

[26] Shawe-Taylor J., Cristianini N. Kernel Methods for Pattern Analysis. – Cambridge UK: Cambridge University Press, 2004.

[27] Scholkopf B. and Smola A. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

[28] Vapnik V.N. Statistical Learning Theory. – New York: Wiley, 1998.

[29] Vovk V. 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.

[30] Vovk V. A game of prediction with expert advice // Journal of Computer and System Sciences – 1998 – V. 56. – No. 2.
P. 153–173.

[31] Vovk V., Watkins C. Universal portfolio selection // Proceedings of the 11th Annual Conference on Computational Learning Theory – New York: ACM Press, 1998. – P. 12–23.

[32] Vovk V. Competitive on-line statistics // International Statistical Review – 2001 – V. 69. – P. 213–248.

[33] Vovk V, Gammerman A., Shafer G. Algorithmic Learning in a Random World. Springer, New York, 2005.

[34] Vovk V., Shafer G. Good randomized sequential probability forecasting is always possible // J. Royal Stat. Soc. B. – 2005 – V. 67 – P. 747–763.

[35] Vovk V., Takemura A., Shafer G. 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.

[36] Vovk V. On-line regression competitive with reproducing kernel Hilbert spaces (extended abstract) // Lecture Notes in Computer Science – Berlin: Springer, 2006. – V. 3959. – P. 452–463.

[37] Vovk V. Predictions as statements and decisions // Theoretical Computer Science – 2008. – V. 405. – No. 3. – P. 285–296.