Кафедра проблем передачи информации и анализа данных (ИППИ РАН). Задачи для второкурсников МФТИ

Михаил Беляев

Введение. Ниже даны некоторые задачи, который в том или ином виде часто возникают в анализе данных. Непосредственно задаче предшествует введение, которое в крайне сжатой форме описывает зачем эту задачу необходимо решать. Предполагается, что на собеседовании это введение будет воспроизведено в достаточном для понимания комиссией варианте и затем описан способ решения поставленной задачи.

Книга Elements of statistical learning может быть скачана на официальном сайте R. Tibshirani, одного из авторов книги.

Вопросы по задачам, замеченные неточности/опечатки, отзывы по уровню сложности задач и их соответствию институтской программе обучения просьба отправлять на email belyaevmichel@gmail.com.

Задача 1. Задача классификации заключается в построении по выборке данных

$$\left\{x_i \in R^d, c_i \in \{1, \dots, K\}\right\}_{i=1}^n$$

решающего правила, которое относит каждую точку $x \in R^d$ к одному из K классов. Выборка данных состоит из пар точка $x_i \in R^d$, метка c_i принадлежности к одному из классов. Линейный дискриминантный анализ — это один из простейших методов классификации. В основе метода лежит предположение, что точки каждого класса порождены многомерным нормальным распределением, причем ковариационная матрица одинакова для всех классов. Для классификации используется байесовский подход, который в рамках таких предположений приводит к линейному решающему правилу для каждой пары классов k и l. Делая выбор между классами k и l точку x относят к классу k, если

$$x^T \Sigma^{-1}(\mu_k - \mu_l) > c_{k,l},$$

где Σ - это выборочная оценка ковариационной матрица, единая для всех классов, μ_k - выборочные оценки среднего точек класса k, для $k = 1, \dots K$, а $c_{k,l}$ некоторая константа. Совокупность таких решающих правил для всех пар классов формирует итоговый классификатор.

Необходимо показать, что в рамках модели линейного дискриминантного анализа и предположения $K < d+1$ существует такая матрица U, что только первые $K-1$ компонент вектора Ux оказывает влияние на выбор класса с помощью указанных выше решающих правил, а остальные $d-K+1$ избыточны и могут быть отброшены. Для частного случая бинарной классификации ($K = 2$) необходимо дать геометрическую интерпретацию этого наблюдения.

Материалы: Elements of statistical learning, раздел 4.3.

Задача 3. Задача регрессии заключается в построении по выборке данных

$$\left\{x_i \in R^d, y_i \in R^1\right\}_{i=1}^n$$

функции $\hat{f} : R^d \to R^1$, которая в некотором смысле хорошо приближает данные обучающей выборки (одна из возможных формализаций "качества" приближения: мы предполагаем, что $y_i = f_0(x_i) + \xi_i$, где $f_0(x_i)$ неизвестная гладкая функция, ξ_i — гауссовский шум, тогда мера качества приближения — близость $\hat{f}$ к f_0, например $\|\hat{f} - f_0\|$). В случае $d = 1$ одним из классических методов решения задачи являются сглаживающие сплайны

$$\hat{f} = argmin_{f \in W_2^2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left(\frac{d^2 f}{dx^2} \right)^2,$$

где параметр регуляризации λ позволяет балансировать между слишком сильным приближением зашумленных значений из обучающей выборки (первое слагаемое) и излишне упрощенным (гладким) поведением модели $\hat{f}$ (второе слагаемое).

Подобные задачи часто сводятся к поиску коэффициентов разложения α по некоторому конечномерному базису $\{\psi_j\}_{j=1}^{p}$с помощью решения задачи квадратичной оптимизации

$$\hat{f} = argmin_{\alpha} \sum_{i=1}^{n} (y_i - \alpha^{\mathrm{T}} \Psi_i)^2 + \lambda \alpha^{\mathrm{T}} \Omega \alpha,$$

где

- матрица Ψ — это $n \times p$ матрица значений базисных функций в точках выборки, $\Psi|_{i,j} = \psi_j(x_i)$,
- p — число этих базисных функций,
- Ψ_i — одна строка матрицы Ψ,
- Ω — симметричная положительно определенная $p \times p$ матрица, состоящая из скалярных произведений вторых производных базисных функций.

Задача 3a. С вычислительной точки, коэфициенты α могут быть найдены явным образом по формуле

$$\alpha_\lambda = \left(\Psi^{\mathrm{T}} \Psi + \lambda \Omega \right)^{-1} \Psi^{\mathrm{T}} y, \tag{1}$$

где y — вектор длины n, состоящий из y_i.

Известно, что обращение $p \times p$ матрицы требует $O(p^3)$ арифметических скалярных операций. При построении модели коэффициенты оцениваются для большого числа разных значений параметров регуляризации λ (происходит некоторый перебор по λ), что приводит к необходимости каждый раз пересчитывать обратную матрицу.

Необходимо доказать корректность формулы (1) и предложить алгоритм вычисления коэффициентов разложения α_λ для набора значений λ, который на этапе инициализации требует $O(np^2 + p^3)$ операций, а затем для каждого значения параметра регуляризации λ всего $O(p^2)$ операций.

Задача 3b. Для выбора параметра регуляризации λ часто используется так называемая ошибка скользящего контроля:

$$Q_{loo} = \sum_{i=1}^{n} (y_i - \alpha_{-i}^{\mathrm{T}} \Psi_i)^2,$$

где параметры α_{-i} оцениваются по выборке без наблюдения $\{x_i, y_i\}$ (то есть из выборки удаляется i-е наблюдение, коэффициенты α_{-i} оцениваются по уменьшенной выборке, состоящей из $n-1$ точек, с использованием (1), а потом на основе полученных коэффициентов α_{-i} оценивается ошибка в i-м наблюдении; такая процедура повторяется для всех точек выборки). Это позволяет честно оценивать ошибку модели, избегая эффекта переобучения.

Необходимо

1. Доказать, что Q_{loo} может быть подсчитана явно без оценки n векторов α_{-i} (то есть без n вычислений по формуле (1)) по следующей формуле

$$Q_{loo} = \sum_{i=1}^{n} \frac{(y_i - \alpha^{\mathrm{T}}\Psi_i)^2}{(1-l_i)^2}, \tag{2}$$

   где

   (a) коэффициенты разложения α оцениваются по формуле (1) (то есть 1 раз по всем n точкам);

   (b) $l_i = \Psi_i^T \left(\Psi^{\mathrm{T}}\Psi + \lambda\Omega\right)^{-1}\Psi_i$.

2. Выписать матричный аналог формулы (2) и на ее основе построить эффективный алгоритм подсчета Q_{loo} для набора значений λ, который на этапе инициализации требует $O(np^2 + p^3)$ операций, а затем для каждого значения параметра регуляризации λ всего $O(p^2)$ операций.

Указание к задачам 3a и 3b: применить разложение (3) к матрицам $\Psi^{\mathrm{T}}\Psi$ и Ω (которые в данном случае являются симметричными положительными определенными матрицами). Вычисление разложения (3) для $p \times p$ матриц требует $O(p^3)$ операций.

Материалы: Elements of statistical learning, разделы 5.1 - 5.3.

Для симметричных положительно определенных A и B существует такая невырожденная матрица X, что справедливо

$$\mathrm{X}^{\mathrm{T}}\mathrm{A}\mathrm{X} = \mathrm{D}_{\mathrm{A}}, \qquad \mathrm{X}^{\mathrm{T}}\mathrm{B}\mathrm{X} = \mathrm{D}_{\mathrm{B}}; \tag{3}$$

где D_{A} и D_{B} диагональные. Более того, $\mathrm{A}x_i = \lambda_i \mathrm{B}x_i$, $\lambda_i = d_i^a / d_i^b$ (что и обуславливает название “обобщенные собственные вектора”).