Обучение с учителем. Классификация. Дискриминантный анализ.

Е. Ларин, Ф. Ежов, И. Кононыхин

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Обучение с учителем

Выборка из генеральной случайной величины

- Для задачи регрессии: $\mathbf{X} \in \mathbb{R}^{n \times p}, \quad \mathbf{y} \in \mathbb{R}^n$
- Для задачи классификации: $\mathbf{X} \in \mathbb{R}^{n \times p}, \quad \mathbf{y} \in \mathbb{A}^n$

Обучение с учителем: формальная постановка

- *Вход*: $\mathbf{X}$ — выборка $\boldsymbol{\xi}$, $\mathbf{y}$ — выборка η. Предполагаем, что существует неизвестное отображение $y^* : \boldsymbol{\xi} \to \eta$ (гипотеза непрерывности или компактности)
- *Задача*: По $\mathbf{X}$ и $\mathbf{y}$ найти такое отображение $\hat{y}^* : \boldsymbol{\xi} \to \eta$, которое приблизит отображение y^*.
- *Оценка*: Функция потерь $\mathfrak{L}(y^*(x), \hat{y}^*(x))$. Здесь x — реализация $\boldsymbol{\xi}$

Классификация

$$\mathbf{X} \in \mathbb{R}^{n \times p}, \quad \mathbf{y} \in \mathbb{A}^n \tag{1}$$

Гипотеза компактности

«Близкие» объекты, как правило, принадлежат одному классу

Понятие близости может быть формализовано, например, так:

$$\rho(\mathbf{x_1}, \mathbf{x_2}) = \left(\sum_{i=1}^{p} w_i |x_1^i - x_2^i|^k \right)^{\frac{1}{k}}$$

Классификация: генеральная постановка

Дано:

- $\boldsymbol{\xi} \in \mathbb{R}^p$ — вектор признаков
- $\eta \in \mathbb{A}$ — классовая принадлежность

Предположение об их зависимсти можно записать в виде 2.

$$\eta = \Phi(\boldsymbol{\xi}, \varepsilon) \tag{2}$$

Обычно на ε накладываются условия

$$E\varepsilon = 0, \quad D\varepsilon = \sigma^2, \boldsymbol{\xi} \perp \varepsilon$$

Задача: найти Φ

Классификация: выборочная постановка

Дано:

- $\mathbf{X} \in \mathbb{R}^{n \times p}$ — матрица признаков
- $\mathbf{y} \in \mathbb{A}^{n}$ — вектор классовой принадлежности

Предположение имеет вид 3.

$$y_i = \Phi(\mathbf{x}_i, \varepsilon_i), \quad i = 1, \dots, n \tag{3}$$

Задача: найти Φ

Классификация: оценка качества

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

На основе этой матрицы есть большое количество разных метрик: *accuracy*, *recall*, *precision*, F_β, *ROC–AUC*

Классификация: типы классов

- По количеству классов:
  - бинарная классификация
  - многоклассовая классификация
- По пересечению классов
  - пересекающиеся
  - непересекающаяся
  - нечёткие

Классификация: этапы обучения модели

- Выбор модели (класс рассматриваемых Ф из 3)
- Выбор метрики
- Выбор метода обучения (способ подбора параметров для минимизации метрики на обучающем множестве)
- Выбор метода проверки (способ оценки качества модели)

Классификация: задача оптимизации

- $\hat{\beta}$ — параметры модели
- $\mathbf{\Phi}(\mathbf{x}, \beta)$ — функционал классификации
- $\mathfrak{L}(\mathbf{\Phi}(\mathbf{x}, \beta), \mathbf{y})$ — функция потерь (метрика)

$$\hat{\beta} = \arg\min_{\beta} \mathfrak{L}(\mathbf{\Phi}(\mathbf{x}, \beta), \mathbf{y})$$

Классификация: общий подход к решению

Как построить функционал Φ?
Общий подход — построить набор f_i, $i = 1, \ldots, K$. Каждая функция $f_i(\mathbf{x})$ показывает меру принадлежности $\mathbf{x}$ классу i.
Таким образом,

$$\Phi(\mathbf{x}) = \arg\max_i (f_i(\mathbf{x})). \tag{4}$$

Дискриминантный анализ

Примем за функции f_i из 4 оценку вероятности принадлежности к i-му классу.

$$\Phi(\mathbf{x}) = \arg\max_i (P(C_i|\mathbf{x})).$$

C_i — класс, состоящий из одного события: $\mathbf{x}$ принадлежит i-му классу.

Дискриминантный анализ

Если известны априорные вероятности получения i-го класса (π_i), применим формулу Байеса

$$P(C_i|\mathbf{x}) = \frac{\pi_i P(\mathbf{x}|C_i)}{\sum_{j=1}^{K} \pi_j P(\mathbf{x}|C_j)}.$$

Отбросим знаменатель

$$f_i = P(C_i|\mathbf{x}) = \pi_i P(\mathbf{x}|C_i).$$

LDA

Предположение:

$$\mathtt{P}(\boldsymbol{\xi}|\eta = A_i) = \mathtt{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$$

Классифицирующая функция:

$$f_i(\mathbf{x}) = \frac{\pi_i}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathtt{T}}\right)$$

После упрощения:

$$h_i(\mathbf{x}) = -0.5\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i^{\mathtt{T}} + \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\mathbf{x} + \log\pi_i$$

QDA

Предположение:

$$\mathtt{P}(\boldsymbol{\xi}|\eta = A_i) = \mathtt{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Классифицирующая функция:

$$f_i(\mathbf{x}) = \frac{\pi_i}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathtt{T}}\right)$$

После упрощения:

$$g_i(\mathbf{x}) = -0.5(\mathbf{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathtt{T}} - 0.5\log|\boldsymbol{\Sigma}_i| + \log\pi_i$$

Лог. регрессия

Зададим модель логистической регрессии следующим образом:

$$\log \frac{P(\eta = G_i|\boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K|\boldsymbol{\xi} = \mathbf{x})} = \beta_{i0} + \boldsymbol{\beta}_i^T \mathbf{x}, i = 1, \cdots, K-1.$$

Перейдем от логитов к вероятностям:

$$P(\eta = G_i|\boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\beta_{i0} + \boldsymbol{\beta}_i^T \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}}}, i = 1, \cdots, K-1,$$

$$P(\eta = G_K|\boldsymbol{\xi} = \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \boldsymbol{\beta}_k^T \mathbf{x}}}.$$

Лог. регрессия: метод максимального правдоподобия

Для оценки параметров воспользуемся методом максимального правдоподобия:

$$l(\theta) = \sum_{i=1}^{N} \log P(\eta = G_k | \boldsymbol{\xi} = \mathbf{x}; \theta),$$

$$\theta = (\beta_{10}, \beta_1^T, \ldots, \beta_{(K-1)0}, \beta_{K-1}^T).$$

Iteratively reweighted least squares (IRLS).

Функция потерь: $\mathfrak{L}(M_i(\beta)) = \log(1 + e^{-y_i \beta^T x_i})$

SVM

Выборка: $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$.
Задача построить классифицирующие правило.

Предположим, что данные - разделимы гиперплоскостью,

$$\mathbf{x}^T \beta - \beta_0 = 0; \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R},$$

$$g(x) = \mathbf{x}^T \beta - \beta_0,$$

$$h(x) = sign(g(x)).$$

SVM

Критерий оптимальности: максимальное расстояние между двумя гиперплоскостями, параллельных данной и симметрично расположенных относительно нее.

Эта пара гиперплоскостей может быть описана парой уравнений:

$$\mathbf{x}^T\beta - \beta_0 = -1,$$

$$\mathbf{x}^T\beta - \beta_0 = 1.$$

Растояние между ними: $\dfrac{2}{||\beta||}$.

SVM

Принадлежность точек обучающей выборки полупространства описывается

$$(\mathbf{x}^T\beta - \beta_0)y_i \geqslant 1$$

Задача сводится к задаче квадратичного программирования с линейными ограничениями:

$$\begin{cases} \frac{1}{2}||\beta||_2^2 \to \min\limits_{\beta,\beta_0} \\ y_i\left(x_i^{\mathsf{T}}\beta - \beta_0\right) \geqslant 1 \end{cases}$$