# Обучение с учителем. Классификация. Дискриминантный анализ.

Е. Ларин, Ф. Ежов, И. Кононыхин

## 1 Обучение с учителем

Рассмотрим задачу обучения с учителем, частным случаем которой являются задачи классификации и регрессии.

Алгоритм в общем виде имеет вид:

- *Вход*: $\boldsymbol{X}$ — выборка $\boldsymbol{\xi}$, случайной величины признаков, $\boldsymbol{y}$ — выборка $\eta$, случайной величны «ответов» (принадлежность к классу для классификации, либо значение функции для регрессии). Предполагаем, что существует неизвестное отображение $y^* : \boldsymbol{\xi} \to \eta$ (гипотеза непрерывности или компактности)
- *Задача*: По $\boldsymbol{X}$ и $\boldsymbol{y}$ 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 $\boldsymbol{\xi}$

## 2 Классификация

Перейдём к задаче классификации. Как и в задаче регрессии, данные должны происходить из некоторой генеральной совокупности.

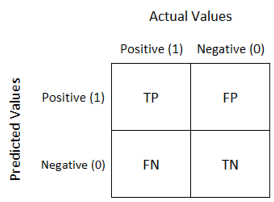
Будем рассматривать выборку признаков и ответов 1

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

Отметим, что множество $\mathbb{A}^n$ не является непрерывным. Размерность этого множества $k \times n$, где $k$ — количество возможных классов.

Для обоснования применения методов классификации используется **гипотеза компактности**:

«Близкие» объекты, как правило, принадлежат одному классу.




## 2.1 Классификация: вероятностная постановка

Поставим задачу классификации в терминах генеральных случайный величин.

*Дано:*

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

Предположение об их зависимости можно записать в виде 2.

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

*Задача:* найти $\Phi$

При переходе к выборкам, случайная величина признаков $\boldsymbol{\xi}$ заменяется на матрицу наблюдений $\boldsymbol{X}$, а случайная величина ответов $\eta$ — на вектор классовой принадлежности $\boldsymbol{y}$.

Предположение принимает вид 3.

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

## 2.2 Классификация: оценка качества

На основе матрицы ошибок 2.2 есть большое количество разных метрик. Приведём некоторые из них:

- $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$,
- $recall = \frac{TP}{TP+FP}$,
- $precision = \frac{TP+TN}{TP+FN}$,
- $F_\beta = (1-\beta^2)\frac{precision \times recall}{(\beta^2 \times precision) + recall}$,
- *ROC–AUC*

### 2.3 Классификация: этапы обучения модели

- Выбор модели (класс рассматриваемых $\Phi$ из 3). Здесь будут рассмотрены модели LDA и QDA.
- Выбор функции потерь. Чаще всего это $\sum_{i=1}^{n}(y_i \neq \widehat{y}_i)$.
- Выбор метода обучения. Выбор способа подбора параметров для минимизации функции потерь на обучающем множестве.
- Выбор метода проверки. Выбор оценки качества модели, например, с помощью метрик из раздела 2.2.

### 2.4 Классификация: общий подход к решению

Как построить функционал $\Phi$?

Общий подход — построить набор классифицирующих функций $f_i$, $i = 1, \ldots, K$. Каждая функция $f_i(\boldsymbol{x})$ показывает меру принадлежности $\boldsymbol{x}$ классу $i$.

Таким образом, решение о принадлежности классу принимается при обнаружении классифицирующей функции с наибольшим значением:

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

## 3 Дискриминантный анализ

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

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

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

Если известны априорные вероятности получения $i$-го класса ($\pi_i$), применим формулу Байеса

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

Отбросим знаменатель

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

### 3.1 LDA

Предположим, что искомые классы имеют многомерное нормальное распределение с равными дисперсиями.

Запишем это в виде формулы:

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

Построим классифицирующие функции:

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

Немного упростим (подробнее было изложено в одном из предыдущих курсов):

$$h_i(\boldsymbol{x}) = -0.5\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \log \pi_i \tag{5}$$

Функции 5 применяются при классификации данным методом.

### 3.2 QDA

Предположим, что искомые классы имеют многомерное нормальное распределение с различными дисперсиями.

Запишем это в виде формулы:

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

Построим классифицирующие функции:

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

Немного упростим (подробнее было изложено в одном из предыдущих курсов):

$$g_i(\boldsymbol{x}) = -0.5(\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} - 0.5 \log |\boldsymbol{\Sigma}_i| + \log \pi_i \tag{6}$$

Функции 6 применяются при классификации данным методом.

## 4 Классификация и регрессия

### 4.1 Регрессия

Обучающая выборка: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$.

1. Модель регрессии:
$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}(\boldsymbol{x}, \beta) = \langle \boldsymbol{x}, \beta \rangle = \sum_{j=1}^{p} \beta_j x_j,\ \beta \in \mathbb{R}^p$$
2. Функция потерь:
$$\mathfrak{L}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = (\hat{\boldsymbol{y}} - \boldsymbol{y})^2$$
3. Метод обучения — метод наименьших квадратов:
$$Q(\beta) = \sum_{i=1}^{n} (\boldsymbol{\Phi}(\boldsymbol{x}_i, \beta) - \boldsymbol{y}_i)^2 \to \min_{\beta}$$

### 4.2 Классификация

Обучающая выборка: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\ y \in \{-1, 1\}$.

1. Модель классификации:
$$\hat{y} = \boldsymbol{\Phi}(\boldsymbol{x}, \beta) = sign\langle \boldsymbol{x}, \beta \rangle = sign \sum_{j=1}^{p} \beta_j x_j,\ \beta \in \mathbb{R}^p$$
2. Функция потерь:
$$\mathfrak{L}(\hat{y}, y) = [\hat{y} y < 0] = [\langle \boldsymbol{x}, \beta \rangle y < 0] \leq \hat{\mathfrak{L}}(\langle \boldsymbol{x}, \beta \rangle y)$$
3. Метод обучения — минимизация эмпирического риска:
$$Q(\beta) = \sum_{i=1}^{n} [\langle \boldsymbol{x}_i, \beta \rangle y_i < 0] \leq \sum_{i=1}^{n} \hat{\mathfrak{L}}(\langle \boldsymbol{x}_i, \beta \rangle y_i) \to \min_{\beta}$$

## 5 Отступы

$\boldsymbol{\Phi}(\boldsymbol{x}, \beta) = sign(g(\boldsymbol{x}, \beta))$ — разделяющий классификатор,
$g(\boldsymbol{x}, \beta)$ — разделяющая функция,
$g(\boldsymbol{x}, \beta) = 0$ — уравнение разделяющей поверхности.

$M_i(\beta) = g(\boldsymbol{x}_i, \beta) y_i$ — отступ объекта $\boldsymbol{x}_i$.
Если $M_i(\beta) < 0$, тогда алгоритм ошибается на $\boldsymbol{x}_i$.

## 6 Отступы

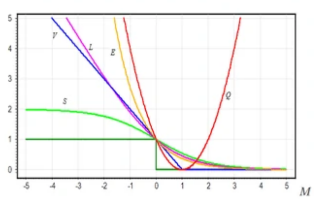



Рис. 7. Непрерывные аппроксимации пороговой функции потерь $[M < 0]$.

$Q(M) = (1 - M)^2$ — квадратичная;
$V(M) = (1 - M)_+$ — кусочно-линейная;
$S(M) = 2(1 + e^M)^{-1}$ — сигмоидная;
$L(M) = \log_2(1 + e^{-M})$ — логистическая;
$E(M) = e^{-M}$ — экспоненциальная.