

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

## Обучение с учителем: формальная постановка

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

# Классификация

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

## 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:*

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

Предположение об их зависимости можно записать в виде 2.

$$\eta = \Phi(\xi, \varepsilon) \quad (2)$$

Обычно на  $\varepsilon$  накладываются условия

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

*Задача:* найти  $\Phi$

# 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.

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

*Задача:* найти  $\Phi$

# 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*,  $F_\beta$ , *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  $\Phi$  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
- $\Phi(\mathbf{x}, \beta)$  — функционал классификации
- $\mathcal{L}(\Phi(\mathbf{x}, \beta), \mathbf{y})$  — функция потерь (метрика)

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