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

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

Обучение с учителем: формальная постановка

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

Классификация

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

<u>Гипотеза комп</u>актности

«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:

- ullet $oldsymbol{\xi} \in \mathbb{R}^p$ 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

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

Задача: найти Ф

Классификация: выборочная постановка

Дано:

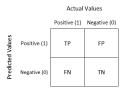
- ullet $\mathbf{X} \in \mathbb{R}^{n imes p}$ матрица признаков
- ullet $\mathbf{y} \in \mathbb{A}^n$ вектор классовой принадлежности

Предположение имеет вид 3.

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

Задача: найти Ф

Классификация: оценка качества



На основе этой матрицы есть большое количество разных метрик: accuracy, recall, precision, F_{β} , $ROC ext{-}AUC$

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

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

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