Лекция 3. Классификация

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

1. Байесовские методы
   - Оптимальный классификатор
   - Параметрическое восстановление плотности

Вероятностная постановка задачи

$\mathbb{X}$ - множество объектов, $\mathbb{Y}$ - множество классов.
$(\mathbb{X} \times \mathbb{Y})$ - вероятностное пространство с совместной плотностью $p(x, y) = P(y)p(x|y)$
$P_y := P(y)$ - **априорные вероятности** классов (prior)
$p_y(x) := p(x|y)$ - **функции правдоподобия** классов (likelihood)

Задачи:

1. По выборке $(X, Y) \in (\mathbb{X}, \mathbb{Y})$ построить оценки распределений $\hat{P}_y$ и $\hat{p_y}(x)$
2. По известным распределениям $p_y(x)$ и P_y построить алгоритм $a : \mathbb{X} \to \mathbb{Y}$ минимизирующий вероятность ошибочной классификации

Функционал среднего риска

Алгоритм $a(x)$ разбивает $\mathbb{X}$ на непересекающиеся области $A_y = \{x \in \mathbb{X} | a(x) = y\}$

Каждой паре $(y, s) \in (\mathbb{Y} \times \mathbb{Y})$ соответствует величина потери λ_{ys} при классификации объекта класса y к классу s, $\lambda_{yy} = 0$ и $\lambda_{ys} > 0$ при $y \neq s$

Функционал среднего риска:

$$R(a) = \sum_{y \in \mathbb{Y}} \sum_{s \in \mathbb{Y}} \lambda_{ys} P_y P(A_s|y)$$

где $P(A_s|y) = \int_{A_s} p_y(x)dx$ - вероятность отнесения к классу s объекта класса y.

Оптимальное байесовское решающее правило

Если известны априорные вероятности P_y и функции правдоподобия $p_y(x)$, то минимум среднего риска $R(a)$ достигается алгоритмом

$$a(x) = \arg\min_{y \in \mathbb{Y}} \sum_{y \in \mathbb{Y}} \lambda_{ys} P_y p_y(x)$$

Если предположить что потери от ошибочной классификации зависят только от истинного класса объекта, т.е. $\lambda_{ys} = \lambda_y$, то алгоритм называется **Байесовским решающим правило**:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \lambda_y P_y p_y(x)$$

Апостериорные вероятности

Вероятность $P(y|x)$ - называется **апостериорной вероятностью** (posterior).

Зная $p_y(x)$ и P_y, то по формуле Байеса:

$$P(y|x) = \frac{p(x,y)}{p(x)} = \frac{p_y(x)P_y}{\sum_{s\in\mathbb{Y}} p_s(x)P_s}$$

Величина ожидаемых потерь на объекте x:

$$R(x) = \sum_{y\in\mathbb{Y}} \lambda_y P(y|x)$$

Принцип максимума апостериорной вероятности

Оптимальный байесовский классификатор через апостериорные вероятности:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \lambda_y P(y|x)$$

Если классы равнозначны ($\lambda_y = \lambda_s \forall y, s \in \mathbb{Y}$), то байесовское решающее правило называют **принципом максимума апостериорной вероятности**.

В случае равновероятных (сбалансированных) классов ($P_y = \frac{1}{|\mathbb{Y}|}$), объект x просто относится к классу с наибольшим значением плотности $p_y(x)$.

Параметрический подход

Имеется выборка $X = (x_1, ..., x_\ell) \in \mathbb{X}$. Предполагается, что плотность, порождающая данные, известна **с точностью до параметра**, $p(x) = \phi(x; \theta)$. Подбор параметров θ приводится по выборке X с помощью **метода максимального правдоподобия**.

Нормальный дискриминантный анализ - случай байесовской классифицакии в предположении о нормальном распределениии всех классов, $p_y(x) \sim N(\mu_y, \sigma_y^2), y \in \mathbb{Y}$.