Лекция 3. Классификация

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

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, $p_y(x) \sim N(\mu_y, \sigma_y^2), y \in \mathbb{Y}$.

Квадратичный дискриминант

Байесовский нормальный классификатор

Линейный дискриминант Фишера

Бинарная классификация

Рассмотрим задачу бинарной классификации,
$X = (x_1, ..., x_\ell) \in \mathbb{X}$, $Y \in \mathbb{Y} = \{-1, 1\}$

Функцию $a(x, \theta) = \operatorname{sign} f(x, \theta)$ будем называть **дискриминантной функцией**.

Если $f(x, \theta) > 0$, то x относится к классу $+1$, $f(x, \theta) < 0$ то -1.

А множество точек $\{x | f(x, \theta) = 0\}$ - **разделяющая поверхность**.

Величина $M_i(\theta) = y_i f(x_i, \theta)$ - **отступ** (margin) классификатора $a(x, \theta) = \operatorname{sign} f(x, \theta)$ относительно объекта x_i.

Если $M_i(\theta) < 0$ то алгоритм a допускает ошибку на объекте x_i. **Чем больше** $M_i(\theta)$ **тем правильнее** и надежнее классификация.

Функция потерь

Требуется подобрать параметры θ при которых классификатор a допускает как можно меньше ошибок:

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[M_i(\theta) < 0] \to \min_{\theta}$$

Однако в таком виде Q - кусочно постоянная функция

Идея - мажорирование (оценка сверху) индикатора ошибки $[M_i(\theta) < 0]$ с помощью "удобной"функцией потерь $\mathcal{L}(M_i)$:

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[M_i(\theta) < 0] <= \widetilde{Q}(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathcal{L}(M_i(\theta))$$

Популярные функции потерь

1. $[M < 0]$ - индикатор ошибки
2. $(1 - M)^2$ - квадратичная
3. $(1 - M)_+$ - кусочно линейная
4. $\frac{2}{1+e^M}$ - сигмоидная
5. $\log_2(1 + e^{-M})$ - логистическая
6. e^{-M} - экспоненциальная

Функция потерь и совместное распределение

Пусть множество $(\mathbb{X} \times \mathbb{Y})$ - вероятностное пространосво. Имея выборку (X, Y) и предполагаемый вид совместной плотности $p(x, y; \theta)$, применим метод максимального правдоподобия

$$L(\theta) = \prod_{i=1}^{\ell} p(x_i, y_i; \theta) \to \max_{\theta}$$

$$\ln L = \sum_{i=1}^{\ell} \ln p(x_i, y_i; \theta) \to \max_{\theta}$$

$$-\ln p(x_i, y_i; \theta) = \mathcal{L}(y_i f(x_i, \theta))$$

По виду плотности $p(x, y; \theta)$ восстанавливается f и $\mathcal{L}$. И обратно, используя некоторые разделяющую поверхность и функцию потерь - предполагаем определенное распределение в данных.

Линейная модель

Случай $f(x,\omega) = \langle x, \omega \rangle$ - класс линейных моделей классификации.

$$a(x,\omega) = \operatorname{sign}\langle x, \omega \rangle$$

Разделяющая поверхность $\operatorname{sign}\langle x, \omega \rangle = 0$ является гиперплосткостью в $\mathbb{R}^n$. Причем объекты по одну сторону от гиперплосткости относятся к одному классу, по другую - к другому.

Метод обучения

Метод минимизации мажорированного эмперического риска

$$\widetilde{Q}(a, X) = \sum_{i=1}^{\ell} \mathcal{L}(\langle x_i, \omega \rangle y_i) \to \min_{\omega}$$

Необходимое условие минимума:

$$\frac{\partial Q}{\partial \omega} = \sum_{i=1}^{\ell} x_i y_i \mathcal{L}'(\langle x_i, \omega \rangle y_i) = 0$$

Логистическая регрессия

Логистическая регрессия - линеный алгоритм бинарной классификации.

При достаточно сильных свойствах обладает свойствами:

- оптимальный байесовский классификатор
- однозначно определена функция потерь
- возможность оценивать вероятности классов

Пусть $(\mathbb{X} \times \mathbb{Y}) = (\mathbb{R}^n \times \{-1, 1\})$ - вероятностное пространтсво с плотностью $p(x, y)$. Выборка (X, Y) получена из этого распределения.

Экспонентный класс распределений

Плотность $p(x), x \in \mathbb{R}^n$ называется экспонентной, если

$$p(x) = \exp(c(\delta)\langle\theta, x\rangle + b(\delta, \theta) + d(x, \delta))$$

где θ -параметр сдвига, δ - масштаба, b, c, d - произвольные числовые функции.

Принадлежат к классу экспонентных:

- Равномерное, Нормальное, Гамма
- Гипергеометрическое, Пуассоновское, Бернулли
- и другие

Обоснование линейной модели

$$a(x) = \mathrm{sign}(\lambda_+ P(+1|x) - \lambda_- P(-1|x)) = \mathrm{sign}(\frac{P(+1|x)}{P(-1|x)} - \frac{\lambda_-}{\lambda_+})$$

Если функции правдоподобия $p(x|y)$ принадлежат экспонентному классу, причем параметры d и δ одинаковы, а отличаются только параметры сдвига θ,

то:

1. байесовский классификатор является линейным: $a(x) = \mathrm{sign}\langle\omega, x\rangle$
2. апостериорная вероятность $p(y|x) = \sigma(\langle\omega, x\rangle y)$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоидная функция, $\sigma(-z) = 1 - \sigma(z)$

Модель оценки вероятностей

Построим модель которая оценивает не сами метки классов, а **вероятности** принадлежности к ним.

$$a(x, \omega) = P(+1|x) = \sigma(\langle w, x \rangle)$$

Другими словами, в каждой точке x величина $y \sim Bernoulli(\sigma(\langle w, x \rangle))$

Задача классификации решается путем выбора порога $t \in [0, 1]$. И тогда итоговый классификатор имеет вид:

$$b(x, t) = \text{sign}(a(x, \omega) - t)$$

Метод максимального правдоподобия

С учетом вероятностной постановки задачи, воспользуемся **методом максимального правдоподобия**:

$$L(\omega) = \prod_{i=1}^{\ell} p(y_i|x_i) = \prod_{i=1}^{\ell} \sigma(\langle w, x_i \rangle y_i) \to \max_{\omega}$$

$$\ln L(\omega) = \sum_{i=1}^{\ell} \ln \sigma(\langle w, x_i \rangle y_i) =$$

$$= -\sum_{i=1}^{\ell} \ln(1 + e^{-y_i \langle w, x \rangle}) \to \max_{\omega}$$

-совпадает с логистической функция потерь

$$Q(\omega) = \sum_{i=1}^{\ell} \ln(1 + e^{-M_i}) \to \min_{\omega}$$

Решение оптимизационной задачи

Имеем

$$Q(\omega) = \sum_{i=1}^{\ell} \ln(1 + e^{-\langle w, x \rangle}) \to \min_{\omega}$$

Аналитического решения нет, поэтому применяются градиентные методы

$$\nabla Q(\omega) = \sum_{i=1}^{\ell} x_i y_i \sigma(-\langle w, x_i \rangle)$$