Лекция 3. Байесовские методы классификации

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.

Минимум среднего риска

Теорема (о минимуме среднего риска)

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

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

достигается алгоритмом

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

Байесовское решающее правило

Теорема (об оптимальном классификаторе)

Если известны априорные вероятности P_y, функции правдоподобия $p_y(x)$ ***и ошибка неправильной классификации зависит только от истинного класса****, т.е. $\lambda_{ys} = \lambda_y, \forall s \neq y$, то минимум среднего риска $R(a)$ достигается алгоритмом*

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

Такой алгоритм называется ***байесовским решающим правилом*** *(оптимальный байесовский классификатор)*

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

Альтернативаня запись оптимального байесовскеого классификатора 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 $p_y(x)$.

Следующая задача оценки плотности

Получен оптимальный байесовский классификатор

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

Но в действительности плотности $P(y|x)$ неизвестны. Чтобы построить итоговый классификатор, ставится задача оценить плотность $\hat{P}(y|x)$ по эмпирическим данным $X \in \mathbb{X}$, $Y \in \mathbb{Y}$.

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

После замены в байесовском решающем правиле $P(y|x)$ на их оценки $\hat{P}(y|x)$, классификатор перестает быть оптимальным.

Восстановление плотности

Имеющаяся выборка $(X, Y) = ((x_i, y_i))_{i=1}^{\ell}$ сгенерирована некоторой плотностью $p(x, y)$.

Оценку совместной плотности можно строить отдельно для P_y и $p_y(x)$, ведь $p(x, y) = P_y p_y(x)$

Оценка P_y строится очень легко:

$$\hat{P}_y = \frac{\ell_y}{\ell},$$

где $\ell_y = |\{(x_i, y_i) \in (X, Y) : y_i = y\}|$

Наивный подход

Суть наивного подхода - предположение о независимости признаков между собой. Это позволяет упростить

$$P(x|y) = \prod_{i=1}^{n} P(x_i|y)$$

Для построения итоговой плотности $P(x|y)$ нужно оценить все индивидуальные распределения признаков $P(x_i), i = 1, \ldots, n$.

Наивный байесовский классификатор

Оценив априорные плотности и индивидуальные функции правдоподобия, получим **наивный байесовский классификатор**

$$a(x) = \arg\max_{y \in \mathbb{Y}} \ln \lambda_y \hat{P}(y) + \sum_{i=1}^{n} \ln \hat{P}(x_i|y)$$

Предположение о независимости признаков является очень сильным и на практике почти никогда не выполняется.

Одномерный непрерывный случай

Пусть $\mathbb{X} = \mathbb{R}$. Эмпирической оценкой плотности есть доля элементов выборки внутри окна шириной h

$$\hat{p}(x) = \frac{1}{2mh} \sum_{i=1}^{m} [|x - x_i| < h]$$

Результат есть кусочно-постоянная функция. Это приводит к появлению зон неуверенности оптимальном классификаторе. Идея состоит в применении ядра

Ядерная оценка плотности

Функция $K(z)$ называется ядром, если она:

- $K(z) = K(-z)$ - четная
- $\int K(z)dz = 1$
- $K(z) \geq 0$

Тогда **ядерная оценка плотности** имеет вид:

$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right)$$

Многомерный случай

Ядерная оценка плотности для многомерной величины $X \in \mathbb{R}^n$

$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{x - x_i}{h_j}\right)$$

Можно обобщить и на случай пространства, где задана функция расстояния между объектами $\rho(z_1, z_2)$

$$\hat{p}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^{m} K\left(\frac{\rho(x, x_i)}{h_j}\right)$$

Метод парзеновского окна

Применяя ядерную оценку плотности в байесовском решающем правиле, получим **метод парзеновского окна**

$$a(x) = \arg\max_{y \in \mathbb{Y}} \lambda_y \sum_{i=1}^{\ell} [y_i = y] K\left(\frac{\rho(x, x_i)}{h}\right)$$

Параметрический подход

Имеется выборка $X = (x_1, \dots, x_\ell) \in \mathbb{X}$. 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, $p_y(x) \sim N(\mu_y, \sigma_y^2), y \in \mathbb{Y}$.

Теорема о разделяющей поверхности

Теорема (о форме разделяющей поверхности)

Если классы имеют n-мерные нормальные плотности распределения

$$p_y(x) = N(x; \mu_y, \Sigma_y), y \in \mathbb{Y}$$

то оптимальный баейсовский классификатор задаёт квадратичную разделяющую поверхность. Она вырождается в линейную, если ковариационные матрица классов равны $\Sigma_y = \Sigma, y \in \mathbb{Y}$

Байесовский нормальный классификатор

Оценим параметры $\hat{\mu}_y$ и $\hat{\Sigma}_y$ n-мерной плотности по имеющимся данным для каждого класса $y \in \mathbb{Y}$.

$$\hat{\mu}_y = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \hat{\Sigma}_y = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

И воспользуемся идеей оптимального байесовского классификатора

$$a(x) = \arg\max_{y \in \mathbb{Y}} \lambda_y \hat{P}(y) N(x; \hat{\mu}_y, \hat{\Sigma}_y)$$

Такой классификатор называется **байесовский нормальный классификатор** или **подстановочный**

Линейный дискриминант Фишера

Предположив, что ковриационные матрицы классов равны $\Sigma_y = \Sigma, y \in \mathbb{Y}$ и применяя подстановочный алгоритм, получим метод линейного дискриминанта Фишера.

В таком случае разделяющая поверхность является линейной (в случае нескольких классов - кусочно линейная) и обладает определенными свойствами устойчивости.

Резюме

Байесовский подход к классификации:

- решает более сложную задачу оценивания плотности и только потом производит классификацию
- знание о распределении позволяет строить доверительные интервалы
- дает понимание вероятностной природы задачи и идеи для некоторых других методов