Лекция 5

Вероятностная классификация

Машинное обучение

Андрей Фильченков / Сергей Муравьев

02.10.2020

План лекции

- Байесовская классификация
- Непараметрическое восстановление плотности распределения
- Параметрическое восстановление плотности распределения
- Мягкая классификация и логистическая регрессия

- В презентации используются материалы курса «Машинное обучение» К.В. Воронцова
- Слайды доступны: **shorturl.at/hjyAX**
- Видео доступны: **shorturl.at/ltVZ3**

План лекции

- Байесовская классификация
- Непараметрическое восстановление плотности распределения
- Параметрическое восстановление плотности распределения
- Мягкая классификация и логистическая регрессия

Задача по COVID-2019

Допустим, сейчас в России COVID-2019 болеет 300 тыс. человек, то есть **0,5%** населения. Допустим, у нас есть некоторый тест, который дает верный диагноз в **99%** случаев. Безговидер сдал тест и получил положительный результат.

Какая вероятность того, что Безговидер на самом деле болеет COVID-2019?

Варианты ответа

Допустим, сейчас в России COVID-2019 болеет **0,5%** населения. Некоторый тест дает верный диагноз в **99%** случаев. Безговидер сдал тест и получил **положительный** результат. Какая вероятность того, что он на самом деле более COVID-2019?

A: $97{,}5\% \leq x \leq 100\%$

B: $95\% \leq x < 97{,}5\%$

C: $92\% \leq x < 95\%$

D: $81\% \leq x < 92\%$

E: $70\% \leq x < 81\%$

F: $55\% \leq x < 70\%$

G: $30\% \leq x < 55\%$

H: $x < 30\%$

Ответ

Допустим, сейчас в России COVID-2019 болеет **0,5%** населения. Некоторый тест дает верный диагноз в **99%** случаев. Безговидер сдал тест и получил **положительный** результат. Какая вероятность того, что он на самом деле более COVID-2019?

$$\Pr(d = 1|t = 1) =$$

$$= \frac{\Pr(t = 1|d = 1)\Pr(d = 1)}{\Pr(t = 1|d = 1)\Pr(d = 1) + \Pr(t = 1|d = 0)\Pr(d = 0)} =$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} = \mathbf{0,33}.$$

Вероятностная классификация

Вместо неизвестно целевой функции $y^*(x)$ будем думать о неизвестном распределении над $X \times Y$ с плотностью $p(x, y)$.

Простой или **независимой одинаково распределенной** (independent identically distributed, i.i.d.) называется выборка, содержащая независимые наблюдения из одного распределения.

Задача: найти алгоритм классификации, который минимизирует вероятность ошибки.

Средний риск

Средний риск a:

$$R(a) = \sum_{y \in Y} \lambda_y \int_{a(x) \neq y} p(x, y) dx$$

λ_y — **потеря от ошибки** для объектов класса y

Они являются частью постановки задачи и обычно определяются экспертно.

Основное уравнение

$$p(X,Y) = p(x)\,\mathrm{Pr}(y|x) = \mathrm{Pr}(y)\,p(x|y)$$

$\mathrm{Pr}(y)$ — **априорная** вероятность класса y.

$p(x|y)$ — **правдоподобие** (likelihood) класса y

$\mathrm{Pr}(y|x)$ — **апостериорная** вероятность класса y.

Две проблемы

Первая проблема: **восстановление плотности распределения**
Дано: $\mathcal{D}$.
Задача: найти эмпирические оценки $\widehat{\Pr}(y)$ и $\hat{p}(x|y)$, $y \in Y$.

Вторая проблема: **минимизация среднего риска**
Дано:
- априорные вероятности $\Pr(y)$,
- правдоподобие $p(x|y), y \in Y$.

Задача: найти классификатор a с минимальным $R(a)$.

Какая из проблем проще?

Оценка апостериорного максимума

Пусть $\Pr(y)$ и $p(x|y)$ известны для всех $y \in Y$.

$$p(x, y) = p(x)\Pr(y|x) = \Pr(y)\, p(x|y).$$

Основная идея: для нового объекта будем возвращать класс, к которому он принадлежит с наибольшей вероятностью.

Оценка апостериорного максимума (maximum a posteriori probability, MAP):

$$a(x) = \operatorname{argmax}_{y\in Y} \Pr(y|x) = \operatorname{argmax}_{y\in Y} \Pr(y)\, p(x|y).$$

Оптимальный байесовский классификатор

Теорема

Если известны $\Pr(y)$ и $p(x|y)$, то минимальный средний риск достигается байесовским классификатором a_{OB}

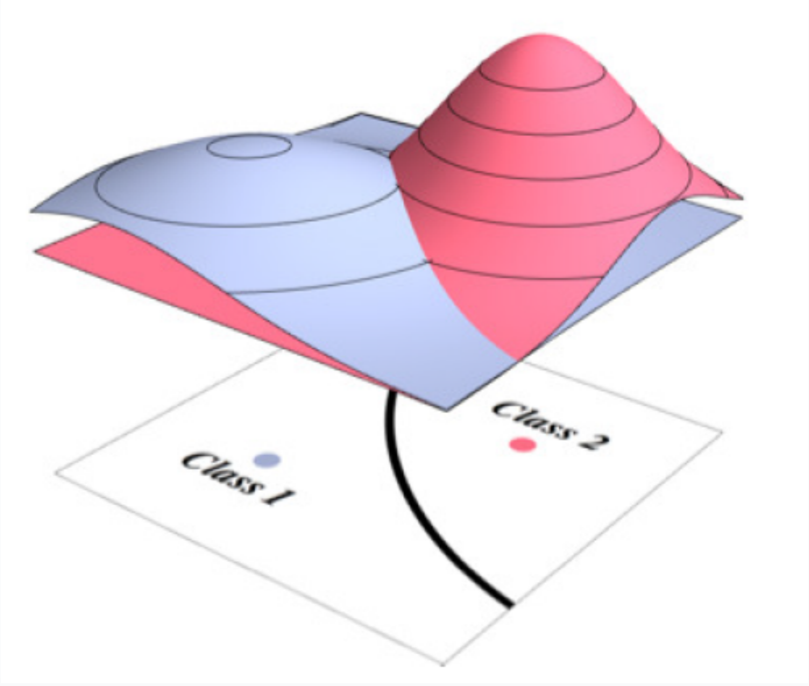
$$a_{OB}(x) = \mathrm{argmax}_{y \in Y} \lambda_y \Pr(y)\, p(x|y).$$

Классификатор $a_{OB}(x)$ называется **оптимальным байесовским классификатором,** а достигаемое им значение $R(a_{OB})$ — **байесовским риском.**

Разделяющая поверхность

Разделяющая поверхность для двух классов y_+ и y_- — это множество точек $x \in X$, на которых достигается равенство в байесовском разделяющем правиле:

$$\lambda_{y_+} \Pr(y_+)\, p(x|y_+) = \lambda_{y_-} \Pr(y_-)\, p(x|y_-).$$



План лекции

- Байесовская классификация
- Непараметрическое восстановление плотности распределения
- Параметрическое восстановление плотности распределения
- Мягкая классификация и логистическая регрессия

Две подзадачи

Необходимо оценить априорные и апостериорные оценки для каждого класса:

$$\widehat{\Pr}(y) = ?$$
$$\hat{p}(x|y) = ?$$

Первую подзадачу можно решить довольно легко:

$$\widehat{\Pr}(y) = \frac{|X_y|}{|\mathcal{D}|}, \qquad X_y = \{x_i, y_i \in \mathcal{D}, y_i = y\}.$$

Вторая подзадача, однако, намного сложнее.

Избавляемся от класса

Мы можем искать

$$\hat{p}(x|y) = ?$$

для каждого класса независимо.

Поэтому вместо $\hat{p}(x|y)$, я буду писать $\hat{p}(x)$, которое нужно восстановить над $\mathcal{D}_s = \left((x_{(1)}, s), \dots, (x_{(m)}, s)\right)$ для каждого $s \in Y$.

Одномерный случай

Если $\Pr([a,b])$ является мерой вероятности на $[a,b]$, то

$$p(x) = \lim_{h\to 0} \frac{1}{2h} \Pr([x-h, x+h]).$$

Эмпирическая оценка вероятности с окном шириной h

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

Окно Парзена — Розенблата

Эмпирическая оценка вероятности с окном h:

$$\widehat{p_h}(x) = \frac{1}{2hm} \sum_{i=1}^{m} \left[\frac{x - x_i}{h} < 1 \right].$$

Оценка Парзена — Розенблатта с окном шириной h:

$$\widehat{p_h}(x) = \frac{1}{hm} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right),$$

где $K(r)$ — некоторая ядерная функция.

$\widehat{p_h}(x)$ сходится к $p(x)$.

Обобщение на многомерный случай

1. Если объекты описываются n вещественными признаками $f_j: X \to \mathbb{R}, j = 1, \dots, n$,

$$\widehat{p_h}(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right).$$

2. Если X — метрическое пространство с мерой расстояния $\rho(x, x')$:

$$\widehat{p_h}(x) = \frac{1}{mV(h)} \sum_{i=1}^{m} K\left(\frac{\rho(x, x_i)}{h}\right),$$

где $V(h) = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ множитель нормализации.

Многомерное парзеновское окно

Оценим $\widehat{p_h}(x)$ окном Парзена – Розенблатта

$$\widehat{p_h}(x) = \frac{1}{mV(h)} \sum_{i=1}^{m} K\left(\frac{\rho(x, x_i)}{h}\right),$$

Окно Парзена:

$$a(x; \mathcal{D}, h) = \arg\max_{y \in Y} \lambda_y \Pr(y) \frac{1}{|\mathcal{D}_y|} \sum_{i: y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

$\Gamma_y(x) = \lambda_y \Pr(y) \frac{1}{|\mathcal{D}_y|} \sum_{i: y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right)$ – близость к классу.

Наивная гипотеза

Гипотеза (наивная): все признаки соответствуют независимым случайным величинам с плотностями вероятностей $p_j(\xi|y), y \in Y,\ j = 1, \dots, n$.

Тогда правдоподобие классов можно расписать следующим образом:

$$p(x|y) = p_1(\xi_1|y) \cdot \ldots \cdot p_n(\xi_n|y), \qquad x = (\xi_1, \dots, \xi_n), y \in Y.$$

Наивный байесовский классификатор

Правдоподобие классов:

$$p(x|y) = p_1(\xi_1|y) \cdot \ldots \cdot p_n(\xi_n|y), \qquad x = (\xi_1, \ldots, \xi_n), y \in Y.$$

Построим классификатор

$$a(x) = \operatorname{argmax}_{y \in Y} \left(\widehat{\Pr}(y) \cdot \prod_{j=1}^{n} \widehat{p_j}(\xi_j|y) \right).$$

Наивный байесовский классификатор (naïve Bayesian classifier):

$$a_{NB}(x) = \operatorname{argmax}_{y \in Y} \left(\ln \lambda_y \widehat{\Pr}(y) + \sum_{j=1}^{n} \ln \widehat{p_j}(\xi_j|y) \right).$$

План лекции

- Байесовская классификация
- Непараметрическое восстановление плотности распределения
- Параметрическое восстановление плотности распределения
- Мягкая классификация и логистическая регрессия

Задача о монетке

Вы нашли монету и подбросили ее 7 раз со следующим результатом:

О Р О О О О Р

Что выпадет в следующий раз:

1. С большей вероятностью выпадет Р
2. Р и О выпадут с одинаковой вероятностью

Задача о монетке

Вы нашли монету и подбросили ее 7 раз со следующим результатом:

О Р О О О О Р

Что выпадет в следующий раз:

1. С большей вероятностью выпадет Р
2. Р и О выпадут с одинаковой вероятностью
3. **С большей вероятностью выпадет О**

Параметры распределения

Как раньше мы выбирали алгоритм из параметрического семейства, так и сейчас будем выбирать распределение из параметрического семейства распределений

Будем искать $p(x, y) \in \{\varphi(x, y, \theta) | \theta \in \Theta\}$ точно так же, как раньше искали $a(x, \theta)$.

Параметрическая нотация

Плотность совместного распределения выборки:

$$p(\mathcal{D}) = p\Big((x_1, y_1), \dots, \big(x_{|\mathcal{D}|}, y_{|\mathcal{D}|}\big)\Big) = \prod_{(x,y)\in\mathcal{D}} p(x, y).$$

Правдоподобие:

$$\mathcal{L}(\theta, \mathcal{D}) = \prod_{(x,y)\in\mathcal{D}} \varphi(x, y, \theta).$$

MAP:

$$a_\theta(x) = \mathrm{argmax}_y \varphi(x, y, \theta).$$

Связь с эмпирическим риском

Возьмем логарифм и поменяем знак:

$$-\ln \mathcal{L}(\theta, \mathcal{D}) = -\sum_{(x,y)\in\mathcal{D}} \ln \varphi(x, y, \theta) \to \min_\theta .$$

Введем функцию потерь на объекте:

$$\mathcal{L}(a_\theta, x) = -|\mathcal{D}| \ln \varphi(x, y, \theta) .$$

Задача минимизации эмпирического риска:

$$\mathcal{L}(a_\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum\nolimits_{x\in\mathcal{D}} \mathcal{L}(a_\theta, x) =$$

$$= -\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} |\mathcal{D}| \ln \varphi(x, y, \theta) = -\sum_{(x,y)\in\mathcal{D}} \ln \varphi(x, y, \theta) \to \min_\theta$$

Принцип максимального правдоподобия

Принцип максимального правдоподобия:

$$\mathcal{L}(\theta; \mathcal{D}_s) = \sum_{x \in \mathcal{D}_s} \ln \varphi(x; \theta) \to \max_\theta,$$

Оптимум θ можно найти в точке, где производная

$$\frac{\partial \mathcal{L}(\theta; \mathcal{D}_s)}{\partial \theta} = 0.$$

Это чаще всего можно искать только итеративно.

Принцип совместного максимального правдоподобия

$$\mathcal{L}(a_\theta, \mathcal{D}) = - \sum_{(x_i, y_i) \in \mathcal{D}} \ln \varphi(x_i, y_i, \theta) \to \min_\theta .$$

$$\varphi(x_i, y_i, \theta) = p(x_i, y_i | w) p(w, \gamma),$$

$p(x_i, y_i | w)$ – вероятностная модель данных,
$p(w, \gamma)$ – априорное распределение параметров модели,
γ – гиперпараметр (в статистическом смысле).

Принцип совместного максимального правдоподобия:

$$\sum_{(x_i, y_i) \in \mathcal{D}} \ln p(x_i, y_i | w) + \ln p(w, \gamma) \to \max_{w, \gamma}$$

Условия квадратичной регуляризации

Пусть $w \in \mathbb{R}^n$ описывается n-мерным нормальным распределением:

$$p(w;\sigma) = \frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{\|w\|^2}{2\sigma}\right),$$

(веса независимы, их матожидания нулевые, их дисперсии равны σ).

Это приводит к квадратичной регуляризации:

$$-\ln p(w;\sigma) = \frac{1}{2\sigma}\|w\|^2 + \mathrm{const}(w).$$

План лекции

- Байесовская классификация
- Непараметрическое восстановление плотности распределения
- Параметрическое восстановление плотности распределения
- Мягкая классификация и логистическая регрессия

Мягкая классификация

Основная идея: вместо того, чтобы возвращать только класс, можно возвращать вектор вероятностей каждого класса:

$$a: X \to Y, \qquad a(x) = y$$

$$b: X \to \mathbb{R}^{|Y|}, \qquad b(x) = \left(q_1, \ldots, q_{|Y|}\right), \sum_i q_i = 1$$

В целом, мягкая классификация более информативна и позволяет использовать более чувствительные функции ошибки.

Перекрестная энтропия

Перекрестная энтропия:

$$H(p,p') = -\sum_{x\in X} p(x)\log p'(x)$$

Функция ошибки перекрестной энтропии (cross-entropy loss):

$$\mathcal{L}(b,x) = -\log \mathrm{q}_{y(x)},$$

где $\mathrm{q}_{y(x)} - y(x)$-й элемент вектора $b(x)$.

Регрессия для предсказания вероятностей

Рассмотрим бинарную классификацию,

$$q_+ + q_- = 1.$$

Можно было бы предсказывать вероятность одного из классов, скажем, q_+.

Это не задача регрессии, потому что $0 \le q_+ \le 1$, а не просто $q_- \in \mathbb{R}$.

Зато $\log \frac{q_+}{1-q_+} \in \mathbb{R}$. Это **логит-преобразование**.

Логистическая регрессия

Возьмем для предсказания линейную регрессию:

$$\log \frac{q_+}{1 - q_+} = \langle w, x \rangle$$

$$q_+ = \frac{1}{1 + e^{\langle w, x \rangle}}$$

Логистическая регрессия:

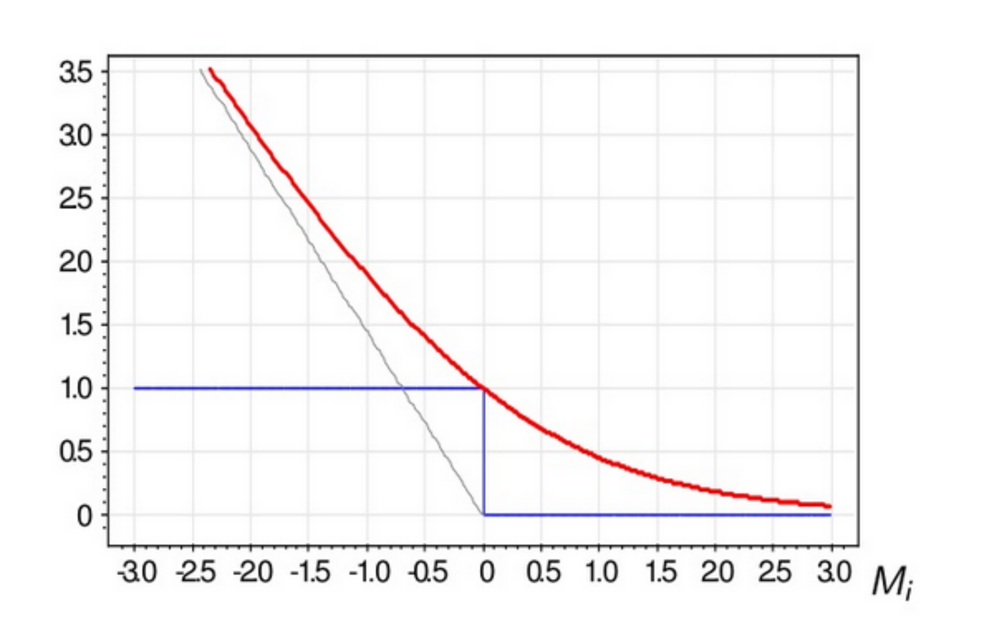
$$a_{\text{LogReg}}(x) = \sigma(\langle w, x \rangle),$$

где $\sigma(x) = \frac{1}{1+e^x}$ – **сигма-функция**.

Логарифмическая функция потерь

Логарифмическая функция потерь:

$$\mathcal{L}(a, \mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} \ln(1 + \exp(-\langle w, x\rangle y)) \to \min_w.$$



Градиентный спуск по $\sigma(x)$

Производная:

$$\sigma'(x) = \sigma(x)\sigma(-x).$$

Градиент:

$$\nabla\mathcal{L}\left(w_{(k)}\right) = -\sum_{i}^{|\mathcal{D}|} y_i x_i \sigma\left(-M_i\left(w_{(k)}\right)\right).$$

Шаг градиентного спуска:

$$w_{(k+1)} = w_{(k)} - \mu y_{(k)} x_{(k)} \sigma\left(-M_i\left(w_{(k)}\right)\right).$$

В следующей серии

- Как вырастить дерево
- Как собрать ансамбль
- Как рандомизировать лес