Анализ данных (Байесовские методы машинного обучения)

М. В. Лебедев

Московский авиационный институт

14 февраля 2022 г.

Материалы к курсу

- Bayesian Methods for Machine Learning (Coursera)
- Байесовские методы машинного обучения
- Лекции Ветрова Д.П.

Условная плотность распределения

Заданы непрерывные СВ X, Y. Тогда $p(y|x) = \frac{p(x,y)}{p(x)}$.

$$p(y|x)p(x) = p(x,y) = p(x|y)p(y) \Rightarrow$$

Формула обращения условной плотности

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

Условная плотность распределения

$$p(y) = \int p(x, y)\, dx = \int p(y|x)p(x)\, dx.$$

Тогда получим формулы маргинализации или правило суммирования (аналог формулы полной вероятности)

$$p(y) = \int p(y|x)p(x)\, dx.$$

Теорема Байеса

Теорема Байеса

$$p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}.$$

$$\text{Апосториорная} = \frac{\text{Правдоподобие} \cdot \text{Априорная}}{\text{Полная}}.$$

Условное распределение

Условная плотность:

$$p(y|x_0) = \frac{p(x_0, y)}{p(x_0)}.$$

Метод максимума апостериорной вероятности

Мотивация (Байесовский подход)

Пусть θ дискретный случайный вектор, принимающий значения θ_i, $i = 1, \dots, m$. У нас есть реализовавшееся наблюдение x_n тогда из формулы Байесса:

$$\mathrm{P}(\theta = \theta_i | X_n = x_n) = \frac{\mathrm{P}(X_n = x_n | \theta_i)\, \mathrm{P}(\theta = \theta_i)}{\mathrm{P}(X_n = x_n)}.$$

В итоге будем выбирать $\hat{\theta}$ из условия максимума

$$\hat{\theta}_{MA} \in \arg\max_{\theta_i} \mathrm{P}(x_n | \theta_i)\, \mathrm{P}(\theta_i).$$

Метод максимума апостериорной вероятности

Для плотности вероятности

$$p(\theta|x_n) = \frac{p(x_n|\theta)p(\theta)}{p(x_n)} \Rightarrow \hat{\theta} \in \arg\max_{\theta\in\Theta} \frac{p(x_n|\theta)p(\theta)}{p(x_n)} \Rightarrow$$

$$\hat{\theta}_{MA} \in \arg\max_{\theta\in\Theta} p(x_n|\theta)p(\theta).$$

Метод максимального правдоподобия

Мотивация

Пусть θ дискретный случайный вектор, принимающий значения θ_i, $i = 1, \ldots, m$. У нас есть реализовавшееся наблюдение x_n тогда из формулы Байесса:

$$\mathrm{P}(\theta = \theta_i | X_n = x_n) = \frac{\mathrm{P}(X_n = x_n | \theta_i)\,\mathrm{P}(\theta = \theta_i)}{\mathrm{P}(X_n = x_n)}.$$

Будем выбирать $\hat{\theta}$ из условия максимума

$$\hat{\theta} \in \arg\max_{\theta_i} \frac{\mathrm{P}(x_n|\theta_i)\,\mathrm{P}(\theta_i)}{\mathrm{P}(x_n)}.$$

Если положить, что $P(\theta_i) = const$, то $\hat{\theta}_{ML} \in \arg\max_{\theta_i} \mathrm{P}(x_n|\theta_i)$.

Метод максимального правдоподобия

Для плотности вероятности

$$p(\theta|x_n) = \frac{p(x_n|\theta)f(\theta)}{p(x_n)} \Rightarrow \hat{\theta} \in \arg\max_{\theta\in\Theta} \frac{p(x_n|\theta)p(\theta)}{p(x_n)} \Rightarrow$$
$$\hat{\theta}_{ML} \in \arg\max_{\theta\in\Theta} p(x_n|\theta).$$

Функция правдоподобия

Пусть неслучайный параметр $\theta \in \Theta \subset \mathbb{R}^m$, тогда функция правдоподобия для x_n в случае непрерывной наблюдаемой СВ X: $L(x_n, \theta) = p_{X_n}(x_n, \theta_1, \ldots, \theta_m) = \prod_{k=1}^{n} p_X(x_k, \theta_1, \ldots, \theta_m)$, а для дискретной СВ X: $L(x_n, \theta_1, \ldots, \theta_m) = \prod_{k=1}^{n} p_X(x_k, \theta_1, \ldots, \theta_m)$, где $p_X(x_k, \theta_1, \ldots, \theta_k)$ — вероятность события $\{X = x_k\}$.

Оценка максимального правдоподобия

Оценка максимального правдоподобия

Оценка максимального правдоподобия (МП-оценка) параметра $\theta \in \Theta$ называется статистика $\hat{\theta}(X_n)$, максимизирующая для каждой реализации x_n функцию правдоподобия, т.е.

$$\hat{\theta}(x_n) \in \arg\max_{\theta \in \Theta} L(x_n, \theta).$$

Оценка максимального правдоподобия

Теорема

Пусть выполнены следующие условия:

1. При каждом $\theta \in \Theta$ $\left|\frac{\partial^{(k)} f(x,\theta)}{\partial\theta}\right| \leq g_k(x)$, $k = 1, 2, 3$, причем $g_1(x)$ и $g_2(x)$ интегрируемы на $\mathbb{R}^1$, $\sup\limits_{\theta\in\Theta} \int\limits_{-\infty}^{\infty} g_3(x)\,dx < \infty$.
2. При каждом $\theta \in \Theta$ функция $i(\theta) = \mathrm{M}_\theta \left\{ \left(\frac{\partial \ln f(x,\theta)}{\partial\theta} \right)^2 \right\}$ конечна и положительна.

Тогда уравнение правдоподобия имеет решение $\hat{\theta}_n$, обладающее следующими свойствами:

1. $\mathrm{M}\{\hat{\theta}_n - \theta_0\} \to 0$ (асимптотическая несмещенность);
2. $\hat{\theta}_n \overset{\text{п.н.}}{\to} \theta_0$, $n \to \infty$ (сильная состоятельность);
3. $\sqrt{n \cdot i(\theta)}(\hat{\theta}_n - \theta_0) \overset{F}{\to} U \sim \mathrm{N}(0, 1)$ (асимптотическая нормальность). θ_0 — истинное значение параметра.
4. Оценка $\hat{\theta}_n$ является асимптотически эффективной.

Эффективность точечных оценок

Регулярное распределение

Распределение $F(x,\theta)$ называется регулярным, если выполнены следующие условия.

- Функция $\sqrt{f(x,\theta)}$ непрерывно дифференцируема по θ на Θ для почти всех x (по мере Лебега).
- Функция $i(\theta) = \mathrm{M}_\theta\left\{\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2\right\} =$

$$= \int\limits_{-\infty}^{\infty}\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)\,dx$$

  конечна положительная и непрерывна по $\theta\in\Theta$.
- $i(\theta)$ — называется информационным количеством Фишера одного наблюдения с плотностью распределения $f(x,\theta)$.

Теорема Рао-Крамера

Пусть выполнено условие регулярности (см. пред. слайд), тогда для любой несмещенной оценки $\hat{\theta}(X_n)$ параметра $\theta \in \Theta \subset \mathbb{R}^1$ справедливо неравенство Рао-Крамер:

$$\mathrm{D}[\hat{\theta}(X_n)] \geq \frac{1}{i(\theta)} = \Delta_n^{min},$$

где $i(\theta)$ — информационное количество Фишера (информация Фишера).

Несмещенная оценка $\hat{\theta}(X_n)$ с.к.-погрешность которой совпадает при всех $n \geq 1$ с нижней границей Δ_n^{min} называется эффективной по Рао-Крамеру (R-эффективной). Эффективная оценка является с.к.-оптимальной на классе всех несмещенных оценок.

Байесовский подход

Преимущества и недостатки

	Частотный	Байесовский
Интерпретация случайности	Объективная неопределенность	Субъективное незнание
Метод вывода	Метод максимального правдоподобия	Теорема Байеса
Оценка	θ_{ML}	$p(\theta\|X)$
Данные	Параметр θ зафиксирован. X — случайный	θ — случайный X – зафиксирован
Применимость	$\|X\| \gg \|\theta\|$	$\forall\|X\|$

Пример (Объединение вероятностных моделей)

Первое измерение

$$p(\theta|x_1) = \frac{p_1(x_1|\theta)p(\theta)}{\int p(x_1|\theta)p(\theta)\,d\theta}.$$

Второе измерение

$$p(\theta|x_1, x_2) = \frac{p_2(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1)\,d\theta}.$$

n-ое измерение

$$p(\theta|x_1, x_2, \ldots, x_n) = \frac{p_n(x_n|\theta)p(\theta|x_1, \ldots, x_{n-1})}{\int p(x_n|\theta)p(\theta|x_1, \ldots, x_{n-1})\,d\theta}.$$

Пример 2

Пример (Линейная регрессия)

$x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $y \in \mathbb{R}$.

- x — признаки.
- y — целевая переменная.
- w — веса линейной регрессии.

Исходные данные

Модель наблюдения:

$$y = w^T x + \varepsilon, \quad w \sim \mathcal{N}(0, I), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

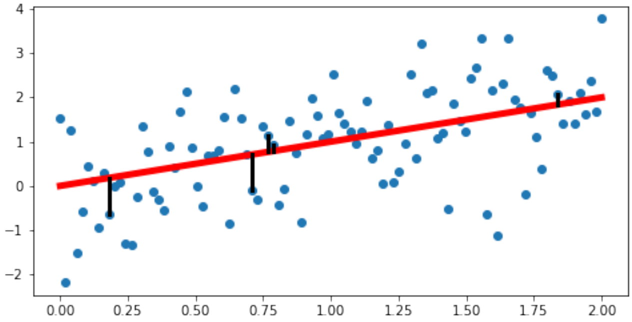
Вероятностная модель

$p(y, w|x) = p(y|x, w)p(w) = \mathcal{N}(y|w^T x, \sigma^2)\mathcal{N}(w|0, I).$

Пусть задана обучающая выборка

$$(X, Y) = (x_i, y_i)_{i=1}^n.$$

$$L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \|w^T X - y\|^2 \to \min_w$$



Максимум апостериорной вероятности

$$p(w|X,Y) = \frac{p(Y,w|X)}{p(Y|X)} \Rightarrow \quad w \in \arg\max_w p(Y,w|X)$$

$$p(Y,w|X) = p(Y|X,w)p(w).$$

$$\arg\max_w p(w|X,Y) = \arg\max_w p(Y|X,w)p(w) =$$

$$= \arg\max_w \prod_{i=1}^n p(y_i|x_i,w)p(w) =$$

$$= \arg\max_w \left\{ log \sum_{i=1}^n p(y_i|x_i,w) + \log p(w) \right\} =$$

$$= \arg\min_w \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^T w)^2 + \frac{1}{2}\|w\|^2 \right\}.$$