Семинар по ММРО "Решение задач на Байесовские методы"

Афанасьев Глеб Ильич

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

17 февраля 2022 г.

Основные понятия

X = $x_1, x_2, ..., x_l$ - обучающая выборка
$\mathbb{X}$ - множество всевозможных объектов
$\mathbb{Y}$ - множество ответов

Основные понятия

X = $x_1, x_2, ..., x_l$ - обучающая выборка
$\mathbb{X}$ - множество всевозможных объектов
$\mathbb{Y}$ - множество ответов

- В байесовском подходе предполагается, что обучающие объекты и ответы на них $(x_1, y_1), ..., (x_{,y_)}$ независимо выбираются из некоторого распределения p(x, y), заданного на множестве $\mathbb{X} \times \mathbb{Y}$

Напоминание

Из курса теории вероятностей:

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

p(y) - априорное распределение
p(x | y) - правдоподобие
p(y | x) - апостериорное распределение
Формула Байеса:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\int p(s)p(x|s)ds}$$

Пусть на множестве всех пар ответов Y × Y задана функция потерь L(y, s)
Функционалом среднего риска называется матожидание функции потерь по всем парам (x, y) при использовании алгоритма a(x):

$$R(a) = \mathbb{E}L(y, a(x)) = \int_Y \int_X L(y, a(x)) p(x, y) dx dy$$

Если распределение p(x, y) известно, то можно найти алгоритм a*(x), оптимальный с точки зрения функционала среднего риска

Классификация

L(y, s) = [y $\neq$ s]
$a^*(x) = argmax_{y \in Y} p(y|x)$ Для произвольного классификатора a(x) запишем:

$$R(a) = \int_Y \int_X L(y, a(x)) p(x, y) dx dy = \sum_{y=1}^{K} \int_X [y \neq a(x)] p(x, y) dx =$$

$$= \int_X \sum_{y=1}^{K} [y \neq a(x)] p(x, y) dx = \int_X \sum_{y \neq a(x)} p(x, y) dx =$$

$$= 1 - \int_X p(x, a(x)) dx >= 1 - \int_X max_s p(x, s) dx = 1 - \int_X max_s p(x, a^*(x)$$

Регрессия

$L(y, s) = (y - s)^2$

Преобразуем функцию потерь следующим образом:

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 =$$
$$= (y - \mathbb{E}(y \mid x))^2 + 2\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big) + (\mathbb{E}(y \mid x) - a(x))$$

Подставим ее в функционал среднего риска, получаем:

$$R(a) = \int_Y \int_{\mathbb{X}} L(y, a(x)) p(x, y) dx dy =$$
$$= \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t \mid x))^2 p(x, y) dx dy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t \mid x) - a(x))^2 p(x, y) dx dy +$$
$$+ 2 \int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t \mid x)\big)\big(\mathbb{E}(t \mid x) - a(x)\big) p(x, y) dx dy.$$

Регрессия

Рассмотрим слагаемое 3:

$$
\begin{aligned}
&\int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t \,|\, x)\big)\big(\mathbb{E}(t \,|\, x) - a(x)\big) p(x, y) dx dy = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t \,|\, x) - a(x)\big) \int_Y \left\{ \big(y - \mathbb{E}(t \,|\, x)\big) p(x, y) \right\} dy dx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t \,|\, x) - a(x)\big) \left\{ \int_Y y p(x, y) dy - \int_Y \mathbb{E}(t \,|\, x) p(x, y) dy \right\} dx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t \,|\, x) - a(x)\big) \left\{ p(x) \int_Y y p(y \,|\, x) dy - \mathbb{E}(t \,|\, x) \int_Y p(x, y) dy \right\} dx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t \,|\, x) - a(x)\big) \underbrace{\{p(x)\mathbb{E}(t \,|\, x) - p(x)\mathbb{E}(t \,|\, x)\}}_{=0} dx = \\
&= 0
\end{aligned}
$$

Регрессия

Получаем:

$$R(a) = \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t \mid x))^2 p(x, y) dx dy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t \mid x) - a(x))^2 p(x, y) dx dy.$$

От алгоритма a(x) зависит только второе слагаемое, и оно достигает своего минимума, если a(x) = E(t| x). Таким образом, оптимальная байесовская функция регрессии для квадратичной функции потерь имеет вид:

$$a_*(x) = \mathbb{E}(y \mid x) = \int_Y y p(y \mid x) dy.$$

Проблема оптимальных байесовских алгоритмов:
Неизвестно распределение $p(x, y)$
Существует два подхода восстановления $p(x, y)$ по обучающей выборке - параметрический и непараметрический. Рассмотрим параметрический.

параметрический метод

Допустим, распределение на парах «объект-ответ» зависит от некоторого параметра : p(x, y | θ). Тогда получаем следующую формулу для апостериорной вероятности:

$$p(y \mid x, \theta) \propto p(x \mid y, \theta) p(y),$$

Для оценивания параметров применяется метод максимального правдоподобия:

$$\theta_* = \arg\max_\theta L(\theta) = \arg\max_\theta \prod_{i=1}^{\ell} p(x_i \mid y_i, \theta),$$

Рассмотрим пример для линейной регрессии.

параметрический метод

Апостериорное распределение-нормальное с параметрами примет вид

$$p(y \mid x, w) = \mathcal{N}(\langle w, x \rangle, \sigma^2).$$

$$L(w) = \prod_{i=1}^{\ell} p(y_i \mid x_i, w) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle w, x_i \rangle)^2}{2\sigma^2}\right).$$

Перейдем к логарифму правдоподобия:

$$\log L(w) = -\ell \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 \to \max_w .$$

Убирая все члены, не зависящие от вектора весов w, получаем задачу наименьших квадратов

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 \to \min_w .$$

параметрический метод

Апостериорное распределение на параметры:
Пусть p() — априорное распределение на векторе параметров .
В качестве функции правдоподобия для данного вектора возьмем апостериорное распределение на ответах p(y | x, θ).
Тогда по формуле Байеса

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta) p(\theta)}{p(y \mid x)}.$$

Вернемся к примеру с линейной регрессией. Введем априорное распределение на векторе весов:

$$p(w_j) = \mathcal{N}(0, \alpha^2), \quad j = 1, \dots, d.$$

параметрический метод

$$p(w \mid y, x) = \prod_{i=1}^{\ell} p(y_i \mid x_i, w) p(w) =$$

$$= \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle w, x_i \rangle)^2}{2\sigma^2}\right) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{w_j^2}{2\alpha^2}\right).$$

Перейдем к логарифму и избавимся от константных членов:

$$\log p(w \mid y, x) = -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 - \frac{\ell}{2\alpha^2} \underbrace{\sum_{j=1}^{d} w_j^2}_{=\|w\|^2}.$$

В итоге получаем задачу гребневой регрессии

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 + \lambda \|w\|^2 \to \min_w,$$

где $\lambda = \frac{\ell}{2\alpha^2}$.