Машинное обучение, ФКН ВШЭ
Семинар №16

1 Байесовские методы машинного обучения

Пусть $X = \{x_1, \ldots, x_\ell\}$ — выборка, $\mathbb{X}$ — множество всех возможных объектов, Y — множество ответов. В байесовском подходе предполагается, что обучающие объекты и ответы на них $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ независимо выбираются из некоторого распределения $p(x, y)$, заданного на множестве $\mathbb{X} \times Y$. Данное распределение можно переписать как

$$p(x, y) = p(y)p(x \mid y),$$

где $p(y)$ определяет вероятности появления каждого из возможных ответов и называется *априорным распределением*, а $p(x \mid y)$ задает распределение объектов при фиксированном ответе y и называется *функцией правдоподобия*.

Если известны априорное распределение и функция правдоподобия, то по формуле Байеса можно записать *апостериорное распределение* на множестве ответов:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int_s p(x \mid s)p(s)ds} = \frac{p(x \mid y)p(y)}{p(x)},$$

где знаменатель не зависит от y и является нормировочной константой.

§1.1 Оптимальные байесовские правила

Пусть на множестве всех пар ответов $Y \times Y$ задана функция потерь $L(y, s)$. Наиболее распространенным примером для задач классификации является ошибка классификации $L(y, s) = [y \neq s]$, для задач регрессии — квадратичная функция потерь $L(y, x) = (y - s)^2$. *Функционалом среднего риска* называется матожидание функции потерь по всем парам (x, y) при использовании алгоритма $a(x)$:

$$R(a) = \mathbb{E}L(y, a(x)) = \int_Y \int_{\mathbb{X}} L(y, a(x))p(x, y)dxdy.$$

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

1.1.1 Классификация

Начнем с задачи классификации с множеством ответом $Y = \{1, \ldots, K\}$ и функции потерь $L(y, s) = [y \neq s]$. Покажем, что минимум функционала среднего риска достигается на алгоритме

$$a_*(x) = \arg\max_{y \in Y} p(y \mid x).$$

Для произвольного классификатора $a(x)$ выполнена следующая цепочка неравенств:

$$\begin{aligned}
R(a) &= \int_Y \int_{\mathbb{X}} L(y, a(x)) p(x, y) dx dy = \\
&= \sum_{y=1}^{K} \int_{\mathbb{X}} [y \neq a(x)] p(x, y) dx = \\
&= \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx = \left\{ \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx + \int_{\mathbb{X}} p(x, a(x)) dx = 1 \right\} = \\
&= 1 - \int_{\mathbb{X}} p(x, a(x)) dx \geqslant \\
&\geqslant 1 - \int_{\mathbb{X}} \max_{s \in Y} p(x, s) dx = \\
&= 1 - \int_{\mathbb{X}} p(x, a_*(x)) dx = \\
&= R(a_*)
\end{aligned}$$

Таким образом, средний риск любого классификатора $a(x)$ не превосходит средний риск нашего классификатора $a_*(x)$.

Мы получили, что оптимальный байесовский классификатор выбирает тот класс, который имеет наибольшую апостериорную вероятность. Такой классификатор называется *MAP-классификатором* (maximum a posteriori).

1.1.2 Регрессия

Напомним, что при выводе разложения на шум, смещение и разброс функционала среднего риска для задачи регрессии и функции потерь $L(y, x) = (y - s)^2$ нами уже была получена формула оптимального алгоритма с точки зрения данного функционала:

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

Иными словами, мы должны провести «взвешенное голосование» по всем возможным ответам, причем вес ответа равен его апостериорной вероятности.

§1.2 Байесовский вывод

Основной проблемой оптимальных байесовских алгоритмов, о которых шла речь в предыдущем разделе, является невозможность их построения на практике, поскольку нам никогда неизвестно распределение $p(x, y)$. Данное распределение можно попробовать восстановить по обучающей выборке, при этом существует два подхода — параметрический и непараметрический. Сейчас мы сосредоточимся на параметрическом подходе.

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

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

где выражение «$a \propto b$» означает «a пропорционально b». Для оценивания параметров применяется *метод максимального правдоподобия*:

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

где $L(\theta)$ — функция правдоподобия. Примером такого подхода может служить *нормальный дискриминантный анализ*, где предполагается, что функции правдоподобия являются нормальными распределениями с неизвестными параметрами $\theta = (\mu, \Sigma)$. Об этом подходе речь пойдет на следующем семинаре, а сейчас рассмотрим более простой пример.

Иногда удобнее сразу задавать апостериорное распределение — например, в случае с линейной регрессией. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом: вычисляется линейная функция $\langle w, x\rangle$, и к результату прибавляется нормальный шум:

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

В этом случае апостериорное распределение примет вид

$$p(y \mid x, w) = \mathcal{N}(\langle w, x\rangle, \sigma^2). \tag{1.1}$$

Задача 1.1. *Покажите, что метод максимального правдоподобия для модели* (1.1) *эквивалентен методу наименьших квадратов.*

Решение. Запишем правдоподобие для выборки $x_1, \ldots, x_\ell$:

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

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

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

■

Байесовский вывод параметров. В некоторых случаях применение метода максимального правдоподобия для поиска параметров приводит к плохим результатам. Например, если имеет место мультиколлинеарность, то функция правдоподобия имеет много минимумов, и решение может оказаться переобученным. Одним из подходов к устранению этой проблемы является введение априорного распределения *на параметрах*.

Пусть $p(\theta)$ — априорное распределение на векторе параметров θ. В качестве функции правдоподобия для данного вектора возьмем апостериорное распределение на ответах $p(y \mid x, \theta)$. 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:

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

Иными словами, мы предполагаем, что веса концентрируются вокруг нуля.

Задача 1.2. *Покажите, что максимизация апостериорной вероятности $p(w \mid y, x)$ для модели линейной регрессии с нормальным априорным распределением эквивалентна решению задачи гребневой регрессии.*

Решение. Запишем апостериорную вероятность вектора весов w для выборки $x_1, \ldots, x_\ell$:

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

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

■

После того, как оптимальный вектор весов w_* найден, мы можем найти распределение на ответах для нового объекта x:

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

Выше мы выяснили, что оптимальным ответом будет матожидание $\mathbb{E}(y \mid x) = \int yp(y \mid x, X, w_*)dy$.

С точки зрения байесовского подхода [**?**] правильнее не искать моду [1] w_* апостериорного распределения на параметрах и брать соответствующую ей модель $p(y \mid x, X, w_*)$, а устроить «взвешенное голосование» всех возможных моделей:

$$p(y \mid x, X) = \int p(y \mid x, w)p(w \mid Y, X)dw,$$

где $X = \{x_1, \ldots, x_\ell\}$, $Y = \{y_1, \ldots, y_\ell\}$.

§1.3 Наивный байесовский классификатор

Как было сказано ранее, при применении байесовского классификатора необходимо решить задачу восстановления плотности $p_y(x)$ для каждого класса $y \in \mathbb{Y}$. Данная задача является довольно трудоёмкой и не всегда может быть решена, особенно в случае большого количества признаков, — в частности, если объектами являются тексты, приходится работать с крайне большим числом признаков, и восстановление плотности многомерного распределения не представляется возможным.

Для разрешения этой проблемы сделаем предположение о независимости признаков. В этом случае функция правдоподобия класса y для объекта $x = (x_1, \ldots, x_d)$ может быть представлена в следующем виде:

$$p(x \mid y) = \prod_{j=1}^{d} p(x_j \mid y),$$

где $p(x_j \mid y)$ — одномерная плотность распределения j-ого признака объектов класса $y \in Y$. В этом случае формула байесовского решающего правила примет следующий вид:

$$a(x) = \arg\max_{y\in Y} p(y \mid x) = \arg\max_{y\in\mathbb{Y}} \left(\ln p(y) + \sum_{j=1}^{d} \ln p(x_j \mid y)\right).$$

[1] Мода — точка максимума плотности.

Предположение о независимости признаков существенно облегчает задачу, поскольку вместо решения задачи восстановления d-мерной плотности необходимо решить d задач восстановления одномерных плотностей. Полученный классификатор называется *наивным байесовским классификатором*.

Плотности отдельных признаков могут быть восстановлены различными способами (параметрическими и непараметрическими). Среди параметрических способов чаще всего используются нормальное распределение (для вещественных признаков), распределение Бернулли и мультиномиальное распределение (для дискретных признаков), благодаря которым получаются различные применяющиеся на практике модели.