Машинное обучение, ФКН ВШЭ
Семинар №17

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 Особенности байесовских алгоритмов

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:

$$a(x) = \arg\max_{y \in Y} p(y)p(x \mid y),$$
$$p(x \mid y) = \mathcal{N}(x \mid \mu_y, \Sigma_y).$$

Параметрами алгоритма являются средние μ_y и ковариационные матрицы классов Σ_y, которые оцениваются по выборке методом максимального правдоподобия.

Если предположить, что ковариационные матрицы классов равны, и оценивать их по всей выборке, то мы получим алгоритм, называемый *линейным дискриминантом Фишера*. Можно показать, что он является линейным:

$$a(x) = \arg\max_{y \in Y}(\langle w_y, x\rangle + w_{0y}),$$

причем $w_y = \Sigma^{-1}\mu_y$. В случае двух классов ($Y = \{-1, +1\}$) классификатор принимает вид

$$a(x) = \operatorname{sign}(\langle w, x\rangle + b) \quad w = \Sigma^{-1}(\mu_2 - \mu_1). \tag{1.1}$$

§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:

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

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.