Машинное обучение, ФКН ВШЭ
Семинар №6

1 Предсказание вероятностей

Разберемся, каким требованиям должен удовлетворять классификатор, чтобы его выход можно было расценивать как оценку вероятности класса.

Пусть в каждой точке $x \in \mathbb{X}$ пространства объектов задана вероятность $p(y = +1 \mid x)$ того, что данный объект относится к классу $+1$, и пусть алгоритм $b(x)$ возвращает числа из отрезка $[0, 1]$. Потребуем, чтобы эти предсказания пытались в каждой точке x приблизить вероятность положительного класса $p(y = +1|x)$.

Разумеется, выполнение этого требования зависит от функции потерь — минимум ее матожидания в каждой точке x должен достигаться на данной вероятности:

$$\arg\min_{b \in \mathbb{R}} \mathbb{E}\left[L(y, b)|x\right] = p(y = +1|x).$$

Задача 1.1. *Покажите, что квадратичная функция потерь $L(y, b) = ([y = +1] - b)^2$ позволяет предсказывать корректные вероятности.*

Решение. Заметим, что поскольку алгоритм возвращает числа от 0 до 1, то его ответ должен быть близок к единице, если объект относится к положительному классу, и к нулю — если объект относится к отрицательному классу.

Запишем матожидание функции потерь в точке x:

$$\mathbb{E}\left[L(y, b)|x\right] = p(y = +1|x)(b - 1)^2 + (1 - p(y = +1|x))(b - 0)^2.$$

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y, b)|x\right] = 2p(y = +1|x)(b - 1) + 2(1 - p(y = +1|x))b = 2b - 2p(y = +1|x) = 0.$$

Легко видеть, что оптимальный ответ алгоритма действительно равен вероятности:

$$b = p(y = +1|x).$$

■

Задача 1.2. *Покажите, что абсолютная функция потерь $L(y, b) = |[y = +1] - b|$, $b \in [0; 1]$, не позволяет предсказывать корректные вероятности.*

Решение. Запишем матожидание функции потерь в точке x:

$$\begin{aligned}\mathbb{E}\left[L(y,b)|x\right] =& p(y=+1|x)|1-b| + (1-p(y=+1|x))|b| = \\ &= p(y=+1|x)(1-b) + (1-p(y=+1|x))b.\end{aligned}$$

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y,b)|x\right] = 1 - 2p(y=+1|x) = 0.$$

Рассмотрим 2 случая:

1. $p(y=+1|x) = \frac{1}{2}$. Тогда $\mathbb{E}\left[L(y,b)|x\right] = \frac{1}{2} \quad \forall b \in [0;1]$, а потому классификатор не позволяет предсказывать корректную вероятность в точке x.

2. $p(y=+1|x) \neq \frac{1}{2}$. В этом случае интервал $(0;1)$ не содержит критических точек, а потому минимум матожидания достигается на одном из концов отрезка $[0;1]$:

$$\begin{aligned}\min_{b\in[0;1]} \mathbb{E}\left[L(y,b)|x\right] = \min\left(\mathbb{E}\left[L(y,0)|x\right], \mathbb{E}\left[L(y,1)|x\right]\right) = \\ \min\left(p(y=+1|x), 1-p(y=+1|x)\right).\end{aligned}$$

   Отсюда $\arg\min_{b\in[0;1]} \mathbb{E}\left[L(y,b)|x\right] \in \{0,1\}$, а потому классификатор также не позволяет предсказывать корректную вероятность в точке x.

■

2 Квантильная регрессия

В некоторых задачах цены занижения и завышения прогнозов могут отличаться друг от друга. Например, при прогнозировании спроса на товары интернет-магазина гораздо опаснее заниженные предсказания, поскольку они могут привести к потере клиентов. Завышенные же прогнозы приводят лишь к издержкам на хранение товара на складе. Функционал в этом случае можно записать как

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} \rho_\tau(y_i - a(x_i)),$$

где

$$\rho_\tau(z) = (\tau-1)[z<0]z + \tau[z\geqslant 0]z = (\tau - \frac{1}{2})z + \frac{1}{2}|z|,$$

а параметр τ лежит на отрезке $[0,1]$ и определяет соотношение важности занижения и завышения прогноза. Чем больше здесь τ, тем выше штраф за занижение прогноза.

Обсудим вероятностный смысл данного функционала. Будем считать, что в каждой точке $x \in \mathbb{X}$ пространства объектов задано вероятностное распределение $p(y\,|\,x)$ на возможных ответах для данного объекта. Такое распределение может возникать, например, в задаче предсказания кликов по рекламным баннерам: один и тот же пользователь может много раз заходить на один и тот же сайт и видеть

данный баннер; при этом некоторые посещения закончатся кликом, а некоторые — нет.

Известно, что при оптимизации квадратичного функционала алгоритм $a(x)$ будет приближать условное матожидание ответа в каждой точке пространства объектов: $a(x) \approx \mathbb{E}[y \mid x]$; если же оптимизировать среднее абсолютное отклонение, то итоговый алгоритм будет приближать медиану распределения: $a(x) \approx \text{median}[p(y \mid x)]$. Рассмотрим теперь некоторый объект x и условное распределение $p(y \mid x)$. Найдем число q, которое будет оптимальным с точки зрения нашего функционала:

$$Q = \int_{\mathbb{Y}} \rho_\tau(y - q) p(y \mid x) dy.$$

Продифференцируем его (при этом необходимо воспользоваться правилами дифференцирования интегралов, зависящих от параметра):

$$\frac{\partial Q}{\partial q} = (1 - \tau) \int_{-\infty}^{q} p(y \mid x) dy - \tau \int_{q}^{\infty} p(y \mid x) dy = 0.$$

Получаем, что

$$\frac{\tau}{1 - \tau} = \frac{\int_{-\infty}^{q} p(y \mid x) dy}{\int_{q}^{\infty} p(y \mid x) dy}.$$

Данное уравнение будет верно, если q будет равно τ-квантили распределения $p(y \mid x)$. Таким образом, использование функции потерь $\rho_\tau(z)$ приводит к тому, что алгоритм $a(x)$ будет приближать τ-квантиль распределения ответов в каждой точке пространства объектов.