Theory task ML

Бунин Кирилл Андреевич, Б05-203

Dec 10, 2024

Bias-Variance decomposition

1 Теория

Допустим, у нас есть некоторая выборка, на которой линейные методы работают лучше решающих деревьев с точки зрения ошибки на контроле. Почему это так? Чем можно объяснить превосходство определённого метода обучения? Оказывается, ошибка любой модели складывается из трёх факторов: сложности самой выборки, схожести модели с истинной зависимостью ответов от объектов в выборке, и богатства семейства, из которого выбирается конкретная модель. Между этими факторами существует некоторый баланс, и уменьшение одного из них приводит к увеличению другого. Такое разложение ошибки носит название разложения на смещение и разброс, и его формальным выводом мы сейчас займёмся.

Пусть задана выборка $X = (x_i, y_i)_{i=1}^n$ с вещественными ответами $y_i \in \mathbb{R}$ (рассматриваем задачу регрессии). Будем считать, что на пространстве всех объектов и ответов $X \times Y$ существует распределение $p(x, y)$, из которого сгенерирована выборка X и ответы на ней.
Рассмотрим квадратичную функцию потерь

$$L(y, a) = (y - a(x))^2$$

и соответствующий ей среднеквадратичный риск

$$R(a) = \mathbb{E}_{x,y}\left[(y - a(x))^2\right] = \int_X \int_Y p(x, y)(y - a(x))^2 dx dy.$$

Данный функционал усредняет ошибку модели в каждой точке пространства x и для каждого возможного ответа y, причём вклад пары (x, y), по сути, пропорционален вероятности получить её в выборке $p(x, y)$. Разумеется, на практике мы не можем вычислить данный функционал, поскольку распределение $p(x, y)$ неизвестно. Тем не менее, в теории он позволяет измерить качество модели на всех возможных объектах, а не только на наблюдённой выборке.

2 Задание

Покажите, что минимум среднеквадратичного риска достигается на функции, возвращающей условное математическое ожидание ответа при фиксированном объекте.

$$a_*(x) = \mathbb{E}[y \mid x] = \int_Y y p(y \mid x) dy = \arg\min_a R(a).$$

Иными словами покажите, что мы должны провести «взвешенное голосование» по всем возможным ответам, при этом веса ответа равны апостериорной вероятности.

3 Решение

Преобразуем функцию потерь:

$$
\begin{gathered}
L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 = \\
= (y - \mathbb{E}(y \mid x))^2 + 2(y - \mathbb{E}(y \mid x))(\mathbb{E}(y \mid x) - a(x)) + (\mathbb{E}(y \mid x) - a(x))^2.
\end{gathered}
$$

Подставляя её в функционал среднеквадратичного риска, получаем:

$$
\begin{gathered}
R(a) = \mathbb{E}_{x,y}[L(y, a(x))] = \\
= \mathbb{E}_{x,y}[(y - \mathbb{E}(y \mid x))^2] + \mathbb{E}_{x,y}[(\mathbb{E}(y \mid x) - a(x))^2] + 2\mathbb{E}_{x,y}[(y - \mathbb{E}(y \mid x))(\mathbb{E}(y \mid x) - a(x))].
\end{gathered}
$$

Разберёмся сначала с последним слагаемым. Перейдём от матожидания $\mathbb{E}_{x,y}[f(x,y)]$ к цепочке матожиданий:

$$
\mathbb{E}_x \mathbb{E}_y[f(x,y) \mid x] = \int_X \left(\int_Y f(x,y) p(y \mid x) dy \right) p(x) dx
$$

и заметим, что величина $(\mathbb{E}(y \mid x) - a(x))$ не зависит от y, и поэтому её можно вынести за матожидание по y:

$$
\begin{gathered}
\mathbb{E}_x \mathbb{E}_y \left[(y - \mathbb{E}(y \mid x))(\mathbb{E}(y \mid x) - a(x)) \mid x \right] = \\
= \mathbb{E}_x \left((\mathbb{E}(y \mid x) - a(x)) \mathbb{E}_y \left[(y - \mathbb{E}(y \mid x)) \mid x \right] \right) = \\
= \mathbb{E}_x \left((\mathbb{E}(y \mid x) - a(x)) (\mathbb{E}_y[y \mid x] - \mathbb{E}_y \mathbb{E}(y \mid x)) \right) = \\
= 0.
\end{gathered}
$$

Получаем, что функционал среднеквадратичного риска имеет вид:

$$
R(a) = \mathbb{E}_{x,y}(y - \mathbb{E}(y \mid x))^2 + \mathbb{E}_{x,y}((\mathbb{E}(y \mid x) - a(x))^2).
$$

От алгоритма $a(x)$ зависит только второе слагаемое, и оно достигает своего минимума, если $a(x) = \mathbb{E}(y \mid x)$. Таким образом, оптимальная модель регрессии для квадратичной функции потерь имеет вид:

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

Что и требовалось показать.