Bias-Variance decomposition

Бунин Кирилл Андреевич, Б05-203

Dec 10, 2024

Задача 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

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

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.

Задание

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

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

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

Решение

Преобразуем функцию потерь:

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

Подставляя её в функционал среднеквадратичного риска, получаем:

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

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:

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

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:

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

Что и требовалось показать.

Задача 2

Теория

Для того, чтобы построить идеальную функцию регрессии, необходимо знать распределение на объектах и ответах $p(x, y)$, что, как правило, невозможно. На практике вместо этого выбирается некоторый *метод обучения* $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to A$, который произвольной обучающей выборке ставит в соответствие некоторый алгоритм из семейства A. В качестве меры качества метода обучения можно взять усредненный по всем выборкам среднеквадратичный риск алгоритма, выбранного методом μ по выборке:

$$L(\mu) = \mathbb{E}_X \left[\mathbb{E}_{x,y} \left[(y - \mu(X)(x))^2 \right] \right] = \tag{1}$$

$$= \int_{(\mathbb{X}\times\mathbb{Y})^\ell} \int_{\mathbb{X}\times\mathbb{Y}} (y - \mu(X)(x))^2 p(x,y) \prod_{i=1}^{\ell} p(x_i, y_i) dx dy dx_1 dy_1 \dots dx_\ell dy_\ell.$$

Здесь матожидание $\mathbb{E}_X[\cdot]$ берется по всем возможным выборкам $\{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$ из распределения $\prod_{i=1}^{\ell} p(x_i, y_i)$.
Обратим внимание, что результатом применения метода обучения $\mu(X)$ к выборке X является модель, поэтому правильно писать $\mu(X)(x)$. Но это довольно громоздкая запись, поэтому будем везде дальше писать просто $\mu(X)$, но не будем забывать, что это функция, зависящая от объекта x.
Среднеквадратичный риск на фиксированной выборке X можно расписать как:

$$\mathbb{E}_{x,y} \left[(y - \mu(X))^2 \right] = \mathbb{E}_{x,y} \left[(y - \mathbb{E}[y \mid x])^2 \right] + \mathbb{E}_{x,y} \left[(\mathbb{E}[y \mid x] - \mu(X))^2 \right].$$

Задание

Подставим это представление в (1):

$$L(\mu) = \mathbb{E}_X \left[\mathbb{E}_{x,y} \left[(y - \mathbb{E}[y \mid x])^2 \right] + \mathbb{E}_{x,y} \left[(\mathbb{E}[y \mid x] - \mu(X))^2 \right] \right] =$$

$$= \mathbb{E}_{x,y} \left[(y - \mathbb{E}[y \mid x])^2 \right] + \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mu(X))^2 \right] \right]. \tag{2}$$

Преобразуем второе слагаемое:

$$\mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mu(X))^2 \right] \right] =$$

$$= \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)] + \mathbb{E}_X[\mu(X)] - \mu(X))^2 \right] \right] =$$

$$= \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)])^2 \right] \right] + \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}_X[\mu(X)] - \mu(X))^2 \right] \right] + \tag{3}$$

$$+ 2\mathbb{E}_{x,y} \left[\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]) (\mathbb{E}_X[\mu(X)] - \mu(X)) \right] \right].$$

Покажите, что последнее слагаемое обращается в нуль.

Решение

Покажем, что последнее слагаемое обращается в нуль:

$$\mathbb{E}_X \left[(\mathbb{E}[y \mid x] - \mathbb{E}_X \left[\mu(X) \right]) (\mathbb{E}_X \left[\mu(X) \right] - \mu(X)) \right] =$$

$$= (\mathbb{E}[y \mid x] - \mathbb{E}_X \left[\mu(X) \right]) \mathbb{E}_X \left[\mathbb{E}_X \left[\mu(X) \right] - \mu(X) \right] =$$

$$= (\mathbb{E}[y \mid x] - \mathbb{E}_X \left[\mu(X) \right]) \left[\mathbb{E}_X \mu(X) - \mathbb{E}_X \mu(X) \right] =$$

$$= 0.$$

Задача 3

Задание

Используя результаты предыдущих задач и подставляя (3) в (2) получите выражение для $L(\mu)$, укажите слагаемые, отвечающие за *смещение*, *шум* и *разброс*.

Решение

Подставим выражение (3) в (2), учитывая результаты предыдущих задач:

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[(y - \mathbb{E}[y \mid x])^2\right]}_{\text{шум}} + \underbrace{\mathbb{E}_x\left[(\mathbb{E}_X[\mu(X)] - \mathbb{E}[y \mid x])^2\right]}_{\text{смещение}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[(\mu(X) - \mathbb{E}_X[\mu(X)])^2\right]\right]}_{\text{разброс}}.$$

Рассмотрим подробнее компоненты полученного разложения ошибки. Первая компонента характеризует *шум* (*noise*) в данных и равна ошибке идеального алгоритма. Невозможно построить алгоритм, имеющий меньшую среднеквадратичную ошибку. Вторая компонента характеризует *смещение* (*bias*) метода обучения, то есть отклонение среднего ответа обученного алгоритма от ответа идеального алгоритма. Третья компонента характеризует *дисперсию* (*variance*), то есть разброс ответов обученных алгоритмов относительно среднего ответа.