Наилучшая модель регрессии. Разложение ошибки на шум, смещение и разброс

Автор материалов: Соколов Евгений Андреевич
Документ составил: Богачев Владимир Александрович

5 января 2024 г.

1 Введение

Рассмотрим задачу регрессии: пусть признаковое описание $\mathcal{X}$ и целевая переменная $\mathcal{Y}$ имеют совместное распределение $(\mathcal{X}, \mathcal{Y}) \sim P_{\mathcal{X},\mathcal{Y}}$.Пусть (X, Y) — выборка мощности l из совместного распределения $P_{\mathcal{X},\mathcal{Y}}$. Необходимо построить функцию $f(x) : \text{supp}(\mathcal{X}) \to \text{supp}(\mathcal{Y})$ из некоторого класса $\mathcal{F}$, оптимизирующую функционал (1)

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathcal{Y})^2 \underset{f\in\mathcal{F}}{\longrightarrow} \min \tag{1}$$

На практике совместное распределение $P_{\mathcal{X},\mathcal{Y}}$ неизвестно, известна лишь реализация выборки $(x_i, y_i)_{i=1}^l$ из данного распределения. Поэтому вычислить точное значение функционала (1) не представляется возможным. На практике задача оптимизации функционала (1) заменяется приближенной задачей оптимизации эмпирического риска (2)

$$\int_{\mathbb{R}^2} (f(x) - y)^2 \hat{F}_l(dx,\, dy) = \sum_{i=1}^{l} (f(x_i) - y_i)^2 \underset{f\in\mathcal{F}}{\longrightarrow} \min \tag{2}$$

Необходимо построить стратегию $\mu(x_1, \ldots, x_l, y_1, \ldots, y_l) : \text{supp}\, X \times \text{supp}\, Y \to \mathcal{F}$, ставящую в соответствие каждой реализации выборки $(x_i, y_i)_{i=1}^l$ решение задачи оптимизации (2).

Решение приближенной задачи зависит не только от совместного распределения $P_{\mathcal{X},\mathcal{Y}}$, но и от реализации выборки $(x_i, y_i)_{i=1}^l$. Поэтому значение эмпирического риска, достигнутое на данном решении, не позволяет оценить качество стратегии μ на всей генеральной совокупности.

Для оценки качества стратегии μ на всей генеральной совокупности возможно использовать усредненный по выборкам среднеквадратичный риск (3).

$$L(\mu) \coloneqq \mathbb{E}_{X,Y}\left[\mathbb{E}_{x,y}\left(y - \mu(X, Y)(x)\right)^2\right] \tag{3}$$

В разделе 2 будет построено решение задачи (1). В разделе 3 будет представлено разложение усредненного среднеквадратичного риска $L(\mu)$, позволяющее сравнить качество решения приближенной задачи (2) с решением (1).

2 Оптимальная модель регрессии

В данном разделе будет построено решение задачи (1). По заданной совместной плотности $P_{\mathcal{X},\mathcal{Y}}$ необходимо выбрать функцию $f \in \mathcal{F}$, оптимизирующую функционал $\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathcal{Y})^2 \underset{f\in\mathcal{F}}{\longrightarrow} \min$.
Далее будем рассматривать класс интегрируемых с квадратом функций $\mathcal{F} = L_2$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathcal{Y})^2 = \\
&= \mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}) + \mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})^2 = \\
&= \mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))^2 + \mathbb{E}_{\mathcal{X},\mathcal{Y}}(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})^2 + \mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})
\end{aligned}$$

Рассмотрим

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y}) = \\
&= \mathbb{E}_{\mathcal{X}}\left[\mathbb{E}_{\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})|\mathcal{X}\right] = \\
&= \mathbb{E}_{\mathcal{X}}\left(\mathbb{E}_{\mathcal{Y}}\left[(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})|\mathcal{X}\right] \cdot (f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))\right) = \\
&= \mathbb{E}_{\mathcal{X}}\left([\mathbb{E}\mathcal{Y} - \mathbb{E}\mathcal{Y}] \cdot (f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))\right) = \\
&= 0
\end{aligned}$$

Следовательно, исходный функционал можно представить в виде суммы неотрицательных слагаемых:

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathcal{Y})^2 = \mathbb{E}_{\mathcal{X},\mathcal{Y}}(f(\mathcal{X}) - \mathbb{E}(\mathcal{Y}|\mathcal{X}))^2 + 2\mathbb{E}_{\mathcal{X},\mathcal{Y}}(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})^2 \tag{4}$$

Функционал (4) достигает минимума при выборе $f(x) = \mathbb{E}(\mathcal{Y}|\mathcal{X} = x)$.

Наименьшее значение ошибки: $\mathbb{E}_{\mathcal{X},\mathcal{Y}}(\mathbb{E}(\mathcal{Y}|\mathcal{X}) - \mathcal{Y})^2$

Как было сказано ранее, на практике совместное распределение $P_{\mathcal{X},\mathcal{Y}}$ неизвестно, известна реализация выборки $(x_i, y_i)_{i=1}^l$ из данного распределения, поэтому построить решение $f(X) = \mathbb{E}(Y|X)$ не представляется возможным. Можно построить решение, основанное на оценке совместного распределения $P_{\mathcal{X},\mathcal{Y}}$ по реализации выборки (например, `KDE`). Однако данные методы имеют высокую вычислительную сложность, а построение оценки совместного распределения может оказаться избыточным. Поэтому встает вопрос о сравнении произвольных моделей машинного обучения с наилучшим решением $f(X) = \mathbb{E}(Y|X)$.

3 Bias-Variance decomposition

В разделе 1 была поставлена приближенная задача (2). Было введено понятие стратегии $\mu : \text{supp}(X) \times \text{supp}(Y) \to \mathcal{F}$. Также был введен усредненный по выборкам среднеквадратичный риск $L(\mu) \coloneqq \mathbb{E}_{X,Y}\left[\mathbb{E}_{x,y}\left(y - \mu(X,Y)(x)\right)^2\right]$.

Далее будет построено разложение $L(\mu)$ на «шум», «смещение» и «разброс»

$$L(\mu) = \mathbb{E}_{X,Y}\Big[\underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y|x]\big)^2\Big]}_{\text{не зависит от } X} + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y|x] - \mu(X,Y)\big)^2\Big]\Big] =$$
$$= \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y|x]\big)^2\Big] + \mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}[y|x] - \mu(X,Y)\big)^2\Big]\Big].$$

Преобразуем второе слагаемое:

$$\mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}[y|x] - \mu(X,Y)\big)^2\Big]\Big] =$$
$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big] + \mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mu(X,Y)\big)^2\Big]\Big] =$$
$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\underbrace{\big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\big)^2}_{\text{не зависит от } X}\Big]\Big] + \mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mu(X,Y)\big)^2\Big]\Big] +$$
$$+ 2\mathbb{E}_{x,y}\Big[\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\big)\big(\mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mu(X,Y)\big)\Big]\Big].$$

Покажем, что последнее слагаемое обращается в нуль:

$$\mathbb{E}_{X,Y}\Big[\big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\big)\big(\mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mu(X,Y)\big)\Big] =$$
$$= \big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\big)\mathbb{E}_{X,Y}\Big[\mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mu(X,Y)\Big] =$$
$$= \big(\mathbb{E}[y|x] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\big)\Big[\mathbb{E}_{X,Y}\big[\mu(X,Y)\big] - \mathbb{E}_{X,Y}\big[\mu(X,Y)\big]\Big] =$$
$$= 0.$$

Тогда $L(\mu)$ можно представить в виде (5)

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[y - \mathbb{E}(y|x)\right]^2}_{\text{шум}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_{X,Y}\,\mu(X,Y) - \mathbb{E}(y|x)\right]^2}_{\text{смещение (bias)}} + \underbrace{\mathbb{E}_x\,\mathbb{E}_{X,Y}\left[\mu(X,Y) - \mathbb{E}_{X,Y}\,\mu(X,Y)\right]^2}_{\text{разброс(variance)}} \tag{5}$$

Из вида (5) следует, что $L(\mu)$ складывается из ошибки наилучшего решения (шума) и неотрицательных компонент `bias` и `variance`

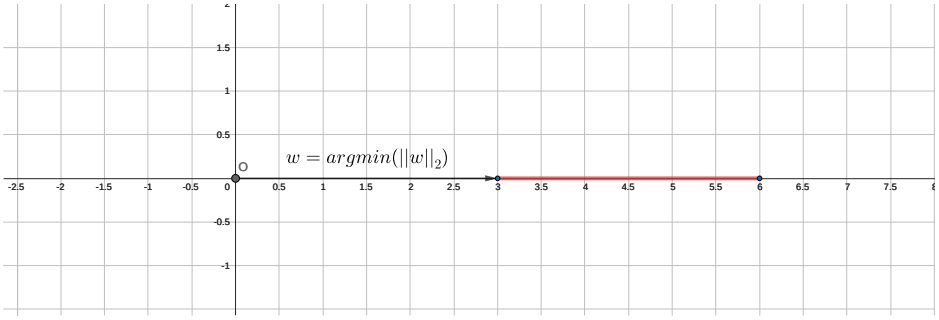
4 Bias-Variance tradeoff

В классической теории `bias` убывает с ростом числа параметров модели, `variance`, наоборот, возрастает при увеличении сложности модели. Однако с развитием глубокого обучения стали появляться модели с миллиардами параметров (например, `LLAMA2` имеет 70 миллиардов параметров). Существование таких моделей не укладывается в классическую теорию. Из-за этого была разработана новая теория([Bel+19]).

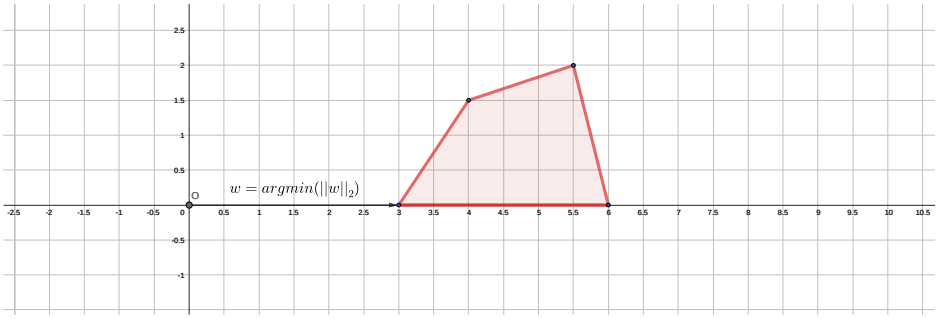
under-fitting При малом количестве параметров увеличение сложности приводит к улучшению качества, так как модель начинает выучивать новые зависимости, присутствующие на всей генеральной совокупности (а следовательно, на `train` и на `test`). Происходит значительное уменьшение `bias` при сравнительно небольшом росте `variance`. Это соответствует участку `under-fitted` на графике 2. Данный участок одинаково описывается и «классической» и «новой» теориями.

over-fitting При большом количестве параметров увеличение сложности приводит к ухудшению качества, так как модель начинает выучивать закономерности, присущие исключительно обучающей выборке. Происходит значительное увеличение `variance` при малом снижении `bias` (так как усреднение происходит не только по обучающей выборке). Это соответствует участку `over-fitted` на графике 2. В «классической» теории эмпирический риск монотонно возрастает, начиная с некоторого значения сложности. В «новой» теории данный участок ограничен.

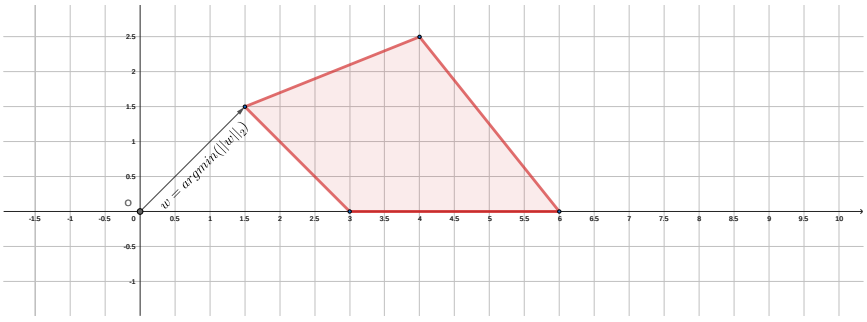
over-parameterized При очень большом количестве параметров появляется множество способов описания обучающей выборки. Регуляризация позволяет выбрать решение с минимальной нормой, следовательно, нет переобучения. (см. рисунок 1) Данный участок описывается только «новой» теорией. Начиная с некоторого значения сложности, эмпирический риск на тестовой выборке начинает улучшаться при неизменно низком эмпирическом риске на тренировочной выборке. На данном этапе основной вклад в эмпирический риск вносит шум.



(a) Одномерный случай



(b) Регуляризация не дает увеличению сложности ухудшить качество



(c) Регуляризация позволяет увеличению сложности улучшить качество

Рис. 1: Влияние регуляризации на качество модели при увеличении сложности модели

На рисунке 1 показано влияние регуляризации на модель при увеличении количества параметров. Если увеличение сложности не позволяет улучшить качество модели, то, благодаря регуляризации, будет выбран набор параметров, не учитывающий добавленную степень свободы (рис. 1b).

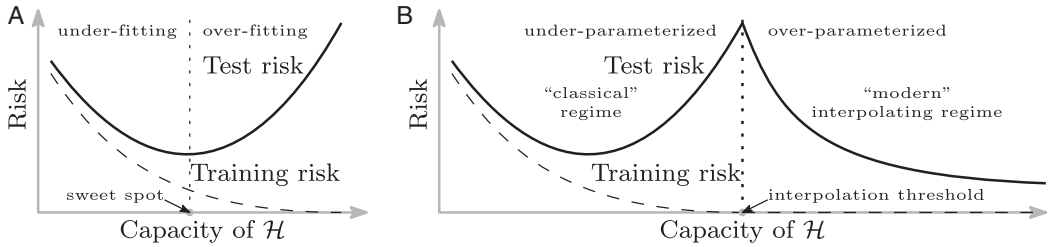


Рис. 2: Bias-variance tradeoff

Для демонстрации данного явления рассмотрим график, опубликованный в [Bel+19]. На рисунке 2A показана зависимость эмпирического риска от сложности модели (*capacity of* $\mathcal{H}$), предсказываемая классической теорией. На рисунке 2B показана аналогичная зависимость, предсказываемая «новой» теорией.

5 Задачи

Будем считать, что объекты x генерируются из нормального распределения $x \sim p(x) = \mathcal{N}(0, \sigma_1^2)$. Правильный ответ на объекте x определяется зашумленной функцией $f(x)$: $y = f(x) + \varepsilon$, $\varepsilon \sim p(\varepsilon) = \mathcal{N}(0, \sigma_2^2)$. Иными словами, $y \sim p(y|x) = \mathcal{N}(f(x), \sigma_2^2)$. Выборка X составляется из ℓ независимых пар (x_i, y_i). Мы будем рассматривать два простых частных случая: $f(x) = ax$, когда модель зависимости отвечает искомой зависимости, и $f(x)$ — четная функция, т. е. $f(-x) = f(x)$.

Задача 1. *Найти шумовую компоненту для одномерной линейной регрессии*

Решение. Так как распределение $p(y|x)$ нормальное, для него легко вычислить мат. ожидание:

$$\mathbb{E}[y|x] = f(x).$$

Тогда

$$\mathbb{E}_{x,y}\big[(y - \mathbb{E}[y|x])^2\big] = \mathbb{E}_{x,\varepsilon}\big[(f(x) + \varepsilon - f(x))^2\big] = \mathbb{E}_\varepsilon \varepsilon^2 = \mathbb{D}\varepsilon + \big(\mathbb{E}\varepsilon\big)^2 = \sigma_2^2 + 0 = \sigma_2^2.$$

□

Задача 2. *Найти смещение алгоритма одномерной линейной регрессии для* $f(x) = ax$ *и для произвольной четной* $f(x)$.

Решение. Для начала найдем «средний» по всем выборкам ответ алгоритма на объекте x:

$$\mathbb{E}_X[\mu(X)(x)] = \mathbb{E}_X[k(X)]\, x.$$

Итак, нам нужно найти «среднее» по всем выборкам значение коэффициента k:

$$\mathbb{E}_X[k(X)] = \int \frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2} \prod_i \big(p(x_i)p(\varepsilon_i)\big)\, dx_1 \dots dx_\ell\, d\varepsilon_1 \dots d\varepsilon_\ell. \tag{6}$$

Здесь записан несобственный интеграл, в котором каждая переменная принимает значения от $-\infty$ до ∞. Значение этого интеграла определяется функцией $f(x)$ и не всегда вычисляется аналитически. В случае, когда истинная зависимость в данных линейная, мы получим:

$$\mathbb{E}_X[k(X)] = \int \frac{\sum_i x_i(a\, x_i + \varepsilon_i)}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon} =$$

$$= a \int \frac{\sum_i x_i^2}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon} + \int \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon}.$$

Мы сократили обозначение для дифференциалов и для плотностей распределений. Первый интеграл равен a (интеграл по всему пространству от плотности распределения). Второй интеграл берется по симметричным относительно нуля интервалам от нечетной по x_i и по ε_i функции, а значит, равен 0. Итак, «средний» коэффициент равен a.
Найдем смещение:

$$\mathbb{E}_x\big[(\mathbb{E}_X[\mu(X)(x)] - \mathbb{E}[y|x])^2\big] = \mathbb{E}_x\big[(a\,x - a\,x)^2\big] = 0.$$

Это интуитивно понятный результат: логично, что перебрав все возможные выборки длины ℓ и усреднив по ним значение k, мы обязательно найдем истинную величину коэффициента.
Теперь найдем «среднее» k для произвольной четной $f(x)$. По аналогии с предыдущим случаем:

$$\mathbb{E}_X[k(X)] = \int \frac{\sum_i (x_i f(x_i))}{\sum_i x_i^2} p(\bar{x})d\bar{x} + \int \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon}.$$

Как мы уже выяснили, второй интеграл равен 0. Первый интеграл тоже равен 0, так как подынтегральное выражение — это нечетная по всем x_i функция. Итак, «среднее» значение коэффициента равно нулю. И это тоже понятный результат, потому что четную функцию логично приближать четной, а единственная четная линейная функция, проходящая через 0 — это $y = 0$.
Найдем смещение:

$$\mathbb{E}_x\big[(\mathbb{E}_X[\mu(X)(x)] - \mathbb{E}[y|x])^2\big] = \mathbb{E}_x\big[(0 - f(x))^2\big] = \mathbb{E}_x f^2(x).$$

Чем меньше четная функция $y = f(x)$ похожа на четную линейную $y = 0$, тем больше будет смещение. Таким образом, если мы пытаемся приблизить нелинейную функцию $f(x)$ в классе линейных, мы получаем большое смещение. Обратите внимание, что если бы $f(x)$ не была четной, мы бы не смогли просто аналитически вычислить интегралы. □

Задача 3. *Найти разброс алгоритма одномерной линейной регрессии для $f(x) = ax$ и для произвольной четной $f(x)$.*

Решение. Для $f(x) = ax$:

$$\mathbb{E}_x\big[\mathbb{E}_X\big[(\mu(X)(x) - \mathbb{E}_X[\mu(X)(x)])^2\big]\big] = \mathbb{E}_x\big[\mathbb{E}_X\big[(ax + \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}x - ax)^2\big]\big] =$$

$$= \Big(\mathbb{E}_x x^2\Big)\Big(\mathbb{E}_X\Big(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\Big)^2\Big) = \sigma_1^2\mathbb{E}_X\Big(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\Big)^2.$$

Мат. ожидание можно немного упростить, раскрыв квадрат суммы в числителе и внеся внутрь суммы мат. ожидание по $\bar{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_\ell)$:

$$\mathbb{E}_X\Big(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\Big)^2 = \mathbb{E}_{\bar{x}}\Big[\frac{\sum_{i\neq j} x_i x_j \mathbb{E}_{\bar{\varepsilon}}[\varepsilon_i\varepsilon_j] + \sum_i x_i^2\mathbb{E}_{\bar{\varepsilon}}\varepsilon_i^2}{(\sum_i x_i^2)^2}\Big].$$

Так как ε_i и ε_j независимы, $\mathbb{E}[\varepsilon_i\varepsilon_j] = 0$, а $\mathbb{E}_{\bar{\varepsilon}}\varepsilon_i^2 = \sigma_2^2$. Тогда

$$\mathbb{E}_X\Big(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\Big)^2 = \mathbb{E}_{\bar{x}}\Big[\frac{\sum_i x_i^2\sigma_2^2}{(\sum_i x_i^2)^2}\Big] = \sigma_2^2\mathbb{E}_{\bar{x}}\Big[\frac{1}{\sum_i x_i^2}\Big],$$

а разброс

$$\mathbb{E}_x\big[\mathbb{E}_X\big[(\mu(X)(x) - \mathbb{E}_X[\mu(X)(x)])^2\big]\big] = \sigma_1^2\sigma_2^2\mathbb{E}_{\bar{x}}\Big[\frac{1}{\sum_i x_i^2}\Big].$$

Последнее мат. ожидание также можно рассчитать, но мы не будем этого делать. Мы получили, что если шум в ответах небольшой, то и разброс модели будет небольшой.

Для четной $f(x)$:

$$\mathbb{E}_x\big[\mathbb{E}_X\big[(\mu(X)(x) - \mathbb{E}_X[\mu(X)(x)])^2\big]\big] = \mathbb{E}_x\big[\mathbb{E}_X\big[(0 - \frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}x)^2\big]\big] =$$

$$= \Big(\mathbb{E}_x x^2\Big)\Big(\mathbb{E}_X\Big(\frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}\Big)^2\Big) = \sigma_1^2\mathbb{E}_X\Big[\frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}\Big]^2.$$

□

Список литературы

[Bel+19] Mikhail Belkin и др. “Reconciling modern machine-learning practice and the classical bias–variance trade-off”. В: *Proceedings of the National Academy of Sciences* 116.32 (2019), с. 15849—15854.