Упражнение 3. Задача 1: Оценка обучающей и внутривыборочной ошибки

Даниил Ковех

2025-10-28

Содержание

1. Теория

Формулируем модель кратко. Матрица признаков $X \in \mathbb{R}^{N \times p}$ детерминирована. Наблюдения $y \in \mathbb{R}^N$ порождает линейная регрессия

$$y = X\beta + \varepsilon,$$

где $\mathbb{E}[\varepsilon] = 0$ и $\text{Var}(\varepsilon) = \sigma^2 I_N$.

Оценка МНК на обучающей выборке T:

$$\hat{\beta}_T = (X^\top X)^{-1} X^\top y_T.$$

Цель: найти математические ожидания трёх величин.

1. Обучающая ошибка: $\mathbb{E}_T \left[\frac{1}{N} (y_T - X\hat{\beta}_T)^\top (y_T - X\hat{\beta}_T) \right]$.
2. Внутривыборочная ошибка: $\mathbb{E}_T \mathbb{E}_y \left[\frac{1}{N} (y - X\hat{\beta}_T)^\top (y - X\hat{\beta}_T) \mid T \right]$.
3. Разница между этими величинами.

Ключевые матрицы:

- Проекционная матрица $\hat{H} = X(X^\top X)^{-1} X^\top$.
- Матрица остатков $I_N - \hat{H}$.

Свойства: $\hat{H}$ симметрична и идемпотента ($\hat{H}^2 = \hat{H}$). Матрица остатков тоже симметрична и идемпотента.

2. Жизненный пример

Представь маркетинг-аналитика Машу. У неё есть N клиентов и p признаков (возраст, доход, история покупок). Маша строит линейную модель, чтобы объяснить средние расходы клиента.

Обучающая ошибка отвечает на вопрос: как точно модель объясняет текущих клиентов. В ней зарыта “самодовольная” оценка качества — мы смотрим на те же данные, которые использовали для подгонки коэффициентов.

Внутривыборочная ошибка — это взгляд вперёд. Предполагаем, что придут клиенты с такими же признаками, но с новой случайной ошибкой в тратах. Мы берём среднее по новым реализациям ответа и получаем честную оценку. В этой величине сидит два источника неопределённости: шум в новом наблюдении и шум, влияющий на оценку коэффициентов. Аналитик Маша знает: модель обычно ошибается на текущих данных меньше, чем на будущих клиентах. Разница между этими величинами — та самая “оптимистичность” оценки.

3. Академическое решение

3.1. Ожидаемая обучающая ошибка

Остатки на обучении:

$$e_T = y_T - X\hat{\beta}_T = (I_N - \hat{H})y_T = (I_N - \hat{H})(X\beta + \varepsilon_T).$$

Поскольку $(I_N - \hat{H})X = 0$, получаем $e_T = (I_N - \hat{H})\varepsilon_T$.

Тогда

$$\frac{1}{N}\mathbb{E}_T[e_T^\top e_T] = \frac{1}{N}\mathbb{E}_T[\varepsilon_T^\top (I_N - \hat{H})\varepsilon_T].$$

Используем тождество: $\mathbb{E}[\varepsilon^\top A\varepsilon] = \sigma^2 \mathrm{tr}(A)$ для симметричной матрицы A и $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Получаем

$$\frac{1}{N}\mathbb{E}_T[e_T^\top e_T] = \frac{\sigma^2}{N}\mathrm{tr}(I_N - \hat{H}) = \sigma^2 \frac{N-p}{N}.$$

3.2. Ожидаемая внутрирыборочная ошибка

Новая выборка $y = X\beta + \varepsilon$ независима от тренировки. Ошибка предсказания на той же сетке X:

$$y - X\hat{\beta}_T = \underbrace{\varepsilon}_{\text{новый шум}} - \underbrace{X(\hat{\beta}_T - \beta)}_{\text{ошибка модели}}.$$

Условное ожидание по новому шуму:

$$\mathbb{E}_y\left[(y - X\hat{\beta}_T)^\top (y - X\hat{\beta}_T) \mid T\right] = \mathbb{E}_y\left[\varepsilon^\top \varepsilon\right] + \|X(\hat{\beta}_T - \beta)\|_2^2,$$

поскольку перекрёстный член исчезает.

Первое слагаемое равно $N\sigma^2$. Второе: выразим ошибку коэффициентов

$$\hat{\beta}_T - \beta = (X^\top X)^{-1} X^\top \varepsilon_T.$$

Тогда

$$\|X(\hat{\beta}_T - \beta)\|_2^2 = \varepsilon_T^\top \hat{H}\varepsilon_T.$$

Берём ожидание по обучающему шуму:

$$\mathbb{E}_T\left[\varepsilon_T^\top \hat{H}\varepsilon_T\right] = \sigma^2 \mathrm{tr}(\hat{H}) = \sigma^2 p.$$

Итого

$$\mathbb{E}_T\mathbb{E}_y\left[\frac{1}{N}(y - X\hat{\beta}_T)^\top (y - X\hat{\beta}_T) \mid T\right] = \frac{1}{N}(N\sigma^2 + \sigma^2 p) = \sigma^2\left(1 + \frac{p}{N}\right).$$

3.3. Разница между ошибками

Вычитаем:

$$\sigma^2\left(1+\frac{p}{N}\right)-\sigma^2\frac{N-p}{N}=\frac{2\sigma^2 p}{N}.$$

Эта величина называется оптимизмом: модель на обучающей выборке кажется на $\frac{2\sigma^2 p}{N}$ лучше, чем на будущих данных.

3.4. Проверяем вычисления в симуляции

```
set.seed(202501)

simulate_gap <- function(n = 250, p = 5, sigma = 1, reps = 5000) {
  X <- matrix(rnorm(n * p), nrow = n, ncol = p)
  XtX_inv <- solve(t(X) %*% X)
  hat_matrix <- X %*% XtX_inv %*% t(X)
  optimism_estimates <- numeric(reps)
  train_error_estimates <- numeric(reps)
  insample_error_estimates <- numeric(reps)

  for (r in seq_len(reps)) {
    eps_train <- rnorm(n, sd = sigma)
    y_train <- X %*% rep(0, p) + eps_train
    beta_hat <- XtX_inv %*% t(X) %*% y_train

    residuals <- y_train - X %*% beta_hat
    train_error_estimates[r] <- mean(residuals^2)

    eps_new <- rnorm(n, sd = sigma)
    y_new <- X %*% rep(0, p) + eps_new
    insample_error_estimates[r] <- mean((y_new - X %*% beta_hat)^2)

    optimism_estimates[r] <- insample_error_estimates[r] - train_error_estimates[r]
  }

  list(
    train_error = mean(train_error_estimates),
    insample_error = mean(insample_error_estimates),
    optimism = mean(optimism_estimates)
  )
}

result <- simulate_gap(n = 200, p = 8, sigma = 1.1, reps = 2000)
result
```

```
## $train_error
## [1] 1.157763
##
## $insample_error
## [1] 1.259242
##
## $optimism
## [1] 0.1014783
```

Теория предсказывает:

$$\mathbb{E}[\text{train}] = \sigma^2 \frac{N-p}{N}, \qquad \mathbb{E}[\text{in-sample}] = \sigma^2 \left(1 + \frac{p}{N}\right), \qquad \text{gap} = \frac{2\sigma^2 p}{N}.$$

```
sigma <- 1.1
n <- 200
p <- 8

theoretical_train <- sigma^2 * (n - p) / n
theoretical_insample <- sigma^2 * (1 + p / n)
theoretical_gap <- 2 * sigma^2 * p / n

c(
  train = theoretical_train,
  insample = theoretical_insample,
  optimism = theoretical_gap
)
```

```
##    train insample optimism
##   1.1616   1.2584   0.0968
```

Симуляция подтверждает формулы. Разница между средней внутривыборочной и обучающей ошибкой устремляется к $\frac{2\sigma^2 p}{N}$.