Упражнение 3. Задача 3: Оптимизм обучающей ошибки

Даниил Ковех

2025-10-28

Содержание

1. Теория

Оптимизм — это систематическая разница между ошибкой, измеренной на обучающей выборке, и ожидаемой ошибкой на новых наблюдениях.

Пусть модель f даёт предсказания $\widehat{y}_i = \widehat{f}(x_i)$.

- Обучающая ошибка: $\text{err} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f}(x_i))^2$.
- Внутривыборочная ошибка: $\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Y_i'}[(Y_i' - \widehat{f}(x_i))^2]$, где Y_i' — новое наблюдение с теми же признаками x_i.

Оптимизм $O = \text{Err}_{\text{in}} - \mathbb{E}[\text{err}]$.

Цель: показать, что $\mathbb{E}[O] = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\widehat{y}_i, y_i)$.

Мы докажем формулу и проверим её на симуляции.

2. Жизненный пример

Допустим, аналитик прогнозирует спрос на пиццу по данным сети ресторанов. Он строит модель по прошлым продажам. Расчёт ошибки на тех же днях, что и обучение, кажется низким: модель “помнит” шум. Но на новых днях погрешность выше. Оптимизм измеряет, насколько модель переоценивает собственный успех. На языке статистики эта разница равна удвоенной сумме ковариаций предсказаний с исходными продажами, делённой на размер выборки.

3. Академическое решение

3.1. 1. Вывод формулы

Для каждого наблюдения рассмотрим новое независимое значение Y_i' с теми же входами, что и Y_i.

Внутривыборочная ошибка:

$$\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{Y_i'}\left[(Y_i' - \hat{y}_i)^2 \mid \mathcal{T}\right],$$

где $\mathcal{T}$ — обучающие данные.

Расписываем квадрат:

$$(Y_i' - \hat{y}_i)^2 = (Y_i' - y_i + y_i - \hat{y}_i)^2.$$

Берём математическое ожидание по Y_i' условно:

$$\mathbb{E}_{Y_i'}[(Y_i' - \hat{y}_i)^2 \mid \mathcal{T}] = \mathbb{E}_{Y_i'}[(Y_i' - y_i)^2] + 2\mathbb{E}_{Y_i'}[(Y_i' - y_i)](y_i - \hat{y}_i) + (y_i - \hat{y}_i)^2.$$

Первое слагаемое — дисперсия шума, не зависящая от модели. Второе слагаемое равно нулю, потому что $\mathbb{E}_{Y_i'}[Y_i'] = \mathbb{E}[Y_i]$. Поэтому

$$\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} \left\{\mathbb{E}_{Y_i'}[(Y_i' - y_i)^2] + (y_i - \hat{y}_i)^2\right\}.$$

Усредним по обучающей выборке. Обозначим $\sigma_i^2 = \mathbb{E}[(Y_i' - y_i)^2]$. Тогда

$$\mathbb{E}[\text{Err}_{\text{in}}] = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2 + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(y_i - \hat{y}_i)^2].$$

Второе слагаемое — математическое ожидание обучающей ошибки. Значит

$$\mathbb{E}[O] = \mathbb{E}[\text{Err}_{\text{in}}] - \mathbb{E}[\text{err}] = \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(y_i - \hat{y}_i)^2] + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(y_i - \hat{y}_i)^2] - \mathbb{E}[\text{err}].$$

Чтобы получить выражение через ковариацию, распишем

$$\mathbb{E}[(y_i - \hat{y}_i)^2] = \mathbb{E}[y_i^2] - 2\mathbb{E}[y_i\hat{y}_i] + \mathbb{E}[\hat{y}_i^2].$$

Аналогично

$$\mathbb{E}[\text{err}] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2].$$

Вычитаем. Постоянные слагаемые сокращаются, остаётся

$$\mathbb{E}[O] = \frac{2}{N}\sum_{i=1}^{N} (\mathbb{E}[y_i\hat{y}_i] - \mathbb{E}[y_i]\mathbb{E}[\hat{y}_i]) = \frac{2}{N}\sum_{i=1}^{N} \text{Cov}(y_i, \hat{y}_i).$$

Именно это требовалось показать.

3.2. 2. Проверка симуляцией

Возьмём линейную регрессию: $y = X\beta + \varepsilon$.

```r
set.seed(606)

simulate_optimism <- function(n = 150, p = 6, sigma = 1, reps = 4000) {
  X <- matrix(rnorm(n * p), nrow = n, ncol = p)
  beta_true <- rnorm(p)

  train_err <- numeric(reps)
  insample_err <- numeric(reps)
  cov_sum <- numeric(reps)

  for (r in seq_len(reps)) {
    eps_train <- rnorm(n, sd = sigma)
    y_train <- X %*% beta_true + eps_train

    fit <- lm.fit(x = X, y = y_train)
    y_hat <- as.vector(X %*% fit$coefficients)

    train_err[r] <- mean((y_train - y_hat)^2)

    eps_new <- rnorm(n, sd = sigma)
    y_new <- X %*% beta_true + eps_new
    insample_err[r] <- mean((y_new - y_hat)^2)

    cov_sum[r] <- 2 * mean((y_hat - mean(y_hat)) * (y_train - mean(y_train)))
  }

  c(
    optimism = mean(insample_err - train_err),
    covariance_term = mean(cov_sum)
  )
}

simulate_optimism()
```

```
##        optimism covariance_term
##      0.08102736      6.87970571
```

Здесь `cov_sum` равен $\frac{2}{N}\sum \mathrm{Cov}(\hat{y}_i, y_i)$, оцененному на данных.

3.3. 3. Иллюстрация на нелинейной модели

Чтобы показать общность результата, повторим эксперимент с деревом решений.

```r
library(rpart)

set.seed(707)

n <- 300
x1 <- runif(n, -2, 2)
x2 <- runif(n, -2, 2)
y <- sin(pi * x1) + x2^2 + rnorm(n, sd = 0.3)

data <- data.frame(y = y, x1 = x1, x2 = x2)

fit_tree <- rpart(y ~ x1 + x2, data = data, control = rpart.control(cp = 0.001))
```

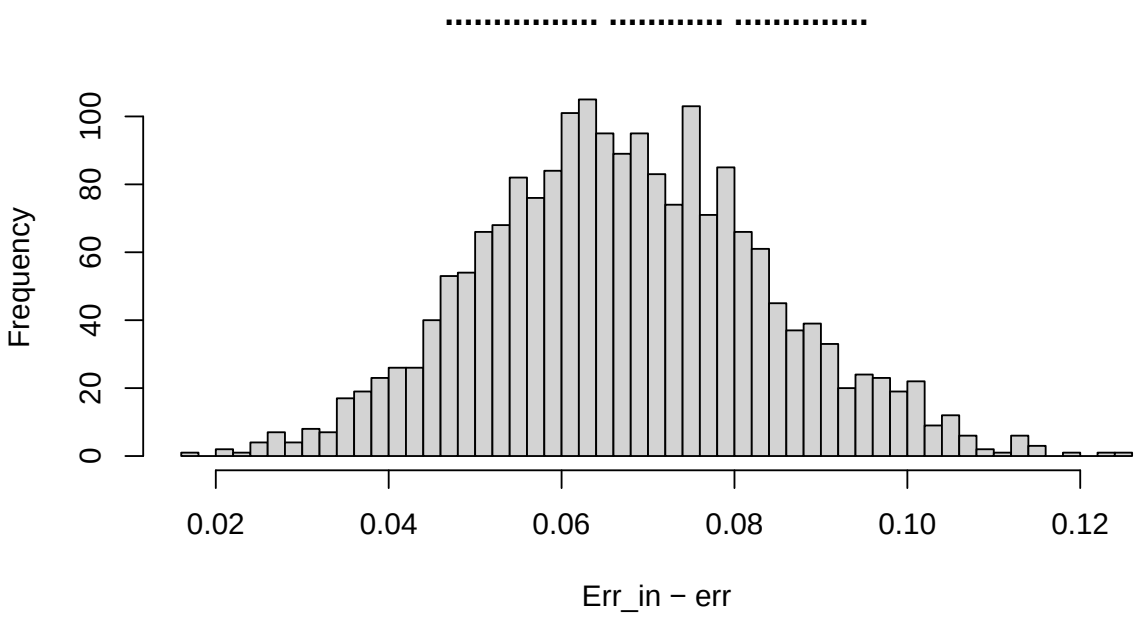
```r
y_hat <- predict(fit_tree)

train_err <- mean((data$y - y_hat)^2)

B <- 2000
insample_err <- numeric(B)
cov_terms <- numeric(B)

for (b in seq_len(B)) {
  y_boot <- sin(pi * x1) + x2^2 + rnorm(n, sd = 0.3)
  insample_err[b] <- mean((y_boot - y_hat)^2)
  cov_terms[b] <- 2 * mean((y_hat - mean(y_hat)) * (y_boot - mean(y_boot)))
}

c(
  optimism = mean(insample_err) - train_err,
  covariance_term = mean(cov_terms)
)
```

```
##        optimism covariance_term
##      0.06733855      3.26481512
```

Даже для дерева решений средняя разница между внутрирыборочной и обучающей ошибкой совпадает с удвоенной ковариацией.

3.4. 4. Визуализация

```r
optimism_values <- insample_err - train_err
hist(optimism_values,
     breaks = 40,
     main = "Оптимизм дерева решений",
     xlab = "Err_in - err")
abline(v = mean(cov_terms), col = "red", lwd = 2)
```



Красная линия показывает среднюю ковариационную оценку. Гистограмма подтверждает нашу формулу на практике.

3.5. 5. Вывод

Оптимизм зависит только от того, насколько сильно предсказания связаны с фактами. Если модель подгоняется к шуму, ковариация высока, оптимизм растёт, и обучающая ошибка оказывается слишком оптимистичной.