Exercise 4. Task 2: Permutation Importance in Linear Regression

Daniil Koveh

2025-11-04

Содержание

1. Теория

Рассмотрим стандартизованную линейную регрессию $y = X\beta + \varepsilon$, где признаки и отклик имеют нулевое среднее и единичную дисперсию. Пусть $\hat{\beta}$ — оценка МНК. Возьмём признак x_j и перетасуем его значения, не меняя остальные. Обозначим через RSS среднюю квадратичную ошибку на обучении до перестановки, а через RSS_j^* — после перестановки при фиксированных коэффициентах $\hat{\beta}$.

Перестановка уничтожает ковариацию между x_j и y, поэтому изменяется только часть ошибки, зависящая от $\hat{\beta}_j$. Формально:

$$\mathrm{RSS} = \frac{1}{N}\|y - X\hat{\beta}\|^2 = \frac{1}{N}\sum_{k=1}^{p}\hat{\beta}_k^2,$$

$$\mathrm{RSS}_j^* = \mathrm{RSS} + \frac{2}{N}\hat{\beta}_j^2,$$

где ожидание по перестановкам даёт:

$$E_P[\mathrm{RSS}_j^* - \mathrm{RSS}] = 2\hat{\beta}_j^2.$$

Таким образом, прирост ошибки после перемешивания признака равен удвоенному квадрату коэффициента. Это классическая база для permutation importance в линейных моделях.

2. Жизненный пример

Представьте, что вы предсказываете итоговый балл студента по нормализованным оценкам экзаменов. Если «перемешать» столбец с математикой, корреляция между математикой и итоговым баллом исчезнет, и ошибка возрастёт. Компонент увеличения ошибки будет пропорционален квадрату веса математики в регрессии: чем важнее был предмет, тем сильнее модель теряет точность. Менее значимые предметы при перемешивании дают минимальный рост ошибки.

3. Академическое решение

3.1. План решения

- Формализуем эффект перестановки отдельного признака в стандартизованной линейной модели и выводим аналитическую формулу прироста ошибки.
- Реализуем симуляцию, где многократно перемешиваем столбец и измеряем средний рост RSS, удерживая коэффициенты фиксированными.
- Сравниваем эмпирику с формулой $2\hat{\beta}_j^2$, чтобы убедиться, что permutation importance совпадает с теорией в линейном случае.

3.2. Аналитические рассуждения

- Поскольку перемешивание разрушает только связь для x_j, остальные части модели не меняются.
- При стандартизации x_j и y получаем простое выражение для RSS через квадраты коэффициентов.
- Используя свойство перестановок, $E_P[x_j y] = 0$ и $E_P[x_j x_k] = 0$ для $k \neq j$, получаем ожидаемый прирост ошибки $2\hat{\beta}_j^2$.

3.3. Симуляционная проверка

```r
set.seed(20250410)
N <- 200; p <- 5
X <- scale(matrix(rnorm(N * p), nrow = N))
beta_true <- runif(p, -1, 1)
y <- scale(X %*% beta_true + rnorm(N, sd = 0.5))

perm_importance(X, y, j = 3)
```

```
##   empirical analytic.x3
##   0.7001106   0.7057437
```

```r
set.seed(20250410) # повторно фиксируем генератор
N <- 200 # размер выборки
p <- 5 # число признаков
X <- scale(matrix(rnorm(N * p), nrow = N)) # генерируем стандартизированный X
beta_true <- runif(p, -1, 1) # задаём истинные коэффициенты
y <- scale(X %*% beta_true + rnorm(N, sd = 0.5)) # формируем отклик
beta_hat <- lm.fit(X, y)$coefficients # оцениваем коэффициенты
2 * beta_hat[3]^2 # аналитический прирост для j = 3
```

```
##        x3
## 0.7057437
```

Симуляция демонстрирует совпадение с аналитическим выражением: среднее увеличение ошибки при перестановке равно $2\hat{\beta}_j^2$, что подтверждает формулу permutation importance для линейной модели.

3.4. Что запомнить

- В стандартизованной регрессии прирост MSE после перемешивания признака равен $2\hat{\beta}_j^2$ — быстрый способ оценить важность без повторного обучения.

- Permutation importance в линейной модели полностью определяется оценённым коэффициентом и не требует дополнительных вычислительных затрат.
- Если перемешивание не повышает ошибку, значит коэффициент близок к нулю, и признак мало влияет на отклик.