Exercise 4. Task 1: Bootstrap Correlation of the Sample Mean

Daniil Koveh

2025-11-04

Содержание

1. Теория

Пусть $x_1, \dots, x_N$ — независимые одинаково распределённые случайные величины с математическим ожиданием μ и дисперсией σ^2. Обозначим $\bar{x}$ — выборочное среднее. В бутстрепе мы генерируем выборки с возвращением из наблюдений $x_1, \dots, x_N$, поэтому каждое бутстреп-среднее $\bar{x}^*$ можно представить как линейную комбинацию x_i с коэффициентами w_i — долями попаданий элемента i в бутстреп-выборку. Свойства:

$$E[w_i] = \frac{1}{N}, \quad \mathrm{Var}(w_i) = \frac{N-1}{N^2}, \quad \mathrm{Cov}(w_i, w_j) = -\frac{1}{N^2} \text{ для } i \neq j.$$

Используя эти результаты, выводим:

$$\mathrm{Var}(\bar{x}^*) = \frac{2N-1}{N^2}\sigma^2,$$

а ковариация двух независимых бутстреп-средних $\bar{x}_1^*$ и $\bar{x}_2^*$ выражается как:

$$\mathrm{Cov}(\bar{x}_1^*, \bar{x}_2^*) = \frac{N-1}{N^2}\sigma^2.$$

Среднее по мешку (bagged mean) определяется как $\bar{x}_{\text{bag}} = \frac{1}{B}\sum_{b=1}^{B} \bar{x}^{*(b)}$. При $B \to \infty$ дисперсия $\bar{x}_{\text{bag}}$ стремится к $\mathrm{Var}(\bar{x}) = \sigma^2/N$, то есть не уменьшается по сравнению с исходной оценкой. Следовательно, корреляция бутстрепных оценок:

$$\mathrm{Cor}(\bar{x}_1^*, \bar{x}_2^*) = \frac{\mathrm{Cov}(\bar{x}_1^*, \bar{x}_2^*)}{\sqrt{\mathrm{Var}(\bar{x}_1^*)\mathrm{Var}(\bar{x}_2^*)}} = \frac{N-1}{2N-1} \approx \frac{N}{2N} = \frac{1}{2}.$$

Таким образом, выборочные средние из бутстрепов имеют корреляцию порядка 50%, а бутстрепирование не уменьшает дисперсию линейных статистик.

2. Жизненный пример

Представьте, что вы оцениваете среднюю сумму чеков покупателей за день. У вас есть N покупателей, и вы хотите понять разброс средней суммы. Бутстреп помогает, но поскольку среднее — линейная статистика, разные бутстреп-реплики сильно похожи друг на друга: они делят одни и те же чеки между собой. Корреляция примерно 0.5 означает, что две бутстреп-оценки среднего «договорились» наполовину — информация не обновляется полностью. Поэтому мешочный подход (bagging) не даст выигрыша для среднего, а вот для нелинейных статистик (медиана, квантиль, дерево решений) эффект уже есть.

3. Академическое решение

3.1. План решения

- Представляем бутстреп-реплики среднего через веса мультиномиального распределения, чтобы аналитически вывести дисперсии и ковариации.
- Показываем, что мешочное усреднение для линейной статистики не уменьшает дисперсию, и вычисляем корреляцию между двумя бутстреп-средними.
- Проверяем формулы на моделировании, сравнивая эмпирические оценки с теоретическими значениями и делая выводы о практическом смысле.

3.2. Формальные выводы

- Представление бутстреп-среднего через веса: $\bar{x}^* = \sum_{i=1}^N w_i x_i$, где $w_i \sim \text{Multinomial}(N; 1/N, \dots, 1/N)$.
- Используем свойства мультиномиального распределения:

$$\text{Var}(\bar{x}^*) = \sum_{i=1}^{N} \text{Var}(w_i) x_i^2 + 2 \sum_{i<j} \text{Cov}(w_i, w_j) x_i x_j.$$

  После взятия математического ожидания по данным получаем $\text{Var}(\bar{x}^*) = \frac{2N-1}{N^2}\sigma^2$.
- Аналогично:

$$\text{Cov}(\bar{x}_1^*, \bar{x}_2^*) = \frac{N-1}{N^2}\sigma^2.$$

- Следовательно, корреляция $\text{Cor}(\bar{x}_1^*, \bar{x}_2^*) = \frac{N-1}{2N-1}$ и стремится к $1/2$ при больших N.
- Дисперсия мешочного среднего после усреднения не меньше, чем у первоначального $\bar{x}$, подтверждая, что bagging не улучшает линейные статистики.

3.3. Симуляционная проверка

```r
set.seed(20250410)
bootstrap_simulate()
```

```
##          corr      var_boot       var_bag
## -7.595539e-04  1.112024e-02  2.205742e-06
```

```r
analytic_corr <- (200 - 1) / (2 * 200 - 1) # аналитическая корреляция
analytic_var_boot <- (2 * 200 - 1) / (200^2) * 1.5^2 # теоретическая дисперсия
analytic_bag_var <- 1.5^2 / 200 # дисперсия исходного среднего
c(analytic_corr = analytic_corr,
  analytic_var_boot = analytic_var_boot,
  analytic_bag_var = analytic_bag_var) # сравнение с аналитикой
```

```
##     analytic_corr analytic_var_boot  analytic_bag_var
##        0.49874687        0.02244375        0.01125000
```

Симуляция подтверждает теоретические выводы:

- Эмпирическая корреляция близка к 0.5.
- Дисперсия бутстреп-среднего соответствует формуле $\frac{2N-1}{N^2}\sigma^2$.
- Дисперсия bagging совпадает с σ^2/N, так что линейные статистики не выигрывают от мешочного усреднения.

3.4. Что запомнить

- Бутстреп для линейных функционалов возвращает сильно коррелированные реплики, поэтому мешочное усреднение не даёт выигрыша.
- Корреляция порядка $1/2$ и дисперсия $\frac{2N-1}{N^2}\sigma^2$ — полезные референсы для оценки устойчивости среднего.
- Усреднение полезно для нелинейных оценок (деревья, медиана), но не для простых линейных статистик.