Моделирование выборок. Бутстреп

Моделирование выборок из совокупностей. Бутстреп: построение доверительных интервалов, проверка гипотез. Метод складного ножа и кросс-валидация.

ПМИ ФКН ВШЭ, 30 ноября 2019 г.

Денис Деркач[1], Алексей Артёмов[1,2]

[1] ФКН ВШЭ [2] Сколтех

Содержание лекции

- Алгоритмы моделирования выборок из совокупностей
- Резервуарная выборка

- Метод складного ножа
- Бутстреп и кросс-валидация
- Бутстреп для построения доверительных интервалов
- Теоретические свойства бутстрепа
- Примеры

Моделирование выборок из больших совокупностей

Моделирование выборок

- Проблема: как выбрать k записей из n имеющихся (без возвращения)?
- Простое случайное сэмплирование `[Meng, 2013]` — это моделирование выборок k различных элементов из совокупности размера n таким образом, что любое подмножество размера k имеет одинаковую вероятность быть выбранным.
- Обзор масштабируемых алгоритмов сэмплирования в Meng, X. (2013, February). Scalable simple random sampling and stratified sampling. In International Conference on Machine Learning (pp. 531-539).

Моделирование выборок

- Пусть $S = \{s_1, \ldots, s_n\}$ — совокупность из n элементов, а $\mathcal{S}_k = \{A \subset S : |A| = k\}$ — множество всех подмножеств S, состоящих из k элементов ($k \leqslant n$)
- По определению, необходимо равновероятно выбрать элемент из $\mathcal{S}_k$, что затруднительно реализовать переборным алгоритмом, так как требуется создать $\binom{n}{k}$ элементов и выбрать всего один
- Необходимо выбрать $A \in \mathcal{S}_k$ без перечисления элементов $\mathcal{S}_k$
- Наивный метод (вы все его знаете):

Моделирование выборок

- Пусть $S = \{s_1, \ldots, s_n\}$ — совокупность из n элементов, а $\mathcal{S}_k = \{A \subset S : |A| = k\}$ — множество всех подмножеств S, состоящих из k элементов ($k \leqslant n$)
- По определению, необходимо равновероятно выбрать элемент из $\mathcal{S}_k$, что затруднительно реализовать переборным алгоритмом, так как требуется создать $\binom{n}{k}$ элементов и выбрать всего один
- Необходимо выбрать $A \in \mathcal{S}_k$ без перечисления элементов $\mathcal{S}_k$
- Наивный метод (вы все его знаете):
  - Выбрать $s_i \in S$
  - Удалить s_i из S
  - Повторять k раз до получения набора $A \in \mathcal{S}_k$
- Нужен случайный доступ в S и удаление случайного $s_i \in S$

Выбор и отклонение

Algorithm 1: Выбор и отклонение `[Fan 1962]`

Set $i = 0$;
for j from 1 to n **do**
  With probability $p_j(i) = \frac{k-i}{n-j+1}$, select s_j and let $i = i + 1$;
end

- Требуется последовательное считывание $s_j \in S$
- Последовательный метод: выбор или невыбор s_j влияет на следующие решения ($p_j = p_j(i)$)
- Не поддерживает параллелизм
- Что, если n неизвестно до конца считываемого потока S (однако задано k или $p = k/n$)?

Случайная сортировка

Algorithm 2: Случайная сортировка `[Sunter 1977]`

Associate each item of S with an independent variable
$X_i \sim U(0, 1)$.;
Sort S in ascending order.;
Select the smallest k items.;

- Сортировка теперь требует $\mathcal{O}(n \log n)$ операцией
- Но она может эффективно выполняться параллельно
- На самом деле необходимо выбирать k наименьших элементов из n, что можно выполнить за линейное время

Резервуарная выборка

- Проблема с объемом совокупности: n очень велико
- Проблема с объемом совокупности: n неизвестно заранее: `The problem arises, for example, in sampling records that reside on a magnetic tape of indeterminate length [Viters 1985].`
- Решение Vitter, J. S. (1985). Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS), 11(1), 37-57.

Резервуарная выборка: псевдокод

```
/* S has items to sample, R will contain the result */
ReservoirSample(S[1..n], R[1..k])
  // fill the reservoir array
  for i = 1 to k
      R[i] := S[i]

  // replace elements with gradually
  // decreasing probability
  for i = k + 1 to n
    j := random(1, i)   // important: inclusive range
    if j <= k
        R[j] := S[i]
```

Резервуарная выборка: объяснение

- Для всех i, i-ый элемент S включается в резервуар с вероятностью $\frac{k}{i} = \mathrm{P}(j \leqslant k | j \sim U(1, i))$
- На каждой итерации j-й элемент резервуара заменяется с вероятностью $\frac{1}{k} \times \frac{k}{i} = \frac{1}{i}$
- Когда алгоритм закончит работу, каждый элемент в S имеет одинаковую вероятность $\frac{k}{|S|}$, что его выберут в резервуар
- Доказательство (индукция по i):
    - После $(i-1)$-го шага вероятность элемента попасть в резервуар равна $p_{i-1} = \frac{k}{i-1}$
    - Вероятность выживания элемента (не быть замененным) на i-ом шаге $q_i = \frac{i-1}{i}$
    - После $(i-1)$-го шага вероятность элемента попасть в резервуар равна $p_i = p_{i-1} \times q_i = \frac{k}{i-1} \times \frac{i-1}{i} = \frac{k}{i}$

Резервуарная выборка: свойства

- Требуется последовательное считывание $s_j \in S$, что долго для больших объемов записей
- Не поддерживает параллелизм в исходном виде
- Требуется резервуар R с возможностью случайной записи
- Не требуется знание n!

Бутстреп: напоминание

Бутстреп: напоминание

- Имеем:
  - $F(x)$ — (истинное) распределение данных
  - $\widehat{F}(x)$ — выборочное распределение данных
  - $T = T(F)$ — интересующая нас величина (неизвестный параметр)
- Нас интересует оценка статистических характеристик T, и вообще — его распределения $\text{Law}(T)$
- Раз $\widehat{F}(x)$ — оценка $F(x)$, а $T = T(F)$, заменим:

$$T = T(F) \quad \to \quad T = T(\widehat{F})$$

Например, если $T(F) = \int x dF(x)$ (т.е. среднее значение), то $T(\widehat{F}) = \int x d\widehat{F}(x) = \frac{1}{n} \sum_i X_i$

Метод складного ножа

Метод складного ножа

- Пусть $T_n = (X_1, \ldots, X_n)$.
- Рассмотрим n подвыборок:
  $T_{n-1}^i = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.
- Пусть $\overline{T}_n = n^{-1} \sum_{i=1}^n T_{n-1}^i$.
- Построим следующую оценку $\mathrm{V}(T_n)$:

$$v_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} (T_{n-1}^i - \overline{T}_n)^2$$

- Тогда оценка стандартной ошибки по методу складного ножа имеет вид $\widehat{se}_{jack} = \sqrt{v_{jack}}$.
- Может быть показано, что $v_{jack} / \mathrm{V}(T_n) \xrightarrow{\mathrm{P}} 1$.

Метод складного ножа

1. Бутстреп — рандомизированная аппроксимация delete-m метода складного ножа.
2. Бутстреп — разные результаты, метод складного ножа — всегда одинаковые
3. Оценка дисперсии функционала vs. оценка всего распределения.
4. Условия гладкости функционала.
5. Метод складного ножа проще применять для сложных схем семплирования.

Бутстреп
и кросс-валидация

Перекрестная проверка

[Слайды MLHEP'19]

Различия между бутстрепом и CV

- Перекрестная проверка с разбиением на k блоков (K-Fold CV) выдает k подвыборок
- Бутстреп тоже выдает k подвыборок (немного другим образом)
- В чем же различия?
- https://datascience.stackexchange.com

Различия между бутстрепом и CV

- Бутстреп моделирует выборки с возвращением (некоторых записей из оригинальной совокупности может быть несколько, а некоторых не быть вовсе)
- Кросс-валидация моделирует выборки без возвращения систематическим проходом по исходной совокупности, при этом каждая запись включается в точности один раз
- Назначение кросс-валидации: измерение качества моделей, бутстрепа — в оценке эмп. ф. распределения для статистик
- Leave-one-out (LOO) аналог бутстрепа — складной нож
- Бутстрепный аналог кросс-валидационных оценок — out-of-bootstrap оценки

Бутстреп: теория и ограничения

Перекрестная проверка

[Слайды Luc Demortier. The Parametric Bootstrap and Particle Physics]