

Прикладная статистика в машинном обучении

Семинар 3

Бутстреп

И. К. Козлов
(Мехмат МГУ)

2022

Интегралы

1 Интегралы

2 Метод складного ножа

3 Медиана

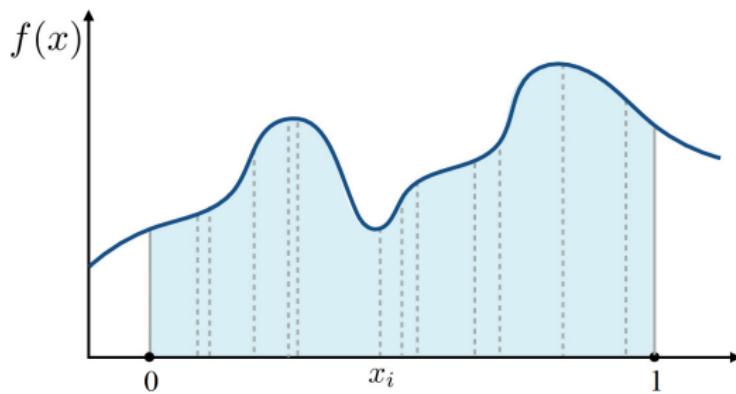
4 LOO-CV

5 Бутстреп на практике

Метод Монте-Карло

На лекции мы обсудили интегрирование методом Монте-Карло.

Оценка интеграла $\int_0^1 f(x)dx$ — берём $f(x_i)$ в случайных точках и усредняем.



Интегрирование по Лебегу

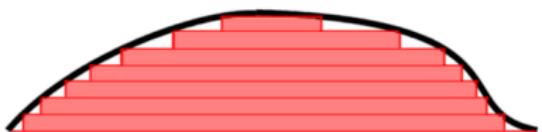
Q: Помните ли Вы, чем отличается интеграл по Жордану от интеграла по Лебегу?

Интегрирование по Лебегу

Q: Помните ли Вы, чем отличается [интеграл по Жордану](#) от [интеграла по Лебегу](#)?



По Жордану



По Лебегу

Интегрирование по Лебегу

Супер-наглядный пример) Как считать деньги?

- По Жордану — суммируем подряд.
- По Лебегу — разбить по номиналу, просуммировать в кучках, потом сложить.

Интегрирование по Лебегу



Интеграл Лебега

Строгое определение [интеграла Лебега](#) см.



A. Н. Колмогоров, С. В. Фомин

Элементы теории функций и функционального анализа.

или любую другую хорошую книжку по теории функций/теории вероятностей.

1 Интегралы

2 Метод складного ножа

3 Медиана

4 LOO-CV

5 Бутстреп на практике

Литература

Подробнее про метод складного ножа см.



Wasserman L.

All of Nonparametric Statistics.

Метод складного ножа

- Пусть $T_n = T(X_1, \dots, X_n)$ — статистика.
- $T_{(-i)}$ — статистика для всех элементов *кроме i-того*:

$$T_{(-i)} = T_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

- Обозначим

$$\bar{T}_n = \frac{1}{n} \sum_{i=1}^n T_{(-i)}.$$

Идея метода складного ножа — оценить смещение и дисперсию через оценки $T_{(-i)}$.

Метод складного ножа

Задача 1. Оценка смещения методом складного ножа.

- Пусть оценка для всех элементов:

$$T_n = \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right).$$

- Q: Какая оценка на $\bar{T}_n = \sum_{i=1}^n T_{(-i)}$?

Метод складного ножа

- Пусть

$$T_n = \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right).$$

- Тогда

$$T_{(-i)} = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Метод складного ножа

- Пусть

$$T_n = \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right).$$

- Тогда

$$T_{(-i)} = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

- Поэтому

$$\bar{T}_n = \sum_{i=1}^n T_{(-i)} = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Q: Как выразить $\text{bias}(T_n) = \frac{a}{n} + O\left(\frac{1}{n^2}\right)$ через T_n и \bar{T}_n ?

Метод складного ножа

Оценка смещения

$$b_{jack} = (n - 1)(\bar{T}_n - T_n)$$

Доказательство. Напомним, что

$$T_n = \theta + \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right).$$

$$\bar{T}_n = \sum_{i=1}^n T_{(-i)} = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Получаем требуемую оценку смещения до 2го порядка малости:

$$b_{jack} = (n - 1)(\bar{T}_n - T_n) = \frac{a}{n} + O\left(\frac{1}{n^2}\right)$$

Метод складного ножа

Оценка дисперсии методом складного ножа.

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

оценка стандартного отклонения $\hat{s}_{\text{e}}_{\text{jack}} = \sqrt{v_{jack}}$.

Это состоятельная оценка на $\mathbb{V}(T_n)$ (при должных условиях)

$$\frac{v_{jack}}{\mathbb{V}(T_n)} \xrightarrow{\mathbb{P}} 1.$$

Метод складного ножа

Задача 2. Найти оценки смещения и дисперсии методом складного ножа для $T_n = \bar{X}_n$.

Метод складного ножа

Задача 2. Найти оценки смещения и дисперсии методом складного ножа для $T_n = \bar{X}_n$.

A1: Оценка смещения

$$b_{jack} = (n - 1)(\bar{T}_n - T_n) = 0.$$

Метод складного ножа

A2: Оценка дисперсии методом складного ножа:

$$\begin{aligned} v_{jack} &= \frac{n-1}{n} \sum_{i=1}^n (T_{(-i)} - \bar{T}_n)^2 = \\ &= \frac{n-1}{n} \sum_{i=1}^n \left(\frac{1}{n-1} (n \cdot \bar{X}_n - X_i) - \bar{X}_n \right)^2 = \\ &= \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n \left(\frac{1}{n-1} (\cdot \bar{X}_n - X_i) \right)^2 = \frac{S_n^2}{n}. \end{aligned}$$

Теорема (Efron, 1982)

- Оценки дисперсии методом складного ножа для квантилей $T(F) = F^{-1}(p)$ несостоятельны.
- Для медианы ($p = \frac{1}{2}$)

$$\frac{v_{jack}}{\sigma_n^2} \xrightarrow{d} \left(\frac{\chi_2^2}{2} \right)^2,$$

где σ_n^2 — асимптотическая оценка дисперсии медианы, а χ_2^2 — распределение хи-квадрат.

Jack of all trades

Jackknife — может применяться везде, полезен почти никогда.

JACK OF
ALL TRADES,
MASTER OF
NONE?



Медиана

1 Интегралы

2 Метод складного ножа

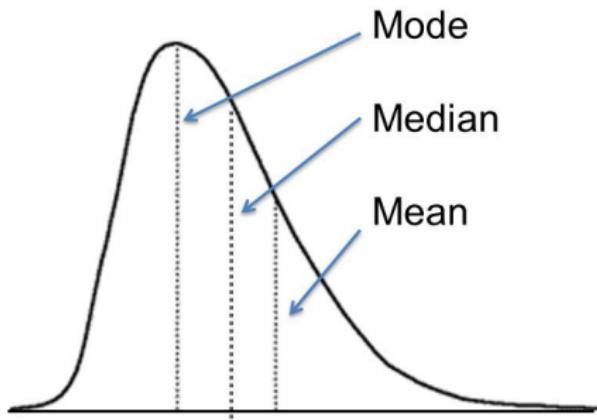
3 Медиана

4 LOO-CV

5 Бутстреп на практике

Медиана и квантили

- Мода — наиболее частое значение.
- Медиана MED = серединное значение ($\mathbb{P}(X < MED) = \frac{1}{2}$).
- Среднее значение = матожидание $\mathbb{E}X$.



Медиана и квантили

Предельное распределение для медианы известно:

Теорема

Пусть элементы выборки имеют плотность $p(x)$, причём $p(x_{\frac{1}{2}}) > 0$. Тогда

$$\sqrt{n} \left(MED - x_{\frac{1}{2}} \right) \rightarrow \xi \sim \mathcal{N} \left(0, \frac{1}{4p^2(x_{\frac{1}{2}})} \right).$$

Медиана и квантили

Предельное распределение для медианы известно:

Теорема

Пусть элементы выборки имеют плотность $p(x)$, причём $p(x_{\frac{1}{2}}) > 0$. Тогда

$$\sqrt{n} \left(MED - x_{\frac{1}{2}} \right) \rightarrow \xi \sim \mathcal{N} \left(0, \frac{1}{4p^2(x_{\frac{1}{2}})} \right).$$

Есть аналогичная теорема для квантилей x_α . Там дисперсия $\frac{\alpha(1-\alpha)}{p^2(x_\alpha)}$.

Медиана и квантили

Подробнее про распределение для квантилей с док-вами — см. Главу 7

 М. Б. Лагутин,
Наглядная математическая статистика.

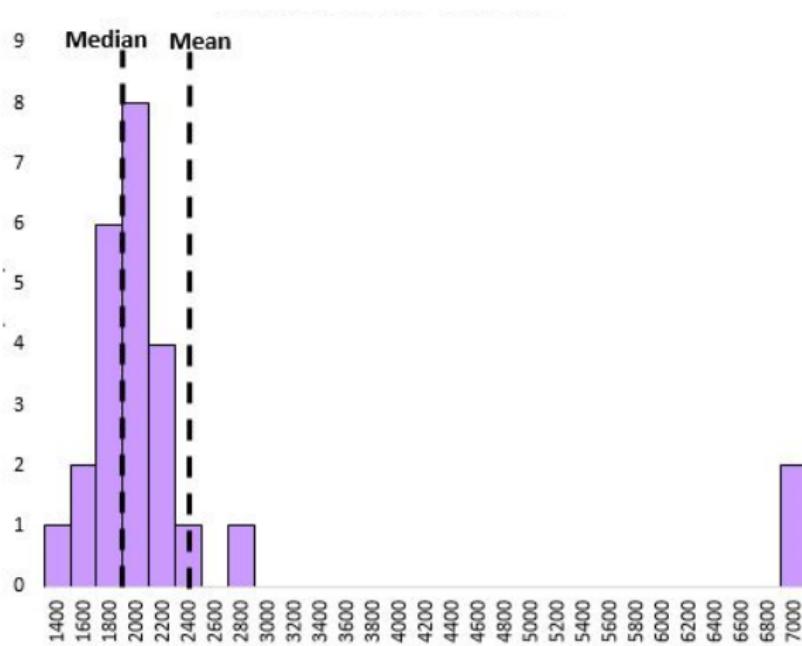
Медиана и матожидание

Q: Какое преимущество имеет медиана перед средним значением?

Медиана и матожидание

Q: Какое преимущество имеет медиана перед средним значением?

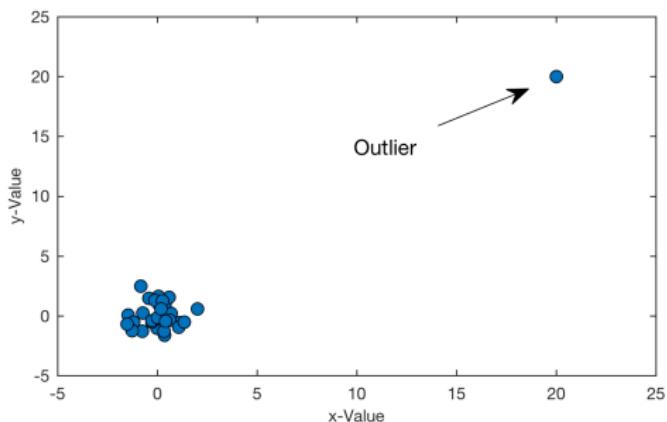
A: Не так сильно подвержены выбросам.



Медиана и матожидание

Выбросы (Outliers) — очень важная проблема в ML.

И априори не понятно — это особые случаи/праздники и т.д. или неверные данные.



Робастность

Робастность (устойчивость к выбросам) обсуждается в Главу 8

 М. Б. Лагутин,
Наглядная математическая статистика.

Робастность

Q: Бутстреп и выбросы — будут проблемы?

Робастность

Q: Бутстреп и выбросы — будут проблемы?

A: Да, будут проблемы — выбросы при сэмплировании окажут ещё большее влияние.



LOO-CV

1 Интегралы

2 Метод складного ножа

3 Медиана

4 LOO-CV

5 Бутстреп на практике

Leave one out cross-validation

Q: Какую ML-технику напоминает Jackknife?

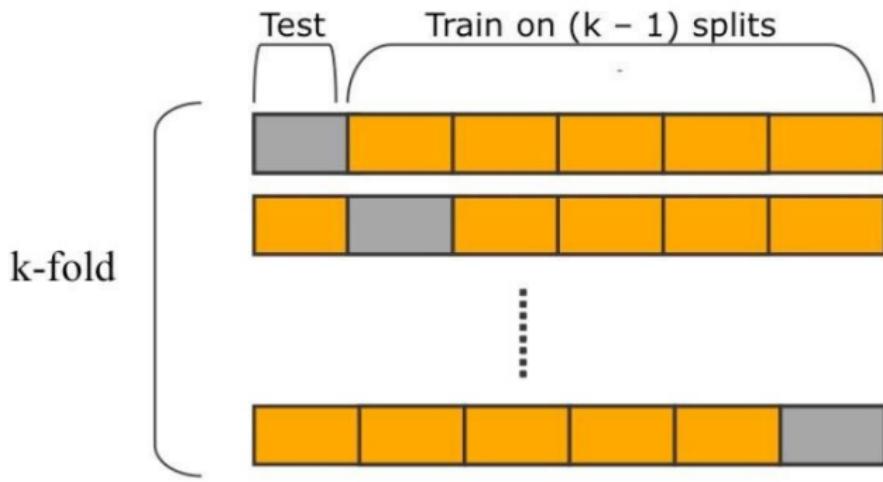
Leave one out cross-validation

Q: Какую ML-технику напоминает Jackknife?

A: Скользящий контроль по отдельным объектам = [Leave one out cross-validation](#).

Leave one out cross-validation

Leave one out cross-validation:

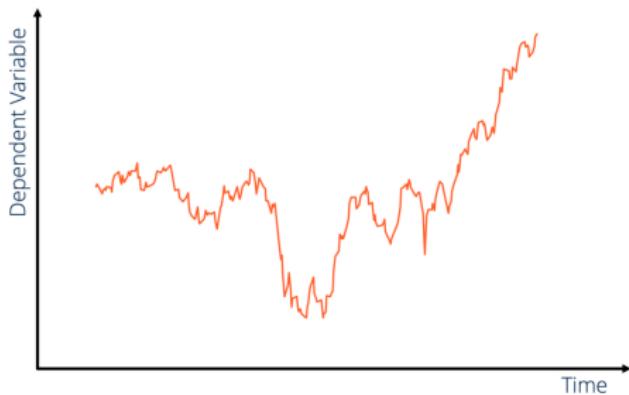


Временные ряды

Временные ряды.

Q: Хорошо ли применять для них кросс-валидацию?

Time-Series Analysis



Временные ряды. Кросс-валидация

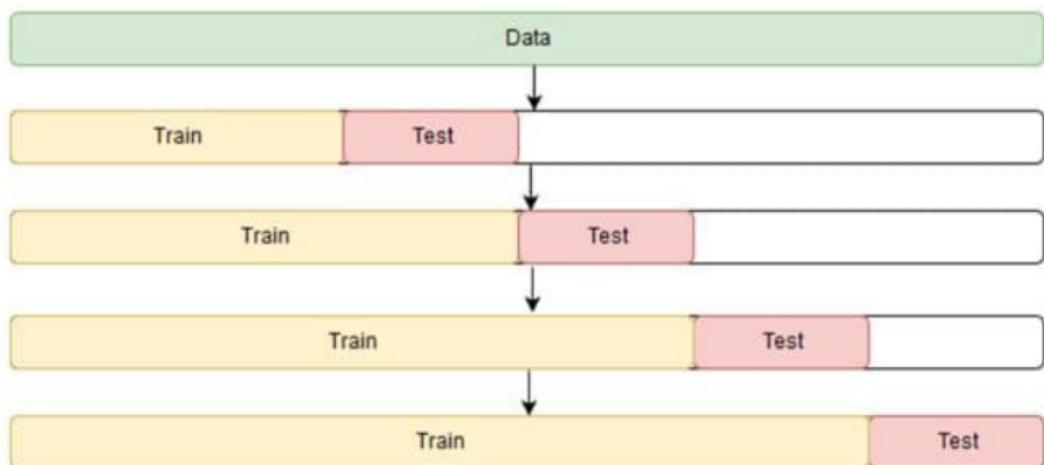
Нельзя допускать утечки данных (**Data Leakage**)!

Нельзя предсказывать прошлое по будущему!

Нельзя допускать утечки данных (Data Leakage)!

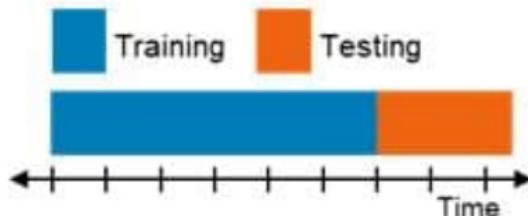
Нельзя предсказывать прошлое по будущему!

Возможная схема кросс-валидации:

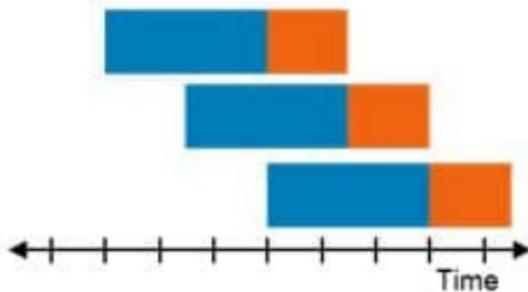


Другая схема кросс-валидации для временных рядов:

Time-based Estimation



Time-based cross-validation



Проблемы бутстрепа

Q: Когда бутстреп плохо применять?

Проблемы бутстрепа

Q: Когда бутстреп плохо применять?

A: Могут быть различные проблемы. Как минимум:

- ➊ Маленькая выборка.
- ➋ Есть выбросы.
- ➌ Теряется структура данных (временные ряды, симметрии, ограничения и т.д.).

Проблемы бутстрепа

Подробнее про проблемы бутстрепа:

[https://notstatschat.tumblr.com/post/156650638586/
when-the-bootstrap-doesnt-work](https://notstatschat.tumblr.com/post/156650638586/when-the-bootstrap-doesnt-work)

Бутстреп на практике

1 Интегралы

2 Метод складного ножа

3 Медиана

4 LOO-CV

5 Бутстреп на практике

Временные ряды. Кросс-валидация

Пример применения бутстрепа.

Scipy — главный пакет для Python по статистике:

[https:](https://)

//docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html