Введение в математическую статистику. Теория оценивания II

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

27 января 2021

- Повторение
- Метод Монте-Карло
- Оценка среднего
- Оценка дисперсии

Повторение

Теория оценивания. Постановка задачи.

- В общем случае задается семейство функций распределения $\{F_\theta(x), \theta \in \Theta\}$, где Θ — множество возможных значений параметра.
- Данные $x_1, \ldots, x_n$ рассматриваются как реализация выборки $X_1, \ldots, X_n$, элементы которой имеют функцию распределения $F_{\theta_0}(x)$ при некотором неизвестном значении $\theta_0 \in \Theta$.
- Задача состоит в том, чтобы оценить (восстановить) θ_0 по реализации $x_1, \ldots, x_n$ наиболее точно.

Повторение

Оценивание θ_0 происходит при помощи некоторых функций $\widehat{\theta}$ от n переменных $x_1, \ldots, x_n$, которые называются оценками или статистиками.

Подставляя в оценку $\widehat{\theta}$ реализацию выборки $x_1, \ldots, x_n$, мы получим число — оценку неизвестного параметра θ_0.

Повторение

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется несмещенной, если

$$\mathbb{E}_\theta \left[\widehat{\theta}(X_1, \ldots, X_n)\right] = \theta \text{ для всех } \theta \in \Theta.$$

Здесь индекс θ у математического ожидания $\mathbb{E}_\theta$ означает, что имеется в виду математическое ожидание случайной величины $\widehat{\theta}(X_1, \ldots, X_n)$, где X_i распределены с функцией распределения $F_\theta(x)$.

Несмещенность означает, что при многократном вычислении оценки на разных данных среднее арифметическое полученных оценок будет стремится к истинному значению параметра θ.

Повторение

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется состоятельной, если для всех $\theta \in \Theta$

$$\widehat{\theta}(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}_\theta} \theta \quad \text{при } n \to \infty.$$

Здесь $\xrightarrow{\mathbb{P}_\theta}$ обозначает «сходимость по вероятности»: для любого $\varepsilon > 0$

$$\mathbb{P}_\theta\left(\left|\widehat{\theta}(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) \to 0 \text{ при } n \to \infty.$$

Состоятельность оценки означает концентрацию оценки около истинного значения параметра с ростом размера выборки n (что устремив $n \to \infty$, оценка сойдется к истинному значению параметра θ).

Повторение

Основная идея методов построения оценок:
чтобы оценить d неизвестных параметров модели, нам необходимо составить d уравнений на них.

Повторение

Метод моментов: d уравнений на неизвестные параметры получаются приравниваем первых d теоретических моментов к их эмпирическим аналогам.

(Теоретическим) моментом k-го порядка случайной величины X называется величина

$$A_k = \mathbb{E}X^k.$$

Выборочным моментом k-го порядка случайной величины X называется величина

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Повторение

Метод максимального правдоподобия: чтобы оценить d неизвестных параметров модели, нам необходимо найти максимум функции правдоподобия (то есть найти частные производные по d параметрам и приравнять их к нулю).

Введем величину:

$$p(u, \theta) = \begin{cases} \mathbb{P}_\theta(X = u) & \text{в дискретном случае,} \\ f_\theta(u) & \text{в непрерывном случае } (f_\theta \text{ — плотность).} \end{cases}$$

Тогда функцией правдоподобия называется величина:

$$L(\theta) = p(x_1, \theta) \cdot \ldots \cdot p(x_n, \theta).$$

Метод Монте-Карло

Пусть дана реализация выборки $x_1, \dots, x_n$ из некоторого распределения X с неизвестным параметром θ.

Иногда интерес представляет получение оценки не для самого параметра θ, а для математического ожидания $\mathbb{E}_\theta[g(X)]$, где $g : \mathbb{R} \to \mathbb{R}$ — некоторая (известная) функция.

Можно, конечно, сначала оценить θ с помощью какой-то оценки $\widehat{\theta}$, а потом посчитать $\mathbb{E}_{\widehat{\theta}}[g(X)]$. Эффективно ли это? Можно ли оценить $\mathbb{E}_\theta[g(X)]$ напрямую?

Метод Монте-Карло

Это можно сделать с помощью оценки Монте-Карло:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Метод Монте-Карло

Оценка Монте-Карло является несмещенной и состоятельной.

1. Несмещенность:

$$\mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right] = \frac{1}{n}\Big(\mathbb{E}_\theta[g(X_1)] + \ldots \mathbb{E}_\theta[g(X_n)]\Big) = \mathbb{E}_\theta[g(X)].$$

2. Состоятельность: согласно закону больших чисел

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \quad \xrightarrow{\mathbb{P}_\theta} \quad \mathbb{E}_\theta[g(X)].$$

Метод Монте-Карло

Примеры:

1. Математическое ожидание:

$$\mathbb{E}_\theta[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i.$$

2. Моменты большего порядка: для $k > 1$

$$\mathbb{E}_\theta[X^k] \approx \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

3. Более сложные функции. Например:

$$\mathbb{E}_\theta[X^3 \sin(X) \log(X)] \approx \frac{1}{n}\sum_{i=1}^{n} x_i^3 \sin(x_i) \log(x_i).$$

Метод Монте-Карло

Оценки Монте-Карло могут быть полезны не только в контексте, когда задано параметрическое семейство.

Их еще можно использовать и тогда, когда

- нам ничего не известно о распределении;
- нам известно распределение, но явное вычисление математического ожидания $\mathbb{E}[g(X)]$ является затратным, а выборку из распределения получить легко.

Метод Монте-Карло

Пример

Давайте сравним методы численного интегрирования и метод Монте-Карло в простой задаче. Пусть дана некоторая функция $g(x)$, у которой первообразную посчитать нельзя. Как вычислить приближенно интеграл?

$$I = \int_0^1 g(u)du$$

Метод Монте-Карло

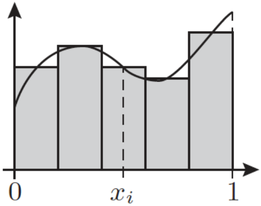
1. Численное интегрирование.

   Простейший способ — метод прямоугольников. Он состоит в оценке I интегральной суммой

$$I_n = \frac{1}{n} \sum_{i=1}^{n} g(x_i),$$

   где $x_i = \frac{i-1/2}{n}$ — это *«узлы» равномерной сетки*, то есть середины интервалов разбиения отрезка $[0, 1]$ на n равных частей.

Метод Монте-Карло



Метод Монте-Карло

2. Метод Монте-Карло.

С помощью метода Монте-Карло I можно оценить через

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} g(x_i),$$

где $x_1, \ldots, x_n$ — реализация выборки из равномерного распределения на отрезке $[0, 1]$.

Данная оценка действительно будет оценивать I:

$$\mathbb{E}\left[\widehat{I}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[g(X_i)] = \int_0^1 g(u) \cdot 1\, du = I.$$

Метод Монте-Карло

Оценка метода Монте-Карло отличается от оценки метода прямоугольников тем, что в качестве «узлов» в ней используются случайные числа $x_1, \ldots, x_n$ из равномерного распределения на [0,1].

Как думаете, какой из двух методов лучше?

Метод Монте-Карло

1. При условии, что $g(x)$ дважды непрерывно дифференцируема, можно показать, что погрешность метода прямоугольников оцениваться сверху так:

$$|I - I_n| \leq \frac{M}{24} \cdot \frac{1}{n^2}, \quad \text{где} \quad M = \max_{x \in [0,1]} |g''(x)|.$$

Метод Монте-Карло

2. Чтобы оценить погрешность метода Монте-Карло, воспользуемся центральной предельной теоремой.

$$\mathbb{E}\left[\widehat{I}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g(X_i)] = \int_0^1 g(u)du = I,$$

$$\mathrm{Var}\left(\widehat{I}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[g(X_i)] = \frac{\sigma^2}{n},$$

где $\sigma^2 = \mathrm{Var}[g(X)]$ по определению.

Метод Монте-Карло

По центральной предельной теореме: для произвольных $a < b$

$$\mathbb{P}\left(a \leq \frac{\sqrt{n}\left(\widehat{I}_n - I\right)}{\sigma} \leq b\right) \approx \mathbb{P}(a \leq Z \leq b),$$

где $Z \sim \mathcal{N}(0, 1)$.

Если положить $a = -3$ и $b = 3$, то мы получим $\mathbb{P}(-3 \leq Z \leq 3) \approx 0.997$ («правило трех сигм»).

В результате:

$$\left|\widehat{I}_n - I\right| \leq 3\sigma \cdot \frac{1}{\sqrt{n}} \quad \text{с вероятностью} \approx 0.997.$$

Метод Монте-Карло

Вывод: Неразумно использовать метод Монте-Карло для вычисления одномерных интегралов — для этого существуют квадратурные формулы, простейшая из которых — рассмотренная формула метода прямоугольников.

Метод Монте-Карло

Тем не менее, метод Монте-Карло или его модификации часто оказываются единственным численным методом, позволяющим решить задачу вычисления интеграла большой кратности.

Дело в том, что чтобы добиться уровня точности $\varepsilon \in (0, 1)$ с помощью квадратурных формул, необходимо взять ε^{-d} «узлов» сетки, где d — кратность интеграла.

Этот феномен называется «проклятием размерности». Грубо говоря, он заключается в том, что большое количество методов «ломаются» в большой размерности.

Метод Монте-Карло

Можно записать, что в многомерном случае:

1. Для метода прямоугольников:

$$\left|I_n - I\right| \leq O\left(\frac{1}{n^{2/d}}\right).$$

2. Для метода Монте-Карло:

$$\left|\widehat{I}_n - I\right| \leq O\left(\frac{1}{n^{1/2}}\right).$$

(Эта запись не совсем корректна, но отражает суть вещей.)

Метод Монте-Карло

Вывод: пусть нам необходимо найти $\mathbb{E}[g(X)]$, где

- X — случайный вектор в $\mathbb{R}^d$ с плотностью $f(u)$,
- $g : \mathbb{R}^d \to \mathbb{R}$ — некоторая функция.

Если размерность d является большой и/или функция g является сложной, то единственным доступным методом решения задачи может оказаться метод Монте-Карло

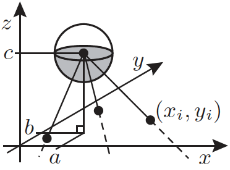
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(u)f(u)du \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i),$$

где $x_1, \ldots, x_n$ — реализация выборки из распределения X.

Оценка среднего

Пример

В некоторой точке пространства с неизвестными координатами (a, b, c) находится источник γ-излучения.



Регистрируются координаты (x_i, y_i) точек пересечения траекторий γ-квантов с поверхностью плоскости $z = 0$. Требуется оценить координаты a и b источника излучения, предполагая, что направления траекторий γ-квантов равномерно распределены.

Оценка среднего

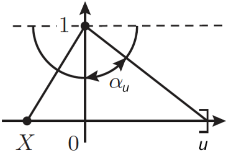
Первое, что приходит в голову, — это усреднить (x_i, y_i).

Ясно, что точки пересечения траекторий с плоскостью $z = 0$ располагаются гуще непосредственно под источником излучения. В подобных случаях прибегают к усреднению данных, чтобы устранить разброс измерений (предполагается, что при этом происходит взаимная компенсация отклонений в разные стороны).

Однако, в данном случае усреднение совершенно бесполезно.

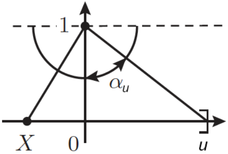
Оценка среднего

Для объяснения, почему это так, рассмотрим одномерный аналог эксперимента.



- Из точки (0, 1) выходит случайный луч, направление которого равномерно распределено на нижней полуокружности с центром $(0, 1)$.
- Пусть случайная величина X — координата пересечения этого луча с осью абсцисс.
- Какой будет плотность $f(u)$ у случайной величины X?

Оценка среднего



Решение. Понятно, что плотность — четная функция. Вычислим ее для $u \geq 0$.

Найдем сначала функцию распределения $F(u) = \mathbb{P}(X \leq u)$:

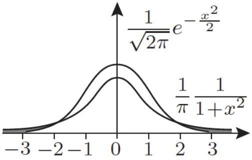
$$F(u) = \mathbb{P}(X \leq 0) + \mathbb{P}(0 < X \leq u) = \frac{1}{2} + \frac{\alpha_u}{\pi} = \frac{1}{2} + \frac{1}{\pi}\arctan(u).$$

Отсюда

$$f(u) = F'(u) = \frac{1}{\pi(1+u^2)}.$$

Оценка среднего

- Плотностью, которую мы получили, — плотность Коши.
- На первый взгляд она похожа на плотность стандартного нормального закона $\mathcal{N}(0, 1)$.



- Однако, они различаются по скорости убывания «хвостов» распределения к нулю. У закона Коши «хвосты» намного «тяжелее».

Оценка среднего

Чем опасны «тяжелые хвосты»?

- Тем, что случайная величина с таким распределением с довольно существенной вероятностью может принимать большие по абсолютной величине значения.
- Поэтому в реализации выборки большого размера из такого закона обязательно появятся одно или несколько наблюдений, которые сильно отличаются от остальных (их называют «выбросами»).
- В этом случае при оценивании «центра» распределения при помощи выборочного среднего X произойдет резкое смещение оценки в сторону наибольшего «выброса».

Оценка среднего

- Из-за слишком «тяжелых хвостов» у закона Коши не существует даже математического ожидания.
- Если бы оно существовало, то по закону больших чисел среднее арифметическое сходилось бы к мат. ожиданию (то, что нам и нужно в этой задаче).
- А что происходит со средним арифметическим для распределения Коши?

  Ответ такой: при любом n среднее арифметическое будет иметь распределение Коши!

  Поэтому оно будет отклоняться от 0 ничуть не меньше значений самих x_i.

Оценка среднего

Поэтому у случайных величин существует несколько характеристик, которые принято называть «средними».

Оценка среднего

Теоретическое среднее	Выборочное среднее
Математическое ожидание: $\mathbb{E}[X]$	Выборочное среднее: $\frac{1}{n}\sum_{i=1}^{n} X_i$

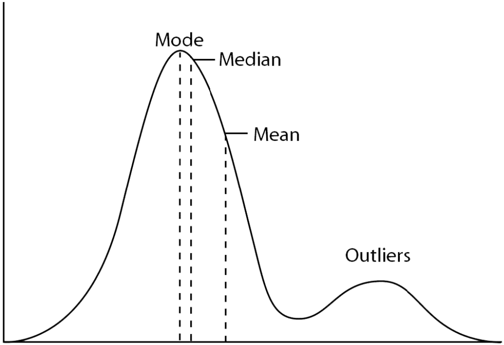
Оценка среднего

Теоретическое среднее	Выборочное среднее
Теоретическая медиана: $x_{1/2}$, которая определяется как решение уравнения $F(u) = 1/2$, где $F(u)$ — функция распределения. Для непрерывной функции $F(x)$ решение всегда существует, но может быть не единственным.	Выборочная медиана: $\mathrm{MED} = \begin{cases} x_{(k+1)}, & n = 2k+1, \\ (x_{(k)} + x_{(k+1)})/2, & n = 2k. \end{cases}$ Здесь $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ это так называемый вариационный ряд, состоящий из упорядоченных по возрастанию элементов реализации выборки $(x_1, \ldots, x_n)$.

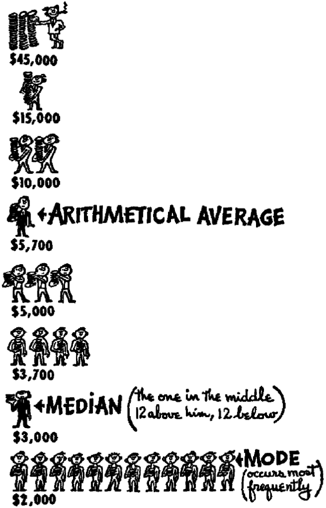
Оценка среднего

Теоретическое среднее	Выборочное среднее
Теоретическая мода: ► В дискретном случае — значение, которое принимаются с наибольшей вероятностью. ► В непрерывном случае — точка максимума функции плотности.	Выборочная мода: ► В дискретном случае — самое распространенное значение реализации выборки. ► В непрерывном случае — нет.

Оценка среднего



Оценка среднего



Оценка среднего

Возвращаясь к оценке среднего для распределения Коши: в данной задаче необходимо было использовать медиану. Она будет и несмещенной, и состоятельной оценкой.

Оценка дисперсии

Пусть нам дана реализация выборки $x_1, \ldots, x_n$ из некоторого распределения X.

Как на основе этих данных оценить дисперсию $\mathrm{Var}(X)$?

Оценка дисперсии

Если бы математическое ожидание $\mathbb{E}[X]$ было бы известным, можно было бы воспользоваться оценкой Монте-Карло:

$$\frac{1}{n}\sum_{i=1}^{n}\big(x_i - \mathbb{E}[X]\big)^2 \quad \approx \quad \mathrm{Var}(X).$$

Но что делать, если $\mathbb{E}[X]$ неизвестно?

Оценка дисперсии

Plug-in principle: если оценка неизвестного параметра требует знания каких-то других неизвестных параметров, то можно попробовать подставить в эту оценку вместо неизвестных параметров их оценки.

При этом, естественно, нет никаких гарантий, что полученная оценка будет хорошей.

Оценка дисперсии

Обозначим оценку для математического ожидания через $\overline{x}$,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Подставим ее в оценку для дисперсии, которая была выше:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2.$$

Данная оценка будет состоятельной, но смещенной.

Оценка дисперсии

Действительно, используя свойства математического ожидания, получаем:

$$\begin{aligned}\mathbb{E}S^2 &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n^2}\sum_{i,j=1}^{n} X_i X_j\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}X_i^2 - \frac{1}{n^2}\sum_{i\neq j}^{n}\mathbb{E}X_i\mathbb{E}X_j \\ &= \mathbb{E}X^2 - \frac{1}{n}\mathbb{E}X^2 - \frac{n-1}{n}(\mathbb{E}X)^2 \\ &= \frac{n-1}{n}\operatorname{Var}X.\end{aligned}$$

Оценка дисперсии

Чтобы устранить смещение у S^2, достаточно домножить ее на $n/(n-1)$. Мы получили несмещенную оценку для дисперсии:

$$S_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Смещенную оценку дисперсии будем обозначать через:

$$S_b^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2.$$

Оценка дисперсии

Среднеквадратическое отклонение (или стандартное отклонение) — это квадратный корень из дисперсии случайной величины:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Оценка дисперсии

Оценка стандартного отклонения по смещённой оценки дисперсии:

$$\widehat{\sigma}_b = \sqrt{S_b^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}.$$

Оценка стандартного отклонения по несмещённой оценки дисперсии:

$$\widehat{\sigma}_u = \sqrt{S_u^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}.$$

Оценка дисперсии

Обе оценки являются смещёнными, так как извлечение квадратного корня «портит» несмещённость. Но при этом обе оценки являются состоятельными.

Термины «среднеквадратическое отклонение» и «стандартное отклонение» обычно применяют к квадратному корню из дисперсии случайной величины, но иногда и к различным вариантам оценки этой величины на основании выборки.

Спасибо за внимание!