Многомерные распределения, оценки параметров.

Центр биоэлектрических интерфейсов, 06 ноября 2018 г.

Денис Деркач, Влад Белавин

Оглавление

Предыдущая лекция

Случайная величина

Определение

Случайная величина — функция, которая связывает пространство состояний системы с множеством вещественных чисел. Для одной и той же системы может существовать несколько случайных величин.

NB: вероятность задаётся на пространстве состояний системы и лишь потом мы можем характеризовать значения (или набор значений) случайной величины какой-то вероятностью.

Вообще говоря, для случайных величин необходимо свойство измеримости, но сейчас для нас оно не важно

Полезные функции

Определение

Пусть ξ — случайная величина. Тогда функцией распределения (cumulative distribution function, cdf) будем называть такую фукнцию $F_\xi(x) : \mathbb{R} \to [0; 1]$:

$$F_\xi(x) = \mathbb{P}(\xi \leq x)$$

Плотностью случайной величины при этом будем называть такую функцию $f_\xi(t)$:

$$F_\xi(x) = \int\limits_{-\infty}^{x} f_\xi(t)\, dt$$

Математическое ожидание и дисперсия

Определение

Для характеристики случайной величины ξ с плотностью $p_\xi(x)$ будем использовать:

- математическое ожидание:

$$\mathbb{E}_\xi(\xi) = \int\limits_{-\infty}^{+\infty} x p_\xi(x)\, dx;$$

- дисперсия:

$$\mathbb{V}ar_\xi(\xi) = \mathbb{E}_\xi\left[(\xi - \mathbb{E}_\xi(\xi))^2\right].$$

Оценки

Обычно мы будем иметь дело с конечным множеством n измерений, **выборкой объёмом** n, из которого пытаемся найти информацию о **генеральной совокупности**. То есть, мы вводим **оценку** (estimator), функцию, которая связывает выборку с интересующим нас параметром. Для отличия оценки от параметра мы используем шляпку: $\widehat{\theta}_n$. Идеальная оценка должна иметь

следующие свойства:

- состоятельность $\widehat{\theta}_n \to \theta$;
- несмещённость $bias = E(\widehat{\theta}_n) - \theta = 0$;
- эффективность $\mathbb{V}ar(\widehat{\theta}_n) \to \min.$

Многомерный случай

Многомерные распределения

В обычной жизни, мы часто встречаемся с ситуациями, когда мы должны анализировать сразу несколько случайных величин. В этом случае нам необходимо анализировать более сложную величину, функцией распределения в этом случае будет:

$$F(x_1, \ldots, x_n) = \mathbb{P}((-\infty, x_1) \times \ldots \times (-\infty, x_n))$$

NB: все слова сказанные про одномерные величины легко обобщаются на многомерные.

Многомерные случайные величины

Несколько случайных величин (или случайный вектор), $\xi = (\xi_1, \dots, \xi_n)$ характеризуются совместной функцией распределения $F_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n) : \mathbb{R}^n \to [0; 1]$:

$$F_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n) = \mathbb{P}(\xi_1 \leq x_1, \dots, \xi_n \leq x_n),$$

аналогично, совместная плотность распределения $p(t_1, \dots, t_n)$:

$$F_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n) = \int\limits_{-\infty}^{x_1} \dots \int\limits_{-\infty}^{x_n} p(t_1, \dots, t_n)\, dt_1 \dots dt_n,$$

Независимость случайных величин

Определение

Пусть случайные величины X и Y имеют совместную плотность $p(x, y)$. X и Y будем называть независимыми, если

$$p(x, y) = p(x) \cdot p(y).$$

NB: когда мы говорим про выборки, обычно имеем в виду набор независимых и одинакого распределённых случайных величин (обозначим i.i.d.).

Условное и маргинальное распределения

Довольно часто мы занимаемся анализом зависимых случайных величин. При этом нам надо переходить к одномерным распределениям. Мы можем это сделать двумя способами:

- условное распределение

$$p(x|y) = \frac{p(x, y)}{p(y)};$$

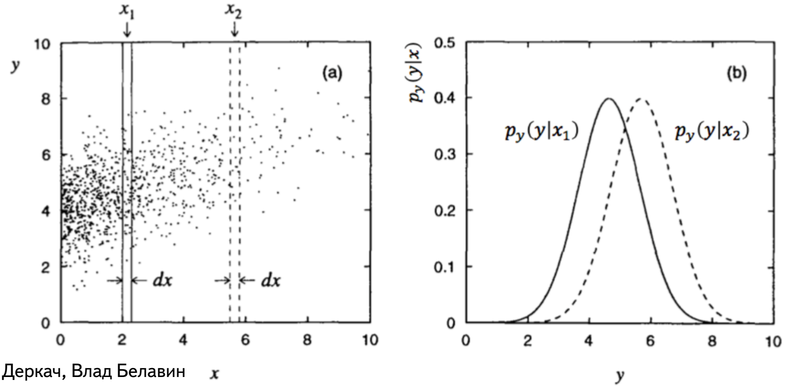
- маргинальное распределение

$$p(x) = \int p(x, y) dy.$$

Условная плотность

Фактически отличается от маргинального, так как значение случайное величины $X = x$ влияет на распределение Y.

$$p(y|x) = \frac{p(x, y)}{p(x)}$$



Теорема Байеса

Из предыдущих рассуждений легко видеть, что

$$p(x \mid Y = y) = \frac{p(y \mid X = x)\, p(x)}{p(y)}.$$

Теорему Байеса можно также переписать в виде вероятностей событий:

$$P(A|B) = P(B|A)P(A)/P(B).$$

Задача 1

Решим задачу 1 из тетрадки.

Корреляция и причинность

Ковариация

Пусть X, Y — две случайные величины.

Определение

Ковариацией X и Y будет называться

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

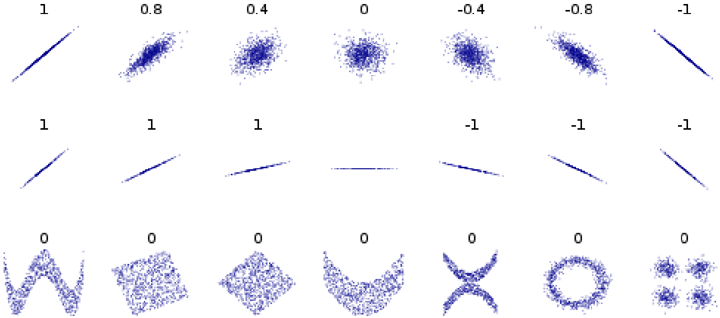
Свойства ковариации

- X и Y независимы, то $\mathrm{cov}(X, Y) = 0$ (обратное неверно).
- $\mathrm{cov}(X, X) = \mathbb{V}ar X.$
- $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
- $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- $\mathrm{cov}(aX, bY) = ab \cdot \mathrm{cov}(X, Y)$.
- $\mathrm{cov}(X + a, Y + b) = \mathrm{cov}(X, Y)$.
- $\mathrm{cov}^2(X, Y) \leq \mathbb{V}ar X \mathbb{V}ar Y.$

Коэффициент корреляции Пирсона

Определение

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}arX\mathbb{V}arY}}$$



Оценка коэффициента корреляции Пирсона

Простая оценка коэффициента корреляции очевидна:

$$\widehat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Стоит помнить, что эта оценка смещена, особенно для малого значения n.

Существуют другие способы оценки.

http://www-bcf.usc.edu/ lototsky/MATH408/SampleCorrCoef.pdf

Решим задачу 3.

Интерпретации вероятности

Априорные и апостериорные суждения

- Предположим, мы хотим узнать значение некоторой неизвестной величины.
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания.
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$.

Два подхода к вероятности

В прикладной статистической науке есть два типа интерепретаций вероятностных процессов:

- классический (или частотный, frequentist) подход считает, что вероятность события X определяется:
  $$P(X) = \lim_{N \to \infty} \frac{n}{N},$$
  где N - количество тестирований, n - количество выпадений X.
- байесовский подход считает $P(X)$ степенью уверенности, что X — верная гипотеза.

Оба подхода удовлетворяют всем требованиям на вероятность.

Частотный подход

- В частотном подходе предполагается, что случайность есть объективная неопределенность.
- При интерпретации случайности как объективной неопределеннсти единственным возможным средством анализа является проведение серии испытаний ($n \to \infty$!). При этом, мы не знаем, когда n является достаточно большой.
- Обычно мы говорим о вероятностях неповторяемых событий (P(дожды|завтра)).

Байесовский подход

- В байесовском подходе предполагается, что случайность есть мера нашего незнания.
- Все величины и параметры считаются случайными. Точное значение параметров распределения нам неизвестно, значит, они случайны с точки зрения нашего незнания.
- В качестве оценок неизвестных параметров выступают апостериорные распределения.
- Построение априорного знания субъективно.

Bayesian vs. Frequnetist

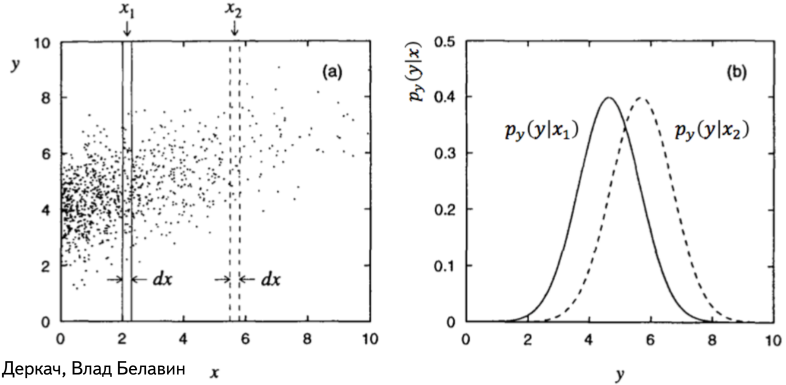
”Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.” — Louis Lyons

Как это влияет на результат?

- Каждая наука рассматривает свой лидирующий подход.
- Методы, которые мы будем описывать далее, в основном разработаны в классическом подходе. Мы будем иногда применять байесовские методы, если они интересны.
- В случае достаточно большой выборки разницы нет :-)

Как это влияет на результат?

- В частотном подходе предполагается, что случайность есть объективная неопределенность.
- В байесовском подходе предполагается, что случайность есть мера нашего незнания.



Пример на броски монетки

Пример

Мы бросили монетку 14 раз, 10 раз выпал орёл. Какие шансы на то, что два следующих броска выпадет орёл?

Фрекентистский подход:

Оценим вероятность успеха: $\widehat{p}_{14} = 10/14 \approx 0.71$. Вероятность двух успехов: $\widehat{p}^2 \approx 0.51$.

Байесовский подход:

Перепишем теорему Байеса:

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)}$$

Пример на броски монетки: байесовское решение

Найдём правую часть:

$$\mathbb{P}(data|p) = \binom{14}{10} p^{10}(1-p)^4,$$

Вероятность данных не зависит от p:

$$\mathbb{P}(data) = \text{const},$$

Мы ничего не знаем о p:

$$\mathbb{P}(p) \sim \text{Uniform}(0,1) \equiv Beta(p,1,1).$$

Тогда

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)} \sim p^{10}(1-p)^4.$$

Пример на броски монетки: байесовское решение

Найдём ответ:

$$\mathbb{P}(HH|data) = \int\limits_0^1 \mathbb{P}(HH|p)\mathbb{P}(p|data)dp = \mathrm{const} \int\limits_0^1 p^2 p^{10}(1-p)^4 dp.$$

Точный подсчёт даст $\mathbb{P}(HH|data) \approx 49\%$.
Отличается от 51% в классическом подходе! Какой правильный?
https://bit.ly/1m54WgZ

Параметрическое оценивание

Статистическая модель

Определение

Статистической моделью будем называть набор $\{P_\theta, \theta \in \Theta\}$ для заданной статистической структуры, где Θ — пространство параметров.

Статистические модели бывают:

1. параметрические;
2. непараметрические;
3. смешанные.

Параметрическая статистическая модель

Определение

Общий вид **параметрической** статистической модели:

$$\mathfrak{F} = \left\{ P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k \right\},$$

где Θ — пространство параметров,
$\theta = (\theta_1, \ldots, \theta_k)$ — вектор параметров, $k \in \mathbb{N}$.

NB: обычно в статистике параметрическими моделями называют те, которые, в отличие от непараметрических, могут быть описаны конечным набором параметров. В некотором смысле разделение нестрогое.

Постановка задачи

Задача параметрического оценивания: необходимо оценить значение $T(\theta)$, где T — некоторая функция параметра θ.

$$\begin{aligned} T: \quad \Theta &\to \mathcal{Y}, \\ \theta &\mapsto T(\theta). \end{aligned}$$

То есть, необходимо построить **оценку** $\hat{T}$ по имеющейся выборке X:

$$\hat{T} : \mathcal{X} \to \hat{\mathcal{Y}}.$$

NB: $\mathcal{Y}$ и $\hat{\mathcal{Y}}$ не обязательно должны совпадать.
NB2: Оценки могут быть детерминированными или рандомизированными.

Пример постановки задачи

Пример

Пусть задана выборка $X_1, ..., X_n$ с распределением $\mathcal{N}(\mu, \sigma^2)$. Необходимо оценить значение μ. В этом случае, $\theta = (\mu, \sigma)$, при этом используется функция $T(\theta) = \mu$, а σ — мешающий параметр. Как определить $\hat{T}$? Какие возможны l?

Метод моментов

Метод моментов

Пусть $\theta = (\theta_1, \ldots, \theta_k)$ — параметр. Для $1 \leq j \leq k$ определим j-й момент согласно формуле:

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}(X^j) = \int x^j dF_\theta(x),$$

и j-й выборочный момент согласно формуле:

$$\widehat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Определение

$\widehat{\theta}_n$ — оценка параметра $\theta = (\theta_1, \ldots, \theta_k)$ на основе метода моментов, если

$$\alpha_1(\widehat{\theta}_n) = \widehat{\alpha}_1,$$
$$\alpha_2(\widehat{\theta}_n) = \widehat{\alpha}_2,$$
$$\ldots$$
$$\alpha_k(\widehat{\theta}_n) = \widehat{\alpha}_k.$$

Пример

Пусть $X_1, \ldots, X_n \sim Bernoulli(p)$, тогда

- $\alpha_1 = \mathbb{E}(X) = p$,
- $\widehat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$,
- Откуда $\widehat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Пример

Пусть $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, тогда

$$\alpha_1 = \mathbb{E}(X_1) = \mu,$$

$$\alpha_2 = \mathbb{E}(X_1^2) = \mathbb{V}ar(X_1) + (\mathbb{E}(X_1))^2 = \sigma^2 + \mu^2,$$

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

$$\widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Решая систему уравнений, получаем, что

$\widehat{\mu} = \overline{X}_n$ и $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$.

Теорема

Пусть $\widehat{\theta}_n$ — оценка параметра θ с помощью метода моментов, тогда (при определенных предположениях о распределении выборки):

1. $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ при $n \to \infty$;
2. Оценка асимптотически нормальна, т. е.
$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \Sigma),$$
где $\Sigma = g\mathbb{E}(XX^T)g^T,\ X = (X^1, X^2, \ldots, X^k)^T$,
$g = (g_1, \ldots, g_k)$ и $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$.

<u>Замечание:</u> последний пункт теоремы можно использовать для нахождения стандартных ошибок и доверительных интервалов.

Метод моментов: пример

Задача 4.

Метод моментов: комментарий

Метод моментов:

- не оптимален;
- прост в использовании;
- полученные с помощью этого метода оценки могут использоваться в качестве начальных значений для более «тонких» алгоритмов.

Функция правдоподобия

Метод максимального правдоподобия

Определение

Пусть задана выборка $X_1, \ldots, X_n \sim F$, при этом у распределения имеется плотность $f(x;\theta)$.

Функция правдоподобия задается формулой:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Логарифмическая функция правдоподобия имеет вид:

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta).$$

Будем рассматривать правдоподобие как функцию параметра $\mathcal{L}_n : \Theta \to [0, \infty)$.

Оценка максимального правдоподобия

Определение

Оценка максимального правдоподобия (ОМП) определяется как такое значение $\widehat{\theta}_n$ параметра θ, которое максимизирует $\mathcal{L}_n(\theta)$.

Пример

Предположим, у нас есть экспоненциальное распределение (например, описывающее распад частицы):

$$f(t,\tau) \sim \frac{1}{\tau} e^{-t/\tau}.$$

И несколько независимых измерений $t_1, \dots, t_n \sim f$.
Функция правдоподобия тогда: $L_n(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t/\tau}$.

Логарифмируем: $\ell_n(\tau) = \sum\limits_{i=1}^{n} n f(t;\tau) = \sum\limits_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau}\right)$.

Найдём максимум L (учитывая, что логарифм монотонная функция):

$$\frac{\partial L_n(\tau)}{\partial \tau} = 0 \rightsquigarrow \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2}\right) = 0 \rightsquigarrow \widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i.$$

Свойства функции правдоподобия:

Определение

1. ОМП состоятельная, то есть $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_*$, где θ_* — реальное значение параметра θ;
2. ОМП не зависит от параметризации, то есть $\widehat{\theta}_n$ — ОМП для θ, тогда $g(\widehat{\theta}_n)$ — ОМП для $g(\theta)$;
3. ОМП асимптотически нормальна: $(\widehat{\theta} - \theta_*)/\widehat{se} \rightsquigarrow \mathcal{N}(0,1)$;
4. ОМП асимптотически оптимальна или эффективна (при достаточно большом объеме выборки ОМП имеет меньшую дисперсию).
5. ОМП приближенно совпадает с байесовской оценкой.

Свойства функции правдоподобия:

Найдём $\widehat{\mu}$ и $\widehat{\sigma}$ для распределения Гаусса для выборки объёмом n:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Снова запищем лог-правдоподобие:

$$\ell_n(\mu,\sigma) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

Производные:

$$\frac{\partial \ell_n}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \quad \frac{\partial \ell_n}{\partial \sigma} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Свойства функции правдоподобия:

Тогда оценки:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

Таким образом ОМП — необладает свойством смещения!

Свойства функции правдоподобия:

Задача 5.