Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Введение. Вероятность. Точечные оценки.

ПМИ ФКН ВШЭ, 7 сентября 2019 г.

Денис Деркач[1], Алексей Артёмов[1,2]

[1] ФКН ВШЭ [2] Сколтех

Оглавление

Вероятность

Что такое вероятность?

- Какая вероятность того, что вы сдадите этот курс на 10?
- Какая вероятность того, что вы сдадите этот курс на 0?
- Какая из этих вероятностей больше? Как Вы это поняли?

Что такое вероятность?

- ВЕРОЯТНОСТЬ — показатель осуществимости тех или иных возможностей при определенных условиях. (Краткий словарь философских терминов).
- Концепция, которая существует без привязки к математичеким объектам.
- Конечно, нам нужны математические термины для правильного количественного описания.

Аксиомы Колмогорова

Для некоторого пространства событий $\mathcal{F}$:

- Событию $A \in \mathcal{F}$ сопоставлено число $\mathbb{P}(A) \geq 0$, которое называется вероятностью .
- Вероятность, что хотя бы одно событие из $\mathcal{F}$ случится: $\mathbb{P}(\mathcal{F}) = 1$.
  - (*) Вероятность пустого набора событий $\mathbb{P}(\emptyset) = 0$.
- Если $X_1 \in \mathcal{F}$ и $X_2 \in \mathcal{F}$ не пересекаются, тогда $\mathbb{P}(X_1 + X_2) = \mathbb{P}(X_1) + \mathbb{P}(X_2)$ (можно распространить для любого счётного числа событий).

Вообще, возможны другие наборы аксиом, однако нас интересует не аксиоматика.

Некоторые свойства вероятности

- Совместные вероятности $P(A \text{ or } B)$ и $P(A \text{ and } B)$:

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B)$$

.

- Полная вероятность:

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \text{ and } B_n)\mathbb{P}(B_n),$$

где всё пространство разделено в наборы B_n,

- Условная вероятность, $\mathbb{P}(A|B)$ означает вероятность, что A произошло, при условии, что B произошло.

Теорема Байеса

- Для совместной вероятности:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

- что означает:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

- используя правило полной вероятности:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|notB)\mathbb{P}(notB)}$$

Пример использования теоремы Байеса

Предположим, что у нас есть детектор определённого вида частиц (K), который в 90% случаях правильно реагирует на пролетающую частицу:

$$P(T^+|K) = 0.9[90\% \text{ acceptance}]$$

и в 1% случаев реагирует на другие частицы:

$$P(T^+|notK) = 0.01[1\% \text{ background}]$$

Предположим, что на прибор среагировал, какая вероятность того, что эта частица K?

Решение примера

Применим теорему Байеса

$$\mathbb{P}(K|T^+) = \frac{\mathbb{P}(T^+|K)\mathbb{P}(K)}{\mathbb{P}(T^+|K)\mathbb{P}(K) + \mathbb{P}(T^+|notK)\mathbb{P}(notK)}$$

. То есть всё зависит от того сколько частиц K есть вокруг детектора, $\mathbb{P}(K)$.

K вокруг	$\mathbb{P}(K) = 1\%$	$\mathbb{P}(K) = 10^{-6}\%$
$\mathbb{P}(K\|T+)$	0.48	10^{-4}
$\mathbb{P}(K\|T-)$	0.01	10^{-7}

- Теорема может быть использована для решения таких задач.
- Этот детектор не очень эффективен, если $\mathbb{P}(K)$ мало.
- Мы никак не интерпретировали, как выглядит $\mathbb{P}$.

Интерпретация вероятностей

Два подхода к интерпретации

Две основные интерпретации:

- классическая (или частотная), вероятность интерпретируется как относительная частота

$$\mathbb{P}(X) = \lim_{N \to \infty} \frac{n}{N},$$

  где N — количество экспериментов, n — количество событий X в этих N экспериментах.
- байесовская считает $\mathbb{P}(X)$ степенью уверенности в исходе X.

Обе интерпретации удовлетворяют аксиомам Колмогорова.

NB: Другие интерпретации также возможны.

NB2: Теорема Байеса $\neq$ байесовской интерпретации.

Классическая интерпретация

- Случайность объективна.
- Выводы построены для большого количества экспериментов: $n \to \infty$.
- Функция правдоподобия — основной инстумент.

NB1: Мы не знаем, когда N становится достаточно большим.
NB2: Обычно мы рассуждаем о следующем событии (например, $\mathbb{P}(rain|tomorrow)$).

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.

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.

NB: Построение априорного знания субъективно.

Есть ли какие-то практические последствия?

- Частотное утверждение: вероятность "наблюдаемых данных" при условии какой-то модели (гипотезы): $\mathbb{P}(data|model)$.
- Байесовское утверждение: вероятность модели (гипотезы) при условии данных $\mathbb{P}(model|data)$.

Проблема:

$$\mathbb{P}(data|model) \neq \mathbb{P}(model|data).$$

Пример:

$$\mathbb{P}(pregnant|woman) \approx 3\%$$
$$\mathbb{P}(woman|pregnant) =?$$

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 @ Phystat 2003 conference

Как это влияет на результат?

- Каждая наука рассматривает свой лидирующий подход.
- Методы, которые мы описывали, в основном разработаны в классическом подходе.
- В случае достаточно большой выборки разницы (почти) нет :-)

Пример на броски монетки

Пример

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

Частотный 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:

$$\mathbb{P}(HH|data) = \int_0^1 \mathbb{P}(HH|p)\mathbb{P}(p|data)dp = \text{const} \int_0^1 p^2 p^{10}(1-p)^4 dp.$$

Точный подсчёт даст $\mathbb{P}(HH|data) \approx 49\%$.

Отличается от 51% в классическом подходе! Какой правильный?

Больше информации по ссылке.

Другие понятия

Случайная величина

Случайная величина — переменная, которая будет принимать разные значения, если мы будем повторять эксперимент. Её значения непредсказуемы за исключением того, что мы знаем по вероятности:

$$\mathbb{P}(data|parameters),$$

при условии, что неизвестным в параметрах заданы некоторые предполагаемые значения.

Плотность вероятности

Для непрерывных величин, распределения вероятности ξ, $\mathbb{P}$, могут быть записаны плотность вероятности или PDF:

$$p_{\xi|parameters}(x)dx = \mathbb{P}(\xi \in [x; x + dx]|parameters).$$

Обычно мы пишем что-то вроде:

$$\mathbb{P}(\xi|parameters) = f(x; parameters)$$

.

NB: точно также для дискретной величины мы можем определить плотность масс.

Характеристики плотности вероятности

Если у нас есть $p_\xi(x)$ случайной величины ξ.

- Математическое ожидание:

$$\mathbb{E}(\xi) = \int x p_\xi dx,$$

- Дисперсия:

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

,

- Высшие моменты:

$$\mu_\xi^k = \mathbb{E}_\xi \left[(\xi - \mathbb{E}_\xi \xi)^3 \right],$$

Свойства математического ожидания и дисперсии

- Математическое ожидание
  - $\mathbb{E}(c) = c$;
  - $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$;
  - Для независимых X и Y: $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.
- Дисперсия
  - $\mathbb{V}ar(c) = 0$;
  - $\mathbb{V}ar(X) \geq 0$;
  - $\mathbb{V}ar(X + c) = \mathbb{V}ar(X)$;
  - $\mathbb{V}ar(cX) = c2\mathbb{V}ar(X)$.

Функция правдоподобия

Можно заметить, что при написании PDF, мы ничего не говорили о параметрах. Что если у нас уже есть данные:

$$\mathbb{P}(data|parameters)\big|_{dataobs.} = \mathcal{L}(parameters)$$

$\mathcal{L}$ называется функцией правдоподобия.
NB: она ненормирована.

Поведение при замене координат

При замене координат, в данных $X \to Y(X)$ или параметрах $\theta\tau(\theta))$.

- Функция правдоподобия остаётся инвариантной:

$$\mathcal{L}(\theta) = \mathcal{L}(\tau(\theta))$$

.

- В случае PDF, инвариантной остаётся интегральная вероятность между точками, то есть, нам необходим учитывать Якобиан замены $X \to Y(X)$:

$$PDF(X) = J(X, Y)PDF(Y)$$

, где Якобиан J определяется $\frac{\partial X}{\partial Y}$ в одном измерении.

Параметрическое оценивание

Постановка задачи

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

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

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.

Смещение оценки

Таким образом, нам необходимо построить оценку параметров на основе нашей ограниченной выборки. Обычно мы обозначаем оценку θ на основании n событий как $\hat{\theta}_n$. Оценка должна быть:

- Состоятельной $\widehat{\theta}_n \to \theta$;
- Несмещённой $bias = E(\widehat{\theta}_n) - \theta = 0$;
- Эффективной $\mathbb{V}ar(\widehat{\theta}_n) \to \min$.

К сожалению, такие оценки не всегда доступны.

Выборочное среднее, выборочная дисперсия

Даже если мы не знаем (не предполагаем) распределение, мы уже можем оценить ранее определенные характеристики. Если у нас есть независимые и одинаково распределенные случайные величины (iid) $X_i \sim f$

- выборочное среднее:

$$\bar{x} = \frac{1}{N} \sum x_i$$

- выборочная дисперсия $\mathbb{V}ar$:

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

NB: хотя S^2 несмещённая оценка σ^2, $S = \sqrt{S^2}$ становится смещённой оценкой σ ($bias = \sigma/4n$).

Пример

Допустим, что измерения $X \sim N(\mu, \sigma^2)$ — интегральная характеристика теста по исследованию крови. Необходимо: вычислить τ — долю наблюдений, для которых характеристика превосходит 1.

Пример

В этом случае также, $\theta = (\mu, \sigma)$ — параметр из пространства параметров $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$.

$$\tau = \mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(\frac{X - \mu}{\sigma} < \frac{1 - \mu}{\sigma}\right) =$$

$$= 1 - \mathbb{P}\left(Z < \frac{1 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{1 - \mu}{\sigma}\right).$$

$\tau = T(\mu, \sigma) = 1 - \Phi((1 - \mu)/\sigma)$ — параметр, который необходимо оценить, Z — случайная величина со стандартным случайным распределением, а Φ — нормальное интегральное распределение.

Пример

Оценить среднее время жизни изотопа.

Пример

Гамма-распределение обычно используется для моделирования времени жизни. Пусть $X \sim Gamma(\alpha, \beta)$, т. е.

$$f(x; \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \text{ где } \alpha, \beta, x > 0.$$

$\Gamma(\alpha)$ — гамма-функция, а $\theta = (\alpha, \beta)$ — параметр.

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy.$$

Для того, чтобы оценить среднее время жизни, то необходимо использовать:

$$T(\theta) = \alpha\beta,$$
$$\hat{T}(\theta) = \mathbb{E}_{\theta}(X).$$

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 $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$.

Замечание: последний пункт теоремы можно использовать для нахождения стандартных ошибок и доверительных интервалов.

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.

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

Свойства функции правдоподобия:

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).

Свойства функции правдоподобия:

Замечание: вышеприведенные свойства ОМП имеют место, если функция $f(x;\theta)$ достаточно регулярная. В «слишком» сложных случаях ОМП «теряет» эти свойства.

Оценка апостериорного максимума

Оценка апостериорного максимума, MAP

Формально, ОМП определяет значения параметров, при которых наши данные наиболее вероятны:

$$f(X;\theta) \sim f(X|\theta).$$

На самом деле, мы обычно задаёмся вопросом, какие значения параметров наиболее вероятны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)},$$

где f, g и h — соответствующие функции распределения.

Оценка MAP

Определение

Оценка апостериорного максимума (MAP) определяется как такое значение $\widehat{\theta}_n$ параметра θ, которое максимизирует $f(\theta|X)$.

Связь с MLE

MAP и MLE очевидно связаны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)} = \frac{\prod_{i=1}^{n} f(X_i;\theta)g(\theta)}{h(X)} \sim \text{const} \prod_{i=1}^{n} f(X_i;\theta)g(\theta)$$

Логарифмируем:

$$\log f(\theta|X) = \log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Получается, что значение MAP оценки и значение оценки MLE совпадают с точностью до априорной оценки $\log(g(\theta))$.

Сопряжённые априорные оценки

Какую $g(\theta)$ выбрать?

- любую;
- но лучше выбирать сопряжённое априорное распределение, для которого функциональная форма совпадает с апостериорным.

Значения параметров сопряжённых распределений имеют смысл предыдущих измерений.

Список из Википедии.

Комментарий о MAP

- позволяет учесть предыдущие знания;
- выдаёт точечную оценку (не совсем байесовский);
- зависит от параметризации;
- при относительно больших n совпадает с MLE (а также в случае $g(X) = \text{const}$!).

Заключение

В этой лекции

- ДЗ делать надо :-)
- Вероятности можно интерпретировать по-разному. Лидирующие интерпретации — Классическая и Байесовская. Интерпретация верна, если она не противоречит аксиомам Коломгорова.
- Для точеченой параметрической оценки существуют несколько способов: метод моментов, метод максимального правдопободобия, метод апостериорного максимума.