Введение в статистику

Гришин Михаил

1 апреля 2016 г.

Полезная информация

Меня зовут - Гришин Михаил

Материалы по лекции доступны в Облаке Mail.ru

Содержание

Вероятность

Эволюция взглядов на вероятность

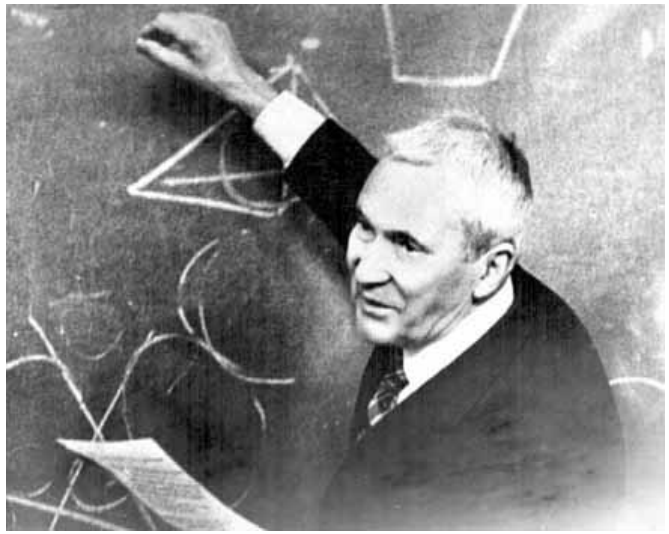
Часто, для описания концепции детерминизма используют следующую цитату П.Лапласа («демон Лапласа»):

> *«Ум, который в данный момент знал бы все силы, действующие в природе [...], охватил бы одной и той же формулой движения крупнейших тел Вселенной и легчайших атомов. Ничто не было бы для него недостоверным, и будущее, как и прошедшее, стояло бы перед его глазами»*

Важно отметить, что для Лапласа, который известен своим вкладом в теорию вероятностей, неопределенность являлась лишь следствием **неполноты** нашего знания.

Современное определение вероятности

Теория вероятностей в современном виде была описана А.Н. Колмоговорым («аксиоматика Колмогорова»).



Объектом нашего рассмотрения будет пространство элементарных исходов Ω. Пример элементарного исхода при игре в карты? В кости?

Сигма-алгебра и ее свойства. Случайные события как элементы сигма-алгебры.

Современное определение вероятности

Измеримость. Вероятность - неотрицательная мера на пространстве случайных событий.

Подмножества пространства элементарных исходов - случайные события. Объединение и пересечение случайных событий.

Свойства вероятности:

- $P(\varnothing) = 0$;
- $P(\Omega) = 1$;
- $P(B \setminus A) = P(B) - P(A), A \subset B$;
- $P(A + B) = P(A) + P(B) - P(A \cap B)$;

Современное определение вероятности

Пример

Вопрос 1: Если сейчас потребуется выбрать случайного студента из аудитории, то какова вероятность, что выберут именно вас?

Вопрос 2: Какова вероятность выбора студента с «отличной успеваемостью» или «днем рождения в апреле»?

Вопрос 3: Менялось ли пространство элементарных исходов в данных двух задачах?

Условная вероятность

Условная вероятность для двух событий A и B :

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

Независимые события

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

$$P(A \cap B) = P(A)\,P(B)$$

Формула Байеса

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

Как правило, используется формула полной вероятности для группы несовместных событий:

$$P(A_j|B) = \frac{P(B|A_j)\,P(A_j)}{\sum_{i=1}^{n} P(B|A_i)\,P(A_i)}$$

Формула Байеса

Пример

Допустим, существет редкое заболевание, которым болеет лишь 0.001 населения. Самый точный тест выявляет всех заболевших (FN = 0), однако в 0.01 случаев он показывает заболевание для здорового индивида (FP = 0.01). Некто X сдает тест и результат положительный. Какова вероятность того, что он болен?

Формула Байеса

Пример

Допустим, существет редкое заболевание, которым болеет лишь 0.001 населения. Самый точный тест выявляет всех заболевших (FN = 0), однако в 0.01 случаев он показывает заболевание для здорового индивида (FP = 0.01). Некто X сдает тест и результат положительный. Какова вероятность того, что он болен?

Ответ

Шанс - всего 9%. Можем ли мы улучшить результат?

Формула Байеса

Пример

Нередко, при вводе запроса в поисковик мы допускаем опечатки, которые поисковик зачастую весьма успешно исправляет. Давайте рассмотрим, как можно построить систему исправления ошибок, пользуясь базовыми законами вероятности - ноутбук Bayes spell correction.ipynb.

Случайные величины

Случайные величины

Случайная величина - измеримая функция $X : \Omega \to \mathbb{R}$, такая что для любого $r \in \mathbb{R}$ событие $\omega : X(\omega) \leq r$ принадлежит сигма-алгебре на пространстве элементарных исходов.

Если случайная величина принимает счетное количество значений, то она называется **дискретной**, в противном случае - **непрерывной** случайной величиной.

Для описания случайной величины, помимо области значений, также необходимо описать её вероятностные свойства. Наиболее общий способ - задать **функцию распределения**.

Функция распределения

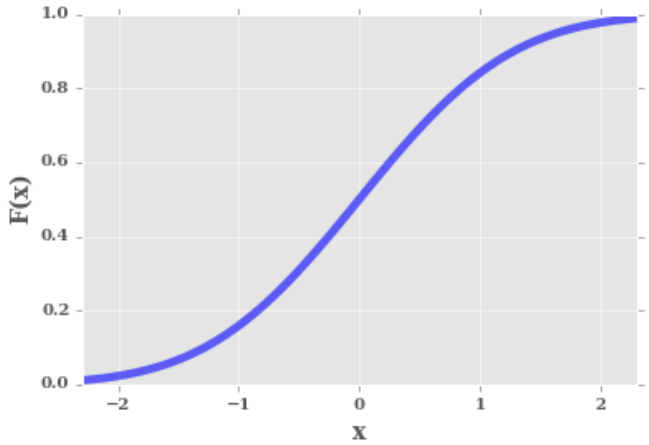
Функция распределения - функция $F_X : \mathbb{R} \to [0, 1]$, определенная как $F_X(x) = P(X \leq x)$. Функция распределения неубывающая и непрерывна справа. Значения функции распределения ограничены $\lim_{x \to +\infty} F_X(x) = 1$, $\lim_{x \to -\infty} F_X(x) = 0$.

Квантильная функция (Quantile function) определяет такое значение случайной величины, что вероятность события «X меньше или равен x» в точности равно p: $Q(p) = \inf\{x \in \mathbb{R} : p \leq F_X(x)\}$. Для непрерывной случайной величины квантильная функция представляет из себя обратную функцию от функции распределения.

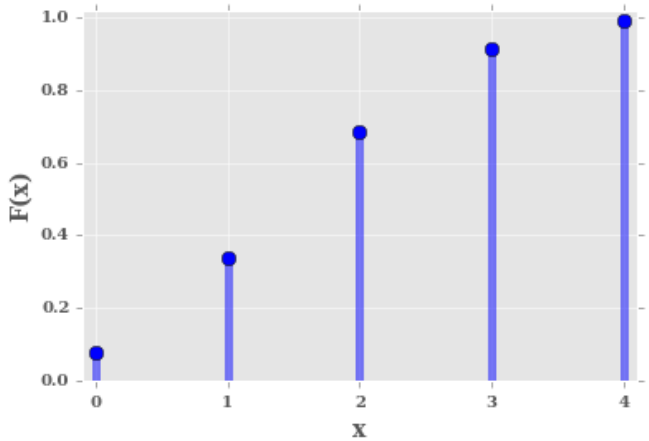
Квантили. Медиана, интерквартильный размах. t-digest.

Пример функции распределения

Определенной на несчетном множестве



Определенной на счетном множестве



Основные законы распределения

Основные сведения

Дискретная случайная величина принимает счетное множество значений. Для каждого из значений ДСВ можем по определению найти вероятность данного значения, поэтому удобно работать с функцией вероятности $p_X(x) = P(X = x)$.

Тогда, функцию распределения можно выразить через функцию вероятности:

$$F_X(x) = \sum_{x_j \leq x} p(x_j).$$

Если мы взвесим все возможные исходы по вероятности их наступления, то получим **математическое ожидание**:

$$E(X) = \mu = \sum_{i=1}^{\infty} p(x_i) x_i.$$

Дискретные случайные величины

Дисперсия: мера разброса значений случайной величины вокруг своего математического ожидания.

$$V(X) = E((X - E(X)^2).$$

Стандартное отклонение: $\sigma = \sqrt{V(X)}$

k-начальный момент:

$$\nu^k = E(X^k) = \sum_{i=1}^{\infty} p(x_i) {x_i}^k.$$

k-центральный момент:

$$\kappa^k = E((X - E(X))^k) = \sum_{i=1}^{\infty} p(x_i)(x_i - E(X))^k.$$

Математическое ожидание - первый начальный момент, дисперсия - второй центральный момент.

Дискретные случайные величины

Коэффициент асимметрии - мера того, насколько «хвосты распределения» отличаются «по длине» друг относительно друга.

$$\gamma_1(X) = \frac{\kappa^3}{\sigma^3}.$$

Коэффициент эксцесса - мера остроты пика распределения случайной величины.

$$\gamma_2(X) = \frac{\kappa^4}{\sigma^4} - 3.$$

Энтропия - способ описать «информацию» о случайной величине, содержащуюся в ее значениях.

$$H(X) = E(-\ln P(X)) = \sum_{i=1}^{\infty} p(x_i) \log p(x_i).$$

Энтропия

Пример

Хорошее объяснение энтропии случайной величины представлено тут - http://colah.github.io/posts/2015-09-Visual-Information/

Дискретные случайные величины

Пример

Давайте ознакомимся с примерами дискретных случайных величин и вспомним основные законы распределения - ноутбук Discrete RV.ipynb.

Непрерывные случайные величины

По аналогии, непрырывная случайная величина прнимает несчетное множество значений. Говорят, что случайная величина является непрырывной, если существует неотрицательная функция $f(x)$, определенная для всех вещественных x, такая что для любого измеримого множества $B \subset \mathbb{R}$ выполняется:

$$P(X \in B) = \int_B f(x)\,dx.$$

Если $B = [a, b]$, то:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Вероятность на множествах меры ноль равна нулю.

$$P(X < a) = P(X \leq a) = F(a) = \int_{-\infty}^a f(x)\,dx.$$

Непрерывные случайные величины

Первые моменты равны:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx.$$

Энтропия непрерывного распределения:

$$H(X) = E(-\ln P(X)) = -\int_{-\infty}^{\infty} P(x) \log P(x)\,dx.$$

Непрерывные случайные величины

Моменты более высоких порядков определены аналогичным (по сравнению с дискретной случайной величиной) образом.

k-начальный момент:

$$\nu^k = E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\, dx.$$

k-центральный момент:

$$\kappa^k = E((X - E(X))^k) = \int_{-\infty}^{\infty} (x - E(X))^k f(x)\, dx.$$

Непрерывные случайные величины

Пример

Давайте ознакомимся с примерами непрерывных случайных величин и вспомним основные законы распределения - ноутбук Continuos RV.ipynb.

Совместная функция распределения

Для двух случайных величин X и Y совместная функция распределения определена как:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

Путем предельного перехода можем получить маржинальную (marginal) функцию распределения:

$$F_X(x) = P(X \leq x, Y < \infty) = \lim_{y \to \infty} F_{XY}(x, y).$$

Совместное распределение ДСВ

Для дискретных случайных величин совместная функция вероятности определена как

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y).$$

или в более общем виде:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2|X_1 = x_1)...$$
$$P(X_n = x_n|X_1 = x_1, X_12 = x_2, ..., X_{n-1} = x_{n-1}).$$

Условные вероятности для ДСВ:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Совместное распределение ДСВ

Исходя из этого определения легко получить:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \sum_{X<x} \sum_{Y<y} P(x, y).$$

И маржинальные вероятности:

$$P(X = x) = \sum_{y} P(X = x, Y = y) = \sum_{y} P(X = x, Y = y).$$

Совместное распределение НСВ

Соответственно, для НСВ функция $f(x, y)$ является функцией плотнотсти вероятности, если $f(x, y) \geq 0$ для всех (x, y) и ограничена:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1.$$

Тогда вероятность для СВ принять значение в области $A \in \mathbb{R} \times \mathbb{R}$:

$$P(X, Y \in A) = \int \int_{A} f(x, y)\, dxdy.$$

Маржинальная функция плотности:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X\,Y}(x, y)\, dy.$$

Совместное распределение НСВ

Ранее, мы отметили, что условная вероятность для ДСВ имеет вид:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Для НСВ мы используем условную функцию плотности вероятности:

$$f_{x|Y=y} = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Соответствующую вероятность для события $X \in A$ можно легко получить как:

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\, dx.$$

Условное распределение СВ

Так как для функции условной вероятности выполняются все необходимые условие, то мы можем говорить об условных моментах случайных величин, например математическом ожидании:

$$E(X|Y=y) = \int_{-\infty}^{\infty} x f_{x|Y=y}\, dx.$$

и условной дисперсии:

$$E(X|Y=y) = \int_{-\infty}^{\infty} (x - E(X|Y))^2 f_{X|Y}\, dx.$$

Задача нахождения ожидаемого значения некоторой функции $Y = f(X) + \epsilon$, где ϵ - отклонение (ошибка), а X - известная нам реализация случайной величины (величин) имеет большое применение в прикладной статистике.

Независимость СВ

Будем говорить, что СВ X, Y независимы, если для любого $A, B \in \mathbb{R}$ верно:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Таким образом, для НСВ условие независимости принимает вид

$$f_{X,Y}(x, y) = f_X(x)f_Y(y),$$

для дискретных случайных величин -

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

В случае, если это условие не выполняется - будем считать, что случайные величины зависимы.

Меры зависимости СВ

Ненормированная мера **линейной** зависимости - ковариация:

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

Корреляция - нормированная мера **линейной** зависимости:

$$corr(X, Y) = \rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

Взаимная энтропия KL для распределений P и Q:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)},$$

если существуют функции плотности вероятности p и q:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)},$$

Случайный вектор

Случайный вектор - набор из n случайных величин $X = (X_1, ..., X_n)$. Случайные величины $(X_1, ..., X_n)$ независимы, если

$$P(X_1 \in A_1, ..., X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

Если каждая из СВ, входящих в вектор X, подчинена одному и тому же закону распределения с функцией вероятности F, то говорят что $(X_1, ..., X_n)$ - **независимые одинаково распределенные** величины или IID (independent identically distributed):

$$X_1, ..., X_n \sim F.$$

Аналогично, будем называть $(X_1, ..., X_n)$ **выборкой** размера n.

Основы теории оценивания

Полезные свойства моментов СВ

Математическое ожидание взвешенной суммы СВ $(X_1, ..., X_n)$:

$$E\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i E(X_i).$$

Если $(X_1, ..., X_n)$ - iid:

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

Дисперсия суммы СВ $(X_1, ..., X_n)$:

$$V\left(\sum_{i=1}^{n} w_i X_i + c_i\right) = \sum_{i=1}^{n} w_i^2 V(X_i).$$

Закон больших чисел

Неравенство Чебышева: для случайной величины X с $E(X) = \mu$ и $V(X) = \sigma^2$:

$$P\left(|X - \mu| \geq t\right) \leq \frac{\sigma^2}{t^2}.$$

Если $(X_1, ..., X_n)$ - iid, причем $E(X_i) = \mu$ для всех i, то выборочное среднее:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

сходится по вероятности к теоретическому среднему:

$$\bar{X}_n \xrightarrow{p} \mu.$$

ЦПТ

Мы видели, что выборочное среднее аппроксимирует математическое ожидание неизвестного распределения при условии, что число наблюдений достаточно **велико**.

Выборочное среднее пример **статистики** - функции от наблюдаемой выборки.
Центральная предельная теорема: пусть $X_1, ..., X_n$ - iid с математическим ожидание μ и дисперсией $V(X)$. Тогда для выборочного среднего $\bar{X}_n$:

$$\frac{\bar{X}_n - \mu}{\sqrt{V(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathbb{N}(0, 1)$$

ЦПТ

Альтернативная формулировка:

$$\bar{X}_n \sim \mathbb{N}(\mu, \sigma^2.)$$

Интерпретация ЦПТ: закон распределения выборочного среднего может быть приближен с помощью нормального распределения.

Оценки и статистики

Теория вероятностей: если задан закон распределения F_X, то как будет выглядеть выборка $X_1, ..., X_n$ и насколька вероятна ее конкретная реализация $x_1, ..., x_n$?

Математическая статистика: имея наблюдаемую выборку $x_1, ..., x_n$ из неизвестного распределения, как мы можем восстановить закон распределения F_X?

Статистика - функция от наблюдаемой выборки.

Оценка - статистика, которая используется для восстановления значений неизвестного параметра θ : $X_1, ...X_n \sim F_\theta$. Можно думать об этом так: кто-то задумал параметр θ и получил по нему реализацию $X_1 = x_1, ..., X_n = x_n$. По наблюдаемой выборке $x_1, ..., x_n$ требуется «отгадать» значение θ.

Выборочные статистики

Пример

Давайте рассмотрим примеры оценок неизвестных параметров, полученных по наблюдаемой выборке - ноутбук Sample Estimates.ipynb.

t-digest

Пример

Зачастую, в данных находятся «выбросы» - аномальные значения, которые могут сильно исказить значения статистики. Обычно, для обозначения устойчивости статистики к выбросам используют термин «робастность». Так, среднее не является робастной статистикой, а медиана - является. Однако, вычисление медианы обычно затруднено - она требует O(n) памяти для хранения выборки. Давайте рассмотрим структуру данных, которая позволяет сократить требования по памяти - t-digest.

Свойства оценок

Для оценок параметров всегда будем использовать **шапку**. Если $X_1, ..., X_n \sim F(\theta)$, то $\hat{\theta}_n = f(X_1, ..., X_n)$.
Несмещенность:

$$E(\hat{\theta}_n) = \theta.$$

Смещение оценки можно записать как:

$$bias(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta.$$

Ранее считалось важным свойством оценок, однако сейчас зачастую используют смещенные оценки которые сходятся к истинному значению θ.

Свойства оценок

Состоятельность:
Говорят, что оценка является состоятельной, если:

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

Распределение $\hat{\theta}_n$ называют выборочным распределением.

Стандартная ошибка оценки:

$$se(\hat{\theta}_n) = \sqrt{V(\theta_n)}.$$

Парадокс Симпсона

Пример

При интерперетации результатов исследований важно не допускать «повседневную» интерпретации причинно-следственных связей. Рассмотрим т.н. парадокс Симпсона в ноутбуке Simpson Paradox.ipynb.

MinHash

Пример

Как мы упоминули ранее, состоятельность оценки для нас является более важным показателем, нежели ее несмещенность. Рассмотрим это на примере алгоритма MinHash в ноутбуке MinHash.ipynb.

Точечные оценки

Назовем оценку асимптотически нормальной, если:

$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \sim \mathbb{N}(0, 1).$$

Точечная оценка - оценка, которая ставит в соответствие выборке единственное число $\hat{\theta}_n = f(X_1, ..., X_n)$.
Интервальная оценка:
$(1 - \alpha)$-доверительный интервал для параметра θ - интервал вида $C_n = (a, b)$, где a и b - статистики, такие что:

$$P(\theta \in C_n) \geq 1 - \alpha.$$

Неформально говоря, (a, b) «накрывает» θ с вероятностью 0.95.

План на следующую лекцию

План на следующую лекцию

- методы получения оценок: ММП и ММ; - интервальные оценки (примеры); - статистические гипотезы; - мощность критерия; - параметрические и непараметрические статистические методы.

Спасибо за внимание!