Введение в математическую статистику II

Computer Science Club, 27 ноября 2021

Данные

Данные d — это реализация случайного элемента D, имеющего (неизвестное) распределение $\mathcal{P}_D$.

Примеры:

- 1 Тб фотографий с кошечками и собачками
- Цена биткоина за последний месяц
- Результат броска монетки
- Количество антител у участников испытания вакцины

PDMI
COMPUTER
SCIENCE
CLUB

Статистическая модель

Статистическая модель — это множество распределений $\mathfrak{P}$, которое, по нашему мнению, адекватно приближает $\mathcal{P}_D$.

Примеры:

- d — результат 10 подбрасываний монетки, $\mathfrak{P} = \left\{B(p)^{\otimes 10} \mid p \in (0,1)\right\}$.
- d — 1 Тб фоточек, $\mathfrak{P}$ — все фоточки независимы, каждая фоточка имеет распределение, приближенно описываемое GAN'ом.

Статистическая модель

Статистические модели делят на

– параметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta | \theta \in \Theta \subset \mathbb{R}^k\}$;

  Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, \sigma^2) | \mu \in \mathbb{R}, \sigma^2 \geq 0\}$

– непараметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta | \theta \in \Theta \subset V\}$, V не обязано быть конечномерным;

  Пример: $\mathfrak{P} = \left\{\mathcal{P}^{\otimes n} \middle| \int_{\mathfrak{X}} x\mathcal{P}(dx) = 0\right\}$

– семипараметрические, если $\mathfrak{P} = \{\mathcal{P}_\theta | \theta \in \Theta \subset \mathbb{R}^k \times V\}$;

  Пример: линейная регрессия $Y = X\beta + \varepsilon$, $\beta \in \mathbb{R}^k$, $\mathbb{E}\varepsilon = 0$, $\mathbb{D}\varepsilon = \sigma^2$

В байесовской статистике в модель также включается априорное распределение на Θ.

Пример: $\mathfrak{P} = \{\mathcal{N}(\mu, 1) | \mu \sim \mathcal{N}(\mu_0, \sigma_0^2)\}$.

PDMI COMPUTER SCIENCE CLUB

Выборка

Если $D = [X_1, \dots, X_n]$, X_i независимы и имеют одинаковое распределение, то D называется выборкой объема n и обозначается как $X_{[n]}$. Распределение $\mathcal{P}_X$, которому подчинены все X_i, называется генеральной совокупностью.

Модель принимает вид $\mathfrak{P} = \left\{\mathcal{P}^{\otimes n} \middle| \mathcal{P} \in \mathfrak{P}_X\right\}$, где $\mathfrak{P}_X$ — модель для $\mathcal{P}_X$.

Предположение о том, что некоторый набор данных является выборкой — это сильное предположение; всегда нужно подумать о том, адекватно ли оно!

PDMI
COMPUTER
SCIENCE
CLUB

Почему выборка? Количество → качество

Для выборки $x_{[n]}$ с одномерным $\mathcal{P}_X$ рассмотрим

— $\mathcal{P}_n^*\left(A \middle| X_{[n]} = x_{[n]}\right) = \frac{1}{n}\sum_{i=1}^n [x_i \in A]$ — эмпирическое распределение,

— $F_n^*\left(x \middle| X_{[n]} = x_{[n]}\right) = \frac{1}{n}\sum_{i=1}^n [x_i < x]$ — эмпирическую функцию распределения.

Теорема Гливенко—Кантелли

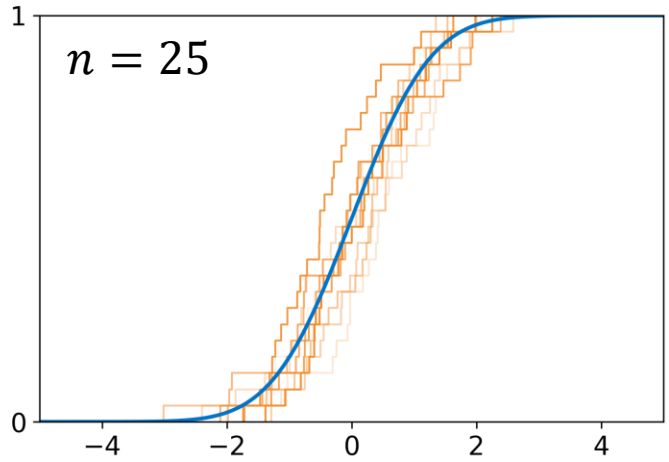
$$\sup_{x\in\mathbb{R}} |F_n^*(x) - F_X(x)| = \sup_{x\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n [X_i < x] - F_X(x)\right| \xrightarrow{\text{п.н.}} 0.$$

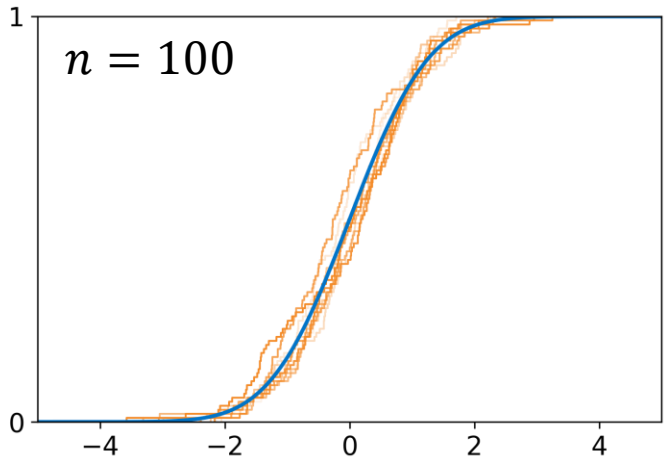
Теорема Колмогорова

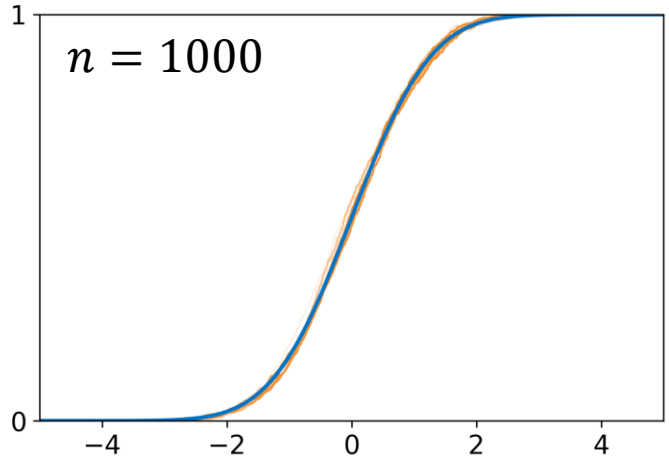
Если F_X непрерывна, то $\sqrt{n}\sup_{x\in\mathbb{R}} |F_n^*(x) - F_X(x)| \xrightarrow{d} K$ — распределение Колмогорова.

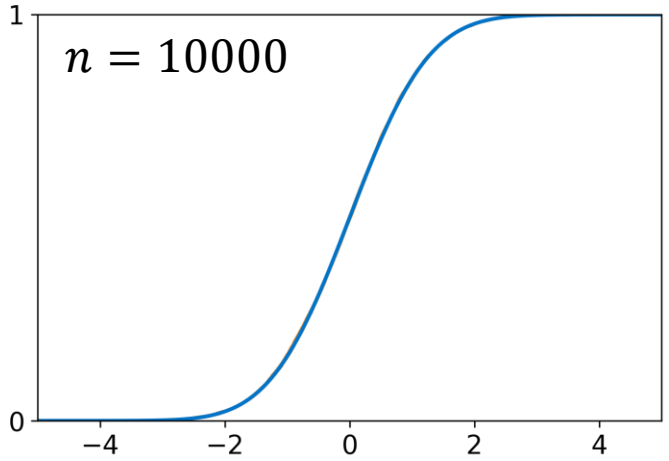
PDMI COMPUTER SCIENCE CLUB

Почему выборка? Количество → качество









PDMI
COMPUTER
SCIENCE
CLUB

Выборочные характеристики: моменты

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — выборочное среднее
- $\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ — выборочная дисперсия, $s^2 = \frac{n}{n-1}\tilde{s}^2$ — исправленная дисперсия
- $\tilde{s}$ — выборочное стандартное отклонение, s — исправленное стандартное отклонение
- $\alpha_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ — k-й выборочный момент, $\bar{x} = \alpha_1^*$
- $\mu_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ — k-й центральный выборочный момент, $\tilde{s}^2 = \mu_2^*$

PDMI
COMPUTER
SCIENCE
CLUB

Вариационный и статистический ряды

Отсортируем выборку: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Это вариационный ряд.

$x_{(k)}$ — k-я порядковая статистика.

Посчитаем уникальные значения: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \to z_{(1)} < \cdots < z_{(k)}$.

	$z_{(1)}$	$z_{(2)}$	…	$z_{(k)}$
кратность	n_1	n_2	…	n_k
частота	n_1/n	n_2/n	…	n_k/n

Это статистический ряд.

PDMI COMPUTER SCIENCE CLUB

Квантили, квартили, медиана

Квантиль уровня α — число x_α (любое) такое, что $\mathbb{P}(X \le x_\alpha) \ge \alpha$ и $\mathbb{P}(X \ge x_\alpha) \ge 1 - \alpha$.

Выборочный квантиль x^*_α:

- если $\frac{k}{n} < \alpha < \frac{k+1}{n}$, то $x^*_\alpha = x_{(k+1)}$,
- если $\alpha = \frac{k}{n}$, то x^*_α — любое число в интервале $[x_{(k)}, x_{(k+1)}]$.

Выборочный квартиль:

$Q_1 = x^*_{0.25}$, $Q_2 = x^*_{0.5}$, $Q_3 = x^*_{0.75}$

Выборочная медиана med^*:

- если $n = 2k + 1$, то $med^* = x_{(k+1)}$
- если $n = 2k$, то $med^* = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right)$

PDMI COMPUTER SCIENCE CLUB

Характеристики положения и разброса

Характеристики положения

- Выборочное среднее
- Выборочная медиана
- Выборочная мода
- Квантили, квартили…

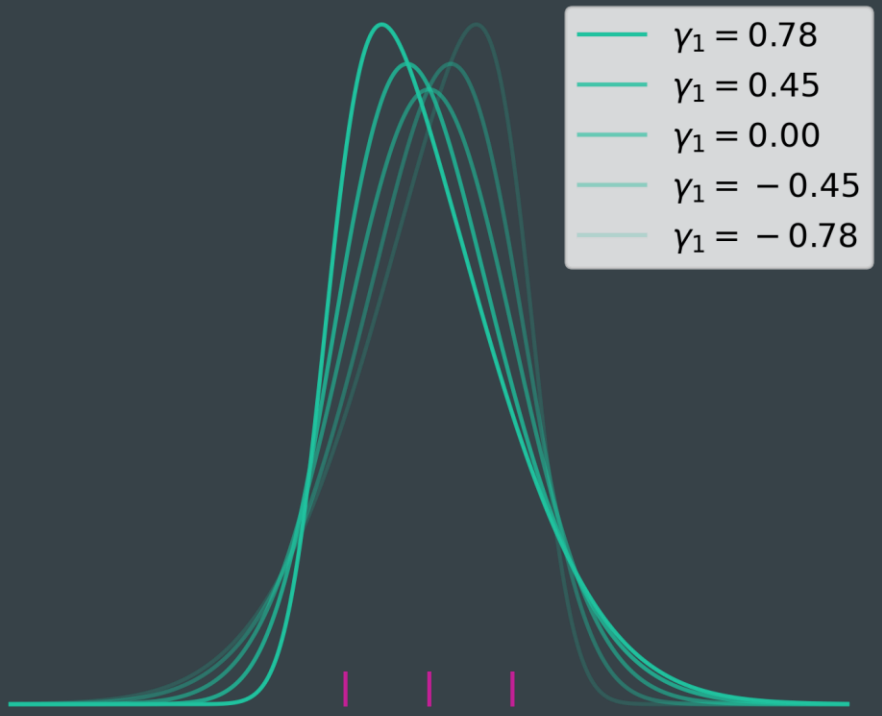
Характеристики разброса

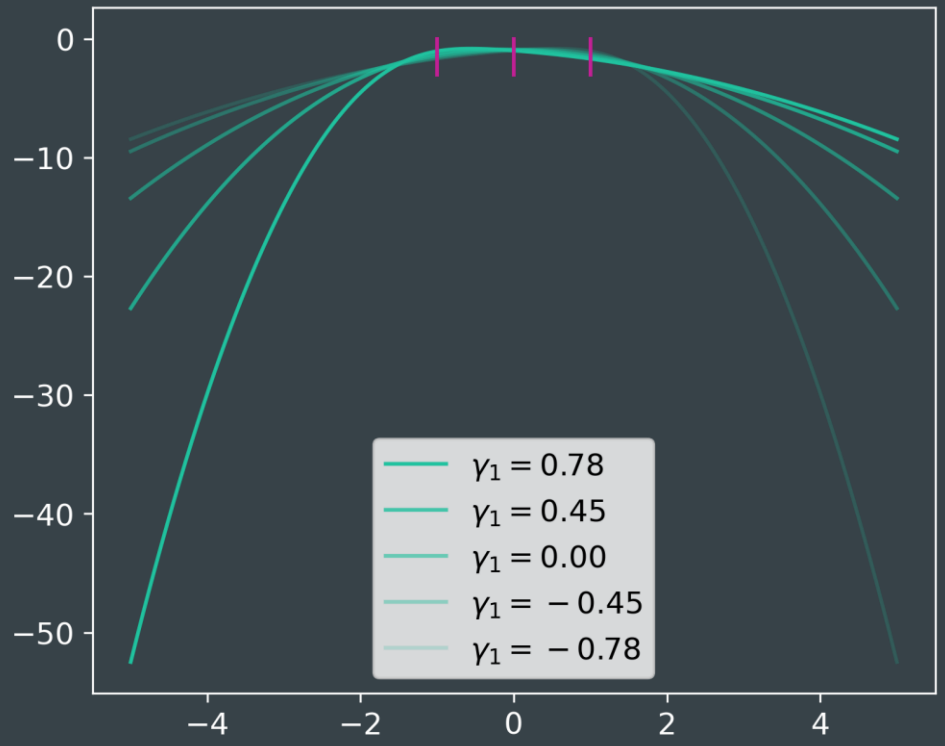
- Размах: $R = x_{(n)} - x_{(1)}$
- Межквартильный размах: $IQR = Q_3 - Q_1$
- Выборочное стандартное отклонение и исправленное стандартное отклонение
- Медианное абсолютное отклонение: $MAD = med(|x_{[n]} - med^*|)$

PDMI COMPUTER SCIENCE CLUB

Коэффициент асимметрии (skewness)

$$\gamma_1 = \frac{\mu_3^*}{(\mu_2^*)^{\frac{3}{2}}} = \frac{\mu_3^*}{\tilde{s}^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\tilde{s}}\right)^3$$

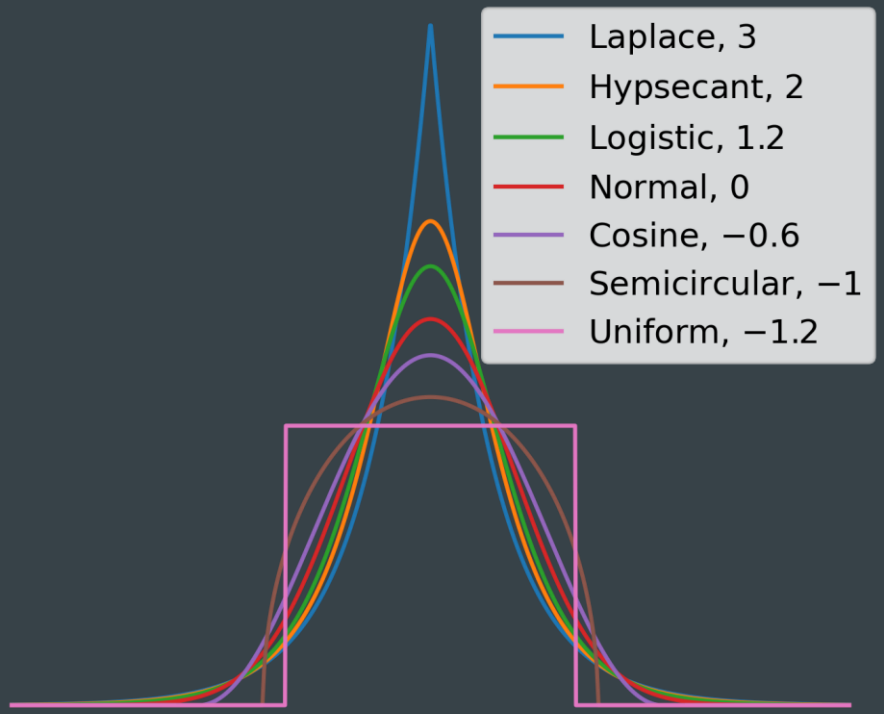


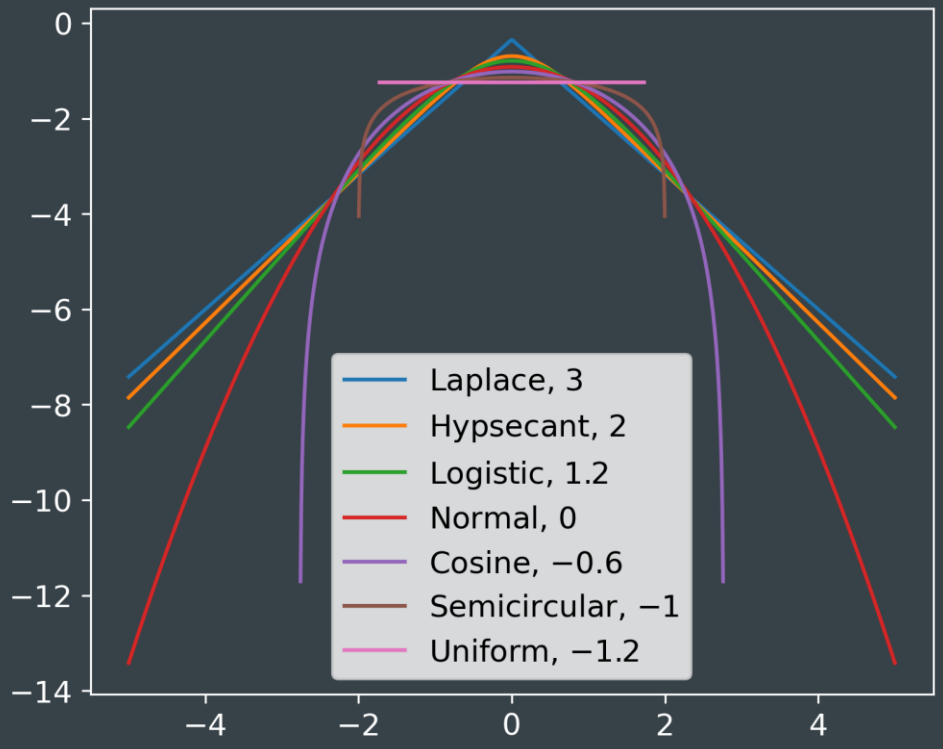


PDMI
COMPUTER
SCIENCE
CLUB

Коэффициент эксцесса (kurtosis)

$$\gamma_2 = \frac{\mu_4^*}{(\mu_2^*)^2} - 3 = \frac{\mu_4^*}{\tilde{s}^4} - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\tilde{s}}\right)^4 - 3$$





PDMI
COMPUTER
SCIENCE
CLUB

Метод Монте-Карло

Задача: Вычислить $\mathbb{E}f(X) = \int f(x)d\mathcal{P}_X(dx)$.

Идея: Если $\mathbb{E}f(X)$ существует, то по закону больших чисел $\frac{1}{N}\sum_{i=1}^{N} f(x_i) \xrightarrow{\text{п.н.}} \mathbb{E}f(X)$, где x_i — независимые реализации X.

Погрешность: Если $\mathbb{D}f(X)$ существует, то по центральной предельной теореме $\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(x_i) - \mathbb{E}f(X)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{D}f(X)\right)$.

Иначе говоря, $\frac{1}{N}\sum_{i=1}^{N} f(x_i) \approx \mathbb{E}f(X) + \xi\sqrt{\frac{\mathbb{D}f(X)}{N}}$, где $\xi \sim \mathcal{N}(0,1)$.

PDMI
COMPUTER
SCIENCE
CLUB

Примеры

- $\mathbb{P}(X \in A) = \mathbb{E}[X \in A]$, а значит для оценки нужно сгенерировать кучу реализаций x_i и посмотреть, какая доля лежит в A.
- $\mathbb{D}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 \approx \tilde{s}$
- Пусть мы не умеем генерировать реализации X, но знаем ее плотность p_X, и умеем генерировать реализации Y с плотностью p_Y. Тогда

$$\mathbb{E}f(X) = \int f(x)p_X(x)dx = \int f(x)\frac{p_X(x)}{p_Y(x)}p_Y(x)dx = \mathbb{E}f(Y)\frac{p_X(Y)}{p_Y(Y)}$$

$$\approx \frac{1}{N}\sum_i f(y_i)\frac{p_X(y_i)}{p_Y(y_i)}$$

PDMI COMPUTER SCIENCE CLUB

Статистический вывод

Введение

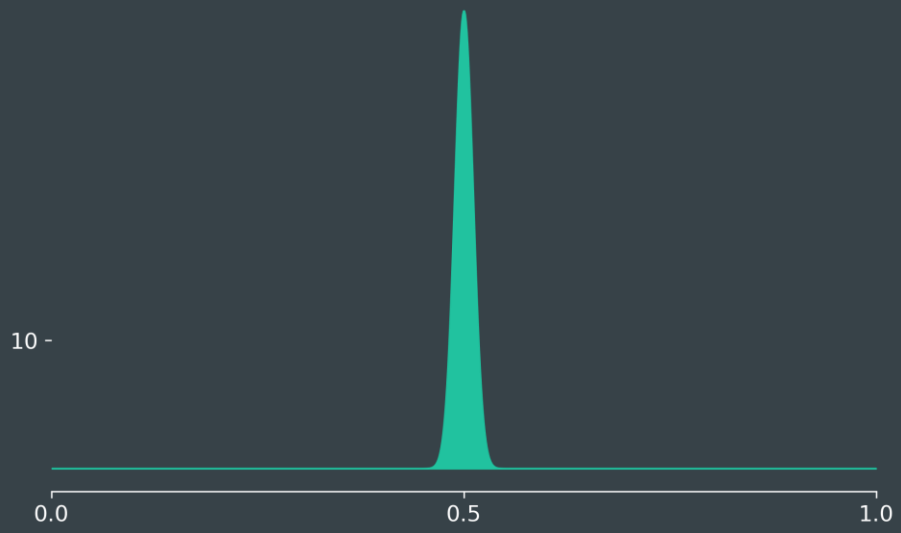
Задачи статистического вывода:

- Оценка параметров (и характеристик): точечные, интервальные
- Проверка гипотез
- Сравнение моделей
- ...

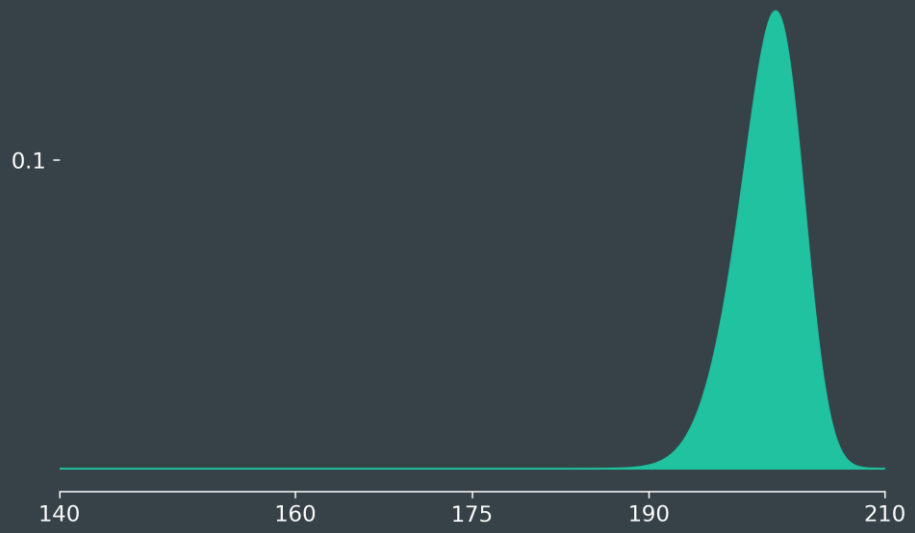
Байесовский подход

Основная идея — степень уверенности в значении величины численно выражаем вероятностью или плотностью.

Чем больше уверенность, тем больше вероятность.



вероятность выпадения орла
у монетки в вашем кармане



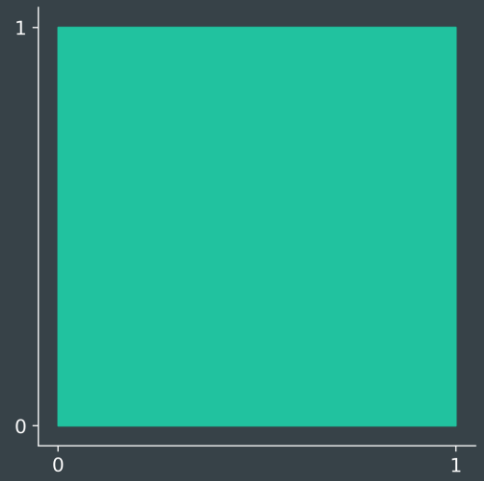
средний рост баскетболиста NBA

PDMI
COMPUTER
SCIENCE
CLUB

Априорное и апостериорное распределения

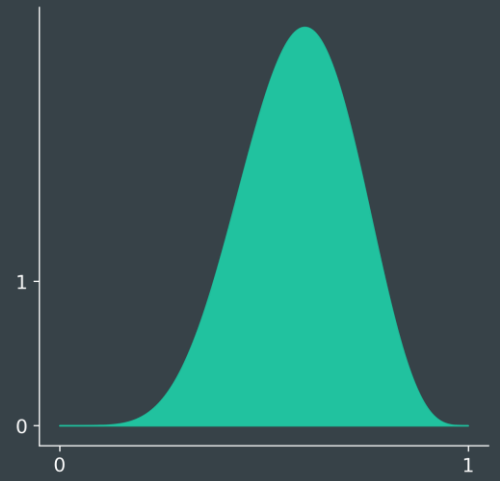
Наше ощущение относительно значения параметра до эксперимента называется априорным распределением.

В результате эксперимента мы получаем данные, которые некоторым образом меняют наше ощущение. Полученное ощущение называется апостериорным распределением.



априорное распределение

$+ \quad d = [0,1,1,0,1,0,1,1,1,0] \quad =$



апостериорное распределение

PDMI COMPUTER SCIENCE CLUB

Формула Байеса

p — вероятность или плотность
d — данные
θ — параметры

правдоподобие
likelihood

априорное распределение
prior

$$p(\theta|d) = \frac{p(d|\theta) \cdot p(\theta)}{p(d)}$$

posterior
апостериорное распределение

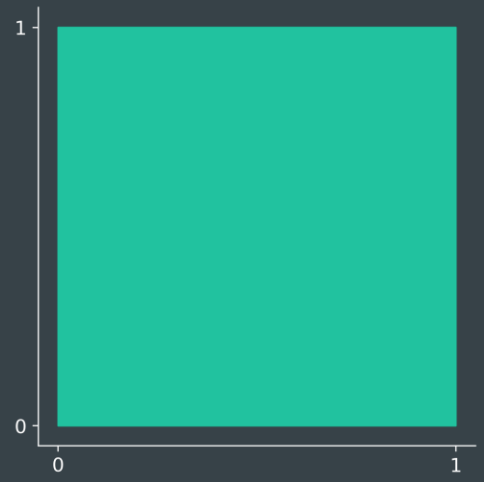
evidence
вероятность данных

нормирующая константа, не зависит от θ!

PDMI COMPUTER SCIENCE CLUB

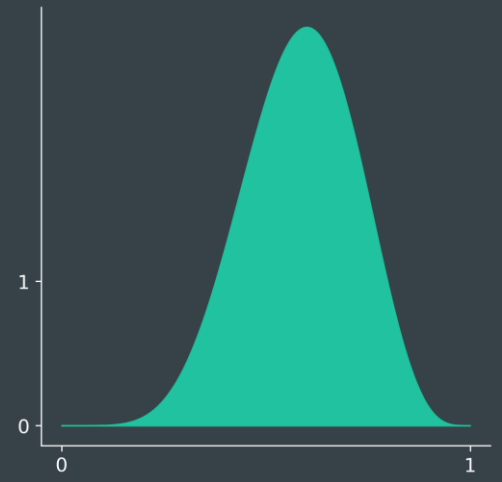
Априорное и апостериорное распределения

$$\Theta = [0,1], p(\theta) \equiv 1 \quad \times \quad p(d|\theta) = \theta^6(1-\theta)^4 \quad \propto \quad p(\theta|d) = \frac{\theta^6(1-\theta)^4}{\mathrm{B}(7,5)}$$



априорное распределение

$$+ \quad d = [0,1,1,0,1,0,1,1,1,0] \quad =$$



апостериорное распределение

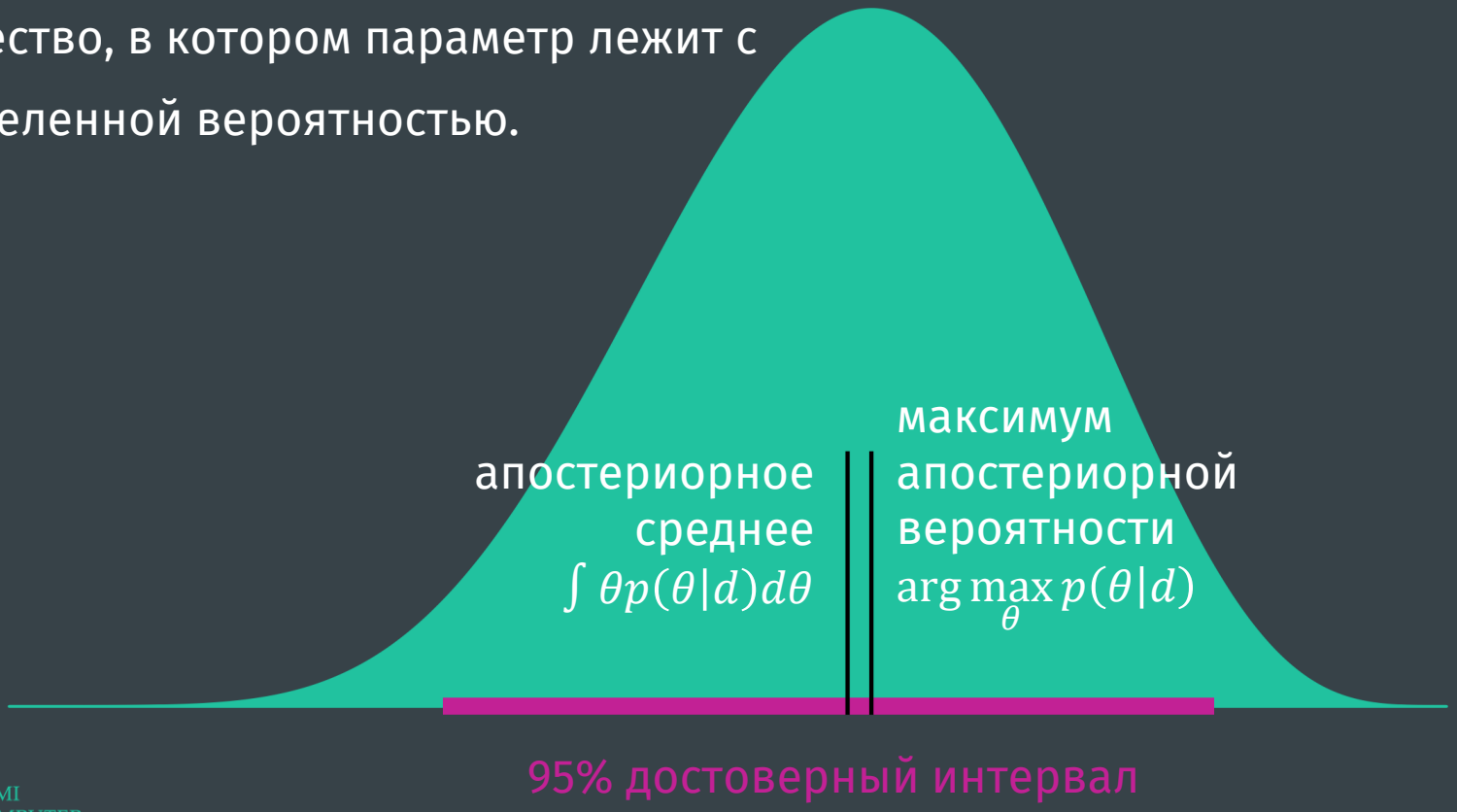
Распределение на $[0,1]$ с плотностью вида $p(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/\mathrm{B}(\alpha,\beta)$, $\alpha, \beta > -1$, называется бета-распределением с параметрами α, β и обозначается как $Beta(\alpha, \beta)$.

PDMI COMPUTER SCIENCE CLUB

Оценка параметров

По апостериорному распределению можно оценивать параметры и погрешности.

Достоверный интервал (credible interval) — множество, в котором параметр лежит с определенной вероятностью.



PDMI COMPUTER SCIENCE CLUB

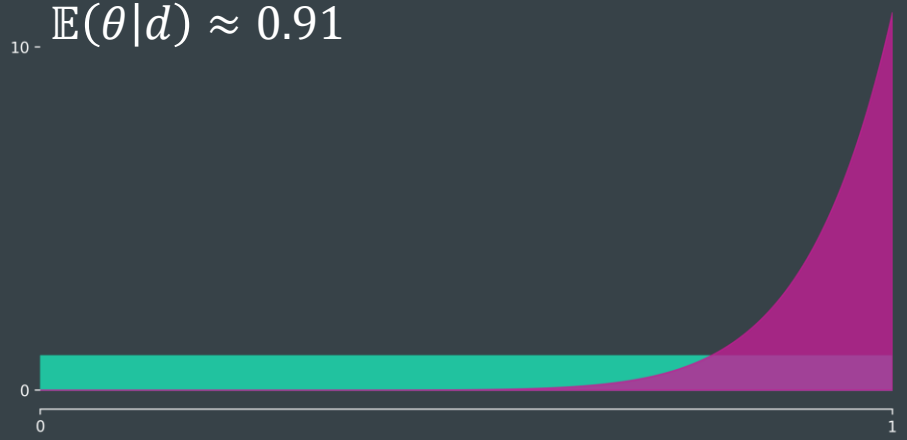
Виды априорных распределений

- Неинформативные — не несут дополнительной информации о значениях параметров; обычно это равномерные распределения на Θ. В таких случаях выводы о параметрах делаются только на основании данных.
- Информативные — несут существенную дополнительную информацию о значениях параметров. Если данных довольно мало, то выводы о параметрах могут в значительной степени определяться априорным распределением.
- Слабо информативные — несут частичную дополнительную информацию о значениях параметров. Основная идея — делать выводы на основании данных, лишь немного корректируя их в случае странных данных. Своего рода регуляризация.

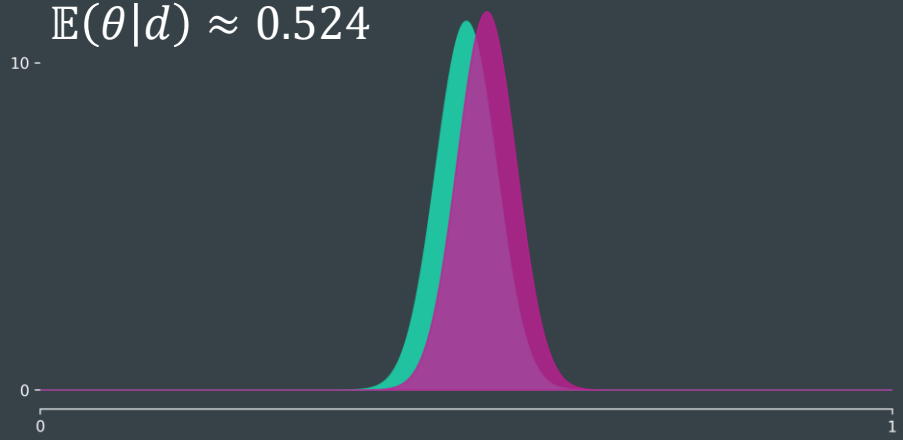
Виды априорных распределений

Пусть $d = [1,1,1,1,1,1,1,1,1,1]$ — десять бросков монетки.

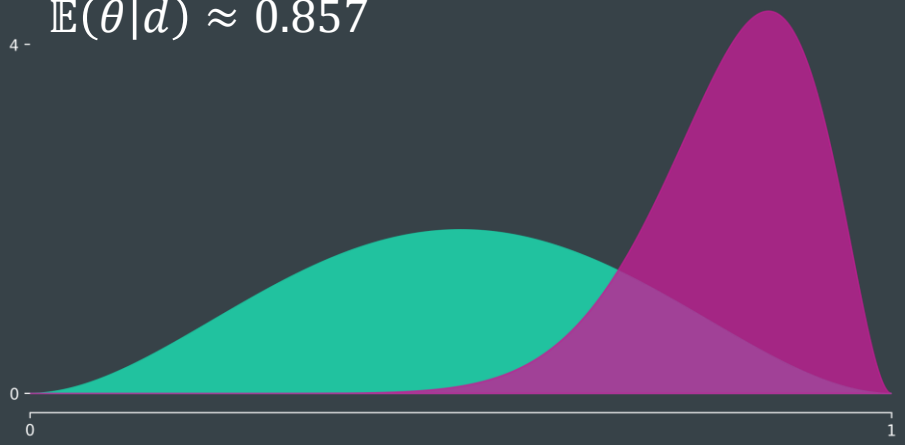
$\mathbb{E}(\theta|d) \approx 0.91$



$\mathbb{E}(\theta|d) \approx 0.524$



$\mathbb{E}(\theta|d) \approx 0.857$



PDMI
COMPUTER
SCIENCE
CLUB

Проверка гипотез

Имеются гипотезы $\mathfrak{P}_1, \ldots, \mathfrak{P}_k$ с априорными вероятностями $\mathbb{P}(\mathfrak{P}_1), \ldots, \mathbb{P}(\mathfrak{P}_k)$, а также данные d.

априорные шансы	$\mathbb{P}(\mathfrak{P}_1) : \mathbb{P}(\mathfrak{P}_2) : \ldots : \mathbb{P}(\mathfrak{P}_k)$
	$\times$
отношение правдоподобия	$\mathbb{P}(d\|\mathfrak{P}_1) : \mathbb{P}(d\|\mathfrak{P}_2) : \ldots : \mathbb{P}(d\|\mathfrak{P}_k)$
	$=$
апостериорные шансы	$\mathbb{P}(\mathfrak{P}_1\|d) : \mathbb{P}(\mathfrak{P}_2\|d) : \ldots : \mathbb{P}(\mathfrak{P}_k\|d)$

Если нас интересует, какая гипотеза скорее всего верна, то смотрим на апостериорные шансы.

Если нас интересует, какая гипотеза наиболее правдоподобна для наших данных, то смотрим на отношение правдоподобия.

PDMI COMPUTER SCIENCE CLUB

Сложная гипотеза

Как посчитать правдоподобие или апостериорную вероятность у сложной гипотезы?

В этом случае на пространстве параметров Θ задается априорное распределение $p(\theta|\mathfrak{P}_i)$ и правдоподобие вычисляется как интеграл по нему:

$$p(d|\mathfrak{P}_i) = \int_\Theta p(d|\theta, \mathfrak{P}_i) p(\theta|\mathfrak{P}_i) d\theta$$

Апостериорная вероятность — это

$$p(\mathfrak{P}_i|d) = \frac{p(d|\mathfrak{P}_i)p(\mathfrak{P}_i)}{\sum_j p(d|\mathfrak{P}_j)p(\mathfrak{P}_j)}.$$

PDMI
COMPUTER
SCIENCE
CLUB

Принятие решений

Задачу проверки гипотез можно рассматривать в рамках теории принятия решений: различные гипотезы соответствуют различным решениям, которые можно принять в данной ситуации, и вместе с этим появляются штрафы за принятие неверного решения.

Байесовская классификация

- Имеется множество объектов X и конечное множество имен классов Y.
- Множество $X \times Y$ является вероятностным пространством с известной плотностью распределения $p(x|y)p(y)$.
- Априорные вероятности классов $p(y)$ известны.
- Функции правдоподобия классов $p(x|y)$ известны.
- Требуется построить алгоритм $a: X \to Y$, минимизирующий потери от неверной классификации.

Воронцов К.В. Лекции по статистическим (байесовским) алгоритмам классификации

Функционал среднего риска

Зная функции правдоподобия можем для $A \subset \mathcal{X}$ посчитать

$$p(A|y) = \int_A p(x|y)dx.$$

Рассмотрим произвольный алгоритм $a: \mathcal{X} \to \mathcal{Y}$. Он разбивает $\mathcal{X}$ на непересекающиеся области $A_y = \{x \in \mathcal{X} | a(x) = y\}$.

Вероятность правильной классификации объекта класса y: $p(A_y|y)p(y)$.

Вероятность отнести объект класса y к классу y': $p(A_{y'}|y)p(y)$.

Каждой паре $(y, y') \in \mathcal{Y} \times \mathcal{Y}$ назначим штраф $\lambda_{y \to y'}$ при отнесении объекта класса y к объекту класса y'. Обычно $\lambda_{y \to y} = 0$ и $\lambda_{y \to y'} > 0$ при $y \neq y'$.

Функционал среднего риска

Примеры:

- $y = 0$, если человек здоров, $y = 1$, если болен: $\lambda_{0\to1} < \lambda_{1\to0}$.
- $y = 0$, если человек невиновен, $y = 1$, если виновен: $\lambda_{0\to1} > \lambda_{1\to0}$.

Средним риском называется ожидаемый штраф классификатора a:

$$R(a) = \sum_{y\in Y}\sum_{y'\in Y}\lambda_{y\to y'}p(A_{y'}|y)p(y).$$

Если $\lambda_{y,y'} = [y \neq y']$, то $R(a)$ это вероятность ошибки алгоритма a.

PDMI COMPUTER SCIENCE CLUB

Оптимальное байесовское правило

Теорема

Если известны $p(y)$ и $p(x|y)$, то минимум $R(a)$ достигается алгоритмом

$$a^*(x) = \arg\min_{y' \in Y} \sum_{y \in Y} \lambda_{y \to y'} p(x|y) p(y).$$

Если к тому же $\lambda_{y \to y} = 0$ и $\lambda_{y \to y'} = \lambda_y$ для $y \neq y'$, то

$$a^*(x) = \arg\max_{y \in Y} \lambda_y p(x|y) p(y).$$

Оптимальное байесовское правило

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y' \in Y} p(x|y')p(y')}.$$

Таким образом риск, связанный с x пропорционален $\sum_{y \in Y} \lambda_y p(y|x)$.

Оптимальный классификатор можно переписать так:

$$a^*(x) = \arg\max_{y \in Y} \lambda_y p(y|x).$$

Если $\lambda_y \equiv 1$ (т.е. $\lambda_{y \to y'} = [y \neq y']$), то данное правило называется принципом максимума апостериорной вероятности.

Если $p(y) \equiv 1/|Y|$, то $a^*(x) = \arg\max_{y \in Y} p(x|y)$ — принцип максимума правдоподобия.

PDMI COMPUTER SCIENCE CLUB

Оценка качества классификатора

Тестирование произвольного классификатора $a(x)$ на модельных данных:

1. задаем $p(x|y)$ и $p(y)$;
2. генерируем из распределения $p(x|y)p(y)$ обучающую и тестовую выборки

   $x_{[n]} = (x_i, y_i)_{i=1}^{n}$, $x_{[k]} = (x_i^t, y_i^t)_{i=1}^{k}$;
3. обучаемся на обучающей выборке,
4. по тестовой вычисляем эмпирическую оценку среднего риска:

   $R^*(a, x_{[k]}) = \sum_{y \in Y} \sum_{y' \in Y} \lambda_{y,y'} \frac{1}{k} \sum_{i=1}^{k} [a(x_i^t) = y' \wedge y_i^t = y]$,
5. сравниваем со средним риском $R(a^*)$ байесовского классификатора.

PDMI COMPUTER SCIENCE CLUB

Предсказательное распределение

Если мы умеем считать апостериорное распределение параметров $p(\theta|d)$, то можем посчитать и апостериорное распределение новых данных $p(d_{new}|d)$, которое называется предсказательным распределением:

$$p(d_{new}|d) = \int_\Theta p(d_{new}, \theta|d)d\theta = \int_\Theta p(d_{new}|\theta, d)p(\theta|d)d\theta\,.$$

Выглядит страшно, но оказывается, генерировать данные из этого распределения сильно проще, чем считать интегралы!

Чтобы сгенерировать новое наблюдение, нужно

1. Сгенерировать новый параметр θ_{new} из апостериорного распределения
2. Сгенерировать новое наблюдение d_{new} из $p(\cdot\,|\theta_{new})$

PDMI COMPUTER SCIENCE CLUB

Оценка качества модели

В байесовском подходе не любят вопрос «Верна ли модель или нет?», так как модель не может быть полностью верна.
Более правильный вопрос: «Оказывают ли недостатки модели существенное влияние на выводы?»

Если модель хорошая, то данные, сгенерированные из предсказательного распределения должны быть похожи на реальные данные: средние должны быть примерно средними, 50% достоверные интервалы должны содержать примерно 50% наблюдений и т.п.

Похожесть можно также определять с помощью функций потерь, как мы это делали раньше.

Оценка качества модели

Усредненный рецепт:

1. Выбираем функцию потерь T (ее еще называют статистикой критерия), которая отслеживает различие между данными и моделью в нужном нам смысле.
2. Считаем вероятность $\mathbb{P}(T(d_{new}) > T(d)|d)$ получить на сгенерированных данных более экстремальное значение, чем то, что мы посчитали на реальных данных d.

Если модель хорошо описывает данные, то эта вероятность должна быть около 0.5, если же она близка к 0 или 1, то это может быть следствием низкого качества модели (и тогда ее стоит доработатать), а может быть случайностью.

Величина $\mathbb{P}(T(d_{new}) > T(d)|d)$ называется апостериорное предсказательное p-значение (или p-value).

PDMI COMPUTER SCIENCE CLUB

Апостериорное предсказательное p-value

Как же считать это p-value? Теоретически, оно считается так:

$$p = \int\int [T(d_{new}) \geq T(d)] p(d_{new}|\theta) p(\theta|d)\; d(d_{new}) d\theta\,.$$

На практике обычно поступают так:

1. Генерируют N параметров $\theta_{new,i}$ из апостериорного распределения
2. Для каждого из них генерируют данные $d_{new,i} \sim p\left(\cdot \,\middle|\, \theta_{new,i}\right)$
3. Для каждого $d_{new,i}$ считают $T\left(d_{new,i}\right)$
4. Значение p приближенно равно доле $T\left(d_{new,i}\right)$, больших $T(d)$

PDMI COMPUTER SCIENCE CLUB

Пример

Пусть мы работаем с последовательностью бинарных величин, которые моделируем как независимые реализации бернуллиевской величины $B(\theta)$, $\theta \sim U([0,1])$. Как мы выяснили ранее, $p(\theta|d) \propto \theta^{\sum d_i}(1-\theta)^{n-\sum d_i}$.

Пусть мы получили такой набор данных:

$$d = [1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0]$$

Длинные последовательности одних и тех же значений вызывают некоторое сомнение в независимости реализаций.

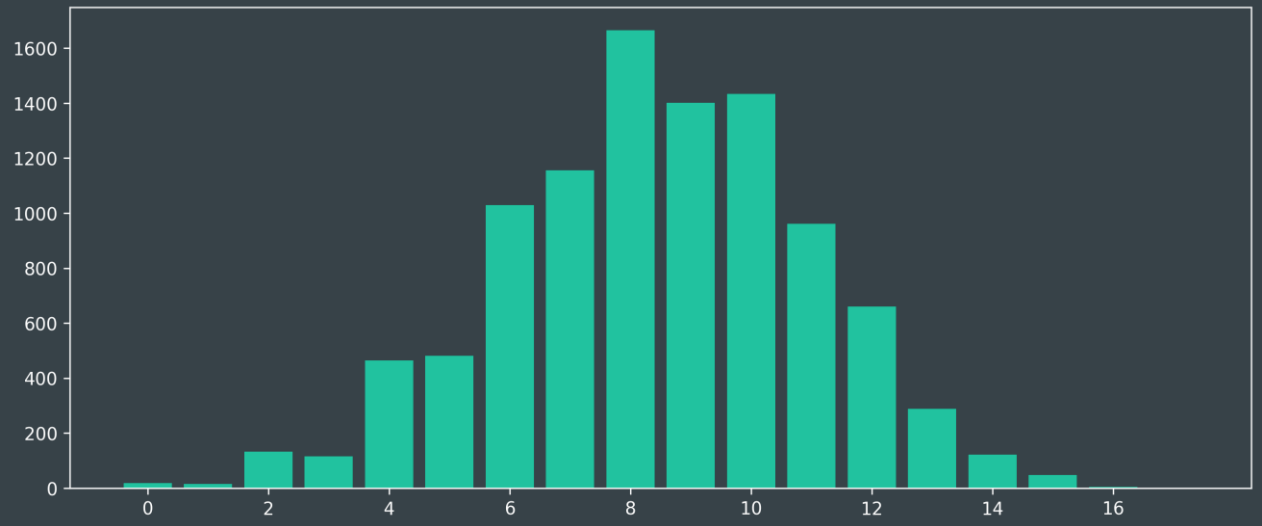
Чтобы разобраться с этим, рассмотрим величину $T =$ число смен 1 и 0.

Пример

Дальше делаем так:

1. Генерируем 10000 параметров $\theta_i \sim Beta(8,14)$.
2. Для каждого из них генерируем d_i: 20 реализаций из $B(\theta_i)$
3. Для каждого набора считаем $T(d_i)$
4. Значение p приближенно равно доле $T(d_i)$, не меньших $T(d) = 3$

Получилось $p \approx 0.983$



PDMI COMPUTER SCIENCE CLUB

Пример

Вспомним игровые автоматы. Попробуем посчитать $\mathbb{P}(\bar{X} < 0.384|d)$.

Рассмотрим Модель 3 с ограничением $\mathbb{E}\, reward(p_1, p_2, p_3) = 0.92$, где $p_i = (p_{i,0}, p_{i,1}, p_{i,2}, p_{i,3}, p_{i,7}, p_{i,c}, p_{i,dd})$ — вероятности окошка i.

Прежде, чем выбрать априорное распределение, посмотрим на функцию правдоподобия:

$$p(d|p_1, p_2, p_3) = \prod_{i=1}^{3} \prod_{x \in [0,1,2,3,7,c,dd]} p_{i,x}^{n_{i,x}},$$

где $n_{i,x}$ — количество картинок x в окошке i. Правдоподобие каждого окошка пропорционально плотности распределения Дирихле с параметрами $\{n_{i,x}\}$.

PDMI COMPUTER SCIENCE CLUB

Распределение Дирихле

Распределение Дирихле является обобщением Бета-распределения на многомерный случай. Оно живет на k-мерном симплексе $\{x \geq 0, \sum_i x_i = 1\}$, а его плотность равна

$$p(x_1, \dots, x_k | \alpha_1, \dots, \alpha_k) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

то есть она возвращает правдоподобие того, что вероятность каждого из k взаимоисключающих событий равна x_i при условии, что каждое событие наблюдалось $\alpha_i - 1$ раз.

PDMI
COMPUTER
SCIENCE
CLUB

Сопряженное семейство распределений

Если в качестве априорного распределения p_i выбрать распределение Дирихле с параметрами $(\alpha_0, \ldots, \alpha_{dd})$, то апостериорное распределение будет снова распределением Дирихле с параметрами $(\alpha_0 + n_0, \ldots, \alpha_{dd} + n_{dd})$.

Семейство априорных распределений, для которых апостериорные распределения принадлежат этому же семейству, называется сопряженным. Если нет какой-то веской причины выбрать специфическое априорное распределение, стараются брать распределение из сопряженного семейства, так как это удобно с вычислительной точки зрения.

Пример

В качестве априорного распределения на каждом окошке возьмем распределение Дирихле с параметрами

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_7, \alpha_c, \alpha_{dd}) = (2,1,1,1,1,0.5,0.5).$$

Тогда апостериорное распределение каждого окошка будет распределением Дирихле с параметрами

	α_0	α_1	α_2	α_3	α_7	α_c	α_{dd}
p_1	61	50	15	7	8	1.5	3.5
p_2	87	9	25	17	5	0.5	1.5
p_3	79	40	7	2	8	3.5	5.5

Ничего не забыли?

PDMI COMPUTER SCIENCE CLUB

Пример

Погодите, но у нас же еще $\mathbb{E}\, reward = 0.92$! Это значит, что наше априорное распределение живет не на симплексе, а на некотором его подмножестве.

Это обстоятельство значительно усложняет генерацию параметров, поэтому мы немного изменим нашу модель и допустим, что $\mathbb{E}\, reward \in [0.915, 0.925]$.

Теперь у нас есть возможность генерировать параметры с помощью генератора распределения Дирихле:

1. Генерируем p_i из своих апостериорных распределений Дирихле.
2. Если $\mathbb{E}\, reward(p_1, p_2, p_3) \in [0.915, 0.925]$ — ура, мы сгенерировали параметр. Иначе выкидываем его и возвращаемся к шагу 1.

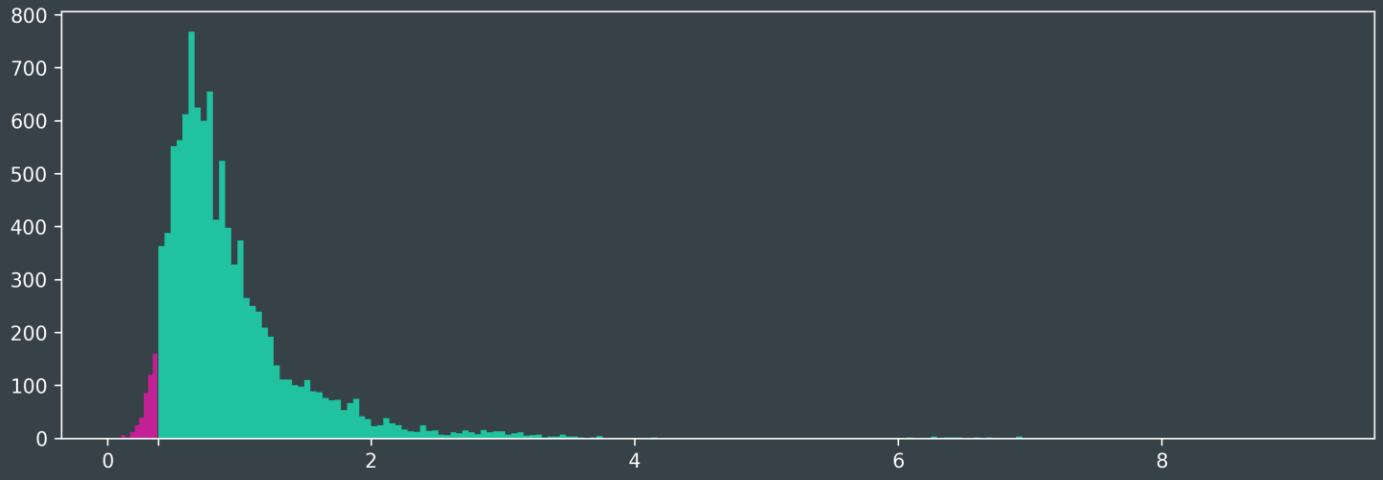
Такой метод называется сэмплированием с отклонением.

Пример

Итак, возвращаемся к исходной задаче:

1. Генерируем $N = 10000$ параметров из честного апостериорного распределения
2. Для каждого из них генерируем выборку из 138 игр в игровом автомате
3. Для каждой выборки считаем средний выигрыш
4. Находим долю средних выигрышей, меньших 0.384

Получилось $p \approx 0.0425$



PDMI
COMPUTER
SCIENCE
CLUB

Проверка на новых данных

Если долго улучшать модель, то можно получить модель, которая описывает скорее имеющиеся данные, нежели распределение, их породившее.

Такое явление называется переобучением.

Чтобы этого избежать, можно проверить модель на новых данных, которые она (и мы!) не «видели». Проверка на новых данных является фактически единственным способом достоверной оценки качества модели.