Байесовское мультимоделирование: распределения и их свойства

Московский Физико-Технический Институт

2021

Что такое случайная величина

Заданы:

- Множество элементарных событий Ω
- Сигма-алгебра $\mathfrak{F}$ на множестве Ω
- Система $\mathfrak{B}(\mathbb{R})$ борелевских множеств на $\mathbb{R}$

Действительная функция $w(\omega) : \Omega \to \mathbb{R}$ называется случайной величиной, если для любого $B \in \mathfrak{B}(\mathbb{R})$:

$$\{\omega : w(\omega) \in B\} \in \mathfrak{F}.$$

Описание случайных величин

Дискретная случайная величина w принимает счётное множество значений $A = \{a_1, a_2, \dots\}$ с вероятностями $p_1, p_2, \dots, \sum_i p_1 = 1$.

$f(a_i) = p(w = a_i) = p_i$ — **функция вероятности**.

Непрерывная случайная величина задаётся с помощью **функции распределения**:

$$F_w(t) = p(w \leq t)$$

или **плотности распределения**:

$$f(w) : \int_a^b f(w)\, dw = p(a \leq w \leq b).$$

Метод максимума правдоподобия

$$\boldsymbol{X} \sim f(\boldsymbol{x}, w),$$

$$L(\boldsymbol{X}, w) = \prod_{x \in \boldsymbol{X}} f(x, w),$$

$$\hat{w} \equiv \arg\max_{w} L(\boldsymbol{X}, w).$$

Удобно прологарифмировать:

$$\log L(\boldsymbol{X}, w) = \sum_{x \in \boldsymbol{X}} \log f(x, w),$$

$$\hat{w} \equiv \arg\max_{w} \log L(\boldsymbol{X}, w).$$

Производные функции правдоподобия

Score function:

$$S(w) \equiv \frac{\partial}{\partial w} \log L(w)$$

ОМП — решение score equation:

$$S(w) = 0$$

Информация Фишера:

$$I(w) \equiv -\frac{\partial^2}{\partial w^2} \log L(w)$$

Дисперсия ОМП:

$$\mathbb{D}\hat{\theta} \approx I^{-1}(\hat{w})$$

Свойства ОМП

- состоятельность:
$$\hat{w}_n \to^p w$$
- асимптотическая нормальность: при $n \to \infty$
$$\hat{w} \sim \mathcal{N}\left(w, I^{-1}(w)\right)$$
- эффективность: ОМП имеют наименьшую дисперсию среди всех состоятельных оценок
- инвариантность: $g\left(\hat{w}\right)$ — ОМП-оценка для $g\left(w\right)$

Максимизация правдоподобия

Максимизация правдоподобия эквивалентна минимизации дивергенции между целевым и истинным распределением:

$$\max_{w} L(\boldsymbol{X}, w) \iff \min KL\left(p^{*}\left(\boldsymbol{X}\right)|p\left(\boldsymbol{X}|w\right)\right).$$

Идея доказательства

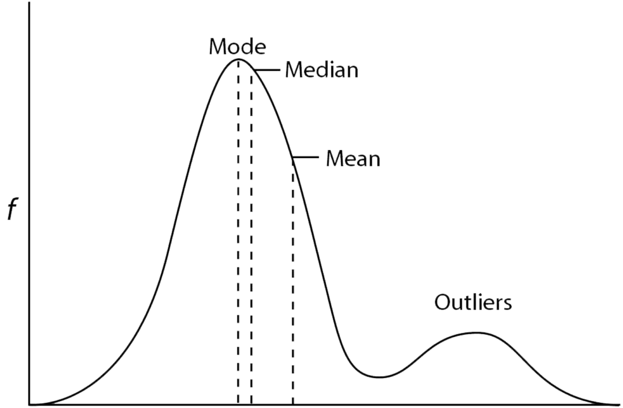
$$KL\left(p^{*}\left(\boldsymbol{X}\right)|p\left(\boldsymbol{X}|w\right)\right) = \mathsf{E}_{x\sim \boldsymbol{p}^{*}(\boldsymbol{X})} \log\left(\frac{p^{*}\left(\boldsymbol{X}\right)}{p\left(\boldsymbol{X}|w\right)}\right) =$$

$$= \mathrm{Const} - \mathsf{E}_{x\sim \boldsymbol{p}^{*}(\boldsymbol{X})} \log p\left(\boldsymbol{X}|w\right) \approx^{\text{ЗБЧ}}$$

$$\approx \mathrm{Const} - L(\boldsymbol{X}, w).$$

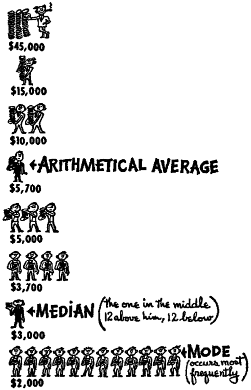
Оценки центральной тенденции

Выборочное среднее — среднее арифметическое по выборке.
Выборочная медиана — центральный элемент вариационного ряда.
Выборочная мода — самое распространённое значение в выборке.



Оценки центральной тенденции

(Huff, 1954):



Медиана

- **квантиль** порядка $\alpha \in (0,1)$:

$$w_\alpha \colon \quad p\,(w \le w_\alpha) \ge \alpha, \;\; p\,(w \ge w_\alpha) \ge 1-\alpha$$

эквивалентное определение:

$$w_\alpha = F^{-1}(\alpha) = \inf\{w \colon F(w) \ge \alpha\}$$

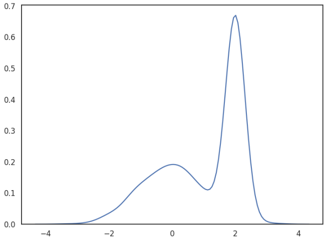
- **медиана** — квантиль порядка 0.5, центральное значение распределения:

$$\text{median}(w) \colon \quad p\,(w \le \text{median}(w)) \ge 0.5, \;\; p\,(w \ge \text{median}(w)) \ge 0.5$$

Мода

Точка максимума функции вероятности или плотности:

$$\mathrm{mode}(w) = \arg\max_{w} f(w)$$



Распределение может иметь несколько мод:

$$w \sim \alpha_1 \mathcal{N}(\mu_1, \sigma_1^2) + \alpha_1 \mathcal{N}(\mu_2, \sigma_2^2).$$

Матожидание

Среднее значение случайной величины w:

$$\mathsf{E}w = \int w\, dF(w).$$

- Может быть не определено;
- линейно;
- не зависит от событий меры нуль;
- ЗБЧ:
$$\bar{w}_n \to_{n\to\infty} \mathsf{E}w;$$
- ЦПТ:
$$\sqrt{n}\frac{\bar{w}_n - \mathsf{E}w}{\sqrt{\mathsf{D}w}} \to \mathcal{N}(0,1).$$

Матожидание: сходимость

Часто решается задача о сходимости матожидания сложной функции от случайно величины:

$$w \to^{\text{d/p/п.н.}} w^*,$$

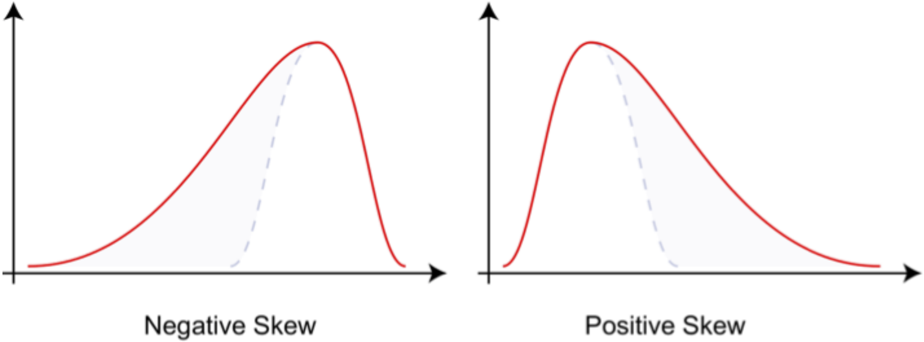
$$\mathsf{E}f(w) \to \mathsf{E}f(w^*)?$$

- f — непрерывная и ограниченная, тогда $\mathsf{E}f(w) \to^d \mathsf{E}f(w^*)$.
- f — непрерывна почти всюду, тогда $f(w) \to^{\text{d/p/п.н.}} f(w^*)$ (Теорема Манна-Вальда).
- Сходимость почти наверное: теорема Лебега о мажорируемой сходимости для переноса предела под знак матожидания.

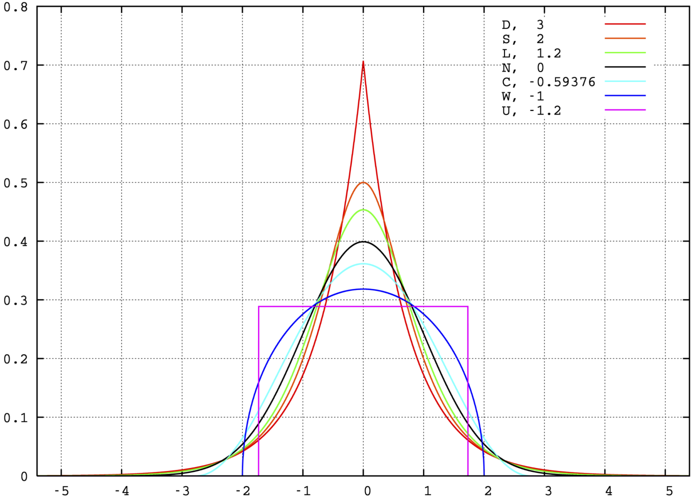
Прочие моменты

- Второй момент (дисперсия): $\mathsf{D}w = \mathsf{E}(w - \mathsf{E}w)^2$ — мера разброса значений случайной величины.
- Третий момент: $\frac{\mathsf{E}(w-\mathsf{E}w)^3}{\mathsf{D}w^{3/2}}$ — коэффициент асимметрии случайно величины.
- Четвертый момент $\frac{\mathsf{E}(w-\mathsf{E}w)^4}{\mathsf{D}w^2} - 3$ — коэффициент эксцесса, мера остроты пика распределения случайной величины.

Прочие моменты

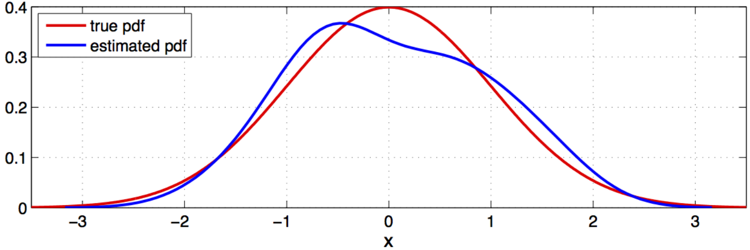


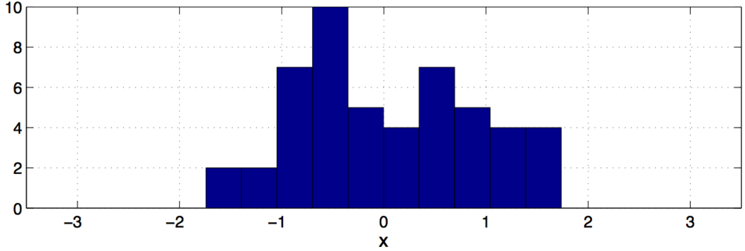
Прочие моменты



(wikipedia)

Оценка плотности распределения

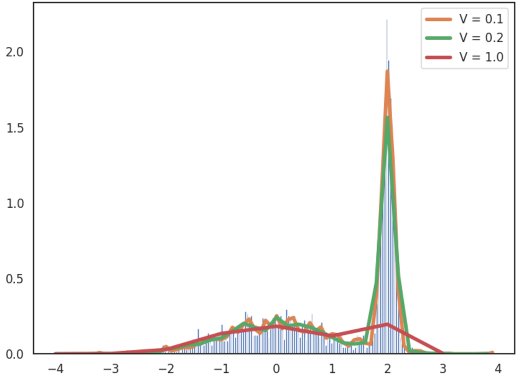




KDE

Оценка строится с использованием ядерной функции k:

$$p(x) = \frac{1}{N} \sum_{x_i \in \boldsymbol{X}} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

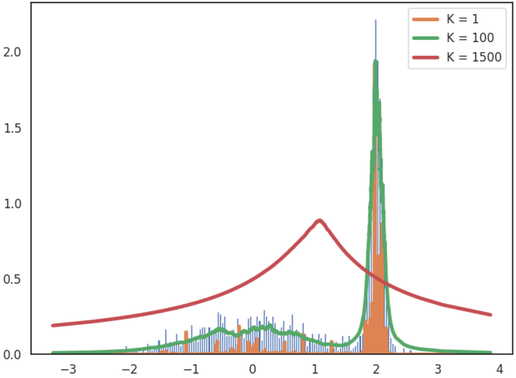


KDE

KNN как обобщение KDE:

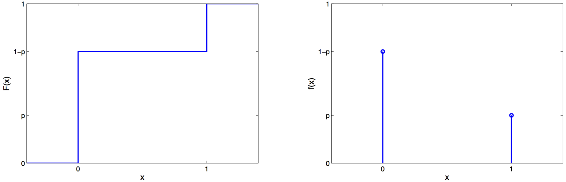
$$p(x) = \frac{1}{N} \sum_{x_i \in \boldsymbol{X}} \frac{1}{h} K\left(\frac{x - x_i}{h_i}\right),$$

$$h_i = \text{dist}\left(x_i, \text{neighbour}\left(x_i, K\right)\right)$$



Распределение Бернулли

$w \in \{0, 1\} \sim Ber\,(p)\,, \quad p \in (0, 1)$



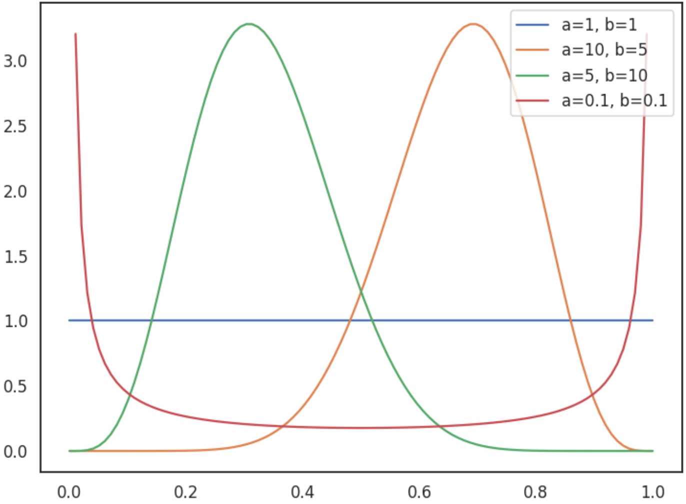
$$F\,(w) = \begin{cases} 0, & w < 0, \\ 1 - p, & 0 \le w < 1, \\ 1, & w \ge 1. \end{cases}$$

$$f\,(w) = \begin{cases} 1 - p, & w = 0, \\ p, & w = 1. \end{cases}$$

- пример: результат подбрасывания монеты

Бета-распределение

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}$$



Бета-распределение

- соответствует *априорным* ожиданиям о распределении Бернулли
- интерпретация: "эффективное количество наблюдений $w = 1, w = 0$"
- при $n \to \infty$ сходится к δ-распределению в точке ОМП распределения Бернулли.

Мультиномиальное распределение

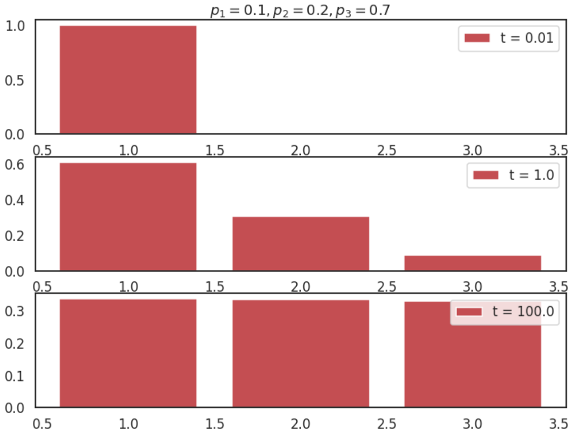
- Обобщение распределения Бернулли на большее количество событий:

$$p(w = w_i | p_1, \dots, p_n) = p_i.$$

- Возможная параметризация: через softmax:

$$\mathbf{softmax}(\log \boldsymbol{p}, t) = \frac{\mathbf{exp}(-\mathbf{log}\boldsymbol{p} \cdot t^{-1})}{\sum_{i=1}^{n} \exp(-logp_i \cdot t^{-1})}$$

Мультиномиальное распределение



Распределение Дирихле

$$p(w_1, \dots, w_K, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} w_i^{\alpha_i - 1}$$

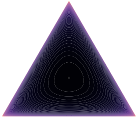
- Обобщение бета-распределения на многомерный случай
- Определяется на $K-1$-симплексе:

$$w_i \geq 0, \sum w_i = 1.$$

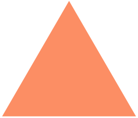
- Интерпретация: вероятность каждого из K взаимоисключающих событий равна w_i при условии, что каждое событие наблюдалось $\alpha_i - 1$ раз
- Альтернативная параметризация параметров:

$$\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}} \cdot t, ||\boldsymbol{\alpha}|| = 1.$$

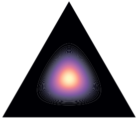
Распределение Дирихле



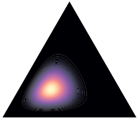
$\bar{\boldsymbol{\alpha}} = [1, 1, 1], t = 0.9$



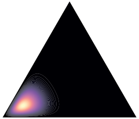
$t = 1.0$



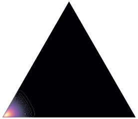
$t = 10.0$



$\bar{\boldsymbol{\alpha}} = [0.5, 0.25, 0.25], t = 30.0$



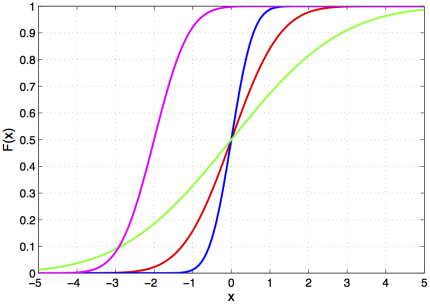
$[0.75, 0.125, 0.125]$

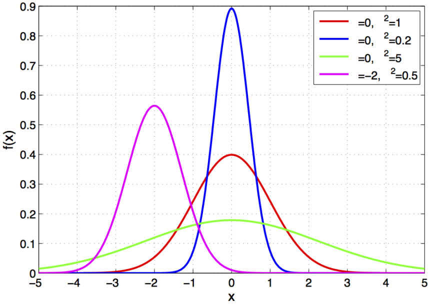


$[0.9, 0.05, 0.05]$

Нормальное распределение

$w \in \mathbb{R} \sim \mathcal{N}\left(\mu, \sigma^2\right), \ \sigma^2 > 0$





$$F\left(w\right) = \Phi\left(\frac{w-\mu}{\sigma}\right)$$

$$f\left(w\right) = \frac{1}{\sigma}\phi\left(\frac{w-\mu}{\sigma}\right)$$

$$\Phi\left(w\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w} e^{-\frac{t^2}{2}}dt$$

$$\phi\left(w\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}$$

Нормальное распределение

- предельное распределение суммы слабо взаимозависимых сл. в.
- $\mathsf{E}w = \mathrm{median}(w) = \mathrm{mode}(w) = \mu$, $\mathsf{D}w = \sigma^2$, все моменты более высокого порядка нулевые
- пусть $w_1, \ldots, w_n$ независимы, $\quad w_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$, тогда $\forall a_1, \ldots, a_n$

$$\sum_{i=1}^{n} a_i w_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

- Нормальное распределение имеет максимальную дифференциальную энтропию среди всех непрерывных распределений, дисперсия которых не превышает заданную величину.
- пример: погрешность измерения

Распределение стьюдента

- $\mathsf{E}w = 0$ при $\nu > 1$, $\mathrm{median}(w) = \mathrm{mode}(w) = 0$ всегда
- пусть $Z \sim N(0,1)$ и $V \sim \chi^2_\nu$ независимы, тогда

$$T = \frac{Z}{\sqrt{V/\nu}} \sim St(\nu)$$

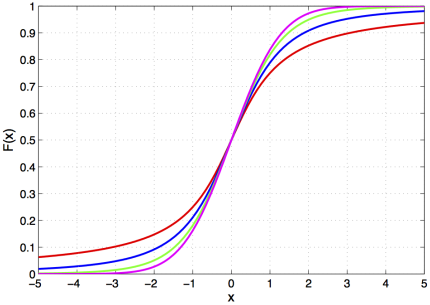
- если $w \sim St(\nu)$, то

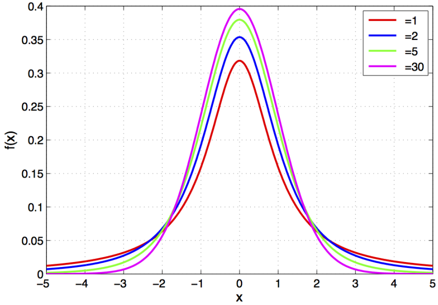
$$Y = \lim_{\nu\to\infty} w \sim \mathcal{N}(0,1)$$

- возникает при оценке среднего значения сл. в. с неизвестной дисперсией

Распределение стьюдента

$X \in \mathbb{R} \sim St\left(\nu\right), \ \nu > 0$





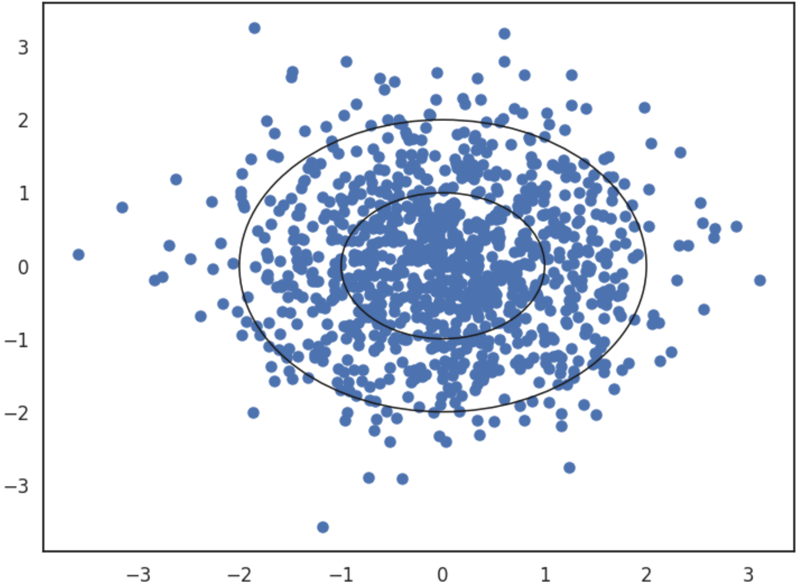
$$F\left(x\right) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right)$$

$$f\left(x\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Периодические распределения

- Наивный подход: использовать произвольное распределение для периодических данных
  - Пример, когда это плохо: выборка состоит из двух объектов на окружности: (1 градус и 359 градусов)
  - $\mathsf{E}w = 180$, $\mathsf{D} = 179$

Периодические распределения



Периодические распределения

- Распределение фон Мизеса:

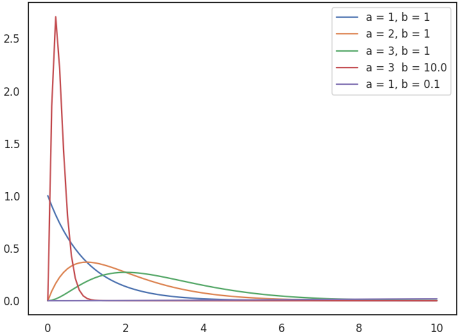
$$p = \frac{\exp(k \cos(x-\mu))}{2\pi I_0(k)}, \quad I_0 \text{ —функция Бесселя.}$$

- Идея распределения:
    - $[w_1, w_2] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
    - Параметризуем их через полярные координаты и ограничим радиус:

$$w_1 = r\cos\phi, w_2 = r\sin\phi, r = \text{Const}.$$

Гамма-распределение

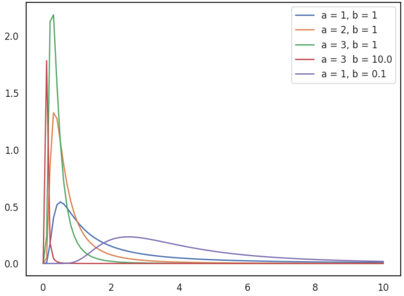
$$p(w, \alpha, \beta) = \frac{\beta^{\alpha} w^{\alpha-1} e^{-\beta}}{\Gamma(\alpha)}$$



- Используется для моделирования априорных предположений о величине, обратной среднеквадратичному отклонению.

Обратное гамма-распределение

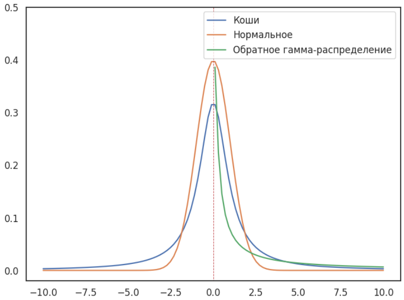
$$p(w, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{-\alpha-1} \exp(-\beta \cdot w^{-1}).$$



- $w \sim \Gamma(\alpha, \beta) \rightarrow w^{-1} \sim \mathrm{Inv} - \Gamma(\alpha, \beta^{-1})$
- Используется для моделирования априорных предположений о среднеквадратичном отклонении.

Распределение Коши

$$p(w, w_0, \gamma) = \frac{\gamma}{\pi\left((w - w_0)^2 + \gamma^2\right)}.$$



- Используется как альтернатива нормальному распределению или обратному гамма-распределению
- Имеет тяжелые хвосты
- Моменты не определены

Литература и прочие ресурсы

- Ширяев А. Н. Вероятность-1. – МЦНМО, 2007. – С. 552-552.
- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- Минимизация KL: https://wiseodd.github.io/techblog/2017/01/26/kl-mle/
- Numpy
- Scipy