Архитектурный документ. Ядро PySATL

Михаил Михайлов, Леонид Елкин

13 сентября 2025 г.

Оглавление

Раздел 1

Введение

1.1. Назначение системы

Вычислительное ядро проекта PySATL предназначено для представления и обработки вероятностных распределений в программной форме. Ядро предоставляет средства для задания распределений и их семейств, выполнения операций над ними и построения более сложных структур путём функциональных и алгебраических преобразований.

Система поддерживает задание как конкретных распределений с определёнными параметрами, так и абстрактных семейств, из которых могут быть получены конкретные экземпляры. Кроме того, предусмотрена возможность определения пользовательских распределений и преобразований, расширяющих базовые возможности.

Ядро служит универсальной основой для статистических и вероятностных вычислений в рамках проекта PySATL и может использоваться другими подсистемами при построении моделей.

1.2. Область применимости

Вычислительное ядро `core` используется во всех подсистемах проекта PySATL, где требуется работа с распределениями вероятностей. Оно предназначено как для непосредственного вычисления характеристик распределений (например, плотности, функции распределения, квантилей), так и для построения и трансформации более сложных моделей на их основе. Оно может быть использовано:

- при определении конкретных распределений, используемых в анализе данных;
- для задания пользовательских распределений, комбинации распределений и создания новых семейств;
- при трансформации распределений через функциональные отображения;
- в задачах символьной или численной обработки распределений.

Вне проекта PySATL ядро может быть применимо в любых системах, где необходима гибкая и расширяемая работа с вероятностными распределениями, особенно в контексте численных симуляций, статистического моделирования и прикладного машинного обучения.

1.3. Общие сведения о системе

Ядро представляет собой модульную систему, реализованную на языке Python, предназначенную для работы с вероятностными распределениями, их семействами и преобразованиями. Архитектура системы построена на разделении функциональности между независимыми компонентами, каждый из которых отвечает за определённый класс задач.

Компоненты взаимодействуют друг с другом через чётко определённые интерфейсы. Такая организация позволяет изолировать ответственность отдельных частей системы, обеспечивать гибкость при расширении функциональности и облегчать сопровождение кода.

Некоторые интерфейсы, предоставляемые одним модулем, используются другими модулями для построения более сложных вычислений или абстракций. Это позволяет комбинировать базовые элементы в составные структуры и формировать цепочки преобразований.

Проект спроектирован таким образом, чтобы допускать расширение без модификации существующих компонентов, в соответствии с принципами модульности и открытости/закрытости. Это обеспечивает стабильную основу для развития ядра и его интеграции с другими частями проекта PySATL.

Раздел 2

Глоссарий

Основным понятиям в статистике и стохастическом моделировании является *распределения случайной величины* или, более общо, *распределения случайного объекта* (далее, под случайной величиной понимается любой случайный объект, реализации которого не обязательно суть вещественные числа) [18]. Ниже изложены основные теоретические сведения касающиеся случайных величин, а также задач в которых они возникают, в соответствии с монографиями [37] и [39].

Случайные величины и способы их задания

Для случайной величины ξ, принимающей значения в некотором пространстве $\mathcal{X}$, её распределением называется (см. [39]) вероятностная мера $\mathbb{P}_\xi(\cdot)$ на $\mathcal{X}$, такая что $\mathbb{P}_\xi(A)$ есть вероятность того что реализация ξ попадет в множество $A \subset \mathcal{X}$[1]. Как правило (см. [37]), выделяют следующие виды случайных величин

- Дискретные случайные величины. В этом случае $\mathcal{X}$ представляет собой некоторое дискретное (конечное или счетное) множество. Например число выпадений монеты орлом при нескольких бросках (биномиальное распределение); уровень образования у случайно выбранного человека (категориальное распределение); случайная величина которая принимает одно значение (вырожденное распределение);
- Одномерные непрерывные случайные величины[2]. В этом случае $\mathcal{X} = \mathbb{R}$ или $\mathcal{X} \subset \mathbb{R}$ ненулевой меры. Такие величины используются для описания случайных времен, расстояний и т.д. Согласно [18] наиболее важными представителями являются: равномерное распределение $\mathcal{U}(a; b)$, нормальное распределение $\mathcal{N}(\mu, \sigma^2)$ и экспоненциальное распределение $\mathrm{Exp}(\lambda)$;
- Многомерные непрерывные случайные величины. В этом случае $\mathcal{X} \subseteq \mathbb{R}^d$, ненулевой меры. Во много многомерные случайные величины являются аналогами одномерных непрерывных случайных величин, однако решение стандартных задач, таких как моделирование или вычисление числовых характеристик затруднено из-за проклятия размерности [7].

Отдельное направление статистики работает с данными о направлении (англ. directional data), в связи с этим часто можно также отдельно выделить следующую категорию случайных величин.

- Случайные геометрические примитивы (англ. geometrical random primitives). Примерами таких случайных величин служат случайные углы или случайные матрицы симметрий. Согласно [29], геометрической случайной величиной называется случайная величина принимающая значения на замкнутой и ограниченной поверхности в евклидовом пространстве (более общо—компактном Римановом многообразии).

С точки зрения ПО, работа с распределением, как с вероятностной мерой, является неудобной, так как компьютер не может работать с произвольными множествами. Однако, как правило, с распределением можно связать некоторую функцию, которая полностью определяет распределение. Так, чтобы идентифицировать распределение дискретной случайной величины, достаточно знать *функцию вероятности*, определяемую равенством (pmf).

$$f_\xi(x) = \mathbb{P}_\xi(\{x\}), \text{ т.е. вероятность того что } \xi = x,\ x \in \mathcal{X} \tag{pmf}$$

Если на пространстве возможных значений $\mathcal{X}$ задана некоторая мера μ, плотностью распределения $\mathbb{P}_\xi$ относительно μ называется[3] такая функция $f_\xi(x)\colon \mathcal{X} \to \mathbb{R}$, что выполнено тождество (pdf):

$$\mathbb{P}_\xi(A) = \int_A f_\xi(x)\, d\mu \tag{pdf}$$

В случае если μ это считающая мера, плотность f_ξ определяется равенством (pmf), в случае если $\mu = m_{\mathrm{Leb}}$, говорят просто о плотности непрерывной случайной величины/случайного вектора.

[1] Строго говоря, $\mathcal{X}$ должно быть снабжено некоторой σ-алгеброй $\mathcal{F}$, и $\mathbb{P}_\xi$ должна быть определена только для $A \in \mathcal{F}$. Иначе говоря, тройка $(\mathcal{X}, \mathcal{F}, \mathbb{P}_\xi)$ должна образовывать вероятностное пространство.

[2] Здесь и далее под непрерывными случайными величинами подразумеваются абсолютно-непрерывные случайные величины, т.е. распределение которых имеет плотность относительно меры Лебега

[3] Условия существования плотности описываются теоремой Радона-Никодима, см. например [21]

Несмотря на то что плотность распределения полностью его характеризует, для того чтобы вычислять вероятности $\mathbb{P}_\xi(A)$ необходимо производить интегрирование (или суммирование), поэтому для некоторых задач представление распределения в виде плотности является неудобным. В частности, если ξ — случайная величина (дискретная или непрерывная) принимающие значения из $\mathbb{R}^d$, довольно часто приходится смотреть на вероятность попадания в некоторую ячейку $\langle \mathbf{a}; \mathbf{b} \rangle$. Под ячейкой подразумевается множество:

$$\langle \mathbf{a}; \mathbf{b} \rangle = \left\{ \begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix} \in \mathbb{R}^d \,\middle|\, a_1 < c_1 \leq b_1, \cdots, a_d < c_d \leq b_d \right\}$$

Для доступа к таким вероятностям эффективнее работать с *функцией распределения случайной величины*, определяемой равенством (cdf).

$$F_\xi(\mathbf{x}) = \mathbb{P}_\xi(\langle -\infty; \mathbf{x} \rangle) \tag{cdf}$$

В этом случае $\mathbb{P}_\xi(\langle a; b \rangle)$ выражается через значения $F_\xi(\cdot)$ с помощью формулы включения-исключения [37].

С понятием функции распределения тесно связано понятие квантильной функции. Для случайной величины ξ со значениями из $\mathbb{R}$, её квантильная функция определяется равенством (ppf) (подробно о различных определениях см. в обзоре [20]).

$$\omega_\xi(p) = \inf_u \{ F_\xi(u) \geq p \} \tag{ppf}$$

Такое определение гарантирует, что случайная величина $\omega_\xi(U), U \sim \mathcal{U}[0; 1]$ имеет такое же распределение как и сама величина ξ. В случае когда $F_\xi(\cdot)$ строго возрастает на всей области определения, квантильная функция является обратной функцией $\omega_\xi(\cdot) = F_\xi^{-1}(\cdot)$. Отдельно следует отметить что существуют обобщения понятия квантильной функции на случай случайных величин со значениями из $\mathbb{R}^d$ [12], однако для их вычисления необходимо решать уравнения в частных производных [10].

Существуют и другие функциональные характеристики распределения, многие из которых приходят из анализа выживаемости [17]. *Функцией выживаемости* называется функция определяемая равенством (sdf).

$$S_\xi(\mathbf{x}) = 1 - F_\xi(\mathbf{x}) \tag{sdf}$$

Для случайных величин со значениями из $\mathbb{R}$, функцией интенсивности отказов и кумулятивной функцией интенсивности отказов называются функции определяемые равенствами (hrdf) и (chdf) соответственно.

$$h_\xi(x) = -\frac{S'_\xi(x)}{S_\xi(x)}; \tag{hrdf}$$

$$H_\xi(x) = -\ln(S_\xi(x)); \tag{chdf}$$

Однако, некоторые распределения, например α-устойчивые распределения [37], не допускают явного задания с помощью плотности или функции распределения, однако допускают задания с помощью так называемых *интегральных преобразований*. Такие распределения все чаще возникают в современных моделях стохастического анализа, (см. например [4]). В случае случайной величины ξ со значениями из $\mathbb{R}$, её

- Характеристической функцией ξ называется преобразование Фурье, определяемое равенством (cf).

$$\varphi_\xi(u) = \int_{\mathbb{R}} \exp(itu) \mathbb{P}_\xi(dt), \quad u \in \mathbb{R} \tag{cf}$$

- Момент-производящей функцией называется преобразование определяемое равенством (mgf).

$$M_\xi(u) = \int_{\mathbb{R}} \exp(tu) \mathbb{P}_\xi(dt), \quad u \in \mathbb{R} \tag{mgf}$$

Характеристическая функция всегда существует и полностью определяет распределение случайной величины [39]. В свою очередь, момент-производящая функция существует не всегда, но в тех случаях когда существует, также однозначно характеризует распределение. В случае когда ξ принимает только неотрицательные значения, определены также преобразование Лапласа и преобразование Меллина, задаваемые равенствами (lt) и (mt) соответственно.

$$\mathcal{L}_\xi(u) = \int_{\mathbb{R}_+} \exp(-tu) \mathbb{P}_\xi(dt), \quad u \in \mathbb{R}_+ \tag{lt}$$

$$\mathcal{M}_\xi(u) = \int_{\mathbb{R}_+} t^u \mathbb{P}_\xi(dt), \quad u \in \mathbb{R}_+ \tag{mt}$$

Эти преобразования также однозначно характеризуют распределение ξ [37], [11]. В случае если ξ многомерная случайная величина, также определяется характеристическая функция (см. [39]), преобразования (mgf), (lt), (mt) в некоторых ситуациях допускают обобщение на многомерный случай, см. например [2].

На рис. 2.1 схематично изображены основные способы задания непрерывных вероятностных распределений, и связь между ними. В связи с тем что в теории многомерных квантилей нет результатов напрямую выражающих квантильные функции через плотности или интегральные преобразования, переходы которые имеют место быть только в одномерном случае, изображены пунктирными стрелками.

Рис. 2.1: Способы задания непрерывных распределений

Замечение. Плотность и функция распределения непрерывной случайной величины со значениями в $\mathbb{R}^d$ связаны соотношениями

$$f_\xi(\mathbf{x}) = \frac{\partial F_\xi}{\partial x_1 \cdots \partial x_d}(\mathbf{x}) \quad F_\xi(\mathbf{x}) = \int\limits_{\langle -\infty;\mathbf{x}\rangle} f_\xi(\mathbf{t})\, d\mathbf{t} \tag{2.1}$$

Формулы обращения для интегральных преобразований представлены в [39], [11]. Отдельно стоит отметить, что в работе [34] показано как можно вычислять квантильную функцию по плотности распределения и наоборот, не прибегая к вычислению функции распределения. Этот подход может оказаться полезным при работе с достаточно сложными плотностями.

Семейства вероятностных распределений

В задачах статистики, как правило, оперируют не с одним каким-то конкретным распределением, а с набором распределений, из которого надо выбрать наиболее подходящее, или проверить какую-то гипотезу. Более строго, *параметрическим семейством распределений* называется некоторое множество $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ распределений, зависящих от скалярного или векторного параметра θ, Θ — множество возможных значений параметра [38].

Для любого распределения $\mathbb{P}_\xi$ случайной величины ξ определено семейство локации и масштаба, т.е. семейство распределений всех аффинных преобразований величины ξ:

$$\mathrm{loc} + \mathrm{scale}\cdot\xi \sim \mathbb{P}^{\xi}_{(\mathrm{loc},\mathrm{scale});} \tag{loc-scale-family}$$

где параметры $\mathrm{loc}, \mathrm{scale} \in \mathbb{R}$ для вещественнозначных случайных величин, и $\mathrm{loc} \in \mathbb{R}^d$, $\mathrm{scale} \in \mathbb{R}^{d\times d}$ для векторозначных случайных величин. Примером такого семейства является семейство нормальных распределений $\mathcal{N}(\mu,\sigma)$, определяемых равенством (normal-family).

$$\mu + \sigma\cdot\xi, \quad \xi\sim\mathcal{N}(0,1), \text{ т.е. } f_\xi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{normal-family}$$

Более общим понятием является понятие семейства замкнутого относительно действия группы, см. [23] и [28].

Другим, в некотором смысле ортогональным, понятием является понятие экспоненциального семейства распределений [3]. Параметрическое семейство распределений $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ относится к экспоненциальному типу, если плотности (или функции вероятностей) которых можно записать в виде (exp-family) [4].

$$f(\mathbf{x}|\theta) = \exp(\langle \mathbf{T}(\mathbf{x}), \vec{\eta}(\theta)\rangle + A(\mathbf{x}) + D(\theta)) \tag{exp-family}$$

Многие распространенные семейства распределений являются экспоненциальными, см. [27] . Для моделей относящихся к экспоненциальным семействам существует богатая теория оценивания параметров [23]. Большой список параметрических семейств и связывающие их соотношения представлены в [22].

Приведенные выше семейства интересны с точки зрения теоретической статистики. С точки зрения прикладной статистики, интерес представляют распределения, которые допускают гибкость в плане оценивания параметров: так, для нормального распределения два параметра не только локацию и масштаб, но и всю форму распределения, причем такое поведение присуще не только нормальному распределению. Для того чтобы решить эту проблему, было предложено несколько гибких семейств распределений, среди которых широко распространены семейство распределений Пирсона [8] и металогическое семейство [16].

[4] При этом требуется чтобы множество точек $\mathbf{x}$, в которых плотность отлична от 0, не зависело от параметра θ

Отдельно стоит отметить, что многие параметрические семейства зачастую имеют несколько параметризаций, каждая из которых может быть удобна в том или ином контексте, например в работе [30] приведены четыре параметризации для обобщенного гиперболического распределения. Множество других различных параметризаций для одних и тех же семейств собраны в базе проекта ProbOnto [35].

Преобразования случайных величин

Во многих моделях распределения могут быть составлены из более простых распределений с помощью различных методов. В [3] отмечается что, в контексте статистического вывода, любая модель для данных может рассматриваться как вероятностное распределение. Существует множество комбинировать и преобразовывать вероятностные распределения, интересная практическая реализация этого взгляда доступна в библиотеке Pomegranate, [32]. Ниже рассмотрены основные способы для непрерывных вещественнозначных случайных величин, большинство которых относят к теории алгебры случайных величин, см. [33].

- *Аффинное преобразование.* Если случайная величина ξ имеет распределение с функцией плотности $f_\xi(x)$, то плотность её линейного преобразования $a\xi + b$ описывается равенством (aff-tr);

$$f_{a+b\xi}(y) = \frac{1}{|a|} f_\xi\left(\frac{y-b}{a}\right), \quad a \neq 0. \tag{aff-tr}$$

- *Биективное преобразование.* Плотность распределения случайной величины $g(\xi)$, где g — строго монотонная функция,определяется через обратную функцию $g^{-1}(y)$ с помощью равенства (bij-tr).

$$f_{g(\xi)}(y) = f_\xi(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right|. \tag{bij-tr}$$

  Уравнение (aff-tr) является частным случаем (bij-tr). Для немонотонных функций распределение вычисляется с использованием разбиения на участки монотонности;

- *Распределение суммы независимых случайных величин* Если случайные величины ξ_1 и ξ_2 независимы[5], то плотность суммы $\xi_1 + \xi_2$ вычисляется с помощью свёртки (sum-rv).

$$(f_{\xi_1} \oplus f_{\xi_2})(z) = \int_{\mathbb{R}} f_{\xi_1}(x) f_{\xi_2}(z-x)\, dx. \tag{sum-rv}$$

- *Распределение произведения независимых случайных величин.* Для независимых случайных величин ξ_1 и ξ_2, плотность их произведения $\xi_1 \cdot \xi_2$ задается с помощью мультипликативной свертки (prod-rv).

$$(f_{\xi_1} \odot f_{\xi_2})(z) = \int_{\mathbb{R}} \frac{1}{|x|} f_{\xi_1}(x) f_{\xi_2}\left(\frac{z}{x}\right) dx. \tag{prod-rv}$$

Другие две важные операции возникают при работе со случайными векторами — маргинализация (взятие проекции) и вычисление порядковых статистик. Умение вычислять такие преобразования позволяет получать точные оценки качества в некоторых моделях статистического вывода.

- *Проекции.* Для случайного вектора $\xi = (\xi_1, \xi_2, \ldots, \xi_d)$ (с возможно зависимыми компонентами) с функцией распределения $F_\xi(\mathbf{x})$, функция распределения случайного подвектора $(\xi_{i_1}, \ldots, \xi_{i_k})$, описывается пределом (proj-tr)

$$F_{\Pr(\xi; i_1, \ldots, i_k)}(x_{i_1}, \ldots, x_{i_k}) = \lim_{\substack{x_i \to +\infty \\ i \neq i_1, \ldots, i_k}} F_\xi(x_1, \ldots, x_d) \tag{proj-tr}$$

- *Длины и углы.* Как правило, под случайными вектором понимается случайный элемент $\mathbb{R}^d$ представимый своими координатами. Однако, в некоторых ситуациях (см. например [13]) куда удобнее оперировать со сферическими или другими координатами. В этом случае работает многомерный аналог формулы (bij-tr).

Отдельно стоит отметить что важную роль играют порядковые статистики (см. [24]). Если компоненты случайного вектора $\xi = (\xi_1, \xi_2, \ldots, \xi_d)$ независимы и одинаково распределены с плотностью распределения $f(x)$ и функцией распределения $F(x)$, то их порядковые статистики $\xi_{(1:d)} \leq \xi_{(2:d)} \leq \ldots \leq \xi_{(n)}$ имеют плотности, описываемые равенством (ord-stat).

$$f_{\xi_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1-F(x)]^{n-k} f(x). \tag{ord-stat}$$

[5] Т.е. $\mathbb{P}(\xi_1 \in B_1 \text{ и } \xi_2 \in B_2) = \mathbb{P}(\xi_1 \in B_1) \cdot \mathbb{P}(\xi_2 \in B_2)$ верно для всех $B_1, B_2 \subset \mathbb{R}$, являющихся борелевскими

Помимо трансформации одного распределения в другое и алгебраических операций над распределениями, еще одним способом образования сложного распределения из нескольких простых является образование *смесей*. Смеси используется для построения сложных вероятностных моделей, в которых присутствуют скрытые параметры и широко применяются в кластерном анализе и для изучения ядерных оценок плотности. Можно выделить два класса смесей; дискретные и непрерывные.

- *Дискретная смесь*. Под дискретной смесью подразумевают комбинацию конечного числа распределений, каждое из которых взвешено определённым коэффициентом. Функция распределения дискретной смеси случайных величин $\xi_1, \ldots, \xi_n$ с весами $\vec{w} = (w_1, \ldots, w_n)$ задаётся равенством (dmix);

$$F_{\mathrm{mix}(\vec{w};\xi)}(x) = \sum_{i=1}^{n} w_i F_{\xi_i}(x), \quad \sum_{i=1}^{n} w_i = 1 \tag{dmix}$$

- *Непрерывная смесь* является непрерывным аналогом дискретной смеси. Пусть $F(x \mid \theta)$ семейство плотностей, зависящее от параметра $\theta \in \Theta \subset \mathbb{R}^n$. Если на Θ задано некоторое распределение параметров с плотностью $\omega(\theta)$, непрерывная смесь плотностей $F(x \mid \theta)$ определяется равенством (cmix)

$$F_{\mathrm{mix}(\omega;F_\xi(\cdot|\theta))}(x) = \int_\Theta F_\xi(x \mid \theta) \cdot w(\theta)\, d\theta, \tag{cmix}$$

  В данном случае параметр θ рассматривается как случайная величина с заданным распределением.

Общая теория смесей в абстрактном случае и конкретные примеры приведены в работе [6].

В приложениях часто возникают понятия цензурированных и урезанных распределений [9]. Для распределения $\mathbb{P}_\xi$ случайной величины ξ принимающей вещественные значения, урезанным называется условное распределение, определяемое равенством (truncated-dist).

$$\mathbb{P}_{\mathrm{Truncated}(\xi,L,R)}(B) = \frac{\mathbb{P}_\xi([L;R] \cap B)}{\mathbb{P}_\xi([L;R])} \tag{truncated-dist}$$

Распределение (truncated-dist) это условное распределение ξ если априори известно, что значение ξ лежит в отрезке $[L;R]$. В свою очередь, цензурированным на отрезке $[L;R]$ распределением называется распределение случайной величины, определяемой равенством (censored-dist).

$$.\mathrm{Censored}(\xi, L, R) = \begin{cases} L & \xi < L \\ \xi & L \le \xi \le R \\ R & R < \xi \end{cases} \tag{censored-dist}$$

Такие распределения часто возникают в задачах регрессионного анализа, см. [5].

Числовые характеристики вероятностных распределений

Для анализа моделей важную роль играют не только функциональные, но и числовые характеристики распределений. Согласно [9], [36], для случайной величины ξ со значениями из $\mathbb{R}$ можно выделить следующие характеристики.

- Меры центральной тенденции, описывающие, вокруг какого значения сконцентрированы реализации случайной величины.
  - *Математическое ожидание*. Для случайной величины ξ, её математическое ожидание определяется равенством (mean);

$$\mathbb{E}[\xi] = \int_{\mathbb{R}} x \mathbb{P}_\xi(dx) \tag{mean}$$

  - *Медиана*. Для случайной величины ξ, её медиана определяется как множество всех значений m, таких что $F_\xi(m) = \frac{1}{2}$, иначе говоря, медиана определяется равенством (med).

$$\mathrm{Med}[\xi] = F_\xi^{-1}\left(\tfrac{1}{2}\right) \tag{med}$$

    В некоторых случаях, в качестве медианы берут какое-то конкретное значение из множества $F_\xi^{-1}(\frac{1}{2})$, такое значение называется *точной медианой* [36];
  - *Мода*. Для случайной величины ξ её мода определяется равенством (mode)

$$\mathrm{Mode}[\xi] = \mathrm{argmax}_{\mathbb{R}}\, f_\xi(x) \tag{mode}$$

    где $f_\xi(x)$ это функция вероятности, если ξ — дискретная случайная величина, и плотность, если ξ непрерывная случайная величина.

- Меры разброса (иногда меры рассеивания) указывают на склонность величины отклоняться от своего центрального значения.
  - *Дисперсия* и *среднеквадратичное отклонение* определяются равенствами (std) и (std) соответственно;

$$\mathbb{D}[\xi] = \mathbb{E}\big[(\xi - \mathbb{E}[\xi])^2\big]; \tag{var}$$

$$\operatorname{std}[\xi] = \sqrt{\mathbb{D}[\xi]} \tag{std}$$

  - *Среднее абсолютное отклонение* определяются равенством (mad).

$$\operatorname{mad}[\xi] = \mathbb{E}[|\xi - \mathbb{E}[\xi]|] \tag{mad}$$

  - *Межквартильный размах* определяется равенством (iqr).

$$\operatorname{IQR}[\xi] = \omega_\xi(0.75) - \omega_\xi(0.25) \tag{iqr}$$

- Меры скоса—мера асимметрии распределения относительно среднего значения.
  - Коэффициент скоса определяется равенством (skew);

$$\operatorname{skew}[\xi] = \frac{\mathbb{E}\big[(\xi - \mathbb{E}[\xi])^3\big]}{\operatorname{std}^3[\xi]} \tag{skew}$$

  - Коэффициент скоса Пирсона определяется равенством (pskew);

$$\operatorname{Skew}^{\mathrm{P}}[\xi] = \frac{\mathbb{E}[\xi] - \operatorname{Med}[\xi]}{\operatorname{mad}[\xi]} \tag{pskew}$$

  - Обобщенный коэффициент скоса Грюневельда определяется равенством (pskew).

$$\gamma(u) = \frac{\omega_\xi(1-u) + \omega_\xi(u) - 2\omega_\xi\left(\frac{1}{2}\right)}{\omega_\xi(u) - \omega_\xi(1-u)} \quad \tfrac{1}{2} < u < 1 \tag{qskew}$$

- Меры эксцесса и тяжести хвостов
  - *Коэффициент эксцесса* измеряет степень остроты вершины распределения и определяется равенством (kurt).

$$\operatorname{kurt}[\xi] = \frac{\mathbb{E}\big[(\xi - \mathbb{E}[\xi])^4\big]}{\operatorname{std}^4[\xi]} - 3 \tag{kurt}$$

  - *Квантильный коэффициент эксцесса* является квантильным аналогом стандартного коэффициента эксцесса [31] и определяется равенством (qkurt);

$$\kappa(u, v) = \frac{\omega_\xi(1-u) - \omega_\xi(u)}{\omega_\xi(v) - \omega_\xi(v)}, \quad 0 < u < v < \frac{1}{2} \tag{qkurt}$$

  - *Экспонента хвоста* определяется для распределений, функция выживания которых убывает согласно степенному закону, как число α при котором верна асимптотическая эквивалентность (tail-idx).

$$S_\xi(x) \sim x^{-\alpha}, \quad x \to \infty. \tag{tail-idx}$$

    В общей ситуации отдельно определяется индекс для левого хвоста и для правого хвоста.

Также, для случайной величины ξ и натурального числа $n \in \mathbb{N}$ определены моменты, центральные моменты (m_n и μ_n соответственно в равенстве (moment)), абсолютные моменты, абсолютные центральные моменты (v_n и ν_n в равенстве (abs-moment)) и факториальные моменты (κ_n в равенстве (fact-moment)) порядка n. На основе этих характеристик можно производить оценку параметров распределения.

$$m_n = \mathbb{E}[\xi^n] \qquad \mu_n = \mathbb{E}[(\xi - \mathbb{E}[\xi])^n] \tag{moment}$$

$$v_n = \mathbb{E}\Big[|\xi|^k\Big], \qquad \nu_n = \mathbb{E}[|\xi - \mathbb{E}[\xi]|^n] \tag{abs-moment}$$

$$\kappa_n = \mathbb{E}[\xi(\xi-1)(\xi-2)\dots(\xi-n+1)] \tag{fact-moment}$$

Обобщением моментов являются так называемые L-моменты [15] и их квантильные аналоги LQ-моменты [26].

Другое семейство числовых характеристик приходит из области теории информации, см. например монографию [19]. Далее, подразумевается что ξ необязательно вещественнозначная случайная величина, со значениями из некоторого пространства $\mathcal{X}$ и под плотностью подразумевается плотность в смысле (pdf).

- *Энтропия* распределения с плотностью p относительно меры μ определяется равенством (entr)

$$H_r(p) = -\int_{\mathcal{X}} p(x) \log_r p(x)\, d\mu \tag{entr}$$

- *Кросс-энтропия* из распределения с плотностью p в распределение с плотностью q определяется равенством (entr)

$$CE_r(p\|q) = -\int_{\mathcal{X}} q(x) \log_r p(x)\, d\mu \tag{cross-enrt}$$

- *KL-дивергенция* является мерой расхождения между двумя распределениями с плотностями p и q и определятся равенством (kl-div)

$$\mathcal{D}(p\|q) = -\int_{\mathcal{X}} q(x) \log_r \frac{p(x)}{q(x)}\, d\mu \tag{kl-div}$$

Методы из теории информации активно применяются в статистике, см. например [1]. В частности, зачастую рассматривают обобщенный вариант KL-дивергенции — *f-дивергенцию*, которая определяется для любой выпуклой функции $f\colon \mathbb{R}_+ \to \mathbb{R}$ равенством (f-div).

$$\mathcal{D}_f(p\|q) = \int_{\mathbb{R}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{f-div}$$

С информационными характеристиками тесно связана информация Фишера. Для параметрического семейства плотностей $f(x;\theta), \theta \in \Theta \subseteq \mathbb{R}$, информация Фишера это функция от параметра, задаваемая равенством (FI).

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_\xi(x;\theta)\right)^2\right] \tag{FI}$$

Способы моделирования вероятностных распределений

Для вычисления метрик качества и моделирования поведения вероятностных систем необходимо уметь производить стохастическое моделирование случайных величин. Согласно [14], для генерации выборок можно выделить три основных метода:

- *Метод обратного преобразования*, который базируется на том факте, что для случайной величины ξ с квантильной функцией $\omega_\xi(p)$, распределение случайной величины $\omega_\xi(U), U \sim \mathcal{U}(0;1)$ будет совпадать с распределением ξ.
- *Метод декомпозиции*, который используется для генерации выборок из смешанных распределений. Общая идея заключается в том, что сначала генерируется значение параметра (например номер кластера), а затем уже случайная величина при условии зафиксированного значения параметра.
- *Метод отбора (rejection sampling)*, который используют когда предыдущие два метода не могут быть использованы. Этот метод значительно медленнее предыдущих и при его использовании возникает много нюансов, но для его использования необходим доступ только к плотности распределения.

При этом дискретные распределения требуют отдельного рассмотрения. Существуют также методы основанные на методе отбора, для генерации выборок из распределений заданных с помощью интегральных преобразований. При этом, нередки ситуации, плотность распределения $f_\xi(x)$ известна только с точностью до нормализующей константы или если требуется производить генерацию в сложных или многомерных пространствах. Для таких случаев разработаны методы на основе марковских цепей Монте-Карло (MCMC). Суть метода заключается в том, что на пространстве реализаций надо завести некоторое случайное блуждание, которое в пределе будет давать желанное распределение, детали см. например в книге [25].

Примеры вероятностных моделей в PySATL

В заключение этого раздела отметим, что поддержка работы с представленными ранее объектами является необходимой для PySATL в рамках существующих и будущих проектов. Пакет `MPEst`[6] использует различные числовые характеристики, такие как L-моменты для оценок параметров в моделях смеси.
Пакет `NMVMEstimation`[7] занимается специальными видами непрерывных смесей и использует различные интегральные преобразования. Библиотека `Experiment`[8] использует базовые характеристики распределений для оценок мощностей статистических тестов методом Монте-Карло. С помощью арифметики распределений можно будет получать точные распределения статистик используемых при проверке гипотез В ближайшем будущем планируется начать разработку библиотек для регрессионного анализа и оценки параметров, где также широко потребуется использование различных свойств и характеристик распределений.

[6] https://github.com/PySATL/MPEst
[7] https://github.com/PySATL/PySATL_NMVM_Module
[8] https://github.com/PySATL/pysatl-experiment

Раздел 3

Заинтересованные лица

Настоящий раздел описывает различные типы заинтересованных сторон, которым будет интересен данный документ, а также их потенциальные опасения/запросы связанные с дизайном ядра.

Мы отметим, что представленные здесь стороны/лица необязательно относятся к пользователям, которые взаимодействуют с системой.

3.1. Разработчики ядра PySATL

Это лица непосредственно принимающие участие в написании кода для ядра. Для них приоритетным является ряд возможностей:

- Добавление новой функциональности или оптимизация уже существующей;
- Покрытие кодовой базы тестами; в частности регрессионными тестами;
- Сопровождение существующей кодовой базы;
- Расширение коллектива разработчиков;

1. Добавление новой функциональности Разработчики ядра расширяют и модифицируют библиотеку в контексте следующих задач:

- Добавление новых параметрических семейств/новых операций над распределениями;
- Добавление новых числовых и/или функциональных характеристик распределений;
- Добавление новых численных методов для вычисления характеристик распределений;
- Добавление новых численных методов для вычисления операций над распределениями;

2. Тестирование Помимо расширения функциональности, разработчики ядра покрывают кодовую базу тестами; Им необходимо иметь ряд механизмов для:

- автоматического тестирования новой функциональности, касающейся свойств добавляемых распределений/семейств/операций
- создания регрессионных тестов для функциональности, оперирующей с псевдо-случайным данными

Пример. При создании нового семейств распределений разработчик указывает что все его представители имеют носитель на отрезке $[0; 1]$. Это свойство распределения которое можно проверить например сгенерировав выборку из данного распределения и проверив что все элементы выборки лежат в этом отрезке.

Пример. Валидационные примеры использования статистических процедур часто служат основной для регрессионых тестов этих самых процедур. Библиотека должна обеспечивать воспроизводимость генерации выборок.

3. Сопровождение кодовой базы и расширения коллектива разработчиков

Сопровждение кодовой базы и расширенние коллектива разработчиков осуществляется засчет нескольких инструментов

- Архитектурная документация; настоящий документ является её частью. В частности, так как ядро планируется активно использовать в других библиотеках, необходимо явно отразить какие части системы являются публично доступными
- Техническая и пользовательская документация; наличие качественной документации, содержащей примеры использования и описывающей роли компонент сисемы

3.2. Разработчики других библиотек в PySATL

К этой категории относятся все разработчики PySATL, которые либо не принимают непосредственного участия в разработке ядра, либо делают это эпизодически, добавляя функциональность под конкретные нужды проектов (в последнем случае они относятся к первой категории заинтерсованных лиц). Их основные интересы по отношению к ядру состоят в следующем.

1 Простота интеграции ядра в другие библиотеки PySATL;
2 Конфигурируемость ядра для типичных сценариев использования;

1. Простота интеграции ядра в другие библиотеки PySATL

Сейчас PySATL активно использует связку Numpy/SciPy для большинства математических вычислений; Это означает что при переходе на ядро, во всех местах использующих эту связку, замена не должна вызвать осложнений. В частности ядро

- Не должно обязывать разработчика других библиотек к конфигурации численных методов, использующихся в ядре;
- Должно иметь всю ту же функциональность, что и модуль `statistics` библиотеки SciPy;
- Должно предоставлять унифицированный интерфейс для прочих математических функций;

Здесь надо пояснить что такое прочие математические функции. При реализации различных оценок, иногда вознкиают довольно нетривиальные объекты (так pysatl-nmvm использует многочлены Белла для вычислений, а mpest - различные методы оптимизации). Так как существует множество математических библиотек, которые могут предоставлять данные возможности, то:

1 Возможно, конечный пользователь библиотеки (не обязательно ядра) захочет чтобы использовались какие-то конкретные математические пакеты (что очень может быть в случае интегрирования/оптимизации)
2 Наличие единного интерфейса для математических утилит не приведет к тому что внутри разных библиотек (или даже одной библиотеки) используются разные математические пакеты

К тому же, многие из математических пакетов, предоставляющих редкие, но нужные функции, являются либо проприетарными, либо с вирусными лицензиями. Использование и тех, и других, не подходит под лицензионные ограничения; создание единого интерфейса для математических утилит позволяет распостранять PySATL под MIT лицензией, используя вирусную лицензию только для реализации этих интерфейсов. Это значительно позволит сократить ресурсы на прототипирование разлчиных библиотек на базе ядра.

2. Конфигурируемость ядра для типичных сценариев использования

3.3. Руководители проекта PySATL

- Представление о планах разработки ядра и его функциональности на каждом этапе
- Гарантии предоставляемые ядром

3.4. Инженеры-исследователи

Программисты, которым нужны только артефакты математической статитистики. Математики, статистики и люди занимающиеся численными методами.

- Поддержка сложных операций над распределениями, смесями и т.п.
- Бенчмаркинг новых алгоритмов

Раздел 4

Ключевые требования, определяющие архитектуру

Раздел 5

Избранные архитектурные точки зрения

5.1. Контекст

TBD

Список литературы

[1] Shun-ichi Amari. *Information geometry and its applications*. T. 194. Springer, 2016.

[2] Irina A Antipova. «Inversion of multidimensional Mellin transforms». B: *Russian Mathematical Surveys* 62.5 (2007), c. 977.

[3] Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

[4] Ole E Barndorff-Nielsen, Fred Espen Benth, Almut ED Veraart и др. *Ambit stochastics*. T. 88. Springer, 2018.

[5] Richard Breen. *Regression models: Censored, sample selected, or truncated data*. 111. Sage, 1996.

[6] Satish Chandra. «On the Mixtures of Probability Distributions». B: *Scandinavian Journal of Statistics* 4.3 (1977), c. 105—112.

[7] David L Donoho и др. «High-dimensional data analysis: The curses and blessings of dimensionality». B: *AMS math challenges lecture* 1.2000 (2000), c. 32.

[8] William Palin Elderton и Norman Lloyd Johnson. «Systems of frequency curves». B: *(No Title)* (1969).

[9] Felix Famoye. *Continuous univariate distributions, volume 1*. 1995.

[10] Alessio Figalli. «On the continuity of center-outward distribution and quantile functions». B: *Nonlinear Analysis* 177 (2018), c. 413—421.

[11] Janos Galambos и Italo Simonelli. *Products of random variables: applications to problems of physics and to arithmetical functions*. CRC press, 2004.

[12] Marc Hallin и Dimitri Konen. «Multivariate Quantiles: Geometric and Measure-Transportation-Based Contours». B: *Applications of Optimal Transport to Economics and Related Topics*. Springer, 2024, c. 61—78.

[13] Daniel Hernandez-Stumpfhauser, F. Jay Breidt и Mark J. van der Woerd. «The General Projected Normal Distribution of Arbitrary Dimension: Modeling and Bayesian Inference». B: *Bayesian Analysis* 12.1 (2017), c. 113—133.

[14] Wolfgang Hörmann, Josef Leydold и Gerhard Derflinger. *Automatic nonuniform random variate generation*. Springer Science & Business Media, 2013.

[15] JRM Hosking. «L-moments». B: *Wiley StatsRef: Statistics Reference Online* (2014), c. 1—8.

[16] Thomas W Keelin. «The metalog distributions». B: *Decision Analysis* 13.4 (2016), c. 243—277.

[17] David G Kleinbaum и Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[18] Donald E Knuth. *The Art of Computer Programming: Seminumerical Algorithms, Volume 2*. Addison-Wesley Professional, 2014.

[19] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[20] Arnaud de La Fortelle. «A study on generalized inverses and increasing functions Part I: generalized inverses». working paper or preprint. Авг. 2015.

[21] Nicolas Lanchier. *Stochastic modeling*. Springer, 2017.

[22] Lawrence M Leemis и др. «Univariate probability distributions». B: *Computational Probability Applications* (2017), c. 133—147.

[23] Erich L Lehmann и George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[24] Erich Leo Lehmann и др. «Statistical methods based on ranks». B: *Nonparametrics. San Francisco, CA, Holden-Day* 2 (1975).

[25] Faming Liang, Chuanhai Liu и Raymond Carroll. *Advanced Markov chain Monte Carlo methods: learning from past samples*. John Wiley & Sons, 2011.

[26] Govind S Mudholkar и Alan D Hutson. «LQ-moments: Analogs of L-moments». B: *Journal of Statistical Planning and Inference* 71.1-2 (1998), c. 191—208.

[27] Frank Nielsen и Vincent Garcia. «Statistical exponential families: A digest with flash cards». В: *arXiv preprint arXiv:0911.4863* (2009).

[28] Luigi Pace и Alessandra Salvan. *Principles of statistical inference: from a Neo-Fisherian perspective*. Т. 4. World scientific, 1997.

[29] Xavier Pennec. «Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements.» В: *NSIP*. Т. 3. 1999, с. 194—198.

[30] Karsten Prause и др. «The generalized hyperbolic model: Estimation, financial derivatives, and risk measures». Дис. ... док. Citeseer, 1999.

[31] David Ruppert. «What is kurtosis? An influence function approach». В: *The American Statistician* 41.1 (1987), с. 1—5.

[32] Jacob Schreiber. «Pomegranate: fast and flexible probabilistic modeling in python». В: *Journal of Machine Learning Research* 18.164 (2018), с. 1—6.

[33] Melvin Dale Springer. *The algebra of random variables*. New York: Wiley, 1979.

[34] György Steinbrecher и William T Shaw. «Quantile mechanics». В: *European journal of applied mathematics* 19.2 (2008), с. 87—112.

[35] Maciej J Swat, Pierre Grenon и Sarala Wimalaratne. «ProbOnto: ontology and knowledge base of probability distributions». В: *Bioinformatics* 32.17 (2016), с. 2719—2721.

[36] Herbert Weisberg. *Central tendency and variability*. 83. Sage, 1992.

[37] Вильям Феллер. *Введение в теорию вероятностей и ее приложения*. Рипол Классик, 2013.

[38] Наталья Исааковна Чернова. *Математическая статистика*. Новосибирский гос. ун-т, 2007.

[39] Альберт Николаевич Ширяев. *Вероятность*. МЦНМО, 2007.