Probability

数据嗨客

基本概念

随机试验 | 指满足以下条件的试验：给定条件不变，可重复进行；结果可穷举不唯一；无法事前确定试验结果.

样本空间 | 指随机试验E的所有可能结果的集合. E的样本空间的子集称为**随机事件**，简称事件.

古典概型 | 指满足以下条件的随机试验：

✓ 随机试验过的样本空间所包含的元素有限；

✓ 随机试验中事件发生的可能性一致.

频率 | 给定条件不变，在n次随机试验中，事件A发生的次数为n_A. 那么，n_A称为频数，$\frac{n_A}{n}$称为A发生的频率.

概率 | 当$n \to \infty$时，事件A发生的频率会逐步稳定在某个数值附近，该数值称为事件A发生的概率，记为$P(A)$. 概率$P(A)$需要满足以下条件，给定样本空间S：

$$P(A) \geq 0 \quad P(S) = 1$$

$$A \cap B = \varnothing, P(A+B) = P(A) + P(B).$$

概率具有以下性质：

✓ $P(\varnothing) = 0$.

✓ 给定任意事件A，$P(A) \leq 1$.

✓ 给定任意事件A，样本空间S以及$A \cup \overline{A} = S$，

$$P(\overline{A}) = 1 - P(A)$$

✓ 给定任意事件A，和B

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

条件概率 | 给定任意事件A和B，$P(A) \neq 0$

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

为事件A发生的情况下事件B发生的概率.

✓ 事件A或者B单独发生的概率，称为**边缘概率**，即$P(A)/P(B)$；

✓ 事件A和B同时发生的概率称为**联合概率**，记为$P(AB)$.

划分 | 给定$\{C_1, C_2, \cdots, C_n\}$为随机试验$E$的样本空间$S$中的一系列事件，如果这些事件满足

$$C_i C_j = \varnothing, i \neq j$$

$$C_1 \cup C_2 \cup \cdots \cup C_n = S$$

那么，$\{C_1, C_2, \cdots, C_n\}$称为$S$的一个划分. 如果$A$为$E$中的一个事件，$P(C_i) > 0$，则有

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(AC_i) \\ &= P(A|C_1)P(C_1) + \cdots + P(A|C_n)P(C_n) \end{aligned}$$

以上的等式称为**全概率公式**.

如果$P(A) > 0, P(C_i) > 0$同时成立，则有

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{P(A)} = \frac{P(A|C_i)P(C_i)}{\sum_{i=1}^{n} P(AC_i)}$$

这样的等式称为**贝叶斯公式**.

独立 | 给定任意事件A和B，二者如果满足

$$P(AB) = P(A)P(B)$$

那么，事件A和B相互独立.

条件独立 | 给定事件ABC，如果在事件C发生的条件下，事件A发生与否和事件B发生与否没有关系，即

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

那么，事件A和B在C下条件独立.

随机变量、概率分布

随机变量 | 给定f是样本空间S上的实值函数，是建立在S中的事件发生的结果与实数的一种映射. 比如，进行三次投掷硬币试验，X表示硬币正面朝上的次数，事件E为“两次正面朝上”，他们的关系表示为

$$X = f(E) = 2, E = \{HHT, THH, HTH\},$$

X称为随机变量. 因此，硬币两次正面朝上的概率为

$$P(X=2) = P(\{HHT, THH, HTH\}) = \frac{3}{2^3} = \frac{3}{8}.$$

如果X的取值为有限个或者可列，那么X称为**离散型随机变量**.

概率分布 | 给定离散随机变量X，

$$P(X = x_k) = p_k, k \geq 1$$

为X的概率分布，也称为**分布律**；$\{p_k\}$为分布列.

✓ 函数$f: \{x_k\} \to \{p_k\}$为**概率质量函数**，即$f(x_k) = p_k$.

离散型随机变量的常见分布：

伯努利分布 | 随机变量$X, X \in \{0, 1, \cdots, n\}$表示$n$次伯努利试验中某事件发生的次数，分布律表示为

$$P(X = k) = \frac{n!}{(n-k)!k!}\theta^k(1-\theta)^{(n-k)}$$

当$n = 1$时，伯努利分布也称为**0-1分布**.

泊松分布 | 随机变量$X \in \{0, 1, 2, \cdots\}$，给定参数$\lambda$，分布律表示为

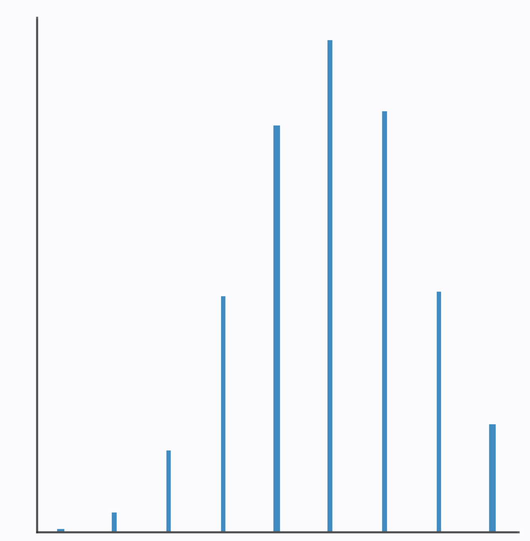
$$P(X = k) = \frac{e^{-\lambda}}{k!}\lambda^k$$

分布函数 | 对于非离散型随机变量X来说，有意义的是X取值落在某个区间的概率. 给定$x \in (-\infty, \infty)$

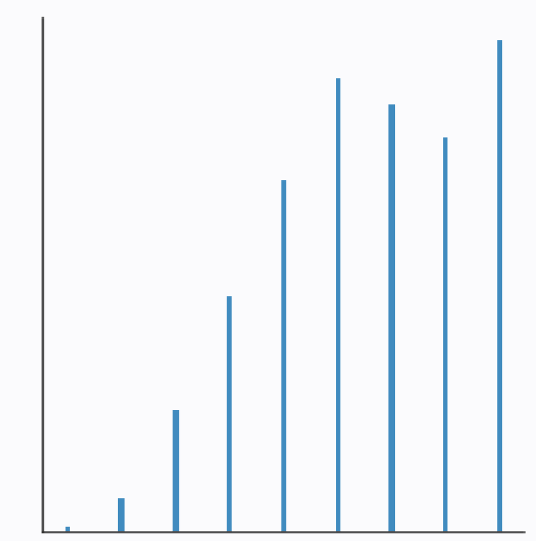
$$F(x) = P(X \leq x)$$

称为X的分布函数. 因此，$\forall x_1, x_2, x_1 < x_2$

$$F(x_2) - F(x_1) = P(X \leq x_2) - P(X \leq x_1) = P(x \in (x_1, x_2])$$



伯努利分布



泊松分布

概率密度 | 给定随机变量X，及其分布函数$F(x)$，如果$\exists f(x) \geq 0$，对于$\forall x$，满足

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

X为**连续型随机变量**，$f(x)$的概率密度函数，简称为概率密度，具有以下性质：

✓ $f(x) \geq 0$ ✓ $\int_{-\infty}^{\infty} f(x)dx = 1$

连续型随机变量的常见分布：

均匀分布 | 给定连续型随机变量X，概率密度表示为

$$f(x) = \begin{cases} \frac{1}{m-n} & x \in (m, n) \\ 0 & \end{cases}$$

X在取值区间(m, n)上服从均匀分布，记为$X \sim \mathcal{U}(m, n)$.

正态分布 | 给定一维连续型随机变量X，概率密度为

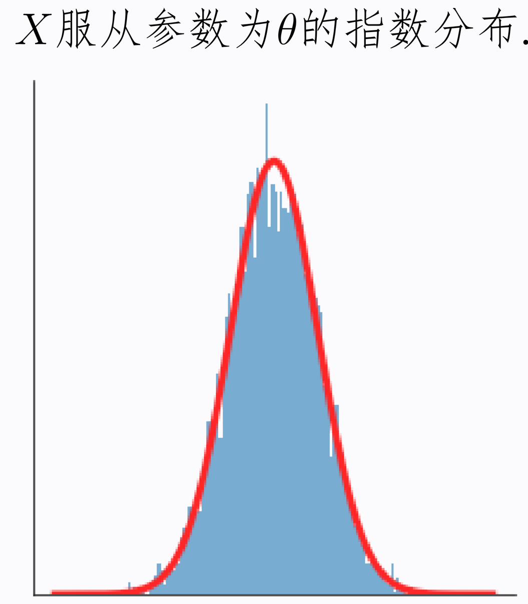
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

X服从正态分布，记为$X \sim \mathcal{N}(\mu, \sigma^2)$.

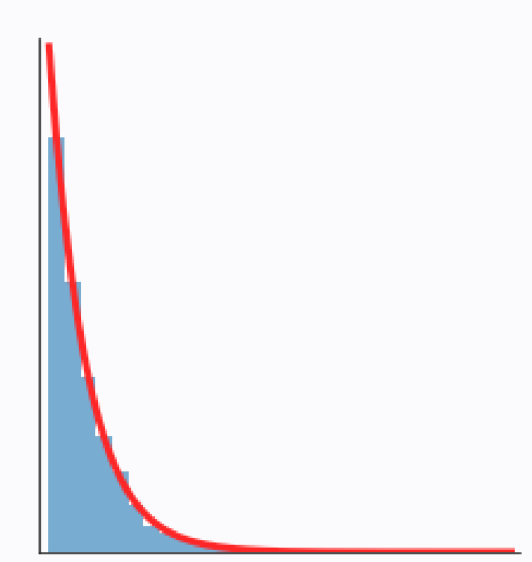
指数分布 | 给定一维连续型随机变量X，概率密度为

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-\frac{x}{\theta}} & x > 0 \\ 0 & \end{cases}$$

X服从参数为θ的指数分布.



正态分布



指数分布

多维正态分布 | 给定随机变量$\boldsymbol{X} \in \mathbb{R}^d$，概率密度表示为

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

其中，$\boldsymbol{\mu}$为均值向量，$\boldsymbol{\Sigma}$为协方差矩阵. 随机变量$\boldsymbol{X}$服从多维正态分布. 若$\{x^i\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，分布参数可由以下等式得到

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x^i, \hat{\boldsymbol{\Sigma}} = \frac{1}{n}(x - \hat{\boldsymbol{\mu}})(x - \hat{\boldsymbol{\mu}})^T$$

其密度函数还具有一种表达形式，称为information form. 约定$\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$，有

$$f(\boldsymbol{x}|\boldsymbol{\eta}, \boldsymbol{\Lambda}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{\eta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} - 2\boldsymbol{x}^T\boldsymbol{\eta}\right\}$$

概率分布推断

给定含有n个样本的数据集 $\mathcal{D} = \{x^1, \cdots, x^n\}$，$p(x|\theta)$为样本的概率分布：

独立同分布 | 给定概率分布参数θ，$\mathcal{D}$的观测样本之间服从统一分部，且互相独立，有

$$p(\mathcal{D}|\theta) = p(x^1, \cdots, x^n|\theta) = \prod_{i=1}^{n} p(x^i|\theta)$$

似然函数 | 关于模型（分布）参数的函数，即在给定分布参数的条件下，数据集$\mathcal{D}$出现的概率.

$$\mathcal{L}(\mathcal{D}|\theta) = f(\theta|\mathcal{D})$$

先验概率 | 与观测数据无关，参数θ的概率分布$f(\theta)$.

后验概率 | 在观测数据集$\mathcal{D}$的条件下参数θ的概率，记为$f(\theta|\mathcal{D})$. 根据贝叶斯公式，给出三者之间的关系

$$\underbrace{p(\theta|\mathcal{D})}_{\text{后验概率}} = \frac{1}{\underbrace{p(\mathcal{D})}_{\text{观测证据}}}\overbrace{p(\mathcal{D}|\theta)}^{\text{似然函数}}\overbrace{p(\theta)}^{\text{先验概率}}.$$

最大化后验概率 | 寻找使$\mathcal{D}$出现的可能性最大化的参数θ.

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmin}} \quad p(\theta|\mathcal{D})$$

极大似然 | 最大化似然函数，则有

$$\theta_{ML} = \underset{\theta}{\operatorname{argmin}} \quad p(\mathcal{D}|\theta)$$

若$p(\theta)$为均匀分布，最大化后验概率等价于极大似然.

KL散度 | 也称为相对熵，用来衡量两个概率分布的差异性. 给定分布$p(\mathcal{D})$和$q(\mathcal{D})$

$$\mathcal{KL}(p|q) = \mathbb{E}_p(\log p - \log q) = \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}$$