CSDN
不 止 于 代 码

第一讲：随机变量及其分布

CSDN学院

2017年7月

AI 101

Roadmap

- 概率公理及推论
- 随机变量及其分布
- 分布的概述
- 常见随机变量分布
- 抽样分布
- 分布的估计

一、概率公理及推论

概率

- 例：硬币正面向上的概率为0.5.
  - 频率学派：多次重复试验（多次投掷硬币），则我们期望正面向上的次数占总实验次数的一半
  - Bayesian学派：我们相信下一次试验中，硬币正面向上的可能性为0.5
    - 概率是我们对事情不确定性的定量描述，与信息有关，而非重复实验
    - 可以用来对不能重复实验的时间的不确定，如明天天晴的概率为0.8，如某个邮件是垃圾邮件的概率

样本空间和事件

- 考虑一个事先不知道输出的试验：
- 试验的样本空间：所有可能输出的集合
  - 例：抛掷两次硬币，则样本空为$\Omega=\{HH, HT, TH, TT\}$
    - 其中H表示正名向上，T表示反面向上

- 事件A是样本空间的子集
  - 上述试验中第一正面向上：$A = \{HH, HT\}$

概率公理

- 对每个事件A ，我们定义一个数字$\mathbb{P}(A)$ ，称为A 的概率。
  概率根据下述三条公理：
  - 1、事件A 的概率是一个非负实数： $\mathbb{P}(A) \geq 0$
  - 2、合法命题的概率为1： $\mathbb{P}(\Omega) = 1$
  - 3、对两两不相交（互斥）事件$A_1, A_2, \ldots$ ，

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

从上述三个公理，可推导出概率的所有的其他性质。

推论

- 不可满足命题的概率为0
  - $\mathbb{P}(\emptyset) = 0$
  - $\mathbb{P}(A \cap \bar{A}) = 0$
- 对任意两个事件A 和 B
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- 事件A的补事件
  - $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$
- 对任意事件A
  - $0 \leq \mathbb{P}(A) \leq 1$

联合概率&条件概率

- 对任意两个事件A 和 B
  - $\mathbb{P}(A\ ,B)=\mathbb{P}(A\cap B)=\mathbb{P}\ (A|B)\ \mathbb{P}(B)$
- 当$\mathbb{P}(B)>0$ 时，给定B时A的条件概率为

$$\mathbb{P}(A\,|\,B)=\frac{\mathbb{P}(A,B)}{\mathbb{P}(B)} \quad \Rightarrow \quad \mathbb{P}(A\ ,B)=\mathbf{P}\ (A|B)\ \mathbf{P}(B)$$ 积规则（Product Rule）

- 给定任意B，若$\mathbb{P}(B)>0$，则$\mathbb{P}(\cdot\,|\,B)$也是一个概率，即满足概率的三个概率公理
  - $\mathbb{P}(A\,|\,B)\geq 0$
  - $\mathbb{P}(\Omega\,|\,B)=1$
  - 当$A_1,A_2,\ldots$不相交时，$\mathbb{P}\left(\bigcup_{i=1}^{\infty}A_i\,|\,B\right)=\bigcup_{i=1}^{\infty}\mathbb{P}(A_i\,|\,B)$

贝叶斯公式

- 全概率公式：令$A_1, \ldots, A_K$为A的一个划分，则对任意事件B，有$\mathbb{P}(B)=\sum_j \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)$。

- 贝叶斯公式：令$A_1, \ldots, A_k$为A的一个划分且对每个k，$k=1,2, \ldots,K$。若$\mathbb{P}(B)>0$，则对每个$\mathbb{P}(A_k)>0$有

先验概率

$$\mathbb{P}(A_k \mid B)=\frac{\mathbb{P}(B \mid A_k)\mathbb{P}(A_k)}{\sum_j \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}$$

后验概率

分子分母形式相同，分母为所有划分之和

Thomas Bayes
(1702-1761)

例：贝叶斯公式应用

- 不同操作系统用户感染病毒的概率稍有不同。已知Macintosh用户感染病毒的概率为0.65，Windows用户感染病毒的概率为0.82，Linux用户感染病毒的概率为0.5。并且用户使用Macintosh、Windows和Linux操作系统的比例分别为0.3, 0.5和0.2。现在发现用户感染了病毒，则该用户为Windows用户的概率是多少？
- 解:令B表示被病毒感染，

A_1表示Macintosh用户，则 $\mathbb{P}(A_1)=0.3$，$\mathbb{P}(B\mid A_1)=0.65$

A_2表示Windows用户，$\mathbb{P}(A_2)=0.5$，$\mathbb{P}(B\mid A_2)=0.82$

A_3表示Linux用户，$\mathbb{P}(A_3)=0.2$，$\mathbb{P}(B\mid A_3)=0.5$

因为$\mathbb{P}(A_1)+\mathbb{P}(A_2)+\mathbb{P}(A_3)=1$，所以$A_1$、$A_{12}$、和$A_3$组成一组完备事件

所以根据贝叶斯公式，

$$\mathbb{P}(A_2\mid B)=\frac{\mathbb{P}(B\mid A_2)\mathbb{P}(A_2)}{\sum_i \mathbb{P}(B\mid A_i)\mathbb{P}(A_i)}=\frac{0.82\times 0.5}{0.65\times 0.3+0.82\times 0.5+0.5\times 0.2}=0.58$$

二、随机变量及其分布

Outlines

- 随机变量
- 随机变量的分布：CDF、pdf
- 分布的概述：
  - 均值、众数、中值、分位数
  - 方差、IQR

随机变量

- 机器学习与数据相关。随机变量就是将随机事件与数据之间联系起来的纽带。

- 随机变量是一个映射/函数 $X:\Omega \to \mathbb{R}$，将一个实数值 $X(\omega)$ 赋给一个试验的每一个输出 ω

- 例1：抛10次硬币，令$X(\omega)$表示序列ω中正面向上的次数，如当 $\omega = \mathrm{HHTHHTHHTT}$，则 $X(\omega) = 6$。
  - X只能取离散值，称为离散型随机变量

随机变量

- 例2：令 $\Omega=\{(x,y):x^2+y^2\le 1\}$ 表示单位圆盘，输出为该圆盘中的一点 $\omega=(x,y)$ ，则有随机变量：

$$X(\omega)=x,\ Y(\omega)=y,\ Z(\omega)=\sqrt{x^2+y^2}$$

  - X取连续值，称为连续型随机变量

数据和统计量

- 数据是随机变量的具体值
- 统计量是数据/随机变量的任何函数
- 任何随机变量的函数仍然是随机变量

15

累积分布函数CDF

- 令X为一随机变量，x为X的一具体值（数据）
- 则随机变量X的累积分布函数 $F:\mathbb{R}\to[0,1]$ (cumulative distribution function, CDF) 定义为

$$F(x)=\mathbb{P}(X\le x)$$

- CDF是一个非常有用的函数：包含了随机变量的所有信息。

► 概率（质量）函数pmf

- 离散型随机变量的概率函数 (probability function or probability mass function, pmf)定义为

$$p(x)=\mathbb{P}(X=x)$$

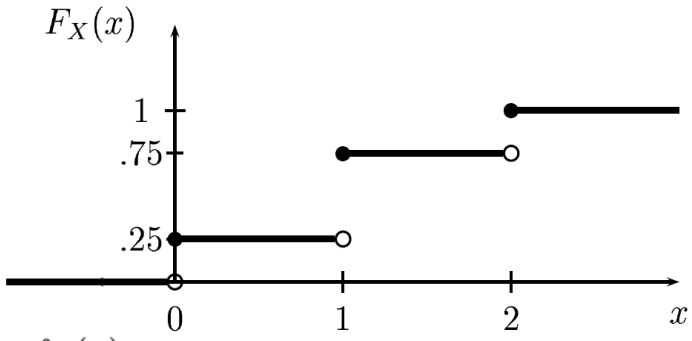
- 对所有的 $x\in\mathbb{R}$，$p(x)\geq 0$
- $\sum_i p(x_i)=1$
- CDF与pmf之间的关系为：$F(x)=\mathbb{P}(X\leq x)=\sum_{x_i\leq x} p(x_i)$

例：离散型随机变量的CDF和pmf

- 例：公正地抛硬币2次，令X表示正面向上的次数，则

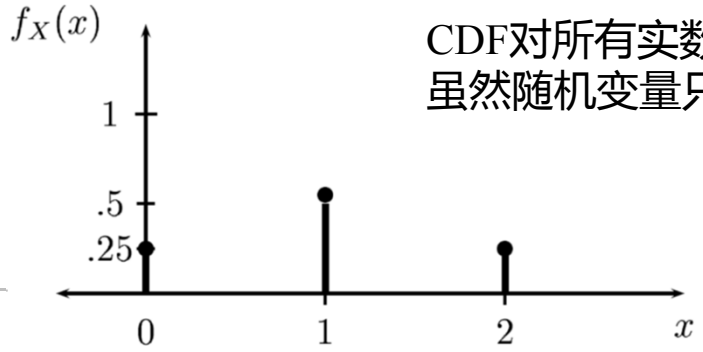
$\mathbb{P}(X=0)=\mathbb{P}(X=2)=1/4$，$\mathbb{P}(X=1)=1/2$，则分布函数如下：

$$F_X(x)=\begin{cases}0 & x<0\\ 1/4 & 0\le x<1\\ 3/4 & 1\le x<2\\ 1 & x\ge 2\end{cases}$$



- 概率函数为：

$$p(x)=\begin{cases}1/4 & x=0\\ 1/2 & x=1\\ 1/4 & x=2\\ 0 & otherwise\end{cases}$$



CDF对所有实数x都有定义，虽然随机变量只取0、1、2

18

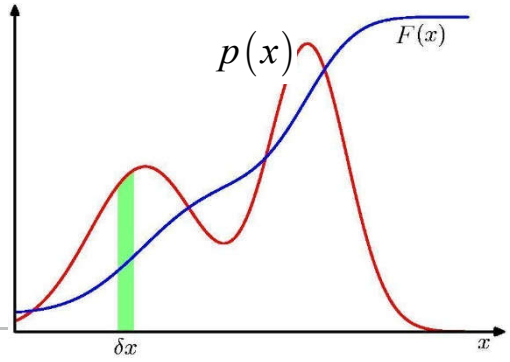
概率密度函数pdf

- 对连续型随机变量X，如果存在一个函数p，使得对所有的x，$p(x)\geq 0$，且对任意 $a\leq b$ 有

$$\mathbb{P}(a<X\leq b)=\int_a^b p(x)dx$$

注意：$p(x)$ 不必<1

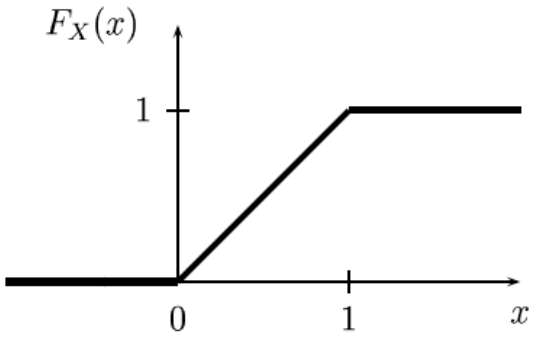
- 函数p被称为概率密度函数 (probability density function, pdf)
- 当F可微时，
  - $F(x)=\int_{-\infty}^{x} p(t)dt$
  - $p(x)=F'(x)$



▶例：连续型随机变量的CDF和pdf

- 例：设X有PDF: $p(x)=\begin{cases}1 & for\ 0\le x<1\\ 0 & otherwise\end{cases}$
- 显然有 $p(x)\ge 0, \int p(x)dx = 1$
- 有该密度的随机变量为(0,1)上的均匀分布：Uniform(0, 1)，即在0和1之间随机选择一个点。
- 其CDF为：

$$F(x)=\begin{cases}0 & x<0\\ x & 0\le x\le 1\\ 1 & x>1\end{cases}$$



分布的概述

- 除了概率分布函数，有时我们也采用一些单值描述量来刻画某个分布的性质。
  - 位置描述：期望/均值、中值、众数、分位数
  - 散布程度描述：方差、四分位矩（ IQR ）

期望

- 期望/均值：随机变量的平均值
  - 概率加权平均

- 如果下列积分有定义的话（$\int |x| dF(x) < \infty$），定义X的期望（均值，一阶矩）为：

$$\mathbb{E}(X) = \mu = \int x dF(x) = \int x p(x) dx$$

- 离散情况下为：$\sum_x x p(x)$

期望的性质

- 线性运算：$\mathbb{E}(aX+b)=a\mathbb{E}(X)+b$
- 加法规则：设$X_1,\cdots X_N$是随机变量，$a_1\cdots a_N$是常量，有

$$\mathbb{E}\left(\sum_i a_i X_i\right)=\sum_i a_i \mathbb{E}\left(X_i\right)$$

- 乘法规则：设$X_1,\cdots X_N$是相互独立的随机变量，有

$$\mathbb{E}\left(\prod_{i=1}^{N} X_i\right)=\prod_{i=1}^{N}\mathbb{E}\left(X_i\right)$$

▶例：期望

- 期望是随机变量的一个很好单值概述
- [**最小距离**] 假设我们用L_2距离度量一个随机变量X与一个常数b的距离，即 $(X-b)^2$。b离X越近，这个量就越小。因此我们可以确定b的值，使得 $\mathbb{E}\left((X-b)^2\right)$最小，$b$可认为是$X$的一个很好预测（不能直接最小化$(X-b)^2$因为结果与$X$有关，对$X$的预测无用）。

$$\mathbb{E}(X-b)^2 = \mathbb{E}\left(X-\mathbb{E}(X)+\mathbb{E}(X)-b\right)^2$$

$$= \mathbb{E}\left((X-\mathbb{E}(X))+(\mathbb{E}(X)-b)\right)^2$$

$$= \mathbb{E}(X-\mathbb{E}(X))^2 + (\mathbb{E}(X)-b)^2 + 2\mathbb{E}\left((X-\mathbb{E}(X))(\mathbb{E}(X)-b)\right)$$

$$= (\mathbb{E}(X)-b)\mathbb{E}(X-\mathbb{E}(X)) = 0$$

$(\mathbb{E}(X)-b)$是常数

$$b^* = \arg\min_b \mathbb{E}\left((X)-b\right)^2 = \arg\min_b (\mathbb{E}(X)-b)^2 = \mathbb{E}(X)$$

众数（mode）

- 众数：设随机变量X有密度$p(x)$，且存在x_0满足

$$x_0 = \arg\max_x p(x)$$

- 则称x_0为X的众数。
  - 刻画随机变量出现次数最多的位置

- 期望、中位数和众数都称为位置参数。
  - 当随机变量的分布为高斯分布时，三者相等

贝叶斯估计中最大后验估计（MAP）就是后验分布的最大值/众数

► 中值（Median）

- 分布的中值可视为分布的中间，即在其上下的概率均为0.5：

$$Median(X) := x^* : P(X \geq x^*) = 0.5$$

- 中值是分布另一个很好单值概述
  - 若我们用L_1距离度量一个随机变量X与一个常数b的距离，则当b为中值时，随机变量X与b的距离最小。
  - Recall: L_2距离度量下，随机变量X与其均值的距离最小。

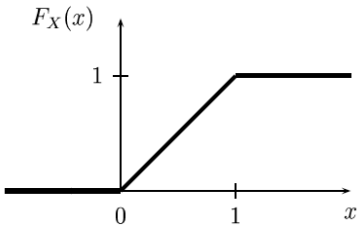
分位函数 (quantile)

- 令随机变量X的CDF为F，CDF的反函数或分位函数(quantile function)定义为

$$F_X^{-1}(\alpha) = \inf\{x : F_X(x) \geq \alpha\}$$

inf:下界

- 其中$\alpha \in [0,1]$。若F严格递增并且连续，则 $F_X^{-1}(\alpha)$ 为一个唯一确定的实数x，使得 $F_X(x) = \alpha$ 。
  - F_X^{-1} 为增函数



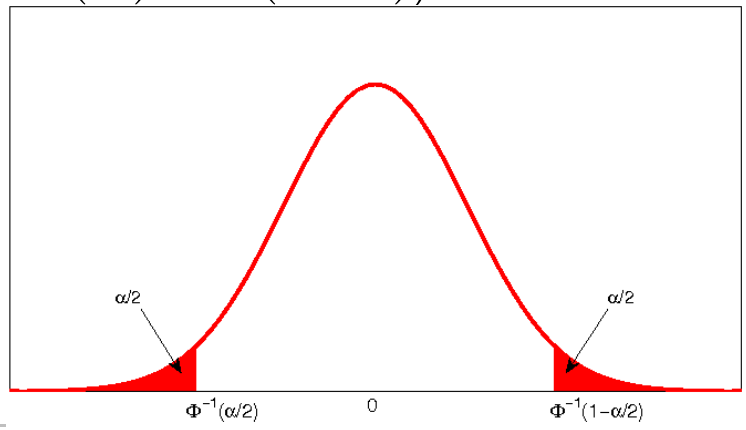
- 中值(median)：$F_X^{-1}(0.5)$
- 上下1/4分位数：$F_X^{-1}(0.25), F_X^{-1}(0.75)$

$$F_X^{-1}(\alpha) = \begin{cases} -\infty & \alpha \leq 0 \\ x & 0 < \alpha \leq 1 \\ 1 & \alpha > 1 \end{cases}$$

分位数 (cont.)

- 对正态分布 $Z \sim \mathcal{N}(0,1)$ ，其CDF的反函数记为 Φ^{-1}
- 当 $\alpha = 0.05$ 时，随机变量95%（1-α）的概率会落在区间：

$$\left(\Phi^{-1}\left(\frac{\alpha}{2}\right), \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = (-1.96, 1.96)$$



方差

- X的k阶矩定义为 $\mathbb{E}(X^k)$ ，假设 $\mathbb{E}(X^k)<\infty$

- 若X 有均值μ，则其方差（二阶中心矩）

$$\sigma^2=\mathbb{V}(X)=\mathbb{E}(X-\mu)^2=\int(x-\mu)^2\,dF(x)$$

- 且标准差 $\sigma=sd=\sqrt{\mathbb{V}(X)}$

$$\mathbb{V}(X)=\mathbb{E}\left[(X-\mu)^2\right]=\mathbb{E}(X^2)-\mu^2$$

- 方差：刻画随机变量围绕均值的散布程度
  - 方差越大，X变化越大；方差越小，X与均值越接近

方差的性质

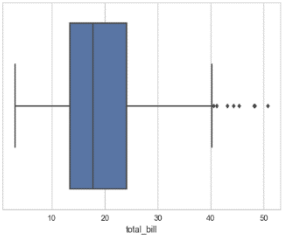
- 设方差有定义，则有以下性质：

1. $\mathbb{V}(X)=\mathbb{E}(X)^2-\mu^2$
2. 当a,b是常数时，$\mathbb{V}(aX+b)=a^2\mathbb{V}(X)$
3. 如果$X_1,\dots X_N$独立，且$a_1,\dots a_N$为常数，则$\mathbb{V}\left(\sum_{i=1}^{N}a_iX_i\right)=\sum_{i=1}^{N}a_i^2\mathbb{V}(X_i)$。

注意：期望的加法规则无需独立条件
不独立随机变量和的方差计算需考虑变量之间的协方差（下节课）

IQR（Interquartile Range）

- 中值是比均值更鲁棒的分布的中心的度量
- 比方差更鲁棒的分布的散布范围的度量是四分位矩（Interquartile Range， IQR）：25%分位数到75%分位数之间的区间
- IQR在boxplot（ seaborn.boxplot）中用到：分布的图形概述
  - 长方形为IQR
  - 中间的线为中值
  - 两头的虚线：1.5IQR
  - 超过上下限的数据为噪声点，用 ‘*’ 或 ‘+’ 等符号表示



三、常见随机变量概率分布

常见离散型随机变量

- 离散型随机变量
  - 贝努利(Bernoulli）分布
  - 二项(binomial)分布
  - 多项（multinomial）分布

Bernoulli分布

- Bernoulli分布又名两点分布或者0-1分布。若Bernoulli试验成功，则Bernoulli随机变量X取值为1，否则X为0。记试验成功概率为θ，即

$$\mathbb{P}(X=1)=\theta, \quad \mathbb{P}(X=0)=1-\theta, \quad \theta\in[0,1]$$

- 我们称X服从参数为θ的Bernoulli分布，记为 $x\sim \mathrm{Ber}(\theta)$

$$p(x|\theta)=\begin{cases}\theta & if\ x=1\\ 1-\theta & f\ x=0\end{cases}$$

$$=\theta^x(1-\theta)^{1-x},\ \text{for}\ x\in[0,1]$$

Jakob I. Bernoulli
(1654－1705)

Bernoulli分布

- Bernoulli分布的均值：$\mu=\theta$
- 方差：$\sigma^2=\theta(1-\theta)$

- 两类分类问题：$y|\mathbf{x}$服从Bernoulli分布，即类别标签y取值为0或1的离散随机变量

二项分布

- 二项(Binomial)分布：在抛掷硬币试验中，若只进行一次试验，则为Bernoulli试验。若进行n次试验，则硬币正面向上的数目X满足二项分布，记为 $x \sim \mathrm{Bin}(n,\theta)$

$$p(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

  - 其中 $\binom{n}{x} \triangleq \frac{n!}{(n-x)!x!}, \quad x \in \{0,\ldots,n\}$

- 二项分布的均值：$\mu=\mathrm{n}\theta$
- 方差： $\sigma^2 = n\theta(1-\theta)$

► 多项分布（ Multinomial 分布）

- 假设抛有K个面的的骰子，其中抛掷到第j面的概率为 θ_j ，令 $\boldsymbol{\theta}=(\theta_1,...,\theta_K)$
- 若一共抛掷n次， $\mathbf{x}=(x_1,...,x_K)$ 为随机向量，其中x_k表示抛掷到第k面的次数，则$\mathbf{x}$的分布为多项分布，即$\mathbf{x}\sim \mathrm{Mu}(n,\boldsymbol{\theta})$

$$p(\mathbf{x}|n,\boldsymbol{\theta})=\binom{n}{x_1\ \ldots\ x_K}\prod_{k=1}^{K}\theta_k^{x_k},\quad \binom{n}{x_1\ \ldots\ x_K}\triangleq\frac{n!}{x_1!\ldots x_K!}$$

- 当n=1时为 $\mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta})$ ，$p(\mathbf{x}|1,\boldsymbol{\theta})=\prod_{k=1}^{K}\theta_k^{x_k}$,

分类分布

Multinomial分布的特例：

- 当n=1时，记为类别型分布（Categorical 分布）：

$$\mathrm{Cat}(\mathbf{x}|\boldsymbol{\theta}) \triangleq \mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta})$$

- 由于$\mathbf{x}$中K维数据中只有一个为1，其余均为0，我们将其写成另一种形式

$$x \sim \mathrm{Cat}(\boldsymbol{\theta}), \text{ 则 } p(x=k|\boldsymbol{\theta}) = \theta_k$$

- 多类分类问题：$y|\mathbf{x}$服从Categorical 分布，即类别标签y取值为0到K之间整数的离散随机变量

常见连续型随机变量

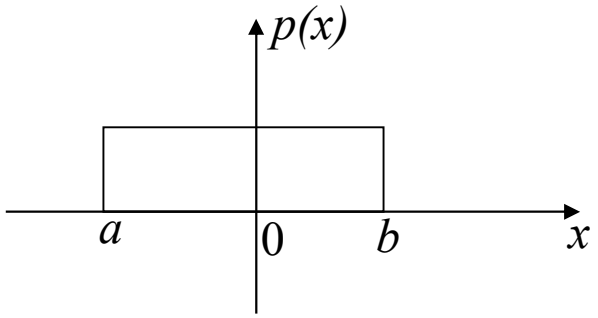
- 连续型随机变量
  - 均匀分布
  - 高斯分布
  - Laplace分布
  - Gamma分布
  - Beta分布

► 均匀分布

一些连续分布的例子：

- 1: 均匀分布： $X \sim \text{Unif}(a,b)$

$$p(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & otherwise \end{cases}$$

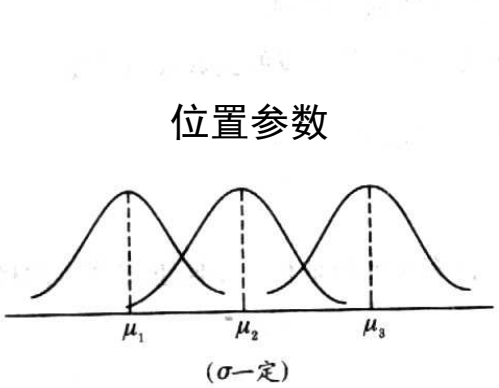


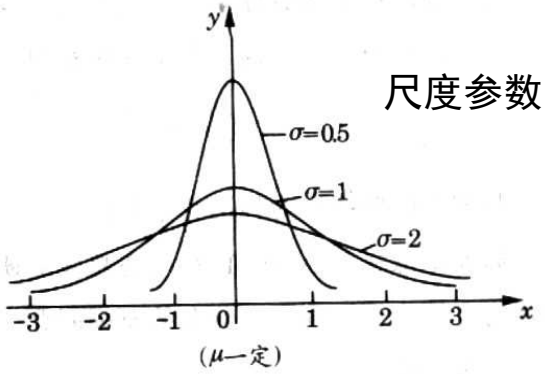
► 高斯分布

- 高斯分布/正态分布： $X \sim \mathcal{N}(\mu, \sigma^2)$

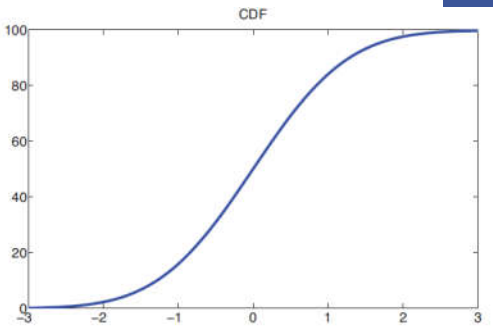
$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ x \in \mathbb{R},\ \sigma > 0$$

  - 其中μ, σ^2分别为高斯分布的均值和方差





高斯分布



- 高斯分布的CDF为 :$\Phi\left(x|\mu,\sigma^2\right)=\int_{-\infty}^{x}\mathcal{N}\left(z|\mu,\sigma^2\right)dz$
- 高斯分布是一种很重要的分布
  - 参数容易解释，也描述了分布最基本的性质
  - 中心极限定理（样本均值的极限分布为高斯分布）
  - 对模型残差或噪声能很好建模（高斯分布的由来）
  - 具有相同方差的所有可能的概率分布中，正态分布具有最大的不确定性（极大熵）

标准正态分布

- 当 $\mu=0,\sigma=1$ 时，称为标准正态分布，通常用Z表示服从标准正态分布的变量，记为 $Z\sim\mathcal{N}(0,1)$
- pdf和CDF分别记为 $\phi(z),\Phi(z)$

- 标准化：
  - 若 $X\sim\mathcal{N}(\mu,\sigma^2)$ ，则 $Z=(X-\mu)/\sigma\sim\mathcal{N}(0,1)$
  - 若 $Z\sim\mathcal{N}(0,1)$ ，则 $X=\mu+\sigma Z\sim\mathcal{N}(\mu,\sigma^2)$

退化的高斯分布

- 当 $\sigma^2 \to 0$ 时，高斯分布退化为无限高无限窄、中心位于μ的“针”状分布：$\lim_{\sigma^2 \to 0} \mathcal{N}(x|\mu,\sigma^2) = \delta(x-\mu)$
- 其中Dirac delta 函数 $\delta(x) = \begin{cases} \infty & if\ x=0 \\ 0 & f\ x \neq 0 \end{cases}$，使得 $\int_{-\infty}^{\infty} \delta(x)dx = 1$
- 有用的性质：将某个信号从求和/积分中筛选出来

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

- 注意：别将Dirac delta函数与Kronecker delta 函数混淆：$\delta_{ij} = \mathbb{I}(i=j) = \begin{cases} 1 & if\ i=j \\ 0 & f\ i \neq j \end{cases}$

经验分布

- Dirac分布经常作为经验分布（empirical distribution）的一个组成部分出现：

$$\hat{p}(x)=\frac{1}{N}\sum_{i=1}^{N}\delta\left(x-x_i\right)$$

  - 将密度 $1/N$ 赋给每一个数据点
  - 只有当定义连续型随机变量的经验分布时，Dirac delta 函数才是必要的；对离散型随机变量，经验分布可以被定义成一个 Multinomial 分布，对每一个可能的输入，其概率可简单地设为在训练集上那个输入值的经验频率。
  - 经验分布也是极大似然估计（使训练数据的出现的概率最大的那个概率密度函数）

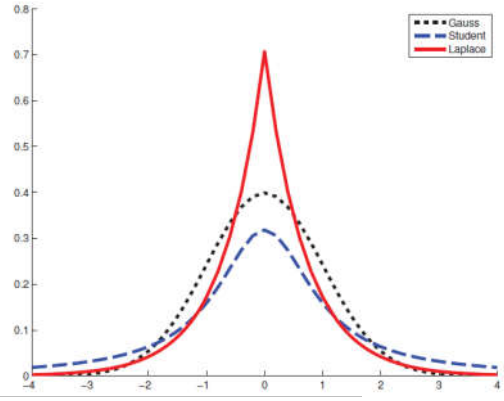
Laplace分布

Pierre-Simon Laplace
（1749－1827）

- Laplace分布是一个有长尾的分布，pdf为

$$\mathrm{Lap}(x \mid \mu, b) = \frac{1}{2b} \exp\left\{-\frac{|x-\mu|}{b}\right\}.$$

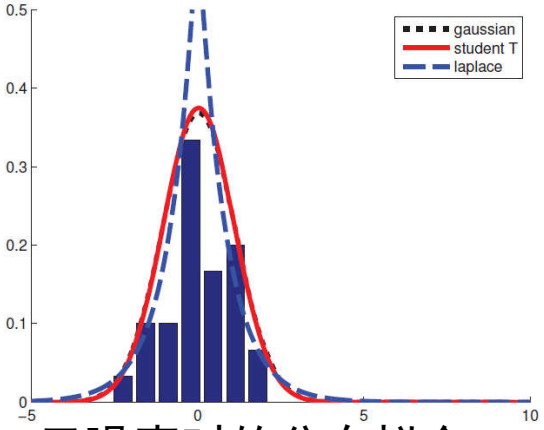
- 相比高斯分布，Laplace分布在0附近附近更集中→稀疏性



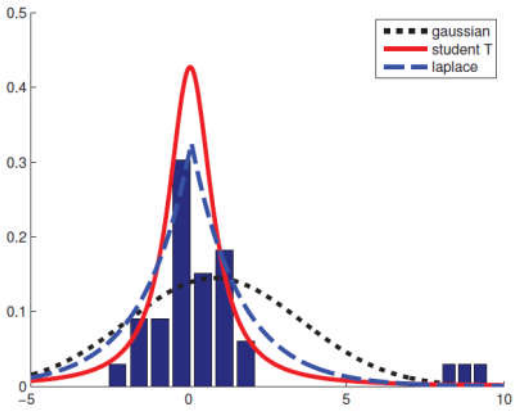
- Laplace 分布的均值：μ
- 方差：$2b^2$

Laplace分布

- 高斯分布对噪声敏感（$\log(p(x))$ 为到中心距离的二次函数 $\frac{1}{2\sigma^2}(x-\mu)^2$ ），而Laplace分布更鲁棒



无噪声时的分布拟合



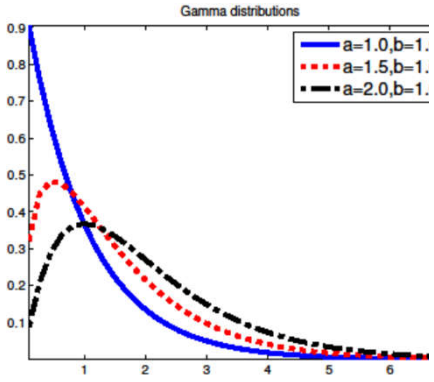
有噪声时的分布拟合

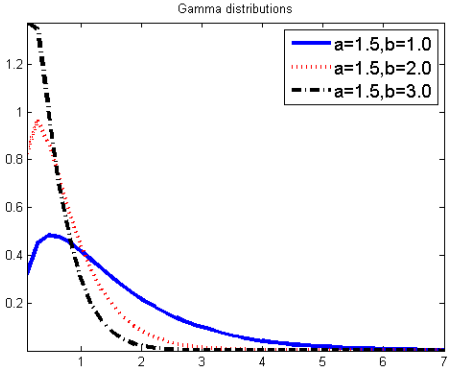
Gamma分布

- 对任意正实数随机变量 $x>0$, Gamma分布为 $x \sim \mathrm{Ga}(\mathrm{shape}=a, \mathrm{rate}=b)$

$$p(x|a,b)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-xb}$$

- 其中 $\Gamma(a)$ 为Gamma函数，a为形状参数，b为比率度参数





对Gamma函数感兴趣的同学可阅读：LDA数学八卦

Gamma分布

- Gamma分布的另一种表示：用尺度参数代替比率参数

$$\mathrm{Ga}(x|\mathrm{shape}=\alpha,\mathrm{scale}=\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}$$

$$=\mathrm{Ga}\left(x|\mathrm{shape}=\alpha,\mathrm{rate}=\frac{1}{\beta}\right)$$

- 反Gamma分布：若 $X\sim\mathrm{Ga}(a,b)$ ，则 $\frac{1}{X}\sim\mathrm{IG}(a,b)$

$$\mathrm{IG}(x|\mathrm{shape}=\alpha,\mathrm{scale}=\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{-\alpha-1}e^{-x/\beta}$$

反Gamma分布用于正态分布方差的共轭先验

Gamma分布

- Gamma分布的均值：a/b

  众数：$(a-1)/b$

  方差：a/b^2

- 反Gamma分布的均值：$b/(a-1)$

  众数：$b/(a+1)$

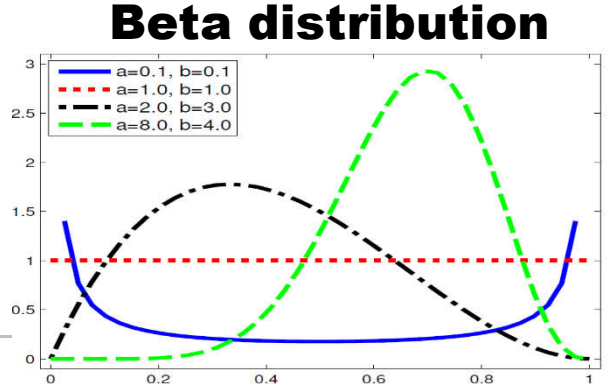
  方差：$b^2/(a-1)^2(a-2)$

Beta分布

- Beta分布的支持区间为[0,1]：$\mathrm{Beta}(x|a,b)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$
- 其中Beta函数$B(a,b):=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
- Beta分布的均值、众数和方差：$\mathbb{E}(x)=\frac{a}{a+b}$, $\mathrm{mode}[x]=\frac{a-1}{a+b-2}$

$$\mathbb{V}(x)=\frac{ab}{(a+b)^2(a+b+1)}$$

- 当 $0<a<1,0<b<1$ 时,在0和1处有两个峰值
- 当 $a>1,b>1$ 时, 有单个峰值
- 当a = b =1 时，为均匀分布



Beta 分布可作为二项分布的参数的共轭先验分布

Dirichlet分布

- 将Beta分布扩展到多维，即得到Dirichlet分布。其pdf为

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

- 其中B函数为 $B(\boldsymbol{\alpha}) := \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\alpha_0)}, \alpha_0 = \sum_{k=1}^{K}\alpha_k$ 。

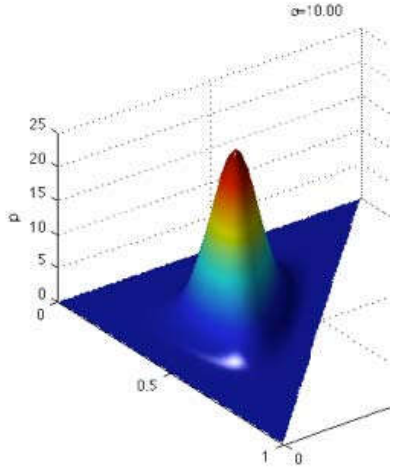
Dirichlet分布在文档分析中的主题模型LDA（ Latent Dirichlet Allocation ）用到。

Johann Peter Gustav Lejeune Dirichlet (1805-1859)

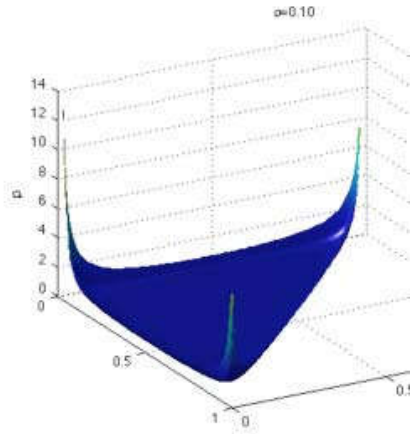
Dirichlet分布

$$\mathbb{E}(x_k)=\frac{\alpha_k}{\alpha_0},\quad \text{mode}[x_k]=\frac{\alpha_k-1}{\alpha_0-K}$$

$$\mathbb{V}(x_k)=\frac{\alpha_k(\alpha_0-\alpha_k)}{(\alpha_0)^2(\alpha_0+1)}$$



- 参数α_0 控制分布的强度（分布有多尖）
α_k 控制峰值出现的地方



分布的混合

- 通过组合一些简单的概率分布来定义新的概率分布也很常见。
- 一种通用的组合方法是构造混合分布（mixture distribution）：混合分布由一些组件 (component) 分布构由哪个组件分布产生的取决于从一个 Multinoulli 分布中采样的结果成。每次实验，样本是：

$$p(x)=\sum_k p(c=k)\,p(x|c=k)$$

  - 其中 $p(c)$ 是对各组件的一个 Multinomial 分布
- 例：经验分布就是以 Dirac 分布为组件的混合分布。

混合高斯模型

- 一个非常强大且常见的混合模型是高斯混合模型（Gaussian Mixture Model，GMM）
  - 组件 $p(x|c=k)$ 是高斯分布
  - 每个组件用自己的参数：均值、方差-协方差矩阵
  - 组件也可以共享参数：每个组件的方差-协方差矩阵相等…

- GMM是概率密度的万能近似器（universal approximator）：任何平滑的概率密度都可以用具有足够多组件的高斯混合模型以任意精度逼近。

► 分布的混合

- 一些有意思的分布可以表示为一组无限个高斯的加权和，其中每个高斯的方差不同

$$p(x)=\int \mathcal{N}(x|\mu,\tau^2)\pi(\tau^2)d\tau^2$$

某个分布

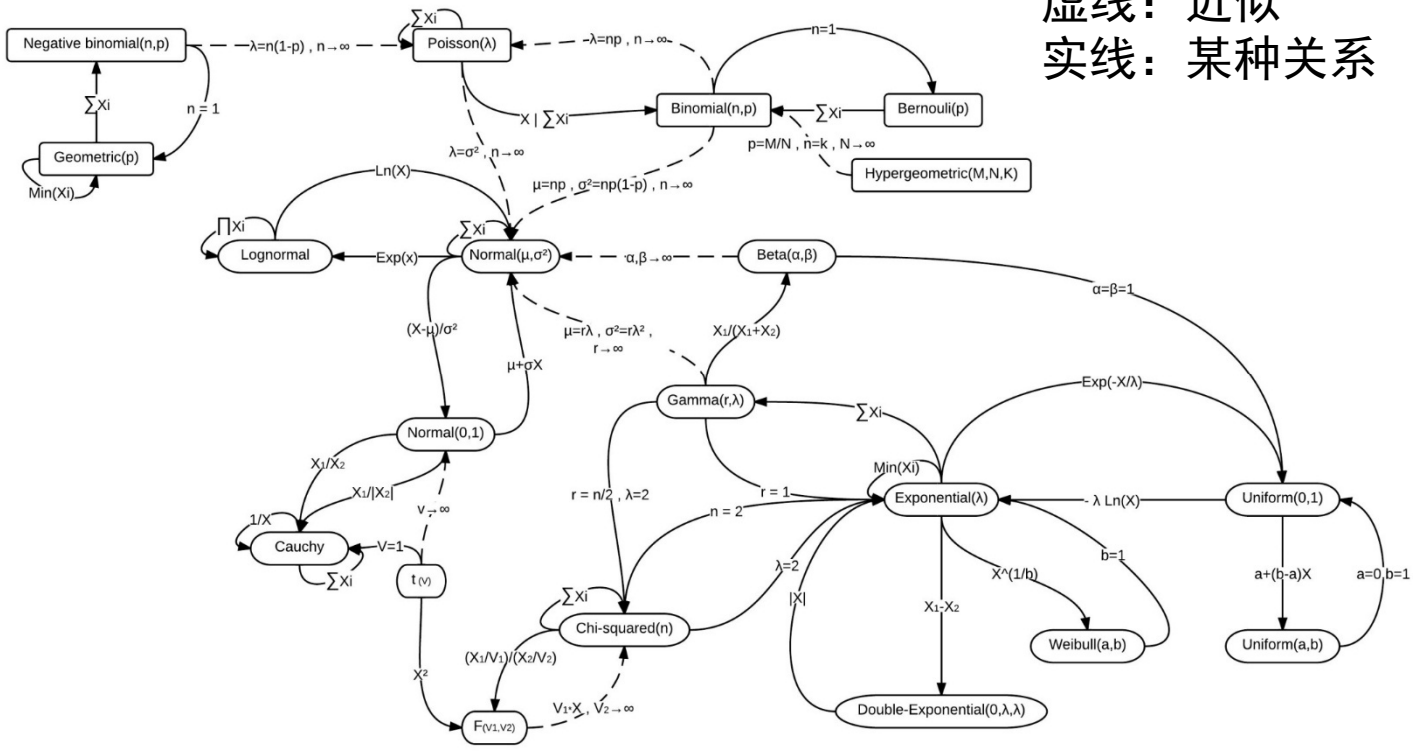
- 如Student分布：

$$\mathcal{T}(x|\mu,\sigma^2,v)=\int_0^\infty \mathcal{N}(x|\mu,\sigma^2/\lambda)\mathrm{Ga}\left(\lambda|\frac{v}{2},\frac{v}{2}\right)d\lambda$$

  - Student 分布和高斯分布很像，但尾巴更长
  - 当自由度$v\to\infty$ 时，极限分布为高斯分布

$$\mathcal{T}(x|\mu,\sigma^2,\infty)=\lim_{v\to\infty}\int_0^\infty \mathcal{N}(x|\mu,\sigma^2/\lambda)\mathrm{Ga}\left(\lambda|\frac{v}{2},\frac{v}{2}\right)d\lambda=\mathcal{N}(x|\mu,\sigma^2)$$

各分布之间的关系



虚线：近似
实线：某种关系

http://www.zhizhihu.com/html/y2012/4117.html

四、抽样分布

IID（Independent Identically Distribution）样本

- 当 $X_1,...,X_N$ 互相独立且有相同的边缘分布F时，记为 $X_1,...,X_N \sim F$
- 我们称 $X_1,...,X_N$ 为独立同分布（ Independent Identically Distribution, IID）样本，表示$X_1,...,X_N$是从相同分布独立抽样/采样，我们也称 $X_1,...,X_N$ 是分布F的随机样本。若F有密度p，也可记为$X_1,...,X_N \sim p$

抽样分布

- 令 $X_1, X_2 ..., X_N$ 为独立同分布样本（IID），其均值和方差分别为μ和σ^2。则样本均值 $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ 为一统计量，也是随机变量，因此也可对其进行分布进行描述，该分布称为统计量的抽样分布。
  - 请不要将X_i的分布与$\bar{X}_N$ 的分布混淆：如X_i的分布是均匀分布，当N足够大时，$\bar{X}_N$ 的分布为正态分布（中心极限定理）

样本均值和样本方差

- 最简单的数据分析问题：如何知道产生数据的分布的期望和方差
- 令 $X_1,\ldots,X_N$ 为IID，样本均值定义为: $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$
- 样本方差定义为: $S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}_N\right)^2$
- 问题：样本均值和样本方差会是分布F真正期望和方差的很好估计？

样本均值和样本方差

- 假设 $X_1,\ldots,X_N$ 为IID，$\mu=\mathbb{E}(X_i)$, $\sigma^2=\mathbb{V}(X_i)$，那么

$$\mathbb{E}(\bar{X}_N)=\mu,\quad \mathbb{V}(\bar{X}_N)=\frac{\sigma^2}{N},\quad \mathbb{E}(S_N^2)=\sigma^2$$

- 即 $\bar{X}_N$ 和 S_N^2 分别为 μ 和 σ 的很好估计（无偏估计）.
  - 样本数 N 越大，$\mathbb{V}(\bar{X}_N)$ 越小，$\bar{X}_N$ 越接近 μ

证明：$\mathbb{E}\left(\overline{X}_N\right)=\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right)=\frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N}X_i\right)=\frac{1}{N}\times\sum_{i=1}^{N}\mathbb{E}\left(X_i\right)=\frac{1}{N}\times N\mu=\mu$

$\mathbb{V}\left(\overline{X}_N\right)=\mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right)=\frac{1}{N^2}\mathbb{V}\left(\sum_{i=1}^{N}X_i\right)=\frac{1}{N^2}\times\sum_{i=1}^{N}\mathbb{V}\left(X_i\right)=\frac{1}{N^2}\times N\sigma^2=\frac{\sigma^2}{N}$

$\mathbb{E}\left(S_N^2\right)=\mathbb{E}\left(\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\overline{X}_N\right)^2\right)=\frac{1}{N-1}\mathbb{E}\left(\sum_{i=1}^{N}\left(X_i-\overline{X}_N\right)^2\right)$

$=\frac{1}{N-1}\mathbb{E}\left(\sum_{i=1}^{N}X_i^2-N\overline{X}_i^2\right)$ $\left[\sum_{i=1}^{N}\left(X_i-\overline{X}_N\right)^2=\sum_{i=1}^{N}\left(X_i^2-2\overline{X}_NX_i+\overline{X}_N^2\right)=\sum_{i=1}^{N}X_i^2-N\overline{X}_N^2\right]$

$=\frac{1}{N-1}\left(\mathbb{N}\mathbb{E}\left(X_1^2\right)-N\mathbb{E}\left(\overline{X}_N^2\right)\right)$

$\mathbb{E}\left(X_1^2\right)=\mathbb{V}\left(X_1\right)+\left(\mathbb{E}X_1\right)^2=\sigma^2+\mu^2$

$\mathbb{E}\left(\overline{X}_N^2\right)=\mathbb{V}\left(\overline{X}_N^2\right)+\mathbb{E}\left(\overline{X}_N^2\right)=\frac{\sigma^2}{N}+\mu^2$

$\mathbb{E}\left(S_N^2\right)=\frac{1}{N-1}\left(\mathbb{N}\mathbb{E}\left(X_1^2\right)-N\mathbb{E}\left(\overline{X}_N^2\right)\right)=\frac{1}{N-1}\left(N\left(\sigma^2+\mu^2\right)-N\left(\frac{\sigma^2}{N}+\mu^2\right)\right)=\sigma^2$

两种收敛的定义

- 令$X_1, X_2 ..., X_N$为随机变量序列，X为另一随机变量，用F_N表示X_N的CDF，用F表示X的CDF
- 1、如果对每个$\varepsilon > 0$，当$N \to \infty$时，

$$\mathbb{P}\left(\left|X_N - X\right| > \varepsilon\right) \to 0$$

- 则X_N依概率收敛于X，记为$X_N \xrightarrow{P} X$。
- 2、如果对所有F的连续点t，有

$$\lim_{N\to\infty} F_N(t) = F(t)$$

- 则X_N依分布收敛于X，记为$X_n \rightsquigarrow X$。

► 弱大数定律（WLLN）

- 独立同分布（IID）的随机变量序列 $X_1, X_2 \ldots, X_N$ ，$\mathbb{E}(X_i) = \mu$ ，方差 $\mathbb{V}(X_i) = \sigma^2 < \infty$ ，则样本均值 $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ 依概率收敛于期望 μ，即对任意 $\varepsilon > 0$

$$\lim_{N\to\infty} \mathbb{P}\left(\left|\overline{X}_N - \mu\right| > \varepsilon\right) = 0$$

- 称 $\overline{X}_N$ 为 μ 的一致估计（一致性）
- 在定理条件下，当样本数目 N 无限增加时，随机样本均值将几乎变成一个常量
- 样本方差也依概率收敛于方差 σ^2

中心极限定理(Central Limit Theorem, CLT)

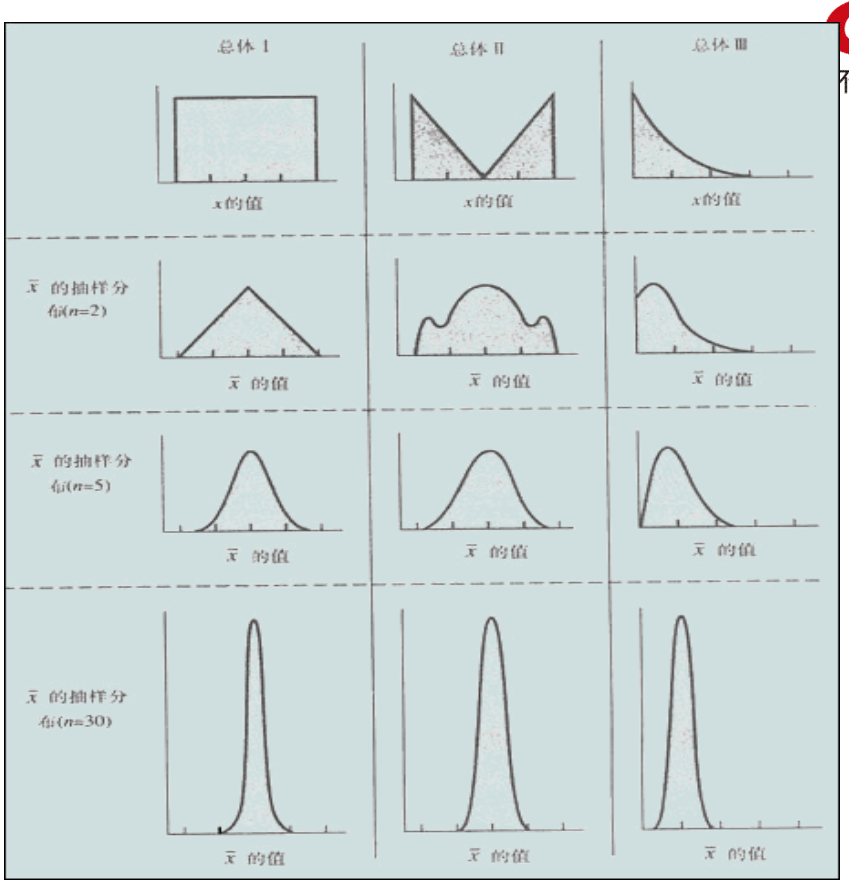
- 独立同分布（IID）的随机变量序列$X_1, X_2 ..., X_N$，$\mathbb{E}(X_i) = \mu$ $\mathbb{V}(X_i) = \sigma^2 < \infty$，则样本均值 $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ 近似服从期望为μ，方差为σ^2/N的正态分布，即

$$Z_N \equiv \frac{\sqrt{N}\left(\overline{X}_N - \mu\right)}{\sigma} \approx Z$$

其中Z为标准正态分布，也记为$\overline{X}_N \approx \mathcal{N}\left(\mu, \sigma^2/N\right)$

- 无论随机变量X为何种类型的分布，只要满足定理条件，其样本均值就近似服从正态分布。正态分布很重要
  - 但近似的程度与原分布有关
  - 大样本统计推理的理论基础

中心极限定理



中心极限定理试验
http://jyjs.gzhu.edu.cn:8080/skills/portal/resources/65995/67826/entryFile/swf/zhongxinjixian.htm

中心极限定理

- 标准差σ通常不知道，可用样本标准差代替，中心极限定理仍成立，即

$$\frac{\sqrt{N}\left(\overline{X}_N-\mu\right)}{S_N}\approx Z$$

- 其中

$$S_N^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\overline{X}_N\right)^2$$

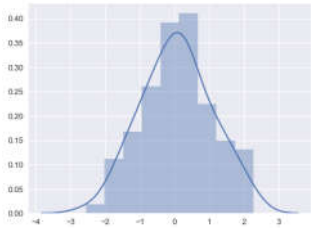
五、分布的估计

分布估计

- 已知分布的类型，但参数未知：参数估计
  - 第3/4节课内容

- 分布类型未知：非参数估计
  - 直方图、核密度估计
  - 根据有限个统计量估计分布：极大熵原理

非参数概率模型



- 一种非参数的概率估计方式是直方图
  - 将输入空间划分为M个箱子(bin), 箱子的宽度为 $h = 1/M$

则这些箱子为

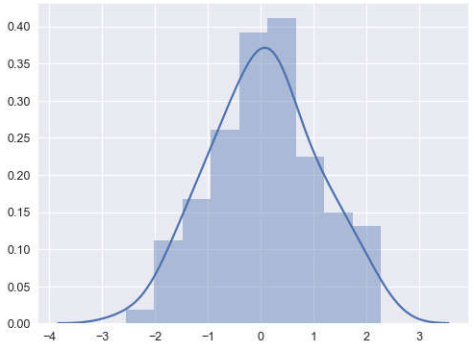
$$B_1 = \left[0, \frac{1}{M}\right), \; B_2 = \left[\frac{1}{M}, \frac{2}{M}\right), \; \cdots, B_M = \left[\frac{M-1}{M}, 1\right)$$

  - 计算落入箱子b中的样本的数目ν_b ,落入箱子b的比率为 $\hat{p}_b = \frac{\nu_b}{N}$
  - 则直方图估计为

$$\hat{p}(x) = \sum_{b=1}^{M} \frac{\hat{p}_b}{h} \mathbb{I}(x \in B_b) = \frac{1}{N} \sum_{b=1}^{M} \frac{1}{h} \mathbb{I}(x \in B_b)$$

直方图估计

- 当箱子数目M为固定值时，该估计为参数模型
  - 参数个数固定/有限
- 通常M与样本数N有某种关系
  - 非参数模型，如何选择M？
    - 交叉验证



核密度估计

- 直方图不连续
  - 箱中每个样本的权重相等

- 核密度估计：更平滑，比直方图收敛更快
  - 基本思想：每个样本的权重随其到目标点的距离平滑衰减

73

核密度估计

直方图估计：

$$\hat{p}(x)=\frac{1}{N}\sum_{b=1}^{M}\frac{1}{h}\mathbb{I}(x\in B_b)$$

- 核密度估计定义为 $\hat{p}(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)$
- 其中参数h称为带宽(bandwidth)，核函数可为任意平滑的函数K，满足

$$K(u)\geq 0,\qquad \int K(u)du=1,$$

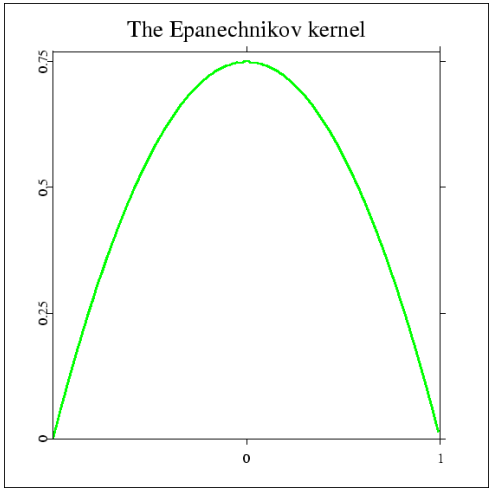
$$\int uK(u)du=0,\quad \sigma_K^2=\int u^2K(u)du>0$$

- 实质：
  - 对样本点施以不同的权，用加权来代替通常的计数

核函数例子

- *Epanechnikov* 核：$K(u)=\frac{3}{4}\left(1-u^2\right)\mathbb{I}\left(|u|\leq 1\right)$
  - 使风险最小的核函数
  - 亦被称为抛物面核或者叫做二次核函数

- 高斯核：$K(u)=\frac{1}{\sqrt{2\pi}}e^{u^2/2}$



核密度估计-带宽

- 为了构造一个核密度估计，我们需要选择核函数K和带宽h
- 带宽的选择比核函数的选择更重要
- 关键是平滑参数— h
  - 小的平滑参数h：估计结果受噪音影响较大，当 $h \to 0$ 得到针状分布
  - 大的平滑参数h：估计结果过分平滑，当 $h \to \infty$ ，趋向于均匀分布

Python Seaborn 数据集分布的可视化

- seaborn包在matplotlib上继承开发，使用更方便
  - http://seaborn.pydata.org/index.html
  - 要安装的包:numpy、sicpy、pandas、matplotlib、seaborn（Anaconda均已集成）

- 快速查看单变量的分布：distplot()函数
  - 默认情况下，将绘制一个直方图，并拟合出核密度估计

直方图

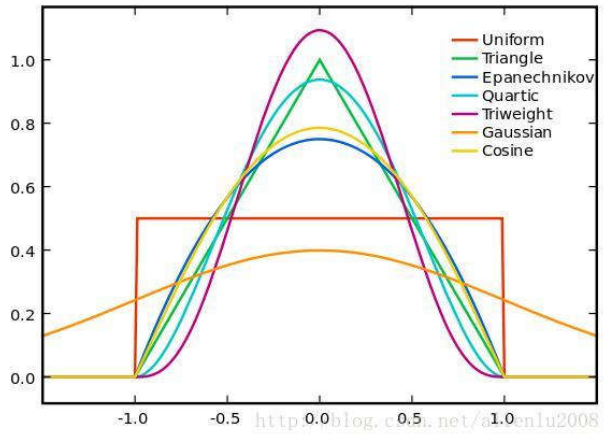
- seaborn.distplot集成了matplotlib的 hist函数 ， seaborn的 kdeplot() 和 rugplot()估计数据的pdf
- 其中hist函数中箱子尺寸采用Freedman-Diaconis Rule：

$$BinSize = 2 \times \mathrm{IQR}(\mathcal{D}) N^{-1/3}$$

  - 其中IQR为数据的四分位矩，$\mathcal{D}$为数据，N为样本数

核密度估计

- seaborn的kdeplot函数支持多种核函数和带宽模式
- kernel参数支持六种核函数
  - gau (Gaussian)　　缺省值
  - cos(Cosine)
  - biw(Quartic(biweight))
  - epa(Epanechnikov)
  - tri （ Tricube)
  - triw (Triweight)



仅用核密度做定性分析的话，任何一种核函数均可

核密度估计（cont.）

- kdeplot函数的bw参数支持四种带宽选择方式
  - scott (斯考特带宽法）　　缺省值
  - silverman （西尔弗曼带宽法）
  - scalar （标量带宽法）　　自己定义各种带宽
  - pair of scalars （标量对带宽法）

► 案例分析：Rent Listing Inquries 数据集

- RelationSVs_RentalListingInquiries.ipynb

极大熵原理

- 参数估计时假设分布的类型已知，实际上这通常很难得到
- 非参数估计在维数较高时又会遇到维数灾难问题

- 极大熵原理：介于二者之间
  - 充分统计量
  - 指数分布族

极大熵原理

- 极大熵原理：1957 年由E. T. Jaynes 提出
- 主要思想：在只掌握关于未知分布的部分知识时，应该选取符合这些知识但熵值最大的概率分布
- 原理的实质：
  - 约束：符合已知知识（特征的统计量）
  - 极大熵：关于未知分布最合理的推断 = 符合已知知识最不确定或最随机的推断
    - 唯一不偏不倚的选择，任何其它的选择都意味着我们增加了其它的约束和假设，这些约束和假设根据我们掌握的信息无法作出

充分统计量

- 统计量：给定数据 $x^N=(x_1,\ldots,x_N)$，统计量为 $t=\phi(x^N)=\phi(x_1,\ldots,x_N)$
- 如 $t_1=\mu=\frac{1}{N}\sum_{i=1}^{N}x_i,\ t_2=\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$
- 充分统计量：统计量（向量）包含了计算参数所需的所有信息

$$\mathbb{P}(\theta|t,x^n)=\mathbb{P}(\theta|t)$$

- 只需记住充分统计量，无需记住大量的样本。

84

概率分布族

- 问题：给定训练样本 $x^N = (x_1, ..., x_N) \sim p(x)$ ，不知道分布的形式，根据数据求密度的估计

- 我们从数据的M个统计量开始: $t_j = \frac{1}{N}\sum_{i=1}^{N}\phi_j(x_j), j = 1, ..., M$

- 如 $\phi_1(x) = 1,$ $\phi_2(x) = x$

  $\phi_3(x) = (x-\mu)^2,$ $\phi_4(x) = \ln x$

85

概率分布族

- 当样本数N增加时，样本均值会接近真正的期望

$$t_j = \frac{1}{N}\sum_{i=1}^{N}\phi_j\left(x_j\right) \approx \int \phi_j\left(x\right) p\left(x\right) dx = \mathbb{E}_p\left(\phi_j\left(x\right)\right)$$

- 我们的目标是计算$q\left(x\right)$ ，一个合理的假设是二者产生相同的期望

$$\int \phi_j\left(x\right) q\left(x\right) dx = \mathbb{E}_q\left(\phi_j\left(x\right)\right) = t_j \approx \mathbb{E}_p\left(\phi_j\left(x\right)\right)$$

- 另外还有一个约束是$\int q\left(x\right) dx = 1$

概率分布族

- 所以共有$M+1$个约束。但这些约束还不足以定义一个密度，因为概率密度的自由度非常大。因此我们选择一个满足这$M+1$个约束的熵最大的密度
  - 这个概率密度是最无偏的，满足约束的最自由的分布

熵：
$$H(X) = -\int p(x)\log p(x)dx$$

- 这样得到一个优化问题

$$q^*(x) = \arg\max_{q(x)}\left(-\int q(x)\log q(x)dx\right)$$

- 满足 $\int \phi_j(x)q(x)dx = t_j,\ j = 1,...,M$

Lagrange乘子法

- $J(q)=-\int q(x)\log q(x)dx$

$$-\sum_{j=1}^{M}\lambda_j\left(\int\phi_j(x)q(x)dx-t_j\right)-\lambda_0\left(\int q(x)dx-1\right)$$

- 令 $\frac{\partial J(q)}{\partial q}=0$，得到

$$-\log q-1-\sum_{j=1}^{M}\lambda_j\phi_j(x)-\lambda_0=0$$

指数分布族

- $q(x \mid \boldsymbol{\lambda}) = \exp\left(-1 - \sum_{j=1}^{M} \lambda_j \phi_j(x) - \lambda_0\right)$

$$= \frac{1}{Z} \exp\left(-\sum_{j=1}^{M} \lambda_j \phi_j(x)\right)$$

- 其中Z为归一化常数，参数 $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_M)$

- 参数可以通过MLE求解。选取的统计量越多，$p(x)$越接近$q(x)$。给定N个有限的数据，$M < N$，否则会过拟合。通常$M = O(\log N)$ 。

► 例：

- 如果我们取两个统计量：

$$\phi_1(x) = x,\ \phi_2(x) = x^2$$

- 我们会得到高斯分布：

$$q(x \mid \boldsymbol{\lambda}) = \frac{1}{Z}\exp\left(-\sum_{j=1}^{2}\lambda_j\phi_j(x)\right) = \frac{1}{Z}\exp\left(\lambda_1 x + \lambda_2 x^2\right)$$

高斯分布：给定均值和协方差矩阵下的最大熵分布

高斯分布很重要：我们能稳定地从数据中估计分布的均值和协方差，我们希望能得到一个分布，既满足均值和协方差的约束，同时又不增加其他额外的假设

► 例：人脸形状

- 极小极大熵

$$p(I;\beta,F)=\frac{1}{Z(\beta,F)}\exp\{-\sum_{j=1}^{K}\sum_{(x,y)}\beta_j(F_j*I(x,y))\}$$

极小熵：用于选择特征
极大熵：用于求分布

$$p^*=\arg\min_{F}\{\max_{\beta}\text{entropy}(p(I;\beta,F))\}$$

ICCV2001: Learning inhomogeneous Gibbs models of faces by minimax entropy (Liu, Zhu, and Shum, ICCV, 2001)

CSDN
不 止 于 代 码

THANK YOU

AI 100