概率论与随机过程

一、概率与分布

1.1 条件概率与独立事件

1. 条件概率：已知 A 事件发生的条件下 B 发生的概率，记作 $P(B \mid A)$，它等于事件 AB 的概率相对于事件 A 的概率，即：$P(B \mid A)=\frac{P(AB)}{P(A)}$ 。其中必须有 $P(A)>0$ 。

2. 条件概率分布的链式法则：对于 n 个随机变量 $X_1, X_2, \cdots, X_n$ ，有：

$$P(X_1, X_2, \cdots, X_n)=P(X_1) \prod_{i=2}^{n} P(X_i \mid X_1, \cdots, X_{i-1})$$

3. 两个随机变量 X, Y 相互独立的数学描述：$P(X, Y)=P(X) P(Y)$ 。记作： $X \perp Y$ 。
4. 两个随机变量 X, Y 关于随机变量 Z 条件独立的数学描述：$P(X, Y \mid Z)=P(X \mid Z) P(Y \mid Z)$ 。
   记作： $X \perp Y \mid Z$ 。

1.2 联合概率分布

1. 定义X 和 Y 的联合分布为：$P(a, b)=P\{X \leq a, Y \leq b\}, \quad -\infty<a, b<+\infty$ 。
   - X 的分布可以从联合分布中得到：
   $$P_X(a)=P\{X \leq a\}=P\{X \leq a, Y \leq \infty\}=P(a, \infty), \quad -\infty<a<+\infty$$
   - Y 的分布可以从联合分布中得到：
   $$P_Y(b)=P\{Y \leq b\}=P\{X \leq \infty, Y \leq b\}=P(\infty, b), \quad -\infty<b<+\infty$$

2. 当 X 和 Y 都是离散随机变量时， 定义 X 和 Y 的联合概率质量函数为： $p(x, y)=P\{X=x, Y=y\}$
   则 X 和 Y 的概率质量函数分布为：
$$p_X(x)=\sum_y p(x, y) \quad p_Y(y)=\sum_x p(x, y)$$

3. 当 X 和 Y 联合地连续时，即存在函数 $p(x, y)$ ，使得对于所有的实数集合 $\mathbb{A}$ 和 $\mathbb{B}$ 满足：
$$P\{X \in \mathbb{A}, Y \in \mathbb{B}\}=\int_{\mathbb{B}} \int_{\mathbb{A}} p(x, y) dx dy$$
   则函数 $p(x, y)$ 称为 X 和 Y 的概率密度函数。
   - 联合分布为：$P(a, b)=P\{X \leq a, Y \leq b\}=\int_{-\infty}^{a} \int_{-\infty}^{b} p(x, y) dx dy$ 。
   - X 和 Y 的分布函数以及概率密度函数分别为：

$$P_X(a)=\int_{-\infty}^{a}\int_{-\infty}^{\infty}p(x,y)dxdy=\int_{-\infty}^{a}p_X(x)dx$$
$$P_Y(b)=\int_{-\infty}^{\infty}\int_{-\infty}^{b}p(x,y)dxdy=\int_{-\infty}^{b}p_Y(y)dy$$
$$p_X(x)=\int_{-\infty}^{\infty}p(x,y)dy$$
$$p_Y(y)=\int_{-\infty}^{\infty}p(x,y)dx$$

.

二、期望和方差

2.1 期望

1. 期望描述了随机变量的平均情况，衡量了随机变量 X 的均值。它是概率分布的泛函（函数的函数）。
   - 离散型随机变量 X 的期望：$\mathbb{E}[X]=\sum_{i=1}^{\infty}x_ip_i$ 。

     若右侧级数不收敛，则期望不存在。
   - 连续性随机变量 X 的期望： $\mathbb{E}[X]=\int_{-\infty}^{\infty}xp(x)dx$ 。

     若右侧极限不收敛，则期望不存在。
2. 定理：对于随机变量 X ，设 $Y=g(X)$ 也为随机变量，$g(\cdot)$ 是连续函数。
   - 若 X 为离散型随机变量，若 Y 的期望存在，则：$\mathbb{E}[Y]=\mathbb{E}[g(X)]=\sum_{i=1}^{\infty}g(x_i)p_i$ 。

     也记做：$\mathbb{E}_{X\sim P(X)}[g(X)]=\sum_x g(x)p(x)$ 。
   - 若 X 为连续型随机变量，若 Y 的期望存在，则 ：$\mathbb{E}[Y]=\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)p(x)dx$ 。

     也记做：$\mathbb{E}_{X\sim P(X)}[g(X)]=\int g(x)p(x)dx$ 。

   该定理的意义在于：当求 $\mathbb{E}(Y)$ 时，不必计算出 Y 的分布，只需要利用 X 的分布即可。

   该定理可以推广至两个或两个以上随机变量的情况。对于随机变量 X,Y，假设 $Z=g(X,Y)$ 也是随机变量， $g(\cdot)$ 为连续函数，则有：$\mathbb{E}[Z]=\mathbb{E}[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)p(x,y)dxdy$ 。也记做：$\mathbb{E}_{X,Y\sim P(X,Y)}[g(X,Y)\int g(x,y)p(x,y)dxdy$ 。
3. 期望性质：
   - 常数的期望就是常数本身。
   - 对常数 C 有 ：$\mathbb{E}[CX]=C\mathbb{E}[X]$ 。
   - 对两个随机变量 X,Y，有：$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$ 。

     该结论可以推广到任意有限个随机变量之和的情况。
   - 对两个相互独立的随机变量，有：$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$ 。

     该结论可以推广到任意有限个相互独立的随机变量之积的情况。

2.2 方差

1. 对随机变量 X ，若 $\mathbb{E}[(X-\mathbb{E}[X])^2]$ 存在，则称它为 X 的方差，记作 $Var[X]$ 。

   X 的标准差为方差的开平方。即：

$$Var[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]$$
$$\sigma=\sqrt{Var[X]}$$

- 方差度量了随机变量 X 与期望值偏离的程度，衡量了 X 取值分散程度的一个尺度。
- 由于绝对值 $|X-\mathbb{E}[X]|$ 带有绝对值，不方便运算，因此采用平方来计算。

  又因为 $|X-\mathbb{E}[X]|^2$ 是一个随机变量，因此对它取期望，即得 X 与期望值偏离的均值。

2. 根据定义可知：

$$Var[X]=\mathbb{E}[(X-\mathbb{E}[X])^2]=\mathbb{E}[X^2]-(\mathbb{E}[X])^2$$
$$Var[f(X)]=\mathbb{E}[(f(X)-\mathbb{E}[f(X)])^2]$$

3. 对于一个期望为 μ， 方差为 $\sigma^2,\sigma\neq 0$ 的随机变量 X，随机变量 $X^*=\frac{X-\mu}{\sigma}$ 的数学期望为0，方差为1。 称 X^* 为 X 的标准化变量。
4. 方差的性质：
   - 常数的方差恒为 0 。
   - 对常数 C ，有 $Var[CX]=C^2Var[X]$ 。
   - 对两个随机变量 X,Y，有： $Var[X+Y]=Var[X]+Var[Y]+2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$

     当 X 和 Y 相互独立时，有 $Var[X+Y]=Var[X]+Var[Y]$ 。这可以推广至任意有限多个相互独立的随机变量之和的情况。
   - $Var[X]=0$ 的充要条件是 X 以概率1取常数。

2.3 协方差与相关系数

1. 对于二维随机变量 (X,Y) ，可以讨论描述 X 与 Y 之间相互关系的数字特征。
   - 定义 $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$ 为随机变量 X 与 Y 的协方差，记作 $Cov[X,Y]=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$。
   - 定义 $\rho_{XY}=\frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$ 为随机变量 X 与Y的相关系数，它是协方差的归一化。
2. 由定义可知：

$$Cov[X,Y]=Cov[Y,X]$$
$$Cov[X,X]=Var[X]$$
$$Var[X+Y]=Var[X]+Var[Y]+2Cov[X,Y]$$

3. 协方差的性质：
   - $Cov[aX,bY]=abCov[X,Y]$，a,b 为常数。
   - $Cov[X_1+X_2,Y]=Cov[X_1,Y]+Cov[X_2,Y]$
   - $Cov[f(X),g(Y)]=\mathbb{E}[(f(X)-\mathbb{E}[f(X)])(g(Y)-\mathbb{E}[g(Y)])]$
   - $\rho[f(X),g(Y)]=\frac{Cov[f(X),g(Y)]}{\sqrt{Var[f(X)]}\sqrt{Var[g(Y)]}}$
4. 协方差的物理意义：
   - 协方差的绝对值越大，说明两个随机变量都远离它们的均值。
   - 协方差如果为正，则说明两个随机变量同时趋向于取较大的值或者同时趋向于取较小的值；如果为负，则说明一个随变量趋向于取较大的值，另一个随机变量趋向于取较小的值。
   - 两个随机变量的独立性可以导出协方差为零。但是两个随机变量的协方差为零无法导出独立性。

因为独立性也包括：没有非线性关系。有可能两个随机变量是非独立的，但是协方差为零。如：假设随机变量 $X \sim U[-1,1]$ 。定义随机变量 S 的概率分布函数为：

$$P(S=1)=\frac{1}{2}P(S=-1)=\frac{1}{2}$$

定义随机变量 $Y=SX$，则随机变量 X,Y 是非独立的，但是有： $Cov[X,Y]=0$ 。

5. 相关系数的物理意义：考虑以随机变量 X 的线性函数 $a+bX$ 来近似表示 Y。以均方误差

$$e=\mathbb{E}[(Y-(a+bX))^2]=\mathbb{E}[Y^2]+b^2\mathbb{E}[X^2]+a^2-2b\mathbb{E}[XY]+2ab\mathbb{E}[X]-2a\mathbb{E}[Y]$$

来衡量以 $a+bX$ 近似表达 Y 的好坏程度。 e 越小表示近似程度越高。

为求得最好的近似，则对 a,b 分别取偏导数，得到：

$$a_0=\mathbb{E}[Y]-b_0\mathbb{E}[X]=\mathbb{E}[Y]-\mathbb{E}[X]\frac{Cov[X,Y]}{Var[X]}$$

$$b_0=\frac{Cov[X,Y]}{Var[X]}$$

$$\min(e)=\mathbb{E}[(Y-(a_0+b_0X))^2]=(1-\rho_{XY}^2)Var[Y]$$

因此有以下定理：

- $|\rho_{XY}|\le 1$ ($|\cdot|$ 是绝对值) 。
- $|\rho_{XY}|=1$ 的充要条件是：存在常数 a,b 使得 $P\{Y=a+bX\}=1$。

6. 当 $|\rho_{XY}|$ 较大时， e 较小，意味着随机变量 X 和 Y 联系较紧密。于是 ρ_{XY} 是一个表征 X、Y 之间线性关系紧密程度的量。

7. 当 $\rho_{XY}=0$ 时，称 X 和 Y 不相关。

- 不相关是就线性关系来讲的，而相互独立是一般关系而言的。
- 相互独立一定不相关；不相关则未必独立。

2.4 协方差矩阵

1. 设 X 和 Y 是随机变量。

- 若 $\mathbb{E}[X^k],k=1,2,\cdots$ 存在，则称它为 X 的 k 阶原点矩，简称 k 阶矩。
- 若 $\mathbb{E}[(X-\mathbb{E}[X])^k],k=2,3,\cdots$ 存在，则称它为 X 的 k 阶中心矩。
- 若 $\mathbb{E}[X^kY^l],k,l=1,2,\cdots$ 存在，则称它为 X 和 Y 的 $k+l$ 阶混合矩。
- 若 $\mathbb{E}[(X-\mathbb{E}[X])^k(Y-\mathbb{E}[Y])^l],k,l=1,2,\cdots$ 存在，则称它为 X 和 Y 的 $k+l$ 阶混合中心矩。

因此：期望是一阶原点矩，方差是二阶中心矩，协方差是二阶混合中心矩。

2. 协方差矩阵：

- 二维随机变量 (X_1,X_2) 有四个二阶中心矩（假设他们都存在），记作：

$$\begin{aligned}c_{11}&=\mathbb{E}[(X_1-\mathbb{E}[X_1])^2]\\c_{12}&=\mathbb{E}[(X_1-\mathbb{E}[X_1])(X_2-\mathbb{E}[X_2])]\\c_{21}&=\mathbb{E}[(X_2-\mathbb{E}[X_2])(X_1-\mathbb{E}[X_1])]\\c_{22}&=\mathbb{E}[(X_2-\mathbb{E}[X_2])^2]\end{aligned}$$

称矩阵

$$\mathbf{C}=\begin{bmatrix}c_{11}&c_{12}\\c_{21}&c_{22}\end{bmatrix}$$

为随机变量 (X_1,X_2) 的协方差矩阵。

- 设 n 维随机变量 $(X_1,X_2,\cdots,X_n)$ 的二阶混合中心矩
  $c_{ij}=Cov[X_i,X_j]=\mathbb{E}[(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])]$ 都存在，则称矩阵

$$\mathbf{C}=\begin{bmatrix}c_{11}&c_{12}&\cdots&c_{1n}\\c_{21}&c_{22}&\cdots&c_{2n}\\\vdots&\vdots&\ddots&\vdots\\c_{n1}&c_{n2}&\cdots&c_{nn}\end{bmatrix}$$

  为 n 维随机变量 $(X_1,X_2,\cdots,X_n)$ 的协方差矩阵。

  由于 $c_{ij}=c_{ji},i\neq j,i,j=1,2,\cdots,n$ 因此协方差矩阵是个对称阵。

3. 通常 n 维随机变量的分布是不知道的，或者太复杂以致数学上不容易处理。因此实际中协方差矩阵非常重要。

三、大数定律及中心极限定理

3.1 切比雪夫不等式

1. 切比雪夫不等式：假设随机变量 X 具有期望 $\mathbb{E}[X]=\mu$， 方差 $Var(X)=\sigma^2$ ，则对于任意正数 ε ，下面的不等式成立：

$$P\{|X-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$$

   - 其意义是：对于距离 $\mathbb{E}[X]$ 足够远的地方（距离大于等于 ε ），事件出现的概率是小于等于 $\frac{\sigma^2}{\varepsilon^2}$。即事件出现在区间 $[\mu-\varepsilon,\mu+\varepsilon]$ 的概率大于 $1-\frac{\sigma^2}{\varepsilon^2}$ 。

     该不等式给出了随机变量 X 在分布未知的情况下， 事件 $\{|X-\mu|\leq\varepsilon\}$ 的下限估计。如：$P\{|X-\mu|<3\sigma\}\geq0.8889$ 。

   - 证明：

$$\begin{aligned}P\{|X-\mu|\geq\varepsilon\}&=\int_{|x-\mu|\geq\varepsilon}p(x)dx\leq\int_{|x-\mu|\geq\varepsilon}\frac{|x-\mu|^2}{\varepsilon^2}p(x)dx\\&\leq\frac{1}{\varepsilon^2}\int_{-\infty}^{\infty}(x-\mu)^2p(x)dx=\frac{\sigma^2}{\varepsilon^2}\end{aligned}$$

2. 切比雪夫不等式的特殊情况：设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，且具有相同的数学期望和方差：$\mathbb{E}[X_k]=\mu,Var[X_k]=\sigma^2$。 作前 n 个随机变量的算术平均： $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$， 则对于任意正数 ε 有：

$$\lim_{n\to\infty}P\{|\overline{X}-\mu|<\varepsilon\}=\lim_{n\to\infty}P\{|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu|<\varepsilon\}=1$$

   证明：根据期望和方差的性质有：$\mathbb{E}[\overline{X}]=\mu$， $Var[\overline{X}]=\frac{\sigma^2}{n}$ 。根据切比雪夫不等式有：

$$P\{|\overline{X}-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{n\varepsilon^2}$$

   则有 $\lim_{n\to\infty}P\{|\overline{X}-\mu|\geq\varepsilon\}=0$ ，因此有：$\lim_{n\to\infty}P\{|\overline{X}-\mu|<\varepsilon\}=1$ 。

3.2 大数定理

1. 依概率收敛：设 $Y_1,Y_2,\cdots,Y_n,\cdots$ 是一个随机变量序列， a 是一个常数。

   若对于任意正数 ε 有 ： $\lim_{n\to\infty}P\{|Y_n-a|\le\varepsilon\}=1$ ，则称序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 依概率收敛于 a 。

   记作： $Y_n\xrightarrow{P}a$

2. 依概率收敛的两个含义：

   - 收敛：表明这是一个随机变量序列，而不是某个随机变量；且序列是无限长，而不是有限长。
   - 依概率：表明序列无穷远处的随机变量 Y_∞ 的分布规律为：绝大部分分布于点 a ，极少数位于 a 之外。且分布于 a 之外的事件发生的概率之和为0。

3. 大数定理一： 设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，且具有相同的数学期望和方差：$\mathbb{E}[X_k]=\mu,Var[X_k]=\sigma^2$。 则序列： $\overline{X}=\frac{1}{n}\sum_{k=1}^nX_k$ 依概率收敛于 μ ， 即 $\overline{X}\xrightarrow{P}\mu$ 。

   注意：这里并没有要求随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 同分布。

4. 伯努利大数定理： 设 n_A 为 n 次独立重复实验中事件 A 发生的次数， p 是事件 A 在每次试验中发生的概率。则对于任意正数 ε 有：

$$\lim_{n\to\infty}P\{|\frac{n_A}{n}-p|<\varepsilon\}=1$$
$$or:\quad\lim_{n\to\infty}P\{|\frac{n_A}{n}-p|\ge\varepsilon\}=0$$

   即：当独立重复实验执行非常大的次数时，事件 A 发生的频率逼近于它的概率。

5. 辛钦定理：设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，服从同一分布，且具有相同的数学期望：$\mathbb{E}[X_k]=\mu$。 则对于任意正数 ε 有：

$$\lim_{n\to\infty}P\{|\frac{1}{n}\sum_{k=1}^nX_k-\mu|<\varepsilon\}=1$$

   - 注意：这里并没有要求随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 的方差存在。
   - 伯努利大数定理是亲钦定理的特殊情况。

3.3 中心极限定理

1. 独立同分布的中心极限定理：设随机变量 $X_1,X_2,\cdots,X_n$ 独立同分布，且具有数学期望和方差：$\mathbb{E}[X_k]=\mu,Var[X_k]=\sigma^2$ ， 则随机变量之和 $\overline{SX_n}=\sum_{k=1}^nX_k$ 的标准变化量：

$$Y_n=\frac{\overline{SX_n}-\mathbb{E}[\overline{SX_n}]}{\sqrt{Var[\overline{SX_n}]}}=\frac{\overline{SX_n}-n\mu}{\sqrt{n}\sigma}$$

   的概率分布函数 $F_n(x)$ 对于任意 x 满足：

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\{Y_n\le x\}=\lim_{n\to\infty}P\{\frac{\sum_{k=1}^nX_k-n\mu}{\sqrt{n}\sigma}\le x\}$$
$$=\int_{-\infty}^x\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x)$$

   - 其物理意义为：均值方差为 μ,σ^2 的独立同分布的随机变量 $X_1,X_2,\cdots,X_n$ 之和 $\overline{SX_n}=\sum_{k=1}^nX_k$ 的标准变化量 Y_n，当 n 充分大时，其分布近似于标准正态分布。

     即：$\overline{SX_n}=\sum_{k=1}^nX_k$ 在 n 充分大时，其分布近似于 $N(n\mu,n\sigma^2)$ 。
   - 一般情况下，很难求出 n 个随机变量之和的分布函数。因此当 n 充分大时，可以通过正态分布来做理论上的分析或者计算。

2. `Liapunov` 定理：设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，具有数学期望和方差：$\mathbb{E}[X_k] = \mu_k, Var[X_k] = \sigma_k^2$ 。记：$B_n^2 = \sum_{k=1}^n \sigma_k^2$。若存在正数 δ，使得当 $n \to \infty$ 时，$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}[|X_k - \mu_k|^{2+\delta}] \to 0$ 。则随机变量之和 $\overline{SX_n} = \sum_{k=1}^n X_k$ 的标准变化量:

$$Z_n = \frac{\overline{SX_n} - \mathbb{E}[\overline{SX_n}]}{\sqrt{Var[\overline{SX_n}]}} = \frac{\overline{SX_n} - \sum_{k=1}^n \mu_k}{B_n}$$

的概率分布函数 $F_n(x)$ 对于任意 x 满足：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\{Z_n \le x\} = \lim_{n\to\infty} P\{\frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n} \le x\}$$
$$= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x)$$

   - 其物理意义为：相互独立的随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 之和 $\overline{SX_n} = \sum_{k=1}^n X_k$的衍生随机变量序列 $Z_n = \frac{\overline{SX_n} - \sum_{k=1}^n \mu_k}{B_n}$ ，当 n 充分大时，其分布近似与标准正态分布。
   - 这里并不要求 $X_1, X_2, \cdots, X_n, \cdots$ 同分布。

9. `Demoiver-Laplace` 定理：设随机变量序列 $\eta_n, n = 1, 2, \ldots$ 服从参数为 (n, p) 的二项分布，其中 $0 < p < 1$ 。则对于任意 x，有：

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2|2} dt = \Phi(x)$$

   该定理表明，正态分布是二项分布的极限分布。当 n 充分大时，可以利用正态分布来计算二项分布的概率。

五、常见概率分布

5.1 均匀分布

1. 离散随机变量的均匀分布：假设 X 有 k 个取值： $x_1, x_2, \cdots, x_k$，则均匀分布的概率密度函数(`probability mass function:PMF`)为：

$$p(X = x_i) = \frac{1}{k}, \quad i = 1, 2, \cdots, k$$

2. 连续随机变量的均匀分布： 假设 X 在 `[a,b]` 上均匀分布，则其概率密度函数(`probability density function：PDF`)为：

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

.

5.2 伯努利分布

1. 伯努利分布：参数为 $\phi \in [0, 1]$。随机变量 $X \in \{0, 1\}$ 。
   - 概率分布函数为：$p(X = x) = \phi^x (1 - \phi)^{1-x}$ ，$x \in \{0, 1\}$ 。
   - 期望： $\mathbb{E}[X] = \phi$ 。方差： $Var[X] = \phi(1 - \phi)$ 。

2. `categorical` 分布：它是二项分布的推广，也称作 `multinoulli` 分布。假设随机变量 $X \in \{1,2,\cdots,K\}$ ，其概率分布函数为：

$$\begin{aligned} p(X=1)&=\theta_1\\ p(X=2)&=\theta_2\\ &\vdots\\ p(X=K-1)&=\theta_{K-1}\\ p(X=K)&=1-\sum_{i=1}^{K-1}\theta_i \end{aligned}$$

其中 θ_i 为参数，它满足 $\theta_i \in [0,1]$ ，且 $\sum_{i=1}^{K-1}\theta_i \in [0,1]$ 。

5.3 二项分布

1. 假设试验只有两种结果：成功的概率为 ϕ ，失败的概率为 $1-\phi$。 则二项分布描述了：独立重复地进行 n 次试验中，成功 x 次的概率。
   - 概率质量函数：

$$p(X=x)=\frac{n!}{x!(n-x)!}\phi^x(1-\phi)^{n-x}, x\in\{0,1,\cdots,n\}$$

   - 期望： $\mathbb{E}[X]=n\phi$。 方差： $Var[X]=n\phi(1-\phi)$ 。

5.4 高斯分布

1. 正态分布是很多应用中的合理选择。如果某个随机变量取值范围是实数，且对它的概率分布一无所知，通常会假设它服从正态分布。有两个原因支持这一选择：
   - 建模的任务的真实分布通常都确实接近正态分布。中心极限定理表明，多个独立随机变量的和近似正态分布。
   - 在具有相同方差的所有可能的概率分布中，正态分布的熵最大（即不确定性最大）。

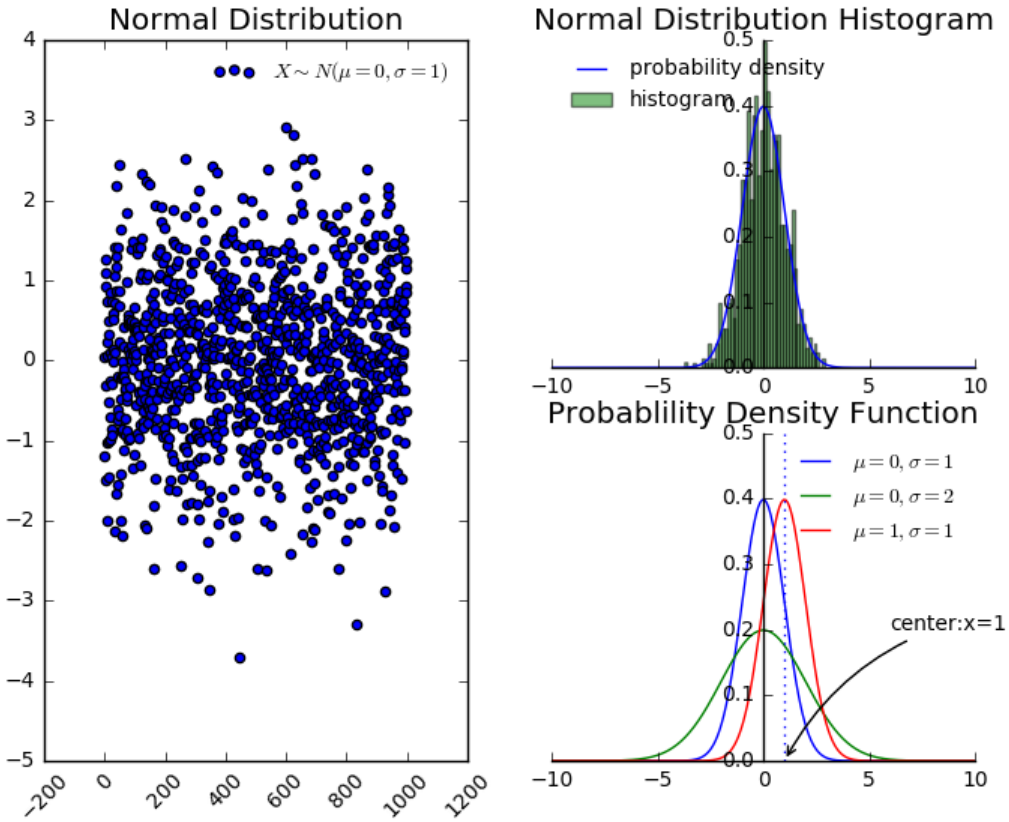
5.4.1 一维正态分布

1. 正态分布的概率密度函数为：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty<x<\infty$$

其中 $\mu,\sigma(\sigma>0)$ 为常数。

   - 若随机变量 X 的概率密度函数如上所述，则称 X 服从参数为 μ,σ 的正态分布或者高斯分布，记作 $X\sim N(\mu,\sigma^2)$。
   - 特别的，当 $\mu=0,\sigma=1$ 时，称为标准正态分布，其概率密度函数记作 $\varphi(x)$，分布函数记作 $\Phi(x)$ 。
   - 为了计算方便，有时也记作：$\mathcal{N}(x;\mu,\beta^{-1})=\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$ ，其中 $\beta\in(0,\infty)$ 。
2. 正态分布的概率密度函数性质：
   - 曲线关于 $x=\mu$ 对称。
   - 曲线在 $x=\mu$ 时取最大值。
   - 曲线在 $x=\mu\pm\sigma$ 处有拐点。
   - 参数 μ 决定曲线的位置； σ 决定图形的胖瘦。



3. 若 $X\sim N(\mu,\sigma^2)$ 则：
   - $\frac{X-\mu}{\sigma}\sim N(0,1)$
   - 期望：$\mathbb{E}[X]=\mu$。方差：$Var[X]=\sigma^2$。
4. 有限个相互独立的正态随机变量的线性组合仍然服从正态分布：若随机变量 $X_i\sim N(\mu_i,\sigma_i^2),i=1,2,\cdots,n$ 且它们相互独立，则它们的线性组合：$C_1X_1+C_2X_2+\cdots+C_nX_n$ 仍然服从正态分布（其中$C_1,C_2,\cdots,C_n$ 不全是为 0 的常数），且：$C_1X_1+C_2X_2+\cdots+C_nX_n\sim N(\sum_{i=1}^n C_i\mu_i,\sum_{i=1}^n C_i^2\sigma_i^2)$。

5.4.2 多维正态分布

1. 二维正态随机变量 (X,Y) 的概率密度为：

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{\frac{-1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]\}$$

   根据定义，可以计算出:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)^2/(2\sigma_1^2)}, -\infty < x < \infty$$
$$p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(y-\mu_2)^2/(2\sigma_2^2)}, -\infty < y < \infty$$
$$\mathbb{E}[X] = \mu_1$$
$$\mathbb{E}[Y] = \mu_2$$
$$Var[X] = \sigma_1^2$$
$$Var[Y] = \sigma_2^2$$
$$Cov[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)p(x,y)dxdy = \rho\sigma_1\sigma_2$$
$$\rho_{XY} = \rho$$

2. 引入矩阵：

$$\vec{\mathbf{x}} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$\mathbf{\Sigma}$ 为 (X,Y) 的协方差矩阵。其行列式为 $\det\mathbf{\Sigma} = \sigma_1^2\sigma_2^2(1-\rho^2)$，其逆矩阵为：

$$\mathbf{\Sigma}^{-1} = \frac{1}{\det\mathbf{\Sigma}} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

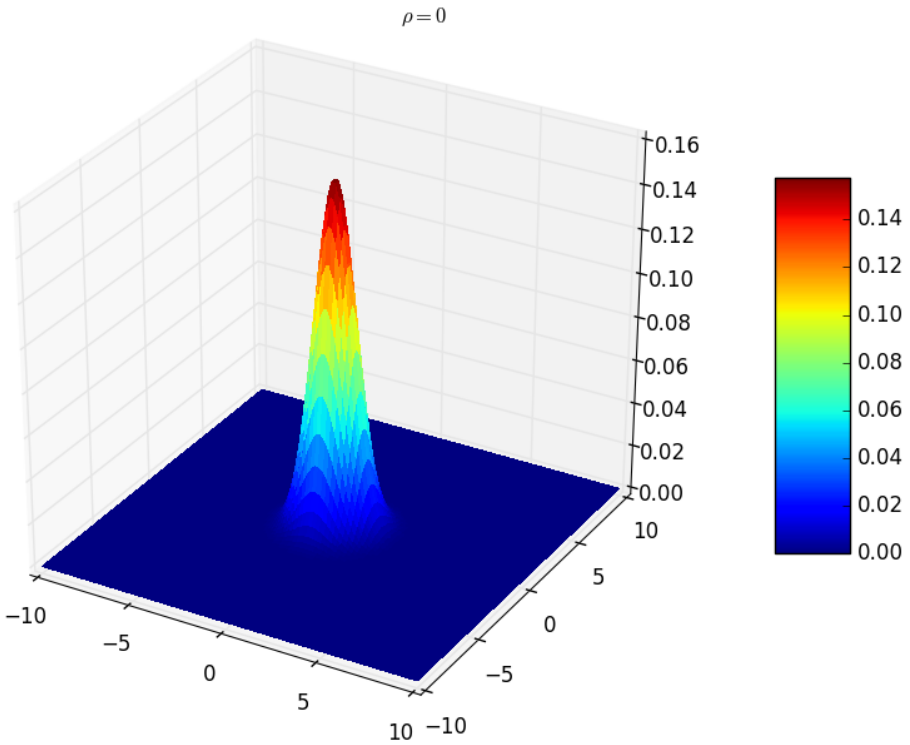
于是 (X,Y) 的概率密度函数可以写作 $(\vec{\mathbf{x}}-\vec{\mu})^T$ 表示矩阵的转置：

$$p(x,x) = \frac{1}{(2\pi)(\det\mathbf{\Sigma})^{1/2}} \exp\{-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\vec{\mathbf{x}}-\vec{\mu})\}$$

其中：

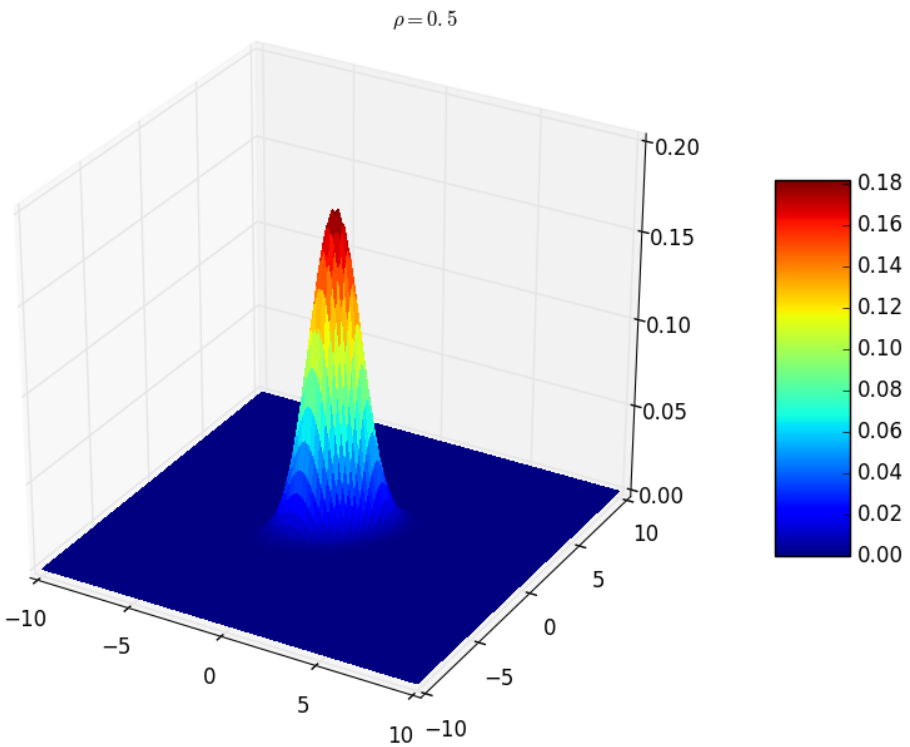
- 均值 μ_1, μ_2 决定了曲面的位置（本例中均值都为0）。
- 标准差 σ_1, σ_2 决定了曲面的陡峭程度（本例中方差都为1）。
- ρ 决定了协方差矩阵的形状，从而决定了曲面的形状。
  - $\rho = 0$ 时，协方差矩阵对角线非零，其他位置均为零。此时表示随机变量之间不相关。

    此时的联合分布概率函数形状如下图所示，曲面在 $z=0$ 平面的截面是个圆形：



- $\rho = 0.5$ 时，协方差矩阵对角线非零，其他位置非零。此时表示随机变量之间相关。

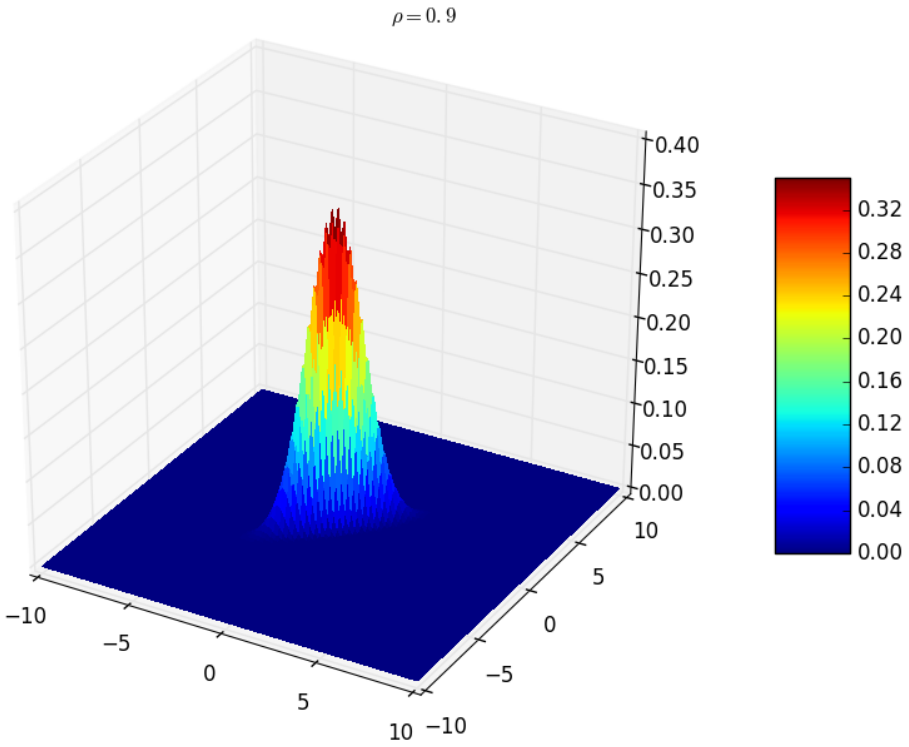
  此时的联合分布概率函数形状如下图所示，曲面在 $z = 0$ 平面的截面是个椭圆，相当于圆形沿着直线 $y = x$ 方向压缩 ：



- $\rho = 1$ 时，协方差矩阵对角线非零，其他位置非零。

  此时表示随机变量之间完全相关。此时的联合分布概率函数形状为：曲面在 $z = 0$ 平面的截面是直线 $y = x$ ，相当于圆形沿着直线 $y = x$ 方向压缩成一条直线 。

  由于 $\rho = 1$会导致除数为 0，因此这里给出 $\rho = 0.9$：



3. 多维正态随机变量 $(X_1, X_2, \cdots, X_n)$ ，引入列矩阵：

$$\vec{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

$\mathbf{\Sigma}$ 为 $(X_1, X_2, \cdots, X_n)$ 的协方差矩阵。则：

$$p(x_1, x_2, x_3, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{1/2}} \exp\{-\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu})^T \mathbf{\Sigma}^{-1}(\vec{\mathbf{x}} - \vec{\mu})\}$$

记做 ：$\mathcal{N}(\vec{\mathbf{x}}; \vec{\mu}, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu})^T \mathbf{\Sigma}^{-1}(\vec{\mathbf{x}} - \vec{\mu})\right)$ 。

4. n 维正态变量具有下列四条性质：

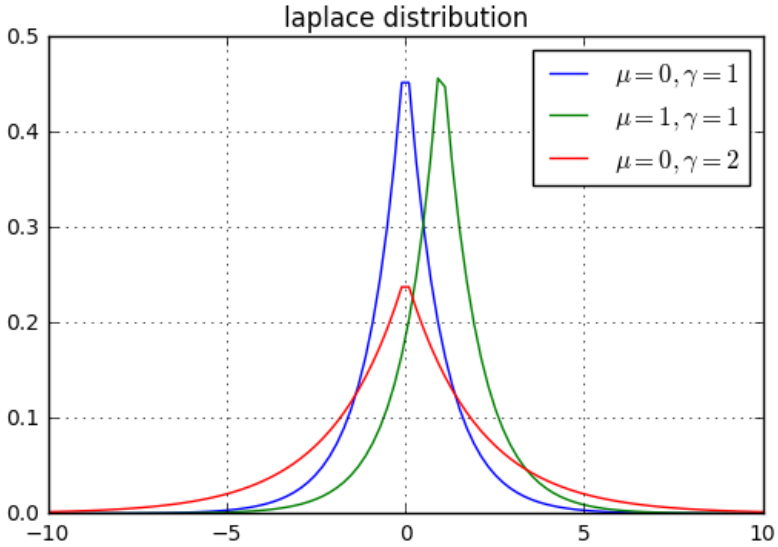
   - n 维正态变量的每一个分量都是正态变量；反之，若 $X_1, X_2, \cdots, X_n$ 都是正态变量，且相互独立，则 $(X_1, X_2, \cdots, X_n)$ 是 n 维正态变量。
   - n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布的充要条件是： $X_1, X_2, \cdots, X_n$ 的任意线性组合： $l_1X_1 + l_2X_2 + \cdots + l_nX_n$ 服从一维正态分布，其中 $l_1, l_2, \cdots, l_n$ 不全为 0 。
   - 若 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布，设 $Y_1, Y_2, \cdots, Y_k$ 是 $X_j, j = 1, 2, \cdots, n$ 的线性函数，则 $(Y_1, Y_2, \cdots, Y_k)$ 也服从多维正态分布。

     这一性质称为正态变量的线性变换不变性。
   - 设 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布，则 $X_1, X_2, \cdots, X_n$ 相互独立 $\Longleftrightarrow$ $X_1, X_2, \cdots, X_n$ 两两不相关。

5.5 拉普拉斯分布

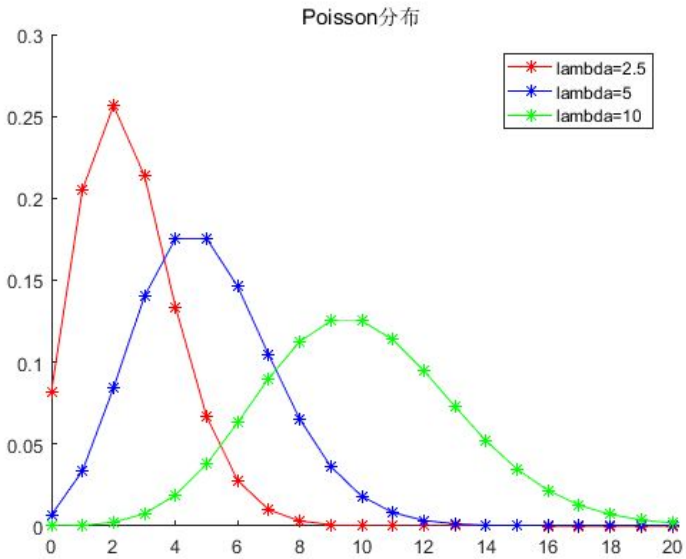
1. 拉普拉斯分布：

- 概率密度函数：$p(x;\mu,\gamma)=\frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)$ 。
- 期望： $\mathbb{E}[X]=\mu$ 。方差： $Var[X]=2\gamma^2$ 。



5.6 泊松分布

1. 假设已知事件在单位时间（或者单位面积）内发生的**平均**次数为 λ ，则泊松分布描述了：事件在单位时间（或者单位面积）内发生的具体次数为 k 的概率。
   - 概率质量函数：$p(X=k;\lambda)=\frac{e^{-\lambda}\lambda^k}{k!}$ 。
   - 期望： $\mathbb{E}[X]=\lambda$。 方差： $Var[X]=\lambda$ 。



2. 用均匀分布模拟泊松分布：

```
def make_poisson(lmd,tm):
    '''
    用均匀分布模拟泊松分布。 lmd为 lambda 参数； tm 为时间
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    count,tm_edges=np.histogram(t,bins=tm,range=(0,tm))#获取每个单位时间内，事件发生的次数
    max_k= lmd *2 # 要统计的最大次数
    dist,count_edges=np.histogram(count,bins=max_k,range=(0,max_k),density=True)
    x=count_edges[:-1]
    return x,dist,stats.poisson.pmf(x,lmd)
```

该函数：

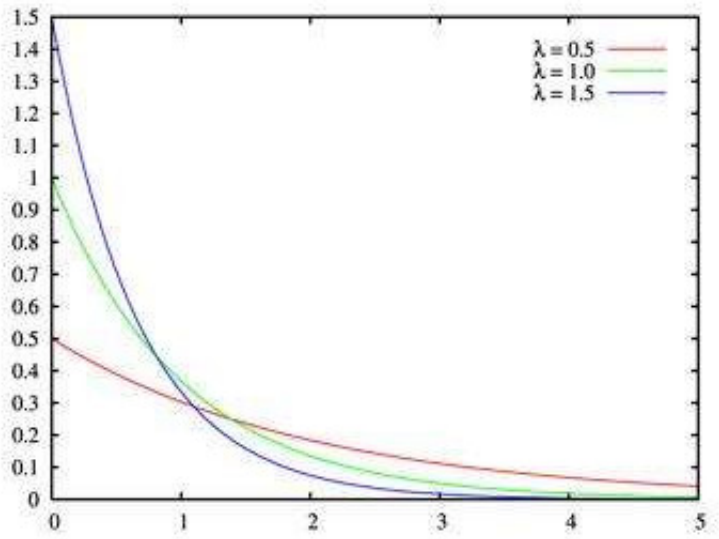
- 首先随机性给出了 `lmd*tm` 个事件发生的时间（时间位于区间 `[0,tm]` ）内。
- 然后统计每个单位时间区间内，事件发生的次数。
- 然后统计这些次数出现的频率。
- 最后将这个频率与理论上的泊松分布的概率质量函数比较。

5.7 指数分布

1. 若事件服从泊松分布，则该事件前后两次发生的时间间隔服从指数分布。由于时间间隔是个浮点数，因此指数分布是连续分布。
   - 概率密度函数：（ t 为时间间隔 ）

$$p(t;\lambda)=\begin{cases}0, & t<0\\ \frac{\lambda}{\exp(\lambda t)}, & t\geq 0\end{cases}$$

   - 期望： $\mathbb{E}[t]=\frac{1}{\lambda}$。方差： $Var[t]=\frac{1}{\lambda^2}$ 。



2. 用均匀分布模拟指数分布：

```
def make_expon(lmd,tm):
    '''
    用均匀分布模拟指数分布。 lmd为 lambda 参数； tm 为时间
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    sorted_t=np.sort(t) #时刻升序排列
    delt_t=sorted_t[1:]-sorted_t[:-1] #间隔序列
    dist,edges=np.histogram(delt_t,bins="auto",density=True)
    x=edges[:-1]
    return x,dist,stats.expon.pdf(x,loc=0,scale=1/lmd) #scale 为 1/lambda
```

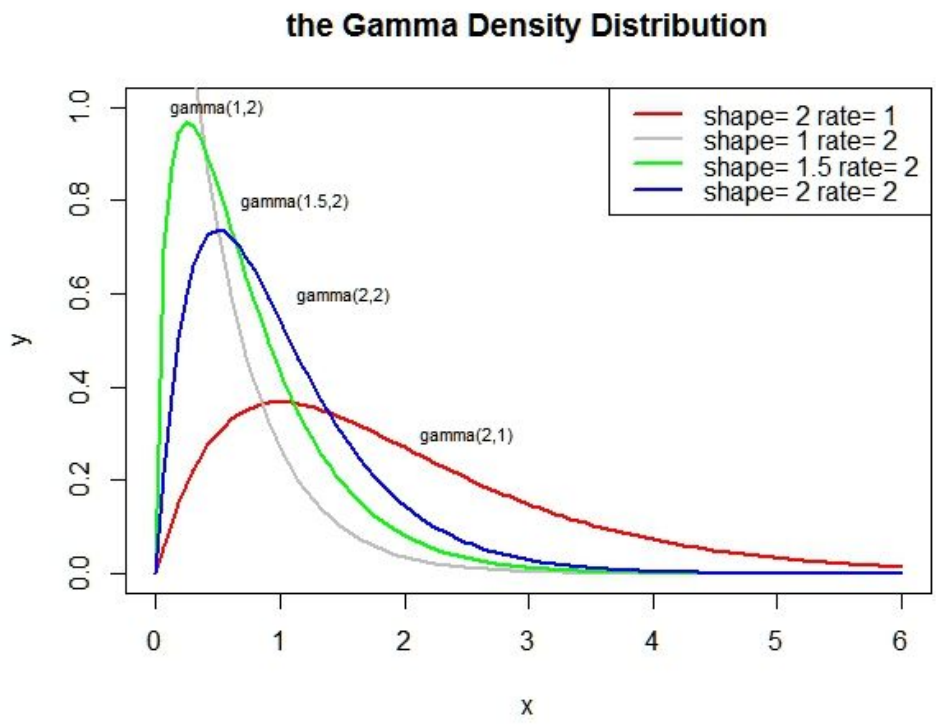
5.8 伽马分布

1. 若事件服从泊松分布，则事件第 i 次发生和第 $i+k$ 次发生的时间间隔为伽玛分布。由于时间间隔是个浮点数，因此指数分布是连续分布。
   - 概率密度函数：$p(t;\lambda,k)=\frac{t^{(k-1)}\lambda^k e^{(-\lambda t)}}{\Gamma(k)}$ ， t 为时间间隔 。
   - 期望： $\mathbb{E}[t]=\frac{k}{\lambda}$ 。方差：$Var[t]=\frac{k}{\lambda^2}$ 。
2. 上面的定义中 k 必须是整数。事实上，若随机变量 X 服从伽马分布，则其概率密度函数为：

$$p(X;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}X^{\alpha-1}e^{-\beta X},\quad X>0$$

记做 $\Gamma(\alpha,\beta)$ 。其中 α 称作形状参数，β 称作尺度参数。

- 期望 $\mathbb{E}[X]=\frac{\alpha}{\beta}$ ，方差$Var[X]=\frac{\alpha}{\beta^2}$ 。
- 当 $\alpha\leq 1$ 时，$p(X;\alpha,\beta)$ 为递减函数。
- 当 $\alpha>1$ 时，$p(X;\alpha,\beta)$ 为单峰函数。



3. 性质：
   - 当 $\beta=n$ 时， 为 `Erlang` 分布。
   - 当 $\alpha=1,\beta=\lambda$ 时，就是参数为 λ 的指数分布。
   - 当 $\alpha=\frac{n}{2},\beta=\frac{1}{2}$ 时，就是常用的卡方分布。
4. 伽马分布的可加性：设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立并且都服从伽马分布： $X_i\sim\Gamma(\alpha_i,\beta)$，则：

$$X_1+X_2+\cdots+X_n\sim\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_n,\beta)$$

5. 用均匀分布模拟伽玛分布：

```
def make_gamma(lmd,tm,k):
    '''
    用均匀分布模拟伽玛分布。 lmd为 lambda 参数； tm 为时间；k 为 k 参数
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    sorted_t=np.sort(t) #时刻升序排列
    delt_t=sorted_t[k:]-sorted_t[:-k] #间隔序列
    dist,edges=np.histogram(delt_t,bins="auto",density=True)
    x=edges[:-1]
    return x,dist,stats.gamma.pdf(x,loc=0,scale=1/lmd,a=k) #scale 为 1/lambda,a 为 k
```

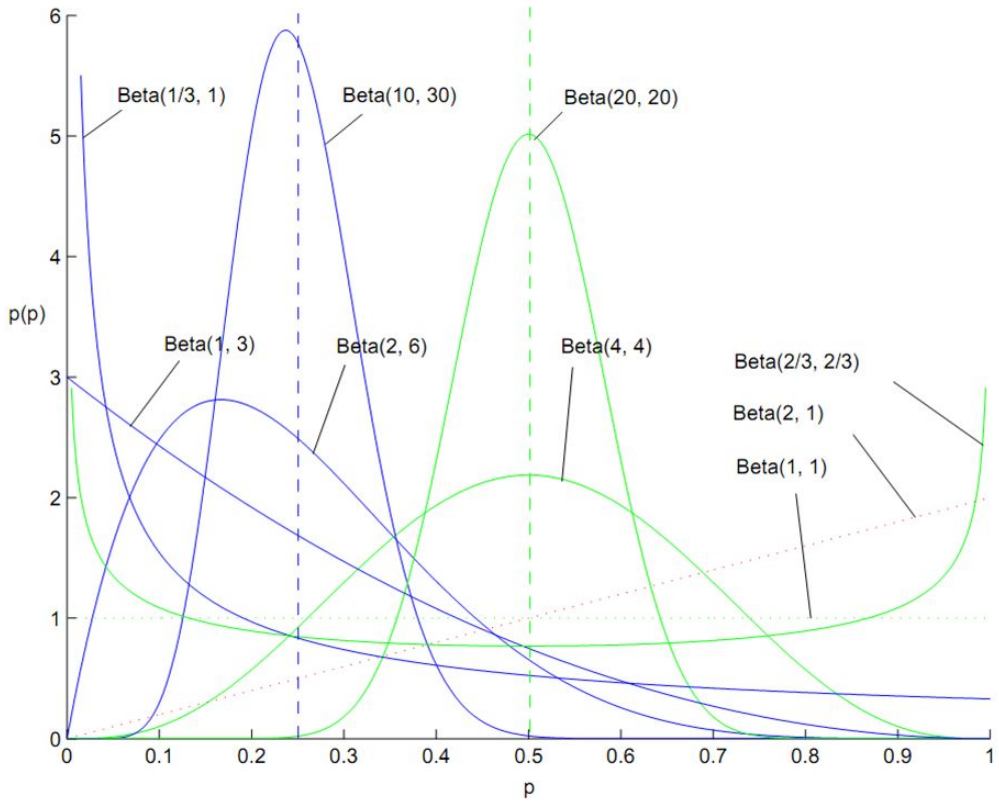
5.9 贝塔分布

1. 贝塔分布是定义在 $(0,1)$ 之间的连续概率分布。

   如果随机变量 X 服从贝塔分布，则其概率密度函数为：

$$p(X,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}X^{\alpha-1}(1-X)^{\beta-1}=\frac{1}{B(\alpha,\beta)}X^{\alpha-1}(1-X)^{\beta-1},\quad 0\leq X<1$$

   记做 $Beta(\alpha,\beta)$ 。
   - 众数为： $\frac{\alpha-1}{\alpha+\beta-2}$ 。
   - 期望为： $\mathbb{E}[X]=\frac{\alpha}{\alpha+\beta}$ ，方差为： $Var[X]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$。



5.10 狄拉克分布

1. 狄拉克分布：假设所有的概率都集中在一点 μ 上，则对应的概率密度函数为：$p(x)=\delta(x-\mu)$ 。

   其中 $\delta(\cdot)$ 为狄拉克函数，其性质为：

$$\delta(x)=0,\forall x\neq 0$$
$$\int_{-\infty}^{\infty}\delta(x)dx=1$$

2. 狄拉克分布的一个典型用途就是定义连续型随机变量的经验分布函数。假设数据集中有样本 $\vec{\mathbf{x}}_1,\vec{\mathbf{x}}_2,\cdots,\vec{\mathbf{x}}_N$ ，则定义经验分布函数：

$$\hat{p}(\vec{\mathbf{x}})=\frac{1}{N}\sum_{i=1}^{N}\delta(\vec{\mathbf{x}}-\vec{\mathbf{x}}_i)$$

   它就是对每个样本赋予了一个概率质量 $\frac{1}{N}$。

3. 对于离散型随机变量的经验分布，则经验分布函数就是`multinoulli`分布，它简单地等于训练集中的经验频率。

4. 经验分布的两个作用：
   - 通过查看训练集样本的经验分布，从而指定该训练集的样本采样的分布（保证采样之后的分布不失真）。
   - 经验分布就是使得训练数据的可能性最大化的概率密度函数。

5.11 多项式分布与狄里克雷分布

1. 多项式分布的质量密度函数：

$$Mult(m_1,m_2,\cdots,m_K;\vec{\mu},N)=\frac{N!}{m_1!m_2!\cdots m_K!}\prod_{k=1}^{K}\mu_k^{m_k}$$

它是 $(\mu_1+\mu_2+\cdots+\mu_K)^{m_1+m_2+\cdots+m_K}$ 的多项式展开的形式。

2. 狄利克雷分布的概率密度函数：

$$Dir(\vec{\mu};\vec{\alpha})=\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\sum_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

3. 可以看到，多项式分布与狄里克雷分布的概率密度函数非常相似，区别仅仅在于前面的归一化项：
   - 多项式分布是针对离散型随机变量，通过求和获取概率。
   - 狄里克雷分布时针对连续型随机变量，通过求积分来获取概率。

5.12 混合概率分布

1. 混合概率分布：它组合了其他几个分量的分布来组成。
   - 在每次生成样本中，首先通过`multinoulli`分布来决定选用哪个分量，然后由该分量的分布函数来生成样本。
   - 其概率分布函数为：

$$p(x)=\sum_i P(c=i)p(x\mid c=i)$$

     其中 $p(c=i)$ 为一个`multinoulli`分布，c 的取值范围就是各分量的编号。
2. 前面介绍的连续型随机变量的经验分布函数就是一个混合概率分布的例子，此时 $p(c=i)=\frac{1}{N}$ 。
3. 混合概率分布可以通过简单的概率分布创建更复杂的概率分布。一个常见的例子是混合高斯模型，其中 $p(x\mid c=i)$ 为高斯模型。每个分量都有对应的参数 $(\vec{\mu}_i,\mathbf{\Sigma}_i)$ 。
   - 有些混合高斯模型有更强的约束，如 $\forall i,\mathbf{\Sigma}_i=\mathbf{\Sigma}$，更进一步还可以要求 $\mathbf{\Sigma}$ 为一个对角矩阵。
   - 混合高斯模型是一个通用的概率密度函数逼近工具。任何平滑的概率密度函数都可以通过足够多分量的混合高斯模型来逼近。

六、先验分布与后验分布

1. 在贝叶斯学派中，`先验分布+数据（似然）= 后验分布` 。
2. 例如：假设需要识别一大箱苹果中的好苹果、坏苹果的概率。
   - 根据你对苹果好、坏的认知，给出先验分布为：50个好苹果和50个坏苹果。
   - 现在你拿出10个苹果，发现有：8个好苹果，2个坏苹果。

     根据数据，你得到后验分布为：58个好苹果，52个坏苹果
   - 再拿出10个苹果，发现有：9个好苹果，1个坏苹果。

     根据数据，你得到后验分布为：67个好苹果，53个坏苹果
   - 这样不断重复下去，不断更新后验分布。当一箱苹果清点完毕，则得到了最终的后验分布。

   在这里：
   - 如果不使用先验分布，仅仅清点这箱苹果中的好坏，则得到的分布只能代表这一箱苹果。
   - 采用了先验分布之后得到的分布，可以认为是所有箱子里的苹果的分布。

- 当采用先验分布时：给出的好、坏苹果的个数（也就是频数）越大，则先验分布越占主导地位。

3. 假设好苹果的概率为 p ，则抽取 N 个苹果中，好苹果个数为 k 个的概率为一个二项分布：

$$Binom(k \mid p;N) = C_N^k p^k (1-p)^{N-k}$$

其中 C_N^k 为组合数。

4. 现在的问题是：好苹果的概率 p 不再固定，而是服从一个分布。

假设好苹果的概率 p 的先验分布为贝塔分布：$Beta(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$ 。

则后验概率为：

$$P(p \mid k;N,\alpha,\beta) = \frac{P(k \mid p;N) \times P(p;\alpha,\beta)}{P(k;N,\alpha,\beta)}$$

$$\propto P(k \mid p;N) \times P(p;\alpha,\beta) = C_N^k p^k (1-p)^{N-k} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto p^{k+\alpha-1}(1-p)^{N-k+\beta-1}$$

归一化之后，得到后验概率为：

$$P(p \mid k;N,\alpha,\beta) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+k)\Gamma(\beta+N-k)} p^{k+\alpha-1}(1-p)^{N-k+\beta-1}$$

5. 好苹果概率 p 的先验分布的期望为：$\mathbb{E}[p] = \frac{\alpha}{\alpha+\beta}$ 。好苹果概率 p 的后验分布的期望为：$\mathbb{E}[p \mid k] = \frac{\alpha+k}{\alpha+\beta+N}$ 。
   - 根据上述例子所述：
     - 好苹果的先验概率的期望为 $\frac{50}{50+50} = \frac{1}{2}$
     - 进行第一轮数据校验之后，好苹果的后验概率的期望为 $\frac{50+8}{50+50+10} = \frac{58}{110}$
   - 如果将 α 视为先验的好苹果数量， β 视为先验的坏苹果数量， N 表示箱子中苹果的数量， k 表示箱子中的好苹果数量（相应的， $N-k$ 就是箱子中坏苹果的数量）。则：好苹果的先验概率分布的期望、后验概率分布的期望符合人们的生活经验。
   - 这里使用先验分布和后验分布的期望，因为 p 是一个随机变量。若想通过一个数值来刻画好苹果的可能性，则用期望较好。

6. 更一般的，如果苹果不仅仅分为好、坏两种，而是分作`尺寸1、尺寸2、...尺寸K`等。则 N 个苹果中，有 m_1 个尺寸1的苹果、 m_2 个尺寸2的苹果....m_K 个尺寸 K 的苹果的概率服从多项式分布：

$$Mult(m_1,m_2,\cdots,m_K;\vec{\mu},N) = \frac{N!}{m_1!m_2!\cdots m_K!} \prod_{k=1}^{K} \mu_k^{m_k}$$

其中苹果为尺寸1的概率为 μ_1 ， 尺寸2的概率为 μ_2 ，... 尺寸 K 的概率为 μ_K ， $N = \sum_{k=1}^{K} m_k$

   - 假设苹果尺寸的先验概率分布为狄利克雷分布：$Dir(\vec{\mu};\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\sum_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1}$ 。

     苹果尺寸的先验概率分布的期望为：$\mathbb{E}[\vec{\mu}] = \left(\frac{\alpha_1}{\sum_{k=1}^{K}\alpha_k}, \frac{\alpha_2}{\sum_{k=1}^{K}\alpha_k}, \cdots, \frac{\alpha_K}{\sum_{k=1}^{K}\alpha_k}\right)$ 。
   - 则苹果尺寸的后验概率分布也为狄里克雷分布：$Dir(\vec{\mu};\vec{\alpha}+\vec{\mathbf{m}}) = \frac{\Gamma(N+\sum_{k=1}^{K}\alpha_k)}{\sum_{k=1}^{K}\Gamma(\alpha_k+m_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$ 。

     苹果尺寸的后验概率分布的期望为：$\mathbb{E}[\vec{\mu}] = \left(\frac{\alpha_1+m_1}{N+\sum_{k=1}^{K}\alpha_k}, \frac{\alpha_2+m_2}{N+\sum_{k=1}^{K}\alpha_k}, \cdots, \frac{\alpha_K+m_K}{N+\sum_{k=1}^{K}\alpha_k}\right)$ 。

七、信息论

1. 信息论背后的原理是：从不太可能发生的事件中能学到更多的有用信息。
   - 发生可能性较大的事件包含较少的信息。
   - 发生可能性较小的事件包含较多的信息。
   - 独立事件包含额外的信息 。
2. 对于事件 $X=x$，定义自信息`self-information`为：$I(x)=-\log P(x)$ 。

   自信息仅仅处理单个输出，但是如果计算自信息的期望，它就是熵：

   $$H(X)=\mathbb{E}_{X\sim P(X)}[I(x)]=-\mathbb{E}_{X\sim P(X)}[\log P(x)]$$

   记作 $H(P)$。
   - 熵刻画了按照真实分布 P 来识别一个样本所需要的编码长度的期望（即平均编码长度）。

     如：含有4个字母`(A,B,C,D)`的样本集中，真实分布 $P=(\frac{1}{2},\frac{1}{2},0,0)$，则只需要1位编码即可识别样本。
   - 对于离散型随机变量 X，假设其取值集合大小为 K，则可以证明： $0\leq H(X)\leq \log K$ 。
3. 对于随机变量 Y 和 X，条件熵 $H(Y\mid X)$ 表示：已知随机变量 X 的条件下，随机变量 Y 的不确定性。

   它定义为：X 给定条件下 Y 的条件概率分布的熵对 X 的期望：

   $$H(Y\mid X)=\mathbb{E}_{X\sim P(X)}[H(Y\mid X=x)]=-\mathbb{E}_{(X,Y)\sim P(X,Y)}\log P(Y\mid X)$$

   - 对于离散型随机变量，有：

     $$H(Y\mid X)=\sum_x p(x)H(Y\mid X=x)=-\sum_x\sum_y p(x,y)\log p(y\mid x)$$

   - 对于连续型随机变量，有：

     $$H(Y\mid X)=\int p(x)H(Y\mid X=x)dx=-\int\int p(x,y)\log p(y\mid x)dxdy$$

4. 根据定义可以证明：$H(X,Y)=H(Y\mid X)+H(X)$ 。

   即：描述 X 和 Y 所需要的信息是：描述 X 所需要的信息加上给定 X 条件下描述 Y 所需的额外信息。
5. `KL`散度（也称作相对熵）：对于给定的随机变量 X，它的两个概率分布函数 $P(X)$ 和 $Q(X)$ 的区别可以用`KL`散度来度量：

   $$D_{KL}(P||Q)=\mathbb{E}_{X\sim P(X)}\left[\log\frac{P(x)}{Q(x)}\right]=\mathbb{E}_{X\sim P(X)}\left[\log P(x)-\log Q(x)\right]$$

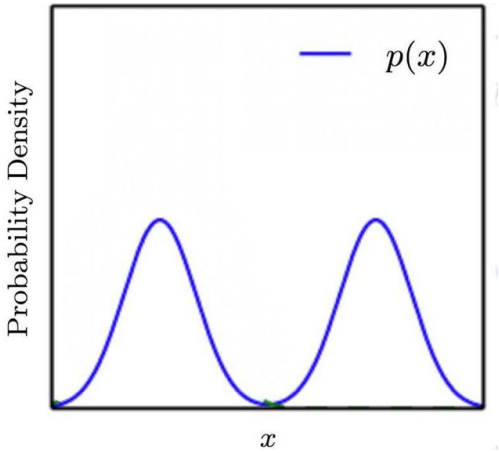
   - `KL`散度非负：当它为 0 时，当且仅当`P`和`Q`是同一个分布（对于离散型随机变量），或者两个分布几乎处处相等（对于连续型随机变量）。
   - `KL`散度不对称：$D_{KL}(P||Q)\neq D_{KL}(Q||P)$ 。

     直观上看对于 $D_{KL}(P||Q)$ ，当 $P(x)$ 较大的地方， $Q(x)$ 也应该较大，这样才能使得 $P(x)\log\frac{P(x)}{Q(x)}$ 较小。

     对于 $P(x)$ 较小的地方， $Q(x)$ 就没有什么限制就能够使得 $P(x)\log\frac{P(x)}{Q(x)}$ 较小。这就是`KL`散度不满足对称性的原因。

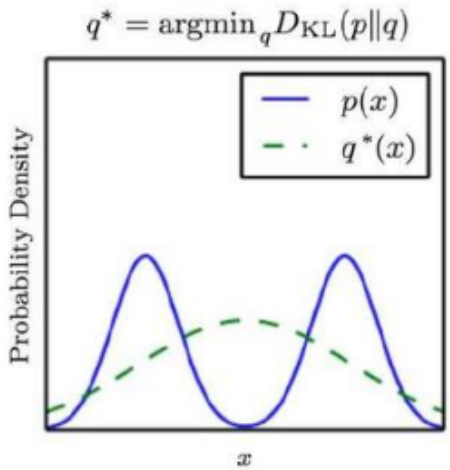
6. 交叉熵`cross-entropy`：$H(P,Q)=H(P)+D_{KL}(P||Q)=-\mathbb{E}_{X\sim P(X)}\log Q(x)$。
   - 交叉熵刻画了使用错误分布Q来表示真实分布 P 中的样本的平均编码长度。
   - $D_{KL(P||Q)}$ 刻画了错误分布Q编码真实分布 P 带来的平均编码长度的增量。
7. 示例：假设真实分布 P 为混合高斯分布，它由两个高斯分布的分量组成。如果希望用普通的高斯分布 Q 来近似 P ，则有两种方案：

$$Q_1^*=\arg\min_Q D_{KL}(P||Q)$$
$$Q_2^*=\arg\min_Q D_{KL}(Q||P)$$



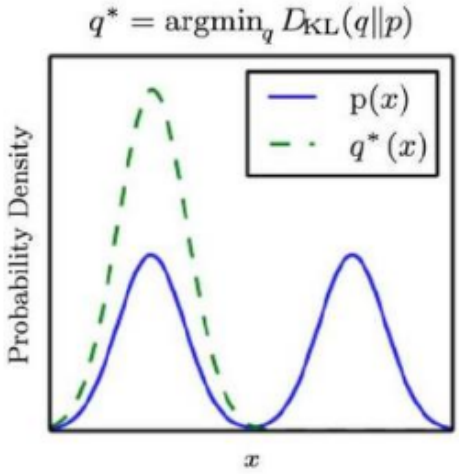
   - 如果选择 Q_1^*，则：
     - 当 $P(x)$ 较大的时候 $Q(x)$ 也必须较大 。如果 $P(x)$ 较大时 $Q(x)$ 较小，则 $P(x)\log\frac{P(x)}{Q(x)}$ 较大。
     - 当 $P(x)$ 较小的时候 $Q(x)$ 可以较大，也可以较小。

     因此 Q_1^* 会贴近 $P(x)$ 的峰值。由于 $P(x)$ 的峰值有两个，因此 Q_1^* 无法偏向任意一个峰值，最终结果就是 Q_1^* 的峰值在 $P(x)$ 的两个峰值之间。



   - 如果选择 Q_2^* ，则：
     - 当 $P(x)$ 较小的时候， $Q(x)$ 必须较小。如果$P(x)$ 较小的时 $Q(x)$ 较大，则 $Q(x)\log\frac{Q(x)}{P(x)}$ 较大。
     - 当 $P(x)$ 较大的时候，$Q(x)$ 可以较大，也可以较小。

     因此 Q_2^* 会贴近 $P(x)$ 的谷值。最终结果就是 Q_2^* 会贴合 $P(x)$ 峰值的任何一个。



- 绝大多数场合使用 $D_{KL}(P||Q)$ ，原因是：当用分布 Q 拟合 P 时我们希望对于常见的事件，二者概率相差不大。

八、其它

1. 假设随机变量 X,Y 满足 $Y=g(X)$，且函数 $g(\cdot)$ 满足：处处连续、可导、且存在反函数。 则有：

$$p_X(x)=p_Y(g(x))\left|\frac{\partial g(x)}{\partial x}\right|$$

或者等价地（其中 $g^{-1}(\cdot)$ 为反函数）：

$$p_Y(y)=p_X(g^{-1}(y))\left|\frac{\partial x}{\partial y}\right|$$

- 如果扩展到高维空间，则有：

$$p_X(\vec{\mathbf{x}})=p_Y(g(\vec{\mathbf{x}}))\left|\det\left(\frac{\partial g(\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}}\right)\right|$$

- 并不是 $p_Y(y)=p_X(g^{-1}(y))$，这是因为 $g(\cdot)$ 引起了空间扭曲，从而导致 $\int p_X(g(x))dx\neq 1$ 。

  根据 $|p_Y(g(x))dy|=|p_X(x)dx|$ ，求解该方程，即得到上述解。

2. 机器学习中不确定性有三个来源：
   - 模型本身固有的随机性。如：量子力学中的粒子动力学方程。
   - 不完全的观测。即使是确定性系统，当无法观测所有驱动变量时，结果也是随机的。
   - 不完全建模。有时必须放弃一些观测信息。

     如机器人建模中：虽然可以精确观察机器人周围每个对象的位置，但在预测这些对象将来的位置时，对空间进行了离散化。则位置预测将带有不确定性。