概率统计上篇

一：事件

1.1 样本空间与事件

1.1.1 随机试验：

在做实验之前，我们知道所有可能出现的结果，但是并不知道每次实验出现的结果是哪一种。例如：

- E1: 抛一枚硬币，分别由H和T表示正面和反面。
- E2: 将一枚硬币连续抛三次，考虑正反面出现的情况
- E3: 掷一颗骰子，可能出现的点数

1.1.2 样本空间：

一个实验所有可能的结果的集合。样本空间中的元素为样本点。

- E1 = {H,T}
- E2 = {HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}
- E3 = {1,2,3,4,5,6}

事件是样本空间的子集，满足某种条件的样本点组成的集合。

样本空间上全空间S，S中的子集叫做事件，即全部样本点中的一部分。

1.1.3 事件关系，频率，概率

事件关系：

- 和事件，并集 $A \cup B$
- 积事件，交集 $A \cap B$,通常记作AB，省略交集的符号
- 对立事件，补集 $\bar{A}$

频率：事件发生的频繁程度，频率稳定与概率

概率：表征事件发生可能性大小

概率的加法公式，对于任意的事件A和B，$P(A \cup B) = P(A) + P(B) - P(AB)$,当A和B的交集为空集时，则最后一项为0。

1.2 贝叶斯公式

1.2.1 条件概率

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

- 条件概率公式表示在A事件发生的条件下，B事件发生的概率。
- 此时的样本空间是A事件的空间。

与前面的概率加法公式类似：

$$P(B1\cup B2|A)=P(B1|A)+P(B2|A)-P(B1B2|A)$$

- 对这个公式的理解方法：当不存在条件概率时，我们是在全空间上进行概率加法运算；而此时，仅仅是样本空间由全空间变成了A事件的空间，对应的概率加法运算只需要指定一下样本空间即可。

由条件概率公式可以得到：

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

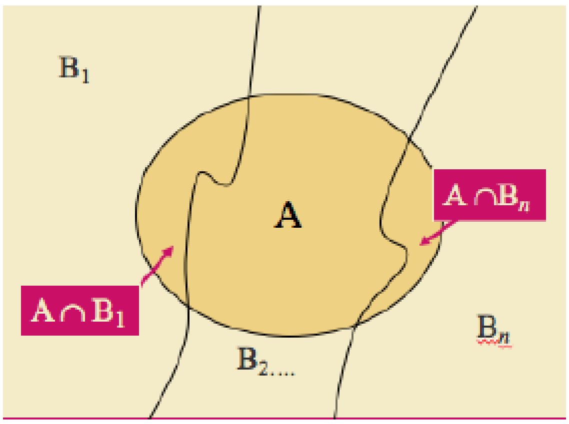
事件独立：

一般情况下 $P(B)\neq P(B|A)$，但是，当A事件和B事件相互独立时， $P(B)=P(B|A)$，因为B事件的发生与A事件发生没有任何关系，A的发生与否对B没有影响。

1.2.2 全概率公式

在条件概率公式中，有时候$P(A)$很难求，这种时候，需要对A的样本空间进行划分。因此得到全概率公式，

$$P(A)=P(A|B1)P(B1)+P(A|B2)P(B2)+P(A|B3)P(B3)+\ldots\ldots$$



1.2.3 贝叶斯公式

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

更一般的情况：

$$P\left(B_i|A\right)=\frac{P\left(A|B_i\right)P\left(B_i\right)}{\sum_{j=1}^{n}P\left(A|B_j\right)P\left(B_j\right)},i=1,2,\cdots,n$$

贝叶斯经典举例：

- 假设吸毒者每次检测呈阳性（+）的概率为 99%．而不吸毒者每次检测呈阴性（-）的概率为 99%．假设某公司对全体雇员进行吸毒检测，已知 0.5% 的雇员吸毒．请问每位检测结果呈阳性的雇员吸毒的概率有多高？
    - $P(D)=0.005$，代表雇员吸毒的概率
    - $P(N)=0.995$，代表雇员不吸毒的概率
    - $P(+|D)=0.99$，代表吸毒者阳性检出率
    - $P(+|N)=0.01$，代表不吸毒者阳性检出率
    - $P(+)$ 代表不考虑其他因素的影响的阳性检出率

$$\begin{aligned}P(D|+)&=\frac{P(+|D)P(D)}{P(+)}\\&=\frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|N)P(N)}\\&=0.3322\end{aligned}$$

1.2.4 朴素贝叶斯

由条件概率公式可以得到如下公式：

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

给定一封邮件，判断这封邮件是否属于垃圾邮件。用D来表示这封邮件，D由N个单词组成。用$h+$来表示垃圾邮件，$h-$来表示正常邮件。

$$P(h+|D)=\frac{P(D|h+)P(h+)}{P(D)}$$
$$P(h-|D)=\frac{P(D|h-)P(h-)}{P(D)}$$

因此，只要比较分子即可。我们假设邮件中每个单词之间没有联系，是相互独立的，因此：

$$P(D|h+) = P(d1|h+)P(d2|h+)P(d3|h+)P(d4|h+)P(d5|h+)\ldots\ldots$$
$$P(D|h-) = P(d1|h-)P(d2|h-)P(d3|h-)P(d4|h-)P(d5|h-)\ldots\ldots$$

$$P(d_1|h+)\times P(d_2|h+)\times P(d_3|h+)\times\cdots = 0.007 * 0.005 * \cdots$$

$$P(d_1|h-)\times P(d_2|h-)\times P(d_3|h-)\times\cdots = 0.00002 * 0.0045 * \cdots$$

	我司	可	办理	正规	发票	......
垃圾	0.0007	0.005	0.002	0.0005	0.006	
正常	0.00002	0.0045	0.0006	0.00008	0.00004	

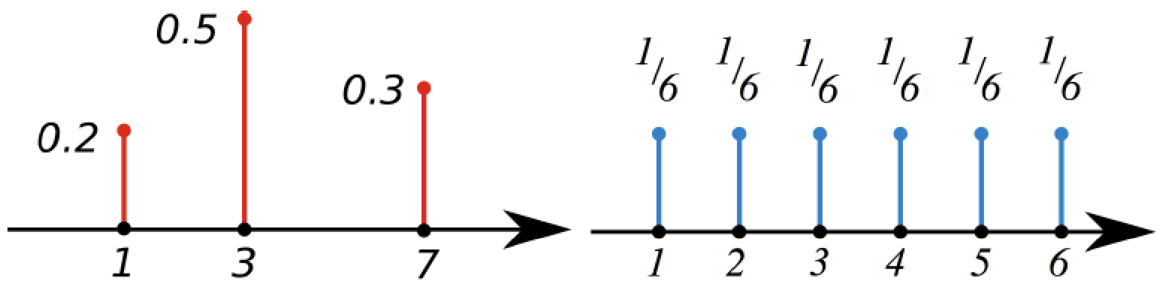
注意：最终的判定规则为：比较$P(h+)\Pi_{i=1}^{n}P(d_i|h+)$和$P(h-)\Pi_{i=1}^{n}P(d_i|h-)$的大小。

二：随机变量

随机变量是随机事件的数量表现。随机变量的取值有离散型和连续型两种。**注意采用对比的方法来理解后续的内容。**

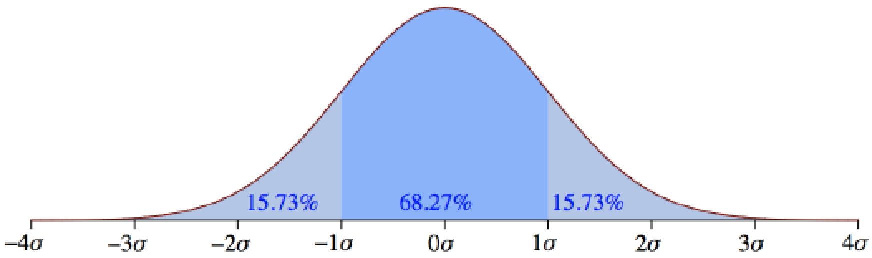
2.1 一维离散随机变量

- 分布率 (Probability mass functions)：$P\{X = x_k\} = p_k, k = 1, 2, \cdots$
  - $p_k \geq 0$
  - $\sum p_k = 1$
- 分布函数 (cumulative distribution function)：$F_X(x) = P\{X \leq x\}$，因此 $P\{x_1 < X \leq x_2\} = P\{X \leq x_2\} - P\{X \leq x_1\} = F_X(x_2) - F_X(x_1)$, 即已知 X 的分布函数，就知道 X 落在任一区间的概率.



2.2 一维连续随机变量

- 概率密度：$f(x) = \frac{d}{dx}F_X(x)$
    - $f_X(x) \geq 0$
    - $\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$
    - $\mathrm{P}(x \leq X \leq x + \Delta x) \approx f_X(x)\,\Delta x$
- $P\{x_1 < X \leq x_2\} = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\, dx$



- 注意：在连续随机变量分布中，某一点的概率为0。

三：随机变量的数字特征

3.1 期望（Expectation)

- 数学期望（离散）：$E(X) = \sum_{k=1}^{\infty} x_k p_k$
- 数学期望（连续）：$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$
- 随机变量的数学期望：$Y = g(X)$
- 离散：$E(Y) = E(g(X)) = \sum_{k=1}^{n} g(x_k) p_{x_k}$
- 连续：$E(Y) = E(g(X)) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$
- $E(a) = a$
- $E(f(x) + g(x)) = E(f(x)) + E(g(x))$
- $E(kX) = kE(X)$
- 当X和Y相互独立时，$E(XY) = E(X)E(Y)$

3.2 方差（Variance)

- 方差（离散）：$D(X) = E\{[X - E(X)]^2\} = \sum_{k=1}^{n} (x_k - E(X))^2 p_{x_k}$
- 方差（连续）：$D(X) = E\{[X - E(X)]^2\} = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x) dx$
- $D(X) = E\{[X - E(X)]^2\} = E(X^2) - E^2(X)$
- $D(X + Y) = D(X) + D(Y)$
- $D(kX) = k^2 D(X)$

- $D(X-Y)=D(X)+D(Y)$

四：人工智能中常见分布

4.1 离散分布

4.1.1 （0-1）分布

在（0-1）分布中，随机变量取1的概率为p，取0的概率为1-p。因此
$P(X=k)=p^k(1-p)^{1-k}, k=0,1$

所以:

$$E(X)=1*p+0*(1-p)=p;$$
$$E(X^2)=1*p+0*(1-p)=p;$$
$$D(X)=p-p^2=p(1-p)$$

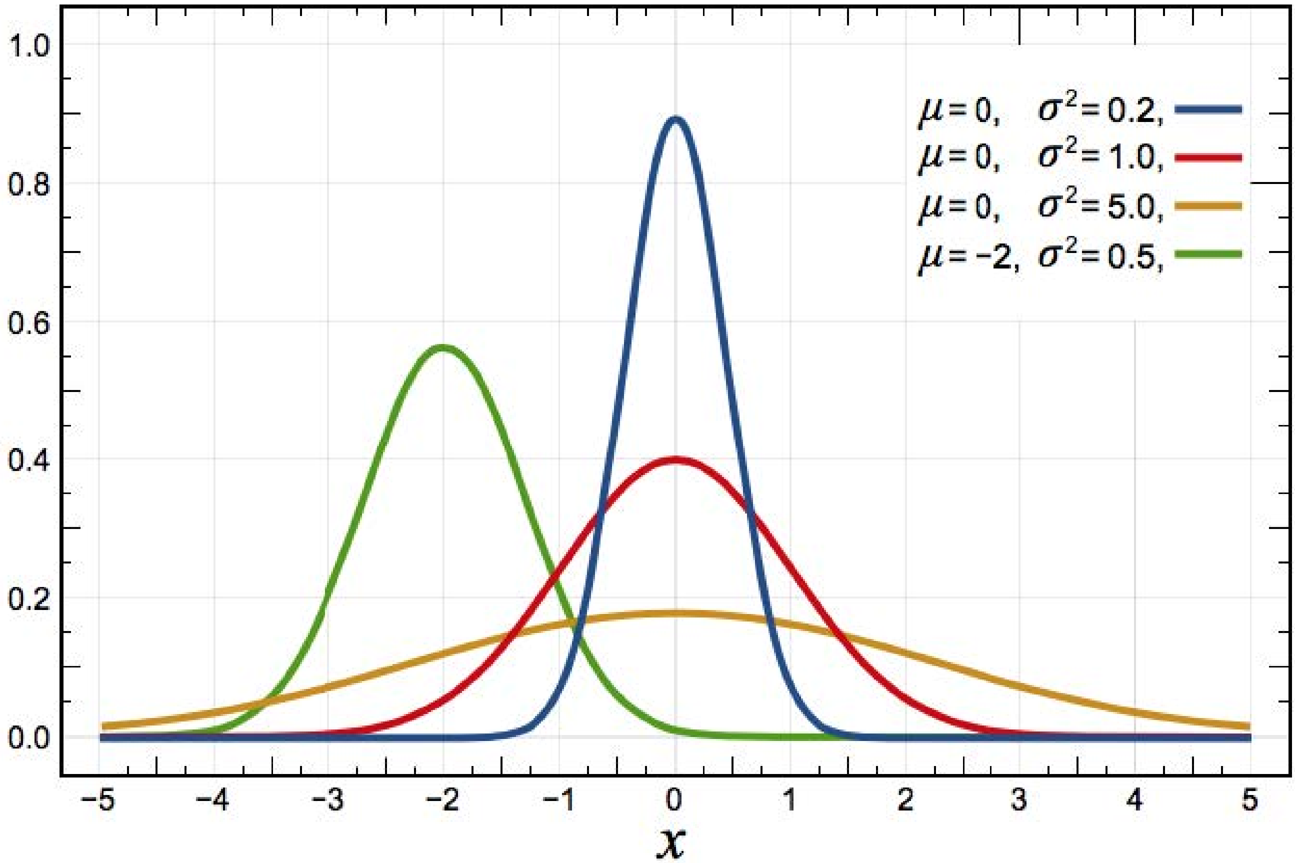
4.1.2 二项分布

将一个试验独立重复进行n次，每次试验中，事件A发生的概率为p，则称这n次实验为n充伯努利实验。以X表示n重伯努利实验中A事件发生的次数，则称X服从参数为(n,p)的二项分布。

$$P(X=k)=C_n^k p^k(1-p)^{n-k}$$
$$E(X)=np;$$
$$D(X)=np(1-p)$$

4.2 高斯分布

$$f\left(x|\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$E(X)=\mu$$
$$D(X)=\sigma^2$$

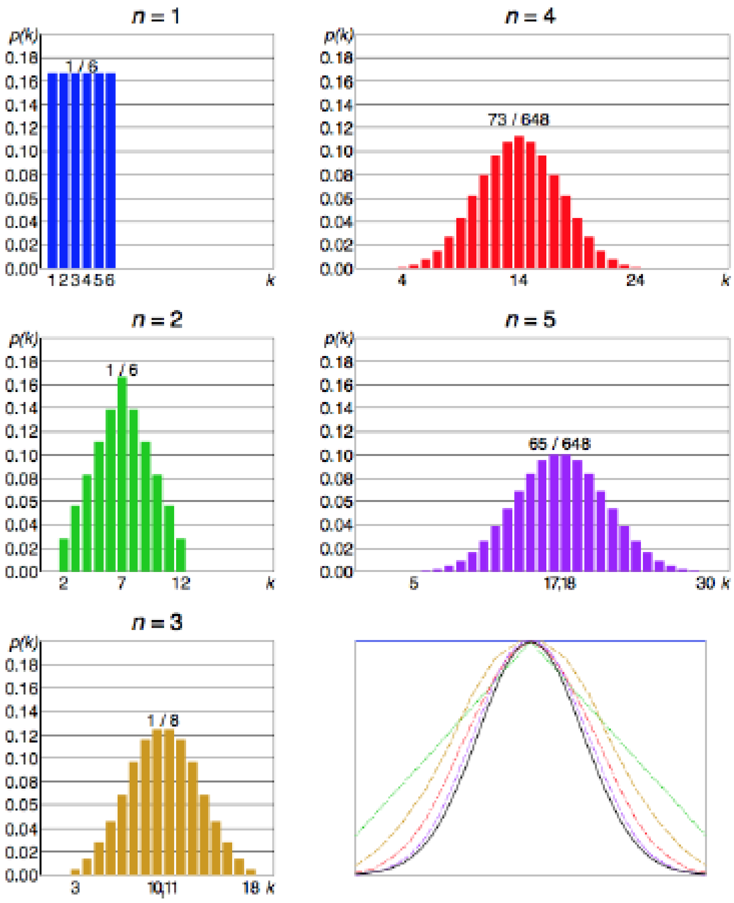


4.3 中心极限定理

设随机变量$X_1,X_2,X_3....X_n$独立同分布，且$E(X_k)=\mu,D(X_k)=\sigma^2$,则当n充分大时，$\bar{X}$（X的均值）近似服从高斯分布：

$$\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k\sim\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$$

以掷骰子为例，展示中心极限定理（变量为点数之和）：



4.4 多维随机变量

以二维随机变量为例：（离散和连续相互对比）

离散：

- 联合分布率：$P(X=x_i,Y=y_j)=p_{ij}$
- 联合分布函数：$F(x,y)=\sum_{x_i\le x,y_j\le y}p_{ij}$
- 边缘分布率：$P(X=x_i)=\sum_{j=1}^{+\infty}p_{ij};P(Y=y_j)=\sum_{i=1}^{+\infty}p_{ij}$
- 条件分布率：$P(X=x_i|Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}$

连续：

- 联合分布密度：$f(x,y)$
- 联合分布函数：$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(x,y)dxdy$
- 边缘概率密度：$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy$
- 条件概率密度：$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$

$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}=E(XY)-E(X)E(Y)$

4.5 多元高斯分布

$$f_{\mathrm{x}}\left(x_1,\ldots,x_k\right)=\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left(-\frac{1}{2}(\mathrm{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathrm{x}-\mu)\right)$$

其中：$\sum$代表协方差，与一元时对照来看，便很容易理解。

$$\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1,X_1] & \cdots & Cov[X_1,X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n,X_1] & \cdots & Cov[X_n,X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2]-E[X_1]E[X_1] & \cdots & E[X_1X_n]-E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1]-E[X_n]E[X_1] & \cdots & E[X_n^2]-E[X_n]E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
&= E[XX^T]-E[X]E[X]^T=\ldots=E[(X-E[X])(X-E[X])^T].
\end{aligned}$$