AI数学基石——概率论篇

目录

01 随机事件和概率

1.1 基础概念

随机试验

试验是指为了察看某事的结果或某物的性能而从事的某种活动. 在概率论与数理统计中，一个试验如果具有以下3个特点：
（1）**可重复性**: 在相同条件下可以重复进行；
（2）**可观察性**: 每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；
（3）**不确定性**: 一次试验之前，不能预知会出现哪一个结果。
就称这样的试验是一个**随机试验**，也简称为**试验**。

样本点和样本空间

每次试验的每一个结果称为**基本事件**，也称作**样本点**，记作$w_1, w_2, \cdots$. 全部样本点的集合称为**样本空间**，记作Ω，则$\Omega = \{w_1, w_2, \cdots\}$。

例子

投掷一颗均匀骰子，观察出现的点数。这是一个随机试验。样本空间$\Omega = \{1，2，3，4，5，6\}$.

随机事件

基本事件是不可再分解的、最基本的事件，其他事件均可由它们复合而成，由基本事件复合而成的事件称为**随机事件**或简称**事件**。常用大写字母A，B，C等表示事件。比如$A = \{出现的点数为偶数\} = \{2，4，6\}$.

1.2 随机事件的概率

随机事件在一次试验中是否发生虽然不能确定，但让人感兴趣的是随机事件在一次试验中发生的可能性有多大。**概率** 就是用来描述随机事件发生的可能性大小的。比如抛硬币的试验，抛得次数越多，出现正面的次数与投掷次数之间的比例（也叫**频**

率)愈加趋于0.5。它的数学定义为：

> 在多次重复试验中，若事件A发生的频率稳定在确定常数p附近摆动，且随着试验次数的增加，这种摆动的幅度是很微小的。则称确定常数p为事件A发生的概率，记作$P(A)=p$.

思考题

设一年有365天，求下列事件A,B的概率:
$A=\{$n个人中没有2人同一天生日$\}$
$B=\{$n个人中有2人同一天生日$\}$

解

显然事件A，B是对立事件，P（B）$=1-P$（A）.
由于每个人的生日可以是365天的任意一天，因此，n个人的生日有365^n种可能结果，而且每种结果是等可能的，因而是古典概型，事件A的发生必须是n个不同的生日，因而A的样本点数为从365中取n个的排列数P_{365}^n，于是

$$P(A)=\frac{P_{365}^n}{365^n}$$

$$P(B)=1-P(A)=1-\frac{P_{365}^n}{365^n}$$

1.3 条件概率

定义

> 设A,B是两个事件，且$P(A)>0$，则称
>
> $$P(B|A)=\frac{P(AB)}{P(A)}$$
>
> 为在事件A发生的条件下，事件B的**条件概率**。

例子

某种元件用满$6000h$未坏的概率是$3/4$，用满$10000h$未坏的概率是$1/2$，现有一个此种元件，已经用过$6000h$未坏，试求它能用到$10000h$的概率。

解

设A表示{用满$10000h$未坏}，B表示{用满$6000h$未坏}，则

$$P(B)=3/4,P(A)=1/2$$

由于$A\subset B$, $AB=A$, 因而$P(AB)=1/2$，故

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{P(A)}{P(B)}=\frac{\frac{1}{2}}{\frac{3}{4}}=\frac{2}{3}$$

1.4 事件的独立性

定义

> 如果事件B发生的可能性不受事件A发生与否的影响，即

$$P(B|A)=P(B),$$

则称**事件B对于事件A独立**.显然，若B对于A独立，则A对于B也一定独立，称事件A与事件B相互独立.

性质

事件A和事件B相互独立的充分必要条件是

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

例子

口袋里装有5个黑球与3个白球，从中有放回地取2次，每次取一个，设事件A表示第一次取到黑球，事件B表示第二次取到黑球，则有

$$P(A)=\frac{5}{8},P(B)=\frac{5}{8},P(AB)=\frac{5}{8}\times\frac{5}{8}=\frac{25}{64}$$

因而

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{5}{8}$$

因此，$P(B|A)=P(B)$，事实上还可以算出$P(B|\overline{A})=P(B)$。这表明不论A发是不发生，都对B发生的概率没有影响。即B与A独立.

02 全概率公式与贝叶斯公式

2.1 全概率公式

定义

如果事件$A_1,A_2,\cdots,A_n$是一个完备事件组，并且都具有正概率，则有

$$\begin{aligned}P(B)&=P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+\cdots+P(A_n)P(B|A_n)\\&=\sum_{i=1}^{n}P(A_i)P(B|A_i)\end{aligned}$$

对于任何事件B，事件$A\overline{A}$构成最简单的完备事件组，根据全概率公式得

$$\begin{aligned}P(B)&=P(AB+\overline{A}B)=P(AB)+P(\overline{A}B)\\&=P(A)P(B|A)+P(\overline{A})P(B|\overline{A})\end{aligned}$$

2.2 贝叶斯公式

定义

设$A_1,A_2,\cdots,A_n$是一完备事件组，则对任一事件B, $P(B)>0$, 有

$$P(A_i|B)=\frac{P(A_iB)}{P(B)}=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n}P(A_i)P(B|A_i)}$$

以上公式就叫贝叶斯公式，可由条件概率的定义及全概率公式证得。

例子

市场上供应的某种商品只由甲、乙、丙3个厂生产，甲厂占45%，乙厂占35%，丙厂占20%。如果各厂的次品率依次为4%，2%，5%。现从市场上购买1件这种商品，发现是次品，试判断它是由甲厂生产的概率。

解

设事件A_1, A_2, A_3, 分别表示“商品为甲、乙、丙厂生产的”，事件B表示“商品为次品”，由题意得到概率

$$P(A_1) = 45\%, P(A_2) = 35\%, P(A_3) = 20\%$$

$$P(B|A_1) = 4\%, P(B|A_2) = 2\%, P(B|A_3) = 5\%$$

根据贝叶斯公式，可得：

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ &= \frac{45\% \times 4\%}{45\% \times 4\% + 35\% \times 2\% + 20\% \times 5\%} \approx 0.514 \end{aligned}$$

在“购买一件商品”这个试验中，$P(A_i)$是在试验以前就已经知道的概率，所以习惯地称为**先验概率**。试验结果出现了次品（即B发生），这时条件概率$P(A_i|B)$反映了在试验以后对B发生的“来源”（即次品的来源）的各种可能性的大小，通常称为**后验概率**。

03 随机变量，期望和方差

3.1 随机变量

把试验的结果与实数对应起来，随试验结果的不同而变化的量就是**随机变量**，包含离散型随机变量和连续型随机变量。

例子

掷一枚匀称的硬币，观察正面、背面的出现情况。这一试验的样本空间为$\Omega = \{H, T\}$，其中，H表示“正面朝上”，T表示“背面朝上”。如果引入变量X，对试验的两个结果进行数值化，将X的值分别规定为1和0，即

$$X = \begin{cases} 1 & \text{if 出现} H \\ 0 & \text{if 出现} T \end{cases}$$

这里的X就叫随机变量，因为它能取的值是离散的，我们就叫它离散型随机变量。

3.2 概率分布

定义

> 设离散型随机变量X的所有可能取值为$x_1, x_2, \cdots, x_n$, 称
>
> $$P\{X = x_k\} = p_k \quad (k = 1, 2, \cdots)$$
>
> 为X的**概率分布**。

离散型随机变量X的分布律具有下列基本性质：

1. $p_k \geq 0, k = 1, 2, \cdots$;

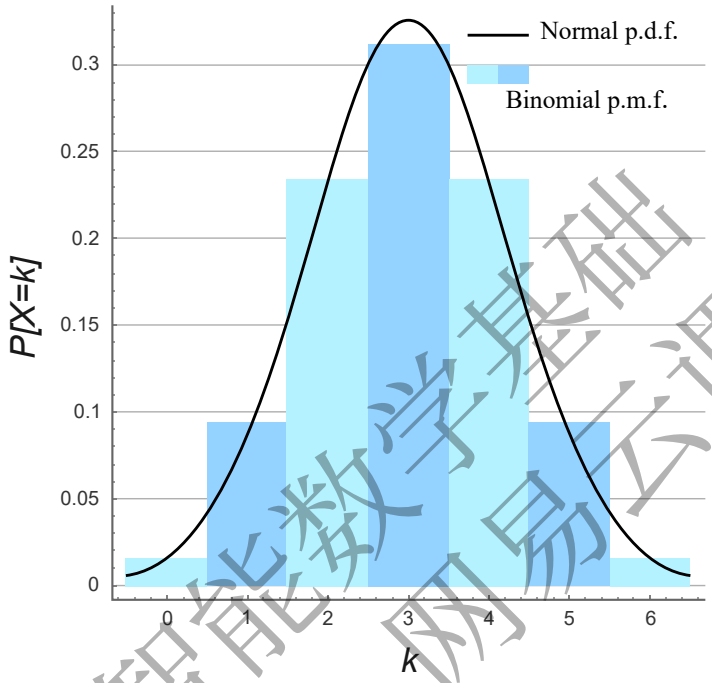
2. $\sum_{i=1}^{+\infty} p_k = 1$

二项分布

二项分布是一种离散型的概率分布。二项代表它有两种可能的结果：成功或者不成功。每次试验必须相互独立，重复n次，并且每次试验成功的概率是相同的，为p；失败的概率也相同，为$1-p$。

掷硬币就是一个典型的二项分布。当我们要计算抛硬币n次，恰巧有x次正面朝上的概率，可以使用二项分布的公式：

$$P\{X=k\} = C_n^k p^k (1-p)^{n-k}$$

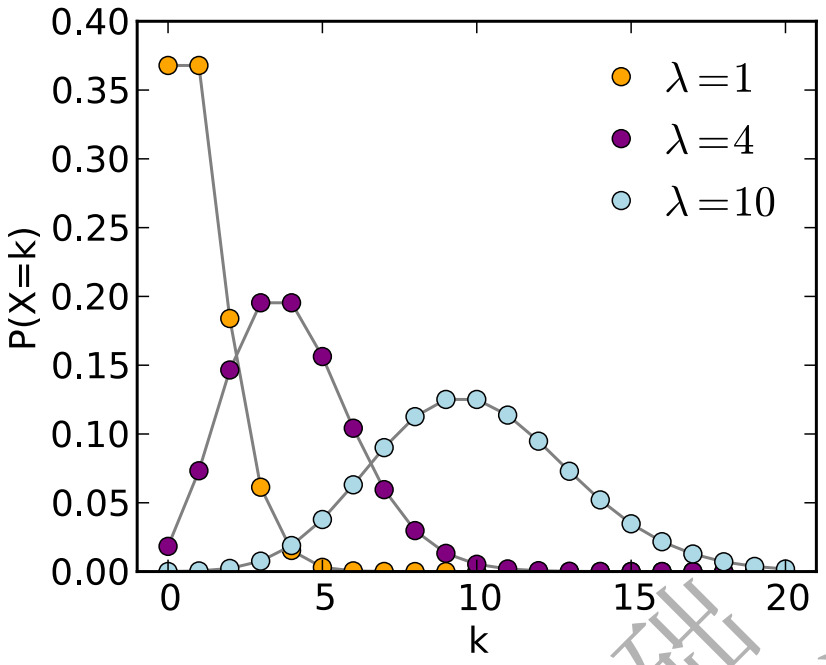


泊松分布

如果随机变量X的概率分布为

$$P\{X=k\} = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \cdots$$

式中，$\lambda > 0$为常数，则称随机变量X服从参数为λ的泊松(Possion)分布，记为$X \sim P(\lambda)$.



3.3 概率密度函数

定义

> 若存在非负函数$f(x)$，使一个连续型随机变量X取值于任一区间$(a，b]$的概率可以表示为
>
> $$P\{a \lhd X \trianglelefteq b\} = \int_a^b f(x)d_x$$
>
> 则称$f(x)$为随机变量X的**概率密度函数**，简称**概率密度**或**密度函数**。

正态分布

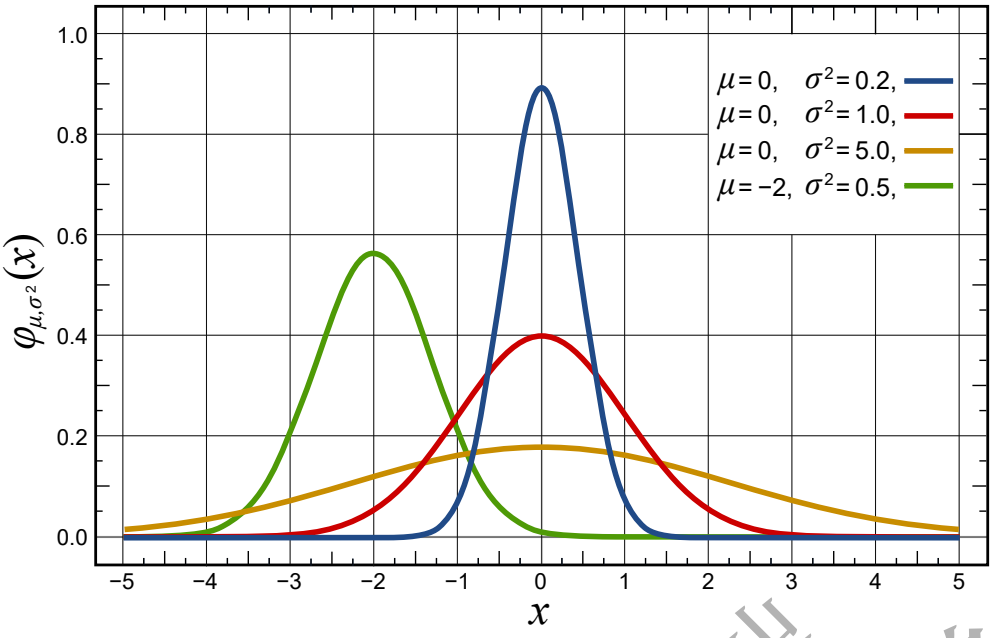
正态分布是概率论中最重要的连续型分布，在19世纪前叶由德国数学家高斯（Gauss）加以推广，故又常称为**高斯分布**。正态分布的概率密度函数曲线呈钟形，概率密度函数为

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

具有两个参数μ和σ^2。第一参数μ是代表服从正态分布的随机变量的均值，第二个参数σ^2是此随机变量的方差。如果一个随机变量服从均值为μ，标准差为σ的正态分布，数学上记作

$$X \backsim N(\mu，\sigma^2)$$

我们通常所说的标准正态分布均值为0, 标准差为1的正态分布。



<以上图片来自于维基百科>

3.4 随机变量的期望

对于一个随机变量，时常要考虑它的平均取什么，期望就是概率中的平均值，对随机变量中心位置的一种度量。

例子

经过长期观察积累，某射手在每次射击中命中的环数X服从分布:

X	0	5	6	7	8	9	10
p_i	0	0.05	0.05	0.1	0.1	0.2	0.5

求这个射手平均命中的环数是多少？

解

一种很自然的考虑是：假定该射击手进行了100次射击，那么，约有5次命中5环，5次命中6环，10次命中7环，10次命中8环，20次命中9环，50次命10，没有脱靶，从而在一次射击中，该射手平均命中的环数为

$$\frac{1}{100}(10\times 50+9\times 20+8\times 10+7\times 10+6\times 5+5\times 5+0\times 0)=8.85$$

所以，我们可以看到离散型的随机变量的期望值可以表示为

$$E(X)=\sum_{i=1}^{+\infty}x_ip_k$$

期望的性质

1. $E(c)=c$
2. $E(X+c)=E(X)+c$
3. $E(kX)=kE(X)$
4. $E(kX+c)=kE(X)+c$
5. $E(X+Y)=E(X)+E(Y)$

3.5 随机变量的方差

方差表示了随机变量的变异性，方差越大，随机变量的结果越不稳定。

定义

> 设X为一随机变量，若
>
> $$E[X-E(X)]^2$$
>
> 存在，则称其为X的方差，记为$D(X)$，即
>
> $$D(X)=E[X-E(X)]^2$$
>
> 而称$\sqrt{D(X)}$为X的**标准差或均方差**。

由方差的定义和数学期望的性质，可以推出方差的计算公式：

$$D(X)=E(X^2)-[E(X)]^2$$

方差的性质

(1) $D(c)=0$
(2) $D(X+c)=D(X)$
(3) $D(cX)=c^2D(X)$

例子

甲、乙两车间生产同一种产品，设1000件产品中的次品数分别为随机变量X，Y，已知他们的分布律如下:

X	0	1	2	3
p_i	0.2	0.1	0.5	0.2

Y	0	1	2	3
p_i	0.1	0.3	0.4	0.2

试讨论甲、乙两车间的产品质量。

解

先计算均值

$$E(X)=0\times0.2+1\times0.1+2\times0.5+3\times0.2=1.7$$

$$E(Y)=0\times0.1+1\times0.3+2\times0.4+3\times0.2=1.7$$

得到：甲、乙两车间次品数的均值相同。
再计算方差

$$D(X)=(0-1.7)^2\times0.2+(1-1.7)^2\times0.1+(2-1.7)^2\times0.1+(2-1.7)^2\times0.5+(3-1.7)^2\times0.2=1.01$$

$$D(Y)=(0-1.7)^2\times0.1+(1-1.7)^2\times0.3+(2-1.7)^2\times0.4+(2-1.7)^2\times0.4+(3-1.7)^2\times0.2=0.81$$

得到$D(X)>D(Y)$，说明乙车间的产品质量较稳定。

04 最大似然估计

概率 vs 统计 （非官方解释）

概率研究的问题是，已知一个模型和参数，怎么去预测这个模型产生的结果的特性（例如均值，方差，协方差等等）。统计研究的问题则相反，它是有一堆数据，要利用这堆数据去预测模型和参数。简单来说，**概率是已知模型和参数，推数据。统计是已知数据，推模型和参数。**

最大似然估计

最大似然估计是一种用来推测参数的方法，属于统计领域的问题。
用通俗的话说，最大似然估计是利用已知的样本结果信息，反推使这个结果出现可能性最大的模型参数值，是一种概率意义下的参数估计，由德国数学家高斯（Gauss）于1821年提出，后来英国统计学家费希尔（R.A.Fisher）于1922年重新发现并作了进一步的研究。

例子

假设有一种特殊的硬币，抛掷这种硬币出现的正反面并不相等的，求它正面出现的概率（记为θ）是多少？

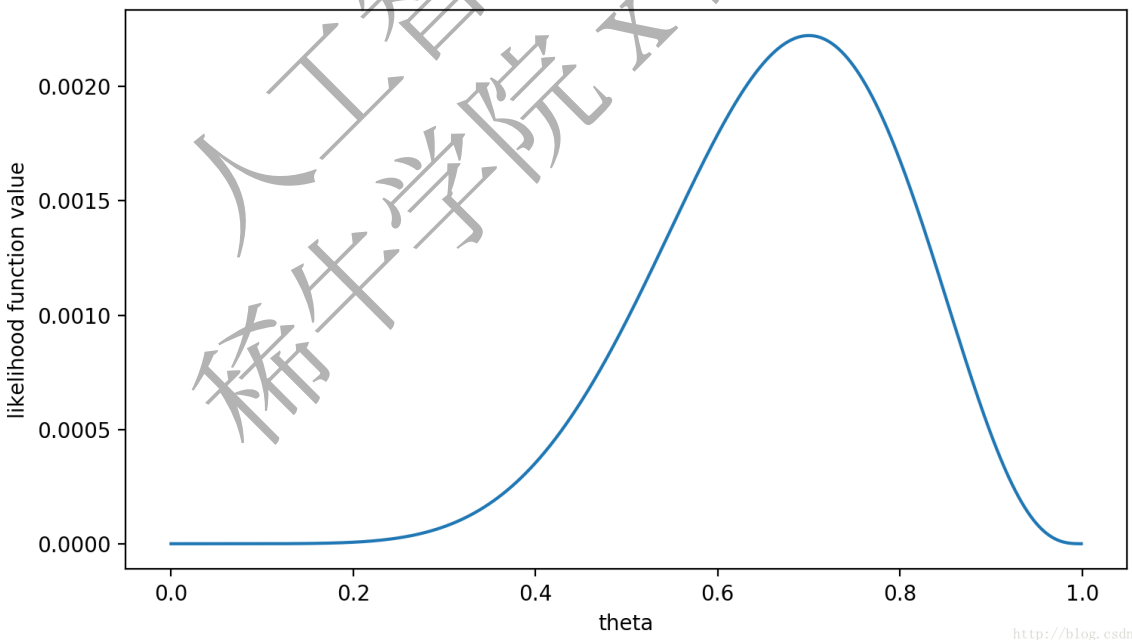
解

首先，这是一个统计问题，回想一下，解决统计问题需要什么？ 数据！

于是我们拿这枚硬币抛了10次，得到的数据（x_0）是：反正正正正反正正正反。我们想求的正面概率θ是模型参数，而抛硬币模型我们可以假设是二项分布。那么，出现实验结果x_0（即反正正正正反正正正反）的似然函数是多少呢？

$$f(x_0,\theta) = (1-\theta) \times \theta \times \theta \times \theta \times \theta \times (1-\theta) \times \theta \times \theta \times \theta \times (1-\theta) = \theta^7(1-\theta)^3 = f(\theta)$$

这是个只关于θ的函数。而最大似然估计，就是要最大化这个函数。于是，我们画出$f(\theta)$的图像：



可以看出，在θ=0.7时，$f(\theta)$取得最大值。

这样，我们已经完成了对θ的最大似然估计。即，抛10次硬币，发现7次硬币正面向上，最大似然估计认为正面向上的概率是0.7。

<以上例子和图片来自于 nebulaf91的csdn博客>

本章要点总结

- 样本空间是一次随机试验出现的所有可能结果的集合
- 随机事件是样本空间的子集
- 概率是用来描述随机事件发生的可能性大小的
- 条件概率公式，全概率公式和贝叶斯公式
- 随机变量是对随机试验的结果数量化
- 随机变量的期望代表了随机变量的平均值，而随机变量的方差刻画了随机变量的取值对于其数学期望的离散程度
- 最大似然估计是一种概率意义下的参数估计, 它利用已知的样本结果信息，反推使这个结果出现可能性最大的模型参数值

参考资料

- 《工程数学 概率论与数理统计》 同济大学出版社
- 详解最大似然估计（MLE）、最大后验概率估计（MAP），以及贝叶斯公式的理解
- 维基百科