贝叶斯的故事

正向概率：由样本分布的概率/情况去计算观测结果出现的概率

逆向概率：由观测结果去预测样本分布的概率/情况

现实世界里，很多本身的数据分布是不确定的，人类的观察能力是局限的，我们会根据观察到的结果去预测数据本身的分布状态。

通过求逆向概率得到本身分布，而求逆向概率可以通过求正向概率得到。

男生：60%　，　总穿长裤

女生：40%　，　一半穿长裤一半穿裙子

正向概率：随机碰见一个学生，穿长裤的概率是多大

逆向概率：随机碰见一个学生穿着长裤，他是女生的概率是多大

穿长裤的男生的人数：U*P(boy)*P(Pants|boy)　　其中 P(Pants|boy) = 100%

穿长裤的女生的人数：U*P(girl)*P(Pants|girl) 其中 P(Pants|girl) = 50%

穿长裤的是女生的概率 P(girl|Pants)：

$$\frac{U*P(girl)*P(Pants|girl)}{U*P(girl)*P(Pants|girl)+U*P(boy)*P(Pants|boy)}$$

P(girl)*P(Pants|girl) = *P(Pants, girl)*　　　　条件概率公式

P(girl)*P(Pants | girl)+P(boy)*P(Pants | boy) = P(Pants, girl)+P(Pants, boy) = *P(Pants)*

$$P(girl|Pants) = \frac{P(Pants\ \ girl)}{P(Pants)}$$　　条件概率公式

$$P(girl|Pants) = \frac{P(girl)*P(Pants\,|\,girl)}{P(Pants)}$$　　贝叶斯公式

贝叶斯公式就是要求逆向概率，可以把问题转换成求正向概率的运算。

补充条件概率：

1、定义

设 A,B 为两个事件，且 $P(A)>0$，在已知事件 A 发生的条件下，事件 B 发生的概率叫做条件概率，用符号 $P(B|A)$ 表示，$P(B|A)$ 读作：A 发生的条件下 B 发生的概率．

$P(A)=\frac{S_A}{S_u}$

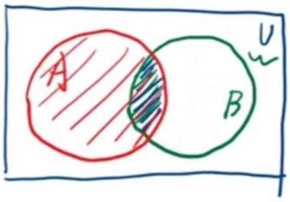
$P(AB)=\frac{S_{AB}}{S_U}$

2、$P(B|A)$、$P(AB)$、$P(A)$ 的区别

$P(B|A)$ 是在事件 A 发生的条件下，事件 B 发生的概率；

$P(AB)$ 是事件 A 与事件 B 同时发生的概率，无附加条件；

$P(A)$ 是事件 A 发生的概率，无附加条件．



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

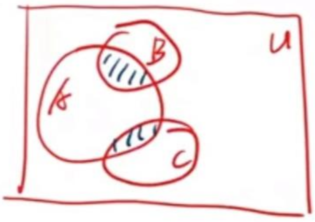
3、条件概率公式： $P(B|A)=\frac{P(AB)}{P(A)}$

4、条件概率公式的变形

公式 $P(B|A)=\frac{P(AB)}{P(A)}$ 揭示了 $P(B|A)$、$P(AB)$、$P(A)$ 的关系，常常用于知二求一

①若 $P(A)>0$，则 $P(AB)=P(A)\cdot P(B|A)$；

②若事件 B 与事件 C 互斥，则 $P(B\cup C|A)=P(B|A)+P(C|A)$



举个现实例子：

新冠病毒，得了新冠的人被检测出为阳性的几率为 90%，未得新冠的人被检测出阴性的几率为 90%，而人群中得新冠的几率为 1%，一个人被检测出阳性，问这个人得新冠的几率为多少？

A 表示事件 “测出为阳性”，B1 表示“得新冠”，B2 表示“未得新冠”

$$P(A|B_1)=0.9, P(A|B_2)=0.1, P(B_1)=0.01, P(B_2)=0.99$$

检测为阳性且得新冠的概率：

$$P(B_1,A)=P(B_1)\cdot P(A|B_1)=0.01\times 0.9=0.009$$

未得新冠且检测出阳性的概率：

$$P(B_2, A) = P(B_2) \cdot P(A|B_2) = 0.99 \times 0.1 = 0.099$$

在检测出阳性的前提下得新冠的概率 $P(B_1|A)$ 就是看被测出为阳性的这 108(9+99)人里（总人数 1000 人），9 人和 99 人分别占的比例是我们要的,也就是需要添加一个归一化(normalization)

$P(B_1|A)$ 为: $\frac{0.009}{0.099 + 0.009} \approx 0.083$，$P(B_2|A)$ 为: $\frac{0.099}{0.099 + 0.009} \approx 0.917$

这个两个条件概率就是贝叶斯统计中的**后验概率**，而人群中患新冠与否的概率 $P(B_1), P(B_2)$ 就是**先验概率，**我们知道了先验概率，根据观测值，是否为阳性，来判断得癌症的后验概率，这就是基本的贝叶斯思想，我们现在就能得后验概率的公式：

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2)}$$

B 事件的分布是离散的，所以在分母用的是求和符号$\sum$。那如果我们的参数 B 的分布是连续的就要用积分，于是贝叶斯公式 ：

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}$$

其中π指的是参数的概率分布，π(θ)指的是先验概率，π(θ|x)指的是后验概率，f(x|θ)指的是我们观测到的样本的分布，也就是似然函数(likelihood)。分母中积分求的区间Ø指的是参数θ所有可能取到的值的域，所以后验概率π(θ|x)是在知道 x 的前提下在Ø域内的一个关于θ的概率密度分布，每一个θ都有一个对应的可能性(也就是概率)。