**全概率公式和贝叶斯公式**

**定义** 设S为试验E的样本空间，B1, B2, …Bn为E的一组事件，若

BiBj≠Ф, i≠j, i, j=1, 2, …,n;

 B1∪B2∪…∪Bn=S

则称B1, B2, …, Bn为样本空间的一个划分。

定理 设试验E的样本空间为，A为E的事件，B1, B2, …,Bn为的一个划分，且P(Bi)>0 (i=1, 2, …n)，则

P(A)=P(A∣B1)P(B1)+P(A∣B2)+ …+P(A∣Bn)P(Bn)

称为全概率公式。

**定理** 设试验E的样本空间为S，A为E的事件，B1, B2, …,Bn为的一个划分，则

P(Bi∣A)=P(A∣Bi)P(Bi)/∑P(B｜Aj)P(Aj)=P(B｜Ai)P(Ai)/P(B)

称为贝叶斯公式。

假定各待分类文本特征变量是相互独立的，如果样本特征不独立，可能影响分类正确性。

贝叶斯文本分类器的任务是将准备分类的文本表示为向量X（x1，x2,x3…xm）分类到与其关系密切的类别Y（y1,y2,y3….yn）中去，

即利用贝叶斯公式根据某对象的先验概率计算出其后验概率，然后选择具有最大后验概率的类作为该对象所属的类。

训练数据的目的就在于获取样本各个特征在各个分类下的先验概率。

假设某样本x有x1,...,xm个属性,文本分类Y有y1,y2…yn种，那么有：

X={X1,x2….xm} 是一个待分类的项，xm为x的特征属性（地位相同、相互独立），共m个。

文本类别Y={y1,y2….yn}

计算各分类项的后验慨率P（y1|x）… P(yn|X)

如果P（yj|x）=max(P(y1|x…P(yn|x)) 则x属于yj类

P（yi|x）=P（x|yi）\*P(yi)/P(X)

Y因为P(x)固定，只要计算分子P（x|yi）\*p（yi）就可以了

根据独立性：

P（x|yi）=p(x1|yi）\*p(x2|yi)\*..p(xm|yi))

由于文本分类是由特征值慨率获得文本属于某一类的慨率，为避免某特征值为0出现整个乘积慨率为0的情况。所有特征值计数都+1.