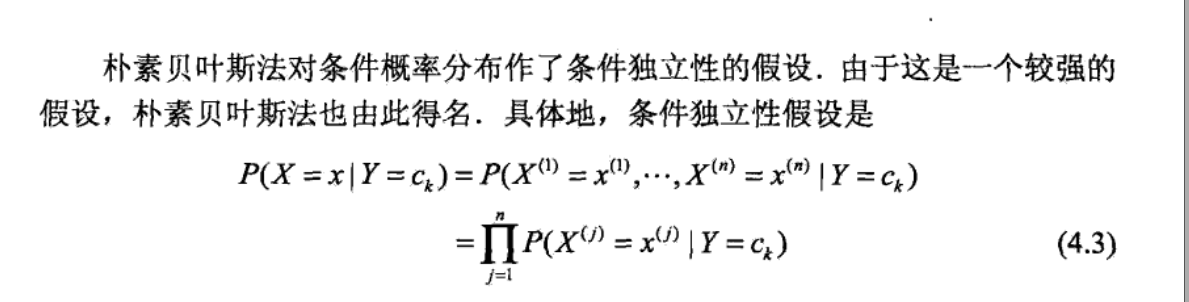
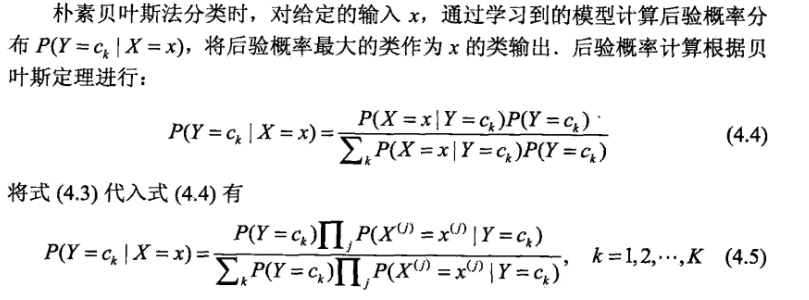
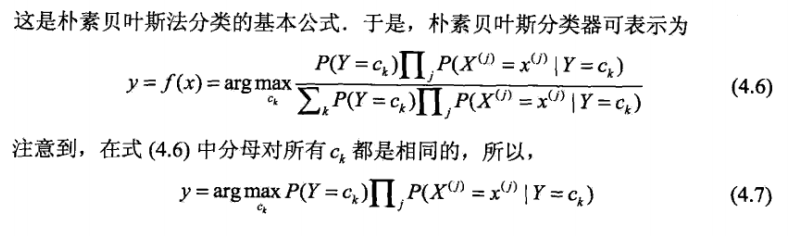
**朴素贝叶斯**

一、

总结：朴素贝叶斯法是假设样本各维度之间是独立同分布的。

二、朴素贝叶斯的推导公式



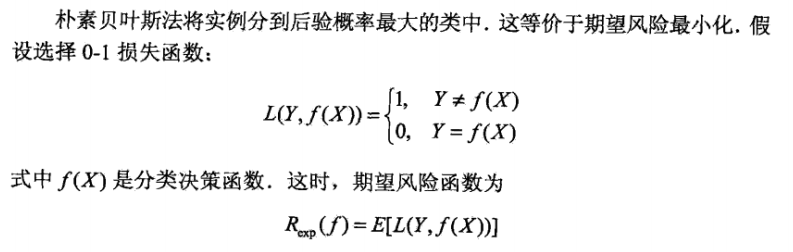


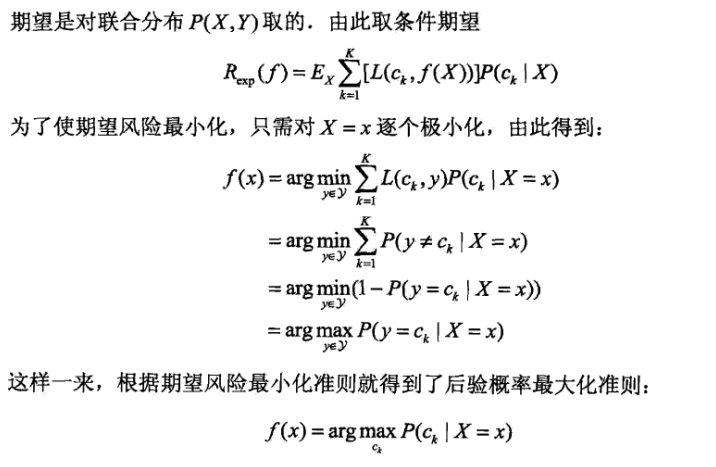
总结过程：

简单说，就是已知样本特征X={x1,x1……xn}；想知道样本X属于第c1类、c2类还是ck类。就是根据后验概率最高的来判断。

而计算公式为式4.5（这是假设了所有样本的特征均为独立同分布的情况下），式4.6即从计算出的k个概率里选出概率最大的那个对应的类别，但式4.6的分母是固定的，为了降低计算量可以将分母去掉就成了式4.7。

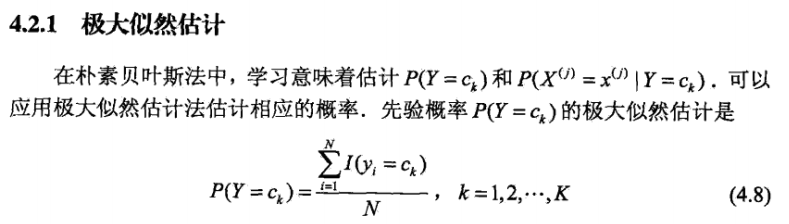
三、后验概率最大化准则

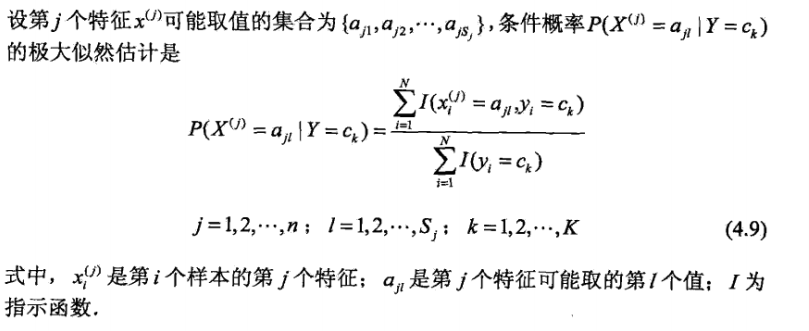




总结：以上推导只为证明一件事，朴素贝叶斯进行分类所遵循的后验概率最大化准则是有依据的，依据就是本书开始提到的经验风险最小化和期望风险最小化中的期望风险最小化准则。也就是说，朴素贝叶斯选择模型的标准是遵循了期望风险最小化的原则。

四、条件概率的计算方法（极大似然估计）





总结：

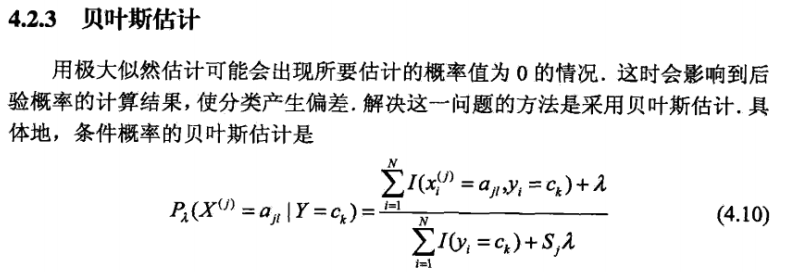
式4.7已经给出了朴素贝叶斯计算最大后验概率的公式。那么公式中的两个部分，先验概率和条件概率如何计算呢。就是式4.8与式4.9。

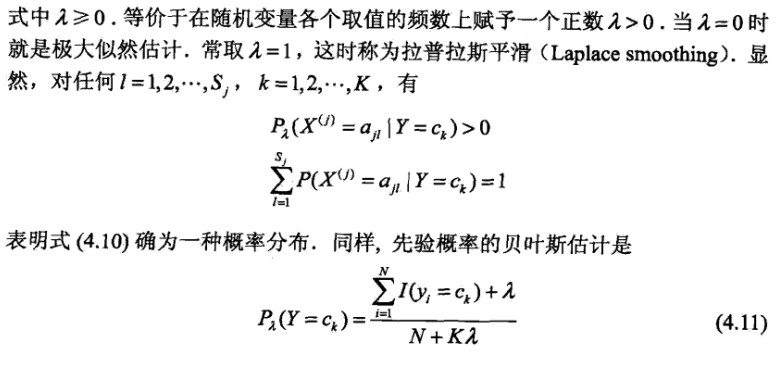
式4.8，是说在已知的样本集中，ck类样本所占的概率有多少；

对于条件概率的总的计算公式，式4.3已经给出，式4.9是对总的条件概率中连乘的每一项的计算公式，式4.3说明的意思是在样本集中特征集合为X的ck类样本占总的ck类样本的比例；式4.9说明的意思是在样本集中特征集合的第j个特征取值为ajl的样本中为ck类的样本占总的ck类样本的比例。

通过4.3、4.9、4.8式的计算，就做好了4.7式的准备条件，通过4.7式就可以算出样本X属于哪一类了。

五、贝叶斯估计





总结：

前面的求条件概率的方法，即式4.9存在缺陷。万一当某个特征取ajl值时的所有样本没有ck类的，将会导致所得的这一项条件概率为0，从而导致最后的此类后验概率为0，意思是说如果某个样本的这个特征如果为ajl则这个样本一定不属于ck类。这样似乎有点太绝对。

为了不要这么绝对，希望所有的后验概率都大于0，不要出现等于0的情况，就在式4.9中加入一项ʎ，直观的意思是说，我给样本集中增加Sjʎ个样本，从而使条件概率大于0。

在先验概率中也用了同样的手法。

这就是贝叶斯估计与前面极大似然估计的区别。

六、我理解的朴素贝叶斯

1、参数：先验概率与条件概率

2、模型：式4.7

3、过程：通过大量的训练集得到先验概率与条件概率，然后通过式4.7进行分类计算，概率最大的为最终分类。

4、缺点：主要在于使用朴素贝叶斯模型就默认了特征之间是独立的，从而会损失一定分类准确率。

朴素贝叶斯特征：

1、生成模型；

2、模型选择目标是：期望风险最小化。