**朴素贝叶斯**

**朴素贝叶斯算法**

设每个数据样本用一个n维特征向量来描述n个属性的值，即：X={x1，x2，…，xn}，假定有m个类，分别用C1, C2,…，Cm表示。给定一个未知的数据样本X（即没有类标号），若朴素贝叶斯分类法将未知的样本X分配给类Ci，则一定是P(Ci|X)>P(Cj|X) 1≤j≤m，j≠i  
 根据贝叶斯定理，由于P(X)对于所有类为常数，最大化后验概率P(Ci|X)可转化为最大化先验概率P(X|Ci)P(Ci)。如果训练数据集有许多属性和元组，计算P(X|Ci)的开销可能非常大，为此，通常假设各属性的取值互相独立，这样先验概率P(x1|Ci)，P(x2|Ci)，…，P(xn|Ci)可以从训练数据集求得。  
 根据此方法，对一个未知类别的样本X，可以先分别计算出X属于每一个类别Ci的概率P(X|Ci)P(Ci)，然后选择其中概率最大的类别作为其类别。朴素贝叶斯算法成立的前提是各属性之间互相独立。当数据集满足这种独立性假设时,分类的准确度较高，否则可能较低。另外，该算法没有分类规则输出。

# **朴素贝叶斯的推导**

  对于样本集:

# 其中 m表示有 m个样本， n 表示有 n 个 特征。



# **先验概率**为：



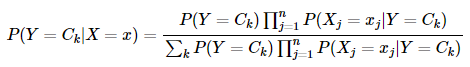
# 条件概率为（依据条件独立假设）：



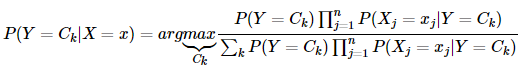
# 则后验概率为：



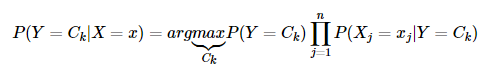
# 将条件概****率****公式带入得：



上式为朴素贝叶斯分类的基本公式。于是，朴素贝叶斯分类器可表示为：



由于分母对所有的Ck都是相同的，所以：



# 

# **朴素贝叶斯算法分析**

# 优点：     （1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。     （2）对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是 数据量超出内存时，我们可以一批批的去增量训练。     （3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。   缺点：     （1）理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下,假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。     （2）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。     （3）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。     （4）对输入数据的表达形式很敏感。