**贝叶斯算法**

在机器学习中有很多算法，而有一种算法有着坚实的数学背景，并且被广泛使用，这种算法就是朴素贝叶斯算法，当然，朴素贝叶斯算法的优点有很多，当然，这种算法的缺点也是我们不能忽视的，那么大家知道不知道朴素贝叶斯算法的优点和缺点是什么呢？下面我们就给大家介绍一下这个问题。

那么什么是朴素贝叶斯算法呢？其实朴素贝叶斯属于生成式模型，也就是关于生成模型和判别式模型，主要还是在于是否需要求联合分布，这种算法是一种比较简单的算法，你只需做一堆计数即可。如果注有条件独立性假设，朴素贝叶斯分类器的收敛速度将快于判别模型，比如逻辑回归，所以你只需要较少的训练数据即可。即使NB条件独立假设不成立，NB分类器在实践中仍然表现的很出色。它的主要缺点是它不能学习特征间的相互作用，用mRMR中R来讲，就是特征冗余。

这种算法的优点有五个，第一就是朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。第二就是对大数量训练和查询时具有较高的速度。即使使用超大规模的训练集，针对每个项目通常也只会有相对较少的特征数，并且对项目的训练和分类也仅仅是特征概率的数学运算而已。第三就是对小规模的数据表现很好，能够处理多分类任务，适合增量式训练（即可以实时的对新增的样本进行训练）。第四就是对缺失数据不太敏感，算法也比较简单，常用于文本分类。第五就是朴素贝叶斯对结果解释容易理解。

当然，朴素贝叶斯算法的缺点也是很明显的，朴素贝叶斯算法的缺点有四点，第一就是

需要计算先验概率。第二就是分类决策存在错误率。第三就是对输入数据的表达形式很敏感。第四就是对由于使用了样本属性独立性的假设，所以如果样本属性有关联时其效果不好。

那么朴素贝叶斯应用领域是什么呢？其实朴素贝叶斯算法在欺诈检测中使用较多。当然，我们还可以用朴素贝叶斯算法来决定一封电子邮件是否是垃圾邮件。还可以用朴素贝叶斯算法判断一篇文章应该的类别，同时也能够使用贝叶斯算法去判断一段文字表达的是积极的情绪还是消极的情绪。从中我们可以看出朴素贝叶斯算法是一个十分实用的算法

P(A)=垃圾邮件的概率，垃圾邮件/所有的邮件。   
P(B) = 邮件中出现词M的概率，出现词M的邮件/所有的邮件。   
P(B|A) = 垃圾邮件中出现词M概率，垃圾邮件中含有词M的邮件数量/所有的垃圾邮件。

可能大家第一次接触可能会有很大的疑问，P(B|A)不是可以用P(A)和P(B)计算出来吗，但是此处的P(B|A)并不是由基础的P(A)和P(B)计算得出，而是由样本中大量的数据统计而来的概率，因此大家在刚刚接触的时候一定要小心。因此在使用贝叶斯定理之前就首先需要有数据库，然后得出基本的三个概率。

因此我们由上述的三个概率可以得出，出现词M的邮件是垃圾邮件的概率P(A|B) = P(A)\*P(B|A)/P(B) =邮件中出现词M的概率X垃圾邮件中出现词M概率/是垃圾邮件的概率。

此应用中对垃圾邮件学习的过程就是计算P(A|B)的过程。一封垃圾邮件中会有多个词被包含在其中，因此需要不断地尝试不同词组的组合尝试，直到找到概率大于预期概率（比如。0.8，0.9等）的词M或一组词M1，M2等。然后就可以用得到的词计算，判断新的邮件是否是垃圾邮件。