# Homework 2: NBC 实验报告

姓名：孙畅畅
学号：201814832

## （一）实验任务

（1）实现朴素贝叶斯分类器，测试骑在 20Newsgroups 数据集上的效果。
（2）实现朴素贝叶斯分类器中的多项式模型、伯努利模型和混合模型。

## （二）实验数据集

（1）数据集种类：20 个种类　　数据集数目：18828

种类：alt.atheism（799）comp.graphics（973）comp.os.ms-windows.misc（985）comp.sys.ibm.pc.hardware（982）comp.sys.mac.hardware（961）comp.windows.x（980）misc.forsale（972）rec.autos（990）rec.motorcycles（994）rec.sport.baseball（994）rec.sport.hockey（999）sci.crypt（991）sci.electronics（981）sci.med（990）sci.space（987）soc.religion.christian（997）talk.politics.guns（910）talk.politics.mideast（940）talk.politics.misc（775）talk.religion.misc（628）（括号内表示各个类别的样本数）

（2）训练集和测试集的划分

考虑到在训练集和测试集中各个类别的数目比例，随机从上诉各个类别中供选择一定数目的样本作为训练集与测试集，各个类别数目依次为：

训练集：

[700,800,800,800,800,800,800,800,800,800,800,800,800,800,800,800,800,800,700,450]

测试集：

[99,173,185,182,161,180,172,190,194,194,199,191,181,190,187,197,110,140,175,78]

训练集与测试集比例约为：4:1

## （三）实验过程

1、实验前的知识准备

（1）条件独立性假设

朴素贝叶斯分类器在估计类条件概率时假设属性之间条件独立，当用一批切分好的词语表示一个文档时，不再考虑这些词语之间的顺序等关系。

（2）平滑假设

在计算每个文本属于某一个类别的概率时，有可能会出现单词表中的单词在该类别中没有出现过的情况，这时候，如果在计算类条件概率时，如果当做零处理，则无法进行分类，这个时候就采用平滑假设。并且对于多项式模型和伯努利模型，平滑假设的公式不同。

（3）无关紧要的词语的删除

在具体实施实验之前，为了删除文本中的停用词，以及对于分类没有太大帮助的词。以此来提高分类的准确率。

2、多项式模型

（1）在多项式模型中，设文档 d={t1,t2,t3,,,,tk},tk 表示文档中出现的词，允许重复。

先验概率：P(c)=类 c 下单词总数/整个训练文本的单词总数

类条件概率 p(tk|c)=(类 c 下单词 tk 在各个文档中出现的次数之和+1)/(类 c 下单词总

数+|V|)。(V 是训练样本中出现的单词，出现多次，仅算一个)

（2）文本属于某个类别的概率为：

$$P(d \mid c) = p(c)\prod_{i=1}^{i=k} p(t_i \mid c)(d = \{t_1, t_2,,, t_k\})$$

（3）分别计算每个文档属于某个类别的概率，取概率最大值，当做分类结果，并与真实的类别进行比较。

3、伯努利模型

（1）在多项式模型中，设文档 d={t1,t2,t3,,,,tk},tk 表示文档中出现的词，不允许重复。

先验概率：P(c)=类 c 下的文件数/整个训练样本文件数

类条件概率：P(tk|c)=(类 c 下包含单词 tk 的文件数+1)/(类 c 下的文件数+2)

（2）文本属于某个类别的概率：

$$P(d \mid c) = p(c)\prod_{i=1}^{i=k} p(t_i \mid c)\prod_{i=k+1}^{i=n}(1 - p(t_i \mid c))(d = \{t_1, t_2,,, t_k\}, \{t_{k+1},,, t_n\}$$表示在单词表中，但不属于文本d的单词)

在这里可以看出，无论是先验概率的计算，还是类条件概率的计算，多项式模型和伯努利模型的计算方法都不同。

4、混合模型

（1）在混合模型中，在训练阶段，考虑重复的词语。在测试阶段，不考虑重复的词语。所以我在编写代码的过程中，将前面两种方法进行综合，在读取文本时，记录需要的数据，然后求解概率。

5、实验技巧

在具体实验求概率时，由于计算乘积时，运算速度慢。所以我使用了 log 函数，将求乘积转化为求和，提高了运算速度。

## （四）实验结果

对比三种模型在分类上的准确率，分析三种模型分类结果的高低。

（1）多项式模型


（2）伯努利模型


（3）混合模型


## （五）实验总结

在本次实验中，使用了朴素贝叶斯对文本进行分类，并分别实现了多项式模型、伯努利模型和混合模型。通过实际的实验，了解了概率的计算以及比较了两种模型的在计算后验概率上的不同。

通过具体的实验，对两种模型有了更加深入的理解，体会到了三者的不同，发现多项式模型和伯努利模型的计算粒度不同，多项式模型是以单词为粒度，伯努利模型是以文件为粒度，并且在伯努利模型中，对于在文本中没有出现的单词，则用 1 减去后验概率出现在最终的概率计算中。

对于文本分类问题，通过以往的实验，发现文本的预处理非常重要。我们选择那些只对文本类别的判断有帮助的词去组成文本的特征向量。