实验报告

一、 简介

本实验是完成贝叶斯分类器在新闻分类上的实现，主要过程分为两部分：

1）对数据集中的文章进行预处理，形成词典

2）通过贝叶斯公式计算文档属于某一个类的概率，选择概率最大的那个类别

二、 方法步骤

1 数据集预处理

预处理部分和上一次 knn 实验基本相同，即大小写统一，去除非字母部分，词干提取，分词等，去掉高频和低频词已减小对内存的负担。

2 贝叶斯分类器

朴素贝叶斯中的朴素一词的来源就是假设各特征之间相互独立。这一假设使得朴素贝叶斯算法变得简单，但有时会牺牲一定的分类准确率。

首先给出贝叶斯公式：$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

换成数据挖掘任务的表达式：$$\text{p(类别|特征)}=\frac{p(\text{特征}|\text{类别})\text{p(类别)}}{\text{p(特征)}}$$

最终需要求得 p(类别|特征)即可得到在这一文章特征下属于这一类别的概率，我们取概率最大的类别当做我们的估计类别，然后和真实类别进行比较确认对错。假设每个单词的出现是相互独立的，我们可以将每个单词看作一个特征属性然后概率相乘，因为分母相同，我们只需求得分子即可

$$y=\arg\max_{c_k} P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)$$

实验结果表明正确率在 0.88 左右，伯努利实验即不考虑词频的情况下正确率在 0.76 左右。