# 编程作业说明：贝叶斯网络

## 任务一：使用朴素贝叶斯过滤垃圾邮件

**（1）问题**：现有 50 封电子邮件，存放在数据集 task1 中，试基于朴素贝叶斯分类器原理，用 Python 编程实现对垃圾邮件和正常邮件的分类。采用交叉验证方式并且输出分类的错误率及分类错误的文档。

**（2）步骤提示：**

**1.收集数据：**这里数据直接提供，一般情况下需要自己手动采集及预分类；

**2.数据整理**：拿到手的数据并不能直接使用，我们需要对数据进行预处理，变成向量的形式（这里采用**词袋模型**）。可以首先把所有字符转换成小写，去掉大小字符不统一的影响；然后构建一个包含在所有文档中出现的不重复的词的列表；获得词汇表后，根据某个单词在一篇文档中出现的次数获得文档向量（这三步的代码可参考给出的 demo）；最后利用构建的分类器进行训练。

**3.训练和测试：**导入文件夹 spam 与 ham 下的文本文件，并将它们解析为词列表。接下来构建一个训练集和测试集，两个集合中的邮件都是随机选出的。经过一次迭代就能输出分类的错误率及分类错误的文档。（注：由于测试集的选择是随机的，所以测试算法时每次的输出结果可能有些差别，如果想要更好的估计错误率，最好将上述过程重复多次，然后求平均值。）

**（3）提交要求：**

任务一需要编写实验报告进行简述其原理，编程思路，数据集的划分方式以及错误率等。

## 任务二：使用朴素贝叶斯对搜狗新闻语料库进行分类

**（1）数据集讲解**：数据集存放在 task2/Database 中，新闻一共分为 9 个类：财经、IT、健康、体育、旅游、教育、招聘、文化、军事，不同种类的新闻分别放在相应代号的文件夹中。NBC.py 中，对新闻进行预处理的模块已经给出，请尝试补全分类器模块的代码（NBC.py 中函数 TextClassifier），自行划分训练集和测试集，对这些新闻进行训练和分类。

**（2）提示：**

**1. 数据预处理：**对语料库中的文本进行分词，并通过词频找到特征词，从而生成相应的训练集和测试集。这部分代码我们已经给出，请尝试看懂并理解它们。

**2. 机器学习库 sklearn：**我们鼓励你手写朴素贝叶斯分类器，但是在本任务中，我们允许你使用机器学习库 sklearn 快速实现分类器，以方便研究不同的参数设置对实验结果的影响，探讨影响朴素贝叶斯分类器分类效果的原因。

**3. 分词工具：**本任务需要你安装分词库 jieba。安装方法：pip install jieba。

**（3）提交要求：**

请尝试在不同的参数下多做几组实验：例如可以在构建词典中尝试删去不同个数的高频词，观察实验结果的变化；也可以在更改不同的特征词数量，观察对分类准确率的影响；或者更改训练集和测试集划分比例，观察不同训练集规模对实验结果的影响；也可以更换特征词的提取方式，如 TF-IDF 等等。编写实验报告，报告中需要包含：实验设置，编程思路，实验结果展示以及自己的思考等。

## 任务三：使用朴素贝叶斯对电影评论分类

**（1）数据集讲解**：

该数据集是 IMDB 电影数据集的一个子集，已经划分好了测试集和训练集，训练集包括 25000 条电影评论 ,测试集也有 25000 条 ,该数据集已经经过预处理，将每条评论的具体单词序列转化为词库里的整数序列 ,其中每个整数代表该单词在词库里的位置。例如，整数 104 代表该单词是词库的第 104 个单词。为实验简单，词库仅仅保留了 10000 个最常出现的单词，低频词汇被舍弃。每条评论都具有一个标签，0 表示为负面评论，1 表示为正面评论。

训练数据在 train_data.txt 文件下，每一行为一条评论，训练集标签在 train_labels.txt 文件下，每一行为一条评论的标签；测试数据在 test_data.txt 文件下，测试数据标签未给出。

**（2）步骤提示：**

**1.训练思路：**每个文档可以看成由 n 个特征构成的文档向量，每个单词表示一个特征维度。统计正样本数和负样本数，可以得到文档分布的先验概率；统计每类样本中，某个单词出现的次数和总单词数的比值，可以得到该特征的条件概率。

**2.拉普拉斯平滑：**朴素贝叶斯用各个特征的条件概率连乘表示某个样本在某个类别的条件概率。然而，如果一个单词没有出现在某个类别的样本中，那么它的条件概率就是 0，导致最后的连乘结果也为 0，从而将不再有文档被分到这一类。因此在训练的过程中，注意使用拉普拉斯平滑处理。

**（3）提交要求**：

将测试数据预测结果，与训练数据标签存储方式相同，存储为 txt 文件，每一行为一条评论的标签。将测试集预测结果的 txt 文件由刘青霞同学负责统一发给助教。实验报告中需要写明具体实验流程，思路。