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：MNIST数据集基于朴素贝叶斯分类器的手写识别**

1. **数据集讲解**：MNIST数据集是由0〜9手写数字图片和数字标签所组成的，由60000个训练样本和10000个测试样本组成，每个样本都是一张28 \* 28像素的灰度手写数字图片。如下图所示：



从MNIST的官网上下载数据集，得到四个文件如下（分别为训练集图片，训练集标签，测试集图片，测试集标签）：

**train-images-idx3-ubyte.gz: training set images (9912422 bytes)**

**train-labels-idx1-ubyte.gz: training set labels (28881 bytes)**

**t10k-images-idx3-ubyte.gz: test set images (1648877 bytes)**

**t10k-labels-idx1-ubyte.gz: test set labels (4542 bytes)**

图片是指0〜9手写数字图片，而标签则是对应该图片之实际数字。

解压后得到四个文件：

**训练集图像：t10k-images.idx3-ubyte  
 训练集标签：t10k-labels.idx1-ubyte  
 测试集图像：train-images.idx3-ubyte  
 测试集标签：train-labels.idx1-ubyte**

1. **步骤提示：**

**图片文件格式说明：**

----------------------------------------

[字节位置] [类型] [值] [描述]

0000 32位整型 2051 幻数

0004 32位整型 60000 图片数

0008 32位整型 28 行数

0012 32位整型 28 列数

0016 无符号字节 ?? 像素

0017 无符号字节 ?? 像素

......

xxxx 无符号字节 ?? 像素

----------------------------------------

**标签文件格式说明：**

----------------------------------------

[字节位置] [类型] [值] [描述]

0000 32位整型 2049 幻数

0004 32位整型 60000 标签数

0008 无符号字节 ?? 标签

0009 无符号字节 ?? 标签

......

xxxx 无符号字节 ?? 标签

----------------------------------------

注：这里的整形指的都是无符号整型

上述的32位整形遵循MSB first，即高位字节在左边，如十进制8，二进制储存形式为1000。幻数是一个固定值，它占据文件的前4个字节，实际上表示的是这个文件储存的是图片还是标签，没有具体用处，我们可以忽略它。图片数与标签数占据文件4~7个字节的位置，在训练集中，它为60,000，表示这个文件有60,000个图片或标签，在测试集中，它为5,000。

行数和列数描述的是每张图片的大小，它们也是固定值，都为28。每张图片有28\*28=784个像素，所以从图片文件第16个字节位置开始，每隔784个字节为一张新图片，其中每个像素的像素值为0~255。从标签文件的第8个字节位置开始，每个字节都对应着一张图片的数字，标签的值为0~9。

task2.py中已经给出了MNIST数据集的读取和图片特征提取的参考代码。

**（3）提交要求：**

请读懂参考代码，并使用训练集训练贝叶斯分类器，在测试集上对训练好的模型进行评估，计算模型的分类精度，截图记录在实验报告中。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。实验报告中需要写明具体实验流程，思路。