**任务一：使用逻辑回归辨别真假钞票**

**（1）数据集讲解**：

钞票数据集（Banknote Dataset）涉及根据给定钞票的数个度量的照片预测是真钞还是假钞。它是一个二分类问题。每个类的观测值数量不均等。4个输入变量和1个输出变量。变量名如下：

变量名1：小波变换图像（连续）

变量名2：小波偏斜变换图像（连续）

变量名3：小波峰度变换图像（连续）

变量名4：图像熵（连续）。

类（0 为真钞，1 为假钞）

要求：**纸质版代码** 和 **电子版.csv结果文件**。**电子版文件**命名包括**姓名\_学号\_班级**。.csv文件**格式**为：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本序号 | 变量名1 | 变量名2 | 变量名3 | 变量名4 | 真钞or假钞 |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |

**任务二：使用逻辑回归对电影评论分类**

**（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。将文档转换为one-hot编码以后再使用逻辑回归模型。

**（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。