华 中 科 技 大 学

《大数据基础》课程

实 验 报 告

实验名称：分类　　实验学时：4

系别：　物理　　专业：　应用物理　　班级：　1502

姓名：　罗颖韬　　学号：U201510146

实验日期：2018 年 5 月 18 日

教师批阅签字：

一、 实验内容：

掌握朴素贝叶斯分类的核心思想，编写程序将给定的文件进行分类

二、实验基本原理和核心算法说明

Bayes.python 为核心算法部分，在于编写贝叶斯分类器。

三、相关程序代码：

```
def trainNB0(trainMatrix, trainCategory):
  numTrainDocs = len(trainMatrix)
  numWords = len(trainMatrix[0])
  pAbusive = sum(trainCategory) / float(numTrainDocs)
  # 初始化概率（这里是评分标准第二部分优化要修改的的地方 1）
  p0Num = zeros(numWords);
  p1Num = zeros(numWords)
  p0Denom = 0.0;
  p1Denom = 0.0
  for i in range(numTrainDocs):
      # 向量相加
      if trainCategory[i] == 1:
```

```
            p1Num += trainMatrix[i]
            p1Denom += sum(trainMatrix[i])
        else:
            p0Num += trainMatrix[i]
            p0Denom += sum(trainMatrix[i])
    # 对每个元素做除法（这里是评分标准第二部分优化要修改的地方 2）
    p1Vet = p1Num / p1Denom
    p0Vet = p0Num / p0Denom
    return p0Vet, p1Vet, pAbusive


def classifyNB(vec2Classify, p0Vec, p1Vec, pClass1):
    # （这里是评分标准第二部分优化要修改的地方 3）
    p1 = sum(vec2Classify * p1Vec) + pClass1
    p0 = sum(vec2Classify * p0Vec) + 1.0 - pClass1
    if p1 > p0:
        return 1
    else:
        return 0
```

四、思考题

1、贝叶斯算法的原理是基于什么的？

随机概率的定义推导而来，是数理统计与概率论知识。

2、贝叶斯算法与实验三的两种聚类算法（KMeans、Canopy）的异同点。

前者是分类，是 supervised learning；后者是聚类算法，是 unsupervised learning。

3、举个例子贝叶斯算法在互联网上的应用（结合 pdf 文档中对于垃圾文件的处理）？

可用于训练学习垃圾邮件的模样，然后去隔离垃圾邮件。

4、本次贝叶斯的是怎么进行优化的，对以后类似的问题有什么启发？

使用 log 函数，使得原本乘法变为加法，避免一个 0 值就把全部整体的值化为 0。

五、本次实验遇到的主要问题及解决方案

没有很难的问题，说实话。

六、对本次实验内容及方法、手段的改进建议

我们不妨把 **bayes** 那一块详细讲一下，以及它的优化原理。

七、实验心得

复习了机器学习基础的贝叶斯分类器。