实验二：聚类与分类

1. 实验目的

掌握对数据使用聚类分析与分类，并理解其在大数据环境下的实现方式。

2. 实验环境

Windows、Linux（推荐），伪分布式 Hadoop 环境（推荐），Java、Python 等

3. 实验内容

3.1 聚类分析

1. 基于 Hadoop 环境，将数据集 US Census Data（https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29）导入 HDFS 中，数据说明与下载地址（USCensus1990.data.txt）详见上述超链接；

2. 基于 Map-Reduce 实现 K-Means 聚类方法，对数据集进行聚类分析（K 值请自行设定），将聚类结果由 HDFS 导出，对聚类结果进行部分可视化，聚类结果可能会影响分数的获得；

3. 完成以上部分最高获得实验二总分的 40%，考虑基于 Map-Reduce 实现其他聚类方法（如 DBSCAN 等），并与 K-Means 结果进行比较，完成此部分最高获得实验二总分 10%的加分；

4. 提示：基于 MR 实现 K-Means 算法，可借助 Configuration 类保存中心点，在 Map 阶段计算点与中心的距离从而决定其所属，Reduce 阶段重新计算中心；

3.2 数据分类

1. 基于 Hadoop 环境，将数据集 SUSY Data Set（https://archive.ics.uci.edu/ml/datasets/SUSY）按 80%:20%的比例拆分为训练数据与测试数据分别导入 HDFS 中，数据说明与下载地址详见上述超链接；

2. 基于 Map-Reduce 实现朴素贝叶斯算法，使用上一步导入的训练数据进行训练，计算算法在测试数据上的正确率并输出（也可以同时输出召回率与精确率等，此处不做要求），正确率可能会影响分数的获得；

3. 完成以上部分最高获得实验二总分的 40%，考虑基于 Map-Reduce 实现其他分类算法（如逻辑回归、SVM 等），并与朴素贝叶斯算法结果进行正确率比较，完成此部分最高获得实验二总分 10%的加分；

4. 提示：算法的运行结果可能与导入数据时数据的顺序有关。基于 MR 实现朴素贝叶斯算法，通常需要多个 MR，分别进行概率的计算、结果的预测等，此处不做 MR 轮数限制。

4. 聚类结果可视化说明

1. 本实验推荐使用 Python 进行结果的可视化，本实验将给出 USCensus1990.data.txt 降至二维后的结果（usdata.pickle），文件顺序与原文件的前 **1** 万条数据一一对应，使用 Python 读取该文件并绘制散点图，即可进行可

视化，也可自行采用其他可视化方法，不对可视化方法强制要求，但要求能直观展现聚类结果，可视化点不应当少于 1 万个。

2. 给出基于 Python3 的可视化代码示例：

```
# 请自行解决依赖问题
import matplotlib.pyplot as plt
import pickle
# 从 pickle 文件读取降维结果
with open("usdata.pickle", "rb") as usdata:
data = pickle.load(usdata)
    y = cluster_result[:10000]  # 这里，y 表示聚类结果（一维向量）
    plt.plot(data[0], data[1], c=y)
    plt.show()
```

5. 参考资料

[1] 大数据课程讲稿-第 6 讲-聚类分析(学生).pdf

[2] 大数据课程讲稿-第 7 讲-分类分析(学生)-修改后 0403.pdf