Big Data Processing

Homework 4

要求

- 实现k-means clustering算法
- 使用k-means clustering算法对iris flower data set进行聚类
- 编写报告
- 两人组队（不组队也可，但不组队这件事自身不会带来任何加分）

Iris数据集介绍

- 原始数据介绍请参考 https://en.wikipedia.org/wiki/Iris_flower_data_set
- 由于原始数据features有4维，不方便可视化，故使用PCA算法进行降维，并在课程主页上提供降维后的数据，同学们直接使用降维后的数据即可。
- 降维后数据介绍如下，第一行是header部分，后续每行为具体数据，一共150个samples，前三列是降维后的features值，第四列代表数据对应的label，label为0,1,2分别对应setosa, versicolor,virginica。
- 每行之间使用\n分割，每列之间使用逗号分隔。

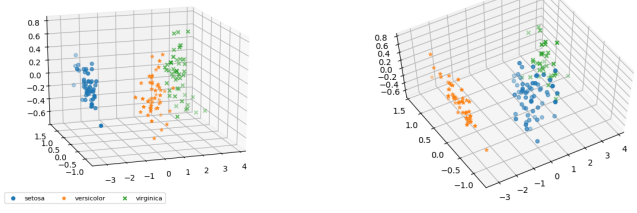
```
feature1,feature2,feature3,label
-2.684207125103950542e+00,3.266073147643868690e-01,-2.151183700196301896e-02,0.000000000000000000e+00
-2.715390615634133198e+00,-1.695568475560270405e-01,-2.035214250054911411e-01,0.000000000000000000e+00
-2.889819539617918931e+00,-1.373456096050286457e-01,2.470924099895660531e-02,0.000000000000000000e+00
-2.746437197308736700e+00,-3.111243157519927305e-01,3.767197528530075168e-02,0.000000000000000000e+00
-2.728592981831317044e+00,3.339245635684536806e-01,9.622969977460874014e-02,0.000000000000000000e+00
```

代码要求&实现内容要求

- 编程语言不限，但是尽可能使用主流语言
- 不允许使用第三方编好的k-means库，必须自己实现k-means的迭代

- 对读入的数据进行可视化
- 尝试clusters的个数为2，3，4
- 对不同clusters个数聚类后不同的结果进行可视化

- 下面左图是原始数据的可视化，右图是聚类后的数据，clusters=3, 仅供参考



报告要求

- 使用Word，Pages, LaTeX或者markdown等编写都可以，但最后提交时转成PDF文件格式
- 报告的内容
  - 自己运行结果的截图
  - 对运行结果进行分析
- 报告的长度控制在五页以内（不要在报告中附上代码）

提交

- 作业提交位置
  - ftp://public.sjtu.edu.cn username: shen_yao password: public
  - 提交到ftp中/upload/CS426/hw4/ 目录下
- 作业提交时间
  - ddl: 5月18号23:59:59
  - 晚交惩罚：每超时24小时，该次作业总分扣除20%成绩，不满24小时按照24小时计算，5月21日23:59:59之后提交的作业一概不接收。
  - 时间根据ftp服务器接收到文件的时间为准。
- 作业命名规则
  - 学号1_姓名1_学号2_姓名2_hw4.zip
  - 压缩包内部文件结构为/report.pdf, /src/你的代码文件

评分标准（满分10分）

- 实现k-means clustering 3分
- 4张图一张一分
- 对实验结果进行分析（分析不同的clusters数对聚类结果的影响, 分析初始时不同的seeds对结果的影响, 分析实际运行时间与clusters数以及sample个数的关系） 3分

完成作业过程遇到任何问题，请发邮件到cs_jerrychen@sjtu.edu.cn