Big Data Processing

Homework 5

要求

- 实现KNN算法和Perceptron算法
- 使用两种算法对iris flower data set进行分类
- 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，此处只选用Sepal Length和Petal Length两维数据，并只对两类数据进行分类（setosa和virginica），

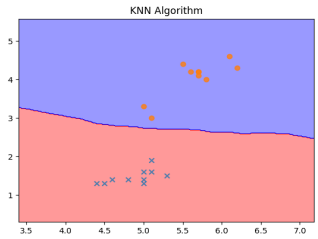
  课程主页上会提供直接可用数据（train数据用于训练，test数据用于测试）
- 数据介绍如下，第一行是header部分，后续每行为具体数据，一共100个samples，前两列是降维后的features值，第三列代表数据对应的label，label为-1,1分别对应setosa, virginica。
- 每行之间使用\n分割，每列之间使用逗号分隔。

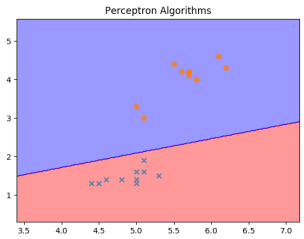
```
feature1 ,feature2,label
5.1,1.4,-1
4.9,1.4,-1
4.7,1.3,-1
4.6,1.5,-1
5,1.4,-1
5.4,1.7,-1
4.6,1.4,-1
```

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库，必须自己实现KNN算法和Perceptron的迭代过程
- 对KNN算法和Perceptron算法的分类结果可视化

- 下左图是KNN算法分类的test结果，右图是Perceptron分类的test结果，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/hw5/ 目录下
- 作业提交时间
  - ddl: 5月25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28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_hw5.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：KNN Algorithm 3分 Perceptron Algorithm：3分
- 两张结果图1张一分
- 对实验结果进行分析（对比分析KNN Algorithm和Perceptron Algorithm的适用场景和局限） 2分

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