ex3/4 实验时间：22/11/07

探索探索再探索 胜利在望✌～

以下实验任选1个完成✅

1.中文手写数字分类

(0)实验准备:文件下载与上传

(1)feture_hog.py文件 //基于skimge实现特征提取

(2)上传hdfs

(3)rdd_logistic.py文件 //基于pyspark.mllib实现回归多分类

(4)spark/spark集群提交任务.

2.鸢尾花聚类

(0)实验准备：安装matplotlib,文件下载与上传

(1)补充iris.py　//数据读取，处理，输出

(2)补充kmeans.py　//实现kmeans功能

(3)spark/spark集群提交任务

3.淘宝回头客预测

(0)实验准备：安装matplotlib,文件下载与上传

(1)数据预处理与上传 //运行sh脚本

(2)补充forecast.py　//基于pyspark.mllib实现SVM

(3)spark/spark集群提交任务

一些文档指路

遇到看不懂or不会用的算子，自主查阅官方文档可以减轻代码理解的困难～👇

Sklearn

https://www.sklearncn.cn/

涉及的算子：train_test_spilt

SKimage

https://scikit-image.org/docs/stable/api/api.html

涉及的算子：skimage.feature.hog

PySpark

https://spark.apache.org/docs/latest/api/python/reference/pyspark.html

涉及的算子：

- pyspark.SparkContext.textFile
- pyspark.SparkConf

PySpark RDD

https://spark.apache.org/docs/latest/api/python/reference/pyspark.html#rdd-apis

涉及的算子：map,filter

PySpark SQL

https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/index.html

涉及的算子：

- pyspark.sql.functions.lit
- pyspark.sql.SparkSession

https://scikit-image.org/docs/stable/api/skimage.feature.html#skimage.feature.hog)

PySpark MLlib(RDD-based)

https://spark.apache.org/docs/latest/api/python/reference/pyspark.mllib.html

涉及的算子：

- pyspark.mllib.regression.LabeledPoint
- pyspark.mllib.linalg.Vectors
- pyspark.mllib.classification.SVMWithSGD
- pyspark.mllib.classification.LogisticRegressionWithLBFGS
- pyspark.mllib.clustering.KMeans
- pyspark.mllib.classification.SVMWithSGD