

#KNN-hadoop

文件结构

```
|——Knn.java // hadoop的mapReduce程序
|——knn_train.txt //训练数据
|——Knn.jar // 运行的包
|——input // 测试数据文件夹
|   |——knn_test.txt // 测试样例2的测试文件
|——run // 生成的class所在的文件夹
|   |——Knn.class
|   |——Knn$DistanceTuple.class
|   |——Knn$KnnMapper.class
|   |——Knn$KnnReducer.class
|   |——Knn$Point.class
|——output // 测试结果文件夹
|   |——part-r-00000 // 测试样例2的测试结果
```

Knn.java代码逻辑

```
|——Point 类 // 用于记录每一个数据。包括数据向量形式，数据的类别..
|——DistanceTuple类 //二元组（类别，距离）的结构，方便KnnReducer对距离进行排序，获取最近的类别
|——readTrainingPoints函数 // 读取训练数据
|——KnnMapper类 // 计算每一个测试样本到所有训练样本的距离。
|——KnnReducer类 // 对每个测试数据的距离数据进行递增的排序，确定前K个点所在类别的出现次数，找到出现次数最多的类别。
```