Advanced Data Mining
Homework 2
Documentation for Map-Reduce KMeans

Yudong Guang (PID 4012835)
yguan004@fiu.edu

February 2, 2014

1 Program functions

The program implements the KMeans clustering method in Hadoop Map-Reduce.

2 Folder structure of this submit

- The folder 'KMeansMRSrc' contains all the java source code
- The test data is in the 'data' folder in the submit. The folder 'rawdata' contains the cluster generator and its original outputs, including a plot that shows the distribution. The folder 'input' and 'nextCenters' can be put to HDFS for test described in 'Usage' section.
- The KMeansMR.jar is a jar that can be used directly, however it is recommended to run through script 'runjob.sh'. For more details, please refer to 'Usage' section.

3 Cluster generator

Along with the submit I have attached a cluster generator under folder 'data/rawdata' written in python, called 'ClusterGen.py'. The usage of the program is like this:

```
python ClusterGen.py xrange yrange number_of_clusters number_of_points number_of_noise
```

This generator can simply generate points in 2-D space, with assigned xy-range and number of points & noise to generate. It will also draw a plot for you to see if the distribution of the data satisfy your requirements. For KMeans clustering method, I chose the data that are distributed homogeneously, with about the same density.

PS. It depends on python package 'matplotlib'. The data I used in the test comes from command:

```
python ClusterGen.py 100 100 4 500 50
```

4 Usage

The usage of the KMeansMR is like this:

1. unzip the zip file
2. create a folder $JOBHOME in your HDFS, suppose your $JOBHOME folder is '/user/hduser/kmeans'
3. create an input file folder '$JOBHOME/input', and upload the input file into it
4. create a folder called '$JOBHOME/nextCenter/iteration_0', and upload the initial center file to it. Center filename must contain 'centers'
5. change file 'runjob.sh', modify variable 'homedir' to your $JOBHOME directory, modify variable 'numiter' to the number of iterations you want to run
6. run command './runjob.sh'. For each run, you don't have to delete the files generated by previous runs because the 'runjob.sh' will do that for you. After the job is finished, it will create folder structure in HDFS like this:

```
$ fs -ls /user/hduser/kmeans
Found 7 items
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:32 /user/hduser/kmeans/currentCenter
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:28 /user/hduser/kmeans/input
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:32 /user/hduser/kmeans/nextCenter
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:31 /user/hduser/kmeans/output1
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:32 /user/hduser/kmeans/output2
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:32 /user/hduser/kmeans/output3
drwxr-xr-x   - hduser supergroup  0 2014-02-02 10:32 /user/hduser/kmeans/output4
```

Where, for a 4-iteration run, '/user/hduser/kmeans/output4' contains the newest result of clustering, and '/user/hduser/kmeans/currentCenter/iteration_3' contains the newest centers used in mapper.

5 Test result

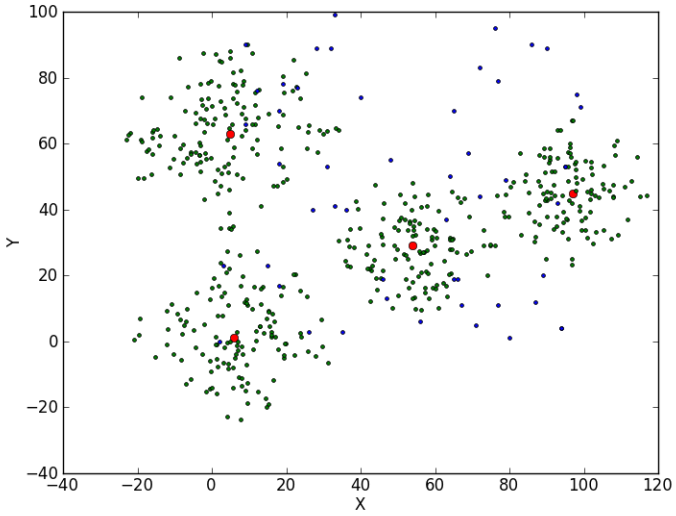

Figure 1: Test data distribution

The red dots represent the centers, the green points are normal points, and the blue points are noise. However in the test I didn't run with noise.

The four centers in the data are:

```
0,54,29
1,6,1
2,5,63
3,97,45
```

I chose the first four points in the dataset as the initial centers:

```
0,101.557954162,39.3268885464,3
1,5.77714596146,78.1036893852,2
2,-4.95814167901,3.35320505419,1
3,45.8160196793,21.6864629838,0
```

After four iterations, the newly found centers are:

```
0,6.915074975090942,1.987097505079365
1,95.81053005873338,45.495967391690385
2,54.42369762973394,28.19728607879002
3,3.2239932359748376,63.785202176776956
```

As we can see, all the new centers are pretty close to the real centers.