CS 481, Homework 2

Out: February 22, 2017, Due: March 08, 2017, Total: 100

Note:

- This homework will carry 8 points towards your final score.
- If a question asks you to write a code, you need to submit a working code through oncourse submission site. It is also your responsibility to test that the code runs in pegasus.cs.iupui.edu machine.
- Start early and if you need help, post your questions on piazza or use instructor/TA's office hour.
- You can use python, Matlab, java or C++ programming language for this work

Questions

1 (80 points). Implement k-Means algorithm for clustering (Algorithm 13.1 in the textbook). Your program should take a comma-separated data file, and the number of clusters(k-value) as command line parameters. There is one optional parameter, which is another filename which lists the id of initial centroids. For example, we may run your program (`mykMeans`) as below:

```
./mykMeans datafile.txt 3
```

In that case, it should cluster the data points in `datafile.txt` into 3 clusters. The data file `datafile.txt` lists one data point in each line. The values of a data points along different dimensions are separated by a comma. No id is given for a data point, but you must assume that the id of a data point is the same as its line number in the input file assuming that the line numbering starts from 1. For initial centroid you should use k random data points. However, we can also run your program with the optional parameter as below:

```
./mykMeans datafile.txt 5 centroidfile.txt
```

the `centroidfile.txt` contains the following integers for a clustering task with k=5:

```
5
19
201
371
390
```

In that case you must use the points with id $5, 19, 201, 371$,and 390 as your initial centroids. For all cases, use $\epsilon = 0.001$ as your stopping condition.

Your program output should consists of the following information:
1. The number of data points in the input file, the dimension, and the value of k
2. The number of iterations the program took for convergence
3. The final mean of each cluster and the SSE score (sum of square error)
4. The final cluster assignment of all the points.
5. The final size of each cluster.

2 (20 points). Execute your program on the `iris.txt` dataset (will be available in piazza resource section). Run it for at least 10 times with different random initializations of centroid (make sure to randomize your random seeds so that different runs of the program starts with a different random set of centroids). Now, evaluate the clustering using purity-based evaluation (see 17.1.1 in textbook) and report the best purity score. Note that for purity based evaluation you need to know the true label of a data points. Here is the true label assignments on the `iris.txt` dataset: 1-50: class 1, 51-100: class 2, and 101-150: class 3.

Deliverables

For question 1, submit a tarzipped folder containing all source code, a makefile (if needed), and a README. For question 2 submit a txt or pdf file.