CSE 4334-002: Spring 2020
Programming Assignment 3
Total Points : 100

(Assignment courtesy of Prof. Vassilis Athitsos)

Guidelines:

The assignment should be submitted via Canvas. Submit a file called P3.zip, containing your answers.pdf document and your code. Your written answers should be in a document called answers.pdf.

For each task requiring code to be submitted, your zip file should contain a folder containing:

- The code for that task.
- A text file called README.txt that fully describes the commands needed to compile (if needed) and to run your code.

Make sure that you use at least 64-bit floating point numbers in your calculations.

Task (100 points)

In this task you will implement k-means clustering

Your zip file should have a folder called k_means, which contains your code and the README.txt file.

Command-line Arguments

Your program will be invoked as follows:
k_means_cluster <data_file> <k> <iterations>

The arguments provide to the program the following information:

- The first argument, <data_file>, is the path name of a file where the data is stored. The path name can specify any file stored on the local computer.
- The second argument, <k>, specifies the number of clusters.
- The third argument, <iterations>, specifies the number of iterations of the main loop. The initialization stage (giving a random assignment of objects to clusters, and computing the means of those random assignments) does not count as an iteration.

The data file will follow the same format as the training and test files in the UCI datasets directory with the assignment. A description of the datasets and the file format can also be found on the directory.

As the description states, **do NOT use data from the last column (i.e., the class labels) as features**. In these files, all columns except for the last one contain example inputs. The last column contains the class label.

Implementation Guidelines

- Use the L_2 distance (the Euclidean distance) for computing the distance between any two objects in the dataset.

Output

After the initialization stage, and after each iteration, you should print the value $E(S_1,S_2,...,S_K)$. For the formula you can have a look at the clustering slides(slide 27) of Prof. Vassilis Athitsos. Remember, in the E value calculation, you are using euclidean distance calculation in stead of SSE(sum of squared error) calculation.

The output should follow this format:

```
After initialization: error = %.4f
After iteration 1: error = %.4f
After iteration 2: error = %.4f
...
```

Output for answers.pdf

In your answers.pdf document, you need to provide the complete output for the following invocations of your program:

k_means_cluster yeast_test.txt 2 5
k_means_cluster yeast_test.txt 3 5

Grading

- 75 points: Correct implementation of k-means clustering.
- 25 points: Following the specifications in producing the required output