CSE 5334

Programming Assignment 3
Total Points : 100

Task (100 points)

In this task you will implement k-means clustering on UCI_datasets. You can't use library functions to implement kmeans clustering.

Dataset Description: The data file will follow the same format as the training 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. In these files, all columns except for the last one contains different features. The last column contains the class label. **DO NOT use data from the last column (i.e., the class labels) as features**.

K-means Algorithm: The basic K-means algorithm works as follows:

1. Initialize 'K', number of clusters to be created.
2. Randomly assign K centroid points.
3. Assign each data point to its nearest centroid to create K clusters.
4. Re-calculate the centroids using the newly created clusters.
5. Repeat steps 3 and 4 until the centroid gets fixed.

Initialization:

We know that k-means clustering suffers from Initial Centroid Problem. Therefore, try different initialization approaches and pick the best initialization approach to deal with this problem. You should implement the initialization method by yourself and not use any library functions. For random state, use 0.

Tasks:

Your program should take one argument <data_file>, which is the path name of a file. The path name can specify any file stored on the local computer.

Given a dataset (yeast/pendigits/satellite), initialize the K-means clustering and run the K-means clustering for a range of K values (2-10). Then, print the Error (**in %.4f format, 4 places after the decimal)** for each k value after 20 iterations.

The Error is calculated as follows:

$$E(\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_K) = \sum_{k=1}^{K} \sum_{x_n \in \boldsymbol{S}_k} \text{Euclidean Distance } (x_n, \mu_k)$$

Finally, draw a graph that shows Error values (on the y-axis) corresponding to the different values of K (on the x-axis)

Example Output:

```
For k = 2 After 20 iterations: Error =
For k = 3 After 20 iterations: Error =
For k = 4 After 20 iterations: Error =
For k = 5 After 20 iterations: Error =
For k = 6 After 20 iterations: Error =
For k = 7 After 20 iterations: Error =
For k = 8 After 20 iterations: Error =
For k = 9 After 20 iterations: Error =
For k = 10 After 20 iterations: Error =
```

Displays the Error vs k chart

Grading:

- 80 points: Correct implementation of k-means clustering
- 20 points: Following the specifications in the assignment