CSE 5334 : Spring 2023
Programming Assignment 3
Total Points : 100

Task (100 points)

In this task you will implement k-means clustering on UCI_datasets.

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, we will use a strategy called "random partition method", where we randomly assign each point in the data to a random cluster ID. Then, we group the points by their cluster ID and take the average (per cluster ID) to yield the initial points.

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 or pendigits or satellite), initialize the K-means clustering algorithm using random partition method and run the K-means clustering for a range of K values (2-10). Then, print the SSE error (**in %.4f format, 4 places after the decimal)** for each k value after 20 iterations.

SSE 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 SSE values (on the y-axis) corresponding to the different values of K (on the x-axis)

Example Output:

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

Displays the SSE vs k chart

Grading:

- 90 points: Correct implementation of k-means clustering for 3 different datasets.
- 10 points: Following the specifications in producing the required output