CSE 572
Data Mining
Instructor- Prof. Arunabha Sen

Assignment 3

Total marks : 20

Submission Deadline : 03.18.2019 11:59 pm AZ time.

- For submission, you should submit codes and a PDF report containing the results in a zipped file (only one submission per group). The PDF file should contain names of all the members. The zipped file name should be in the following format:

  GroupName_GroupID.zip [eg.- DM_12.zip]

- Refer to the 'group formation sign-up' sheet in the blackboard for group ID and GroupName (Group name should be the First name of Member 1. Group ID can be obtained from the first column).

- For coding you can use both Matlab and Python. For Matlab and Python codes, include .m and .py files in the zipped folder respectively.

In this Assignment, you need to implement three algorithms **from scratch**:

Algorithm (a): K-means (Initialize k cluster centers by randomly picking up k points among all data points in the dataset).

Algorithm (b): A clustering technique of distributing all the data-points in the dataset into k groups (clusters) such that diameter of the largest cluster is minimum among all possible ways of creating k clusters out of these data-points (Diameter = Euclidean distance between the two farthest points in a cluster).

Algorithm (c): Spectral-Clustering (Use a Gaussian kernel for computing affinity score between two points. Use k-nearest neighbor for graph construction **(set k=5)**. You may use libraries for sub-tasks in spectral-clustering, for example- computing *diagonal Degree matrix, Eigen-vectors & Eigen-values*).

Task 1)

Implement Algorithms (a), (b) and (c) on Dataset-1 ('Dataset_1.csv' contains 3 columns - 1st, 2nd columns represent features and 3rd column represents class information for each observation). Set **number of clusters = 2**.

Generate the following plots:
(i) Plot all the data values (first 2 columns of 'Dataset_1.csv') as points in a 2-Dimensional space. Represent them in different colors according to their class labels.
(ii) Plot the clustered results. Total number of plots = 3 (1 plot for each algorithm).

Deliverables:

[1] Code. **[2]** 4 Plots.

Task 2)

Implement Algorithms (a), (b) and (c) on Dataset-2 ('Dataset_2.csv' contains 3 columns - 1st, 2nd columns represent features and 3rd column represents class information for each observation). Set **number of clusters = 2**.

Generate the following plots:
(i) Plot all the data values (first 2 columns of 'Dataset_2.csv') as points in a 2-Dimensional space. Represent them in different colors according to their class labels.
(ii) Plot the clustered results. Total number of plots = 3 (1 plot for each algorithm).

Which algorithm gave the worst performance? Why.

Deliverables:

[1] Code. **[2]** 4 Plots. **[3]** Justification.

Task 3)

Implement Algorithms (a), (b) and (c) on Dataset-3 ('Dataset_3.csv' contains 3 columns - 1st, 2nd columns represent features and 3rd column represents class information for each observation). Set **number of clusters = 3**.

Generate the following plots:
(i) Plot all the data values (first 2 columns of 'Dataset_3.csv') as points in a 2-Dimensional space. Represent them in different colors according to their class labels.
(ii) Plot the clustered results. Total number of plots = 3 (1 plot for each algorithm).

Which algorithm gave the best performance? Why.

Deliverables:

[1] Code. **[2]** 4 Plots. **[3]** Justification.