CS4038D Data Mining
Assignment No: 1
Due Date: 28.09.2020

Topic: Familiarization of similarity and dissimilarity measures and implementation of a clustering algorithm using a standard data set.

1) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters:
   A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9):
   The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only
   (a) The three cluster centers after the first round execution
   (b) The final three clusters

2) Write a program to implement k-means clustering algorithm by using Iris data set (available in UCI Machine learning repository) and find the followings:
   i) Clusters of the Iris data set (final clustering solution).
   ii) Sum of the Intra-Cluster Distances (SICD) or Sum of the Squired Error (SSE) values of the obtained clustering solution.
   iii) Graphical representation of the obtained clusters.
   iv) Give 150 iterations of the implemented k-means by changing the initial centroids and see the changes in the clustering solution (SICD values).
   v) Plot iteration vs. SICD values.

Students are requested to take care of the following points:

1) Download the dataset from UCI repository, do a shuffling of the records/ tuples/rows/objects/data points of the dataset such that same class data points or objects are not placed one after another and then remove the class label attribute.
2) Please don't use inbuilt functions available in python or in some other platforms for distance measures and coding etc.
3) Save your solution document as a .pdf file. (Format: "Reg.NumberYour NameExercise No.pdf" (Ex: B190585CS JohnE01.pdf).
4) Submit your pdf file to Eduserver before deadline (**28.09.2020, 23.59 PM)**.
5) Late submission may incur penalty of **-0.5** marks.

6) Evaluation will not be conducted for those who don't submit their solution on time.

5) Viva (Interactions) will be conducted through WebEx/Google meet. (Share your screen and show your execution)

6) Due to unpredictable situations, students who are not able to appear for online viva sessions, may get chance to attend the oral viva through mobile phone.

   Note: Only genuine cases will be entertained and students will have to get prior permission from me for the same.

7) WebEx join meeting link details will be sent to the students one day earlier.

8) You may use any programming language of your choice.

9) Please take snapshots of all your executions and put them in a word file and save it as a .pdf file for submission. If any doubts on the uploaded marks and uploaded exercises, please mail me (praneshdas@nitc.ac.in).

****