भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 360: Machine Learning Lab
Evaluation Assignment 3
Total Marks:30

Instructions: This assignment is for evaluation and marks will be awarded. You need to complete execution by 12 PM. The saved results folder (consisting .csv files, plots, word file) has to be submitted through a Google form, which will be shared by Teaching Assistant.

1. Download the "New York City Taxi Trip Duration" dataset from mail.
   Write a program to do the followings:
   (a) Read the dataset.
   (b) Extract the "pickup_latitude" and "pickup_longitude" features.
   (c) Normalize the features.
   (d) Plot a scatter plot using the two attributes and visualize the presence of dense or sparse regions of trip/ pickup, and analysis and report your findings.
   (e) Apply DBSCAN clustering to detect regions with maximum pickup. Consider the followings for this:
      - minpts → minimum taxis to form a cluster. Possible values to consider are {5, 10, 15}.
      - epsilon → maximum distance to form a cluster(in degrees or meters). Apply K-distance graph , w.r.t. the optimal minpts-value obtained from above, to determine the optimal value for *epsilon*.
   (f) For each generated cluster, determine number of pickup points, cluster centroids and average trip duration, and report the results.
   (g) From the results of (d), detect the outliers / noise points (pickup points), and report the results.
   (h) Analysis the results of 1.(e), 1.(f) and 1.(g) to determine the regions with high or low demand of trips and report your analysis.
   (i) Determine the quality of the generated final clusters using Silhouette score.

2. Download the Pima Indians Diabetes Database from mail.
   Write a program to do the followings:
   (a) Read the dataset.
   (b) Extract the "Glucose", "BMI", "BloodPressure", and "Age" features.
   (c) Normalize the features, if necessary.

(d) Apply Fuzzy C-means clustering technique, with C=3 clusters, denoting as below to group the patients into any of these clusters
Cluster 1: patients with low risk of diabetes.
Cluster 2: Patients with borderline risk of diabetes.
Cluster 3: patients with high risk of diabetes.
Consider, fuzziness parameter. m-values as $\{2,4\}$.

(e) After convergence, assign the patients to cluster with highest membership value, and then, visualize the generated clusters, and determine the Silhouette score for each value of fuzziness parameter. m.

(f) Compute the cluster centroids and report it.

(g) Considering the best case (based on Silhouette score), for 10 random patients, check the membership values to each cluster, and based on that, report the clinical interpretation.

(h) Determine the patients with highest membership ≤ 0.5. Consider this gives a scenario of uncertain clustering.

(i) Based on the value obtained from (g), compute the followings:

1. Number of patients with uncertain clustering.
2. Average of clinical parameters (”Glucose”, ”BMI”, ”BloodPressure”), and how these are differing from patients with certain clustering (i.e., highest membership ≥ 0.8).

(j) Report and analysis the results of 2(e)- 2(i).