Machine Learning: Programming Assignment 4: K-Means Clustering

Problem Statement:

Write a program to cluster a set of points using K-means. Consider, K=3, clusters. Consider Euclidean distance as the distance measure. Randomly initialize a cluster mean as one of the data points. Iterate for 10 iterations. After iterations are over, print the final cluster means for each of the clusters.

Use the ground truth cluster label present in the data set to compute and print the Jacquard distance of the obtained clusters with the ground truth clusters for each of the three clusters.

Data Set Description:

Data Filename: *data4_19.csv*

The data set contains 150 data points, there are three clusters where each cluster refers to a type of iris plant. The first four columns represent the attributes listed below. Note that only the first four columns should be used as attributes. The last column is the ground truth cluster name and is to be used for evaluating the cluster quality.

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. Ground truth cluster name:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Submission Guidelines:

You may use one of the following languages: c/C++/Java/Python. You should name your file as <rollnumber_4.extension> (e.g., 14CS10001_4.c). Your program should be standalone and should not use any *special purpose* library. numpy may be used. You should submit the program file only. The submitted single program file *should* have the following header comments:

Roll # Name # Assignment number # Specific compilation/execution flags (if required)

*Please submit the program in moodle by **November 11, 2019 midnight** (hard deadline). Copying from friends/ web will lead to strict penalties.*