Machine Learning: Programming Assignment 7: K-Means Clustering

Problem Statement:

Write a program to cluster a set of points using K-means. Consider, K=2, clusters. Also, 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 numbers for each of the data points.

Data Set Description:

Data Filename: *data7.csv*

Data File Format: Boolean input attributes (x1, x2, … , x8) in first 8 columns. There are 20 training instances.

Please STRICTLY follow the program input/output format specified below.

Input Format: Assume the data files data7.csv is present in the same directory and contains the training and test data. Thus, your program should not require any input from user and should read from these files. Strictly use these filenames only.

Output Format: The cluster labels (1/2) for the data points exactly in the order in which the instances are present in the data file. Put a blank space between printed the cluster labels. (e.g., output 1 1 2 2 …, if the cluster labels are - Data Instance 1: 1, Data Instance 2: 1, Data Instance 3: 2, Data Instance 4: 2 …). Output, in above format, should be printed to the file: *rollnumber_7.out* (e.g., 14CS10001_7.out). Strictly use this filename format.

Submission Guidelines:

You may use one of the following languages: c/C++/Java/Python. You should name your file as <rollnumber_7.extension> (e.g., 14CS10001_7.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 and not the output/input file. 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 7, 2018 midnight*** *(hard deadline). Copying from friends/web will lead to strict penalties.*