Machine Learning: Project 3: K-Means Clustering

Problem Statement:

We have data for recent customers of a supermarket. We would like to segment the customers into a number of groups for targeted marketing. Write a program to cluster a set of points using the K-means clustering algorithm. You can choose whatever distance function you want. You may also use some cluster quality index to find out the optimal value of K. Similarly, you may use any initialization technique.

Data Set Description:

Data Filename: *Project3.csv*

Data File Format: Attribute descriptions are mentioned in the file P3_data_legend.xlx

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 3 …, if the cluster labels are - Data Instance 1: 1, Data Instance 2: 1, Data Instance 3: 2, Data Instance 4: 3 …). Output, in above format, should be printed to the file: *rollnumber_P3.out* (e.g., 20CS10001_P3.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_P3.extension> (e.g., 20CS10001_P3.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. You should also submit a brief report (2-3 pages) outlining your approach. 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 MS Teams by **April 3, 2022 midnight** (hard deadline). Copying from friends/web will lead to strict penalties.*