CIS 635 Data Mining

Homework 9

Description

In this homework you will be writing an R script to implement the kmeans clustering algorithm. You will also be using the kmeans in R to describe clusters.

Instructions

Part 1 – analysis

The data file hw09data.txt contains information about 50000 people and their yearly spending by category. The data has 8 natural, roughly equal sized clusters. You are tasked with finding the clusters and describing them. It will take some modifications to the data in order to correctly identify the groups. You should use the techniques discussed in lecture. A high quality description of the clusters is one that is clear, succinct and easy for a neophyte (someone without knowledge of data mining) to understand.

Part 2 – R script

Write a script to implement the kmeans clustering algorithm. It should use two parameters, the data matrix and the number of clusters desired. You should make use of the r functions that are posted under documents/R commands. Suggestions:

- name your function something other than kmeans since there is already one with that name. Maybe use myKmeans.
- be aware that my functions are not as efficient at the ones built into R so it will take much longer to get results with your own kmeans algorithm compared to the one from the stats library. I would suggest printing out the sse each iteration so that you can track how the algorithm is progressing.
- After you have finished, and you get the clusters you can call calcCent one last time to see the centers. You might want to compare the results of your own algorithm to the one from stats. Don’t worry if you get different centroids – remember that kmeans is dependent on the initial centroids.