CS685:  Special Topics in Data Mining

Homework 2: Due March 3rd

**Goal:**  This homework will reinforce the understanding of two basic clustering algorithms - k-means and hierarchical clustering algorithms. Through the homework, you will learn how to use k-means and hierarchical clustering and how to evaluate the clustering results including determining the number of clusters.

**Description of the homework:**

Your homework will be an application of existing algorithms. Only the report needs to be submitted. Your report should contain answers for the following questions.

**Algorithms:**

Two algorithms you are supposed to work with include K-means and hierarchical clustering. Both algorithms are available as functions in SciPy (http://docs.scipy.org/doc/scipy/reference/cluster.html).

Note: You can use other software, such as R and Weka, as long as you do the same analysis

**Datasets:**

User Knowledge Modeling Data Set

            [http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling](http://archive.ics.uci.edu/ml/datasets/User%2BKnowledge%2BModeling)

The dataset contains the information of 403 users in order to identify the users’ knowledge class. Each user is characterized by 5 attributes (columns in the excel sheet) and their knowledge class is reported as the last column in the excel sheet.

Please Read through the data description carefully to understand the particular application.

**Questions to answer in the report:**

**1.**     **Write a script to compare two sets of clustering results**

Assume  C1 and C2 are two sets of clustering results on the same set of data. C1 = (S1, S2, …, Sn) contains n clusters, where each S represents a cluster and C2 = (T1, T2, …, Tn) contains n clusters as well, where each T is a cluster.

In order to compute the similary between C1 and C2 , you will need to do the following.

      sim = 0

   For each S in C1,

        sim = sim + the highest similarity between S and any cluster T in C2  measured by [Jaccard Coefficient](http://en.wikipedia.org/wiki/Jaccard_index).

   End

   avgsim = sim / n;

**2.**     **Please follow the instructions in SciPy for K-means clustering to do the following.**

a)     Write the script to do k-means clustering. Check out the following function.

 ---Kmeans

b)    run the program 10 times  with  k = 4. Check how the clustering algorithm changes by checking out sumdistance. Please report your result.

c)     Use k = 2, 4, 6, 8, 10 and check how the sumdistance changes. Please report your result.

d)    Pick one of the results as the best (please argue why), and compare it with real clusters.

**3.**     **Please follow the instructions in matlab for hierarchical clustering to do the following**

a)     Write the script to do hierarchical clustering. The following functions might be useful

--load

--Pdist

--Linkage

--Squareform

--Dendrogram

--Cluster

b)    Choose one of the existing dissimilarity measures to compute the dissimilarity matrix for hierarchical clustering. Please reason whether it is appropriate for the particular application and suggest another that might also work.

c)     Run the hierarchical clustering on the datasets using single linkage, complete linkage and maximum linkage.  Generate 6 clusters for each run and Compare the change of the clustering results.

d)    Pick one of the results as the best (please argue why), and compare it with real clusters.

**4.**     **Compare the best (3(d)) of the hierarchical clustering results generated in (c) with the best generated k-means clustering when k=6 (2(d)).**

**5.**     **Please include all your matlab scripts for submission.**

**6.**     **Please discuss the pros and cons of k-means and hierarchical clustering. Give examples where either k-means or hierarchical clustering will work better than the other.**

**7.**     **Please report your impression with the clustering algorithms and your experience with the assignments.**