Intelligent Data Analysis

Homework #3

Due Date: April 5th, 2017, 7PM

Consider the data file attached with this homework. It contains scores for some students in four different subjects (Physics, Maths, English, and Music). Perform the following tasks with this data set.

1. Perform k-means clustering with this dataset for values of k to be 5, 6, 7, 8, 9 and 10. For each case of k run the clustering algorithm with three different initial cluster centers and select the one with the lowest total SSE value of all clusters in the clustering. Report the following in the submitted work: (Use Matlab kmeans function or any other similar toolbox)
   1. Show the cluster centers, SSE values of the clusters, and the total SSE value for the clustering for each value of k.
   2. Plot the total SSE value against the values of k.
   3. Show a plot of the silhouette coefficients for the data points in any two of the clusterings. (Each value of k results in one clustering)
   4. How many clusters would you form in this dataset? Justify your answer. For your choice of the best number of clusters,report the centroids of all the clusters and their SSE values (Call this as Clustering-1).
   5. Generate 100 random 4-dimensional random data points such that each attribute can take values between 0 and 100. With this dataset form the same number of clusters as selected by you in (d) above. Report the centroids and populations of the clusters. Compare the total SSE for this random dataset with the SSE for the clustering of the provided dataset. Compare and comment on the differences between the two total SSE values.
2. Perform hierarchical clustering for the students’ scores dataset. Generate and show dendrograms for the cases (i) Single-Linkage clustering (Clustering-2), and (ii) Complete-Linkage clustering (Clustering-3). Use Euclidean distance for computing distance between data points. Report the following in the submitted work: (Use Matlab functions pdist and linkage, or any other similar toolbox.) Make sure the dendrogram shows all points at its lowest level.
   1. Dendrograms for the two clusterings (Clustering-2 and Clustering-3)
   2. Cluster compositions for each case when we need only four clusters. Write ALL the data points included in each cluster and compute their centroids.
   3. Comment on any differences in the cluster centers and cluster compositions for the two different clusterings as performed in (b) above.
   4. Compute Rand Index for the comparison of Clustering-2 and Clustering-3 and show the counts a, b, c, and d as determined for computing the Rand index. Explain the meaning of each count and why such counts have been obtained for this dataset and their clusterings.
3. Compute Rand Index for the comparison of Clustering-1 and Clustering-2 and show the counts a, b, c, and d as determined for computing the Rand index. Explain the meaning of each count and why such counts have been obtained for this dataset and these clusterings in this comparison.
4. Show the execution tree for the CHARM algorithm for finding all the closed itemsets for the dataset containing the following transactions: ABCDEH, ACDHJM, ABD, ABCDM, BDM, ACDEFJ.
5. For the same data as in #4 above, show execution of the algorithm for finding all the maximal itemsets.