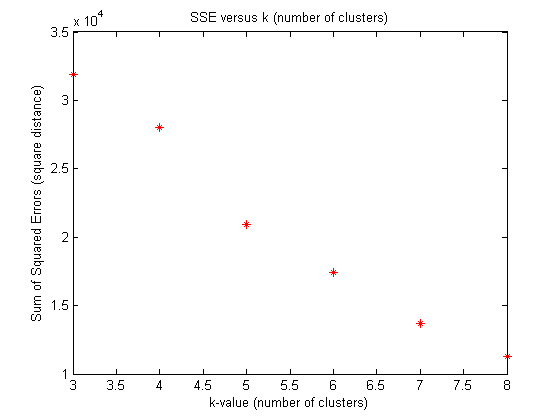
Intelligent Data Analysis

Homework #3

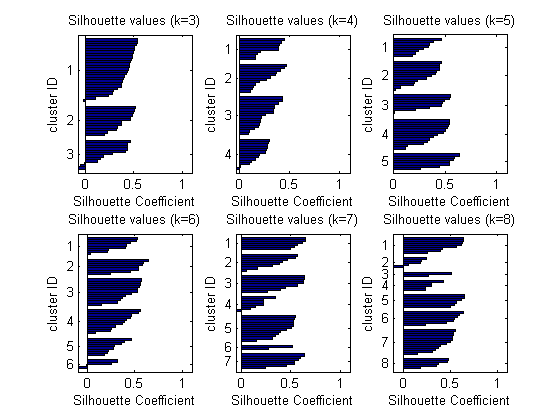
Due Date: Nov 14th, 2016, 7PM

Consider the data file attached with this homework. It contains scores for fifty 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 3, 4, 5, 6, 7, and 8. For each case of k run the clustering algorithm with three different initial cluster centers and select the one with the lowest SSE value. Plot the SSE against the values of k. Report the following in the submitted work: (Use Matlab kmeans function or any other similar toolbox)
   1. A plot of the SSE values against the values of k.

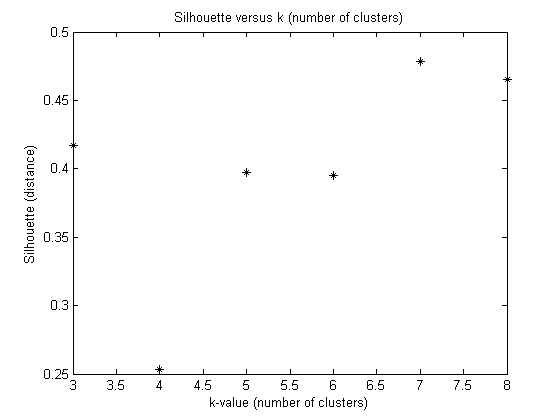


* 1. A plot of the silhouette coefficients for the data points in each clustering. (Each value of k results in one clustering)



* 1. What is the best number of clusters for this dataset? Justify your choice for the best number of clusters.

I would choose k=5 as the best clustering



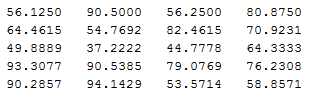
For k = 6, 7, and 8, there are at least a few clusters which only contain a few data points and should be considered anomalous.

K=1 does not look so bad, but it does have some data points which are closer to other clusters than to their own (see coefficients which are less than zero).

To determine whether k=4 or k=5 is a better clustering, let us also look at a total median silhouette for each clustering. Notice that k=4 has a much lower median silhouette than k=5.

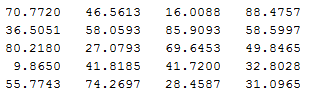
* 1. For your choice of the best number of clusters report the centroids of all the clusters (Call this as Clustering-1).

Every row is a centroid (Physics, Math, English, Music)



* 1. Generate 50 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 (c) above. Report the centroids and populations of the clusters. Compare the SSE for this dataset with the SSE for the provided dataset. Comment on the differences between the two values.

Centroids:



Population of clusters 1, 2, 3, 4, and 5 (respectively) for random data:

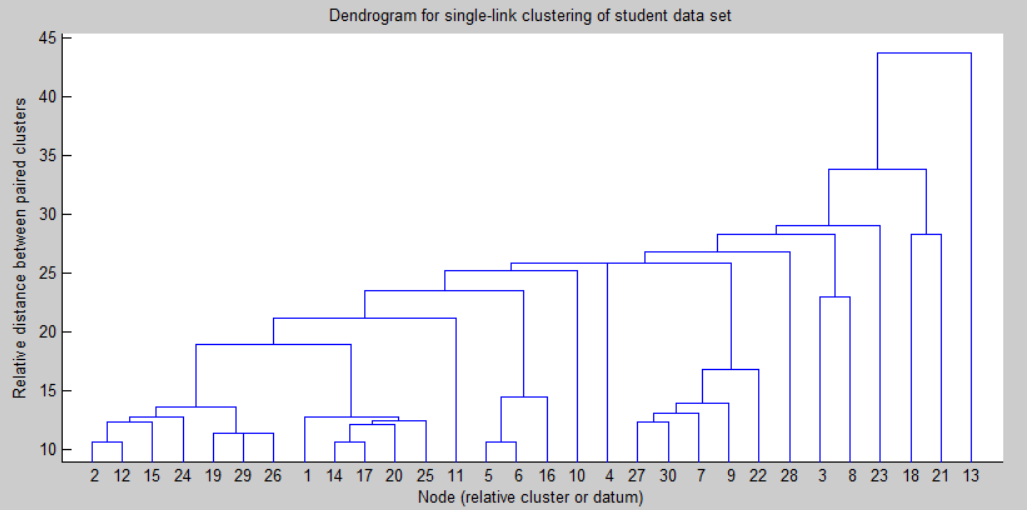


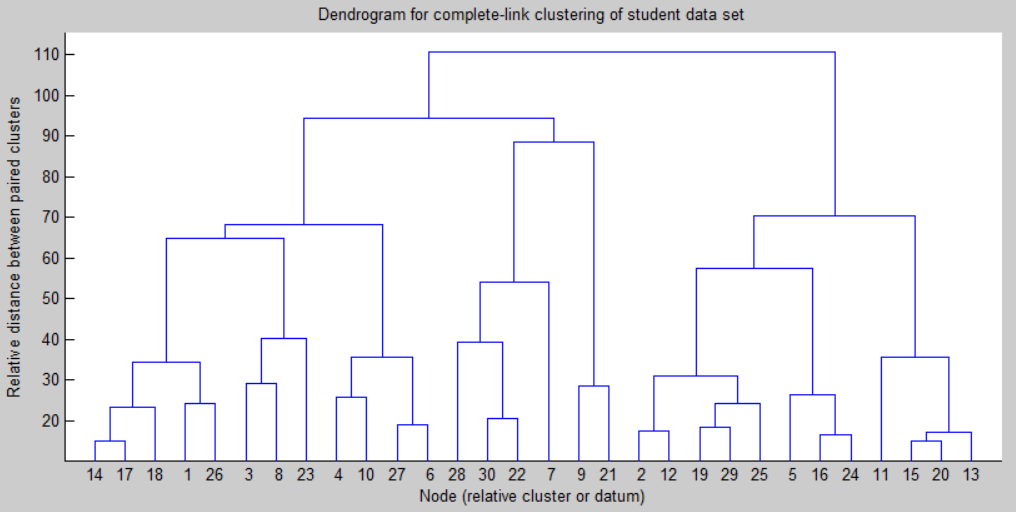
SSE for random data = 7.1622 e4

SSE for real data = 2.0927 e4

It is not surprising that the SSE for real data is lower than the SSE for random data. Because both the random data and real data fall within similar ranges and contain the same number of data, we can conclude from the SSE’s that the random data is less clustered (i.e., more evenly spread out) than the real data. Therefore, we can expect that the real data exhibits more clustering and does not follow a uniform random distribution.

1. 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.)
   1. Dendrograms for the two clusterings (Clustering-2 and Clustering-3)





* 1. Cluster compositions for each case when we need only four clusters. Write the data points included in each cluster and compute their centroids.

Below are the number of data for each cluster (sorted by most populous clusters):

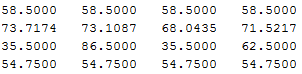
Single link: 46, 2, 1, 1

Complete link: 25, 18, 5, 2

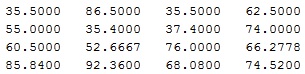
Notice how clusters in the complete link clustering are more evenly distributed

Centroids:

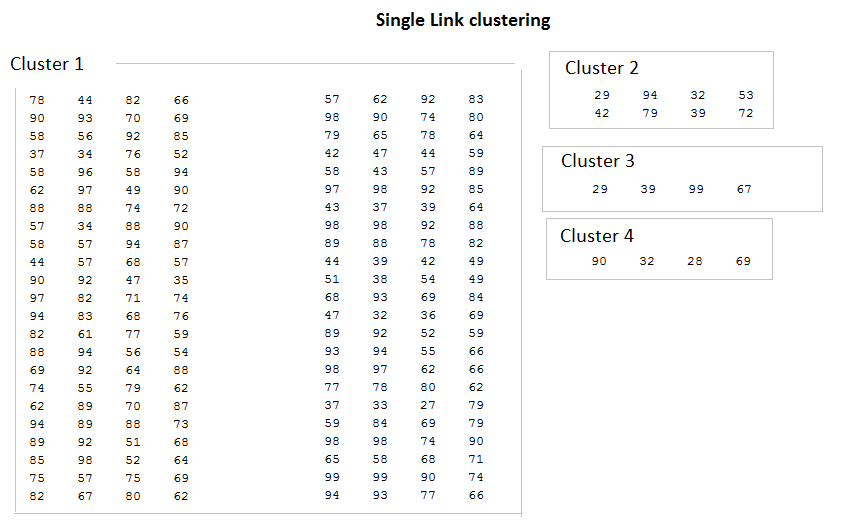
Single link:



Complete link:



Data points included in each cluster (per clustering):



* 1. Comment on any differences in the cluster centers and cluster compositions for the two different clusterings as performed in (b) above.
  2. 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 these clusterings in this comparison.

1. 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.