Clustering

Exercises

1. Mark the following statements as true or false.

   a. k-means clustering assumes that your data is normally distributed.

   b. Hierarchical clustering assumes that your data is normally distributed.

   c. k-means clustering automatically determines the number of clusters.

   d. Hierarchical clustering can make it easy to determine a number of clusters once the procedure is run.

   e. Hierarchical clustering using single linkage is the same as the minimum spanning tree clustering method (from Dr. Healey's text mining notes).

2. Suppose we have four observations and we compute the following distance matrix for them:

$$\begin{pmatrix} 0 & 0.2 & 0.3 & 0.6 \\ 0.2 & 0 & 0.5 & 0.7 \\ 0.3 & 0.5 & 0 & 0.4 \\ 0.6 & 0.7 & 0.4 & 0 \end{pmatrix}$$

   a. Using this distance matrix, sketch the dendrogram that results from hierarchical clustering of these four observations using complete linkage.

   b. Repeat part (a) using single linkage.

   c. Suppose we cut the dendrograms from parts (a) and (b) to create two clusters. Which observations are in each cluster?

3. Explain how the k-means algorithm works.

List of Key Terms

Hard vs. Fuzzy Clustering

Hierarchical Clustering

Single Linkage

Average Linkage

Complete Linkage

k-means Clustering

SSE measure of Clusters