Machine Learning exercise 1: Clustering

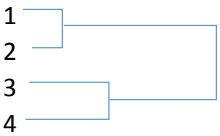
1. . Mark the following statements as true or false.
   a. k-means clustering assumes that your data is normally distributed. F
   b. Hierarchical clustering assumes that your data is normally distributed. F
   c. k-means clustering automatically determines the number of clusters. F
   d. Hierarchical clustering can make it easy to determine a number of clusters once the procedure is run. T
   e. Hierarchical clustering using single linkage is the same as the minimum spanning tree clustering method (from Dr. Healey's text mining notes). T
2. A.

$$\begin{pmatrix} 0 & 0.2 & 0.3 & 0.6 \\ 0.2 & 0 & 0.5 & 0.7 \\ 0.3 & 0.5 & 0 & 0.4 \\ 0.6 & 0.75 & 0.4 & 0 \end{pmatrix}$$

Step 1: obs 1 and 2 form cluster 1

Step 2: obs 3 and 4 form cluster 2

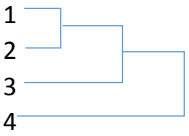
Step 3: cluster 1 and cluster 2 form 1 cluster
Dendrogram:



b. Single linkage
Step 1: obs 1 and 2 form cluster 1

Step 2: cluster 1 and obs 3 form cluster 2

Step 3: cluster 2 and obs 4 form 1 cluster
Dendrogram:



c. Using complete linkage:
Cluster 1: observations 1 and 2
Cluster 2: observations 3 and 4
Using single linkage:
Cluster 1: observations 1, 2 and 3
Cluster 2: Observation 4

3. K-means tries to minimize the sum of squared distances to each point to its cluster centroid. The algorithm begins deciding the number of clusters (k) and the seed points. The distance to each point to each seed is calculated and the points get clustered together according to which seed is the closest. The seeds get relocated in the centroid of each cluster, and the distances are recalculated, forming new clusters. This process continues until the clusters do not change.