STAT 636, Fall 2015 - Assignment 7

Due Wednesday, Nov. 25, 11:55pm central

Online Students: Submit your assignment through WebAssign.

On-Campus Students: Email your assignment to the TA.

1. Consider the matrix of distances for four items

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0 \end{bmatrix}$$

   For each of single, complete, and average linkage:

   (a) List all intermediate distance matrices involved in the hierarchical clustering routine.

   (b) Draw the dendrograms and compare the results of the different linkage methods.

2. Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are as follows:

	Observations	
Item	x_1	x_2
A	5	4
B	1	-2
C	-1	1
D	3	1

   (a) Starting with the initial groups (AB) and (CD), write out all steps to the K-means clustering routine with $K = 2$. Use Euclidean distance on the unstandardized variables.

   (b) Repeat, now starting with the initial groups (AC) and (BD). Compare the results with those in the part (a).

3. Consider the breakfast cereal data in textbook Table 11.9 ($n = 43$ cereals and $p = 8$ variables). In what follows, use Euclidean distance on the unstandardized variables.

   (a) Carry out complete linkage hierarchical clustering of the cereal brands. Report a dendrogram. How many clusters would you say there are? Comment on the composition of those clusters.

   (b) Carry out K-means clustering of the cereal brands, with $K = 3$. Plot the first two principal components and color-code by cluster membership. How do the K-means results compare with the hierarchical clustering results?

4. Consider the pottery data in textbook Table 12.8 ($n = 6$ sites and $p = 4$ variables).

   (a) Using Euclidean distance, carry out multidimensional scaling to find the "best" representation of the data in $q = 2$ dimensions; you can use the `cmdscale` function in `R` for this. Make a scatterplot of the resulting variables, using the site names as plotting characters. Comment.

   (b) Construct a biplot and interpret. How does the biplot compare to the MDS plot in part (a)?