STATS 415 Homework 11

Due Thursday April 12, 2018

Please include your name, uniqname, and lab section (number or time or GSI). A point will be taken off homework without the section info. Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

1. Textbook Section 10.7, p. 413, Conceptual exercise 2.
2. Consider the `USArrests` data from the textbook.
   (a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Plot the dendrogram.
   (b) Cut the dendrogram at a height that results in three distinct clusters. Report the states belonging to each of the three clusters. Make a silhouette coefficient plot and comment on any interesting features.
   (c) Repeat questions (a) and (b) using single linkage instead.
   (d) Perform K-means clustering on the data with $K = 3$ and report which states belong to which clusters. Report how you initialized the algorithm. Make a slihouette coefficient plot and comment on any interesting features.
   (e) Scale all the variables to have mean 0 and standard deviation 1. Repeat questions (a) - (d) using the scaled data.
   (f) What effect does scaling the variables have on hierarchical clustering? On K-means clustering? In your opinion, should the variables be scaled before clustering in this example? Explain your reasoning.

Please limit your solutions to at most 6 pages (including figures).