Clustering HW

For all problems, if you have to produce a plot in order to answer a question, submit the plot alongside the answer unless instructed otherwise.

Problem 1: Hierarchical Clustering

A) Using R, apply hierarchical clustering to *Prob1Data* with the "complete linkage" method. Without looking at any direct plots of the data, examine the output tree and choose the three best levels you would cut at to isolate meaningful clusters. List these three levels from best to worst, and explain your reasoning in picking them.

B) Now apply each cut and plot the results. How well did your guesses go? How does the algorithm perform when a cluster number is specified that is fewer than the obviously correct one? What about when you use the obviously correct one?

C) Instead of the "complete linkage" method, apply the "single linkage" (aka friends-of-friends) method. How does the tree look now compared to before? Why?

D) Plot the results of using 3, 6, and 15 clusters with this method. Which seemed better at low cluster number, this method or the previous one? Even at the right number of clusters, how does this algorithm perform? Why? Lastly, what does this whole exercise tell you about the human brain's classificatory ability?

Problem 2: Linear Discriminant Analysis

The data for this problem are measurements of three different species of irises (in the files, the columns are: sepal_length, sepal_width, petal_length, petal_width, and species). Using R's MASS library, perform LDA with *Prob2DataFirst* as your training set, then apply the classification to *Prob2DataSecond*. Compare the results of using different pairs of the variables in your analysis - as in, try LDA with sepal length and sepal width, and compare that to LDA with sepal width and petal length. Pick any one of the plots to turn in.

Problem 3: Why Not Both?

Prob3DataBoth is a list of the discovery methods, orbital periods (in days) and estimated masses (in Jupiter masses) of over 1000 confirmed exoplanets. Note that there are two big clusters - Radial Velocity and Transit - but also several smaller ones, such as Microlensing and Imaging. *Use log-log space,* try both LDA and hierarchical clustering separately to reconstruct the classification by discovery method, then compare the three plots you get. How do the two methods fare in this case? Why - is it the fault of the data or the methods?

Note: For LDA, you can use Prob3DataFirst to train on, and Prob3DataSecond to test on.