CSCE/STAT 587 Clustering Homework
Due: Tuesday, February 6

Part 0

Answer the following questions:

1) (**5 points**) Name a data type that can not be used for KMeans clustering.

2) (**5 points**) Is KMeans clustering deterministic? In other words, if you perform the clustering with the same value of K on the same data are you guaranteed to get the same clusters? Explain/justify your answer.

3) (**grad/honor students only: 5 points**) When using the "elbow method", what two properties of the clusters are we looking to optimize?

Part 1 (70 points)

Step 0: Review the material from the in-class lab we did on K-means clustering.

Step 1: Download the dataset "Data587.csv" from
https://cse.sc.edu/~rose/587/CSV/Data587.csv

using **wget** if you are on the VM (or your browser if you are using your own installation of RStudio). Load this data set into rstudio using the "import data" button in the environment tab. Be sure to select "From Text (base)…" in the pulldown menu.

Step 2: Plot this data set. **Save the plot to a pdf file.**

Step 3: set the random number generator seed to 888 using the command: set.seed(888). **Do this before each call to the kmeans() function so that the Grader can easily tell if your results are correct.**

Step 4: We want to explore different numbers of clusters in order to select a good value for K. As in class, calculate the within-sum-of-squares values for k=1 to 20. *Since we are using a for-loop to do this, be sure to set the random number generator seed to* ***before*** *each call to kmeans().* Plot these sum-of-squares values. **Save the plot to a pdf file.**

Step 5: From step 4, above, it is clear that K=1 is not a good number of clusters? Choose the first reasonable K based on the results from step 4. Use the kmeans() function with this number for K. Plot the results such that *each cluster is plotted in a different color* (as we did in class). **Save the plot to a pdf file.**

Step 6: Normalize the data set using the normalization functions that we created during the K-means lab.

Step 7: repeat steps 4 using the normalized data.

Step 8: repeat step 5 using the normalized data.

Step 9. Compare results from steps 4 and 7. Did the plots change enough to cause you to select different values for k? Why or why not?

Submit your plots from steps 2, 4, 5, 7, and 8 to Teams. Be sure to also document and submit your commands/code for steps 2, 3, 4, 5, 6, 7, 8 in the form of a .R file. By document, I want you to label to which step each group of commands or code correspond. Finally, do not forget to submit your analysis from step 9.

Part 2 (20 points) Next, consider hierarchically clustering the same data:

Step 1: Review the material from the in-class lecture on hierarchical clustering (slides 15 - 20). Create the distance matrix from your **normalized data of step 6** in the previous part and then create the hierarchy with the hclust() function using the "ward.D2" method. Plot the hierarchy with the 'plot' method. **Save the plot to a pdf file.**

Step 2: Visually analyze your plot from the previous step. Decide what the best number of clusters should be based on the lengths of the branches. Remember, long branches mean widely separated clusters. Assuming you decide on *m* clusters, outline the *m* cluster in the plot similar to what we did in class. **Save the plot to a pdf file.**

Submit your plots from step 1, and step 2 of this part to Teams.