STSCI 5010 HW5

Due: 11:59PM, 12/5/2018

General instruction:

- Put your SAS outputs and your answers in an MS Word file first and then convert it into a PDF file. You are required to submit both your SAS code (named HW5_LastName_FirstName.sas) and the PDF file containing your results (named HW5_Outputs_LastName_FirstName.pdf). Compress the above two files into a single file (HW5_ LastName_FirstName.7z) and submit it to the course website.
- Create a libref called HW5 to hold all your SAS data sets.
- Use SAS comments to clearly mark the beginning of each question in your SAS code.
- Use title statements to mark the beginning of the output of each question.

A dataset (flowers.sas7bdat) about some measures of three related flower species is provided. You are asked to cluster these flowers based on these measures. Do the following:

1. Perform a variable selection procedure. Set proportion =0.9. Report the R-square table showing the 1-R**2 ratios (Hint: the table is available from the Results Panel in SAS). Select the variable representatives based on your results. (15 points)
2. Produce a horizontal tree with above results; set _propor_ as the height. (13 points)
3. Do a principle component analysis (PCA) on the data set just by using the variables selected above, and then estimate the number of clusters using a PCA plot, for which the plot symbols should be directly labeled with the flower species (A, B and C) and color coded (Hint: Refer to the lecture notes of Preparation for Clustering). (30 points)
4. Standardize the dataset by setting method=range (see the example in the lecture notes); only use the variable representatives selected above. (7 points)
5. Use the Average Linkage method to cluster the standardized data. Determine the number of clusters with the values of CCC, PSF and PST2 (Hint: These values can be found by clicking “Cluster History” under “Cluster” on the Results Panel; you need to scroll down to the bottom of the “Cluster History.” You only need the last portion of the cluster history to determine the number of clusters). Comment on how you used these values to decide the cluster numbers. (25 points)
6. Plot the tree. (10 points)