Homework 5

STAT 448 - Advanced Data Analysis

Due: Friday, November 17, 2017 at 5:00 pm

For all exercises, use the `leaf.csv` file and the code in `HW5Data.sas` in the course space to obtain the data set. The data set and original variables are described on http://archive.ics.uci.edu/ml/datasets/Leaf, and additional information about the species can be found in (leaf_data_description.pdf) in HW5 folder.

1. (3 parts) Perform a principal components analysis on the characteristics of the leaves (all variables except **Species** and **SpecimenNumber**).
   (a) Determine how many components you would keep to retain at least 85% of the total variation from the original variables. Also comment on how many components would be chosen based on the average eigenvalue and scree plot methods.

   (b) For the components you would keep based on the 85% criterion in part a), explain what features these components pick out of the data (e.g. what leaf features or contrasts are they picking up on?).

   (c) Create score plots for the components kept and label with the **Species** values. Comment on any species that are extreme with respect to any of the components and what the principal component values for those species tell us about features of those species leaves.

2. (4 parts) Repeat Exercise 3, but now perform a covariance-based principal components analysis.
   (a) Determine how many components you would keep to retain at least 85% of the total variation from the original variables. Also comment on how many components would be chosen based on the average eigenvalue and scree plot methods.

   (b) For the components you would keep based on the 85% criterion in part (a), explain what features these components pick out of the data.

   (c) Using the covariance-based PCA from part (a), create score plots for the components kept and label with the Species values. Comment on any species that are extreme with respect to any of the components and what the principal component

values for those species tell us about features of those species leaves.

(d) Comment on any differences between the correlation-based and covariance-based PCA results.

3. Perform an average linkage cluster analysis on the leaf characteristics for the first 8 species (do not standardize the clustering variables)

   (a) Comment on how many clusters you would choose based on the dendrogram, and the CCC, pseudo F and pseudo t^2 statistics.

   (b) Compare the clusters with the species. Which species are separated out well by the clustering and which species seem to be grouped together? How do the clusters differ with respect to the leaf characteristics and what do the results tell us about the species whose leaves were well separated by the clustering?