Homework 6

STAT 448 - Advanced Data Analysis

Due: Thursday, April 26 at 5:00 pm

The data sets are provided in the **HW6Data.sas** file in the Homework 6 folder on the course website. For exercise 1-2, use the **glassid** data set. The data set is based on the **Glass Identification data** from the UCI Machine Learning Repository. The original data and variables are described on `http://archive.ics.uci.edu/ml/datasets/Glass+Identification`. The **glassid** data set omits the **Id** number, and the **Type** of glass variable is replaced by a **groupedtype** variable which merges several combinations of glass types.

For exercise 3-4, use the **wine.txt** file. The data set and original variables are described on `http://archive.ics.uci.edu/ml/datasets/Wine`, and additional information about the wine data can be found here:`http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names`. The alcohol variable in the data set classifies the wine into three groups based on cultivars (the types of grapes used to make the wine). All other variables are physiochemical attributes of the wines.

1. Consider grouping observations into clusters using cluster analysis. Recall that you will need to set an option to get a dendrogram for data sets containing more than 200 observations by default.

   树图

   (a) Use average linkage on the standardized glass oxide levels. Do not include **Ri** (refractive index) or **groupedtype** (glass type variable) as clustering variables, but do copy them into the results for use in further analysis. What do the dendrogram, pseudo F, pseudo t^2, and CCC statistics suggest about the number of clusters. How many clusters should you choose?

   (b) Comment on how well the clusters do (or do not) match up with the glass type groupings, and which types of glass have similar chemical compositions based on the clustering.

2. Now consider an analysis of variance (ANOVA) for refractive index as a function of cluster using the results from Exercise 1. In a good clustering, observations in very small clusters will tend to have extreme differences from other clusters. For this analysis, leave out observations from clusters that contain less than 5 observations to focus on the larger clusters.

(a) Fit the ANOVA model for refractive index for clusters containing 5 or more observations. Perform homogeneity of variance test for the analysis, comment on the significance of the model, and the variation described by the model.

(b) Perform the Tukey's pairwise test for comparing all pairwise differences of means, and comment on any significantly different refractive index means across clusters. Interpret what these results tell us about differences of refractive index across clusters. Given R^2 from part (a) and the mean differences observed for this model, how useful would this model be for predicting refractive index?

3. Use the **wine** data and answer the following questions.

   (a) Perform a principal components analysis on the attributes of wine (all variables except **alcohol**), and keep only the first 2 principal components. Make a scatter plot using these two principal components and comment on how the different alcohols separate visually.

   (b) Perform an average linkage cluster analysis on the first two principal components, and comment on how many clusters you would choose based on the dendrogram, CCC, pseudo F and pseudo t^2 statistics.

   (c) Comment on the separation performance of the combined method (PCA + average linkage cluster analysis) on the wine data.

4. Now we use the **wine** data for discriminant analysis.

   (a) We want to perform a discriminant analysis for **alcohol** with a function of continuous variables chosen in a stepwise discrimination. Consider all continuous variables as candidates for the model selection. Comment on which predictors are chosen based on the stepwise discrimination procedure.

   (b) Perform a discriminant analysis for alcohol as a function of all the continuous variables. Test whether LDA or QDA is more appropriate. Comment on what the MANOVA tests tell us about possibility to discriminate between alcohol types based on these variables.

   (c) Comment on the cross-validation error results and how well the discrimination matches the groups. Comment on similarities and dissimilarities between the classification results and the cluster frequency analysis results from Exercise 3 part (c).