Statistical and Machine Learning (Spring 2018)

Mini Project 5

Instructions:

- Due date: April 12, 2018.
- Total points = 40.
- Submit a typed report.
- It is OK to discuss the project with your old partner and other students in the class, but each person must write their own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- **Explain all assertions, interpret all results, and justify all conclusions**.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider the planet data stored in the `planet.csv` file available on eLearning. These data give values of three features for 101 exoplanets discovered up to October 2002. We are interested in clustering the exoplanets on the basis of these features. Note that an exoplanet is a planet located outside the solar system. You may read about them at `https://en.wikipedia.org/wiki/Exoplanet`. The features recorded are — `Mass` (in Jupiter mass), `Period` (in Earth days), and `Eccentricity`.

   (a) Perform an exploratory analysis of the data. Be sure to examine the univariate distributions of the variables and their bivariate relationships using appropriate plots and summary statistics.

   (b) Do you think standardizing the variables before clustering would be a good idea?

   (c) Would you use metric-based or correlation-based distance to cluster the exoplanets?

   (d) Regardless of your answers in (b) and (c), standardize the variables and hierarchically cluster the exoplanets using complete linkage and Euclidean distance. Display the results using a dendogram. Cut the dendogram at a height that results in three distinct clusters. Summarize the cluster-specific means of the three variables (on the original scale) in a tabular form. Also, make pairwise scatterplots of the three variables (on the original scale) and show the three clusters in different colors.

   (e) Repeat (d) using K-means clustering with $K = 3$. (Of course, you won't have a dendogram in this case.) Compare the conclusions with (d).

2. Consider the data presented in the file `track-records-women.csv` on eLearning. The data consist of national track records for women in 54 countries in a certain year. Note that the records for 100-m, 200-m, and 400-m, runs are measured in seconds, whereas those for 800-m, 1500-m, 3000-m, and marathon runs are measured in minutes. We are interested in performing a principal components analysis of these data.

   (a) Perform an exploratory analysis of the data.

   (b) Do you think standardizing the variables before performing the analysis would be a good idea?

   (c) Regardless of your answer in (b), standardize the variables, and perform a principal components analysis of the data. Summarize the results using appropriate tables and graphs. How many principal components would you recommend?

   (d) Focus on the first two principal components obtained in (c). Prepare a table showing the correlations of the standardized variables with the components and the cumulative percentage of the total variability explained by the two components. Also, display the scores on the two components and the loadings on them using a biplot. How can we interpret the two components? Can the first be interpreted as a measure of athletic excellence of a given nation and the second as the relative strength of a nation at various running distances?

   (e) Rank the nations based on their score on the first principal component. Does this ranking correspond with your intuitive notion of athletic excellence for the various countries?

3. Consider the `OJ` dataset which is a part of the `ISLR` package. It consists of `Purchase` as a binary response variable and a number of other variables as predictors.

   (a) Divide the data into a training set consisting of the first 870 observations and a test set consisting of the last 200 observations.

   (b) Fit a tree to the training data. Summarize the results and display them graphically. Explicitly describe the regions corresponding to the terminal nodes that provide a partition of the predictor space (i.e., provide expressions for the regions $R_1, \ldots, R_J$).

   (c) Use cross-validation to determine whether pruning is helpful and determine the optimal size for the pruned tree. Display the pruned tree and compare with the un-pruned tree. Which predictors seem to be the most important? Report the test MSE for the pruned tree.

   (d) Use a bagging approach to analyze these data with $B = 1000$. Compute the test MSE. Which predictors seem to be the most important?

   (e) Repeat (d) with a random forest approach with $B = 1000$ and $m = \sqrt{p}$

   (f) Repeat (d) with a boosting approach with $B = 1000$, $d = 1$, and $\lambda = 0.01$.

   (g) Repeat (d) with a KNN approach with K chosen optimally.

   (h) Compare the results from the various methods. Which method would you recommend?