STAT 6340 (Statistical and Machine Learning, Spring 2020)

Mini Project 6

Instructions:

- Due date: Apr 27, 2020.
- Total points = 30.
- Submit a typed report.
- It is OK to discuss the project with others and to search on the internet, but each student must write their own code and answers. If your 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.
- You must use the following template for your report:

  Mini Project #
  Name
  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.

- Section 1 of the report must be limited to **3** pages. Also, only those output should be provided in this section that are referred to in the report.

---

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 `College` data from the previous project. Recall that we would like to predict the number of applications received using the other variables. All data are taken as training data.

(a) Fit a tree to the data. Summarize the results. Unless the number of terminal nodes is large, display the tree graphically and explicitly describe the regions corresponding to the terminal nodes that provide a partition of the predictor space (i.e., provide expressions for the regions $R_1, \ldots, R_J$). Report its MSE.

(b) Use LOOCV to determine whether pruning is helpful and determine the optimal size for the pruned tree. Compare the pruned and un-pruned trees. Report MSE for the pruned tree. Which predictors seem to be the most important?

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

(d) Repeat (c) with a random forest approach with $B = 1000$ and $m \approx p/3$.

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

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