STAT 6340 (Statistical and Machine Learning, Spring 2021)

Mini Project 6

Instructions:

- Due date: May 3, 2021.
- Total points = 30
- Submit a typed report.
- It is OK to discuss the project with other students in the class, 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 `Hitters` dataset from the previous project. It consists of 20 variables measured on 263 major league baseball players (after removing those with missing data). Take log(`Salary`) as response (due to skewness in `Salary`) and the remaining 19 variables as predictors. All data will be taken as training data. For all the models below, use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

   (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 estimated test MSE.

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

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

   (d) Use a random forest approach to analyze the data with $B = 1000$ and $m \approx p/3$. Compute the estimated test MSE. Which predictors seem to be the most important?

   (e) Use a boosting approach to analyze the data with $B = 1000$, $d = 1$, and $\lambda = 0.01$. Compute the estimated test MSE. Which predictors seem to be the most important?

   (f) Compare the results from the various methods. Which method would you recommend? How does your recommendation compare with the method you recommended in the previous project?

2. Consider the diabetes dataset from Mini Projects 3 and 4. As there, we will take `Outcome` as the binary response, the remaining 8 variables as predictors, and all the data as training data. For all the models below, use 10-fold cross-validation to compute the estimated test error rates and also to tune any hyperparameter that requires tuning.

   (a) Fit a support vector classifier to the data with cost parameter chosen optimally. Summarize key features of the fit. Compute its estimated test error rate.

   (b) Fit a support vector machine with a polynomial kernel of degree two and cost parameter chosen optimally. Summarize key features of the fit. Compute its estimated test error rate.

   (c) Fit a support vector machine with a radial kernel with both γ and cost parameter chosen optimally. Summarize key features of the fit. Compute its estimated test error rate.

   (d) Compare results from the above three methods and also from the method you recommended for these data in Mini Projects 3 and 4. Which method would you recommend now?