

Homework 4

Due on 04/08/2021

This problem involves the `Prostate` data in the `lasso2` package (see `L8_2021.Rmd`).

- (a) Fit a regression tree with `lpsa` as the response and the other variables as predictors. Use cross-validation to determine the optimal tree size. Which tree size corresponds to the lowest cross-validation error? Is this the same as the tree size obtained using the 1 SE rule?
- (b) Create a plot of the final tree you choose. Pick one of the terminal nodes, and interpret the information displayed.
- (c) Perform bagging and report the variable importance.
- (d) Perform random forest and report the variable importance.
- (e) Perform boosting and report the variable importance.
- (f) Which of the above models will you select to predict PSA level?

2. This problem involves the `OJ` data in the `ISLR` package. The data contains 1070 purchases where the customers either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of customers and products are recorded. Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

- (a) Fit a classification tree to the training set, with `Purchase` as the response and the other variables as predictors. Use cross-validation to determine the tree size and

create a plot of the final tree. Predict the response on the test data. What is the test classification error rate?

- (b) Perform random forest on the training set and report variable importance. What is the test error rate?
- (c) Perform boosting on the training set and report variable importance. What is the test error rate?