Homework 5

STAT 601

Instructions

Answer all questions stated in each problem. Discuss how your results address each question.

Submit your answers as a pdf, typeset (knitted) from an Rmd file. Include the Rmd file in your submission. You can typeset directly to PDF or typeset to Word then save to PDF In either case, both Rmd and PDF are required. If you are having trouble with .rmd, let us know and we will help you. If you knit to Word, check for any LaTeX commands that will not be compatible with Word.

This file can be used as a template for your submission. Please follow the instructions found under "Content/Begin Here" titled **Homework Formatting**. No code should be included in your PDF submission unless explicitly requested. Use the `echo = F` flag to exclude code from the typeset document.

For any question requiring a plot or graph, answer the question first using standard R graphics (See ?graphics). Then provide a equivalent answer using `library(ggplot2)` functions and syntax. You are not required to produce duplicate plots in answers to questions that do not explicitly require graphs, but it is encouraged.

You can remove the `Instructions` section from your submission.

Exercises

1. (Ex. 9.1 pg 186 in HSAUR, modified for clarity) The **BostonHousing** dataset reported by Harrison and Rubinfeld (1978) is available as a `data.frame` structure in the **mlbench** package (Leisch and Dimitriadou, 2009). The goal here is to predict the median value of owner-occupied homes (`medv` variable, in 1000s USD) based on other predictors in the dataset.

   a) Construct a regression tree using rpart(). Discuss the results, including these key components:

      - How many nodes does your tree have?
      - Did you prune the tree? Did it decrease the number of nodes?
      - What is the prediction error (MSE)?
      - Plot the predicted vs. observed values.
      - Plot the final tree.

   b) Apply bagging with 50 trees. Report the prediction error (MSE) and plot the predicted vs observed values.

   c) Apply bagging using the randomForest() function. Report the prediction error (MSE). Was it the same as (b)? If they are different what do you think caused it? Plot the predicted vs. observed values.

   d) Use the randomForest() function to perform random forest. Report the prediction error (MSE). Plot the predicted vs. observed values.

   e) Include a table of each method and associated MSE. Which method is more accurate?