DATA SCI 415 - Homework 7

Due Wednesday, April 2, 2025

1. This exercise relates to the `spam` data set, which can be found in the files under the `Assignments/Data` directory. The data set is a collection of 4601 emails of which 1813 were considered "spam", i.e., unsolicited commercial email. The data set consists of 58 variables of which 57 are continuous predictors and one is a class label that indicates whether the email was considered spam (1) or not (0). Among the 57 prediction variables are: percentage of the word "free" in the email, percentage of the exclamation marks in the email, etc. See file `spam-names` for the full list of variables. The goal is, of course, to predict if an email is "spam" or not.

   (a) The data set has been divided at random into two parts `spam-train` and `spam-test` that contain 2/3 and 1/3 of the original data respectively. Fit a classification tree using only the training set. Find the percentage of emails in the test set that were misclassified by your optimal tree. Of all the spam emails of the test set what percentage was misclassified and of all the non-spam emails of the test set what percentage was misclassified?

   (b) Plot a subtree of the optimal tree that has at most 8 terminal nodes. What are some of the variables that were used in tree construction?

   (c) Try (a) again using Random Forest. Use the "importance()" function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.