ANLY512: Homework 5

Due May 5 11:59pm

Problem 1 - Random Forest to predict median home value `medv` in the Boston data

Split the data into train, dev, and test (70/15/15)

Part A: Use the `Boston` data using `mtry=6` with `ntree=25` and `ntree=500`.

Part B: Create a plot displaying the dev set error resulting from random forests on this data set for a more comprehensive range of values for mtry and ntree. Make sure you use a vector of mtry and ntree values and use a for loop or function to run each model. Use Figure 8.10 from ISLR as a guide for this plot. In addition, show a plot of train error.

Part C: Describe the results obtained (is there evidence of overfitting?) and recommend the choice for `mtry` and `ntree` based on the dev test error. Report the performance of your chosen model on the test dataset.

Part D: Produce a variable importance plot for this chosen model and discuss the predictors that influence your final result.

Problem 2 - Clustering MNIST Data

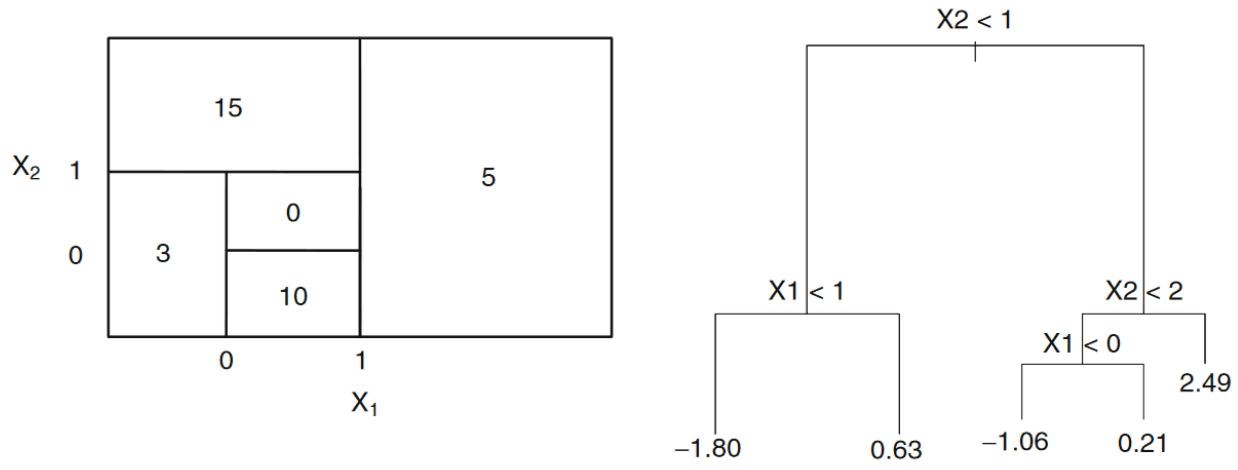
Use the file `mnist__all.RData` to create a dataframe of the train dataset, just like HW4.

Part A: Use k-means clustering with two clusters. Can you tell which digits tend to be clustered together?

Part B: Apply k-means clustering with 10 clusters. How well do the cluster labels agree with the actual digit labels? Use a confusion matrix to answer this question.

Problem 3 - Delving into Decision Trees

Figure:



Part A: Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of the figure above. The numbers inside the boxes indicate the mean of Y within each region.

Part B: Create a diagram similar to the left-hand panel of the figure above, using the data from the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.