EMIS 3309: Information Engineering

Lab 10: Decision Tree, Bagging and Random Forest

For this lab, you will need to go to 'UCI Machine Learning Repository' website. Use the link provided below and download 'adult.data'. Read description on the website to understand features and response values stored in the dataset.

Source of the dataset: `https://archive.ics.uci.edu/ml/datasets/Adult`

The data is from 1994 Census database. The goal of the problem is to apply tree-based methods using attained feature values to predict whether the income of a given household is greater than (or less than) $50,000.

1. Import data from the downloaded file to R. Save the data imported from 'adult.data' as **adult**. Note that the file does not contain header (names of the features).

   Hint: use **read.table( )** function to import data from a comma-separated text file. To learn more about the function, try **?read.table** in R console.

2. Remove the feature that contains native-country values from 'adult'.

3. Assign column-names to the dataset, adult, using **colnames( )** function.

4. Use **summary( )** function to check the summary of values in each column, and to scan whether the dataset contains empty values (NA) or outliers.

5. Set seed of 100 for the random number generator, then split 'adult' into training set (containing 20,000 samples) and testing set (containing the remaining samples).

6. Build a tree model using all features. Plot the resulting tree model.

7. Apply 5-fold cross validation on the above tree model to decide the number of terminal nodes. What is the size of the tree (number of terminal nodes) that yields the minimum error?

8. Prune the tree with the size selected in the previous problem. Using the pruned tree, make prediction on the testing dataset. What is the misclassification error rate?

   Hint: Make sure to use **type = "class"** inside the **predict( )** function to make classification-prediction.

9. Set random number generator seed as 100. Based on bootstrapping method, build a tree model using all features. What is the misclassification rate of the model on testing data?

10. Set random number generator seed as 100. Build a random forest model. Make sure that the model is built by selecting a feature out of a set of 4 randomly selected features whenever branching is considered. What is the misclassification rate of the model on testing data?