ML 4375 – Intro to Machine Learning – Summer 2018 – Mazidi

Homework 6: Decision Trees

This homework compares the accuracy of logistic regression versus decision trees in two different data sets.

Problem 1: Default data (25 points)

Instructions:

1. Set up the data. (5 points)
   1. Load the ISLR library and attach the Default data set.
   2. Run the dim() and names() functions on Default
   3. Use function set.seed(2017) so your results are reproducible.
   4. Divide the data into 80% for training and the rest for a test set.
2. Logistic regression model (5 points)
   1. create a logistic regression model on the training data where **default** is predicted by all other variables
   2. run a summary of the model
   3. predict Yes/No on the test data
   4. compute the accuracy
3. Decision tree model (10 points)
   1. create a decision tree model on the training data
   2. run a summary of the model
   3. make predictions for the test set
   4. compute the accuracy
   5. print the tree with labels
   6. display the tree in nested decision form by just using the tree name
4. Try Random Forest or Boosting (5 points) and compare your results.

Problem 2: Heart data (45 points)

Instructions:

1. Set up the data. (5 points)
   1. Download the heart data to your machine from Piazza.
   2. Load the data into R and attach it
   3. Remove the “X” column
   4. Set up train and test sets with 80% for training again using seed 2017
2. Logistic regression model (5 points)
   1. create a logistic regression model on the training data where **AHD** is predicted by all other variables
   2. run a summary of the model
   3. predict Yes/No on the test data
   4. compute the accuracy
3. Decision Tree Model (5 points)
   1. create a decision tree model on the training data
   2. run a summary of the model
   3. make predictions for the test set
   4. compute the accuracy
4. Display the tree (5 points)
   1. print the tree with labels
   2. display the tree in nested decision form by just using the tree name
5. Cross validation (10 points)
   1. create a new tree from the cv.tree() function
   2. look at the $dev and $size variables by displaying the tree using its name
   3. plot in a 1x2 format:
      1. $size and $dev
      2. $k and $dev
6. Prune the tree (10 points)
   1. create a new pruned tree using best=n where n is the optimal size indicated in step 5
   2. plot the new pruned tree with labels
   3. Using the pruned tree, make predictions on the test set
   4. Compute the accuracy
7. Try Random Forest or Boosting (5 points) and compare your results

Answer the questions at the bottom of your Rmd file. Upload the Rmd file to TurnItIn.

Questions: (30 points)

Problem 1

1. In the logistic regression model, which variables were important and which were not?

2. What was the accuracy of the logistic regression model and the decision tree model?

3. In the tree, why might you have a branch where both branches are No?

4. Write a simple if/else statement that summarizes the Yes/No values in the decision tree.

5. Is it a good idea to prune this tree? Why or why not?

Problem 2

1. Which variables were important (2 or 3 \*\*) in the logistic regression model?

2. Which variables were used to create the decision tree?

3. Compare the accuracy of the logistic regression model and the decision tree.

4. What was the accuracy on the pruned tree?

5. Which model (logistic regression, decision tree) might be more meaningful to a doctor, and why.