Grant Ikehara, Cameron Healy, Dominic Soares

CPSC 310

HW6

Descriptions of Processes and Results

Step 1:

1. We made a function that takes the tables and determines their length to figure out how many instances will be in the test and remainder set.
2. We made a bootstrapping function that can take a remainder set of instances and the N variable that specifies the number of trees, which is the number of bootstrap lists. When creating the tree, we use the bootstrap sample, and randomly selecting attributes, we end up with 20 different trees.
3. We wrote a function to test each tree and find the tree’s accuracy when it is used on the test set. The M variable for the number of trees we want is passed in and we take the M best trees and store them.
4. Using the trees that were stored, we use them to classify the test set by using majority voting from the 7 results per instance, and get the overall accuracy after prediction on the test set.

Step2:

We passed in the titanic and auto data tables with N = 20, M = 7, and F = 2 to the random forest function to find the accuracy. Each of the predictions and results are put into the confusion matrix. We also made a normal tree that uses the entropy values to order the tree to compare the confusion matrix and accuracy against the random forest’s.

Step 3:

We ran our program with different values of the three parameters N,M,F to determine if the values of the different random forests differ greatly from each other. The most accurate trees are highlighted in the results. This result is also compared to the normal tree’s accuracy.

Step 4:

We ensured that our functions for the random forest on the titanic and auto data can also be used on the Wisconsin data, including the read csv function. Using the functions, we ran the Wisconsin text file through the forests algorithms to find the accuracy of the random forest and the normal decision tree for comparison.

Test Results:

Normal Decision Tree

|  |  |  |
| --- | --- | --- |
|  | Accuracy % | Error Rate % |
| Auto Data | 45-55 | 45-55 |
| Titanic | 70-80 | 20-30 |
| Wisconsin | 89-93 | 7-11 |

Random Forest - N=20, M=7, F=2

|  |  |  |
| --- | --- | --- |
|  | Accuracy % | Error Rate % |
| Auto Data |  |  |
| Titanic |  |  |
| Wisconsin |  |  |

Random Forest - N=20, M=7, F=2

|  |  |  |
| --- | --- | --- |
|  | Accuracy % | Error Rate % |
| Auto Data |  |  |
| Titanic |  |  |
| Wisconsin |  |  |

Test Results Screenshots: