Compsci 361 Assignment 1

Hasnain Cheena  
190411106  
hche737

## Part A

Obscured A is the real file while Obscured B is the shuffled dataset.

## Part B:

*Method 1: Examing nodes and tree structure*

The first method used was to examine the structure of the decision tree and nodes of the decision tree produced. The decision tree created by Obscure A had a size of 1636 and had 1623 leaves. In comparision the decision tree created by Obscure B had a size of 1 and had 1 leaf. Therefore, as the tree fitted to Obscure B had no internal nodes meaning that no attributes were used during the classification. Therefore, the tree found no useful features. Hence it was concluded that Obscure B did not contain a signal and was the shuffled dataset.

*Method 2: Re-shuffle the data*

The second method involved fitting a decision tree on both datasets with 10-fold cross validation. Then shuffling both datasets, rerunning the training process with 10-fold cross validation and comparing the results to before shuffling. Prior to shuffling the accuracy for Obscure A was 92.1% and for Obscure B was 88.6%. After shuffling the accuracy for Obscure A dropped down to 88.6% and for Obscure B the average accuracy remained at 88.6%. Therefore, as the accuracy for Obscure A decreased when the dataset was randomly shuffled, Obscure A is the real dataset. Furthermore, as the accuracy for Obscure B remained similar before and after randomisation, Obscure B does not contain a signal and was the shuffled dataset.

## Part C

Assessing the nodes and leaves of the decision tree’s produced (method 1) and re-shuffling the data (method 2) work well even as the datasets are scaled down. These techniques work well even when the datasets are scaled down because the distribution target labels when scaled down is similar to the distribution of target labels in the original dataset.

Results are summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Total Nodes** | **Number of Leaves** | **Accuracy** | **Randomised Accuracy** |
| Obscure A | 1636 | 1623 | 92.1% | 88.6% |
| Obscure A-50 | 1630 | 1620 | 92.0% | 88.2% |
| Obscure A-25 | 604 | 600 | 91.7% | 87.4% |
| Obscure B | 1 | 1 | 88.6% | 88.6% |
| Obscure B-50 | 1 | 1 | 88.4% | 88.5% |
| Obscure B-25 | 1 | 1 | 88.4% | 88.5% |

From the results of method 1 you can see clear differences in the structure of the decision trees fitted on Obscure A (and its scaled variants) and Obscure B (and its scaled variants). As the number of total instances was scaled down the size of the overall tree’s produced got smaller, however the differences in the size of the trees remained significant. The trees produced by all scaled variants of Obscure B contained a single node, indicating they found no useful features and thus allowed us to easily conclude that Obscure B-50 and Obscure B-25 were the shuffled datasets.

From the results of method 2 you can see the as the Obscure A dataset is scaled down the unrandomised accuracy decreases slightly. However, the difference between the unrandomised and randomised accuracies of all variants of Obscure A is signficiant enough to indicate that Obscure A must be the dataset with a signal. In comparision the unrandomised and randomised accuarcies of Obscure B when scaled down remains similar, indicating that Obscure B (and its variants) is the shuffled dataset.

## Part D

Method 1 (examing the tree structure) is more reliable than Method 2 (re-shuffling data). The key reason for this is when randomly shuffling there is the chance that shuffling does not truly simulate randomness. For example in the case where shuffling only affects the majority class labels. In this scenario, the randomised and unrandomised accuracy of the Obscure A dataset would be similar, rendering method 2 with insufficient evidence to decide which dataset has been shuffled.