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: Examine size and structure of the tree

The first method involved looking at the internal nodes and leaves of the decision trees produced. By examining the tree's themselves, it was possible to assess whether attributes were actually used during the classification process.

The decision tree created by Obscured A had a size of 1636 and had 1623 leaves. In comparision the decision tree created by Obscured B had a size of 1, the only node being a leaf. The tree fitted to Obscured B had no internal nodes meaning that no attributes were used to classify instances. This means that the decision tree found no useful features during the pruning process. Hence it was concluded that Obscured 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 and re-running the process. The shuffled dataset will produce similar accuracy scores before and after the randomisation, while the true dataset will show a decrease in accuracy.

Prior to shuffling the accuracy for Obscured A was 92.1% and for Obscured B was 88.6%. After shuffling the accuracy for Obscured A dropped down to 88.6% and for Obscured B the accuracy remained at 88.6%. Therefore, as the accuracy for Obscured A decreased when the dataset was randomly shuffled, Obscured A is the real dataset. Furthermore, as the accuracy for Obscured B remained similar before and after randomisation, Obscured B does not contain a signal and was the shuffled dataset.

Part C

Both methods (described above) were passed scaled down versions of Obscured A and Obscured B. The results are summarized below:

Dataset	**Tree Size**	**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%

As can be seen from the results, the methods both work well to distinguish between the scaled-down variants of Obscured A and Obscured B. The reason for this is because the scaled-down datasets are representitive samples of the original datasets and they contain enough data to detect a signal. However, it is evident from the results that both the size of tree and the gap between the randomised/unrandomised accuracy are becoming smaller. The reason for this behaviour is as the datasets shrink, the decision tree begins overfitting to the training examples. Therefore, as the datasets become smaller (same number of variables, less instances) there will come a point where neither method will be any good and it will be very difficult to tell the true dataset from the randomised one.

Part D

Method 2 (re-shuffling the data) is more reliable than Method 1 (examine tree structure). This is because as the datasets are scaled down (same number of variables, less instances), the tendency of a decision tree to overfit will increase. Therefore, in Method 1 both datasets (shuffled and normal) will produce overfitted decision trees, making it difficult to distinguish the true dataset from the shuffled one. In contrast the 10-fold cross-validation in Method 2 should alleviate the overfitting problem, making it possible to distinguish between the randomised and normal datasets. However, as aforementioned there will come a point where the datasets are small enough and it will be impossible to distinguish between Obscured A and Obscured B, using either method.