Report

Zihao Xue, Saifullah Ijaz, Alexandros Panagiotidis and Samuel Andresen

1 Decision Trees without Pruning

1.1 Confusion Matrix

The confusion matrices have been calculated by splitting the data into 10 folds and using 1 fold for testing and 9 for training. The training data is then turned into a decision tree, which we then evaluate against our test data to obtain the confusion matrix.

$$\text{Unpruned Confusion Matrix Clean} = \begin{pmatrix} 50 & 0 & 1 & 0 \\ 0 & 48 & 2 & 0 \\ 0 & 2 & 46 & 1 \\ 0 & 0 & 1 & 49 \end{pmatrix}$$

$$\text{Unpruned Confusion Matrix Noisy} = \begin{pmatrix} 40 & 2 & 3 & 4 \\ 3 & 40 & 4 & 2 \\ 3 & 5 & 41 & 3 \\ 4 & 2 & 3 & 41 \end{pmatrix}$$

1.2 Accuracy

$$\textit{Unpruned Clean Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{50 + 48 + 46 + 49}{200} = 0.965$$

$$\textit{Unpruned Noisy Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{40 + 40 + 41 + 41}{200} = 0.81$$

1.3 Recall and Precision Rates

Unpruned Clean Data

Class	Room1	Room2	Room3	Room4
Recall Rate	0.98	0.96	0.94	0.98
Precision Rate	1	0.96	0.92	0.98

Unpruned Noisy Data

Class	Room1	Room2	Room3	Room4
Recall Rate	0.82	0.82	0.79	0.82
Precision Rate	0.80	0.82	0.80	0.82

1.4 F1 Measures

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{\beta^2 * precision + recall}$$

Class	Room1	Room2	Room3	Room4
F_1 *Clean*	0.99	0.96	0.93	0.98
F_1 *Noisy*	0.81	0.82	0.79	0.82

$$Unpruned\ Avg\ Clean\ F_1 = \frac{0.99 + 0.96 + 0.93 + 0.98}{4} = 0.965$$

$$Unpruned\ Avg\ Noisy\ F_1 = \frac{0.81 + 0.82 + 0.79 + 0.82}{4} = 0.81$$

1.5 Result Analysis

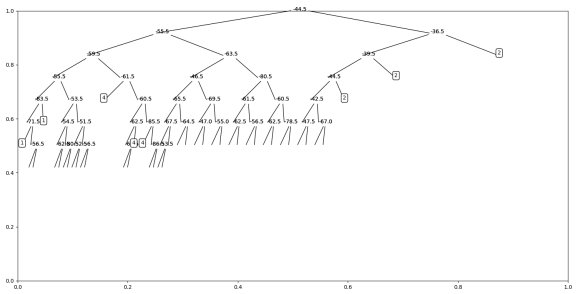
The confusion matrix, precision and recall rates show that for the clean dataset, room 1 has the best accuracy, followed by room 4, room 2 and finally room 3 with the lowest accuracy. Rooms 2 and 3 are sometimes confused with each other, possibly due to the close proximity of several Wi-Fi emitters around both of these rooms. For the noisy dataset, the accuracies are a lot lower overall but every room has a similar accuracy, with all of the rooms being confused with each other more often.

1.6 Dataset Differences

The F1-measures show that the decision tree performs better on the clean dataset, with the average accuracy being 0.965 compared with an average accuracy of 0.81 for the noisy dataset. It is expected that the tree performs worse with the noisy dataset since the decision tree attempts to classify noisy samples which results in overfitting. This reduces the accuracy of the tree since it does not generalise well to new data and will more often misclassify samples, as shown by the cross-validation metrics.

1.7 BONUS - Plot of Clean Dataset

In this plot, we've only plotted 50 branches since after this the plot becomes very clustered.



2 Decision Trees with Pruning

2.1 Confusion Matrix

$$\text{Pruned Confusion Matrix Clean} = \begin{pmatrix} 50 & 0 & 1 & 0 \\ 0 & 47 & 2 & 0 \\ 0 & 3 & 48 & 1 \\ 0 & 0 & 0 & 49 \end{pmatrix}$$

$$\text{Pruned Confusion Matrix Noisy} = \begin{pmatrix} 42 & 3 & 3 & 4 \\ 2 & 41 & 3 & 2 \\ 2 & 4 & 42 & 2 \\ 3 & 2 & 2 & 43 \end{pmatrix}$$

2.2 Accuracy

$$Pruned\ Clean\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{50 + 47 + 48 + 49}{200} = 0.97$$

$$Pruned\ Noisy\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{42 + 41 + 42 + 43}{200} = 0.84$$

2.3 Recall and Precision Rates

Pruned Clean Data

Class	Room1	Room2	Room3	Room4
Recall Rate	0.98	0.96	0.92	1
Precision Rate	1	0.94	0.94	0.98

Pruned Noisy Data

Class	Room1	Room2	Room3	Room4
Recall Rate	0.81	0.85	0.84	0.86
Precision Rate	0.86	0.82	0.84	0.84

2.4 F1 Measures

Class	Room1	Room2	Room3	Room4
F_1 *Clean*	0.99	0.94	0.93	0.99
F_1 *Noisy*	0.83	0.83	0.84	0.85

$$Pruned\ Avg\ Clean\ F_1 = \frac{0.99 + 0.94 + 0.93 + 0.99}{4} = 0.96$$

$$Pruned\ Avg\ Noisy\ F_1 = \frac{0.83 + 0.83 + 0.84 + 0.85}{4} = 0.84$$

2.5 Result Analysis After Pruning

The accuracy for the clean dataset set only improves by 0.5% after pruning, whereas the accuracy for the noisy dataset improves by 3% after pruning. Pruning reduces the depth of the tree so that it does not try and classify noisy samples which means over-fitting is less likely to occur. Since the noisy dataset has more noisy samples than the clean dataset, pruning is more effective on the noisy dataset so the overall increase in accuracy is higher for the noisy dataset.

2.6 Depth Analysis

For the unpruned tree on the clean dataset, the average depth was 13. For the pruned tree on the clean dataset, the average depth was 12. For the unpruned tree on the noisy dataset, the average depth was 15. For the pruned tree on the noisy dataset, the average depth was 13. The tree will have a certain depth at which the prediction accuracy will be at its highest because hyperparameter tuning will result in the best classifier. Depths below or above this point will result in a decrease in accuracy.