Machine Learning Coursework 1 - Decision Trees

David Cormier, Thomas Loureiro van Issum,
Petra Ratkai, Simon Staal

October 2021

1 Output of the tree visualisation function

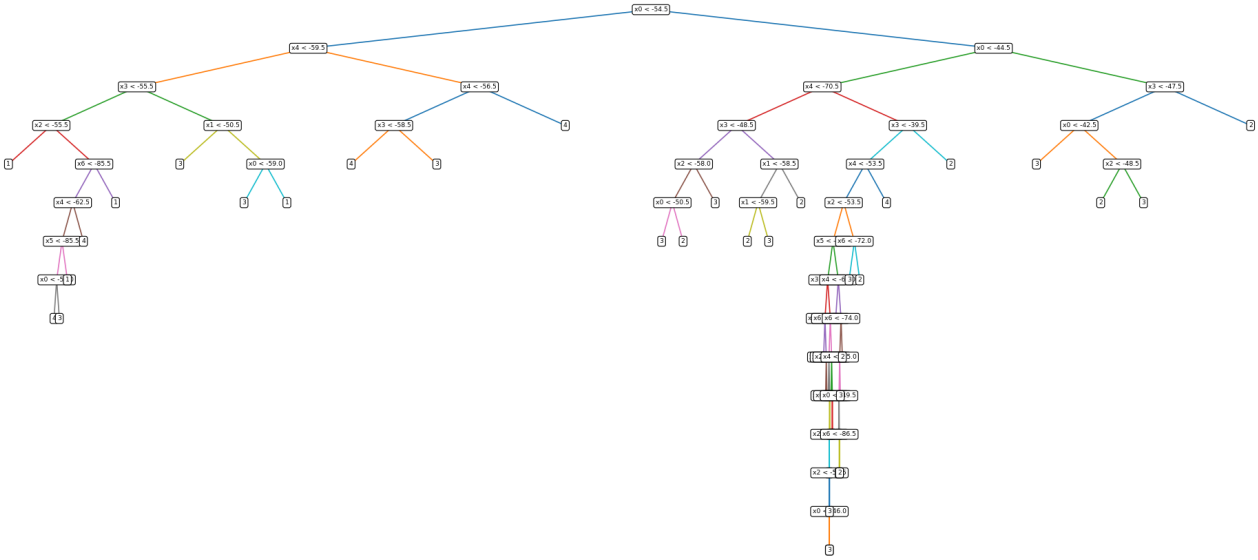

Figure 1: Clean tree visualised

2 Cross Validation Classification Metrics

2.1 Clean Dataset Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.2	0	0.5	0.3
Room 2 Actual	0	48	2	0
Room 3 Actual	0	2.2	47.5	0.3
Room 4 Actual	0.4	0	0.2	49.4

2.2 Clean Dataset Metrics

Accuracy : 0.9705

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.9919	0.9562	0.9462	0.9880
Recall Rates	0.9840	0.9600	0.9500	0.9880
F1 Measures	0.9880	0.9581	0.9481	0.9880

Macro Averaged F1: 0.9705

2.3 Noisy Dataset Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	38.2	3.2	3.4	4.2
Room 2 Actual	2.6	39.8	4.1	3.2
Room 3 Actual	2.5	4.2	41.6	3.2
Room 4 Actual	3.9	2.6	3.3	40.0

2.4 Noisy Dateset Metrics

Accuracy : 0.7980

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.8093	0.7992	0.7939	0.7905
Recall Rates	0.7796	0.8008	0.8078	0.8032
F1 Measures	0.7942	0.8000	0.8008	0.7968

Macro Averaged F1: 0.7979

3 Result Analysis

Room 4 has the most accurate readings across both data sets, even having a recall rate of 0.988 in the clean data set. Rooms 2 and 3 are noticeably less accurate and have a relatively high frequency of being confused between one another in both data sets. Finally, Room 1 had the widest range of performance between sets but was most confused with Room 4.

4 Dataset Analysis

There is around a 10% loss of accuracy between the clean and noisy data sets. Rooms 1 and 4 were affected more than 2 and 3 despite originally having greater accuracy. The noise means that the data set is less precise leading to the models formed being able to generate less accurate divisions. This causes more incorrect predictions as more data points are split incorrectly.

5 Cross Validation Classification Metrics After Pruning

5.1 Clean Dataset Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.8	0	0.2	0
Room 2 Actual	0	47.8	2.2	0
Room 3 Actual	0.3	2.2	47.1	0.4
Room 4 Actual	0.5	0	0.3	49.2

5.2 Clean Dataset Metrics

Accuracy : 0.9695

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.9842	0.9560	0.9458	0.9919
Recall Rates	0.9960	0.9560	0.9420	0.9840
F1 Measures	0.9901	0.9560	0.9439	0.9880

Macro Averaged F1: 0.9705

5.3 Noisy Dataset Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	44.5	1.0	1.4	2.1
Room 2 Actual	1.8	43.9	2.8	1.2
Room 3 Actual	2.4	3.1	44.4	1.6
Room 4 Actual	2.2	1.5	1.9	44.2

5.4 Noisy Dataset Metrics

Accuracy : 0.8850

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.8743	0.8869	0.8792	0.9002
Recall Rates	0.9082	0.8833	0.8621	0.8876
F1 Measures	0.8909	0.8851	0.8706	0.8938

Macro Averaged F1: 0.8851

6 Result Analysis After Pruning

From the metrics tables, the performance of the tree trained on clean data changes very little after pruning, whereas in the case of the noisy set all metrics are increased by 4 to 11%. Models trained on noisy data will fit to random noise and generalise poorly to other sets, so test set performance is improved by removing nodes that worsen performance on the validation set.

7 Depth Analysis

The average unpruned tree maximal depth is 12.4 and 19 for trees trained on clean and noisy sets. The corresponding pruned depths are 5.6 and 10.2. This is because noise increases the random spread of data, leading to trees with more decision rules. Pruning overfit trees will improve performance, but the clean set metrics show that depth is not directly linked to accuracy.