Machine Learning Coursework 1 - Decision Trees

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

October 2021

1 Cross Validation Classification Metrics

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

Clean Dataset Metrics:
Accuracy : 0.9705

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.99193548	0.9561753	0.94621514	0.988
Recall Rates	0.984	0.96	0.95	0.988
F1 Measures	0.98795181	0.95808383	0.94810379	0.988

Macro Averaged F1: 0.9705348579948536

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

Noisy Dateset Metrics:
Accuracy : 0.798

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.80932203	0.79919679	0.79389313	0.79051383
Recall Rates	0.77959184	0.80080483	0.80776699	0.80321285
F1 Measures	0.79417879	0.8	0.80076997	0.796812275

Macro Averaged F1: 0.7979403785772152

2 Result Analysis

Room 1 has the most accurate readings across both data sets, even having a recall rate of 0.996 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 4 had the widest range of performance but was most confused with Room 1.

3 Dataset Analysis

There is a 10-15% loss of accuracy between the clean and data sets across all rooms measured. Rooms 1 and 4 were affected to a greater extent than Rooms 2 and 3 despite originally having greater accuracy. The presence of noise means that the data set is less precise leading to the models formed being able to generate less specific divisions of data. This leads to a greater number of incorrect predictions as more data points are divided incorrectly.

4 Cross Validation Classification Metrics After Pruning

Clean Dataset Confusion Matrix:

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.8	0	0.1	0.1
Room 2 Actual	0	48.2	1.8	0
Room 3 Actual	0.4	2.8	46.4	0.4
Room 4 Actual	0.5	0	0.2	49.3

Clean Dataset Metrics:
Accuracy : 0.9685

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.98224852	0.94509804	0.95670103	0.98995984
Recall Rates	0.996	0.964	0.928	0.986
F1 Measures	0.98907646	0.95445545	0.94213198	0.98797595

Macro Averaged F1: 0.9684099604726415

Noisy Dataset Confusion Matrix:

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	43.1	1.6	1.8	2.5
Room 2 Actual	2.7	42.8	2.6	1.6
Room 3 Actual	2.0	3.8	43.8	1.9
Room 4 Actual	3.1	1.7	2.5	42.5

Noisy Dataset Metrics:
Accuracy : 0.861

	Room 1	Room 2	Room 3	Room 4
Precision Rates	0.84675835	0.85771543	0.86390533	0.87628866
Recall Rates	0.87959184	0.861167	0.85048544	0.85341365
F1 Measures	0.86286286	0.85943775	0.85714286	0.8646999

Macro Averaged F1: 0.8610358423200841

5 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.

6 Depth Analysis

The average unpruned tree maximal depth is around 12.4 and 19 for trees trained on clean and noisy sets. The corresponding pruned depths are 5.5 and 10. 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.