Imperial College
London

DECISION TREE COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Introduction to Machine Learning

Authors:
Henry Hausamann, Pablo Romo, Yusuf Salim, Lun Tan

Date: November 4, 2022

1 Visualisation of the decision tree

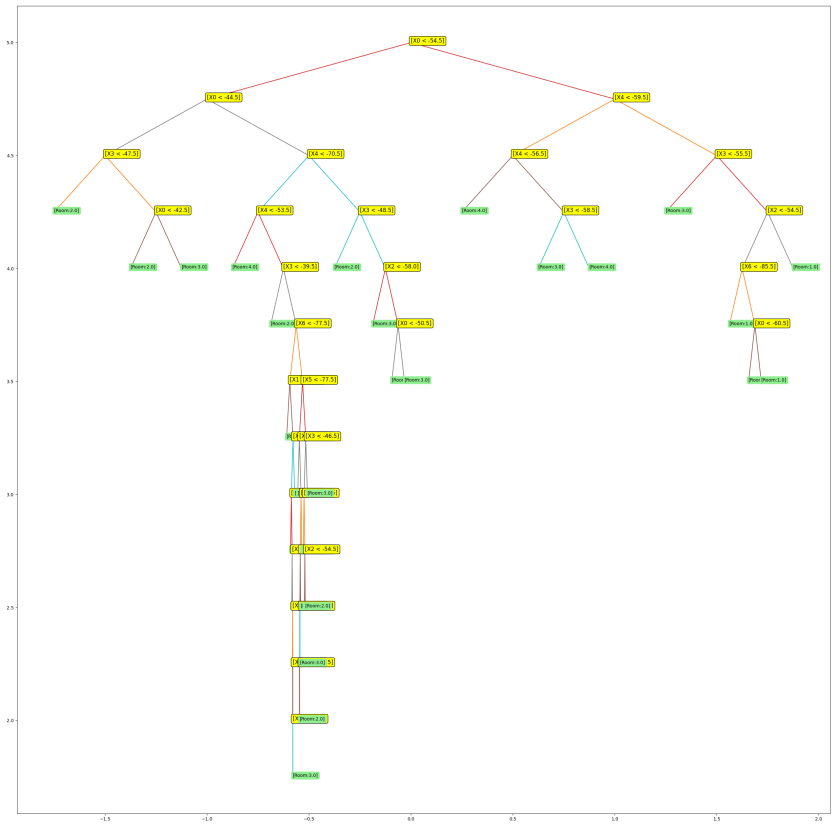

Figure 1: This is a tree trained on the entire clean dataset.

2 Cross validation classification metrics

2.1 Clean and noisy datasets' confusion matrices and metrics

The below results were obtained using 10-fold cross-validation.

Table 1: Average clean dataset confusion matrix

Room	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 1	49.2	0	0.1	0.7
Actual: 2	0	48.1	1.9	0
Actual: 3	0.6	1.7	47.6	0.1
Actual: 4	0.5	0	0.1	49.4

Table 2: Average clean dataset metrics

Metric	Room: 1	Room: 2	Room: 3	Room: 4
Accuracy	97.15%			
Precision	0.979	0.966	0.957	0.984
Recall	0.985	0.962	0.953	0.988
F1	0.982	0.963	0.955	0.986

Table 3: Average noisy dataset confusion matrix

Room	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 1	37.6	2.9	3.9	4.6
Actual: 2	2.4	41.3	3.6	2.4
Actual: 3	2.6	3.7	41	4.2
Actual: 4	3.1	2.3	4.1	40.3

Table 4: Average noisy dataset metrics

Metric	Room: 1	Room: 2	Room: 3	Room: 4
Accuracy	80.10%			
Precision	0.824	0.825	0.782	0.780
Recall	0.770	0.830	0.795	0.808
F1	0.795	0.827	0.787	0.792

2.2 Result analysis

Rooms 2, 3 and 4 are strongly recognised across both datasets. All 4 Rooms are recognised almost perfectly in the clean dataset, all having precision and recall above 95%. Room 1's classification performance drops the most between the datasets and is most commonly confused with Room 4. Room 2 is sometimes incorrectly classified as Room 3, whereas Room 3 is incorrectly classified as either Room 2 or 4.

2.3 Dataset differences

Room are classified almost perfectly in the clean dataset. With the noisy dataset, performance degrades by around 15-20% . Room 1 is most affected by noise, whereas Rooms 2 and 3, despite having worse performance initially, are less affected by it. In the noisy dataset, the model overfits to the noisy data, causing less accurate splits/decisions to be made so more rooms are incorrectly predicted.

3 Cross validation classification metrics after pruning

3.1 Clean and noisy dataset's confusion matrices and metrics after pruning

The below results were obtained using 10-fold nested cross-validation.

Table 5: Average pruned clean confusion matrix

Room	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 1	$50.2\overline{3}$	0	$0.1\overline{4}$	$0.2\overline{8}$
Actual: 2	0	$47.8\overline{7}$	0.9	0
Actual: 3	$0.1\overline{3}$	$0.9\overline{8}$	$48.4\overline{5}$	0.2
Actual: 4	$0.3\overline{2}$	0	$0.1\overline{6}$	$50.2\overline{8}$

Table 6: Average pruned clean metrics

Metric	Room: 1	Room: 2	Room: 3	Room: 4
Accuracy	98.43%			
Precision	0.990	0.980	0.975	0.990
Recall	0.991	0.983	0.974	0.991
F1	0.991	0.981	0.974	0.990

Table 7: Average pruned noisy matrix

Room	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 1	$41.2\overline{1}$	$2.2\overline{7}$	$2.2\overline{7}$	$3.0\overline{1}$
Actual: 2	$1.5\overline{3}$	$44.2\overline{7}$	$2.5\overline{1}$	$1.4\overline{5}$
Actual: 3	$1.4\overline{8}$	$2.7\overline{8}$	$45.7\overline{8}$	$2.0\overline{4}$
Actual: 4	$2.4\overline{2}$	$1.5\overline{1}$	$2.0\overline{3}$	$43.3\overline{6}$

Table 8: Average pruned noisy metrics

Metric	Room: 1	Room: 2	Room: 3	Room: 4
Accuracy	87.32%			
Precision	0.882	0.871	0.872	0.872
Recall	0.848	0.887	0.880	0.879
F1	0.863	0.878	0.874	0.874

3.2 Result analysis after pruning

Pruning marginally increases the performance on the clean dataset. On the noisy dataset, pruning increases performance by around 10%. The noisy dataset causes

the model to overfit to the noise, leading to trees that do not accurately model more general datasets. By pruning on nodes that give the same or better performance after evaluation on the validation set, the decision tree can generalise better.

3.3 Depth analysis

The average depths of the clean unpruned, clean pruned, noisy unpruned and noisy pruned trees are: 12.7, $11.\overline{8}$, 19.6 and $18.0\overline{6}$ respectively. Trees trained on clean data are close to the optimal depth where the classification rate peaks. Trees trained on noisy data have a higher than optimal depth which generalises poorly to unseen data, leading to lower accuracy, suggesting accuracy is linked to depth, but not linearly.