Introduction to Machine Learning - Coursework 1

Cornelius Braun, Ryan Lail, Ben Boyd, Shayaan Sindhoo

3rd November 2021

Tree Visualisation

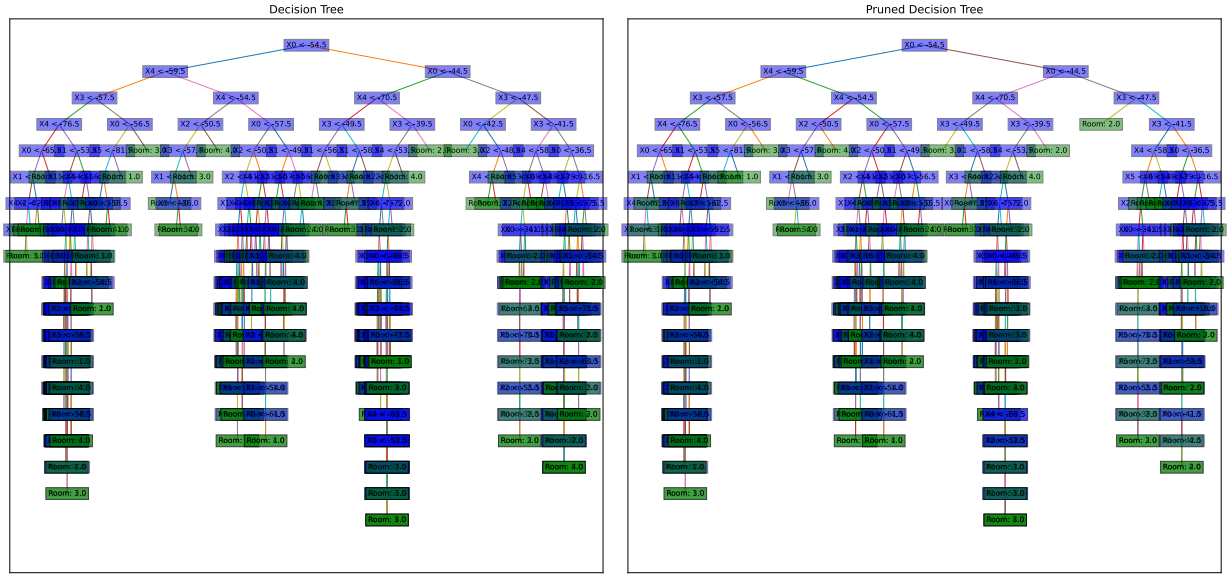

Figure 1: Visualisation of both unpruned (left) and pruned (right) decision trees.

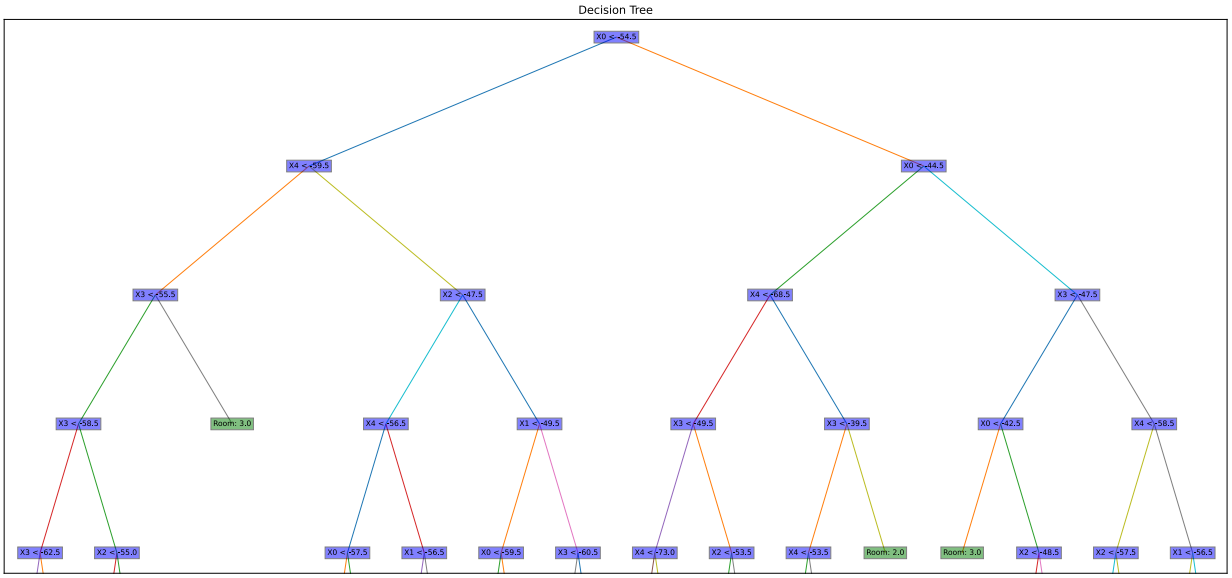

Figure 2: Close up of the base of a decision tree.

Question 3: Evaluation before Pruning

To evaluate the performance of our decision tree classifier, we report several metrics:

3.1 Cross-Validation Classification Metrics

actual \ predicted	**1**	**2**	**3**	**4**
1	49.3	0.00	0.20	0.50
2	0.00	47.90	2.10	0.00
3	0.10	1.90	47.70	0.30
4	0.40	0.00	0.20	49.40

(a) Clean dataset

actual \ predicted	**1**	**2**	**3**	**4**
1	38.80	2.80	3.00	4.40
2	2.80	40.10	4.50	2.30
3	3.10	3.80	41.20	3.40
4	3.20	2.80	4.00	39.80

(b) Noisy dataset

Table 1: Confusion matrices for the clean dataset (left) and noisy dataset (right).

	Clean Dataset				**Noisy Dataset**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Precision	0.99	0.96	0.95	0.98	0.81	0.81	0.78	0.80
Recall	0.99	0.96	0.95	0.99	0.79	0.81	0.80	0.80
F1	0.99	0.96	0.95	0.99	0.80	0.81	0.79	0.80
Accuracy	0.97				0.80			

Table 2: Performance metrics of our decision tree for the clean and noisy dataset.

3.2 Result Analysis

Table 2 shows that our classifier has an accuracy of 0.97 for the clean dataset. Rooms 1 and 4 perform best with F1 scores of 0.99. Rooms 2 and 3 have lower F1 scores as they are commonly mistaken for each other, as reflected in Table 1, leading to the reduced precision and recall shown in Table 2. The noisy dataset has a reduced collective accuracy of 0.80. Each room scores the same across all performance metrics when using the noisy data.

3.3 Dataset Differences

The classification on the clean dataset is near-perfect, whereas accuracy on the noisy dataset is worse by 0.17. The reason for this is because the classifier trained on the clean dataset learns only the causal relations in the data to make the prediction, whereas the classifier that is trained on the noisy dataset makes false relations due to random noise. These false relations lead to the misclassification of test cases which have randomly skewed signals different from the training set.

Question 4: Evaluation after Pruning

To evaluate the performance of our decision tree classifier with pruning, we report several metrics:

4.1 Cross Validation Classification Metrics After Pruning

actual \ predicted	**1**	**2**	**3**	**4**
1	49.53	0.00	0.21	0.26
2	0.00	47.84	2.16	0.00
3	0.39	1.74	47.6	0.27
4	0.44	0.00	0.21	49.34

(a) Clean dataset with pruning

actual \ predicted	**1**	**2**	**3**	**4**
1	39.70	2.76	3.18	3.37
2	2.71	40.79	4.04	2.16
3	2.48	3.57	42.23	3.02
4	3.73	2.11	3.11	40.84

(b) Noisy dataset with pruning

Table 3: Confusion matrices for the clean dataset (left) and noisy dataset (right) after pruning.

	Clean Dataset				**Noisy Dataset**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Precision	0.98	0.96	0.95	0.99	0.82	0.83	0.80	0.83
Recall	0.99	0.96	0.95	0.99	0.81	0.82	0.82	0.82
F1	0.99	0.96	0.95	0.99	0.81	0.82	0.81	0.82
Accuracy	0.97				0.82			

Table 4: Performance metrics of our decision tree for the clean and noisy dataset.

4.2 Result Analysis After Pruning

Comparing Table 2 and Table 4 we see that the accuracy of the clean dataset is unchanged as a result of pruning whilst the accuracy of the noisy dataset increases by 0.02. Our pruning strategy removes redundant nodes in the tree which reduces overfitting. Despite this, we still expect the performance on a noisy dataset to be worse than a clean dataset as noise is unpredictable, which leads to misclassification.

4.3 Depth Analysis

	Clean Dataset	**Noisy Dataset**
Not Pruned	13.10	20.40
Pruned	12.29	18.71

Table 5: Comparison of the average depth of trees before and after pruning.

The depth of a tree reflects the number of decisions needed to segment the data. We expect the depth of a noisy tree to be deeper than that of a clean tree as unnecessary decisions are made due to the noise. When pruning nodes in clean trees we prune decisions that result in overfitting, while in noisy trees we can also prune the decisions made due to the noise. Hence, why in Table 5 we see a larger decrease in the maximal depth of the noisy tree compared to the clean tree.