Introduction to Machine Learning Coursework One

Decision Trees

Chenghong Ren, Daniel Ong, Arthika Sivathasan

October 2022

Contents

1 Output Visualisation of the Decision Tree

Task: Show in your report the output of this function with a tree trained on the entire clean dataset

Below is the visualisation of the unpruned decision tree, trained on the clean dataset without any cross validation.

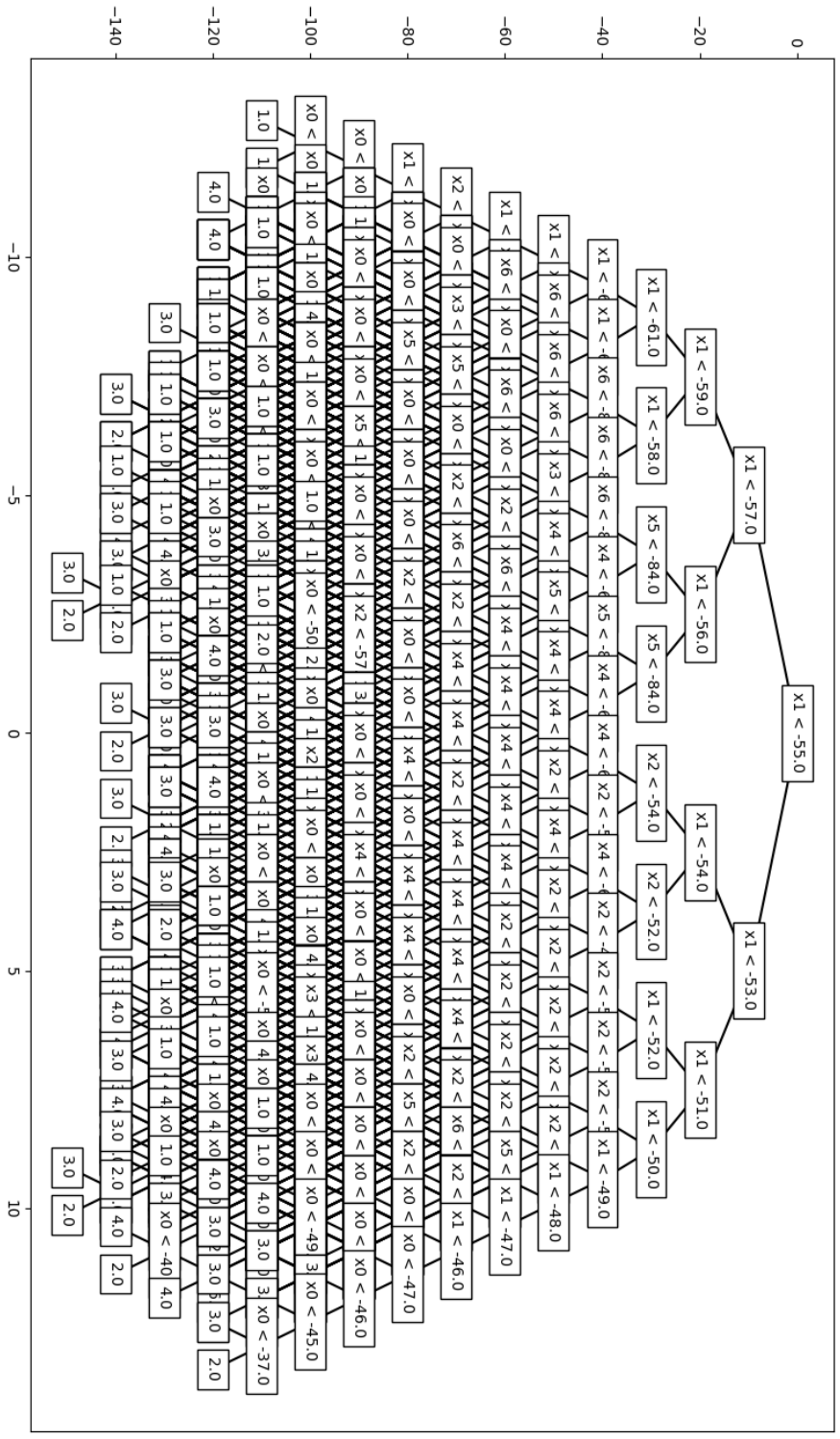

Figure 1: Visualisation of the Decision Tree (Rotated Image)

2 Evaluation

2.1 Cross Validation Classification Metrics

Task: By computing the average over all the test folds, report the following cross-validation classification metrics for both clean and noisy data

2.1.1 Confusion Matrix

Below is the confusion matrix for the clean dataset, averaged across the 10 test folds:

	Room 1 (Predicted)	Room 2 (Predicted)	Room 3 (Predicted)	Room 4 (Predicted)
Room 1 (Actual)	42.3	0.2	1.2	6.3
Room 2 (Actual)	6.8	28.9	14.2	0.1
Room 3 (Actual)	5.8	0.6	33.8	9.8
Room 4 (Actual)	7.2	0.0	1.0	41.8

Table 1: Confusion Matrix for the Clean Dataset

Below is the confusion matrix for the noisy dataset, averaged across the 10 test folds:

	Room 1 (Predicted)	Room 2 (Predicted)	Room 3 (Predicted)	Room 4 (Predicted)
Room 1 (Actual)	35.4	1.0	4.3	8.3
Room 2 (Actual)	8.4	26.3	12.5	2.5
Room 3 (Actual)	7.5	3.5	28.6	11.9
Room 4 (Actual)	9.3	1.1	3.9	35.5

Table 2: Confusion Matrix for the Noisy Dataset

2.1.2 Classification Metrics

Below are tables summarising and comparing the different classification metrics for the noisy and clean dataset:

	Clean Dataset	Noisy Dataset
Accuracy	0.734	0.629

Table 3: Accuracy for the Clean Dataset and Noisy Dataset

Recall	Clean Dataset	Noisy Dataset
Room 1	0.846	0.722
Room 2	0.578	0.529
Room 3	0.676	0.555
Room 4	0.836	0.713

Table 4: Recall for the Clean Dataset and Noisy Dataset

Precision	Clean Dataset	Noisy Dataset
Room 1	0.681	0.584
Room 2	0.973	0.824
Room 3	0.673	0.580
Room 4	0.725	0.610

Table 5: Precision for the Clean Dataset and Noisy Dataset

F1 Measure	Clean Dataset	Noisy Dataset
Room 1	0.755	0.646
Room 2	0.725	0.644
Room 3	0.674	0.567
Room 4	0.777	0.657

Table 6: F1 Measure for the Clean Dataset and Noisy Dataset

2.2 Result Analysis

Task: Comment for both datasets which rooms are recognized with high/low accuracy, and which rooms are confused. 5 lines max.

For both datasets, Room 1 has the highest recall but relatively low precision. Yet, Room 2 has the highest precision but the lowest recall. This suggests false negatives for Room 2 exist, which could be due to Room 2 being sometimes confused as Room 3. But, any predictions labelled as Room 2, are likely to be correct. Comparing F1 values, we conclude that Room 4 is more accurate on the whole and Room 3 is recognised with the lowest accuracy.

2.3 Dataset Differences

Task: Is there any difference in the performance when using the clean and noisy datasets? If yes/no explain why. 5 lines max.

As expected, the noisy data has reduced the accuracy of the decision tree. Similarly, the precision and recall values for the tree have also decreased, which has resulted in reduced F1 measure values. Therefore, overall, the performance of the model has decreased. This is because the noisy data has introduced instances where the data does not correlate to the other information provided and so the model struggles to identify the correct label.

3 Pruning

3.1 Cross Validation Classification Metrics After Pruning

Task: Report the performances of your trees after pruning by using a nested 10-fold cross-validation ("option 2") to compute the metrics defined in the previous section for both datasets.

3.1.1 Confusion Matrix

Below is the confusion matrix for the clean dataset, averaged across the 10 test folds:

	Room 1 (Predicted)	Room 2 (Predicted)	Room 3 (Predicted)	Room 4 (Predicted)
Room 1 (Actual)	43.0	0.2	1.4	5.4
Room 2 (Actual)	0.8	45.4	3.8	0.0
Room 3 (Actual)	4.2	0.9	39.1	5.8
Room 4 (Actual)	3.7	0.0	1.2	45.1

Table 7: Confusion Matrix for the Clean Dataset

Below is the confusion matrix for the noisy dataset, averaged across the 10 test folds:

	Room 1 (Predicted)	Room 2 (Predicted)	Room 3 (Predicted)	Room 4 (Predicted)
Room 1 (Actual)	38.0	1.7	3.7	5.6
Room 2 (Actual)	3.0	39.9	4.4	2.4
Room 3 (Actual)	6.2	2.4	35.8	7.1
Room 4 (Actual)	5.1	1.9	2.4	40.4

Table 8: Confusion Matrix for the Noisy Dataset

3.1.2 Classification Metrics

Below are tables summarising and comparing the different classification metrics for the noisy and clean dataset:

	Clean Dataset	Noisy Dataset
Accuracy	0.863	0.771

Table 9: Accuracy for the Clean Dataset and Noisy Dataset

Recall	Clean Dataset	Noisy Dataset
Room 1	0.863	0.773
Room 2	0.905	0.803
Room 3	0.781	0.695
Room 4	0.903	0.813

Table 10: Recall for the Clean Dataset and Noisy Dataset

Precision	Clean Dataset	Noisy Dataset
Room 1	0.831	0.723
Room 2	0.977	0.866
Room 3	0.861	0.774
Room 4	0.801	0.727

Table 11: Precision for the Clean Dataset and Noisy Dataset

F1 Measure	Clean Dataset	Noisy Dataset
Room 1	0.844	0.745
Room 2	0.938	0.832
Room 3	0.819	0.730
Room 4	0.847	0.764

Table 12: F1 Measure for the Clean Dataset and Noisy Dataset

3.2 Result Analysis After Pruning

Task: Comment the difference in performance before and after pruning for both datasets. Briefly explain these performance differences. 5 lines max.

Looking at the data above, we can conclude that the performance of the model has increased after pruning, with regards to all the classification metrics. This is because the pruned tree no longer overfits the training data and is a more generalised model. More generalised models are better at classifying unseen data and are more robust against noise, which improves the overall performance of the model.

3.3 Depth Analysis

Task: Comment on the average depth of the trees that you generated for both datasets, before and after pruning. What can you tell about the relationship between maximal depth and prediction accuracy? 5 lines max.

Surprisingly, the average depth of the tree, before and after pruning, has remained the same: 15 for the clean dataset and 14 for the noisy dataset. But the prediction accuracy of the model has increased, probably due to removal of nodes within the width of the tree as opposed to depth. This is because pruning removes redundant and disadvantageous branches which will ultimately reduce the size of the tree without affecting performance.