Decision Tree Coursework - Report

Leonardo Garofalo, Karim Khairaz, Nicholas Pfaff, Bradley Stanley-Clamp

Contents

1 (Bonus points: Output of the tree visualization function) 2

2 Step 3 - Evaluation 2
2.1 Cross validation classification metrics . . . 2
2.1.1 Confusion Matrix . . . 2
2.1.2 Accuracy per dataset . . . 3
2.1.3 Recall rates per class . . . 3
2.1.4 Precision rates per class . . . 3
2.1.5 F1-measures per class . . . 3
2.2 Result analysis . . . 3
2.3 Dataset differences . . . 4

3 Step 4 - Pruning (and evaluation again) 4
3.1 Cross validation classification metrics after pruning . . . 4
3.1.1 Accuracy per dataset . . . 4
3.1.2 Recall rates per class . . . 4
3.1.3 Precision rates per class . . . 5
3.1.4 F1-measures per class . . . 5
3.2 Result analysis after pruning . . . 5
3.3 Depth analysis . . . 5

1 (Bonus points: Output of the tree visualization function)

The visual representation of a decision tree trained on the whole of the clean data set can be seen below in two images. The first is the full tree, and the second is a zoomed in section the tree.

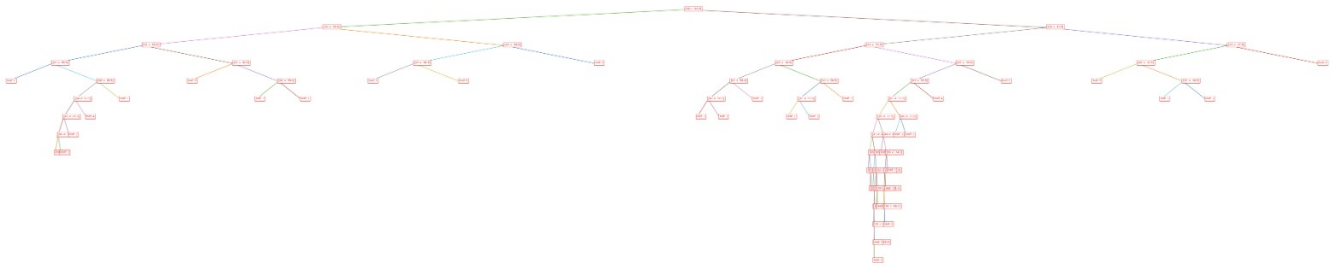

Figure 1: A visual representation of a decision tree

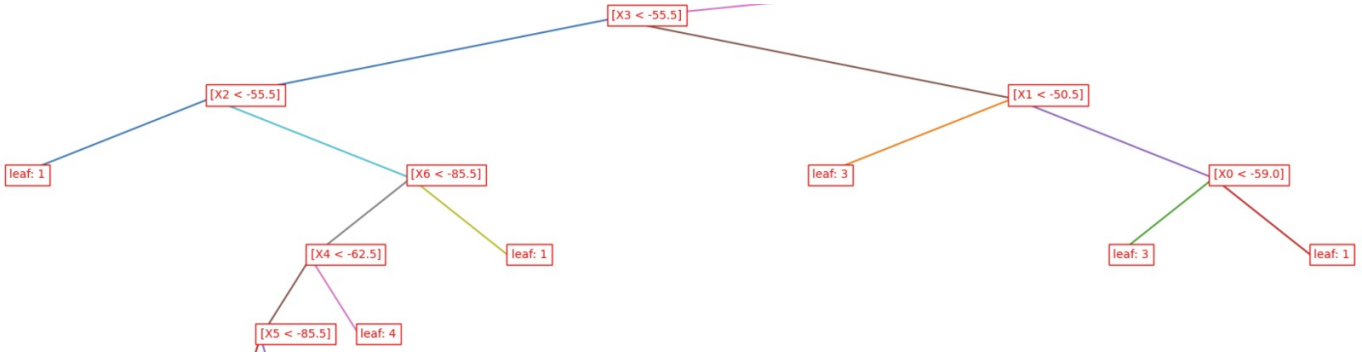

Figure 2: A zoomed section of a decision tree

2 Step 3 - Evaluation

2.1 Cross validation classification metrics

All evaluation metric data values have been corrected to 4 decimal places.

2.1.1 Confusion Matrix

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.6000	0.0000	0.2000	0.2000
Room 2 Actual	0.0000	47.8000	2.2000	0.0000
Room 3 Actual	0.2000	1.8000	47.7000	0.3000
Room 4 Actual	0.5000	0.0000	0.1000	49.4000

Table 1: Confusion Matrix for Clean Dataset (Unpruned Tree)

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	38.6000	3.1000	2.8000	4.5000
Room 2 Actual	2.8000	40.3000	4.3000	2.3000
Room 3 Actual	2.5000	3.9000	41.3000	3.8000
Room 4 Actual	4.0000	2.4000	3.7000	39.7000

Table 2: Confusion Matrix for Noisy Dataset (Unpruned Tree)

2.1.2 Accuracy per dataset

	Accuracy Rate
Clean Dataset	0.9725
Noisy Dataset	0.7995

Table 3: Accuracy rate per dataset for Unpruned Tree

2.1.3 Recall rates per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9920	0.956	0.9540	0.9880	0.9725
Noisy Dataset	0.7878	0.8109	0.8019	0.7972	0.7994

Table 4: Recall rates per class for Unpruned Tree

2.1.4 Precision rates per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9861	0.9637	0.9502	0.9900	0.9725
Noisy Dataset	0.8058	0.8109	0.7927	0.7893	0.7997

Table 5: Precision rates per class for Unpruned Tree

2.1.5 F1-measures per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9890	0.9598	0.9521	0.9890	0.97252
Noisy Dataset	0.7967	0.8109	0.7973	0.7932	0.7995

Table 6: F1-measure per class for Unpruned Tree

2.2 Result analysis

Rooms are recognized with higher accuracy if their F1 measure is higher than the macro-averaged F1-measure and vice versa. The F1 measure was chosen to measure "class accuracy" as it takes into account with equal weights both recall and precision. This analysis indicates that for the clean dataset, room 1 and 4 are recognized with higher accuracy than room 2 and 3. For the noisy dataset, this is less clear: all rooms are recognised with similar accuracy, with room 2 having slightly higher accuracy than the others.

The confusion matrix shows whether rooms are commonly confused with each other. For the clean dataset, it can be seen that rooms 2 and 3 are sometimes confused. For the noisy dataset, all rooms are sometimes confused.

2.3 Dataset differences

The decision tree algorithm achieves lower performance (accuracy, macro-averaged recall, macro-averaged precision and macro-averaged F1-measure) on the noisy dataset than on the clean dataset.

The lower performance on the noisy dataset can be explained by overfitting: The decision tree models the training set very closely, which contains noise, and hence generalizes poorly. This is less of a problem for the clean dataset as it does not contain noise.

3 Step 4 - Pruning (and evaluation again)

All evaluation metric data values have been corrected to 4 decimal places.

3.1 Cross validation classification metrics after pruning

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	49.7444	0.0000	0.2556	0.0000
Room 2 Actual	0.0000	47.8000	2.2000	0.0000
Room 3 Actual	0.5556	2.0333	47.1444	0.2667
Room 4 Actual	0.5111	0.0000	0.2667	49.2222

Table 7: Confusion Matrix for Clean Dataset (Pruned Tree)

	Room 1 Predicted	Room 2 Predicted	Room 3 Predicted	Room 4 Predicted
Room 1 Actual	44.1778	1.0333	1.5889	2.2000
Room 2 Actual	2.1000	43.8444	2.5111	1.2444
Room 3 Actual	2.2222	3.4556	44.0667	1.7556
Room 4 Actual	2.3778	1.4778	2.1333	43.8111

Table 8: Confusion Matrix for Noisy Dataset (Pruned Tree)

3.1.1 Accuracy per dataset

	Accuracy Rate
Clean Dataset	0.9696
Noisy Dataset	0.8795

Table 9: Accuracy rate per dataset for Pruned Tree

3.1.2 Recall rates per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9949	0.9560	0.9429	0.9844	0.9696
Noisy Dataset	0.9016	0.8822	0.8557	0.8797	0.8798

Table 10: Recall rate per class for Pruned Tree

3.1.3 Precision rates per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9790	0.9592	0.9454	0.9946	0.9696
Noisy Dataset	0.8683	0.8802	0.8760	0.8939	0.8796

Table 11: Precision rate per class for Pruned Tree

3.1.4 F1-measures per class

Class:	Room 1	Room 2	Room 3	Room 4	Macro-averaged
Clean Dataset	0.9869	0.9576	0.9441	0.9895	0.9695
Noisy Dataset	0.8846	0.8812	0.8657	0.8868	0.8796

Table 12: F1-measure per class for Pruned Tree

3.2 Result analysis after pruning

The performance measures (accuracy, macro-averaged recall, macro-averaged precision and macro-averaged F1-measure) for the clean dataset remain approximately unchanged with pruning. The same performance measures for the noisy dataset improve significantly after pruning.

Pruning reduces overfitting by decreasing the complexity of the model. Overfitting occurs in the noisy dataset as we are modelling the training dataset very closely and hence the noise contained within it. However, overfitting does not occur in the clean dataset as there is no noise to model. Consequently, pruning improves the performance on the noisy datset but does not improve performance on the clean dataset.

3.3 Depth analysis

Our implementation of the pruning algorithm prunes a node if the validation error decreases or stays the same.

Pruning decreases the average depth for both datasets, but slightly more for the noisy dataset. The tree depth decreases more for the noisy dataset as due to overfitting the model benefits from decreased complexity. There is not much overfitting for the clean dataset and hence it is not possible to decrease the validation error significantly by pruning.

As we decrease the maximum depth of the tree through pruning, the model's accuracy increases or stays the same. However, this trend does not continue indefinitely as beyond a certain point decreasing tree depth decreases accuracy due to underfitting. Nevertheless, the pruning algorithm should not decrease depth beyond this point.

Average tree depth:	Without pruning	With pruning
Clean dataset	12.7333	9.2111
Noisy dataset	19.4444	14.3889

Table 13: Average tree depth for Pruned and Unpruned tree