Decision Tree Report

Alan Liang, Mark Chen, Leyang Yu, Xinan Xu

30th Oct 2022

1 Loading Data

2 Creating Decision Trees

2.1 Bonus part

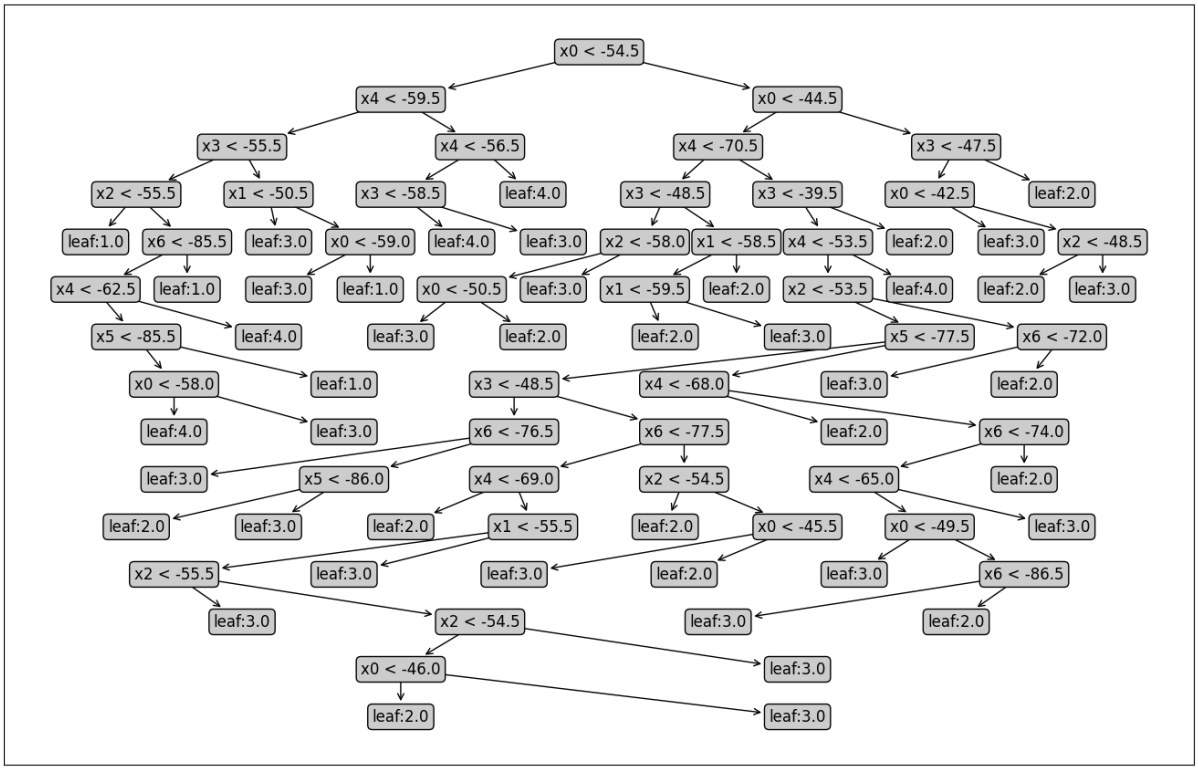

Figure 1: Visualization of a tree trained on the entire clean dataset, before pruning

3 Evaluation

3.1 Cross validation classification metrics before pruning

3.1.1 Confusion Matrix

Clean dataset:

classes\predictions	1	2	3	4
1	49.4	0.0	0.4	0.2
2	0.0	48.1	1.9	0.0
3	0.3	1.8	47.8	0.1
4	0.5	0.0	0.2	49.3

Noisy dataset:

classes\predictions	1	2	3	4
1	37.5	3.4	3.9	4.2
2	2.7	40.4	3.8	2.8
3	2.6	3.8	41.7	3.4
4	3.8	2.8	3.7	39.5

3.1.2 Accuracy

The macro mean accuracy of the clean dataset is 0.973.
The macro mean accuracy of the noisy dataset is 0.7955.

3.1.3 Recall and Precision

Recall by class:

class	1	2	3	4
Clean dataset	0.9870	0.9629	0.9587	0.9862
Noisy dataset	0.7666	0.8115	0.8076	0.7933

Precision by class:

class	1	2	3	4
Clean dataset	0.9846	0.9638	0.9518	0.9940
Noisy dataset	0.8079	0.8010	0.7890	0.7925

3.1.4 F1 measures

F1 measures by class:

class	1	2	3	4
Clean dataset	0.9856	0.9631	0.9548	0.9899
Noisy dataset	0.7852	0.8052	0.7966	0.7920

The average F1 measure of the clean dataset is 0.9734.
The average F1 measure of the noisy dataset is 0.7948.

3.2 Result analysis

For the clean dataset, the rooms from the highest prediction accuracy to the lowest(based on F1 measure) are room 4, 1, 2 and 3. Room 2 and 3 are most often confused based on the number of false positives/false negatives in the confusion matrix. For the noisy dataset, the rooms from the highest prediction accuracy to the lowest are room 2, 3, 4 and 1. All rooms are sometimes confused, yet confusions between room 1 and 4, room 2 and 3 are more often.

3.3 Dataset differences

The accuracy of the clean dataset is 17.75% higher than the accuracy of the noisy dataset. The average depth of the trees generated from noisy dataset(19.3) is larger by 7 layers than that of the clean dataset(12.3). There are more noisy data in the noisy dataset and our decision tree models the noise. Hence, overfitting on the training set would result in a larger error on test set for the noisy dataset.

4 Pruning

4.1 Cross validation classification metrics after pruning

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.

4.1.1 confusion matrix

Clean dataset after pruning:

classes\predictions	1	2	3	4
1	49.6	0.0	0.3	0.1
2	0.0	47.9	2.1	0.0
3	0.6	2.5	46.6	0.3
4	0.5	0.0	0.4	49.1

Noisy dataset after pruning:

classes\predictions	1	2	3	4
1	44.2	1.2	1.5	2.2
2	2.0	43.9	2.6	1.2
3	2.2	3.3	44.1	1.8
4	2.3	1.6	2.0	43.9

4.1.2 accuracy

The macro mean accuracy of the clean dataset after pruning is 0.9663.
The macro mean accuracy of the noisy dataset after pruning is 0.8806.

4.1.3 recall and precision

Recall by class:

class	1	2	3	4
Clean dataset	0.9920	0.9592	0.9336	0.9822
Noisy dataset	0.9007	0.8809	0.8545	0.8821

Precision by class:

class	1	2	3	4
Clean dataset	0.9799	0.9515	0.9452	0.9925
Noisy dataset	0.8719	0.8810	0.8806	0.8926

4.1.4 F1 measures

F1 measures by class:

class	1	2	3	4
Clean dataset	0.9857	0.9548	0.9384	0.9871
Noisy dataset	0.8851	0.8799	0.8661	0.8867

The F1 measure of the clean dataset after pruning is 0.9665.
The F1 measure of the noisy dataset after pruning is 0.8794.

4.2 Result analysis after pruning

For the clean dataset, the accuracy after pruning drops by 0.0067. The difference might just be random statistical noise. For the noisy dataset, the accuracy increases by 0.0852, which is quite significant. This shows that pruning reduces overfitting and makes our model more robust on unseen dataset. Pruning reduces the depth, hence reducing the high model capacity, leading to lower variance and better performance.

4.3 Depth analysis

After pruning, the mean depth of the noisy dataset dropped by 28% from 19.3 to 13.9, while for the clean dataset, it falss by 33% from 12.3 to 8.3. The mean depth of the noisy dataset is larger than that of the clean dataset but its accuracy is lower. While an overly small maximal depth means low model capacity and high bias, an overly high maximal depth might model some noise in dataset and result in high variance and lower prediciton accuracy.