Decision Tree Coursework Report

TszHang Wong, Quix Fan, Xuan Cai, Kaiwen Liu

November 4, 2022

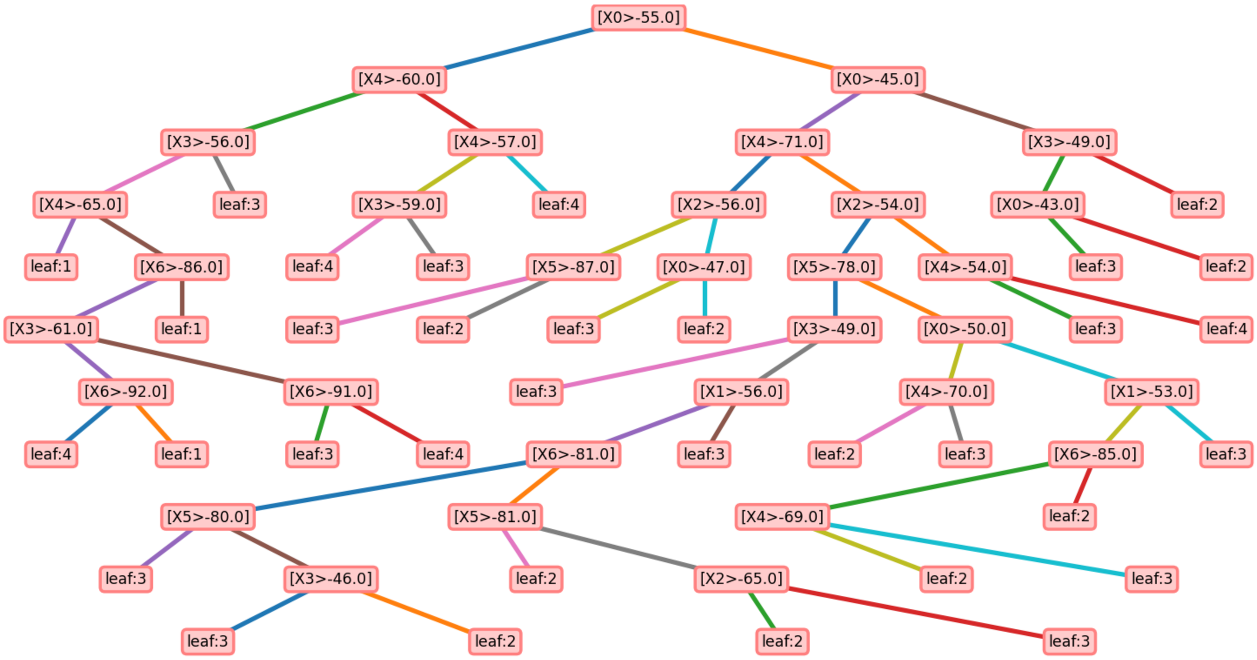

Figure 1: Tree Diagram of clean data set

1 Evaluation

1.1 10-Folds Cross validation classification metrics

	Room1 Predicted	Room2 Predicted	Room3 Predicted	Room4 Predicted
Room1 Actual	49.7	0	0.2	0.1
Room2 Actual	0	48.3	1.7	0
Room3 Actual	0.2	1.6	48	0.2
Room4 Actual	0.4	0	0.2	49.4

Table 1: Clean set confusion matrix.

	Room1 Predicted	Room2 Predicted	Room3 Predicted	Room4 Predicted
Room1 Actual	38.5	3.4	3.4	3.7
Room2 Actual	2.9	40	3.8	3
Room3 Actual	3.3	3.6	41.9	2.7
Room4 Actual	3.5	2.4	3.3	40.6

Table 2: Noisy set confusion matrix.

	Precision	Recall	F1-score	Accuracy	$\sigma_{Accuracy}$
Room1	0.9880	0.9940	0.9910	0.9770	0.0090
Room2	0.9679	0.9660	0.9670		
Room3	0.9580	0.9600	0.9590		
Room4	0.9939	0.9880	0.9910		

Table 3: Clean set average metrics.

	Precision	Recall	F1-score	Accuracy	$\sigma_{Accuracy}$
Room1	0.7988	0.7857	0.7922	0.8050	0.0256
Room2	0.8097	0.8048	0.8073		
Room3	0.7996	0.8136	0.8065		
Room4	0.8120	0.8153	0.8136		

Table 4: Noisy set average metrics.

1.2 Result analysis

Among the four rooms in the clean set, Room2 and Room3 are relatively low in precision, recall, and f1-score. These two rooms are confused as their confusing cases are the highest beside the diagonal shown in the confusion matrix. In comparison, Room4 has the highest performance among all the metrics. Regarding the noisy set, Room1 and Room4 has the least accuracy and are confused, as most false predictions lie between them in the confusion matrix. Moreover, the same performance is shown in predicting Room3 and Room2.

1.3 Dataset differences

In contrast to the noisy dataset, all the rooms in the clean dataset perform much better in all metrics. According to the training algorithm, a decision tree must separate the training set into subsets with a pure label to reach the optimal information gain. Hence, each noisy data is considered and captured by the model. Different labels may mix up in high-dimensional space due to noise. Thus more nodes are needed to separate them, resulting in a more complex tree, which leads to overfitting and poor generalization performance.

2 Pruning - evaluation

2.1 10-Folds Cross validation classification metrics after pruning

		Room1 Predicted	Room2 Predicted	Room3 Predicted	Room4 Predicted
Room1 Actual	Before	49.30	0	0.37	0.33
	After	**49.60**	0	0.38	0.01
Room2 Actual	Before	0	**47.90**	2.10	0
	After	0	47.62	2.38	0
Room3 Actual	Before	0.20	1.90	**47.51**	0.39
	After	0.69	2.09	46.8	0.46
Room4 Actual	Before	0.37	0	0.18	**49.50**
	After	0.45	0	0.26	49.29

Table 5: Clean set confusion matrix before and after pruning.

		Room1 Predicted	Room2 Predicted	Room3 Predicted	Room4 Predicted
Room1 Actual	Before	38.3	3.43	3.22	4.03
	After	**44**	1.30	1.44	2.26
Room2 Actual	Before	3	39.79	4.07	2.84
	After	1.90	**44.04**	2.54	1.21
Room3 Actual	Before	2.87	3.66	41.83	3.14
	After	2.22	3.31	**44.08**	1.89
Room4 Actual	Before	3.72	2.48	3.43	40.17
	After	2.52	1.43	1.71	**44.13**

Table 6: Noisy set confusion matrix before and after pruning.

		Precision	Recall	F1-score		Accuracy	$\sigma_{Accuracy}$	Depth$_{avg}$
Room1	Before	**0.9886**	0.9860	**0.9873**	Before	**0.9708**	0.0144	11.88
	After	0.9775	**0.9922**	0.9848				
Room2	Before	**0.9618**	**0.9580**	**0.9599**				
	After	0.9580	0.9524	0.9552				
Room3	Before	**0.9473**	**0.9502**	**0.9487**	After	0.9664	0.0157	8.29
	After	0.9395	0.9353	0.9374				
Room4	Before	0.9856	**0.9891**	0.9873				
	After	**0.9906**	0.9858	**0.9882**				

Table 7: Clean set average metrics before and after pruning.

		Precision	Recall	F1-score		Accuracy	$\sigma_{Accuracy}$	Depth$_{avg}$
Room1	Before	0.7998	0.7819	0.7907	Before	0.8005	0.0288	18.28
	After	**0.8688**	**0.8980**	**0.8831**				
Room2	Before	0.8062	0.8006	0.8034				
	After	**0.8793**	**0.8862**	**0.8828**				
Room3	Before	0.7960	0.8123	0.8041	After	**0.8813**	0.0288	13.42
	After	**0.8855**	**0.8559**	**0.8704**				
Room4	Before	0.8003	0.8066	0.8034				
	After	**0.8918**	**0.8862**	**0.8870**				

Table 8: Noisy set average metrics before and after pruning.

2.2 Result analysis after pruning

A slight decrease in average accuracy is shown after pruning for the clean set because of the loss of non-noisy information captured within the train set. In contrast, the accuracy increased considerably in the noisy dataset after pruning. Pruning eliminated the captured noise by replacing nodes with the dominant label set when accuracy remains reasonable, resulting in a less overfitted decision tree with a higher bias and lower variance than before.

2.3 Depth analysis

The tree trained on the noisy dataset is shown to have a much larger depth than the clean dataset because more nodes are needed to separate data under the noisy data pattern. Pruning resulted in a tree with less average depth on both datasets, with prediction accuracy increasing in the noisy dataset and a mighty decrease in the clean set. Overall, the tree's generalization performance increases with less complexity and smaller maximum depth.