Introduction to ML - Decision Tree Coursework

Luca Chammah, Anthony Jones, Chris Myers, Rohan Gandhi

October - November 2022

Contents

1 Introduction

This report presents the results of decision tree classification algorithms to predict the room location of a mobile phone based on WiFi signal strengths. The given data is a 2000x8 array, with each row containing seven WiFi signal strengths and the corresponding room number label. The following sections analyse the performance of models trained on both clean and noisy datasets, and for each before pruning and after pruning. The performance metrics analysed include accuracy, precision, recall and F1-measures, visualised in confusion matrices.

2 Tree visualization

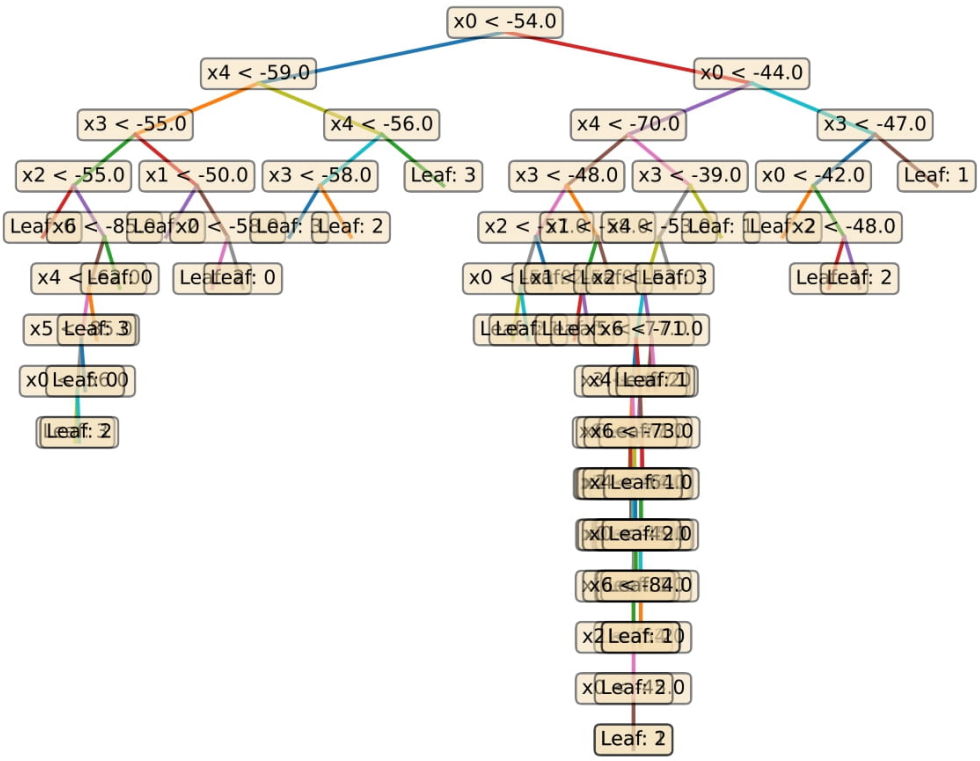

Figure 1: Visualisation of a tree trained on the entire clean data set.

3 Evaluation before pruning

3.1 Cross validation classification metrics

3.1.1 Confusion matrix

Clean dataset

$$\begin{bmatrix} 49 & 0 & 0 & 0 \\ 0 & 48 & 2 & 0 \\ 0 & 2 & 48 & 0 \\ 0 & 0 & 0 & 49 \end{bmatrix}$$

Noisy dataset

$$\begin{bmatrix} 39 & 3 & 3 & 4 \\ 3 & 40 & 4 & 3 \\ 3 & 4 & 41 & 3 \\ 4 & 2 & 4 & 40 \end{bmatrix}$$

Table 1: Confusion matrix for clean and noisy datasets

3.1.2 Accuracy

Clean dataset	Noisy dataset
97.37%	79.76%

Table 2: Accuracy percentage for clean and noisy datasets

3.1.3 Recall and precision per class

	Recall		Precision	
Room	Clean dataset	Noisy dataset	Clean dataset	Noisy dataset
1	98.83%	78.57%	98.70%	79.50%
2	96.13%	80.88%	96.75%	80.89%
3	95.62%	79.81%	95.42%	78.55%
4	98.80%	79.62%	99.00%	80.48%

Table 3: Precision and recall for clean and noisy datasets per class

3.1.4 F1-measures

Room	Clean dataset	Noisy dataset
1	98.76%	78.87%
2	96.41%	80.69%
3	95.48%	79.02%
4	98.89%	79.85%

Table 4: F1-measures for clean and noisy datasets per class

3.2 Result analysis

In the clean dataset, all rooms are highly accurately predicted, and the confusion matrix demonstrates confusion between rooms 2 and 3, shown by their lower F1-measures. This can be attributed to the WiFi point that exists between these rooms. However, the noisy data has an even spread of mispredictions, demonstrating the noise in this dataset. Room 4 has the highest precision due to a WiFi point being near room 4 only. Room 2 has the highest recall because there are several WiFi points around it and so has the most data to classify it.

3.3 Dataset differences

There is a noticeable difference in the accuracies of the trees trained on the two datasets, with the clean tree achieving 97.37% accuracy, but the noisy tree achieving only around 79.76%. This has been attributed to the clean dataset being representative of the system's actual behaviour; therefore, a model trained on this data would be very well fitted to the system. The noisy dataset, however, introduces an element of error to the WiFi readings, and so a model trained on this data cannot be as accurate.

4 Evaluation after pruning

4.1 Cross validation classification metrics

4.1.1 Confusion matrix

Clean dataset	Noisy dataset
$\begin{bmatrix} 50 & 0 & 0 & 0 \\ 0 & 48 & 2 & 0 \\ 0 & 2 & 47 & 0 \\ 0 & 0 & 0 & 49 \end{bmatrix}$	$\begin{bmatrix} 44 & 1 & 2 & 2 \\ 2 & 43 & 3 & 1 \\ 2 & 4 & 44 & 2 \\ 3 & 2 & 2 & 44 \end{bmatrix}$

Table 5: Confusion matrix for clean and noisy datasets

4.1.2 Accuracy

Clean dataset	Noisy dataset
96.89%	87.43%

Table 6: Accuracy percentage for clean and noisy datasets

4.1.3 Recall and precision per class

	Recall		Precision	
Room	**Clean dataset**	**Noisy dataset**	**Clean dataset**	**Noisy dataset**
1	99.26%	89.29%	98.31%	86.52%
2	95.55%	87.41%	95.95%	87.23%
3	94.25%	85.11%	94.39%	87.11%
4	98.39%	87.75%	99.37%	89.53%

Table 7: Precision and recall for clean and noisy datasets per class

4.1.4 F1-measures

Room	Clean dataset	Noisy dataset
1	98.77%	87.78%
2	95.71%	87.19%
3	94.28%	85.98%
4	98.87%	88.43%

Table 8: F1-measures for clean and noisy datasets per class

4.2 Result analysis

The accuracy of the clean tree decreased by 0.48% after pruning. This is because an unpruned tree overfitted to the clean data is more representative of the system than a pruned tree (explained in Section 3.3). However, without knowing the data quality, one should prune to generalise the model. Pruning the noisy tree improves accuracy by 7.67% on average, due to pruning reducing overfitting. Rooms 2 and 3 have the most WiFi points around them, so are most affected by pruning reducing specificity; this can be seen in the clean dataset where their F1-measures drop.

4.3 Depth analysis

For the clean dataset, pruning reduces the depth of the tree by an average of 3.52 nodes, from 12.84 nodes to 9.32. For the noisy dataset, pruning has a larger impact, with an average reduction of 4.30 nodes, from 19.36 nodes to 15.06. The noisy trees are much deeper due to overfitting, with deep trees containing decisions more specific to the training set, but depth reduction generalises the trees. Figure 2 demonstrates how accuracy in the noisy tree improves with pruning. It is clear that the greater the maximal depth, the worse the accuracy.

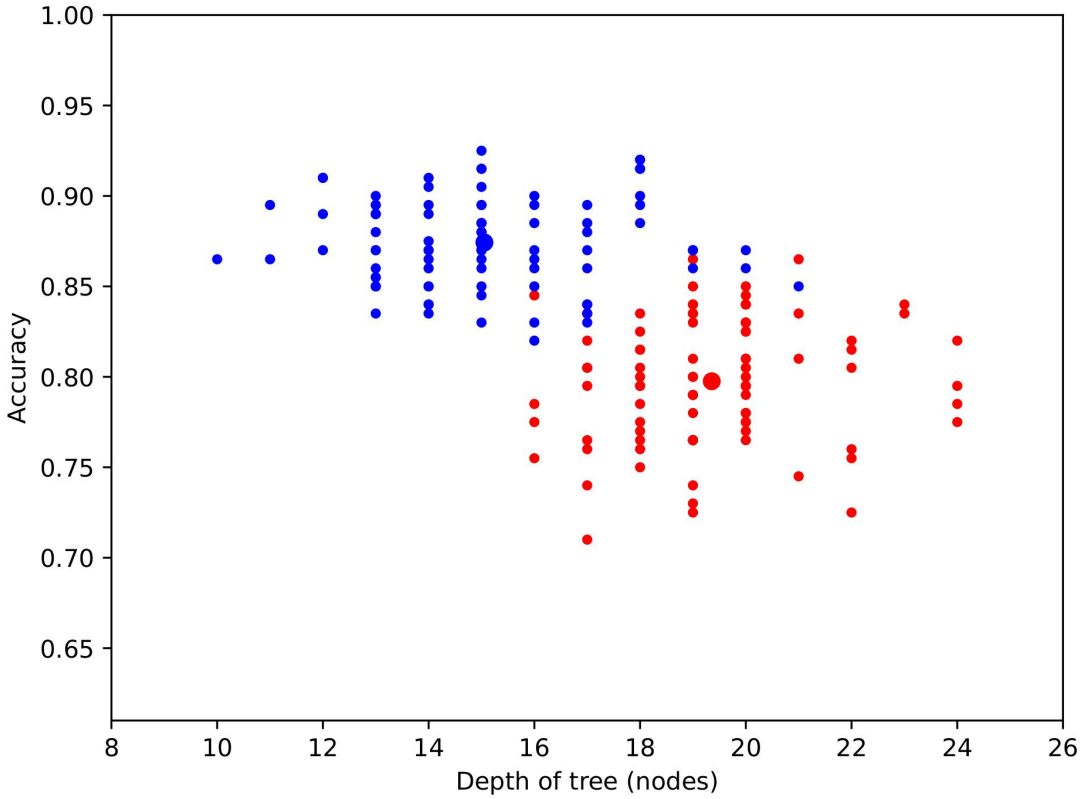

Figure 2: The relationship between maximum depth
and accuracy of a tree trained on the noisy data set
(red is unpruned, blue is pruned; the larger points are the averages per set).