Imperial College London

Report

Imperial College London

Department of Computing

Machine Learning Decision Trees Coursework

Authors:
Hyunjoon Jeon, Michael Kyriakou, Pranav Bansal, Raghav Viswakumar

Date: 5th November 2021

1 Tree Visualisation

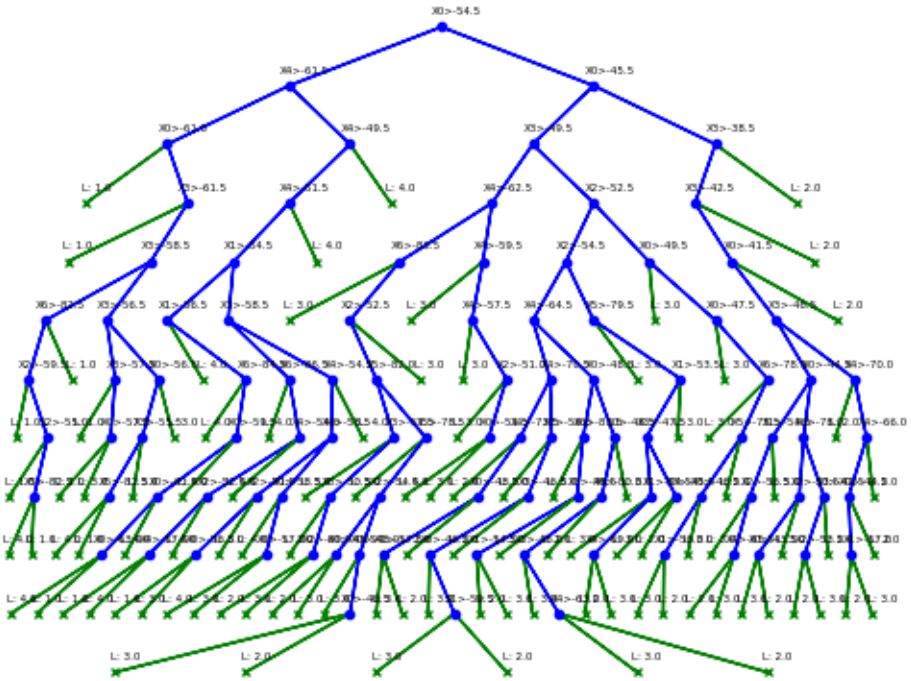

Figure 1: Clean Entire Trained Tree

Command to reproduce

```
> python3 main.py prune=true db=wifi_db/clean_dataset.txt seed=666
folds=10 visualise_cnt=1
```

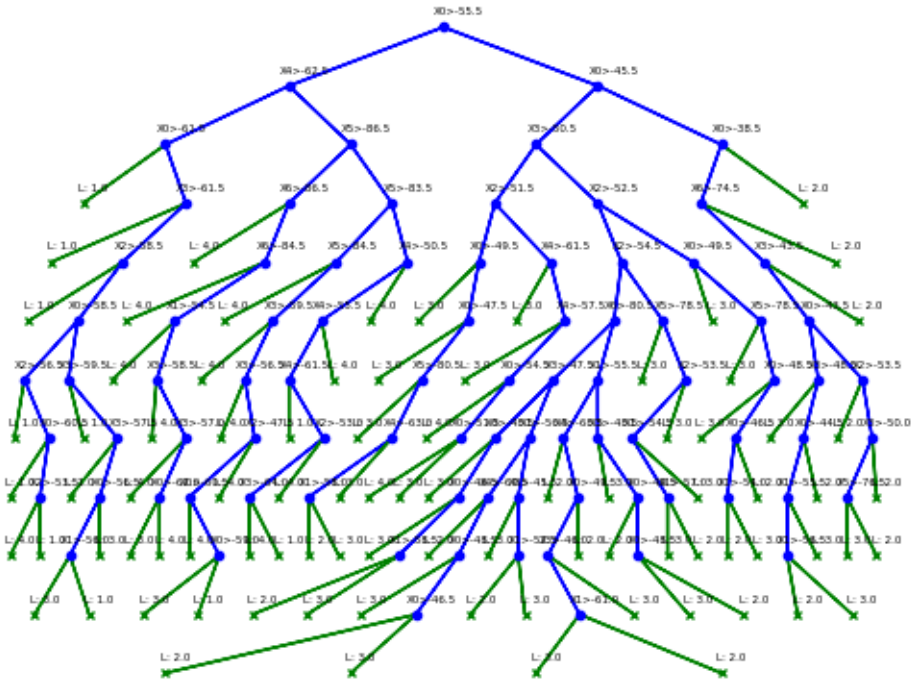

Figure 2: Clean Dataset Unpruned Tree

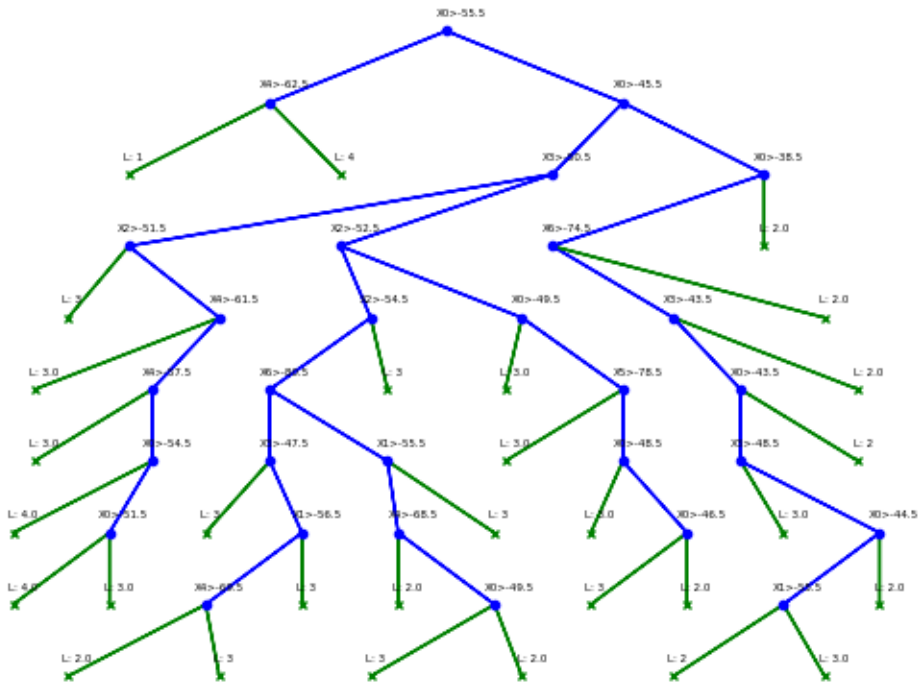

Figure 3: Clean Dataset Pruned Tree

2 Evaluation

2.1 Cross Validation Classification Metrics

```
Commands to reproduce
> python3 main.py db=wifi_db/clean_dataset.txt latex=true folds=10 seed=666
> python3 main.py db=wifi_db/noisy_dataset.txt latex=true folds=10 seed=666
```

2.1.1 Clean Dataset

$$\begin{bmatrix} 49.3 & 0. & 0.2 & 0.5 \\ 0. & 47.9 & 2.1 & 0. \\ 0.5 & 2.1 & 46.9 & 0.5 \\ 0.3 & 0. & 0.1 & 49.6 \end{bmatrix} \tag{1}$$

	Room 1	Room 2	Room 3	Room 4
Accuracy		0.9684999999999999		
Recall	0.986	0.958	0.938	0.992
Precision	0.98403194	0.958	0.95131846	0.98023715
F1 measure	0.98501499	0.958	0.94461229	0.9860835

2.1.2 Noisy Dataset

$$\begin{bmatrix} 40. & 2.9 & 2.4 & 3.7 \\ 2.1 & 39.7 & 5.2 & 2.7 \\ 3. & 3.8 & 40.7 & 4. \\ 3.5 & 1.9 & 4. & 40.4 \end{bmatrix} \tag{2}$$

	Room 1	Room 2	Room 3	Room 4
Accuracy		0.804		
Recall	0.81632653	0.79879276	0.79029126	0.81124498
Precision	0.82304527	0.82194617	0.77820268	0.79527559
F1 measure	0.81967213	0.81020408	0.78420039	0.80318091

2.2 Result Analysis

The diagonal entries of the confusion matrix represent the average number of true positive cases associated with each room. From the confusion matrix of the clean dataset, it can be observed that room 1 and 4 have relatively high accuracy and F1 measures while room 2 and 3 have relatively low accuracy and F1 measures. Another thing to note is that the average number of false positive cases is relatively high between room 2 and 3, showing that these two rooms are confused. From the confusion matrix of the noisy dataset, the error is more evenly distributed as the accuracy and F1 measures of all rooms are similar. In the noisy dataset, rooms 2 and 3 and additionally rooms 1 and 4, 3 and 4 have the highest confusion between one another.

2.3 Dataset Differences

Higher accuracy is obtained using the "clean" dataset, compared to the "noisy" dataset. This is due to noisy datapoints corrupting our decision tree due to the increase of outliers in our data, making it more difficult to accurately split our data. This results in a greater chance of incorrectly classifying test dataset values, thus lowering our accuracy and overall performance which is evidenced through the F1 and accuracy metrics.

3 Pruning

```
Commands to reproduce
> python3 main.py db=wifi_db/clean_dataset.txt latex=true folds=10 seed=666
prune=true
> python3 main.py db=wifi_db/noisy_dataset.txt latex=true folds=10 seed=666
prune=true
```

3.1 Cross Validation Classification Metrics after Pruning

3.1.1 Clean Dataset

$$\begin{bmatrix} 49.51111111 & 0. & 0.25555556 & 0.23333333 \\ 0. & 47.32222222 & 2.67777778 & 0. \\ 0.67777778 & 1.45555556 & 47.41111111 & 0.45555556 \\ 0.38888889 & 0. & 0.32222222 & 49.28888889 \end{bmatrix} \tag{3}$$

	Room 1	Room 2	Room 3	Room 4
Accuracy		0.9676666666666668		
Recall	0.99022222	0.94644444	0.94822222	0.98577778
Precision	0.97891037	0.97015945	0.93574561	0.9862161
F1 measure	0.9845338	0.95815523	0.9419426	0.98599689

3.1.2 Noisy Dataset

$$\begin{bmatrix} 44.08888889 & 1.22222222 & 1.37777778 & 2.31111111 \\ 1.94444444 & 43.43333333 & 3.15555556 & 1.16666667 \\ 2.53333333 & 3.07777778 & 44.07777778 & 1.81111111 \\ 2.32222222 & 1.41111111 & 1.87777778 & 44.18888889 \end{bmatrix} \tag{4}$$

	Room 1	Room 2	Room 3	Room 4
Accuracy		0.8789444444444444		
Recall	0.89977324	0.87391013	0.85587918	0.88732709
Precision	0.86637555	0.88378928	0.87301937	0.89310577
F1 measure	0.88275862	0.87882194	0.86436431	0.89020705

3.2 Result Analysis after Pruning

Varying amount of performance increases are noticed, based on seed value which determines the values in each fold. With the clean dataset, there are only minor improvements or declines in accuracy due to the fact that the dataset is generalised well by the initial decision tree before pruning. With the noisy dataset, we see a much greater improvement in accuracy and F1 values due to pruning the data which has been over fitted on the training set. By reducing the nodes/decisions in our decision tree the model becomes more generalised, thus showing a higher accuracy on the test dataset. This increase in accuracy is more pronounced on the noisy dataset due to the initial tree being overfitted more with the noisy dataset compared to the clean dataset.

3.3 Depth Analysis

Clean

```
With pruning:  9.811111111111112
Without pruning:  11.133333333333333
```

Noisy

```
With pruning:  11.033333333333333
Without pruning:  12.755555555555556
```

Average maximum depth of trees decrease as a result of pruning by definition. Large depths are an indication of overfitting. When the data is overfitted or has a higher depth, we can see that a decrease in maximum depth yields to a larger improvement in accuracy. Although the decrease in depth is comparable with the clean and noisy dataset, it is more effective on the noisy dataset due to the higher average depth both before and after pruning.