Imperial College London

Department Of Computing
Introduction To Machine Learning

Decision Tree Coursework

TEAM MEMBERS:

SALIM AL-WAHAIBI

ALICIA LAW

MARCOS-ANTONIOS CHARALAMBOUS
WEI JIE CHUA

Contents

1 Introduction

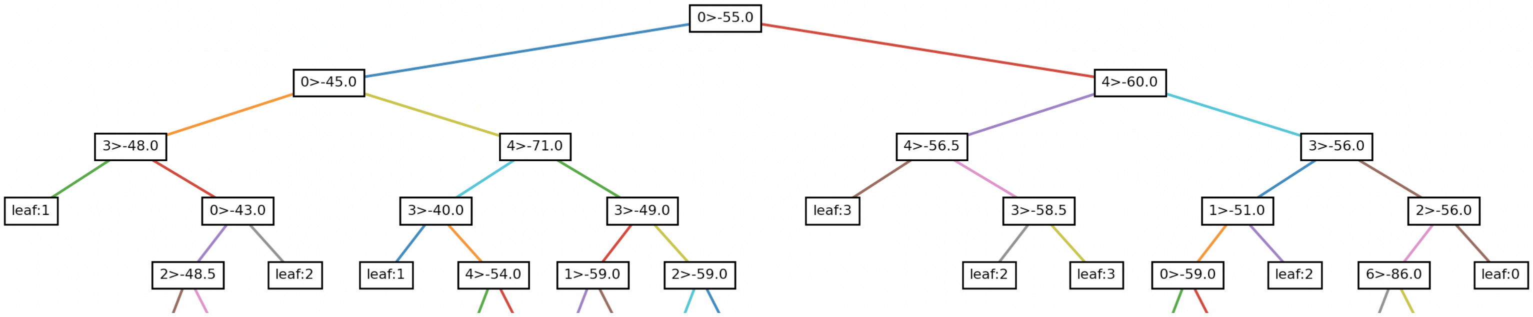
In this project, we have implemented a decision tree algorithm to identify an individual's indoor room location based on WIFI strength signals collected from their mobile phone relative to different routers. We have split the report into multiple sections.

The exact specification of the implementation can be found on GitHub and the README.md file.

2 Step 2 - Plotting of tree

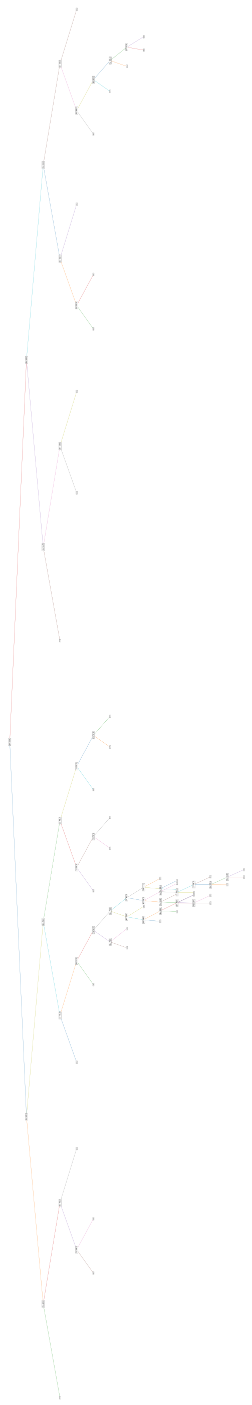
2.1 Up close view of tree

Tree trained on entire clean data is visualized.



2.2 Complete tree

The full tree is printed on the next page.
To zoom in on the tree, the full tree can be found in project folder as dt_bonus.png.



3 Step 3 - Evaluation (Before pruning)

3.1 Clean dataset

3.1.1 Confusion Matrix

We have highlighted the average true positive rate for each class in green.

	Prediction			
	Class 1	Class 2	Class 3	Class 4
Class 1	49.5	0.0	0.2	0.3
Class 2	0.0	48.1	1.9	0.0
Class 3	0.2	2.4	47.1	0.3
Class 4	0.4	0.0	0.2	49.4

Figure 1: Confusion matrix for clean dataset

3.1.2 Accuracy, Recall and Precision rate

Accuracy: 97.0%

	Class 1	Class 2	Class 3	Class 4
Recall	99.0%	96.2%	94.2%	98.8%
Precision	98.8%	95.2%	95.3%	98.8%
F1-score	98.9%	95.7%	94.8%	98.8%

Figure 2: Accuracy, recall, and precision rate for each class

Macro recall: 97.1%
Macro precision: 97.0%
Macro F1-score: 97.0%
Average tree depth: 12.2

3.2 Noisy dataset

3.2.1 Confusion Matrix

We have highlighted the average true positive rate for each class in green.

	Prediction			
	Class 1	Class 2	Class 3	Class 4
Class 1	38.5	3.3	3.3	3.9
Class 2	2.9	40.1	4.1	2.6
Class 3	2.8	3.5	41.8	3.4
Class 4	3.9	2.9	2.7	40.3

Figure 3: Confusion matrix for noisy dataset

3.2.2 Accuracy, Recall and Precision rate

Accuracy: 80.4%

	Class 1	Class 2	Class 3	Class 4
Recall	78.6%	80.7%	81.2%	80.9%
Precision	80.0%	80.5%	80.5%	80.3%
F1-score	79.3%	80.6%	80.9%	80.6%

Figure 4: Accuracy, recall, and precision rate for each class

Macro recall: 80.3%
Macro precision: 80.3%
Macro F1-score: 80.3%
Average tree depth: 19.1

3.3 Result Analysis

In clean data, class 1 and class 4 are recognized with higher accuracy. Both recall and precision for class 1 and class 4 are highest, with an approximate 3% - 4% difference to class 2 and class 3. However in the noisy data, the accuracy difference between the 4 classes are not obvious. There will be higher misclassification between class 1 and class 4, as shown in the confusion matrix.

3.4 Dataset differences

	Clean data	Noisy data
Accuracy	97.0	80.4
Macro recall	97.0	80.3
Macro precision	97.0	80.3
Macro F1-score	97.0	80.3

Figure 5: Noisy data macro evaluation metrics comparison

The tree's accuracy on clean data is 97.0%, whereas the accuracy on noisy data is 80.4%, resulting in a 16.6% difference. This difference in performance is due to the nature of the decision tree, where it tries to maximize information gain and reduce entropy in order to obtain a homogeneous leaf. This leads to overfitting of the tree on the noise of the training data, greatly reducing its ability to generalize and hence, reducing its performance and accuracy on unseen test data.

4 Step 4 - Evaluation (After pruning)

4.1 Clean dataset

4.1.1 Confusion Matrix

We have highlighted the average true positive rate for each class in green.

	Prediction			
	Class 1	Class 2	Class 3	Class 4
Class 1	49.8	0.0	0.1	0.1
Class 2	0.0	47.5	2.5	0.0
Class 3	0.5	2.1	47.1	0.3
Class 4	0.5	0.0	0.3	49.2

Figure 6: Confusion matrix for clean dataset

4.1.2 Accuracy, Recall and Precision rate

Accuracy: 96.8%

	Class 1	Class 2	Class 3	Class 4
Recall	99.6%	95.0%	94.2%	98.4%
Precision	98.0%	95.8%	94.2%	99.2%
F1-score	98.8%	95.4%	94.2%	98.8%

Figure 7: Accuracy, recall, and precision rate for each class

Macro recall: 96.8%
Macro precision: 96.8%
Macro F1-score: 96.8%
Average tree depth: 7.3

4.2 Noisy dataset

4.2.1 Confusion Matrix

We have highlighted the average true positive rate for each class in green.

	Prediction			
	Class 1	Class 2	Class 3	Class 4
Class 1	44.1	1.2	1.4	2.3
Class 2	1.8	44.0	2.7	1.2
Class 3	2.1	3.2	44.3	1.9
Class 4	2.2	1.5	1.7	44.4

Figure 8: Confusion matrix for noisy dataset

4.2.2 Accuracy, Recall and Precision rate

Accuracy: 88.4%

	Class 1	Class 2	Class 3	Class 4
Recall	90.0%	88.5%	86.0%	89.2%
Precision	87.8%	88.2%	88.4%	89.2%
F1-score	88.9%	88.4%	87.2%	89.2%

Figure 9: Accuracy, recall, and precision rate for each class

Macro recall: 88.4%
Macro precision: 88.4%
Macro F1-score: 88.4%
Average tree depth: 13.3

4.3 Result Analysis After Pruning

4.3.1 Macro metrics comparison

	Before pruning	After pruning
Accuracy	97.0	96.8
Macro recall	97.1	96.8
Macro precision	97.0	96.8
Macro F1-score	97.0	96.8

Figure 10: Clean data macro metrics comparison

	Before pruning	After pruning
Accuracy	80.4	88.4
Macro recall	80.3	88.4
Macro precision	80.3	88.4
Macro F1-score	80.3	88.4

Figure 11: Noisy data macro metrics comparison

4.3.2 Comments

For clean data, the accuracy after pruning reduced slightly by 0.2% due to information loss in pruning. Clean data is not vulnerable to noise, and hence the training data built a low bias and variance tree. Cross validation in step 4 also reduced the available training data by 9% (90% x 10% for validation), potentially lowering performance. For noisy data, the accuracy after pruning improved by 8% as it removed the noise in the dataset, hence reducing overfitting of the tree.

4.4 Depth Analysis

4.4.1 Depth comparison

	Before pruning	After pruning
Clean data	12.2	7.3
Noisy data	19.1	13.3

Figure 12: Depth comparison

4.4.2 Comments

Before pruning, the average depth for noisy data is 19.1 compared to 12.2 for clean data. This 6.9 mean difference in depth is due to the effect of noise, which caused overfitting of the decision tree and its reduced generalisation. After pruning, the average depth reduced for both noisy and clean data, now 13.3 and 7.3 respectively. For clean data, this does not impact the model significantly. However for noisy data, the reduced depth greatly increases its accuracy.