Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Decision Trees

Authors:
Alina Zhu (az421)
Arnav Kohli (sk1421)
Bhavya Sharma (bs921)
Johan Jino (jj21)

Supervisor:
Dr. Antoine Cully

Submitted in partial fulfillment of the requirements for the Introduction to Machine Learning (COMP60012) coursework of Imperial College London.

October 2023

Contents

Chapter 1

Tree visualization function

Visualization of the decision tree was facilitated using the *matplotlib* library. A specific function, `plot_tree` took as argument the decision tree generated and plotted it on a graph using *border boxes* and *dynamic spacing*.

1.1 Clean Data

While using the entire clean data-set as input, a decision tree of depth 14 was produced. The results of this can be seen below:

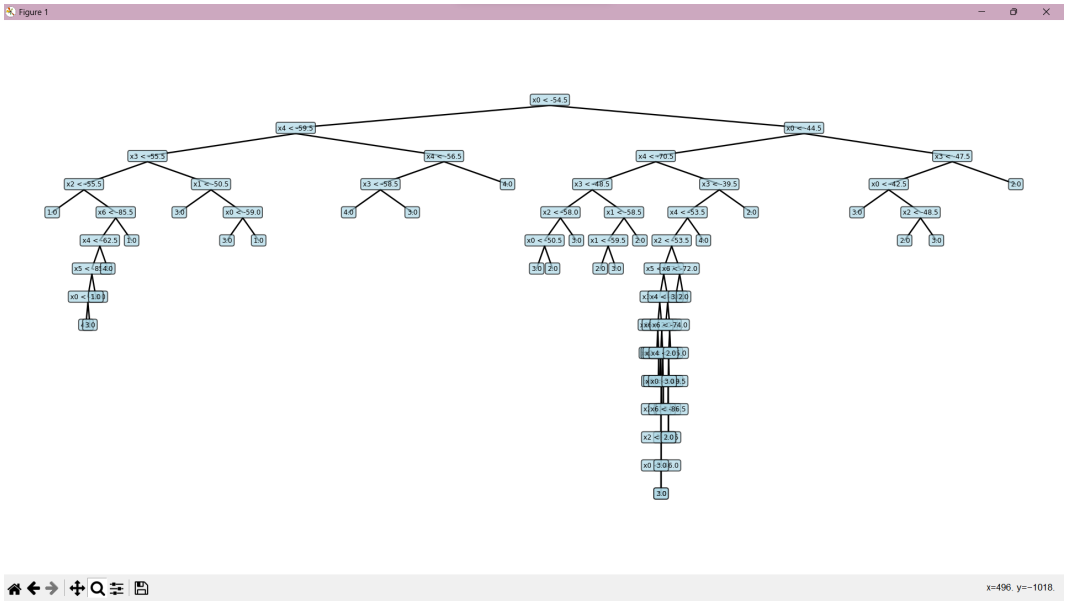


Figure 1.1: Clean data-set tree

The entire decision tree could not be rendered on this image, and a screen recording of the best representation of the entire decision tree has been saved with the source code.

1.2 Noisy Data

Similar to the clean data-set, the noisy data-set was also used as an input. The depth of this tree was 18 (higher than clean data-set), the reason for which will be explained in the result analyses. The tree generated was:

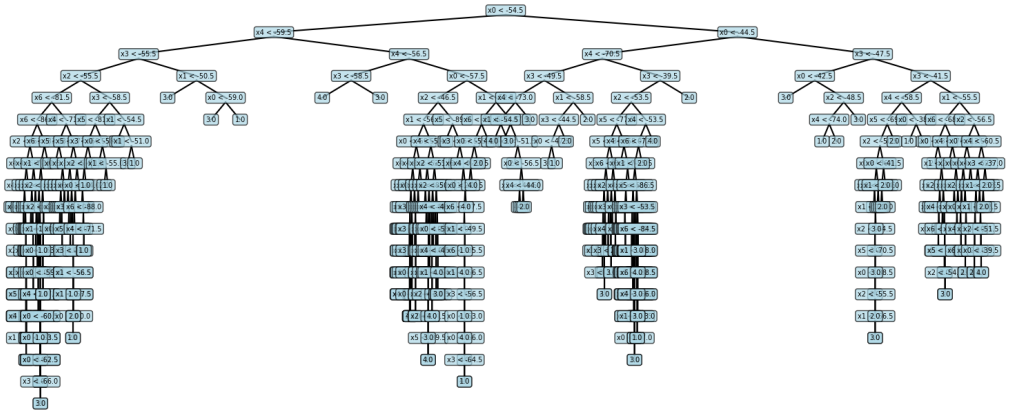


Figure 1.2: Noisy data-set tree

Chapter 2

Evaluation

k-*fold cross validation* was used as the evaluation metric. The data-set was split into k parts, $k-1$ of which were used as training sets and 1 of these was used as the test set in an iterative manner. This was repeated k times and the average across all iterations was computed. Throughout the evaluation process, the value of k was set to 10. To achieve reproducible (and thus, verifiable) results, a seed of 63000 was arbitrarily provided to the *numpy* library's `default_rng()` function. Evaluation was performed separately on the clean and noisy data-sets.

2.1 Clean Data

2.1.1 Cross validation classification metrics

Confusion Matrix

The following figure is a heatmap that visualizes the confusion matrix (4x4, due to their being 4 possible rooms - 4 possible choices for the prediction and the actual label) generated from a clean data-set. It provides insights into the model's accuracy by comparing predicted and actual class labels. The matrix represents the average of the results obtained over the k folds. Along the diagonal are the correct predictions (out of 50).

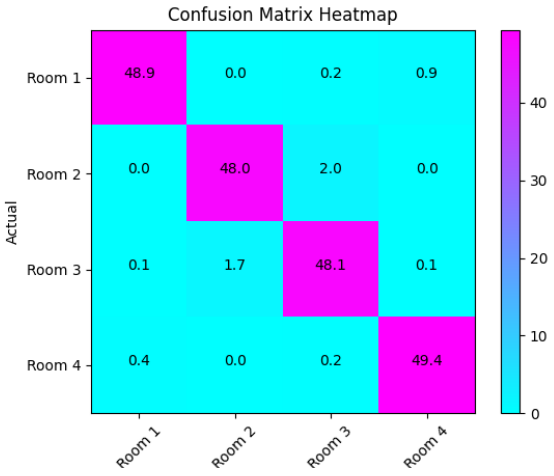


Figure 2.1: Clean data-set confusion matrix heat map

Color intensity within the cells indicates the number of instances, with darker colors representing higher numbers.

Metrics

The other metrics calculated for the clean data-set can be seen below:

```
Obtained Values:

Precisions [0.99 0.97 0.95 0.98]
Recalls [0.98 0.96 0.96 0.99]
Accuracy 0.97
F1 [0.98 0.96 0.96 0.98]
```

Figure 2.2: Clean data-set metrics

2.1.2 Result analysis

From the confusion matrix heatmap, we can see the tree can mostly correctly predict the rooms, especially for room 1 and 4. Room 2 and 3 are confused with each other on rare occasions, indicated by 1.7 and 2.0 instances out of 50 predicted incorrectly. An accuracy of 0.97 shows that the error rate was relatively low.

2.2 Noisy Data

2.2.1 Cross validation classification metrics

Confusion Matrix

Similar to the clean data-set, a confusion matrix heatmap was produced for the noisy data-set which can be seen in figure 2.3.

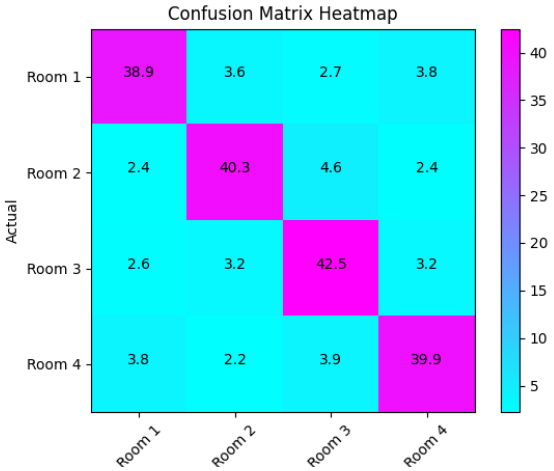


Figure 2.3: Noisy data-set confusion matrix heat map

Metrics

The other metrics calculated for the noisy data-set can be seen in figure 2.4

```
Obtained Values:

Precisions [0.82 0.82 0.79 0.81]
Recalls [0.79 0.81 0.83 0.8 ]
Accuracy 0.81
F1 [0.8  0.82 0.81 0.81]
```

Figure 2.4: Noisy data-set metrics

2.2.2 Result analysis

From the heatmap we can conclude that room 3 has the highest accuracy in correct prediction, however room 2 and 3 are also confused with each other the most, with an average of 4.6 out of 50 instances of wrong classification. An accuracy of 0.81 shows that the error rate was relatively low.

2.3 Data-set differences

The average accuracy of the noisy data-set across folds was lower than the clean data-set when tested using cross validation. Noisy data has a higher number of outliers causing the tree to be over-fitted according to anomalous data points. Clean data displayed a lower degree of over-fitting due to higher consistency across data points. The effect of noise can be seen by the depths of the trees as well, deeper trees tend to be precursors to noisier data-sets.

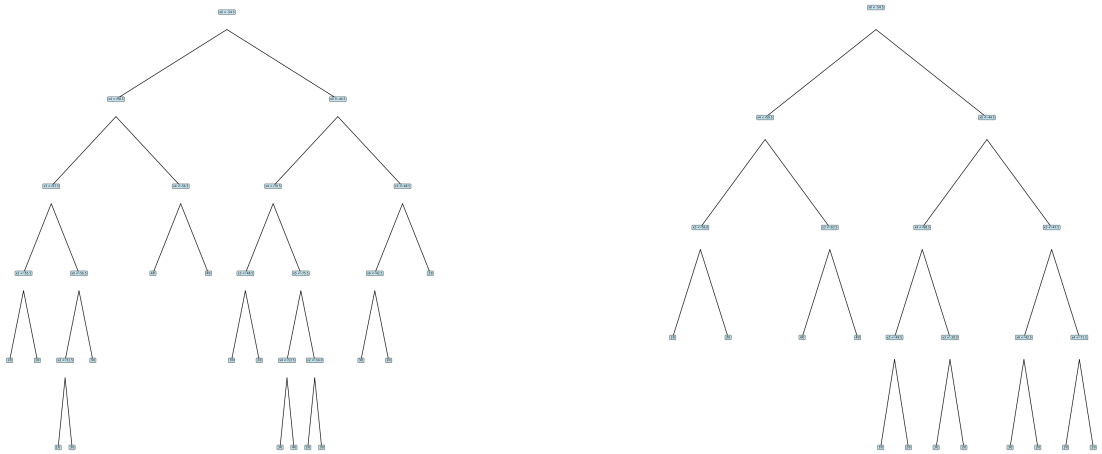
Chapter 3

Pruning

The idea of pruning is to essentially remove nodes that would potentially improve the overall accuracy on the validation set. Although splitting the data to train, validation and split reduces training data, it improves generalisation of the decision tree.

3.1 Tree Visualization

The tree generated after pruning can be seen below:

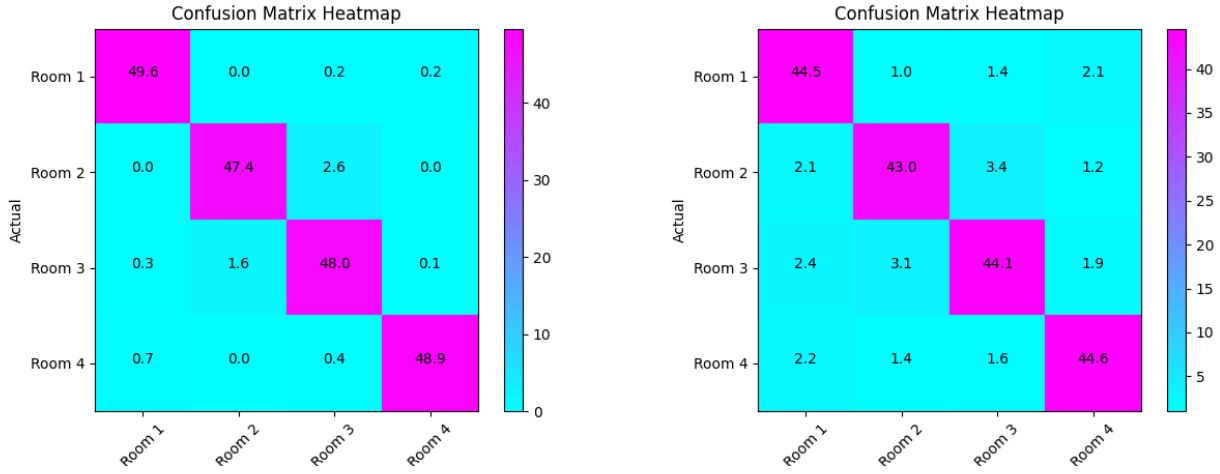


((a)) Clean data-set **((b))** Noisy data-set

Figure 3.1: Tree after pruning

3.2 Evaluation

The confusion matrix for the same as been given below:



((a)) Clean data-set **((b))** Noisy data-set

Figure 3.2: Tree after pruning

The evaluation metrics for the same as been given below:

```
Obtained Values:

Precisions [0.98 0.97 0.94 0.99]
Recalls [0.99 0.95 0.96 0.98]
Accuracy 0.97
F1 [0.99 0.96 0.95 0.99]
```

((a)) Clean data-set

```
Obtained Values:

Precisions [0.87 0.89 0.87 0.9 ]
Recalls [0.91 0.87 0.86 0.9 ]
Accuracy 0.88
F1 [0.89 0.88 0.86 0.9 ]
```

((b)) Noisy data-set

Figure 3.3: Tree after pruning

3.2.1 Observation

We can observe that in case of the clean data-set the accuracy slightly drops, while in the case of the noisy data-set the accuracy has improved significantly. This is to be expected since in the case of the noisy data-set the pre-pruned tree overfits and learns the errors in the data-set, which is removed during pruning. While in the case of clean, the accuracy drops slightly due to less training data to train on since we need a validation set as well.
We also see that the rooms that are confused with each other also have reduced errors, meaning the mis-prediction is reduced.