Imperial College London

Department of Computing

IMPERIAL

Coursework 1: Decision Trees

Introduction to Machine Learning ELEC60012

Authors:

Gurjan Singh Samra	02288570
Arav Swati Abhay	02242706
Vishesh Mongia	02288840
Alex Seferidis	02269571

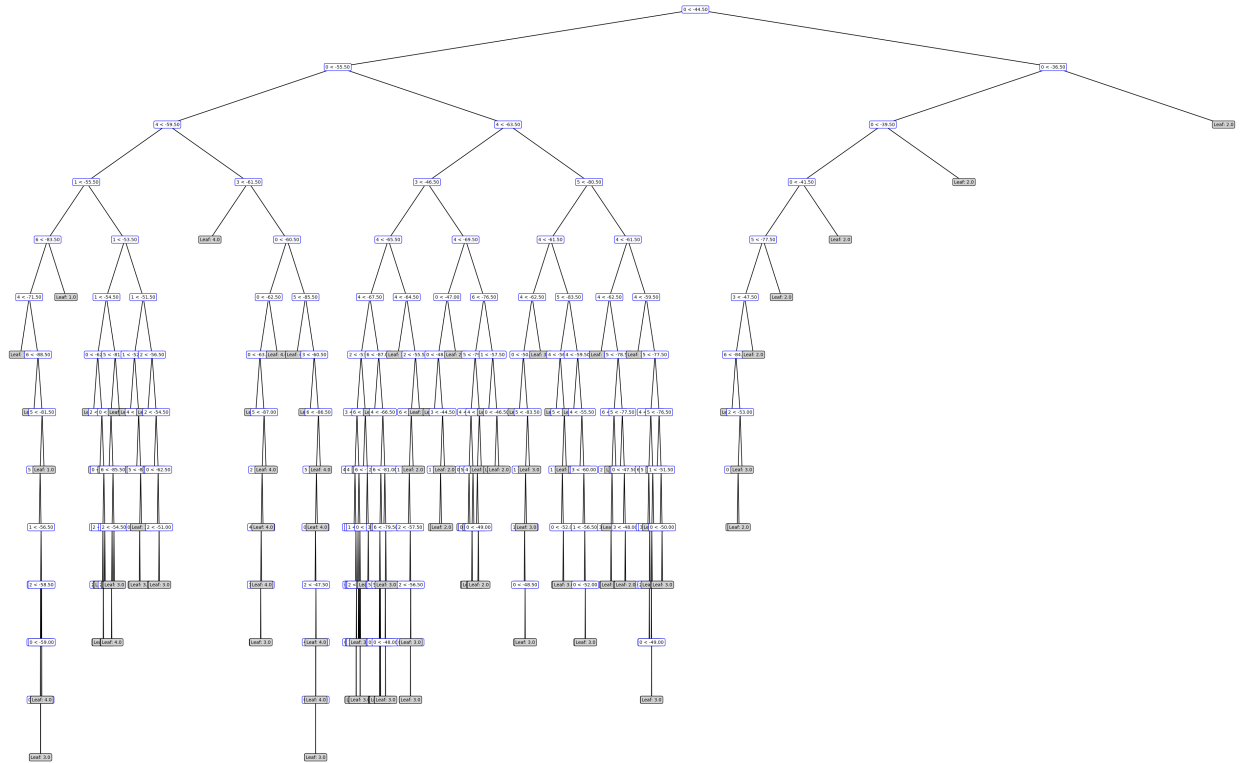Autumn 2024

Visualisation

Figure 1.1: Tree Visualisation Using Clean Dataset

Evaluation

2.1 Crossvalidation Classification Metrics

2.1.1 Confusion Matrices

Below in Tab. 2.1 and Tab. 2.2 are the confusion matrices for the clean and noisy datasets respectively. These matrices are produced as a **total** of all K folds, rather than an average of all folds. The trace of the clean dataset's confusion matrix is 1921, in contrast to 1629 for the noisy dataset's confusion matrix.

Table 2.1: Confusion Matrix - Clean Dataset

Actual / Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	492	0	3	5
Room 2	0	473	27	0
Room 3	4	23	467	6
Room 4	3	0	8	489

Table 2.2: Confusion Matrix - Noisy Dataset

Actual / Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	398	24	30	38
Room 2	26	408	38	25
Room 3	28	31	426	30
Room 4	42	18	41	397

2.1.2 Metrics

The macro-averaged accuracy for both clean and noisy datasets can be seen in Tab. 2.3. The recall, precision and F1 per class can also be seen in Tab. 2.4 and Tab. 2.5 for the clean and noisy datasets respectively.

Table 2.3: Macro-Averaged Accuracy for Clean and Noisy Dataset

	Clean Dataset	Noisy Dataset
Accuracy	0.9605	0.8145

Table 2.4: Recall, Precision, and F1 Score per Class - Clean Dataset

Class	Recall	Precision	F1 Score
Room 1	0.984	0.986	0.985
Room 2	0.946	0.954	0.950
Room 3	0.934	0.925	0.930
Room 4	0.978	0.978	0.978

Table 2.5: Recall, Precision, and F1 Score per Class - Noisy Dataset

Class	Recall	Precision	F1 Score
Room 1	0.812	0.806	0.809
Room 2	0.821	0.848	0.834
Room 3	0.827	0.796	0.811
Room 4	0.797	0.810	0.804

2.2   Result Analysis

Comment for both datasets on which rooms are correctly recognized, and which rooms are confused with others

The clean dataset's rooms are all well recognised, evident by the large values on the diagonal, Room 1 is the most correctly recognised, followed by Room 4. Rooms 2 and 3 are confused the most with each other. The noisy dataset's rooms are less well recognised, with the best being Room 3. Rooms 1 and 4 appear to be confused often with each other, as well as Room 4 being predicted as Room 3. Overall there is a wide and consistent spread of confusion with the noisy dataset.

2.3   Dataset Differences

Is there any difference in the performance when using the clean and noisy datasets? If yes/no explain why

Overall, the noisy dataset performs **worse** than the clean dataset. This is evidenced by the noticeable drop with the noisy dataset in macro-averaged accuracy, the confusion matrix's trace and per class metrics (recall, precision and F1). This is due to attempts to fit around the noise rather than exclude it, producing inaccurate / misleading patterns in the model.