Intro. to ML Coursework 1: Decision Tree Report

November 5, 2021

Members

- **Kaiyan Fan**
  kf619@ic.ac.uk
- **Xuan Zhao**
  xz1919@ic.ac.uk
- **Zelin Deng**
  zd419@ic.ac.uk
- **Sudarshan Sreeram**
  ss8119@ic.ac.uk

Before Pruning

Cross Validation Classification Metrics On Clean Data

Confusion Matrix

Actual \ Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	495	0	1	4
Room 2	0	478	22	0
Room 3	0	21	477	2
Room 4	5	0	4	491

Recall and Precision of Each Class

	Room 1	Room 2	Room 3	Room 4
Recall	0.99	0.956	0.954	0.982
Precision	0.99	0.946	0.95	0.984

Accuracy: 0.97
Macro-averaged F-1 Score: 0.971
Average Tree Depth: 6.686

Cross Validation Classification Metrics On Noisy Data

Confusion Matrix

Actual \ Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	400	32	26	32
Room 2	33	402	40	22
Room 3	30	34	414	37
Room 4	36	27	33	402

Recall and Precision of Each Class

	Room 1	Room 2	Room 3	Room 4
Recall	0.816	0.809	0.804	0.807
Precision	0.802	0.812	0.807	0.815

Accuracy: 0.809
Macro-averaged F-1 Score: 0.809
Average Tree Depth: 10.637

Result Analysis (Pre-Pruning)

On the clean data set, Room 1 is recognised with both the recall and precision being the highest, followed by Room 4 which has the second highest recall and precision. This means that 99% of the Room 1 data is being correctly recognised, and 99% of the predicted Room 1 data are actually Room 1. The decision is the most confused at Room 2, whose precision is the lowest, and Room 3, whose recall is the lowest. On the dirty data set, Room 1 has the highest recall and Room 4 has the highest precision.

Dataset Difference

The performance of decision tree in every metric is higher on the clean dataset than on the dirty dataset. The reason is that the noisy dataset has a less uniform distribution than the clean dataset: if a set of data points with similar attribute values in the clean dataset all correspond to a single room, those attribute values might actually correspond to other rooms as well in the noisy dataset. In terms of the shape of the decision tree, noisy dataset produces a tree with greater depth and more branches than the tree produced by clean dataset. This indicates that the decision tree algorithm needs to generate more complex rules in order to comprehensively categorise/fit each data points.

After Pruning

Cross Validation Classification Metrics On Clean Data

Confusion Matrix

Actual \ Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	499	0	1	0
Room 2	0	473	27	0
Room 3	3	18	475	4
Room 4	5	0	3	492

Recall and Precision of Each Class

	Room 1	Room 2	Room 3	Room 4
Recall	0.998	0.946	0.95	0.982
Precision	0.984	0.963	0.939	0.992

Accuracy: 0.97
Macro-averaged F-1 Score: 0.97
Average Tree Depth: 5.375

Cross Validation Classification Metrics On Noisy Data

Confusion Matrix

Actual \ Predicted	Room 1	Room 2	Room 3	Room 4
Room 1	444	11	13	22
Room 2	18	441	27	11
Room 3	21	35	439	20
Room 4	24	13	15	446

Recall and Precision of Each Class

	Room 1	Room 2	Room 3	Room 4
Recall	0.906	0.887	0.852	0.896
Precision	0.876	0.882	0.889	0.894

Accuracy: 0.885
Macro-averaged F-1 Score: 0.885
Average Tree Depth: 7.375

Result Analysis (Post-Pruning)

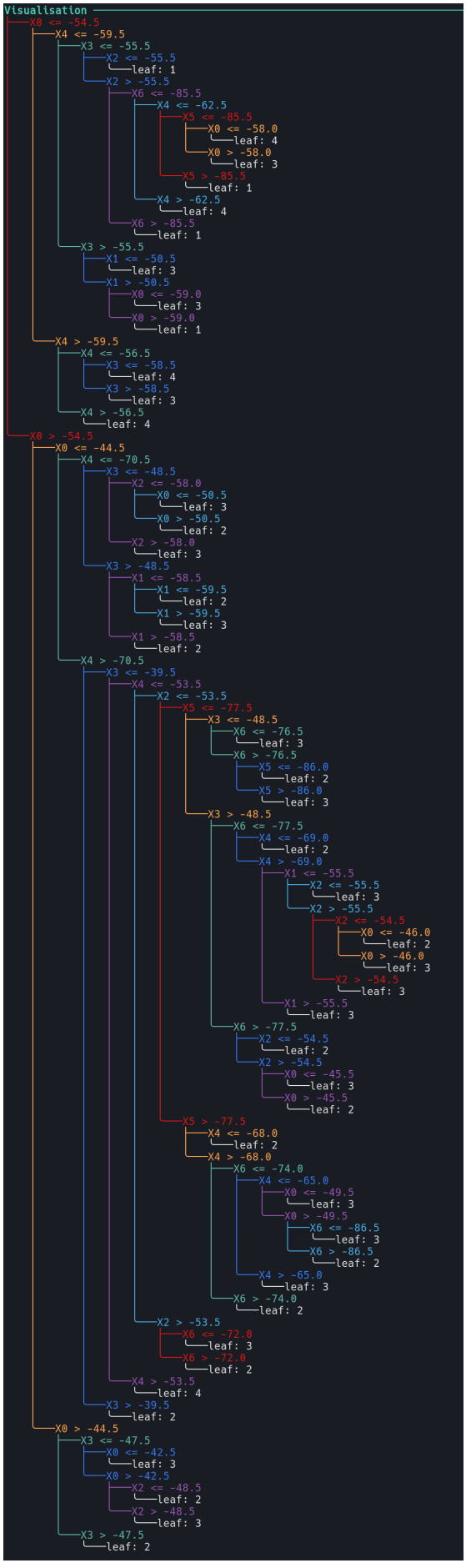
Pruning doesn't affect the clean dataset significantly as the accuracy and F-1 score remain almost identical, but the accuracy on the noisy dataset improves significantly by 8%. This is because for the clean dataset, the accuracy and F-1 score is high enough such that there might not exist a branch/twig on which pruning will produce better result. For noisy dataset, pruning actually works better because there are branches or twigs that appear to be the result of over-fitting the training dataset, thus there is opportunity to improve accuracy performance by pruning.

As a comparison, a tree trained on the full set of clean data achieved an accuracy of 91.8% on noisy data, our resultant accuracy of 88.5% comes reasonably close to this idea situation.

Depth Analysis

The average depth (only counting leaf depth) of the decision tree on clean and noisy dataset are 6.686 and 10.637 before pruning, and 5.375 and 7.375 after pruning. The maximum depth of unpruned tree on clean and noisy dataset are 12 and 19, respectively. After pruning, the maximum depth of the decision tree for clean and noisy dataset becomes 10 and 12, respectively. With pruning, the prediction accuracy will improve with lower maximal depth. This is because some of the long branches are actually the result of over-fitting the training set, and a good decision strategy should not involve too many complex rules to categorise data points.

Decision Tree Visualization



This is the visualization of decision tree trained on the entire clean dataset before pruning