Lab 1: Decision Trees

DD2421 Machine Learning

André Silva, Jinisha Bhanushali

September 16, 2020

Assignment 0

Based on the concepts of True, MONK-2 should be the most difficult dataset to learn for a decision tree algorithm as it depends on all six variables, while MONK-1 and MONK-3 only depend on three variables.

- `MONK 1` - Should require a depth of 3, as it only depends on 3 variables.
- `MONK 2` - Should require a depth of 6, as it depends on all variables and they are independent.
- `MONK 3` - Should require a depth of 3.

Assignment 1 - Entropy of the training datasets

Dataset	Entropy
MONK-1	1.0
MONK-2	0.957117428264771
MONK-3	0.999806132804711

Assignment 2

- Uniform distribution: every outcome has the same probability. Maximizes the entropy function.
- Non-uniform distribution: at least two outcomes have different probabilities. Entropy function is not maximized.

Assignment 2 - Example

Given a basket with 100 balls, either coloured black or white, the entropy of drawing a ball depends on the number of balls of each colour. If the basket contains more black than white, say 80 black and 20 white, the entropy would be lower considering the reduced randomness. On the other hand, if the basket has 50 black and 50 white balls, then the entropy is high as the predictability reduces.

Assignment 3 - Information gain

The attribute that should be used for splitting the examples at the root node is the one which maximizes information gain. So for MONK-1 and MONK-2 it would be a_5 while for MONK-3 a_2.

Dataset	a_1	a_2	a_3	a_4	a_5	a_6
MONK-1	0.07527	0.00584	0.00471	0.02631	0.28703	0.00076
MONK-2	0.00376	0.00246	0.00106	0.01566	0.01728	0.00625
MONK-3	0.00712	0.29374	0.00083	0.00289	0.25591	0.00708

Assignment 4

The entropy of each subset can be higher or lower than the entropy of the original set, but the weighted average of the subset entropies is lower than the entropy of the original set.
By using information gain as a heuristic we want to maximize, we are, step by step, creating a decision tree which can more accurately split and classify samples.

Assignment 5 - Train and test set errors

- Since we built our trees with no practical depth limit the trees were overfit to the training subsets, which explains why the error 0.0 when checking these subsets. When checking the test subsets we see that the error values are significant.
- MONK-2 was, as we expected, the hardest dataset to learn.
- The depths of the trees did not correspond to the expectations, as they weren't perfect trees.

	E_{train}	E_{test}
MONK-1	0.0	0.171
MONK-2	0.0	0.308
MONK-3	0.0	0.056

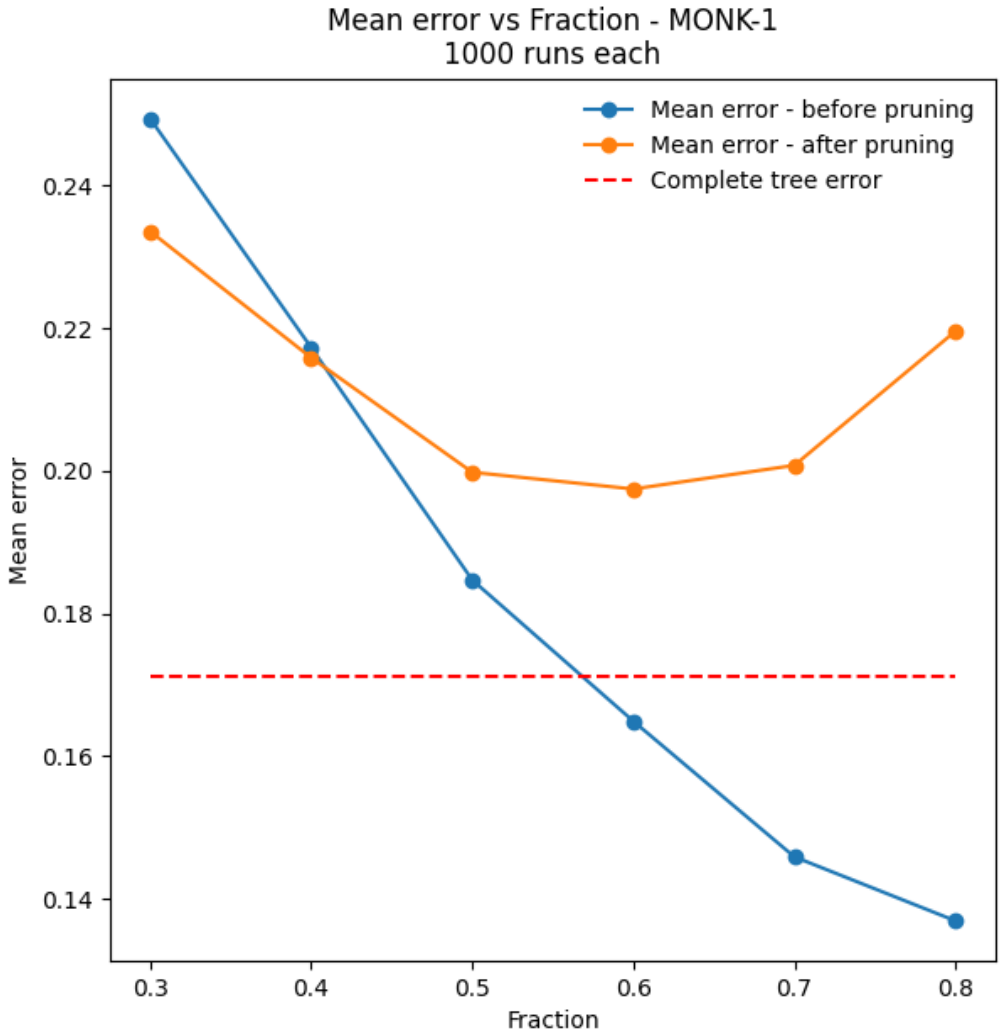
Assignment 6

- Decision trees are, by nature, a low bias and high variance technique. This means that, when building them, they tend to be overfitted to the training data, or in other words, generalize poorly to real data. Small changes in the testing data will produce big changes in the outcome.
- Pruning is used to reduce the complexity of a decision tree, i.e. reduce the variance or prevent overfitting. By doing so, we are also increasing bias, making the algorithm lose prediction accuracy. This is why we need to find a good balance between bias and variance while performing pruning.

Assignment 7 - Table

Fraction	0.3	0.4	0.5	0.6	0.7	0.8
MONK-1 - Mean error - before pruning	0.24937	0.21728	0.1847	0.16491	0.14589	0.13692
MONK-1 - Standard deviation - before pruning	0.04969	0.04972	0.04985	0.0502	0.04792	0.04988
MONK-1 - Mean error - after pruning	0.23357	0.21585	0.1998	0.19745	0.20077	0.21957
MONK-1 - Standard deviation - after pruning	0.03795	0.03876	0.0427	0.04231	0.04147	0.03908
MONK-3 - Mean error - before pruning	0.11249	0.09403	0.08317	0.07665	0.07109	0.06521
MONK-3 - Standard deviation - before pruning	0.06188	0.04717	0.03627	0.03102	0.02622	0.02168
MONK-3 - Mean error - after pruning	0.08054	0.05801	0.04242	0.034	0.03244	0.03882
MONK-3 - Standard deviation - after pruning	0.05592	0.04436	0.03496	0.03009	0.02651	0.03206

Assignment 7 - MONK-1 plot



Assignment 7 - MONK-3 plot