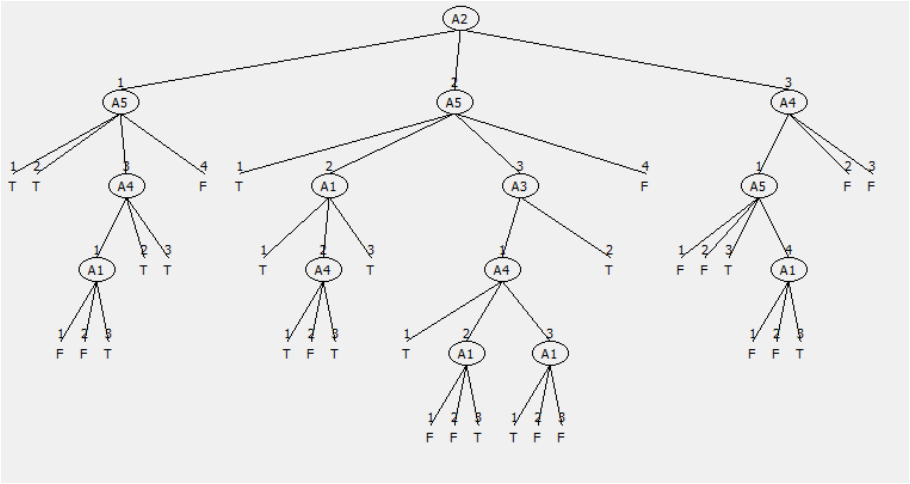
Machine Learning - Lab 1 (Decision Trees)

hfag@kth.se

November 8, 2022



Assignment 0

Each one of the datasets has properties which makes them hard to learn. Motivate which of the three problems is most difficult for a decision tree algorithm to learn.

MONK-1: $(a_1 = a_2) \vee (a_5 = 1)$

To verify that $a_1 = a_2$ it is hard to split the tree. For checking $a_5 = 1$ in case $a_1 \neq a_2$ it require a tree with a depth of three.

MONK-2: $a_i = 1$ *for exactly two* $i \in \{1, 2, ..., 6\}$

Here we need to check all the possible numbers that the attributes can assume and we therefore need a tree with a depth of six to, since all points needs to be checked. It is therefore the most difficult dataset to examine.

MONK-3: $(a_5 = 1 \wedge a_4 = 1) \vee (a_5 \neq 4 \wedge a_2 \neq 3)$

Here one can dismiss many of possible combinations. Should the first part, $(a_5 = 1 \wedge a_4 = 1)$, be proven not to be true, one would only need to check $a_2 \neq 3$. Thus, a tree with a depth of three is sufficient. This is probably the simplest dataset to examine.

Assignment 1

The function entropy in dtree is called repeatedly for the three dataset and presented in the following table:

Dataset	**Entropy**
MONK-1	1.0
MONK-2	0.95712
MONK-3	0.99981

Assignment 2

Entropy is a measurement of the average level of uncertainty for a variable's different outcomes, meaning how difficult it is to predict an outcome. The higher, the more unpredictable.

A uniform distribution describe a case where all possible outcomes have the same probability of happening, i.e. equally likely. Some examples of such cases could be a fair coin toss or a fair dice toss.

A non-uniform distribution describe cases where the outcomes have different probability of happening. Such cases could be a spinning wheel where some

"cake-pieces" are of different size, or a biased coin.

During the lecture it was shown that by using the following equation for calculating the entropy one could establish that the entropy for the non-uniform case is lower than for the uniform case.

$$\text{Entropy } = \sum_i -p_i \log_2 p_i \tag{1}$$

Assignment 3

Average gain for all the datasets is calculated and the information gained for each attribute is presented in the following table:

Dataset	a_1	a_2	a_3	a_4	a_5	a_6
MONK-1	0.0753	0.0058	0.0047	0.0263	0.2870	0.0007
MONK-2	0.0037	0.0025	0.0011	0.0157	0.0173	0.0062
MONK-3	0.0071	0.2937	0.0008	0.0029	0.2559	0.0071

The attributes selected as the first branch is the ones with the highest information gain for each dataset. For MONK-1 we select a_5, for MONK-2 we select $a5$, and for MONK-3 we select a_2.

Assignment 4

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{k \in \text{values}(A)} \frac{|S_k|}{|S|} \text{Entropy}\,(S_k)$$

From equation above, one can state that maximizing the information gain is equal to minimizing the following equation:

$$\sum_{k \in \text{values}(A)} \frac{|S_k|}{|S|} \text{Entropy}\,(S_k)$$

$|S|$ is constant, which means that we want to minimize the following equation:

$$\sum_{k \in \text{values}(A)} |S_k| \,\text{Entropy}\,(S_k)$$

We should therefore minimize the *Entropy(S_k)*. This is the reason for the whole idea to split on the attribute that with the most information gained.

If we use the information gain as a heuristic then we can ensure that by each split we have maximized the information gained, i.e reduced entropy the most.

Assignment 5

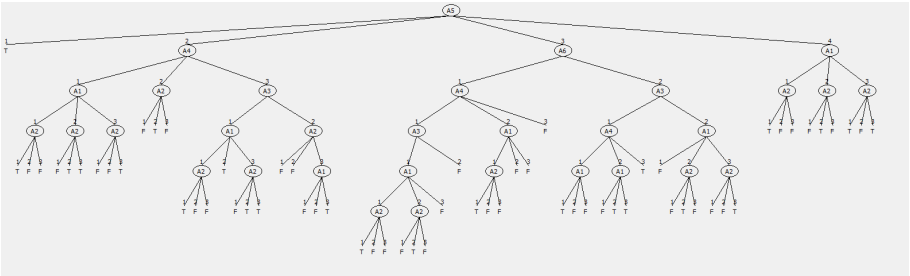


Figure 1: Decision Tree for MONK-1.

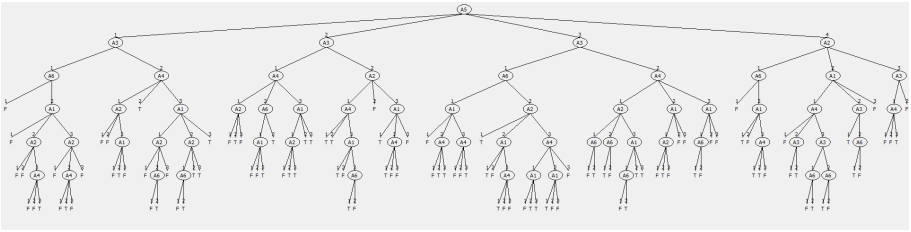


Figure 2: Decision Tree for MONK-2.

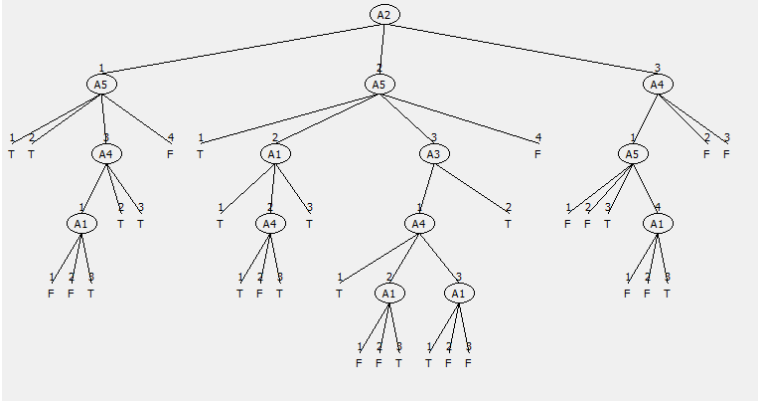


Figure 3: Decision Tree for MONK-3.

Train and test set errors for all datasets:

Dataset	E_{train}	E_{test}
MONK-1	$1.0 - 1.0 = 0$	$1.0 - 0.8287 = 0.1713$
MONK-2	$1.0 - 1.0 = 0$	$1.0 - 0.6921 = 0.3079$
MONK-3	$1.0 - 1.0 = 0$	$1.0 - 0.9444 = 0.0556$

The errors for the various datasets prove the previous assumptions for the difficulty of the datasets.

Assignment 6

Pruning is a technique for trying to reduce model complexity and simplifying the process by removing branches in our decision tree. It is done to to reduce the variance and to avoid overfitting, makes the model more usable and practical. We know however when decreasing variance leads to a higher bias.

Assignment 7

For a smaller sample of the training data, the machina can not appropriately learn the "generality" of the dataset and is the reason for the large error at the start. That's why we increase the fraction of training test vs test set. Seen in the plots, this leads to a improved accuracy.

One can read from the plots that the pruned version performs better, as we have repeatedly checked the accuracy for the training set, this is then generalised to the test set.

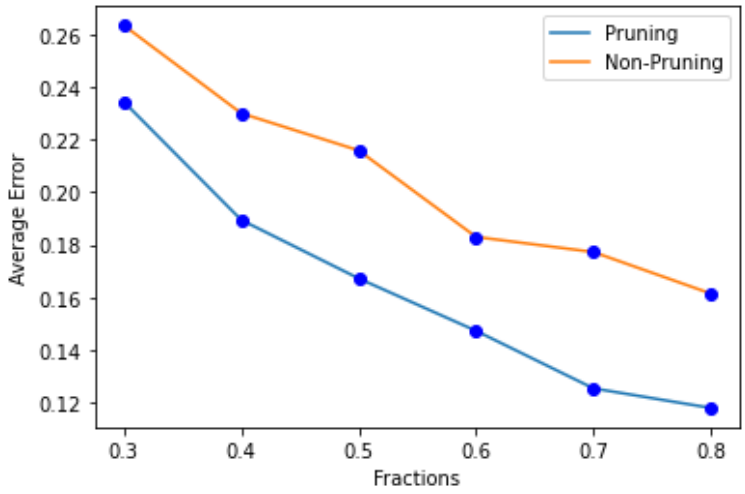


Figure 4: Average error before and after pruning, data-set MONK-1.

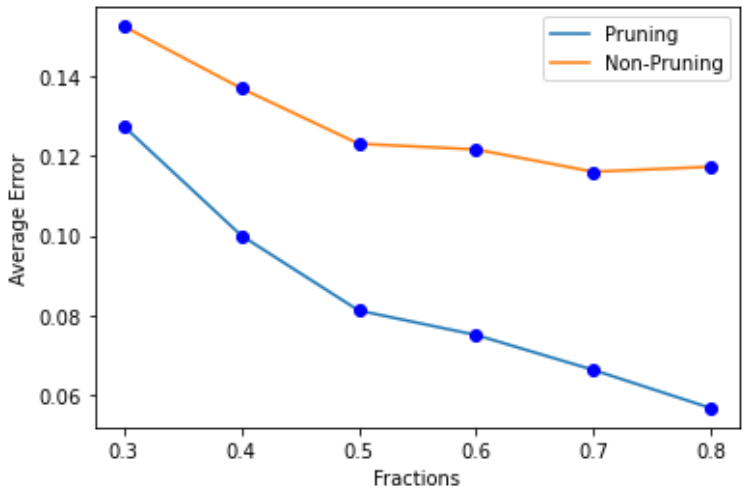


Figure 5: Average error before and after pruning, data-set MONK-3.