# DD2421 Machine Learning - Lab 1: Decision Trees

Rohit Kini & Shrivatsan Raghuram

**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.
The most difficult problem is MONK-2 as all six attributes are independent and thus to solve the decision tree, we need all six to find the outcome. While for MONK-1 and MONK-3 it is 3, but for the later there is also presence of noise which makes it more difficult compared to former.

**Assignment 1:** The file dtree.py defines a function entropy which calculates the entropy of a dataset. Import this file along with the monks datasets and use it to calculate the entropy of the training datasets.

| **Dataset** | **Entropy** |
|---|---|
| Monk-1 | 1.0 |
| Monk-2 | 0.957117428264771 |
| Monk-3 | 0.9998061328047111 |

Table 1: Entropy of each MONK dataset

**Assignment 2:** Explain entropy for a uniform distribution and a non-uniform distribution, present some example distributions with high and low entropy.
In the case of uniform distribution, the probability of each outcome is same and therefore the entropy is higher whereas in the case of a non-uniform distribution the probability of the occurrence of one outcome is more likely. There is more bias in the case of non-uniform distribution which reduces the uncertainty or entropy. This can be related to the example from class on the case of an unbiased (Entropy =2.58) and biased die (Entropy=2.16).

**Assignment 3:** Use the function averageGain (defined in dtree.py) to calculate the expected information gain corresponding to each of the six attributes. Note that the attributes are represented as instances of the class Attribute (defined in monkdata.py) which you can access via m.attributes[0], ..., .attributes[5]. Based on the results, which attribute should be used for splitting the examples at the root node?

| Dataset | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| Monk-1 | 0.07527256 | 0.00583843 | 0.00470757 | 0.0263117 | **0.28703075** | 0.00075786 |
| Monk-2 | 0.00375618 | 0.0024585 | 0.00105615 | 0.01566425 | **0.01727718** | 0.00624762 |
| Monk-3 | 0.00712087 | **0.29373617** | 0.00083111 | 0.00289182 | 0.25591172 | 0.00707703 |

Table 2: Information Gain of each MONK dataset

For MONK-1 and MONK-2 attribute A5 should be used for splitting at the root node while attribute A2 should be used for MONK-3

**Assignment 4:** For splitting we choose the attribute that maximizes the information gain, Eq.3. Looking at Eq.3 how does the entropy of the subsets, $S_k$, look like when the information gain is maximized? How can we motivate using the information gain as a heuristic for picking an attribute for splitting? Think about reduction in entropy after the split and what the entropy implies.
The Information gain and the entropy are inversely proportional. Thus, when information gain is high, entropy is low which means non-uniform distribution of data thus more certain for picking an attribute for classification.

**Assignment 5:** Build the full decision trees for all three Monk datasets using buildTree. Then, use the function check to measure the performance of the decision tree on both the training and test datasets.
Compute the train and test set errors for the three Monk datasets for the full trees. Were your assumptions about the datasets correct? Explain the results you get for the training and test datasets.

| Dataset | Error (Train) | Error (Test) |
|---|---|---|
| Monk-1 | 0 | 0.17129629629629628 |
| Monk-2 | 0 | 0.30787037037037035 |
| Monk-3 | 0 | 0.05555555555555558 |

Table 3: Train and Test errors of each MONK dataset

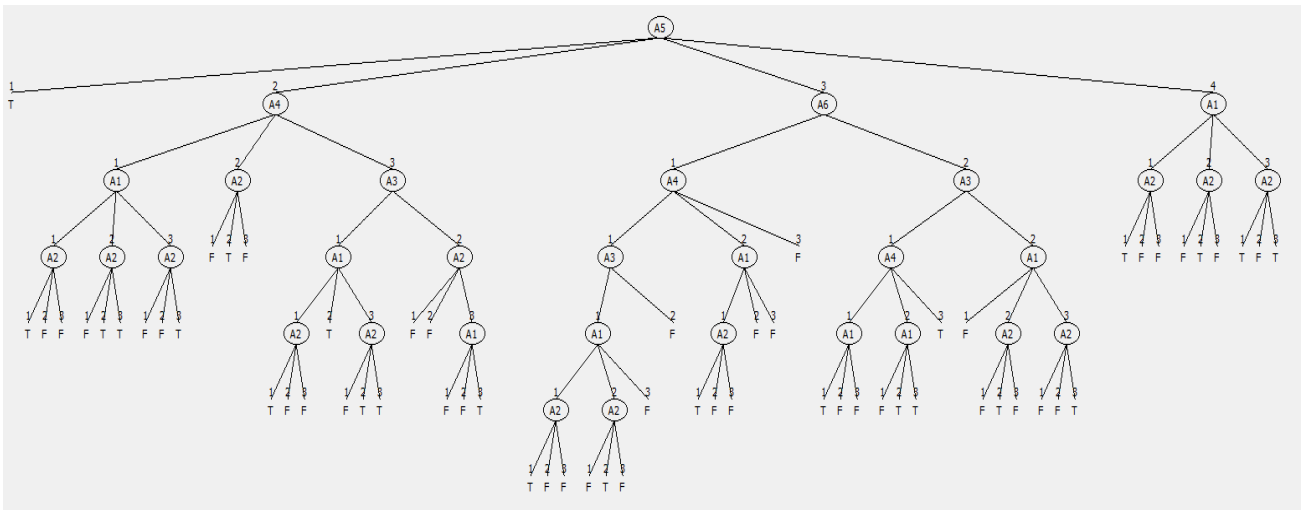

Figure 1: Decision tree of MONK-1 Dataset

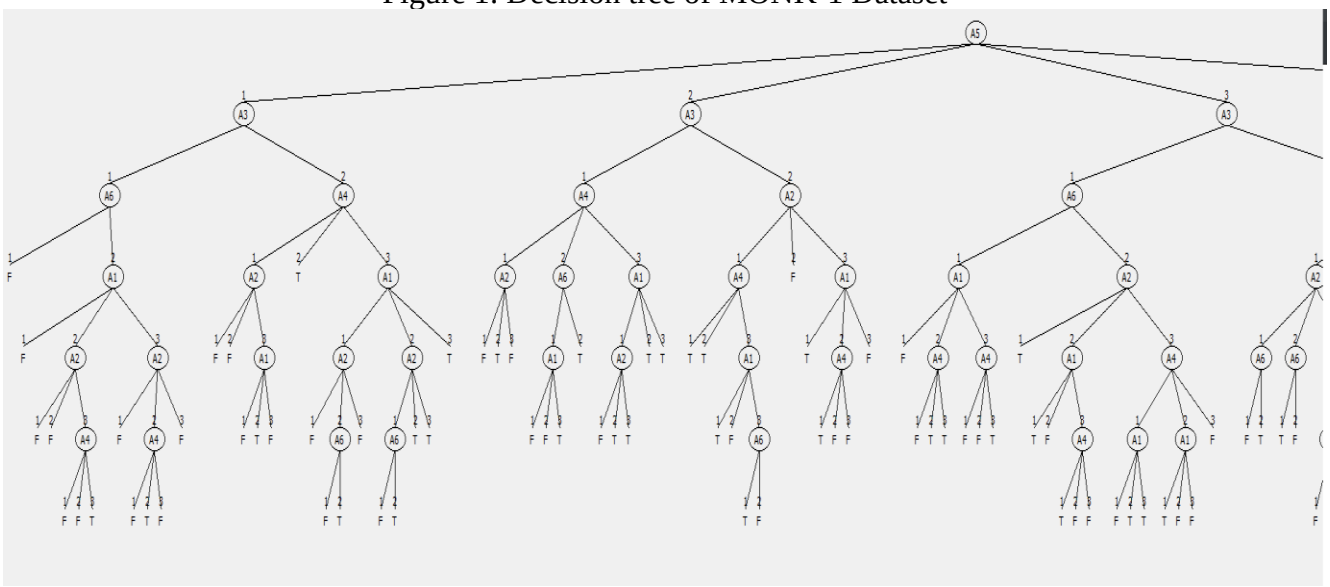

Figure 2: Decision tree of MONK-2 Dataset

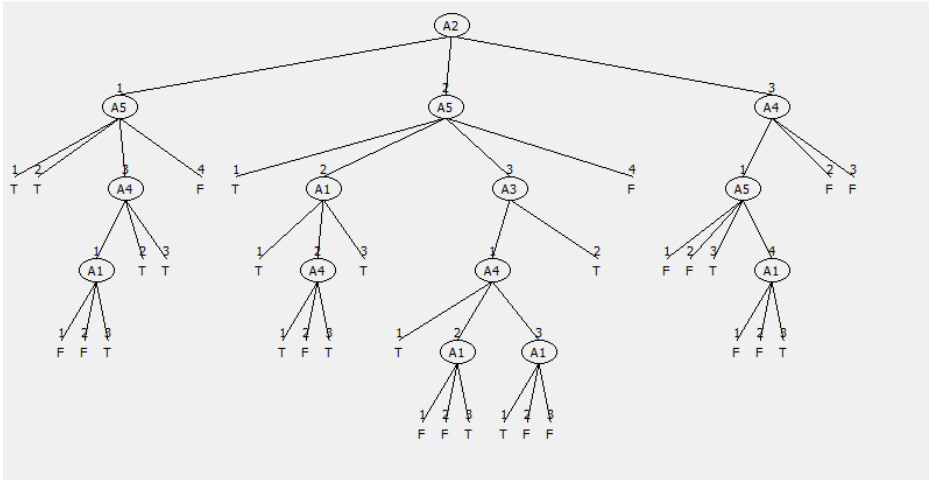

Figure 3: Decision tree of MONK-3 Dataset

**Assignment 6:** Explain pruning from a bias variance trade-off perspective.
Pruning leads to simpler model compared to the original model and thus the bias is high while the variance is low.

**Assignment 7:** Evaluate the effect pruning has on the test error for the monk1 and monk3 datasets, in particular determine the optimal partition into training and pruning by optimizing the parameter fraction. Plot the classification error on the test sets as a function of the parameter fraction.
Note that the split of the data is random. We therefore need to compute the statistics over several runs of the split to be able to draw any conclusions. Reasonable statistics includes mean and a measure of the spread. Do remember to print axes labels, legends and data points as you will not pass without them.

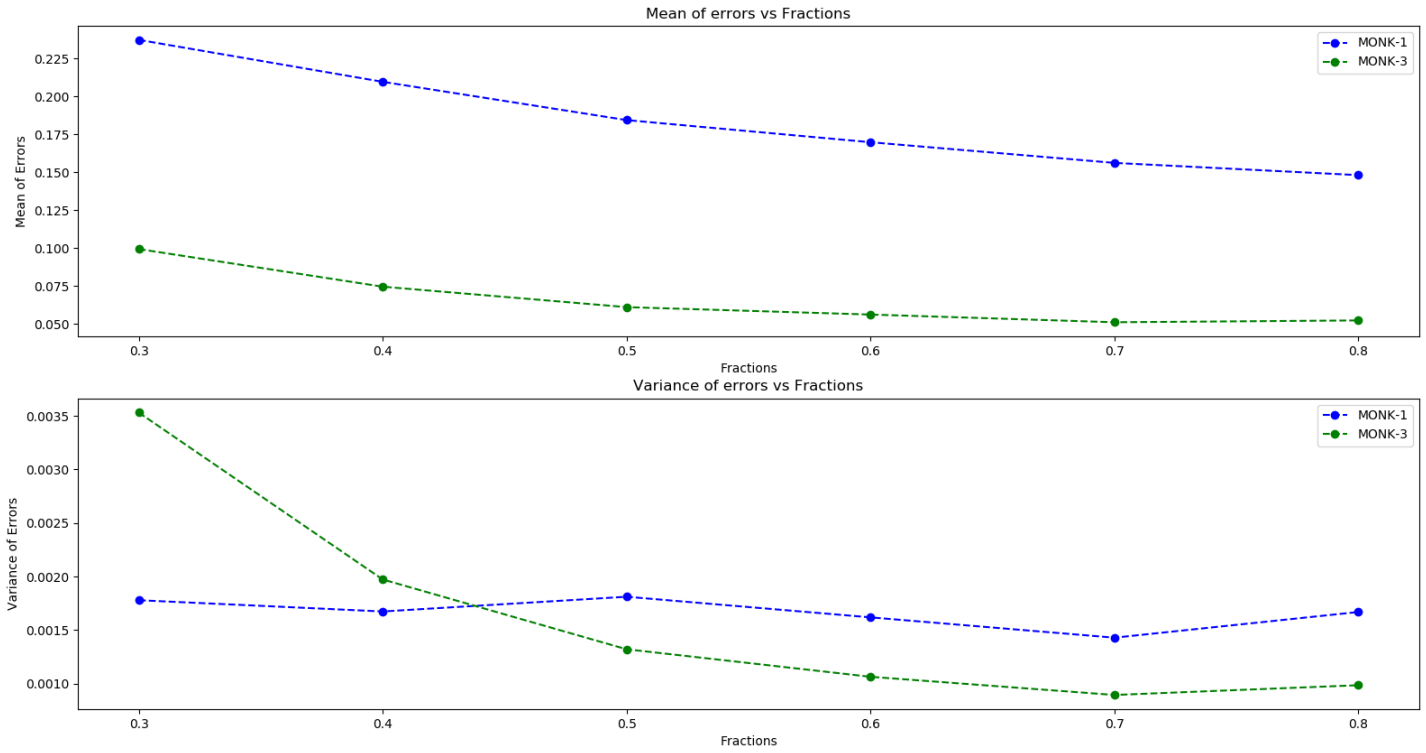


Figure 4: Error and Variance vs Fraction of MONK-1 and MONK-3

As the fraction of the splitting increases the training data increases which increases the accuracy and thus reduces the error.