Decision trees

Daniel Bányay, Muhammad Bilal Akhtar

September 16, 2018

Assignment 0 - Problems with datasets

1. Number of Training Samples: Monk 2 has maximum number of samples for training, hence it should give lowest amount of error.
2. Noise: Monk 3 training set is contaminated by noise, hence it should give highest amount of error.
3. Condition for True state: Monk 2 has a very generalized condition statement for Class "True" and turns out to be most difficult decision tree formation.
4. Data in MONK2 is not equally distributed as frequency of ones is less than 0.5, hence it should be biased and produce worst error rate

Assignment 1 - Entropy

Dataset	Entropy	Frequency of 1's
MONK-1	1.0000	62 / 124
MONK-2	0.9571	64 / 169
MONK-3	0.9998	60 / 122

- Uniform distribution sets an upper limit on entropy value.
  Entropy for Uniform Distribution = N x (1/N) x (-log_2(1/N)) = log_2(N)
  Example of a fair dice has entropy 2.58
- A non-uniform distribution gives entropy that is always less than the entropy for uniformly distributed dataset.
  An unfair dice with probability of 6 as 0.5 and all others as 0.1 gives an entropy of 2.16.
  While an unfair dice with probability of 6 as 0.95 and all others as 0.01 gives an entropy of 0.4025.

Assignment 3 - Information Gain

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	0.0008
MONK-2	0.0038	0.0025	0.0011	0.0157	0.0173	0.0062
MONK-3	0.0071	0.2937	0.0008	0.0029	0.2559	0.0071

Based on these results attribute to be used for splitting at the root node for:

- MONK-1 dataset is a_5
- MONK-2 dataset is a_5
- MONK-3 dataset is a_2

Assignment 4 - Maximizing information gain

Looking at Eq.3 how does the entropy of the subsets S_k look like when the information gain is maximized?

$$\mathrm{Gain}(S, A) = \mathrm{Entropy}(S) - \sum_{k \in \mathrm{values}(A)} \frac{|S_k|}{|S|} \mathrm{Entropy}(S_k) \quad (1)$$

How can we motivate using the information gain as a heuristic for picking an attribute for splitting?

- By choosing maximizing information gain we are decreasing entropy
- Smaller entropy leads to less randomness
- Less randomness leads to more prior knowledge about the outcome

Assignment 5 - Two level decision tree of MONK-1

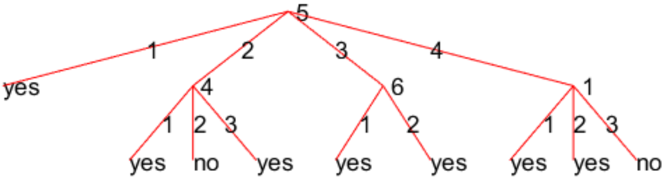

Figure 1: MONK-1 2 level decision tree with majority class as leaf nodes

Information Gain for all the attributes = 0 when a_5 is 1 in MONK-1.

Assignment 5 - Full Decision Tree

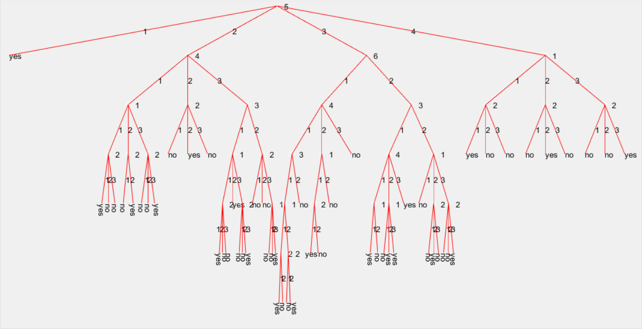

Figure 2: Full Decision Tree for MONK-1

- Total number of leaf nodes is 55 for MONK-1
- Majority of the decisions can be made on traversing 5 levels.

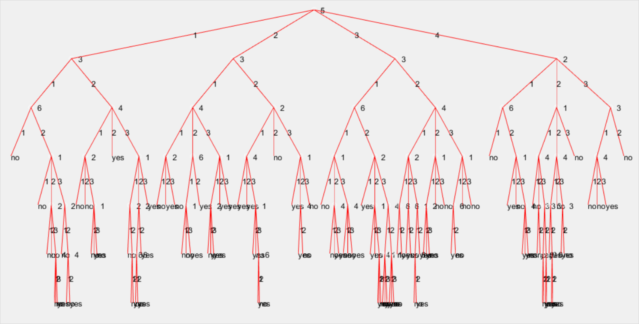

Figure 3: Full Decision Tree for MONK-2

- Total number of leaf nodes is 105 for MONK-2
- Majority of the decisions can be made on traversing 5 levels.

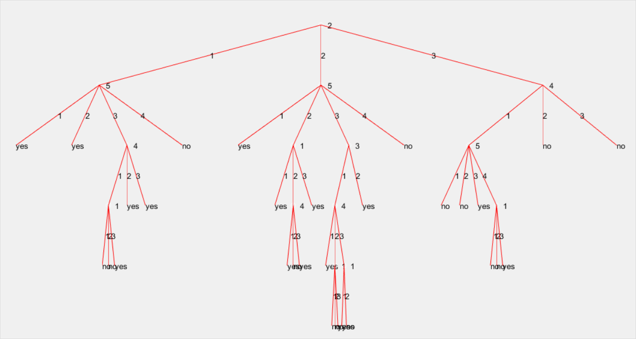

Figure 4: Full Decision Tree for MONK-3

- Total number of leaf nodes is 30 for MONK-3
- Majority of the decisions can be made on traversing 4 levels.

Assignment 5 - Error

	E_{Train}	E_{Test}
`MONK-1`	0.000000	0.129630
`MONK-2`	0.000000	0.303241
`MONK-3`	0.000000	0.055556

Training set gives zero error because we are continuously narrowing down on the data. The tree becomes highly tuned to the data present in the training set.

Assignment 5 - Discussion on Accuracy of Models

- MONK-2 gives the most errors and MONK-3 gives the least errors on Test data.
- Accuracy of model depends on the characteristic / nature of the problem for which the dataset is provided.
- MONK-3 has the simplest problem definition (a very well defined logical expression) hence its full decision tree has least number of leaf nodes. Whereas, MONK-2 has a very complex model.
- If the number of levels is too high i.e a complicated decision tree, the model tends to overfit.
- For complex decision trees even if one of the parameters deviates slightly, the condition will not be met and it will take the wrong branch leading to high amount of error.

Assignment 6 - Bias, Variance

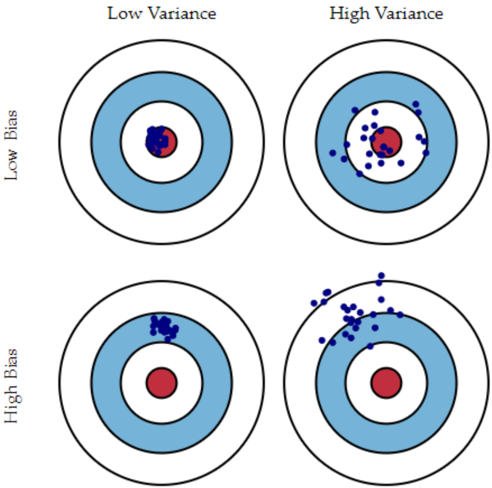

Figure 5: Bias and Variance on a 2D model

Effects of pruning

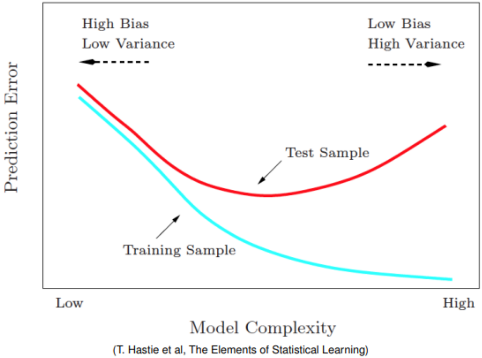

Figure 6: Bias-Variance tradeoff

- Depth of a tree determines complexity, and complex models have low bias but high variance
- With pruning, we can reduce variance by reducing the model complexity

Assignment 7 - Effects of pruning, bias error

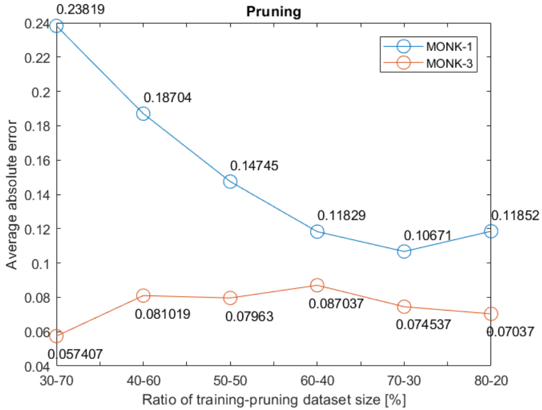

Figure 7: Effect of changing dataset size, bias

In this graph, after increasing the number of iterations, we observe that the error decreases all the way on increasing training size from 30 - 100

Assignment 7 - Effects of pruning, variance

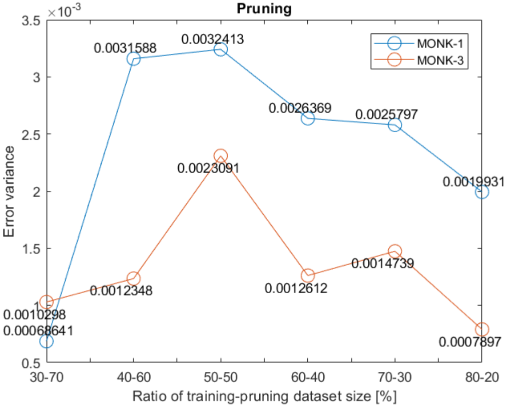

Figure 8: Effect of changing dataset size, bias, variance