Machine Learning Lab 1

Cesare De Cal & Shounak Chakraborty

Really good book to study on the assignment material here (Chapter 3).

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.

Answer: MONK-2 because it will require many samples in the training dataset to understand the pattern. For an instance of MONK-2 to be evaluated true we need *exactly* two attributes (any attribute) with value 1. Instead, the other datasets seem to be easier to learn because the attribute index is fixed and the rules are simpler, thus it's easier to see the pattern.

Answer(S) : Monk1 : Check if a1= a2, if true , we arrive at the leaf node. If not, check if a5=1, then again we arrive at the leaf node. The algorithm will be able to learn the condition faster ,and therefore more information gain. a3, a4, a5 doesn't contribute to the decision tree. Only 3 required to know the outcome.
Monk2: First check with a3 and a6, as they have a smaller domain, if they contain 1. Then we have to check each of the attributes if they contain 1. All the attributes i.e 6 need to be considered to understand the pattern. And all the outcomes will depend on the previous outcomes. Thus the hardest.
Monk3: We have to check either of the given two conditions. Only 3 attributes are used. With some noise it may get harder.

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.

Answer:

Dataset	Entropy
MONK-1	1.0
MONK-2	0.957117428264771
MONK-3	0.9998061328047111

Assignment 2

Explain entropy for a uniform distribution and a non-uniform distribution, present some example distributions with high and low entropy.

Answer: given a binary classification, entropy is zero is all if all members of a collection belong to the same class. Entropy is one when a collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1.

Answer(S) : In a uniform distribution the outcomes have an equal probability. Therefore the Entropy is high or 1. For example a fair coin.
In a non-uniform distribution, one of the values has more probability than the other therefore giving us more information and low Entropy . For example a fake dice.

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], ..., m.attributes[5]. Based on the results, which attribute should be used for splitting the examples at the root node?

Information Gain:

Dataset	a1	a2	a3	a4	a5	a6
MONK-1	0.075	0.0058	0.0047	0.026	**0.29**	0.00076
MONK-2	0.0038	0.0025	0.0011	0.016	**0.017**	0.0062
MONK-3	0.0071	**0.29**	0.00083	0.0029	0.26	0.0070

Based on the results, which attribute should be used for splitting the examples at the root node?
Root of MONK-1: a5
Root of MONK-2: a5
Root of MONK-3: a2

Because they have the highest information gain across all attributes.

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.

Answer: The information gain is given by the difference between the entropy of the original collection S, and the expected value of the entropy after S is partitioned using a given attribute. By appending the node with the highest information gain to the decision tree at each step (greedy approach), we therefore minimize the entropy.

Answer(S) : The Entropy of the subsets are lower in order to maximise the information gain.
By using the highest information gain, an attribute can be selected to perform a split. The entropy of the subset will decrease. This would lead to a limited number of classified samples.

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.

Answer:

Dataset	E_{train}	E_{test}
MONK-1	0% (100% accuracy)	17% (83% accuracy)
MONK-2	0% (100% accuracy)	30% (70% accuracy)
MONK-3	0% (100% accuracy)	6% (94% accuracy)

The accuracy for the training test is 100% since the decision tree was built from it. As we initially assumed, the MONK-2 dataset is the hardest to learn. The Monk 3 dataset has the least error and Monk2 dataset with the maximum error.

Assignment 6

Explain pruning from a bias variance trade-off perspective.

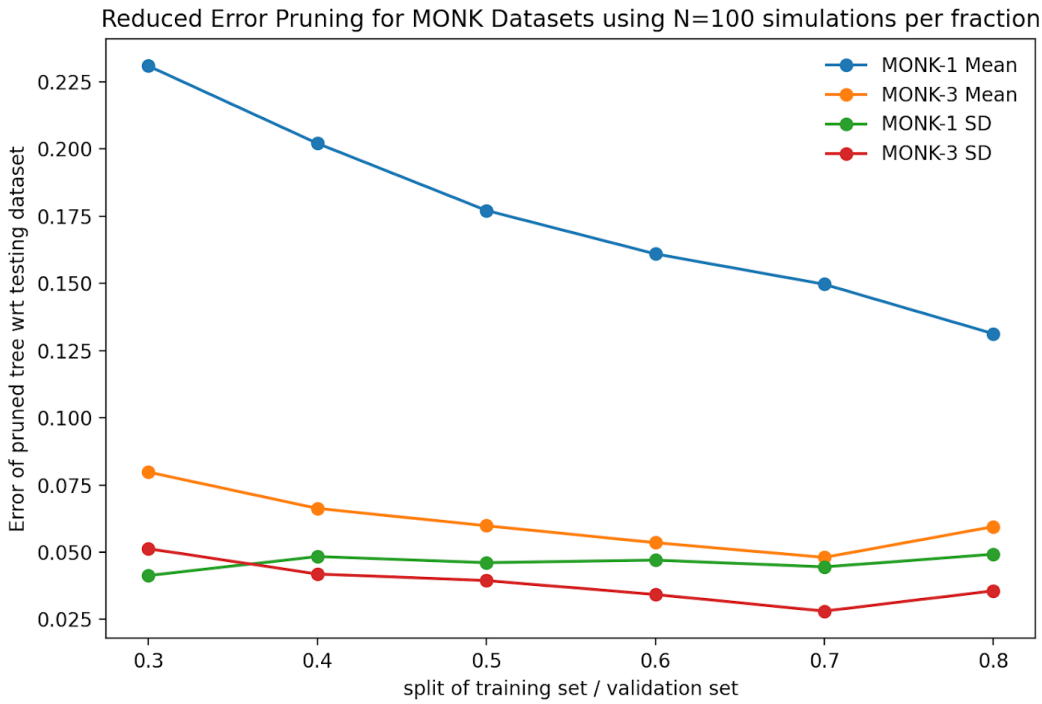
Answer: the Classification trees are usually complex models thereby contributing to higher variance. It tends to fit perfectly for a given set of data. If we include pruning and remove overfitting nodes, it may reduce the variance and increase the bias.

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 $\in$ *{0.3, 0.4, 0.5, 0.6, 0.7, 0.8}.*

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.

Answer:


Reduced Error Pruning for MONK Datasets using N=100 simulations per fraction
MONK-1 Mean
MONK-3 Mean
MONK-1 SD
MONK-3 SD
Error of pruned tree wrt testing dataset
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.3
0.4
0.5
0.6
0.7
0.8
split of training set / validation set