DD2421 HT20 - Lab 1: Decision Trees

Authors: Erfan Wu, Yage Hao

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.

Answers:

MONK-2 classification will be the hardest for a decision tree to learn. Since the classifying conditions for MONK-1 and MONK-3 are static and explicit, requiring only some attributes to be examined. However, the classification problem of MONK-2 requires observing the current status of all attributes.

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.

Answers:

Dataset	Entropy
MONK-1	1
MONK-2	0.9571
MONK-3	0.9998

Assignment 2: Explain entropy for a uniform distribution and a non-uniform distribution, present some example distributions with high and low entropy.

Answers:

Intuitively, entropy implies the uncertainty and randomness of a distribution. A uniform distribution means the realization of the variable lies arbitrarily between an interval, indicating the entropy reaches maximum. The entropy for a non-uniform distribution is always less than a uniform one’s.

For example, an extreme example is that there is only one possible result among different choices. This could be classified as a kind of non-uniform distribution, and it's entropy is 0

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?

Answers:

Dataset	a1	a2	a3	a4	a5	a6
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

For monk-1 training dataset, a5 gives most information gain. Thus at the root node, split examples based on a5.
For monk-2 training dataset, a5 gives most information gain. Thus at the root node, split examples based on a5.
For monk-3 training dataset, a2 gives most information gain. Thus at the root node, split examples based on a2.

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, Sk, 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.

Answers:

According to Eq.3, the definition of information gain, while we maximize the information gain, the entropy of the subsets, Sk, will be minimized.

According to the definition of entropy, it represents the uncertainty and unpredictability of the system or the dataset. Entropy = 1 indicates the worst classification, while Entropy = 0 indicates a complete classification.

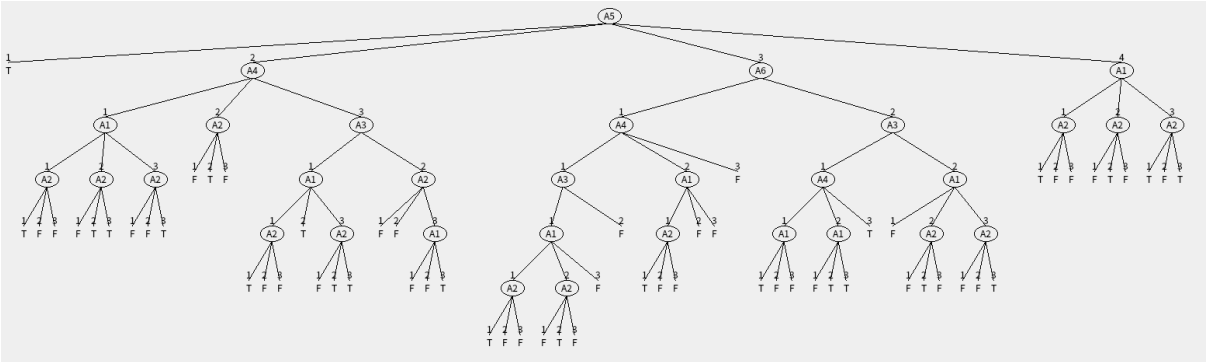
Therefore, using the IG as a heuristic and picking an attribute that maximizes the expected reduction of the entropy indicates we find the attribute that best classifies data into different categories and reduces the uncertainty.
The procedure of reducing and converging entropy from 1 to 0 means the improvement in classification correction.

Assignment 5: 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.

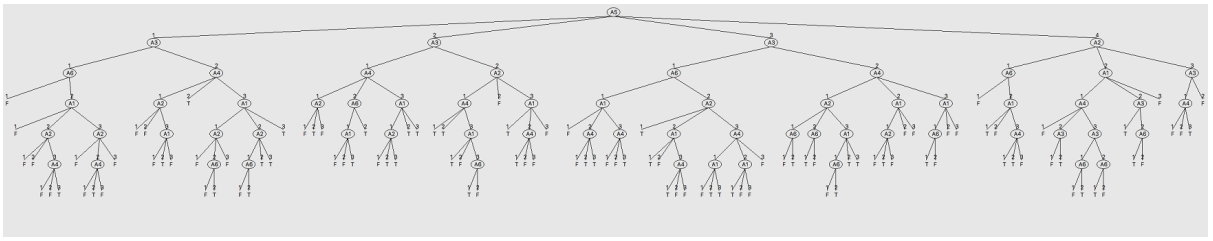
	Etrain	Etest
MONK-1	0	0.1713
MONK-2	0	0.3079
MONK-3	0	0.0556

In Assignment 0, it is predicted that MONK-2 is the most challenging dataset, and it is confirmed here. With the buildTree function, all the three training datasets produce decision trees with no error in training data because they are used to form the trees. However, on the test dataset, MONK-2 decision tree has the largest error, implying a poor classification correction.

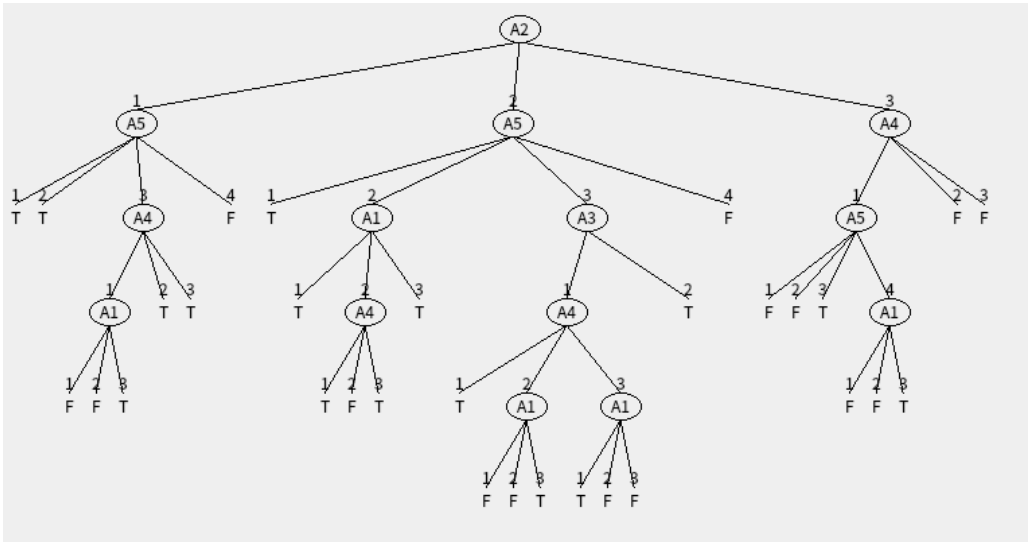
Monk1 - Tree



Monk2 - Tree



Monk3 - Tree



Assignment 6: Explain pruning from a bias variance trade-off perspective.

Answers:

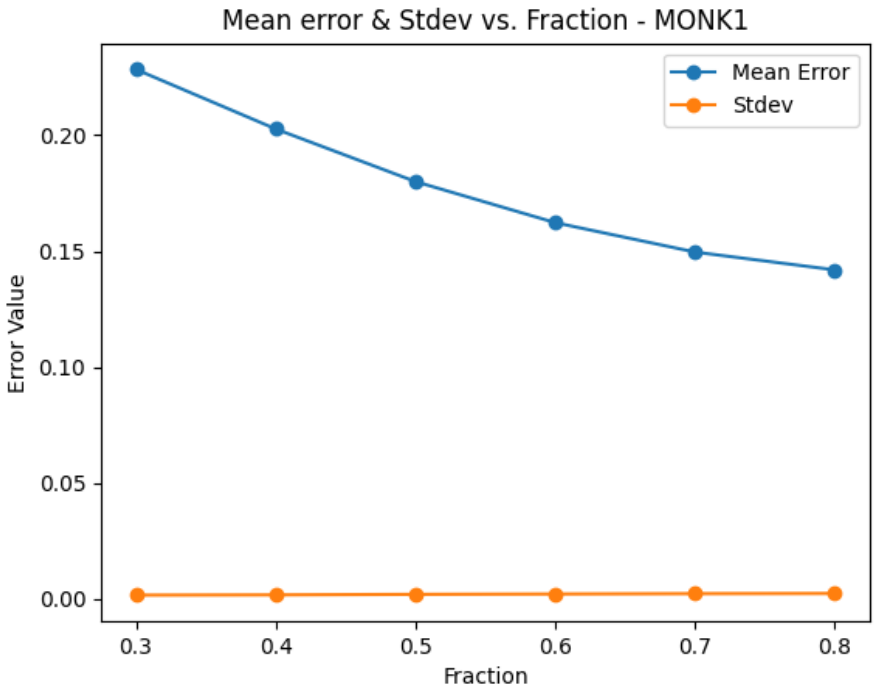
Pruning provides an approach that avoids overfitting.
In this case, firstly grow the full decision tree, and then do a reduced-error pruning. (Specifically, evaluate impact on validation set of pruning each possible node and greedily remove the one that most improves validation set accuracy.)

Bias perspective: The trees before and after pruning should have little differences in bias since we manually set the pruning threshold.

Variance perspective: As the degrees of freedom of the decision tree reducing, the variance should be reduced and thus avoid overfitting.

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

Monk1:



Monk3: