ML Decision tree

Kildo
Andrej

January 2020

Assignments

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 true concepts behind dataset MONK-2 are not related to any specific attribute but instead that the value of two of the attributes should be equal to one and all the others not equal to one. Since it is ambiguous which attributes are equal to one for individual samples the decision tree has to ask questions on all attributes to be able to classify with reasonable probability. This makes MONK-2 the hardest to learn using the decision tree algorithm.

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	0.95712
MONK-3	0.99981

Table 1: Entropy

Assignment 2

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

A uniform distribution has higher entropy than a non-uniform distribution because there is more uncertainty in a uniform. For example as in the lectures, a binary uniform coin toss has a maximum entropy of one. A uniform dice roll with six outcomes has a entropy of 2.58. A dice roll that is non-uniform has a entropy of 2.16, lower than a uniform distribution of the same scenario but still higher than a uniform one with lower number of outcomes (the coin toss).

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 in- stances 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?

Dataset	a_1	a_2	a_3	a_4	a_5	a_6
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

Table 2: Information gain

The best attribute is a_5 for MONK-1, a_5 for MONK-2 and a_2 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 tells how much the entropy decreases after the split for each attribute. Equation 3 suggest the entropy of the subset is minimized when the information gain is maximized. Low entropy implies low uncertainty.

Assignment 5

Dataset	E_{train}	E_{test}
MONK-1	0	0.171
MONK-2	0	0.307
MONK-3	0	0.056

Table 3: Train and test set errors for the three Monk datasets

Assignment 6

Explain pruning from a bias variance trade-off perspective.
The pruning simplifies the model by removing some nodes that result in a tree that does not perform worse over the validation set. Simplifying the model gives a lower model complexity which gives a lower variance but at the same time gives a higher 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]

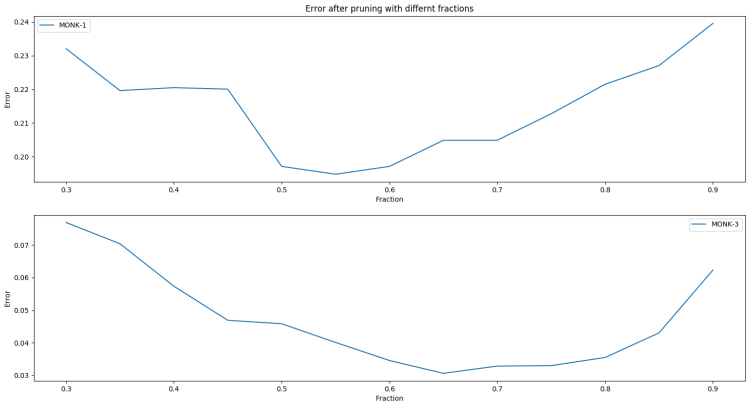

Figure 1: Mean error of performance with different fractions.

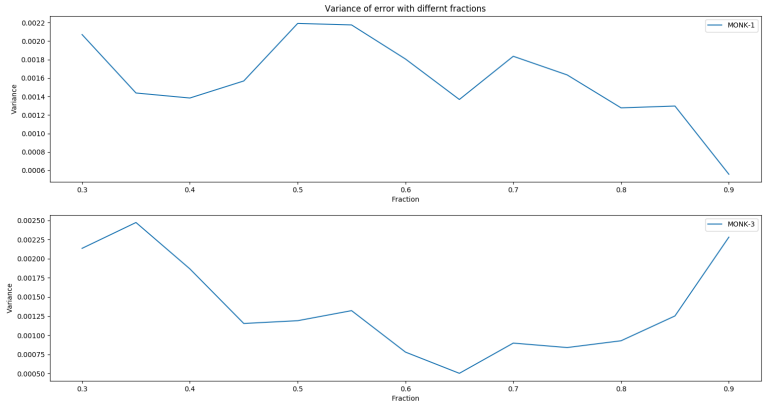

Figure 2: Variance of error