LAB1. DECISION TREES

Andrés Toledo (aatc@kth.se) aatc@kth.se

Flavia García (flaviagv@kth.se) flaviagv@kth.se

September 2019

Assignment 0

Problem MONK-2 contains information that makes it harder to learn than MONK-1 and MONK-3; the latter problems require a depth of two levels in order to process the complete dataset, meanwhile MONK-2 would require two attributes being equal to 1, resulting in a total of 15 combinations and equal number of levels for the complete dataset to be processed.

Assignment 1

The entropy of the different training datasets, rounded to the third decimal, is shown in the following table:

Dataset	Entropy
MONK-1	1.0
MONK-2	0.957
MONK-3	0.998

Assignment 2

Entropy is a measure of uncertainty of a dataset. In other words, by calculating the entropy you can measure the $'impurity'$ of your set. If a dataset has different classes with the same probability (uniform distribution) its entropy will be high. On the other hand, if you have a greater probability of having elements of one class than of the other (non-uniform distribution), its entropy will be low.

$$Entropy(S) = -p_0 log_2 p_0 - p_1 log_2 p_1$$

The entropy can be explained more clearly with an example:

Example 1. *Suppose we have a set S with apples and pears. If the probability of picking an apple is the same as picking a pear, our set is uniformly distributed. Therefore, the entropy of this set will be:*

Having:

$$P(x \in Pear) = 0.5$$
$$P(x \in Apple) = 0.5$$

Then:

$$Entropy(S) = -0.5log_2(0.5) - 0.5log_2(0.5) = 1.0$$

In contrast, if the probability of picking a pear is greater than picking an apple, the distribution of the set is non-uniform. Consequently, the entropy of this set will be:

Having:

$$P(Pear) = 0.9$$
$$P(Apple) = 0.1$$

Then:

$$Entropy(S) = -0.9log_2(0.9) - 0.1log_2(0.1) = 0.46$$

Therefore, the entropy of a uniform distribution is greater than the entropy of a non-uniform distribution.

Assignment 3

The expected information gain corresponding to each of the six attributes, rounded to the third decimal, is presented in the following table:

Dataset	a_1	a_2	a_3	a_4	a_5	a_6
MONK-1	0.075	0.006	0.005	0.026	0.287	0.001
MONK-2	0.004	0.002	0.001	0.0157	0.017	0.006
MONK-3	0.007	0.0294	0.001	0.003	0.256	0.007

Based on the results of the previous table, the attribute used for splitting the examples at the root node should be: in MONK-1 a_5, in MONK-2 a_5 and in MONK-3 a_2.

Assignment 4

According to the information gain equation:

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

Maximizing the information gain would require that the entropy of the subsets, S_k, is minimized. This is the Decision Tree algorithm target.

The information gain measures the amount of entropy drop between the set before the splitting and the subsets generated after. This split is based on one attribute, therefore the attribute with greater information gain will be the one that generates the best partition. In other words, by splitting the set based on this attribute, the tree will have less impure subsets in each branch than if the set is split with another attribute.

Assignment 5

The train and test set errors for the three MONK datasets trees were computed. These values, rounded to the second decimal, are shown in the following table:

	E_{train}	E_{test}
MONK-1	0.00	0.17
MONK-2	0.00	0.31
MONK-3	0.00	0.06

Our assumptions about the datasets were correct. We stated that the most complex dataset was MONK-2 and it is shown in the above table, that this dataset has the greatest test error. Therefore, MONK-2 true concept is more complex than the others.

The three trees generated by learning each training dataset fit perfectly. This is the reason why the error of predicting the class of the training data is zero. The error over the test data is greater than the one over the training data. This is very common. We can see that the MONK-1 and MONK-3 decision trees generalize pretty good, while the MONK-2 does not. We can state that the MONK-2 decision tree overfits the training data because the decision tree created is very complex.

Assignment 6

In Machine Learning, bias variance trade-off is very important. When the bias is low, the variance is high, so the model is accurate in predicting the training sets but it does not generalize properly

for not seen data. This is what it is called overfitting. On decision trees, pruning is done to avoid overfitting. Pruning consists of removing sections of the tree that do not improve the classification. By pruning, the variance is reduced but the bias is increased. Consequently, the pruned tree will not make so accurate classifications but it would generalize better.

Assignment 7

To prune the decision tree, we divided the dataset in train and validation sets. The percentage of data in each set (parameter $fraction$) has an effect on the pruning and consequently, on the test error. To show this effect, we have calculated the mean and standard deviation of this error over 600 iterations. We calculate the test error in more than one iteration because the dataset splitting is done randomly, so we need more than one iteration to be able to draw any conclusions.

In Figure 1 we can see the test error per each value of the parameter $fraction$ for $monk1$ dataset. The best $fraction$ value for $monk1$ dataset, based on the test error, is 0.6 and 0.7.

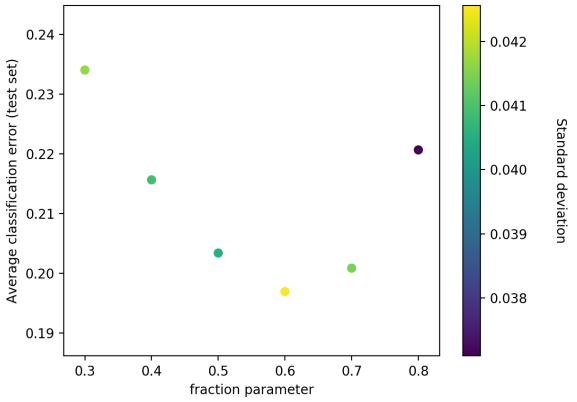

Figure 1: Scatter plot of the $monk1$ dataset test error against $fraction$ value

We have also calculated the test error depending on the $fraction$ parameter for $monk3$ dataset (2). It can be stated that the best $fraction$ value for $monk3$ dataset, based on the test error, is 0.7.

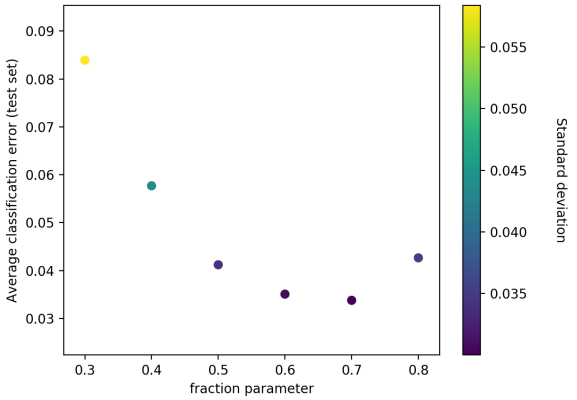

Figure 2: Scatter plot of the *monk*3 dataset test error against $fraction$ value