DD2421: Machine Learning

Lab 1 : Decision Trees

Amna Irshad (amnai@kth.se)
Aravind Nair (aanair@kth.se)

February 2, 2023

KTH Royal Institute of Technology
Stockholm, Sweden

Assignment 0

Each one of the datasets has properties which makes them hard to learn. Motivate which of the three problems is most diffcult for a decision tree algorithm to learn.

- The biggest challenge in all of the MONK dataset is the limited number of training data samples. Also the the test set is significantly larger than the training set. This makes the decission tree prone to over-fitting and poor generalisation.
- In Monk1 dataset, it is difficult to split between attributes a_1 and a_2 as they are the same in many samples.
- In Monk2 dataset, there are always two attributes having the same value as 1. This makes it difficult for the decision tree to find the best split based on a single attribute and thereby increasing the depth of the tree. This is the most difficult dataset to examine comparatively.
- Monk3 dataset has the least amount of training data and has an additional 5% noise. This makes the classification problem really difficult.

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	0.957117428264771
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.

Entropy(E) is the measure of randomness present in a dataset. The higher the entropy, the harder it is to draw any conclusions from that dataset. Consider a dataset of N classes, the entropy for this dataset can be calculated as,

$$E = -\sum_{i=1}^{N} p_i \cdot log_2(p_i) \tag{1}$$

where p_i is the probability of an element in class i.

- **Uniform Distribution:** A distribution in which all of the possible outcomes have the same probability of occurrence is called uniform distribution. In such distributions, the entropy is maximised as all the p_i are equal. Tossing of a fair dice is an example of uniform distribution.
- **Non-Uniform Distribution:** A distribution in which the possible outcomes have different probability of occurrence is called non-uniform distribution. Tossing of a biased dice is an example of non-uniform distribution.

Assignment 3

Use the function averageGain (defined in dtree.py) to calculate the expected information gain corresponding to each of the six attributes. 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	0.075273	0.005838	0.004708	0.026312	0.287031	0.000758
MONK-2	0.003756	0.002458	0.001056	0.015664	0.017277	0.006248
MONK-3	0.007121	0.293736	0.000831	0.002892	0.255911	0.007077

The largest information gain indicates the suitability for the splitting of dataset. In this way, for MONK-1, fifth attribute a_5 has the highest information gain and is used for splitting. For MONK-2, a_5 is suitable for splitting. For MONK-3, a_2 is suitable for splitting.

Assignment 4

Maximising the information gain reduces the entropy of the subset S_k. Hence, we can optimize the complexity of a tree by selecting an attribute with highest information gain to operate a split.

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.

Dataset	E_{train}	E_{test}
MONK-1	0	0.17129
MONK-2	0	0.30787
MONK-3	0	0.0555

Highest error rate is observed in case of MONK-2, then MONK-1 and MONK-3.

Assumptions vs Outcomes:

- It was assumed that Monk-2 dataset would be the most difficult one to classify. This assumption proved to be correct.
- Monk-3 dataset was assumed to be more challenging than Monk-1 as the former had less training samples and an additional 5% noise. However, the results show that this assumption was wrong.
- Due to the nature of the attributes, it was assumed that the Monk-2 dataset will result with the largest decision tree as there were no correlation within the attributes. This assumption proved to be correct.
- It was assumed that the error will be zero when the decision tree is trained and tested on the same dataset as it is able to classify the data perfectly. Also the test results confirm that the models are overfitted on the training dataset.

Assignment 6

Pruning is a technique to reduce the decision tree complexity, by discarding its least important nodes. Pruning helps in reducing the variance and make the tree less prone to overfitting. As variance and bias are inversely proportional to each other, reducing the variance results in increasing the bias.

Assignment 7

Dataset	E_{test} on Complete Tree	Best mean E_{test} on Pruned Tree	Best Fraction
MONK-1	0.17129	0.137	0.8
MONK-3	0.0555	0.044	0.7

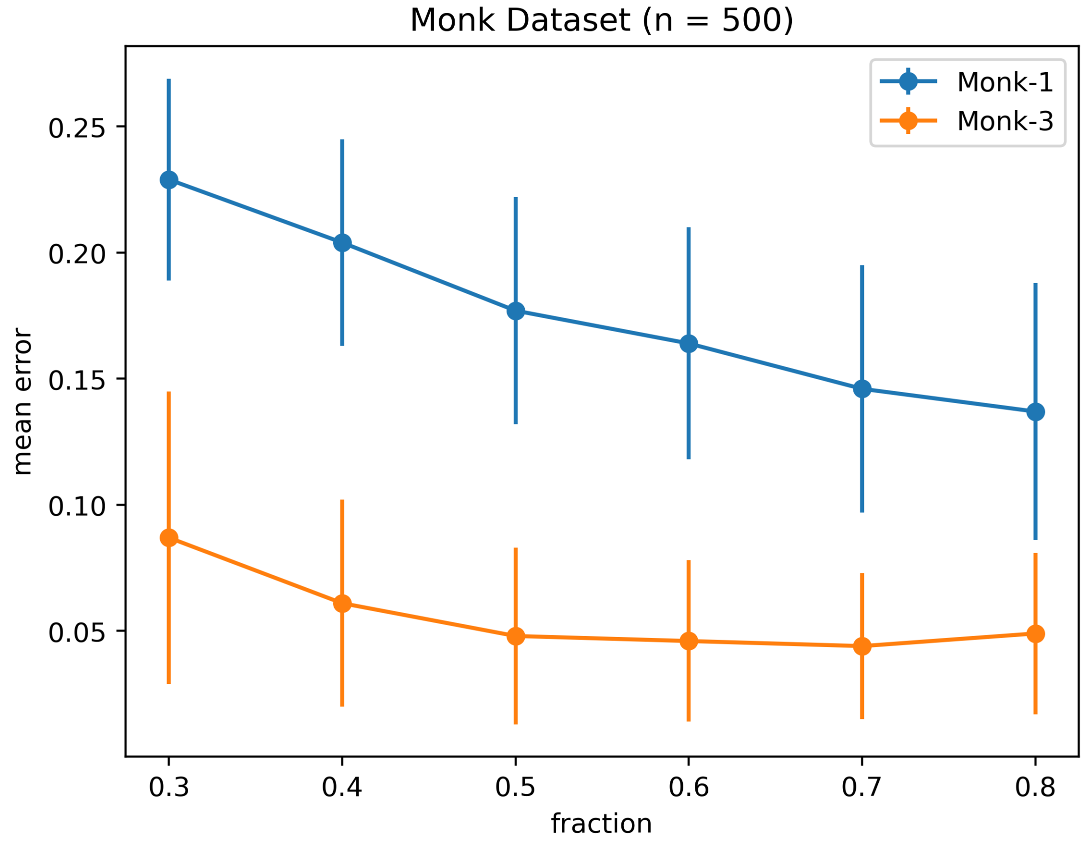
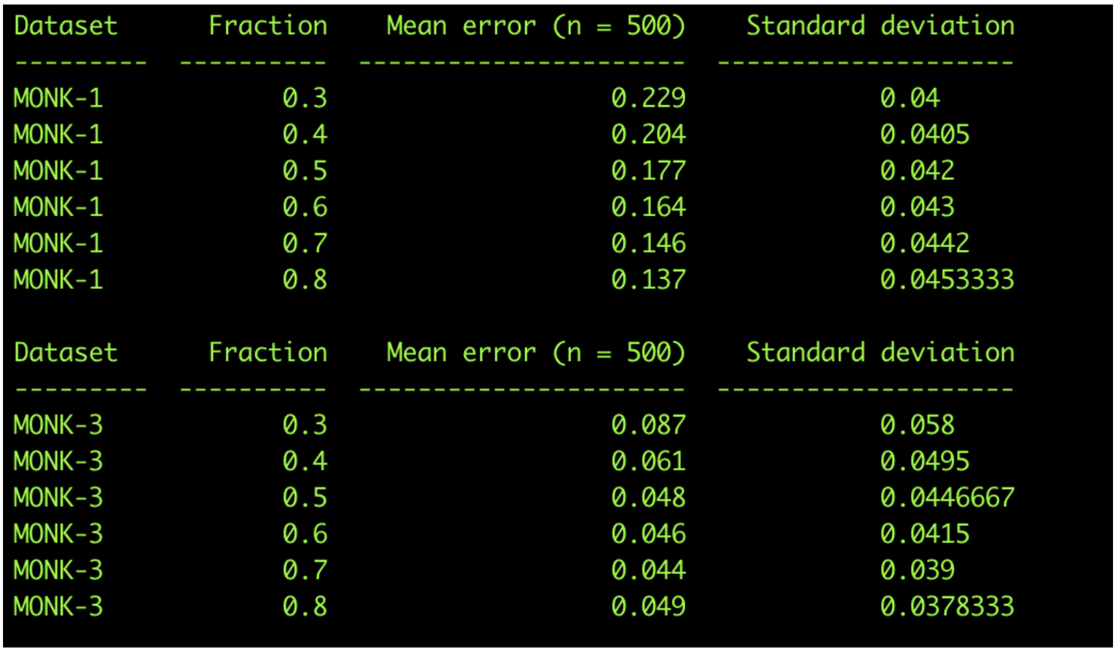


Figure 1: Error bar representation of pruned decision trees on Monk-1 and Monk-3 test dataset



```
Dataset      Fraction    Mean error (n = 500)    Standard deviation
---------  ----------  ----------------------  --------------------
MONK-1            0.3                   0.229             0.04
MONK-1            0.4                   0.204             0.0405
MONK-1            0.5                   0.177             0.042
MONK-1            0.6                   0.164             0.043
MONK-1            0.7                   0.146             0.0442
MONK-1            0.8                   0.137             0.0453333

Dataset      Fraction    Mean error (n = 500)    Standard deviation
---------  ----------  ----------------------  --------------------
MONK-3            0.3                   0.087             0.058
MONK-3            0.4                   0.061             0.0495
MONK-3            0.5                   0.048             0.0446667
MONK-3            0.6                   0.046             0.0415
MONK-3            0.7                   0.044             0.039
MONK-3            0.8                   0.049             0.0378333
```

Figure 2: Detailed results of pruned decision tree on Monk-1 and Monk-3 test dataset