1. **Assignment 0**

**Each one of the datasets has properties which makes them hard to learn.**

Monk 1 seems to be the easiest data set to predict since it involves a few attributes

Monk 2 involve all attributes and it’s the combination of at least 2 attributes to have response True

Monk 3 looks like a tree, but it involves many attributes and has noise

**Motivate which of the three problems is most difficult for a decision tree algorithm to learn.**

Monk 2 is the most difficult to learn because the function need to take into account all attributes at a time and the interaction between attributes.

1. **Assignment 1**

Calculating entropy

|  |  |
| --- | --- |
| **Dataset** | **Entropy** |
| MONK-1 | 1.0 |
| MONK-2 | 0.95712 |
| MONK-3 | 0.99981 |

1. **Assignment 2**

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

Uniform distribution is the where the probability of each event is the same.

Non-uniform distribution is where the probabilities of events are not the same.

Given a random variable that has n events

Entropy is maximum in case of uniform distribution, i.e., pi = 1/n

Entropy is minimum = 0 in case of a distribution with an absolutely certain event.

Ex: Having 3 six-side dices A, B, C

Dice A: pi = 1/6 with i= (1,2, … ,6)

Entropy = -6\*1/6\*log2(1/6) = 2.585

Dice B: p1 = 1/12; p2 = 2/12; p3 = 3/12; p4 = 3/12; p5 = 2/12; p6 =1/12

Entropy = 2.459

Dice 3: p1 = 6/12; p2 = 1/12; p3 = 1/12; p4 = 1/12; p5 = 1/12; p6 =2/12

Entropy = 2.125

1. **Assignment 3**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Information gain** | | | | | |
| **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 |

Based on the results, which attribute should be used for splitting the examples at the root node?

Monk1 = A5

Monk2 = A5

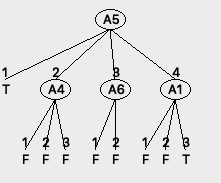
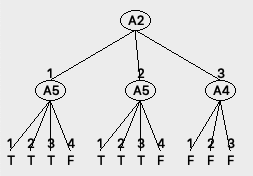
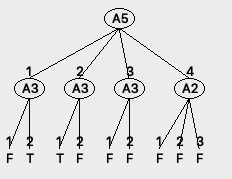
Monk3= A2

1. **Assignment 4**

**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 (short term solution without optimal rationale) for picking an attribute for splitting? Think about reduction in entropy after the split and what the entropy implies.**

When the information gain is maximized, entropy of Sk is minimum. Entropy implies the total unknown information that we want to know, each attribute of the data set contains some of the total unknown information. When choosing an attribute to ask, we want to choose the attribute that has most information, i.e., after knowing about that attribute (splitting data based on that attribute) the left-over unknown information is smallest. When the information gain is maximised, the reduction in entropy is maximized.

1. **Assignment 5**



Monk 1

Monk 2

Monk 3

Compute the train and test set errors for the three Monk datasets for the full trees.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Etrain** | **Etest** |
| MONK-1 | 0 | 0.171 |
| MONK-2 | 0 | 0.305 |
| MONK-3 | 0 | 0.056 |

**Were your assumptions about the datasets correct?**

Yes, about Monk 2 is the most difficult to learn.

No, about Monk 1 is the easiest

**Explain the results you get for the training and test datasets.**

In the 2 training datasets, the accuracy rate = 1, i.e., the full trees can classify correctly on 100% observations.

In the test datasets, Monk1 has accuracy rate = 83%, i.e., 17% observations in this dataset were misclassified. Monk2 has accuracy rate = 69%, i.e., 31% observations were misclassified. Monk3 has accuracy rate = 94%, i.e., 6% observations were misclassified.

1. **Assignment 6**

Explain pruning from a bias variance trade-off perspective.

The inability for a machine learning method to capture the true relationship is called bias

When the method is simple, it might be biased but it’s consistent when trying in different data sets

The difference in fits between the data sets is called variance

When the method has high variance, it’s difficult to predict how well it will do in other datasets

Low Bias = can accurately model the true relationship

Low variability = producing consistent prediction across different datasets

1. **Assignment 7**

Evaluate the effect pruning has on the test error for the monk1 and monk3 datasets

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Since we did the resampling, we could use the bootstrap CI which is the 25th and 75th ranked mean value as lower and upper boundaries.