# 1

# 2

Each one of the datasets has properties which makes them hard to learn. Can you guess which of the three problems is most difficult for a decision tree algorithm to learn?

* MONK-2 since it does not have any specific attribute that makes it true of false.

# 3

**Assignment 1:**

|  |  |
| --- | --- |
| **Dataset** | **Entropy** |
| **MONK-1** | **1.0** |
| **MONK-2** | **0.957** |
| **MONK-3** | **0.9998** |

# 4

**Assignment 2:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **a\_1** | **a\_2** | **a\_3** | **a\_4** | **a\_5** | **a\_6** |
| **MONK-1** | **0.0753** | **0.0058** | **0.0047** | **0.0263** | **0.2870** | **0.0008** |
| **MONK-2** | **0.0037** | **0.0025** | **0.0010** | **0.0157** | **0.0173** | **0.0062** |
| **MONK-3** | **0.0071** | **0.2937** | **0.0008** | **0.0029** | **0.2559** | **0.0070** |

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

* MONK-1 should have a5 as root node, MONK-2 should have a4 and MONK-3 should have a2.

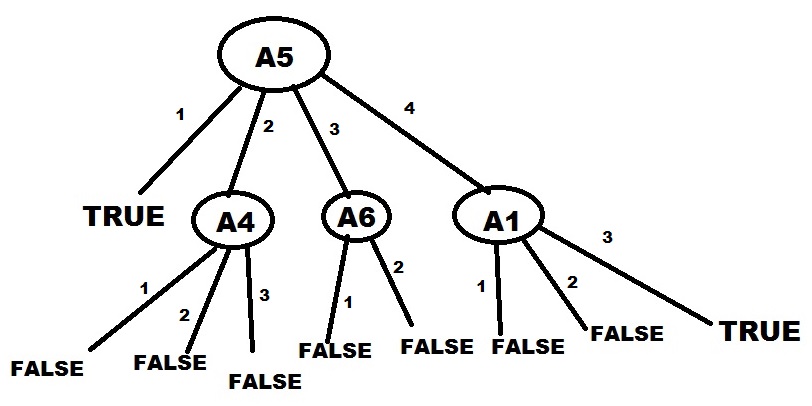
# 5

Split the monk1 data into subsets according to the selected attribute using the function select (again, defined in dtree.py) and compute the information gains for the nodes on the next level of the tree. Which attributes should be tested for these nodes?

* Gains for the second level were calculated by using the *select* function to divide the attributes on a5. Selection a5=1 gave entropy = 0 --> all true in this case. The gains can be seen in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **a\_1** | **a\_2** | **a\_3** | **a\_4** | **a\_5** | **a\_6** |
| **a5 = 2** | **0.0402** | **0.0151** | **0.0373** | **0.0489** | **0.0000** | **0.0258** |
| **a5 = 3** | **0.0331** | **0.0022** | **0.0180** | **0.0191** | **0.0000** | **0.0451** |
| **a5 = 4** | **0.2063** | **0.0340** | **0.0259** | **0.0759** | **0.0000** | **0.0033** |

The leaf nodes where calculated by *mostCommon*



**Assignment 3:**

|  |  |  |
| --- | --- | --- |
|  | **E\_train** | **E\_test** |
| **MONK-1** | **0%** | **17%** |
| **MONK-2** | **0%** | **31%** |
| **MONK-3** | **0%** | **6%** |

# 6

**Assignment 4 (error in %):**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9(extra)** |
| **MONK-1** | **25** | **21** | **19** | **18** | **17** | **14** | **15** |
| **MONK-3** | **48** | **45** | **44** | **48** | **45** | **21** | **26** |

0.8 -> 80% of the available data is training data; hence 20% is validation data.

Figur . Pruning