Lab 1 - Decision Trees

Martin Pettersson
martinp4@kth.se

October 12, 2015

1 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	0.957117428265
MONK-3	0.999806132805

2 Assignment 2

Use the function averageGain (defined in dtree.py) to calculate the expected information gain corresponding to each of the six at- tributes. Note that the attributes are represented as instances of the class Attribute (defined in monk-data.py) which you can access via m.attributes[0], ..., m.attributes[5].

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	0.255912	0.007077

Based on the results, which attribute should be used for splitting the examples at the root node?
Answer: According to wiki, you should choose an attribute with high mutual information, which in this case is a_5.

3 Assignment 3

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

	E_{train}	E_{test}
MONK-1	1.0	0.828703703704
MONK-2	1.0	0.69212962963
MONK-3	1.0	0.944444444444

4 Assignment 4

```
def partition(data, fraction):
    ldata = list(data)
    random.shuffle(ldata)
    breakPoint = int(len(ldata) * fraction)
    return ldata[:breakPoint], ldata[breakPoint:]

monk1train, monk1val = partition(m.monk1, 0.6)

def prune_tree(tree, validation):
    pruned_trees = d.allPruned(tree)
    pruned_trees_performance = [0 for x in range(len(pruned_trees))]
    for candidate in pruned_trees:
        index = pruned_trees.index(candidate)
        pruned_trees_performance[index] = d.check(candidate, validation)
    if d.check(tree, validation) <= max(pruned_trees_performance):
        tree = pruned_trees[pruned_trees_performance.index
                                        (max(pruned_trees_performance))]
        tree = prune_tree(tree, validation)
    return tree
```

Dataset	$f = 0.3$	$f = 0.4$	$f = 0.5$	$f = 0.6$	$f = 0.7$	$f = 0.8$
MONK-1	0.8125	0.814	0.768	0.754	0.771	0.708
MONK-3	0.958	0.935	0.977	0.963	0.944	0.949