
Exercise Sheet No 10

July 04, 2002

Deadline: July 12, 2002, before the lecture

1 bonus point

Exercise 10.1

(4 points)

Specify decision trees representing the following boolean functions:

1. $A \text{ XOR } B$
2. $(A \wedge B) \vee (C \wedge D)$

Exercise 10.2

(4 points)

Here we will practice the basic information-theoretical concepts used to build decision trees. Consider the following set of training examples

a_1	a_2	Classification
T	T	+
T	T	+
T	F	-
F	F	+
F	T	-
F	T	-

What is the entropy of this collection of training examples with respect to the target function *classification*? What is the information gain of a_2 relative to these training examples?

Exercise 10.3

(6 points)

In this exercise you will construct two learning curves using C4.5 which is a software extension of the decision tree learner discussed in the lecture. The software is downloadable from <http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz>. A C4.5 tutorial can be found at

<http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html>

Consider the series of training sets `monk1_20.names`, `monk1_20.data`, `monk1_20.test`, `monk1_40.names`, `monk1_40.data`, `monk1_40.test...` (see `monk.tgz`) containing 20, 40, 60, and 80 examples. For each of these training sets construct a decision tree using `c4.5`, i.e.

```
c4.5 -u -g -f monk1_20
```

```
...
```

Test the accuracy (percentage of correctly classified examples) of the unpruned decision trees on the corresponding training and test sets. Plot the number of examples vs. the accuracies, and discuss the resulting learning curves in a nutshell.

Exercise 10.4

(6 points)

Consider the task of learning the target concept “Japanese Economy Car” from the following sequence E of examples:

<i>Country of Origin</i>	<i>Manufacturer</i>	<i>Color</i>	<i>Decade</i>	<i>Type</i>	<i>ExampleType</i>
Japan	Honda	Blue	1980	Economy	Positive
Japan	Toyota	Green	1970	Sports	Negative
Japan	Toyota	Blue	1990	Economy	Positive
USA	Chrysler	Red	1980	Economy	Negative
Japan	Honda	White	1980	Economy	Positive

The attribute *ExampleType* indicates whether or not the example describes a Japanese economy car. The task is to learn to predict the value of *ExampleType* for an arbitrary car based on the values of its other attributes. Give the sequence of the S and G boundary sets computed by the CANDIDATE-ELIMINATION algorithm if it is given the sequence E of training examples. Now, add the following two examples

<i>Country of Origin</i>	<i>Manufacturer</i>	<i>Color</i>	<i>Decade</i>	<i>Type</i>	<i>ExampleType</i>
Japan	Toyota	Green	1980	Economy	Positive
Japan	Honda	Red	1990	Economy	Negative

to the end of the sequence E . Let the resulting sequence of training examples be E' . Give the sequence of the S and G boundary sets computed by the CANDIDATE-ELIMINATION algorithm if it is given the new sequence E' of training examples.