Exercises for MI

Exercise sheet 7

Thomas Dyhre Nielsen

Note: Some of the exercises below asks you to solve the exercises using Weka. If you feel adventurous (or perhaps would like to get some hands-on programming experience) you are also most welcome to solve these exercises using other (programming) tools such as scikit-learn, which support decision tree learning.

When you have completed the exercises below, continue with the remaining exercises from the last session (if any) or the decision tree related questions from the last exams.

Exercise 1* Give decision trees to represent the following Boolean functions:

- $A \vee \neg B$
- $A \wedge (B \vee C)$
- $A \ XOR \ B$
- $(A \vee B) \wedge (B \vee C)$

Exercise 2 Download and install the WEKA data-mining toolbox:

`http://www.cs.waikato.ac.nz/ml/weka/`

WEKA provides several user-interfaces. Select the 'Explorer' interface from the 'Applications' menu, and try the following:

- Load the 'Iris' dataset. This dataset contains measurements from 150 individual plants of the genus Iris, belonging to 3 different species 'Iris setosa', 'Iris versicolor', and 'Iris virginica'. The machine learning task associated with this dataset is: predict the species from the four measurement values.
- Use the 'Visualize' tab to get an overview of the attribute values and their relation to the class label. Sketch by hand a small decision tree for predicting the class label.

- Use WEKA's decision tree construction methods to build a decison tree (under the 'Classify' tab select e.g. J48 or the SimpleCart classifier). Compare with your own proposed decision tree.

Exercise 3

- Download the Pregnancy dataset. Note that the format of this file does not follow the standard file-format used by Weka. When trying to load the file you will therefore have to use the 'converter' suggested by Weka.
- Construct a decision tree for classification. Try to reason about the structure of the tree. Hint: have a look at the underlying Bayesian network model (which can be found here) that we have previously looked at in the course.

Exercise 4* Consider a database of cars represented by the five training examples below. The target attribute *Acceptable*, which can have values `yes` and `no`, is to be predicted based on the other attributes of the car in question. These attributes indicate a) the age of the car (*Age* having values < 5 `years` and ≥ 5 `years`), b) the make of the car (*Make* having states `Toyota` and `Mazda`), c) the number of previous owners (*#Owners* having values `1`, `2` and `3`), d) the number of kilometers (*#Kilometers* having values $> 150k$ and $\leq 150k$) and e) the number of doors (*#Doors* having values 3 and 5).

	Attributes					Target
	Age	*Make*	*#Owners*	*#Kilometers*	*#Doors*	*Acceptable*
1	< 5	`Mazda`	`1`	$> 150k$	`3`	`yes`
2	≥ 5	`Mazda`	`3`	$> 150k$	`3`	`no`
3	≥ 5	`Toyota`	`1`	$\leq 150k$	`3`	`no`
4	≥ 5	`Mazda`	`3`	$> 150k$	`5`	`yes`
5	≥ 5	`Toyota`	`2`	$\leq 150k$	`5`	`yes`

a) Calculate the entropy for the attribute *#Owners*.[1]

b) Show the decision/classification tree that would be learned by the learning algorithm assuming that it is given the training examples in the database.

c) Show the value of the information gain for each candidate attribute at each step in the construction of the tree.

Exercise 5 (part of it: *) Solve Exercise 7.3 (except sub-question f) in PM.

[1] Note that $\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$.