TDT4171 Artificial Intelligence Methods

Exercise 4

Mar 2nd, 2012

- **Delivery deadline: Mar 13, 2012** by 25:59.
- Required reading for this assignment: Chapter 18 (the parts in the curriculum)
- Deliver your solution on *It's Learning*
- Students can NOT work in groups. Each student can only submit solution individually.
- This homework counts for 3% of the final grade.
- Cribbing from other students (koking) is not accepted, and if detected will lead to the assignment being failed.

Introduction

In this exercise you will implement a decision tree learning algorithm. Please make sure to read the exercise completely before starting the implementation.

You will implement the decision tree learning algorithm (pseudo-code in Figure 18.5 in the textbook). You can do this using the programming language you want, but make sure to deliver your source file, a print-out of the results, and appropriate discussion. Your source-files should be ready to compile; any non-standard dependencies must be specified and supplied. **Note also that just submitting your source code is not a sufficient answer - make sure to answer specific questions in the text!**

Tasks

You should implement two different versions of CHOOSE-ATTRIBUTE (see Fig. 18.5):

1. A random selection of the possible attributes
2. Selection based on information gain, as discussed in the lecture.

To compare the two versions of CHOOSE-ATTRIBUTE, you should examine each of them by doing the following steps:

1. Learn a decision tree from the data in `training.txt`
2. Document the tree you got in your report
3. Classify all examples in the test-set (given in the text-file `test.txt`), and calculate the accuracy of the learner by comparing to the correct classification of the examples in the test-set.

Discuss your findings:

- What can you conclude about the results you have obtained? Which CHOOSE-ATTRIBUTE is better, and why?
- Do you get the same result if you run the random CHOOSE-ATTRIBUTE several times?
- What happens when you run the learner based on Information Gain several times?

Experiment Data

Two data-files are in the ZIP-file you have downloaded. The two data files have the same format: Each line describes an object, the first seven numbers are the attributes, the last number is the class of that object. All attributes as well as the class take values `1` or `2`, and you can take advantage of this to simplify your code if you want.

Example:
The first line in the training data is:

`1 1 2 2 1 1 1 1`.

This means that for this object, we have

- Attribute 1 = `1`,
- Attribute 2 = `1`,
- Attribute 3 = `2`,
- Attribute 4 = `2`,
- Attribute 5 = `1`,
- Attribute 6 = `1`,

- Attribute 7 = `1`,
- ... and that the object comes from class `1`.

The test data has the same format. The class label for the test data is not to be used by the decision tree during learning or classification, only to quantify the learning algorithm's accuracy afterwards.