Assignment 1

CSG220, Spring 2007
Due: Thursday, Jan. 25

1. Do Exercise 2.4 (p. 48).

2. Do Exercise 3.1 (p. 77).

3. Carry out the information gain computations by hand to verify that the subtree under the the Outlook=Rain branch is indeed correct according to the ID3 algorithm. (The PlayTennis tree is shown in Figure 3.1 on p. 53.)

4. Given any pure conjunctive concept (with discrete attribute values), describe the general shape of a simple decision tree that always gives the same classifications as this concept description.

5. Run a decision tree learning program on the PlayTennis data of the textbook (Table 3.2, p. 59). You may either write one from scratch yourself, use the Common Lisp program provided at the textbook web site (by following the "Software and data" link there), or use the one in the suite of Common Lisp programs provided by me under the "Homegrown Machine Learning Software" link. If you use the one at the textbook site you'll see that the PlayTennis data is already included with the program. If you use my software suite, you'll see that there is a file `play-tennis-data.lisp` containing this data in a format suitable for use with these programs. For help in using my suite of programs, consult the `README` file. (Yet another option: download another decision tree program from elsewhere. But for this assignment be sure your program does not apply any pruning to the tree.)

a. First run it on all 14 examples and confirm that it generates the tree shown in Figure 3.1 (p. 53) of the texbook.

b. Next, randomly divide the 14 examples into two halves, one for training and one for testing. Show the tree obtained. What are the training set and test set accuracies you get in this experiment?

What to turn in for this problem: (a) any source code for your programs that you wrote or modified yourself, or a description of what code you used and where you obtained it if none of it was written by you; (b) output data obtained during (or upon completion of) the running of your programs; (c) any input data or script files used in the running of your programs (Lisp dribble files are very helpful here to capture both input and output); and (d) appropriate written commentary recording your observations and answers to the questions above.