Machine Learning 2003

Assignment 2: Optical character recognition using decision trees.

- Due by: Implementation: 23h59 on 06/03/2003; Experiments and report: 23h59 on 13/03/2003.
- Marking: 50 marks for implementing a working algorithm; 50 marks for application to optical character recognition, experimental evaluation and report (including description of algorithm, presentation of results, discussion and conclusions).
- Remember that discussion and conclusions are the most important part of an assignment!
- Note: this is **not** a team project. Your implementation and experiments should be your own work.

1 Getting acquainted with decision trees

Implement the ID3 algorithm for building decision trees (Mitchell, p. 56). Use Table 3.2 (Mitchell, p. 59) to test your implementation of the ID3 algorithm.

2 Application: Optical character recognition

You will test decision trees for optical character recognition. The following files provide training and test data for discriminating the letter B (encoded as class 0) from the letter R (encoded as class 1). For upper case letters, this is a difficult discrimination problem.

Download the following two files from the course website:

1. br.data is a file of training examples that you will use to construct a decision tree.
2. br.test is a file of test examples that you will use to test the decision tree.

Each of these files has the following format:

number-of-examples number-of-classes number-of-features
class0 x01 x02 ... x0n
class1 x11 x12 ... x1n
class2 x21 x22 ... x2n
...
classm xm1 xm2 ... xmn

Where classi is either 0 (B) or 1 (R), and the values of the features xij are numbers.

Write a program to do the following:

1. Grow a decision tree using the ID3 algorithm shown on page 56.
2. Display a decision tree.
3. Extract rules from the decision tree.
4. Evaluate a decision tree on a data set by finding a learning curve as follows: (1) Construct a series of subsets of br.data containing 40, 80, 160, 320, and 615 examples. You may simply use the first n examples from br.data, (2) construct a decision tree from each individual data set, and (3) test the accuracy of the decision tree on br.test. Report the error rates for each individual training data set.

3 Deliverables

1. Working program (demonstration)
2. Learning curve (in report)
3. For each experiment, display the decision tree (in report)
4. Discussion of results (in report)
5. Conclusions (in report)