

CS 410/510
Machine Learning
Spring, 2007

Homework 2: Decision Trees

Due Tuesday, April 17.

In this homework assignment you will experiment with C4.5, an implementation of ID3 written by Ross Quinlan and available on the Internet.

You will be working with your training examples of spam and non-spam e-mail messages, and the features you extracted from these examples in Homework 1.

For this homework, you need to perform the following steps:

1. Download C4.5: Download C4.5 and documentation from the following web site:

<http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html>

This version of C4.5 works on Linux systems. If you have any trouble getting it to compile and work, let me know.

For this assignment, you will need only the manual pages for “c4.5” and for “verbose c4.5”. For the format of the input files, see one of the sets of example files given on this web page (e.g., “Golf”).

2. Create training and test data sets for your spam data: Use some of the spam and “easy ham” examples from Homework 1 to create TrainingSetEasy and some of the spam and “hard ham” examples to create TrainingSetHard. Also use some of the spam, easy ham and hard ham examples to create a single, separate, independent test set (i.e., no overlapping examples between the training and test sets). The training and test sets can be about the same size. All of these sets should contain both spam and ham examples. (Note that you do not need a third, “validation”, set for this assignment.)

Have your feature-extraction program create these files in the format required

by C4.5. Your files should be named something like *spam.data* and *spam.test*. You also need to create a *spam.names* file that gives the names and types of your attributes (see the examples on the C4.5 web page).

3. Run C4.5 on your data: (a) Run C4.5 on your spam/easy-ham training set, and test the resulting trees on your test set. This can be done with the command:

```
c4.5 -f spam -u
```

Put the following information in your write-up: A table of results like the one we created in class, giving tree size (unpruned and pruned), and training and test error (unpruned and pruned) for the following experiments:

1. Training set = TrainingSetEasy
2. Training set = one-half of TrainingSetEasy (but containing both spam and ham examples)
3. Training set = one-fourth of TrainingSetEasy (but containing both spam and ham examples)
4. Training set = TrainingSetEasy with 5% noise (incorrect classifications)
5. Training set = TrainingSetEasy with 1% noise (incorrect classifications)
6. A sixth experiment that you design

(b) Now run c4.5 using information gain instead of gain ratio:

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

and give the same information as above in your write-up.

4. Hard training set: Repeat step 3(a) using your TrainingSetHard, and the same test set as before.

5. Add an attribute or attributes: What additional attribute(s) do you think would improve the performance of C4.5 on the test data? Add it (or

them) to your set of attributes, and repeat step 3(a).1 (entire training set, no noise) for both easy and hard training sets.

6. Summarize your results: Write a summary of your results from the steps above. Here are some questions to answer in your summaries.

- Which attributes seemed to be the most useful for classifying spam/ham? Why were these the most important?
- Were any of your attributes not used in the pruned trees? If so, why do you think they were omitted?
- What are the effects of training set size, noise, and “easy” or “hard” ham on the accuracy of decision trees? Explain why you think these effects occurred.
- What are the effects of using information gain versus gain ratio in the various experiments you did? Why do you think you got these effects?
-
- What new attribute(s) did you include and why? Did it (they) have the effect you expected? If not, any ideas why not?
- Include anything else that you found noteworthy in your results.

Grading

You will be graded on the completeness and clarity with which you report and explain your results on the steps listed above.