CSCI/ARTI 8950 Machine Learning

Assignment Number 2: Due Thursday 2/7/2008 **(in class)**

For this assignment you will need to use a decision tree learning package. I would recommend the See5/C5.0 package that you can download from **http://www.rulequest.com/see5-info.html**. See5 is for Windows and C5.0 is for Unix machines. They are the successors of C4.5 which in turn is the successor of ID3. RuleQuest is Ruth Quinlan's (author of ID3 and C4.5) company and it offers free trial versions of the codes which can handle datasets of no more than 400 instances. You can also use the decision tree Lisp code package available in the text book's web pages at **http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/decision-trees.html** or the Weka package downloadable from

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

.

1. [**30 points**] For this part you will experiment with the *PlayTennis* data from Table 3.2 in the textbook.

   (a) Use your decision tree learner to learn a decision tree based on all 14 instances. Do you get the same decision tree as the one in Figure 3.1 in the text book? If not, find out why (it may be because the learner is using an information gain measure that is different from the one the book uses).

   (b) Use the leave-one-out cross-validation method described on page 235 in the text book to estimate the error of the decision tree. You can do this manually by removing one example at a time and training on the remaining 13 and then testing on the one you removed, or you can use the package to do this for you automatically by asking for a 14-fold cross-validation if the package supports cross-validation. The See5/C5.0 package supports cross-validation.

2. [**50 points**] For this part you should use a data set with at least **100** instances. You should choose one of the data sets in the UCI repository at

   `http://archive.ics.uci.edu/ml/`

   . You may use any data set from this web page provided that you tell me which one you used! Many of the data sets in the repository have missing attribute values but See5/C5.0 can handle this automatically. For larger datasets, you may pick a subset of instances and use it instead of the full dataset but please try to use as many instances of the set as your software would allow (400 instances if you use See5/C5.0).

   (a) Use your full dataset to learn a decision tree. Give the tree, the error on the training set and the time needed for learning (if your package gives the learning time). **Do not use pruning**

(b) Use 10-fold cross-validation to estimate the error (again without pruning).

(c) Repeat the learning on the full set and the 10-fold cross-validation with pruning allowed. You may use any pruning method you like but you should describe it. Compare in a table between the error on the training set and the 10-fold cross-validation with and without pruning. Comment on the time needed for learning with and without pruning (if your package gives the learning time)