Program 4: Decision Tree Induction
CSC 4240
Due Date: Dec. 5, 2003

You are to write a decision tree program based on the ID3 algorithm described in class and in the book (section 18.3 – page 658). You can assume that there are no missing values and that all of the attributes are nominal (not continuous or ordered in any way). You do not need to implement pruning of the decision tree, but if you do, you can receive up to 10 bonus points on the program for it.

Your program should do the following:

1. Read in a file containing the training data (described below).
2. Use the training data to construct a decision tree.
3. Read in a file containing the test data (described below).
4. Classify each instance using the decision tree and determine whether or not the classification was correct.
5. Output the total number of correct and incorrect classifications the occurred during testing.

For each data domain, you should perform 10-fold cross-validation. I have already partitioned the data into appropriate training/testing pairs. For each domain, run one experiment for each pair. Then add up the results from the 10 experiments and compute a final error rate (number of incorrect divided by total number).

Turn in the code and a short description of the results. For each data domain, tell me the total correct and incorrect for each of the 10 experiments and report the final error rate.

If you implement pruning, include results for both the pruned tree and the unpruned tree.

Data files:

Each file contains one data item per row. Values are comma delimited, the last column is the attribute to be predicted.

soybean data domain
files for test *n*:
1. soy_train_*n*.txt
2. soy_test_*n*.txt

cars data domain
files for test *n*:
1. cars_train_*n*.txt
2. cars_test_*n*.txt