

**CSE 5800 Mining/Learning and the Internet HW1**  
**Due 6:30pm, Sep 9, 2009**  
**Submit Server: course=cse5800 , project=hw1**

Programming assignment:

1. Decision Tree Learning (DTL) algorithm in the handout
2. Allow more than binary outcomes and continuous-valued attributes
3. Allow the option of pruning to handle noise (handout)
4. Three data sets:
  - (a) Restaurant in the handout and on the course web site
  - (b) Intrusion detection on the course web site
  - (c) your own data set [for example, <http://www.ics.uci.edu/~mlearn/MLRepository.html> ]
5. Separate the data set into a training set and a test set, report the accuracy on the two disjoint sets (with and without pre-pruning).
6. For the second data set, corrupt the class labels of training examples from 0% to 50% (5% increment), by changing from the correct class to another class. Compare the accuracy of the tree with and without pre-pruning on uncorrupted test data.
7. Implementation:
  - (a) Use one of these programming languages: C, C++, Java, or LISP.
  - (b) input files: attributes description, training data, test data
  - (c) You would have two (maybe three) executables:
    - i. Miner/Learner: input training examples/instances, output a tree
    - ii. Classifier/predictor: input a tree and labeled instances, output the classifications/predictions and how accurate the tree is with respect to the correct labels (% of correct classifications).
    - iii. Tree printer: if the output from the learner is human-readable, no need for a tree printer; otherwise, build a tree printer so that we can see the built tree.