

## CSE/CBS 572 Project: Decision Tree

Due Date: March 4, 2008 (midnight)

This project focuses on the decision tree algorithm. We will use the tic-tac-toe data set from UCI in this project (<http://archive.ics.uci.edu/ml/machine-learning-databases/tic-tac-toe/>). There are 958 instances in the data set with 9 features and two classes. Randomly partition the data into a training set (use 75% of the data from each of these two classes) and a test set (the rest 25% of the data from each of these two classes).

Task 1. (70 points) Implement a decision tree classification algorithm. For attribute test condition, you can use either binary splitting or multi-branch splitting (i.e., as many branches as the number of values of the splitting attribute). To decide the best splitting attribute, Information Gain will be used as measure of node impurity. Use the provided training data to build the fully-grown tree (without pruning). Then apply the tree on the provided test data to determine the class labels and the generalization error rate on the test data.

The following should be contained in the output of your program:

- 1) The optimistic error rate on the training data.
- 2) The generalization error rate on the test data.

Task 2. (30 points) Prune the fully-grown decision tree from Task 1, and then make predictions of the class labels for the test data and determine the new generalization error rate. You use post-pruning and choose generalization error on the validation set as pruning criteria to prune the fully-grown tree.

Deliverables: an electronic copy of the program source code, a README file and a short report including the sample outputs for both tasks **should be compressed in one single file** (The names of both project members should appear in the file) and submitted to the Blackboard (through digital drop box). Program source code should be well commented. README file should clearly describe how to compile the source and run the executables.