4100/5100 Assignment 4: Decision Trees and Overfitting
Due June 7 9PM

In this assignment, we'll experiment with decision trees and observe the phenomenon of overfitting. *We won't use HackerRank this time* - submit your code directly to Blackboard, along with a PDF containing the answers to the questions in the final step.

1) Download the DecisionTree.java code, filling out the DecisionTree constructor, its classify method, and any other methods as necessary. Do not prune at this stage; only create a leaf when all examples agree.

2) You can now check that running on test1.txt and test2.txt produce the outputs test1.out and test2.out, respectively. You may want to produce other tests and verify that they work the way you expect.

3) Run your code on the adult.data.csv dataset. This dataset (adapted from the "Adult" dataset of the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/adult) contains a variety of features for people measured in the 1994 U.S. Census, predicting whether the person in question makes more than $50,000 per year. Make a note of the precision, recall, and accuracy on the training set and the test set for step 7.

4) Now add a flag that, if set, forces the decision tree learning to use chi-square pruning on each node as the tree is built, creating a leaf instead of a decision when the best feature's relationship with the target is not significantly different from chance.

5) You can verify that test1.txt and test2.txt continue to produce the expected results, while testSig.txt produces a single leaf, as shown in testSig.out.

6) Run the pruned decision tree code on adult.data.csv. Make a note of the training and test precision and recall.

6) In a PDF submitted on Blackboard along with your code, answer the following:

a) Report the precision, recall, and accuracy for training and test data for both the unpruned and pruned decision trees.

b) Explain what we'd expect of the precision and recall of the pruned versus unpruned versions, and the training versus test. What should be better than what, and why? Do the experimental results match our expectations?

c) Take a look at the trees constructed for the pruned decision tree on the adult.data.csv dataset. Are there any decisions in the tree that you don't trust are real predictors of high income? If so, what?

d) On a different dataset that I decided not to use for this assignment -- predicting whether people would default on loans, an infrequent and hard-to-predict occurrence -- the solution classifier got 90% accuracy but 0% precision and recall. How is that possible?

Decision trees by themselves typically aren't the best performing machine learning algorithm, but they are the most easily interpreted. Later, we'll cover how we can improve their performance while sacrificing their readability with *random forests,* one of the most popular machine learning methods among biologists.