CS4811: Homework 4 --- Learning Decision Trees

Due: Wednesday, March 23, 2016, 9:00pm.
(Assigned: Wednesday, March 2, 2016.)

Reminder: This is an individual assignment. All the work should be the author's and in accordance with the university's academic integrity policies. You are allowed to use any written source in preparing your answers, but if you use any other source than the textbook and the class notes, you should specify it on your assignment.

Problem:

In this assignment, you will implement the basic algorithm for learning decision trees.

You should implement your own code from scratch. You may consult existing implementations but you may not build on them. The textbook's web site has implementations in a range of languages, if you'd like to take a look.

Task:

Implement a decision tree learner that can use the two heuristics we discussed in class:

- probability of error
- entropy

Run your program with the following dataset with each heuristic.

http://archive.ics.uci.edu/ml/datasets/Mushroom

It is fine for your program to work on the mushroom dataset only. In other words, you are not required to write a general purpose decision tree program. You may hardcode that the dataset has 22 attributes, and also the possible values for each of the attributes. However, in addition to the original dataset, we will test your program with a subset of it. Therefore, don't hardcode the number of examples and make it so that they are read from a file.

For each run, the program should print a trace of how the decision tree is generated. For each node of the decision tree, the trace should include information about each attribute tested and its heuristic value, and the attribute that was chosen. This can be similar to the slides, but no graphics is needed.

When the final decision tree is generated, print the tree in a way that is readable and is convenient to you. For example, you can traverse the tree in a depth first fashion and print it as text, indenting each level by a few characters. Printing must be part of the program because we will test your code with different datasets.

According to the description file the dataset has missing values for attribute 11, the stalk root. Simply replace the "?" with an "m" for missing, and use is as the value.

Write a short report that describes your implementation, and presents a discussion of the results:

- Did you encounter any problems while preparing the data set for processing, or while processing it?
- Are the trees generated with the two heuristics different?
- Are the trees different than what you expected initially?
- Provide a drawing of the trees (in the report only)

In your submission include a file named README that contains full instructions on how to execute your code. We should be able to test it with another set of examples from the mushroom domain. In your documentation explain how we can feed another data file to your program, and what kind of output to expect.

You are free to use any language that runs on the CS department Linux platforms.

Submit the following on Canvas in a single (g)zip or (g)tar file. Hardcopies are not needed.

- The fully commented code
- The README file
- The program output for each of the two heuristics
- The report in its original format and in pdf format.