CSE 5523 Homework 2: Decision Trees

Alan Ritter

In this assignment you will implement the ID3 decision tree learning algorithm and apply it to a dataset of poisonous and edible mushrooms[1].

Data

The data is provided in a CSV file format:

```
e,x,s,y,t,l,f,c,b,g,e,c,s,s,w,w,p,w,o,p,k,n,g
e,f,s,n,f,n,a,c,b,o,e,?,s,s,o,o,p,n,o,p,b,v,l
p,k,s,e,f,f,f,c,n,b,t,?,k,k,p,p,p,w,o,e,w,v,d
e,f,f,g,f,n,f,w,b,k,t,e,s,f,w,w,p,w,o,e,k,s,g
e,x,f,n,t,n,f,c,b,w,t,b,s,s,p,w,p,w,o,p,n,v,d
e,f,y,n,t,l,f,c,b,w,e,r,s,y,w,w,p,w,o,p,k,s,p
p,x,y,g,f,f,f,c,b,h,e,b,k,k,p,n,p,w,o,l,h,v,g
p,f,s,w,t,n,f,c,b,w,e,b,s,s,w,w,p,w,t,p,r,v,m
e,x,f,g,t,n,f,c,b,w,t,b,s,s,w,w,p,w,o,p,n,y,d
...
```

Each row corresponds to a mushroom. The first column is the label indicating whether the mushroom is edible (e) or poisonous (p). (This is the output that we wish to predict from the other columns). Information about the meaning of the other columns is listed in the file `agaricus-lepiota.names`. Starter code is provided to read in this file in `Data.py`.

Starter Code

First, try running the provided starter code from the command line. Your output should look something like the following.

[1] `https://archive.ics.uci.edu/ml/datasets/Mushroom`

```
bash-3.2$ python id3.py train test 1
>FeatureVal(feature=1, value='x')       (100, 0)        0
0.513863216266
```

The provided code simply "learns" a decision tree consisting of a root node with no children that always predicts "Yes". The program prints out this tree (consisting of a single node with no children), and then prints its evaluated accuracy on the test set which is about 0.51. The command line arguments are: the training, test file and a threshold on the minimum information gain (this is used to determine when to stop growing the tree).

Your job is to edit the file `id3.py` to implement the functions `InformationGain()` and `ID3()` to grow a decision tree from the data.

What to Turn In

Please turn in the following to the dropbox on Carmen:

1. Your code
2. A brief writeup (text file format) that includes the output of your program with the minimum information gain set to the following values: 1, 0.1, 0 (grows the full tree).