Markiyan Varhola

9016820171

Decision Tree and ID3 Implementation

The main goal of the program was to create an algorithm involving decision trees that could classify each item from a testing set using what is learned from a learning set. The program is written in such a way that it could be applicable to any scenario, and can be used with multiple data sets. Although this uses a baseline ID3 heuristic, I found the accuracy to be satisfactory for both data sets (100% accuracy on computer purchase set, and 91.3% accuracy on vehicle safety). The program is split into three parts: dt.py, dtree.py, and id3.py. The dt.py file is the main program, which uses helper functions in dtree.py and id3.py files. The id3.py file contains helper functions that make up the ID3 algorithm (calculating entropy and gain). In summary, the algorithm works in the following way:

1. Calculate the entropy of every attribute using the data set S
2. Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes.

The dtree.py file contains functions to generate a decision tree which will use the ID3 algorithm output as the main way to classify data. The main program, dt.py, parses the input file, uses the other libraries to generate a decision tree, and then generates classifies each item in the testing file and creates an output file in the correct format. The output file can then be tested using the supplied testing program.

Instructions on running script:

The script was tested on Python 2.7.3, and running it on Python 3> will result in an error. In order to run the script, use the format as outlined in the original assignment document.

To run the script, use the following format:

```
python dt.py {input training file} {input testing file} {output file}
```

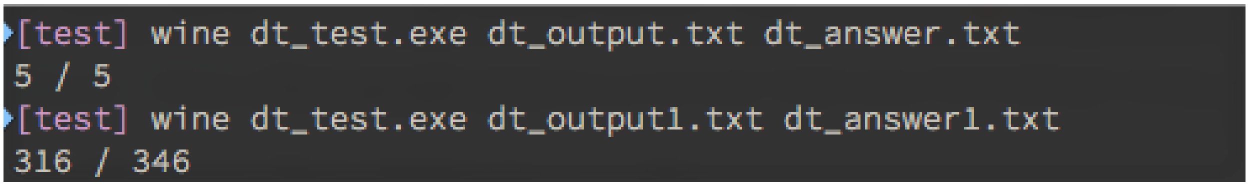
For example:

```
python dt.py data/dt_train.txt data/dt_test.txt test/dt_output.txt
```

This command will produce a file called dt_output.txt in the ./test directory with the properly formatted data. Use output file as an input for the testing program to view the results.

Current script output:

**The first test passes with a 5/5 (100% accuracy), and the second test passes with a 316/346 (~91.3% accuracy).



```
[test] wine dt_test.exe dt_output.txt dt_answer.txt
5 / 5
[test] wine dt_test.exe dt_output1.txt dt_answer1.txt
316 / 346
```