Cs: 365
Group Members: Maniz Shrestha, Nirdesh Bhandari
Lab-C Decision Trees

Final Report

For this lab we have implemented the DECISION-TREE-LEARNING algorithm as seen in chapter 18. Our program reads in a tab-delimited dataset and outputs the decision tree in the screen along with the training set accuracy and the cross validation accuracy. Our program uses numpy arrays to pass outcomes and attributes along to functions. We decided this was the best way to maintain consistency and make best use of some very helpful numpy libraries for counting unique values, transposing lists, partitioning subsets and frequency calculations.

First, we generated the trees using all the examples and tested the accuracy of our tree for each of the examples. The accuracy score we received for this were as follows:

Pets.txt - 0.86 (86%)	**no of nodes: 43**
Tennis.txt - 1.0 (100%)	**no of nodes: 12**
Titanic2.txt – 0.69 (69%)	**no of nodes: 36**

Then we used **Leave-one-out cross-validation** method to check the accuracy of our decision trees by testing it against n examples for each dataset. For this we picked out a single test example whose attributes were recorded (incase they weren't observed in other examples) and the tree was generated using the rest of the examples. The test case was then used to check if the decision tree could accurately predict the outcome. The accuracy scores we received were as follows:
Pets.txt - 0.47 (47%)
Tennis.txt - 0.79 (79%)
Titanic2.txt – 0.69 (69%)

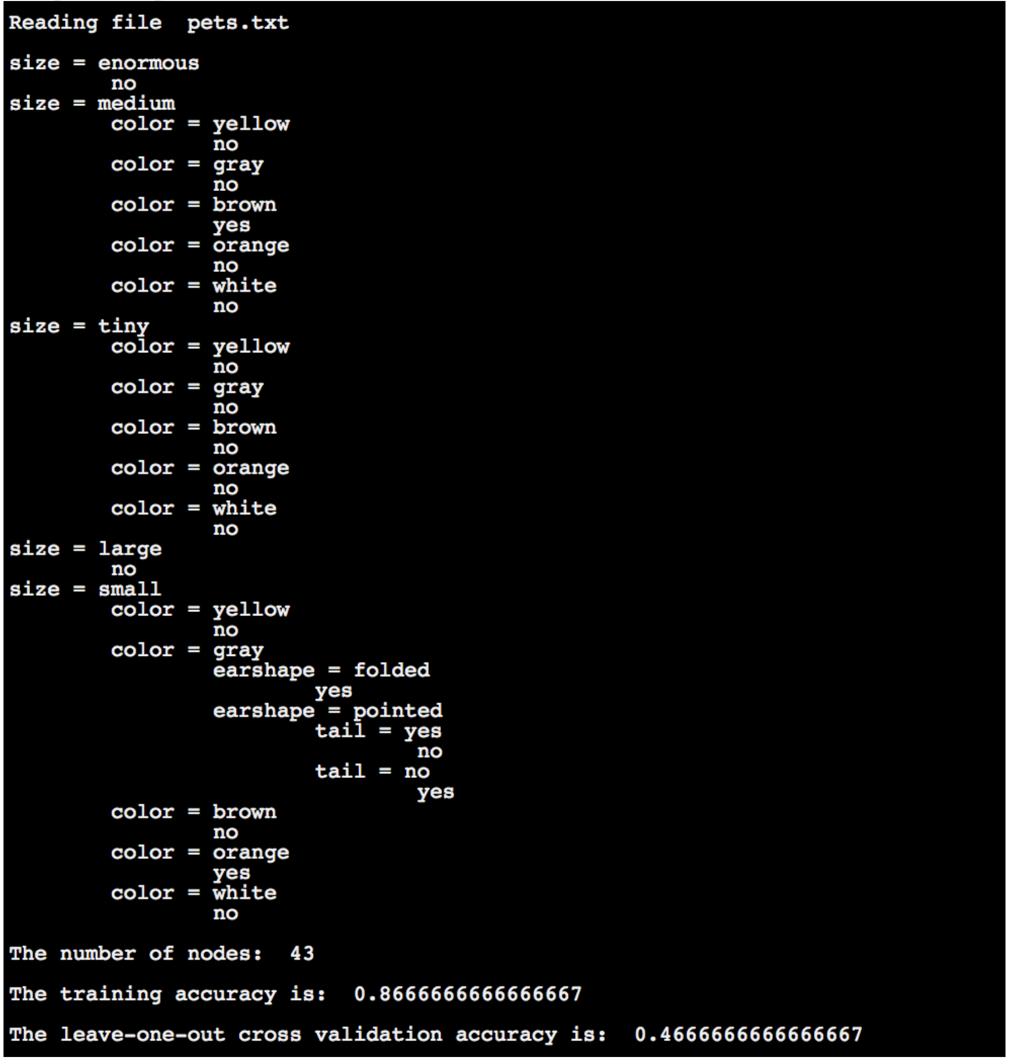
Discussion

Our accuracy measures were higher, especially for pets and tennis, when were using a tiebreaker to deal with equal probabilities. However, since the assignment specifically asked that ties be assigned a 'no', our scores dropped for pets and tennis but went up titanic2, probably because it was composed of twice as many 'no' compared to 'yes'.

References:

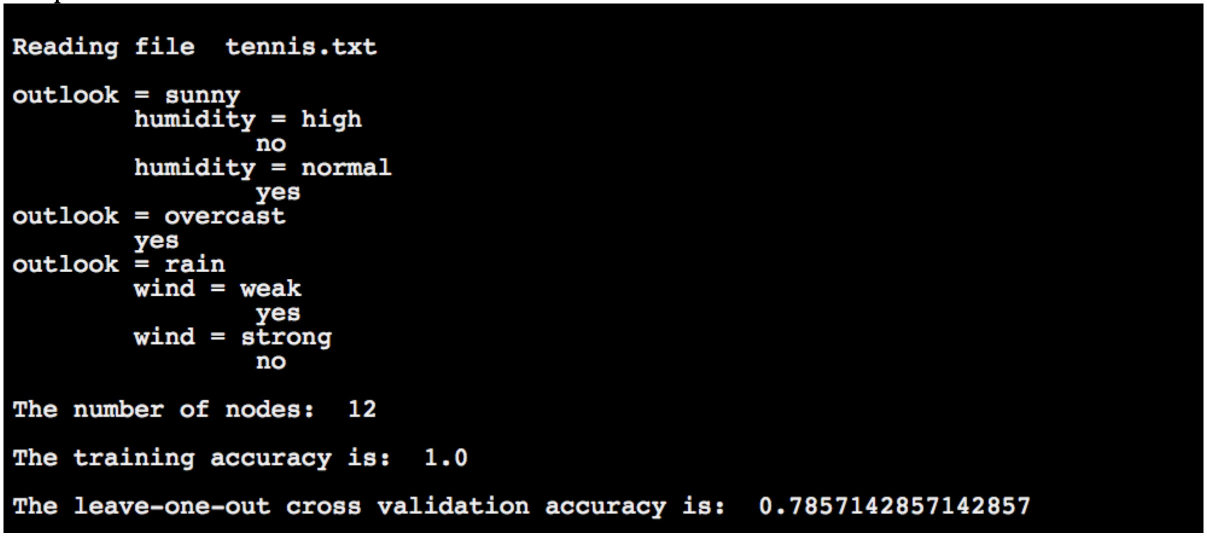
For pretty printing: https://stackoverflow.com/questions/3229419/how-to-pretty-print-nested-dictionaries
http://gabrielelanaro.github.io/blog/2016/03/03/decision-trees.html

Output for pets.txt



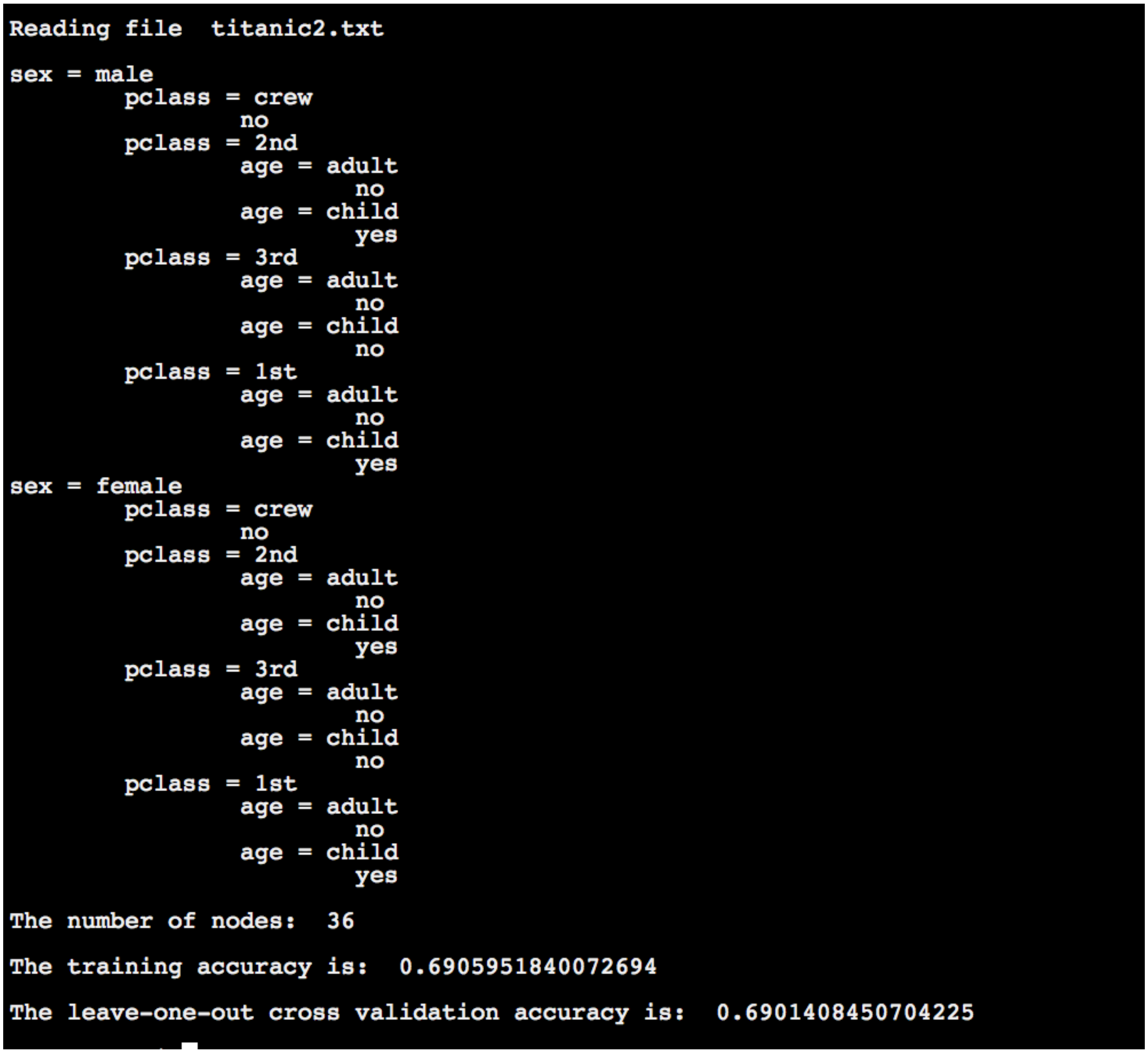
```
Reading file  pets.txt

size = enormous
        no
size = medium
        color = yellow
                no
        color = gray
                no
        color = brown
                yes
        color = orange
                no
        color = white
                no
size = tiny
        color = yellow
                no
        color = gray
                no
        color = brown
                no
        color = orange
                no
        color = white
                no
size = large
        no
size = small
        color = yellow
                no
        color = gray
                earshape = folded
                        yes
                earshape = pointed
                        tail = yes
                                no
                        tail = no
                                yes
        color = brown
                no
        color = orange
                yes
        color = white
                no

The number of nodes:  43

The training accuracy is:  0.8666666666666667

The leave-one-out cross validation accuracy is:  0.4666666666666667
```

Output for tennis.txt:



```
Reading file  tennis.txt

outlook = sunny
        humidity = high
                no
        humidity = normal
                yes
outlook = overcast
        yes
outlook = rain
        wind = weak
                yes
        wind = strong
                no

The number of nodes:  12

The training accuracy is:  1.0

The leave-one-out cross validation accuracy is:  0.7857142857142857
```

Output for titanic.txt :



```
Reading file  titanic2.txt

sex = male
        pclass = crew
                no
        pclass = 2nd
                age = adult
                        no
                age = child
                        yes
        pclass = 3rd
                age = adult
                        no
                age = child
                        no
        pclass = 1st
                age = adult
                        no
                age = child
                        yes
sex = female
        pclass = crew
                no
        pclass = 2nd
                age = adult
                        no
                age = child
                        yes
        pclass = 3rd
                age = adult
                        no
                age = child
                        no
        pclass = 1st
                age = adult
                        no
                age = child
                        yes

The number of nodes:  36

The training accuracy is:  0.6905951840072694

The leave-one-out cross validation accuracy is:  0.6901408450704225
```