Vibhor Jain
3/17/13
CSE 176
Prof. David Noelle
TA Jimei Yang

Programming Assignment #3

Program Description:

The program was run in MATLAB. The user runs the main program ‘dtl_driver.m’ by typing ‘dtl_driver’ in the console window. This program makes a decision tree on a given dataset by executing the ID3 algorithm, which is called in ‘ID3_sample.m.’ In order to successfully execute the ID3 algorithm, the algorithm calculates the entropy and information gain of a given subset of the dataset (as seen in ‘entropy_func.m’). In the end, the program displays the learned decision tree through the function ‘print_dtree.m.’

The main program takes in a configuration file, a training set data file, and a test set data file and outputs the results of the algorithm in a results text file. Before running the program, the user must make sure that ‘dtl_driver.m’, ‘entropy_func.m’, ‘ID3_sample.m’, ‘print_dtree.m’, the configuration file, the training data file, and the test data file are in the same directory.

A trace is given on an example case:

```
>> dtl_driver
Enter configuration file name (in this directory)
restaurant-dtl.cfg
>>
```

The results as produced in the log file are given below:

```
THE FOLLOWING DECISION TREE WAS LEARNED FROM EXAMPLES:

TEST FEATURE NUMBER 4:
   IF FEATURE IS FALSE, THEN ...
      TEST FEATURE NUMBER 2:
         IF FEATURE IS FALSE, THEN ...
            ITEM IS NON-TARGET
         IF FEATURE IS TRUE, THEN ...
            TEST FEATURE NUMBER 9:
               IF FEATURE IS FALSE, THEN ...
                  ITEM IS TARGET
               IF FEATURE IS TRUE, THEN ...
                  ITEM IS NON-TARGET
   IF FEATURE IS TRUE, THEN ...
      TEST FEATURE NUMBER 6:
         IF FEATURE IS FALSE, THEN ...
            ITEM IS TARGET
         IF FEATURE IS TRUE, THEN ...
            TEST FEATURE NUMBER 3:
               IF FEATURE IS FALSE, THEN ...
```

```
                  ITEM IS NON-TARGET
               IF FEATURE IS TRUE, THEN ...
                  TEST FEATURE NUMBER 7:
                     IF FEATURE IS FALSE, THEN ...
                        ITEM IS NON-TARGET
                     IF FEATURE IS TRUE, THEN ...
                        ITEM IS TARGET
```

```
TESTING SET PERFORMANCE:

1.000000     0.000000    0.000000    1.000000    0.000000    0.000000
0.000000    1.000000    0.000000    1.000000    1.000000    0.000000
0.000000    0.000000    0.000000    0.000000    1.000000    1.000000
0.000000
1.000000     0.000000    0.000000    1.000000    0.000000    1.000000
1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
1.000000    0.000000    1.000000    0.000000    0.000000    0.000000
0.000000
0.000000     1.000000    0.000000    0.000000    0.000000    0.000000
1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
0.000000    0.000000    0.000000    0.000000    1.000000    1.000000
0.000000
1.000000     0.000000    1.000000    1.000000    0.000000    1.000000
1.000000    0.000000    1.000000    0.000000    0.000000    0.000000
1.000000    1.000000    0.000000    0.000000    1.000000    1.000000
0.000000
1.000000     0.000000    1.000000    0.000000    0.000000    1.000000
0.000000    1.000000    0.000000    1.000000    1.000000    0.000000
0.000000    0.000000    0.000000    1.000000    0.000000    0.000000
0.000000
0.000000     1.000000    0.000000    1.000000    0.000000    0.000000
0.000000    0.000000    1.000000    1.000000    0.000000    1.000000
0.000000    0.000000    0.000000    0.000000    1.000000    1.000000
0.000000
0.000000     1.000000    0.000000    0.000000    1.000000    0.000000
1.000000    0.000000    1.000000    0.000000    0.000000    0.000000
0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
0.000000
0.000000     0.000000    0.000000    1.000000    0.000000    0.000000
0.000000    0.000000    1.000000    1.000000    0.000000    0.000000
1.000000    0.000000    0.000000    0.000000    1.000000    1.000000
0.000000
0.000000     1.000000    1.000000    0.000000    0.000000    1.000000
1.000000    0.000000    1.000000    0.000000    0.000000    0.000000
0.000000    0.000000    0.000000    1.000000    0.000000    0.000000
0.000000
1.000000     1.000000    1.000000    1.000000    0.000000    1.000000
0.000000    1.000000    0.000000    1.000000    0.000000    1.000000
0.000000    1.000000    0.000000    0.000000    0.000000    0.000000
0.000000
0.000000     0.000000    0.000000    0.000000    1.000000    0.000000
1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
0.000000
```

```
1.000000     1.000000    1.000000    1.000000    0.000000    1.000000
1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
0.000000    0.000000    1.000000    0.000000    1.000000    1.000000
0.000000
```

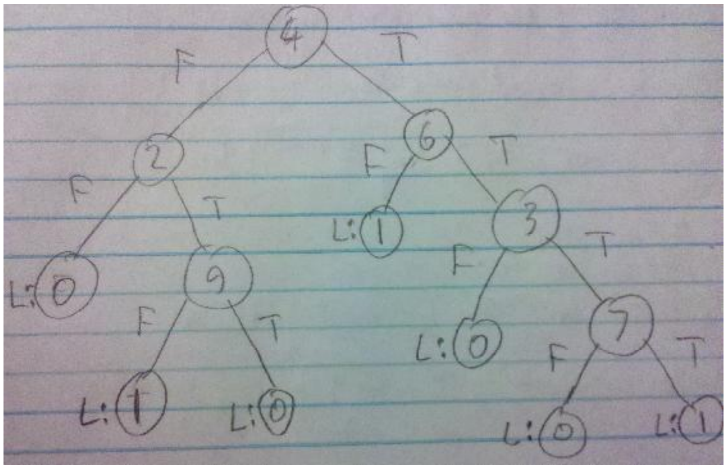
Another trace is given on a test case:

```
>> dtl_driver
Enter configuration file name (in this directory)
lenses-dtl.cfg
>>
```

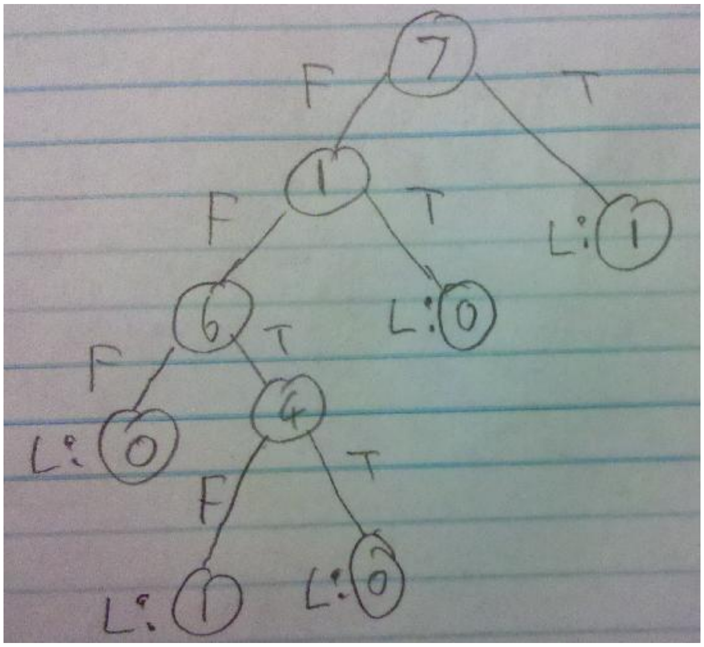
The results produced in the log file are also given:

```
THE FOLLOWING DECISION TREE WAS LEARNED FROM EXAMPLES:

TEST FEATURE NUMBER 7:
   IF FEATURE IS FALSE, THEN ...
      TEST FEATURE NUMBER 1:
         IF FEATURE IS FALSE, THEN ...
            TEST FEATURE NUMBER 6:
               IF FEATURE IS FALSE, THEN ...
                  ITEM IS NON-TARGET
               IF FEATURE IS TRUE, THEN ...
                  TEST FEATURE NUMBER 4:
                     IF FEATURE IS FALSE, THEN ...
                        ITEM IS TARGET
                     IF FEATURE IS TRUE, THEN ...
                        ITEM IS NON-TARGET
         IF FEATURE IS TRUE, THEN ...
            ITEM IS NON-TARGET
   IF FEATURE IS TRUE, THEN ...
      ITEM IS TARGET



TESTING SET PERFORMANCE:

0.000000     0.000000    1.000000    1.000000    0.000000    0.000000
0.000000    0.000000    1.000000    1.000000
0.000000     0.000000    1.000000    1.000000    0.000000    0.000000
1.000000    1.000000    1.000000    0.000000
0.000000     0.000000    1.000000    1.000000    0.000000    1.000000
1.000000    1.000000    1.000000    0.000000
0.000000     1.000000    0.000000    0.000000    1.000000    1.000000
1.000000    1.000000    1.000000    0.000000
0.000000     1.000000    0.000000    1.000000    0.000000    0.000000
0.000000    0.000000    0.000000    0.000000
0.000000     1.000000    0.000000    1.000000    0.000000    1.000000
0.000000    0.000000    0.000000    0.000000
1.000000     0.000000    0.000000    0.000000    1.000000    0.000000
1.000000    1.000000    1.000000    0.000000
1.000000     0.000000    0.000000    1.000000    0.000000    0.000000
0.000000    0.000000    0.000000    0.000000
```

The tree produced from the restaurant data files is given below:

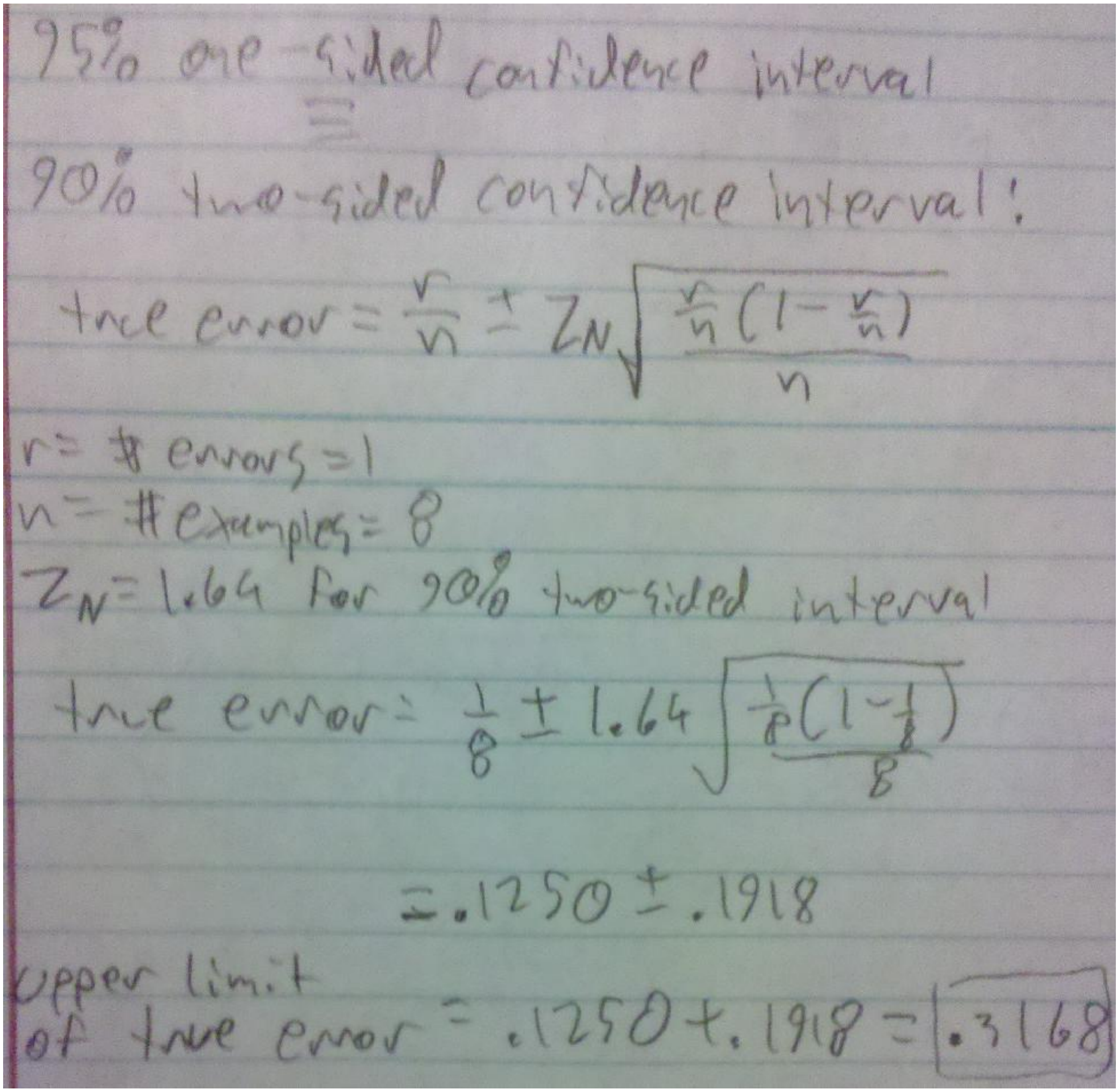


The tree produced from the lenses data files is given below:



The nodes labeled with numbers reflect the attribute number that is being tested. The nodes with the expression 'L:' next to them reflect the labels at the end of the trees. The number 0 is associated with 'ITEM IS NON-TARGET' and the number 1 is associated with 'ITEM IS TARGET.' The 'T' and 'F' links reflect whether the given feature of the attribute being tested is true or false.

The calculation of the error of the learned decision tree of the lenses data set is shown below:



95% one-sided confidence interval

=

90% two-sided confidence interval:

$$\text{true error} = \frac{r}{n} \pm Z_N \sqrt{\frac{\frac{r}{n}\left(1-\frac{r}{n}\right)}{n}}$$

r = # errors = 1

n = # examples = 8

Z_N = 1.64 for 90% two-sided interval

$$\text{true error} = \frac{1}{8} \pm 1.64 \sqrt{\frac{\frac{1}{8}\left(1-\frac{1}{8}\right)}{8}}$$

$$= .1250 \pm .1918$$

upper limit of true error = $.1250 + .1918 = \boxed{.3168}$

The upper limit (.3168) would reflect the worst-case scenario where the algorithm would produce the maximum amount of error. The variation in the error (.1918) is greater than the initial error (.1250). This reflects the fact that because we ran the decision tree over so few examples, there would be more uncertainty in the calculated error over the few test examples. This means that we can only be highly certain of a *large* margin of error in our calculated error of .1250. Running the algorithm over more test examples would give a smaller variation in the error, and thus we would be able to calculate a more precise error of the algorithm with a *smaller* margin of error.

What was Learned:

The decision tree serves as another way for achieving dimensionality reduction. Depending on the training dataset, the tree that is learned would be capable of filtering through attributes of the dataset that do not have much effect on the classification of the data. This was seen when the algorithm was applied to the restaurant data set (and the lenses data set). Even though the data came with 16 attributes (7 attributes for the lenses dataset), the algorithm was able to identify the 6 relevant attributes for the restaurant dataset and 4 relevant attributes for the lenses dataset (see the trees above). The downside is that the learned tree would be sensitive to noise in the data.

Acknowledgements:

The ID3 algorithm as wells as the calculation procedure for the ID3 algorithm's error was taken from Tom Mitchell's *Machine Learning*. Parts of codes from Programming Assignments one and two were also used for this code.