Exploring Machine Learning:
The ID3 algorithm
User Manual

Mohammed Ibrahim

Computer Science Department

College of Science, Swansea University

Swansea, SA2 8PP, UK

May 25, 2012

Contents

Abstract

This User Manual details all of the features of the two programs with the screen shots, and briefly explained how to run both the programs.

1 Installaion

To run both the programs user must download and install Java. The user can be able to run the program using any number of tools like example, Eclipse, Netbeans, Textpad etc.
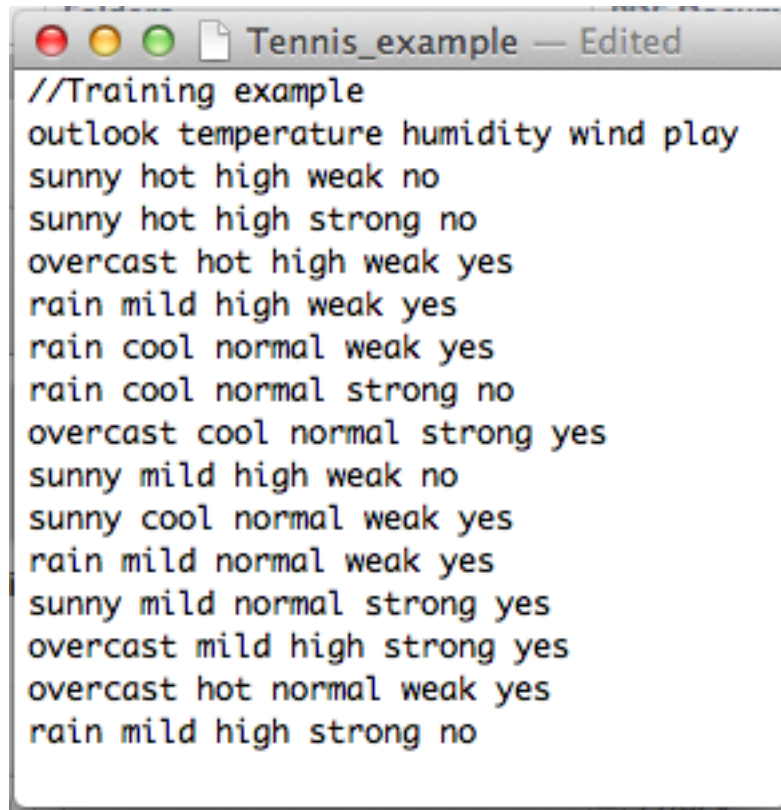
2 Operation

2.1 Training set

As to check the program how it works, user must need have a data sets. The data sets examples have to follow our specified file format. Below I have given the specified input file format. By having input file format we eliminate the requirement that the program should sort the input attributes in ordered to be processed. Any input file should follow the following input file format otherwise the program generates the error to the user that the file format is invalid and can't be processed.

The input file format rules are:

1. Comments are ignored in the file by using double slash (//) symbol.

2. Blank line also be ignored in the data file.

3. The attributes and values are separated by using single space.

4. Every file should have end of line.

5. The very first recognizable line of the data file have to be an attributes.

6. The number of attributes is inferred from the number of words in this line.

7. The last word of the attributes is taken as the $target$ attribute.

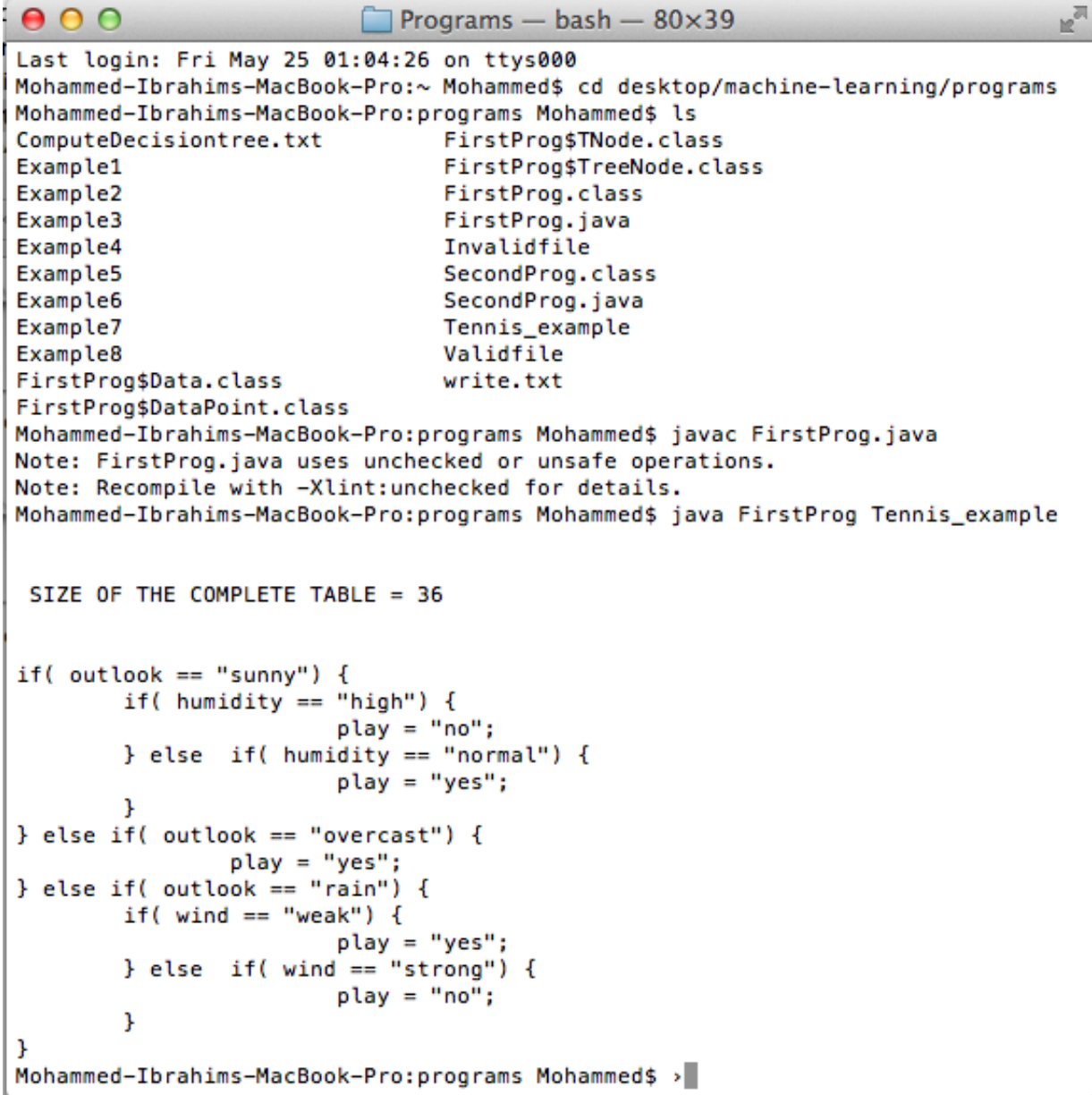8. In the data file each subsequent line contains the values of attributes for a data point.

//Training example
outlook temperature humidity wind play
sunny hot high weak no
sunny hot high strong no
overcast hot high weak yes
rain mild high weak yes
rain cool normal weak yes
rain cool normal strong no
overcast cool normal strong yes
sunny mild high weak no
sunny cool normal weak yes
rain mild normal weak yes
sunny mild normal strong yes
overcast mild high strong yes
overcast hot normal weak yes
rain mild high strong no

Figure 1: Valid training set

The above figure shows the valid training set file that includes 4 attributes and a target attributes. There are 14 possible values corresponding to each attributes.

2.2 Compile and Run First Program

When the data sets are ready, the user can be able input the data in first program. Following steps have to be followed to compile and run the programs. To Compile in any terminal **javac firstProg.java** To Run the program in terminal **java firstProg Tennis_example**

```
Last login: Fri May 25 01:04:26 on ttys000
Mohammed-Ibrahims-MacBook-Pro:~ Mohammed$ cd desktop/machine-learning/programs
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$ ls
ComputeDecisiontree.txt          FirstProg$TNode.class
Example1                         FirstProg$TreeNode.class
Example2                         FirstProg.class
Example3                         FirstProg.java
Example4                         Invalidfile
Example5                         SecondProg.class
Example6                         SecondProg.java
Example7                         Tennis_example
Example8                         Validfile
FirstProg$Data.class             write.txt
FirstProg$DataPoint.class
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$ javac FirstProg.java
Note: FirstProg.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$ java FirstProg Tennis_example


 SIZE OF THE COMPLETE TABLE = 36


if( outlook == "sunny") {
        if( humidity == "high") {
                        play = "no";
        } else  if( humidity == "normal") {
                        play = "yes";
        }
} else if( outlook == "overcast") {
                play = "yes";
} else if( outlook == "rain") {
        if( wind == "weak") {
                        play = "yes";
        } else  if( wind == "strong") {
                        play = "no";
        }
}
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$ ▸
```

Figure 2: Decision Tree for the above training set

The above figure shows the decision tree in the form of if,else clause for the above tennis example. It also shows the size of the complete table.

2.3 Compile and Run Second Program

Second program checks the produced decision tree with the help of validation sets.
Validation set does not include the target attribute.
User can be able to add more possible values to check the validation set.
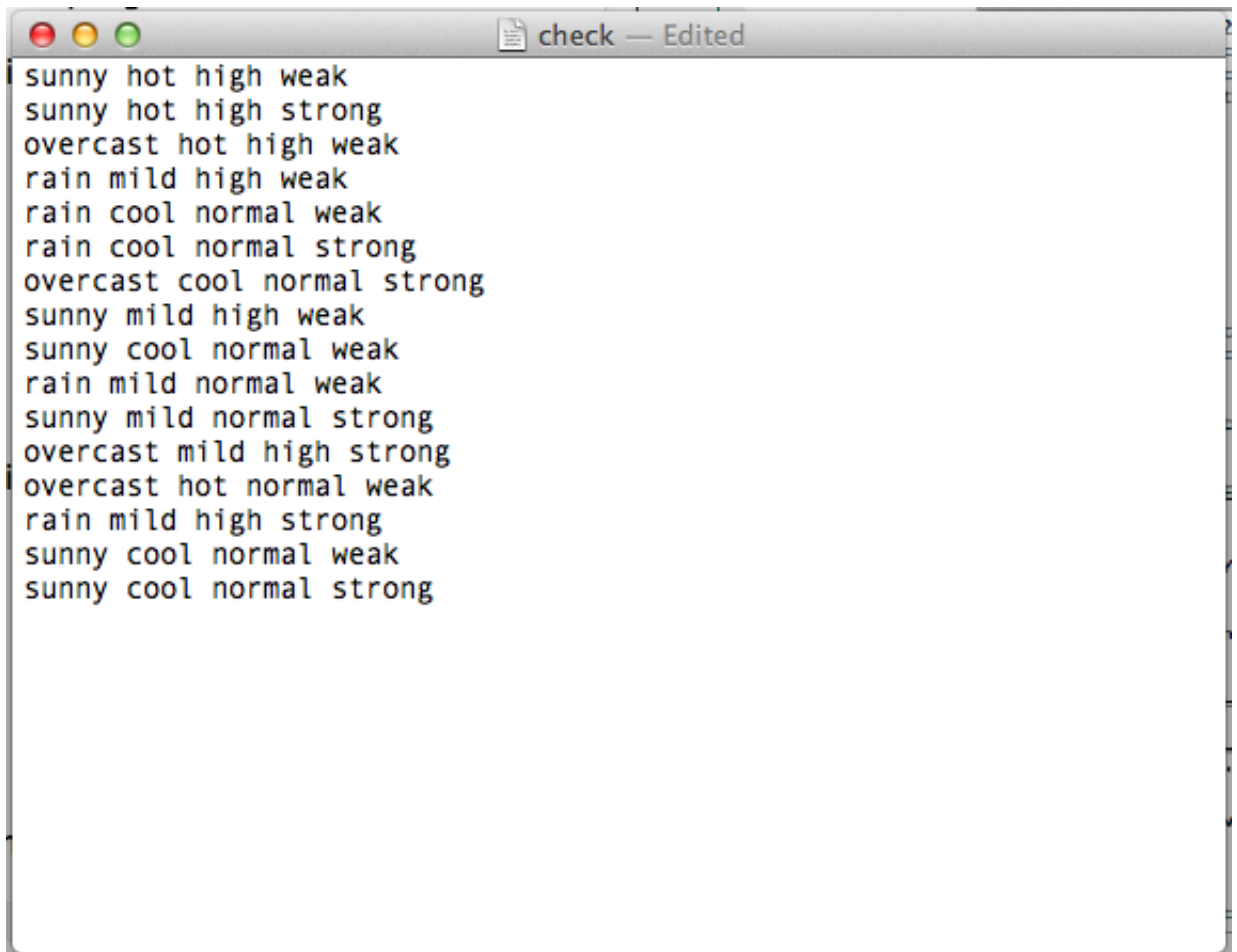


```
sunny hot high weak
sunny hot high strong
overcast hot high weak
rain mild high weak
rain cool normal weak
rain cool normal strong
overcast cool normal strong
sunny mild high weak
sunny cool normal weak
rain mild normal weak
sunny mild normal strong
overcast mild high strong
overcast hot normal weak
rain mild high strong
sunny cool normal weak
sunny cool normal strong
```

Figure 3: Validation set

The above figure shows the validation set has 16 possible values with out the target attribute. We give this validation set then the program outputs the target attribute.

```
 ● ● ●                    Programs — bash — 80×27
  PREDICTED ATTRIBUTE :   yes

  PERCENTAGE OF ERROR RATE= 0.0

Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$
Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$ java SecondProg check.txt
  PREDICTED ATTRIBUTE :   no
  PREDICTED ATTRIBUTE :   no
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   no
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   no
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   no
  PREDICTED ATTRIBUTE :   yes
  PREDICTED ATTRIBUTE :   yes

  PERCENTAGE OF ERROR RATE= 0.0

Mohammed-Ibrahims-MacBook-Pro:programs Mohammed$
```

Figure 4: Second program Output

The above figure shows the predicted attribute with error rate 0. In the validation set we have added two rows which includes possible values. Both the results are 'Yes' so the error rate is 0.

The below figure shows that the validation file has two extra possible values in which one predicted attribute is given as 'no'. The program gives the right predicted attributes for both the values and it calculate the percentage of error rate. Error rate is a simple formula to give an idea about the accuracy of the decision tree. It calculates the proportion of examples that have been misclassified by the decision tree.

Figure 5: Validation set and Output