Decision Tree Learning

Mallika Subramanian 2018101041

20 April 2020

1 Introduction

A decision tree represents a function that takes as input a vector of attribute values and returns a "decision"—a single output value. The input and output values can be discrete or continuous.
Each leaf node in the tree specifies a value to be returned by the function.

A Boolean decision tree is logically equivalent to the assertion that the goal attribute is true if and only if the input attributes satisfy one of the paths leading to a leaf with value true. Writing this out in propositional logic, we have

$$Goal \Leftrightarrow (Path1VPath2V...) \tag{1}$$

where each Path is a conjunction of attribute-value tests required to follow that path.

2 Dataset Provided

The Data set that has been used in this problem to construct the corresponding Decision Tree is as follows :

No.	Eye-Colour	Height(cm)	Hair-Length	Ideal
1.	Black	170	Long	N
2.	Black	179	Long	N
3.	Black	169	Short	Y
4.	Black	177	Short	Y
5.	Brown	176	Long	N
6.	Brown	164	Short	Y
7.	Brown	176	Short	N
8.	Brown	178	Short	N
9.	Brown	164	Long	Y
10.	Black	177	Long	N

Table 1: Initial Data-set

As we can observe from the table, there are *three* major attributes that determine the decisions made - **Eye-Colour** , **Height** and **Hair-Length**.

The Hair-Length and Eye-Colour attributes are **discrete** valued attributes. That is they can take one of the labels from a *finite* set of labels.

- Hair-Length can take **2** values : Long, Short
- Eye-Colour can take **2** values : Black, Brown

On the other hand, the Height(cm) attribute is a **continuous** valued attribute. That is it can take any real valued number as the Height. Therefore there are infinite labels that this attribute can pick from.

3 Algorithm Overview

In order to build the decision tree, we need to pick a deciding attribute at each stage so as to branch the examples from the data-set and place them in different buckets.

Entropy: At each level in the tree, entropy is the measure of randomness in the data. It is a measure of the uncertainty of the random variable. Throughout this report the entropy is represented in terms of *B(q)* where q is the fraction of examples whose result is positive. That is

$$B(q) = -qlog(q) - (1-q)log(1-q) \tag{2}$$

since the random variable under consideration here is a "Y/N" Boolean random variable.

Now in order to choose the deciding attribute based on which the data-set is to be partitioned, we require a deciding function. We know that acquisition of information corresponds to a reduction in entropy. An attribute A with d distinct values divides the training set of examples E into subsets $\mathrm{E_1}$,..., $\mathrm{E_d}$.
In order to compute the expected entropy remaining after testing attribute A we have

$$Remainder(A) = \sum \frac{(p_{\mathrm{k}} \ + n_{\mathrm{k}})}{p+n} B\left(\frac{p_{\mathrm{k}}}{p_{\mathrm{k}} + n_{\mathrm{k}}}\right) \tag{3}$$

That is, a *weighted average* of the different labels for a single attribute are used to compute the *Remainder(A)*. Once we obtain this remainder, we can calculate the Information Gain corresponding to each attribute and thereby select the deciding attribute.

Information Gain: for an attribute A is the expected reduction The attribute with the maximum information gain will help arrive at a decision faster. It is critical to select the right attribute for the partitioning since otherwise the depth of the tree may increase.

$$Gain\,(A) \ = \ B\left(\frac{p}{p+n}\right) \ - \ Remainder\,(A) \tag{4}$$

This same process is repeated at every intermediate node to determine the best attribute based on which the examples at that node must be split.

4 Building the Tree

As mentioned earlier, analysing the data set we see that there are 2 discrete valued attributes and 1 continuous valued attribute.
The ideal split point for the continuous valued "Height" attribute can be obtained via trial and error. This will in turn reflect on the information gain for the Height attribute and the structure of the tree can thus change.

Whilst building the tree, for the continuous valued attributes there were at max *example-number* of distinct initial split points that I could have considered for the Height attribute :

- Split at $H < 172$ and $H >= 172$
- Split at $H < 170$ and $H >= 170$
- Split at $H < 179$ and $H >= 179$
- Split at $H < 169$ and $H >= 169$
- Split at $H < 177$ and $H >= 177$
- Split at $H < 176$ and $H >= 176$
- Split at $H < 164$ and $H >= 164$
- Split at $H < 176$ and $H >= 176$
- Split at $H < 178$ and $H >= 178$

Naturally the tree generated in each of these cases would be different. As I proceeded further, there were intermediate nodes as well, where a selection was to be made regarding the continuous valued attribute. At each stage by trial and error, I made appropriate changes to the structure of the tree and the one with the least depth is show in the following sections.

Note: At every stage, the set of examples from the data set that are relevant to that node are tabulated and shown alongside the explanation. The entropy is also calculated at each step.

4.1 Choosing the first deciding attribute:

In the beginning the entire data set is considered. That is :

Calculate the entropy

The number of $\mathbf{Y} = 4$ The number of $\mathbf{N} = 6$ Therefore $q = \frac{4}{10}$

Entropy : $B\left(\frac{4}{10}\right) = B\left(\frac{2}{5}\right) = \frac{-2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right)$

$$B\left(\frac{2}{5}\right) = 0.9709505945 \tag{5}$$

No.	Eye-Colour	Height(cm)	Hair-Length	Ideal
1.	Black	170	Long	N
2.	Black	179	Long	N
3.	Black	169	Short	Y
4.	Black	177	Short	Y
5.	Brown	176	Long	N
6.	Brown	164	Short	Y
7.	Brown	176	Short	N
8.	Brown	178	Short	N
9.	Brown	164	Long	Y
10.	Black	177	Long	N

Table 2: Initial Data-set

Gain(Eye-Colour) : $B\left(q\right) \;-\; \left[\frac{1}{2}B\left(\frac{2}{5}\right)+\frac{1}{2}B\left(\frac{2}{5}\right)\right]$

$$G\left(E\right) \;=\; 0 \tag{6}$$

Gain(Hair-Length) : $B\left(q\right) \;-\; \left[\frac{1}{2}B\left(\frac{1}{5}\right)+\frac{1}{2}B\left(\frac{3}{5}\right)\right]$

$$B\left(q\right)-\left[\frac{1}{2}\left(-\frac{1}{5}\log\left(\frac{1}{5}\right)-\frac{4}{5}\log\left(\frac{4}{5}\right)\right)+\frac{1}{2}\left(-\frac{3}{5}\log\left(\frac{3}{5}\right)-\frac{2}{5}\log\left(\frac{2}{5}\right)\right)\right] \tag{7}$$

$$G\left(Len\right) \;= 0.9709505945 - 0.8464393447 = 0.12451124978 \tag{8}$$

Gain(Height) : Consider split at $H < 170$ and $H >= 170$ (since it has a maximum gain value)

$$B\left(q\right) \;-\; \left[\frac{3}{10}B\left(\frac{3}{3}\right)+\frac{7}{10}B\left(\frac{1}{7}\right)\right] \tag{9}$$

$$B\left(q\right)-\left[0+\frac{7}{10}\left(-\frac{1}{7}\log\left(\frac{1}{7}\right)-\frac{6}{7}\log\left(\frac{6}{7}\right)\right)\right] \tag{10}$$

$$G\left(Ht\right) \;= 0.9709505945 - 0.414170945 = 0.5567796495 \tag{11}$$

The gains for the remaining split points are as follows :

- $H < 179$ and $H >= 179$: $Gain = 0.07898214060026854$
- $H < 169$ and $H >= 169$: $Gain = 0.3219280948873623$

- $H < 177$ and $H >= 177$: $Gain = 0.0464393446710154$
- $H < 176$ and $H >= 176$: $Gain = 0.256425891682003$
- $H < 164$ and $H >= 164$: $Gain = 0.0$
- $H < 178$ and $H >= 178$: $Gain = 0.17095059445466854$

From Eqns (6), (8) and (11) we see that the Gain for the **Height** attribute is the maximum, and this maximum is for the splitpoint H=170.
Furthermore, we have successfully been able to distinguish three examples under the **YES** category. The remaining still have to be classified further.

Hence that is selected as the first attribute for the partition and the structure of the tree looks like

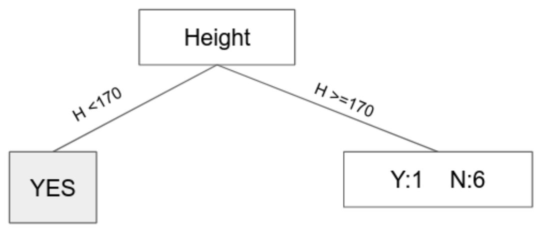

Figure 1: Stage 1

4.2 Next deciding attribute:

After splitting on the Height, for the examples with $H >= 170$ we have the following data

No.	Eye-Colour	Height(cm)	Hair-Length	Ideal
1.	Black	170	Long	N
2.	Black	179	Long	N
3.	Black	177	Short	Y
4.	Brown	176	Long	N
5.	Brown	176	Short	N
6.	Brown	178	Short	N
7.	Black	177	Long	N

Table 3: Stage2 Data-set

Calculate the entropy

The number of $\mathbf{Y} = 1$ The number of $\mathbf{N} = 6$ Therefore $q = \frac{1}{7}$

Entropy : $B\left(\frac{1}{7}\right) = \frac{-1}{7}\log\left(\frac{1}{7}\right) - \frac{6}{7}\log\left(\frac{6}{7}\right)$

$$B\left(\frac{1}{7}\right) = 0.5916727786 \tag{12}$$

Gain(Eye-Colour) : $B\left(q\right) \;-\; \left[\frac{4}{7}B\left(\frac{1}{4}\right) + \frac{3}{7}B\left(\frac{0}{3}\right)\right]$

$$B\left(q\right) - \left[\frac{4}{7}\left(-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right) + 0\right] \tag{13}$$

$$0.5916727786 - \;0.4635874997 = 0.1280852789 \tag{14}$$

Gain(Hair-Length) : $B\left(q\right) \;-\; \left[\frac{4}{7}B\left(0\right) + \frac{3}{7}B\left(\frac{1}{3}\right)\right]$

$$B\left(q\right) - \left[0 + \frac{3}{7}\left(-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right)\right] \tag{15}$$

$$0.5916727786 - .3935553575 = 0.1981174211 \tag{16}$$

Gain(Height) : Consider split at $H < 177$ and $H >= 177$ (since it has the maximum gain)

$$B\left(q\right) \;-\; \left[\frac{3}{7}B\left(0\right) + \frac{4}{7}B\left(\frac{1}{4}\right)\right] \tag{17}$$

$$B\left(q\right) - \left[0 + \frac{4}{7}\left(-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right)\right] \tag{18}$$

$$0.5916727786 - \;0.4635874997 = 0.1280852789 \tag{19}$$

The gains for the remaining split points are as follows :

- $H < 170$ and $H >= 170$: $Gain = 0.0$
- $H < 179$ and $H >= 179$: $Gain = 0.03451070288373825$
- $H < 176$ and $H >= 176$: $Gain = 0.03451070288373825$
- $H < 178$ and $H >= 178$: $Gain = 0.0760098536627829$

From Eqns (14), (16) and (19) we see that the Gain for the **Hair-Length** attribute is the maximum.
Once again, we have successfully been able to distinguish four examples under the **NO** category. The remaining still have to be classified further.

Hence that is selected as the second attribute for the partition and the structure of the tree looks like

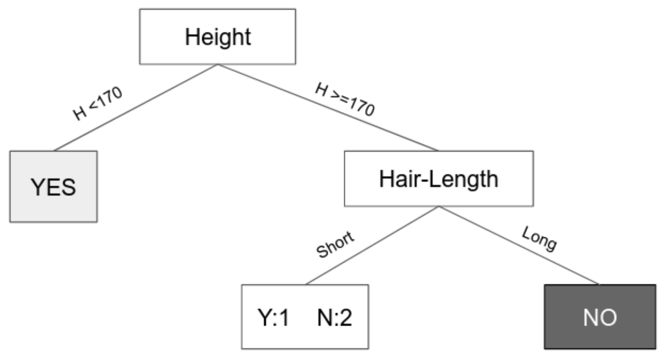

Figure 2: Stage 2

4.3 Next deciding attribute:

After splitting on the Height, for the examples with $H >= 170$ and $Length == Short$ we have the following data

No.	Eye-Colour	Height(cm)	Hair-Length	Ideal
1.	Black	177	Short	Y
2.	Brown	176	Short	N
3.	Brown	178	Short	N

Table 4: Stage2 Data-set

Calculate the entropy

The number of $\mathbf{Y} = 1$ The number of $\mathbf{N} = 2$ Therefore $q = \frac{1}{3}$

$\underline{\text{Entropy}}: B\left(\frac{1}{3}\right) = \frac{-1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)$

$$B\left(\frac{1}{3}\right) = 0.9182958341 \tag{20}$$

$\underline{\text{Gain(Eye-Colour)}}: B(q) \; - \; \left[\frac{1}{3}B(1) + \frac{2}{3}B(0)\right]$

$$B(q) - [0 + 0] \tag{21}$$

$$0.9182958341 - \ 0 = 0.9182958341 \tag{22}$$

Gain(Hair-Length) : $B\left(q\right) \ - \ \left[1B\left(\frac{1}{3}\right)\right]$

$$B\left(q\right) - \left[1\left(-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right)\right] \tag{23}$$

$$0.9182958341 - \ 0.9182958341 = 0 \tag{24}$$

Gain(Height) : Consider split at $H < 177$ and $H >= 177$

$$B\left(q\right) \ - \ \left[\frac{1}{3}B\left(0\right) + \frac{2}{3}B\left(\frac{1}{2}\right)\right] \tag{25}$$

$$B\left(q\right) - \left[0 + \frac{2}{3}\left(-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)\right)\right] \tag{26}$$

$$0.9182958341 - \ 0.6666666667 = 0.2516291674 \tag{27}$$

The gains for the remaining split points are as follows :

- $H < 176$ and $H >= 176$: $Gain = 0.0$
- $H < 178$ and $H >= 178$: $Gain = 0.2516291673878229$

From Eqns (20), (24) and (27) we see that the Gain for the **Eye-Colour** attribute is the maximum.
Finally we can see now that this attribute has partitioned the remaining set also into one **YES** and one **NO** leaf node in the tree. This indicates that we have finished the construction of the decision tree for this dataset.
Hence that is selected as the third attribute for the partition and the structure of the tree looks like :

With this we have completed constructing the decision tree for the given dataset.
The tree has a depth of 4
And there are 4 leaf nodes that classify the examples into YES and NO buckets.

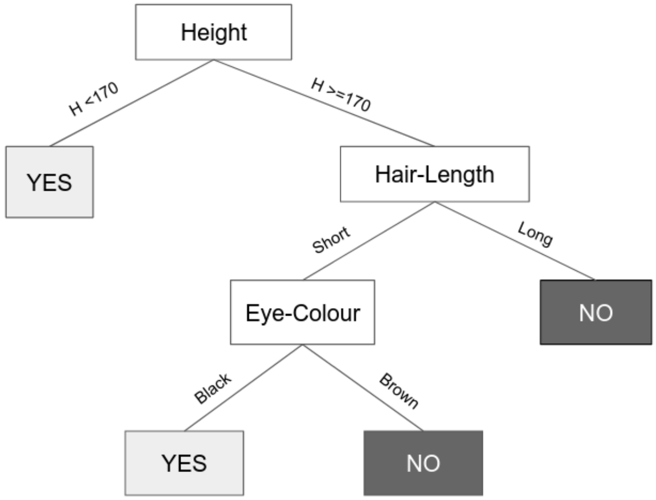

Figure 3: Stage 3

5 Inferences and Conclusion

The final structure of the final decision tree for the given data-set is shown in Section 4.3-Figure:3

The result that is expected of the algorithm is a tree that is consistent and in agreement with the examples and has the least depth. However, it is impossible to find the smallest tree as there is no polynomial time algorithm to traverse through all the 2^{2^n} candidate trees.

Therefore the the algorithm uses a *greedy* strategy. This *greedy, divide and conquer* algorithm ensures that maximum information is gained at every stage, thus picking the attribute that gives us the least Remainder value. That is the most important attribute is always tested first.