DMML Coursework C

Joseph Davidson
071729468
MEng Software Engineering

1 Introduction

This is the report for DMML 2010 coursework c set by Professor De Wilde. This coursework concerned itself with the construction of a decision tree that is used to select Facebook adverts for users based on certain attributes. To this end, a training set of 14 records was selected, each of which having one of 3 adverts. 4 attributes that seemed the most relevant to the adverts were also chosen.

The 3 adverts selected were:

- A dating website.
- A car insurance website.
- An advert for coupons that offer deals to Edinburgh students.

The 4 attributes selected:

- Relationship Status.
- Current Location.
- Education.
- Age[1].

Table 1: The 14 records selected

Record	RS	CL	Education	Age	Selected Advert
1	Single	Edinburgh	University	20	Edinburgh Student Deals
2	Relationship	London	University	19	Car Insurance
3	Single	Glasgow	University	21	Dating Website
4	Single	Edinburgh	University	20	Edinburgh Student Deals
5	Relationship	Glasgow	School	18	Car Insurance
6	Single	Edinburgh	School	18	Dating Website
7	Single	Edinburgh	University	19	Dating Website
8	Relationship	London	None	21	Car Insurance
9	Relationship	Glasgow	University	19	Car Insurance
10	Single	London	None	21	Dating Website
11	Single	Edinburgh	None	19	Dating Website
12	Relationship	Edinburgh	School	18	Edinburgh Student Deals
13	Single	Glasgow	School	18	Car Insurance
14	Relationship	London	University	20	Car Insurance

2 Calculating the root

All of the calculations for building the tree were entropy and gain calculations. These formulae were taken from the lecture notes and are reiterated below.

[1]While Age itself is not an attribute, the age can be inferred by the birthday attribute that is part of a Facebook page.

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i log_2 p_i \tag{1}$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

First the attribute of the root had to be calculated. This was done by working out the entropy over the entire set ($Entropy([3_{ES}, 6_{CI}, 5_{DS}]) = 1.53$ where ES = Edinburgh Student Deals, CI = Car Insurance and DS = Dating Site) and then calculating the gain from sorting on a particular attribute.

2.1 Gain on Relationship Status

$Entropy(S_{Single}) \equiv Entropy([2_{ES}, 1_{CI}, 5_{DS}]) = 1.298$
$Entropy(S_{Relationship}) \equiv Entropy([1_{ES}, 5_{CI}, 0_{DS}]) = 0.65$
$Gain(S, RS) = 0.509$

2.2 Gain on Current Location

$Entropy(S_{Edinburgh}) \equiv Entropy([3_{ES}, 0_{CI}, 3_{DS}]) = 1$
$Entropy(S_{Glasgow}) \equiv Entropy([0_{ES}, 3_{CI}, 1_{DS}]) = 0.811$
$Entropy(S_{London}) \equiv Entropy([0_{ES}, 3_{CI}, 1_{DS}]) = 0.811$
$Gain(S, CL) = 0.638$

2.3 Gain on Education

$Entropy(S_{University}) \equiv Entropy([2_{ES}, 3_{CI}, 2_{DS}]) = 1.556$
$Entropy(S_{School}) \equiv Entropy([1_{ES}, 2_{CI}, 1_{DS}]) = 1.5$
$Entropy(S_{None}) \equiv Entropy([0_{ES}, 1_{CI}, 2_{DS}]) = 0.918$
$Gain(S, Edu) = 0.1267$

2.4 Gain on Age

$Entropy(S_{18}) \equiv Entropy([1_{ES}, 2_{CI}, 1_{DS}]) = 1.5$
$Entropy(S_{19}) \equiv Entropy([0_{ES}, 2_{CI}, 2_{DS}]) = 1$
$Entropy(S_{20}) \equiv Entropy([2_{ES}, 1_{CI}, 0_{DS}]) = 0.918$
$Entropy(S_{21}) \equiv Entropy([0_{ES}, 1_{CI}, 2_{DS}]) = 0.918$
$Gain(S, Age) = 0.4222$

The best gain experienced by sorting on attributes was from using Current Location. This is why it was selected as the root attribute.

3 The Second Level

For the second level of the tree, I had to do entropy and gain calculations for each of the remaining attributes paired with each leaf of the root node. Which happened to be $3^2 = 9$. Each leaf now had a reduced set of S, which is shown below:

$S_{Edinburgh} = [D1, D4, D6, D7, D11, D12]$

$S_{Glasgow} = [D3, D5, D9, D13]$

$S_{London} = [D2, D8, D10, D14]$

The specific calculations for determining the next node for each branch is below, but here are the results. The bolded results are the ones that were selected as the next node in each branch.

$Gain(S_{Edinburgh}, RS) = 0.196$
$Gain(S_{Edinburgh}, Edu) = 0.2$
$\mathbf{Gain(S_{Edinburgh}, Age) = 0.666}$

$Gain(S_{Glasgow}, RS) = 0.311$
$Gain(S_{Glasgow}, Edu) = 0.311$
$\mathbf{Gain(S_{Glasgow}, Age) = 0.811}$

$\mathbf{Gain(S_{London}, RS) = 0.811}$
$Gain(S_{London}, Edu) = 0.311$
$Gain(S_{London}, Age) = 0.311$

3.1 Relationship Status Calculations

3.1.1 Edinburgh

$Entropy(S_{Edinburgh}) \equiv Entropy([3, 0, 3]) = 1$

$Entropy(S_{Edinburgh/Single}) \equiv Entropy([2_{ES}, 0_{CI}, 3_{DS}]) = 0.970$
$Entropy(S_{Edinburgh/Relationship}) \equiv Entropy([1_{ES}, 0_{CI}, 0_{DS}]) = 0$
$Gain(S_{Edinburgh}, RS) = 0.196$

3.1.2 Glasgow

$Entropy(S_{Glasgow}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{Glasgow/Single}) \equiv Entropy([0_{ES}, 1_{CI}, 1_{DS}]) = 1$
$Entropy(S_{Glasgow/Relationship}) \equiv Entropy([0_{ES}, 2_{CI}, 0_{DS}]) = 0$
$Gain(S_{Glasgow}, RS) = 0.311$

3.1.3 London

$Entropy(S_{London}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{London/Single}) \equiv Entropy([0_{ES}, 0_{CI}, 1_{DS}]) = 0$
$Entropy(S_{London/Relationship}) \equiv Entropy([0_{ES}, 3_{CI}, 0_{DS}]) = 0$
$Gain(S_{London}, RS) = 0.811$

3.2 Education Calculations

3.2.1 Edinburgh

$Entropy(S_{Edinburgh}) \equiv Entropy([3, 0, 3]) = 1$

$Entropy(S_{Edinburgh/University}) \equiv Entropy([2_{ES}, 0_{CI}, 1_{DS}]) = 0.918$
$Entropy(S_{Edinburgh/School}) \equiv Entropy([1_{ES}, 0_{CI}, 1_{DS}]) = 1$
$Entropy(S_{Edinburgh/None}) \equiv Entropy([0_{ES}, 0_{CI}, 1_{DS}]) = 0$
$Gain(S_{Edinburgh}, Edu) = 0.2$

3.2.2 Glasgow

$Entropy(S_{Glasgow}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{Glasgow/University}) \equiv Entropy([0_{ES}, 1_{CI}, 1_{DS}]) = 1$
$Entropy(S_{Glasgow/School}) \equiv Entropy([0_{ES}, 2_{CI}, 0_{DS}]) = 0$
$Entropy(S_{Glasgow/None}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Gain(S_{Glasgow}, Edu) = 0.311$

3.2.3 London

$Entropy(S_{London}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{London/University}) \equiv Entropy([0_{ES}, 2_{CI}, 0_{DS}]) = 0$
$Entropy(S_{London/School}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}]) = 0$ - This evaluated as 0.
$Entropy(S_{London/None}) \equiv Entropy([0_{ES}, 1_{CI}, 1_{DS}]) = 1$
$Gain(S_{London}, Edu) = 0.311$

3.3 Age Calculations

3.3.1 Edinburgh

$Entropy(S_{Edinburgh}) \equiv Entropy([3, 0, 3]) = 1$

$Entropy(S_{Edinburgh/18}) \equiv Entropy([1_{ES}, 0_{CI}, 1_{DS}]) = 1$
$Entropy(S_{Edinburgh/19}) \equiv Entropy([0_{ES}, 0_{CI}, 2_{DS}]) = 0$
$Entropy(S_{Edinburgh/20}) \equiv Entropy([2_{ES}, 0_{CI}, 0_{DS}]) = 0$
$Entropy(S_{Edinburgh/21}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Gain(S_{Edinburgh}, Age) = 0.666$

3.3.2 Glasgow

$Entropy(S_{Glasgow}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{Glasgow/18}) \equiv Entropy([0_{ES}, 2_{CI}, 0_{DS}]) = 0$
$Entropy(S_{Glasgow/19}) \equiv Entropy([0_{ES}, 1_{CI}, 0_{DS}]) = 0$
$Entropy(S_{Glasgow/20}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Entropy(S_{Glasgow/21}) \equiv Entropy([0_{ES}, 0_{CI}, 1_{DS}]) = 0$
$Gain(S_{Glasgow}, Age) = 0.811$

3.3.3 London

$Entropy(S_{London}) \equiv Entropy([0, 3, 1]) = 0.811$

$Entropy(S_{London/18}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Entropy(S_{London/19}) \equiv Entropy([0_{ES}, 1_{CI}, 0_{DS}]) = 0$
$Entropy(S_{London/20}) \equiv Entropy([0_{ES}, 1_{CI}, 0_{DS}]) = 0$
$Entropy(S_{London/21}) \equiv Entropy([0_{ES}, 1_{CI}, 1_{DS}]) = 1$
$Gain(S_{London}, Edu) = 0.311$

Now there is a second level for the tree. The age attribute is evaluated at two points, the Edinburgh branch and the Glasgow one. At the London branch, the relationship status attribute is evaluated.

4 Third Level

Only one path on the tree has any entropy left: selecting Edinburgh then 18 will result in another decision needing to be made. So I will have to evaluate the Relationship Status and Education attributes at this point.

The reduced dataset of S that is gained by sorting on Edinburgh and 18 is below as well as the entropy for that set:

$S_{Edinburgh/18} = [D6, D12]$
$Entropy(S_{Edinburgh/18}) \equiv Entropy([1_{ES}, 0_{CI}, 1_{DS}]) = 1$

4.1 Relationship Status

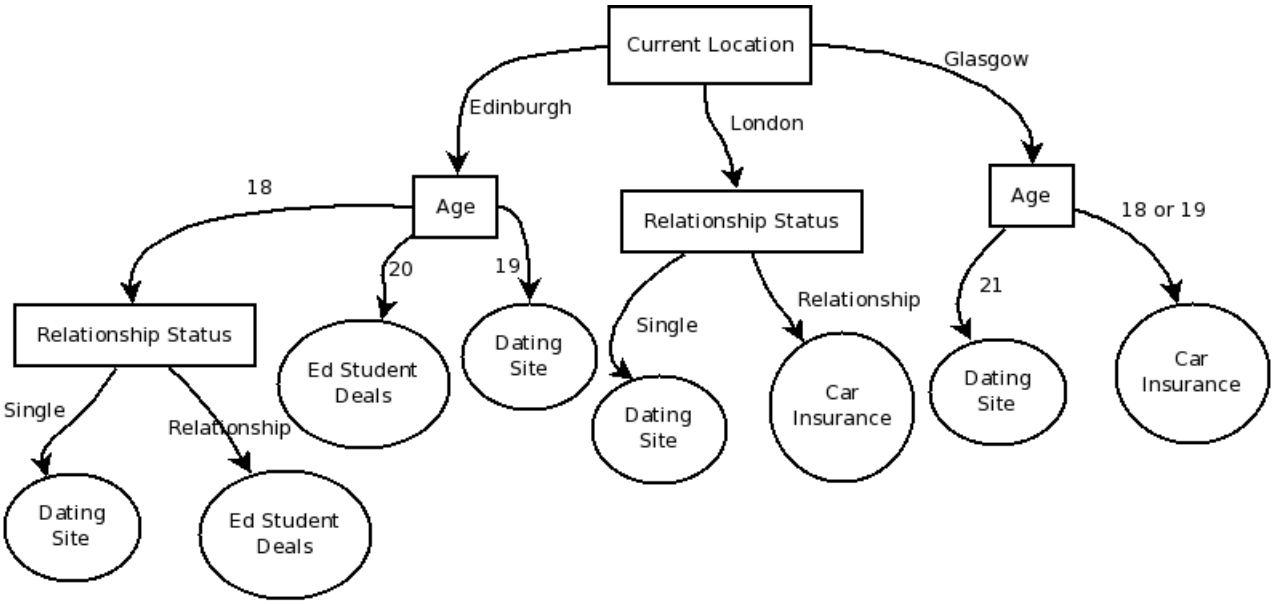
$Entropy(S_{Edinburgh/18/Single}) \equiv Entropy([0_{ES}, 0_{CI}, 1_{DS}]) = 0$
$Entropy(S_{Edinburgh/18/Relationship}) \equiv Entropy([1_{ES}, 0_{CI}, 0_{DS}]) = 0$
$Gain(S_{Edinburgh/18}, RS) = 0$

4.2 Education

$Entropy(S_{Edinburgh/18/University}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Entropy(S_{Edinburgh/18/School}) \equiv Entropy([1_{ES}, 0_{CI}, 1_{DS}]) = 0$
$Entropy(S_{Edinburgh/18/None}) \equiv Entropy([0_{ES}, 0_{CI}, 0_{DS}])$ - This evaluated as 0.
$Gain(S_{Edinburgh/18}, Edu) = 1$

As can be seen, sorting on Education gives no gain but sorting on Relationship Status gives us a gain of 1. We choose it for the next level thereby completing the tree.

5 The Final Tree



6 An Unclassified Example

Below, we have the essential details of a person on Facebook. The year of their birth is assumed to be the same as mine (1990).

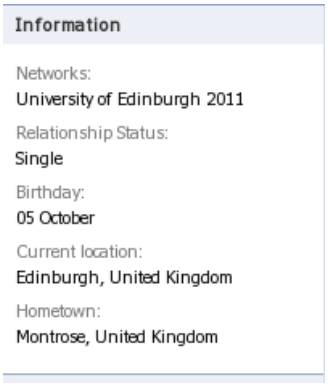


Figure 1: A Facebook pages statistics

This person makes up a record that conforms to the tree built with my original training set. Their information in record form is below:

Table 2: An Unclassified Record

Record	**RS**	**CL**	**Education**	**Age**	**Selected Advert**
15	Single	Edinburgh	University	20	?

Classifying them with the tree will follow this path (with the subscript showing the choice made at each decision point):

$$\text{Current Location}_{Edinburgh} \rightarrow \text{Age}_{20} \rightarrow \text{Edinburgh Student Deals} \tag{3}$$

This shows that the tree correctly classifies the instance as a page that should be displaying the Edinburgh Student Deals advert.

7 Final Comments

This tree uses a continuous data field for classification. This may have been a mistake because I don't think that the techniques I used in creating the tree were really meant for handling fields with continuous values. For instance, if the example record had an age of 21, the tree wouldn't have been able to classify it. Another classification strategy should really have been used.

As far as the rest of the coursework goes, I enjoyed it all immensely and even though I ran ID3 by hand to make the tree, it wasn't as much effort as I thought it would be.