Machine Learning and Pattern Recognition: CA2 Task 1

Submitted By : Vaibhav Tiwari
Student Number : 10621051
Lecturer: Courtney Ford

Task 1 : We have solved the assignment using manual steps and by using code, This document contains the steps performed manually. The link for the code is - https://colab.research.google.com/drive/1kelswJtYJJsSn_xCW5PE9c7NWdgx2ChL?usp=sharing

Dataset in csv format for importing in the code -

entropy.csv

Suppose the following dataset is about the properties of 14 people where the attribute "Won" shows whether a person will win the fashion competition or not. The attribute "Won" is the dependent attribute with two values (Won = 'yes', Won = 'no').

Age	Hair_Size	Brown_Eye	Sex	Won
youth	long	no	male	no
youth	long	no	female	no
middle_age	long	no	male	yes
senior	medium	no	male	yes
senior	short	yes	male	yes
senior	short	yes	female	no
middle_age	short	yes	female	yes
youth	medium	no	male	no
youth	short	yes	male	yes
senior	medium	yes	male	yes
youth	medium	yes	female	yes
middle_age	medium	no	female	yes
middle_age	long	yes	male	yes
senior	medium	no	female	no

There are independent variables to determine the dependent variable. The independent variables are Age, Hair_Size, Brown_Eye, and Sex. The dependent variable is whether the person Won or not.

As the first step, we have to find the parent node for our decision tree. For that follow the steps:

Find the entropy of the class variable('Won'):

E(S) = -[(9/14)log(9/14) + (5/14)log(5/14)] = 0.94

Note: Here typically we will take log to the base 2. In total there are 14 yes/no. Out of which 9 are yes and 5 are no. We have calculated the probability based on these values from the table above.

From the above data for 'Age' we can arrive at the following table easily:

Won				
		yes	no	total
Age	youth	4	0	4
	middle_age	3	2	5
	senior	2	3	5
				14

Now we have to calculate average weighted entropy.

ie, we have found the total of weights of each feature multiplied by probabilities.

E(S, Age) = (4/14)*E(4,0) + (5/14)*E(3,2) + (5/14)*E(2,3) = (4/14)(-(4/4)log(4/4)-(0/4)log(0/4))+ (5/14)(-(3/5)log(3/5)-(2/5)log(2/5)) + (5/14)(-(2/5)log(2/5)-(3/5)log(3/5)) = 0.693

The next step is to find the information gain.

It is the difference between parent entropy and average weighted entropy we found above.

IG(S, Age) = 0.94 - 0.693 = 0.247

Similarly find Information gain for Hair_Size, Brown_Eye, and Sex.

Won				
		yes	no	total
Hair_Size	short	3	1	4
	medium	4	2	6
	long	2	2	4
				14

E(S, Hair_Size) = (4/14)*E(3,1) + (6/14)*E(4,2) + (4/14)*E(2,2) = (4/14)(-(3/4)log(3/4)-(1/4)log(1/4))+ (6/14)(-(4/6)log(4/6)-(2/6)log(2/6)) + (4/14)(-(2/4)log(2/4)- (2/4)log(2/4)) = 0.911

IG(S, Hair_Size) = 0.940 - 0.911 = 0.029

Won				
		yes	no	total
Brown_Eyes	yes	6	1	7
	no	3	4	7
				14

E(S, Brown_Eyes) = (7/14)*E(6,1) + (7/14)*E(3,4) = (7/14)(-(6/7)log(6/7)-(1/7)log(1/7))+ (7/14)(-(3/7)log(3/7)-(4/7)log(4/7)) = 0.788

IG(S, Brown_Eyes) = 0.940 - 0.788 = 0.152

Won				
		yes	no	total
Sex	male	6	2	8
	female	3	3	6
				14

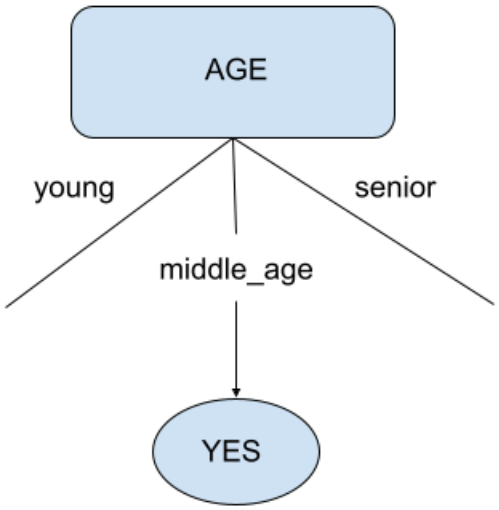
E(S, Sex) = (8/14)*E(6,2) + (6/14)*E(3,3) = (8/14)(-(6/8)log(6/8)-(2/8)log(2/8))+ (6/14)(-(3/6)log(3/6)- (3/6)log(3/6)) = 0.892

IG(S, Sex) = 0.940 - 0.892 = 0.048

Now select the feature having the largest information gain: Here it is Age. So it forms the first node(root node) of our decision tree. Now our data look as follows:

Age	Hair_Size	Brown_Eye	Sex	Won
youth	long	no	male	no
youth	long	no	female	no
youth	medium	no	male	no
youth	short	yes	male	yes
youth	medium	yes	female	yes
Age	Hair_Size	Brown_Eye	Sex	Won
middle_age	long	no	male	yes
middle_age	short	yes	female	yes
middle_age	medium	no	female	yes
middle_age	long	yes	male	yes
Age	Hair_Size	Brown_Eye	Sex	Won
senior	medium	no	male	yes
senior	short	yes	male	yes
senior	short	yes	female	no
senior	medium	yes	male	yes
senior	medium	no	female	no

Since the value ‘middle_age’ contains only examples of class ‘Yes’ we can set it as yes. That means the person is middle aged, he will win. Now our decision tree looks as follows.



The next step is to find the next node in our decision tree. Now we will find one under the value ‘youth’. We have to determine which of the remaining three columns (Hair_Size, Brown_Eye, Sex) has a higher information gain.

Age	Hair_Size	Brown_Eye	Sex	Won
youth	long	no	male	no
youth	long	no	female	no
youth	medium	no	male	no
youth	short	yes	male	yes
youth	medium	yes	female	yes

Calculate parent entropy E(youth)

E(youth) = (-(3/5)log(3/5)-(2/5)log(2/5)) = 0.971.

Now Calculate the information gain of ‘Hair_Size’ : IG(youth, Hair_Size)

Won				
		yes	no	total
Hair_Size	short	1	0	1
	medium	1	1	2
	long	0	2	2
				5

E(youth, Hair_Size) = (1/5)*E(1,0) + (2/5)*E(1,1) + (2/5)*E(0,2)=2/5=0.4

Now calculate information gain.

IG(youth, Hair_Size) = 0.971–0.4 =0.571

Similarly, we get:

Won				
		yes	no	total
Brown_Eye	Yes	2	0	2
	No	0	3	3
				5

E(youth, Brown_eye) = (2/5)*E(2,0) + (3/5)*E(0,3) = (2/5)(-(2/2)log(2/2)-(0/2)log(0/2))+

(3/5)(-(0/3)log(0/3)- (3/3)log(3/3)) =0.0

IG(youth, Brown_eye) = 0.971

Won				
		yes	no	total
Sex	male	1	2	3
	female	1	1	2
				5

E(youth, Sex) = (3/5)*E(1,2) + (2/5)*E(1,1) =

(3/5)(-(1/3)log(1/3)-(2/3)log(2/3))+ (2/5)(-(1/3)log(1/3)- (1/3)log(1/3)) =0.973

IG(youth,Sex) = -0.0020

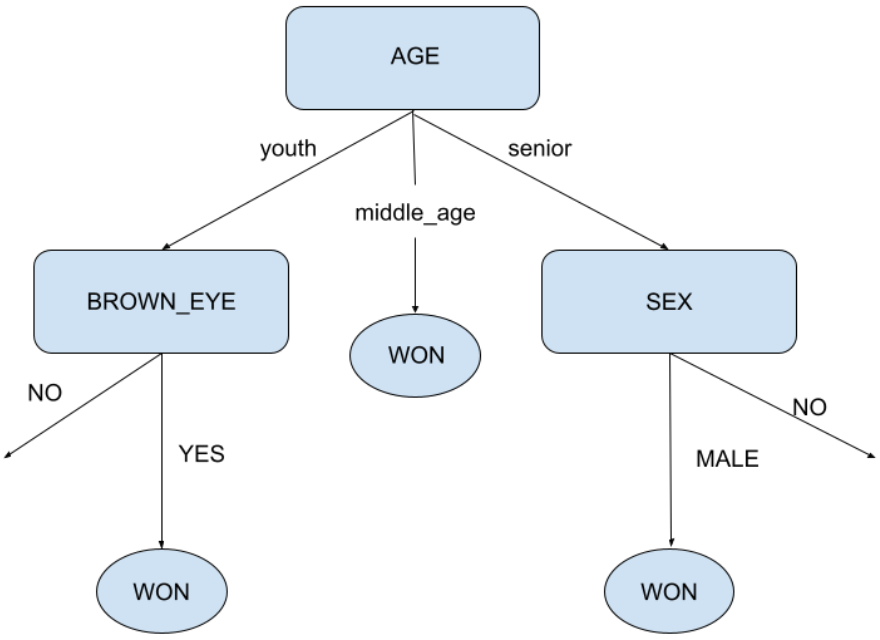
Here IG(youth, Brown_Eye) is the largest value. So Brown_Eye is the node that comes under youth.

For Brown_Eye column from the above table, we can say that person will win if they have brown eye but will lose if they don't have brown eye. Similarly, find the nodes under senior.

Age	Hair_Size	Brown_Eye	Sex	Won
senior	medium	no	male	yes
senior	short	yes	male	yes
senior	short	yes	female	no
senior	medium	yes	male	yes
senior	medium	no	female	no

For Sex column from the above table, we can say that person will win if they are Senior and their sex is Male. Otherwise, if the sex is female, the person will not win.

Finally, our decision tree will look as below:



CONCLUSION :

The completion of the decision tree using the method of Entropy helps us to answer the questions indicated in the CA two :
a) Which one of these features is the most important feature?
b) What is the best Information Gain (IG)?

We can say that **Age** is the most important feature of the dataset because it has the best Information Gain (0.247) and hence it was selected as the root node while deriving the decision tree shown above.

REFRENCES :
To plot manually -
Moodle - Week 7 – Decision Tree Example word Document
To plot using code -
https://elearning.dbs.ie/mod/folder/view.php?id=1295470
001_Decision_Tree_PlayGolf_ID3.ipynb