10/11/2020

Decision Tree Assignment

By: Mohammed Aadil (IIT2018179)

Q1) Consider two features, age and heart disease to create a decision tree with gini impurity.

Ans) As age is a continous value, we will take the average weight of the adjacent examples (After sorting).

I will do the first 5 examples:

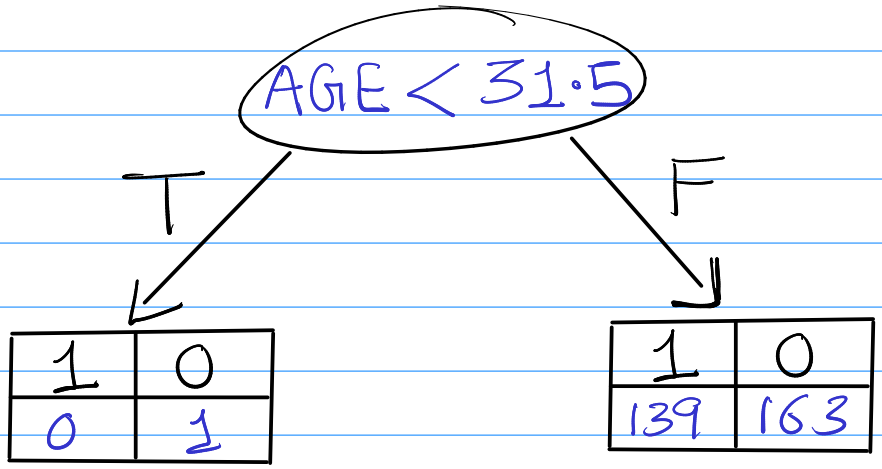
age	heart Disease
29	0
34	0
34	0
35	0
35	0

avg_age
→ 31.5
→ 34
→ 34.5
→ 35

using these values we make the decision trees with their Gini impurity

Note: Here I have only 5 examples but in the Dataset there are 303 examples.

1)



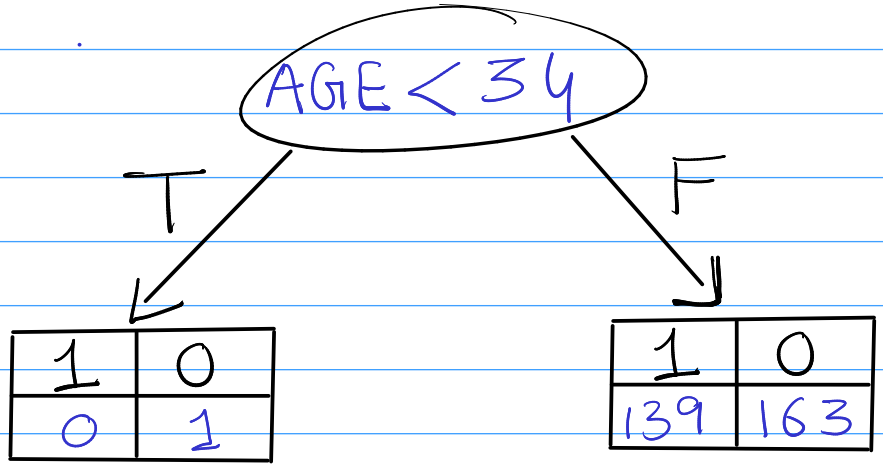
P(1) = 0/1
P(0) = 1/1
GINI = 0

P(1) = 139/302
P(0) = 163/302
GINI = 0.49684

Avg-GINI = $(1/303) * 0 + (302/303) * 0.49684$

= 0.49520

2)



1	0
0	1

1	0
139	163

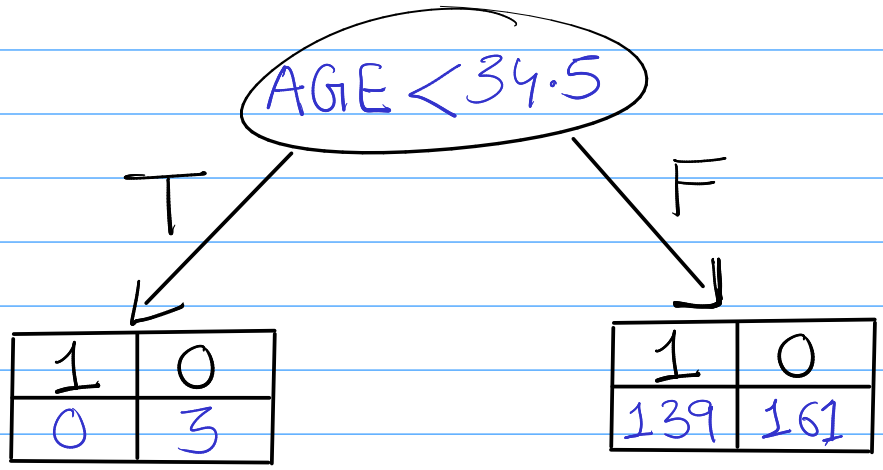
P(1) = 0/1
P(0) = 1/1
GINI = 0

P(1) = 139/302
P(0) = 163/302
GINI = 0.49684

Avg-GINI = $(1/303) * 0 + (302/303) * 0.4968$

= 0.49520

3)



1	0
0	3

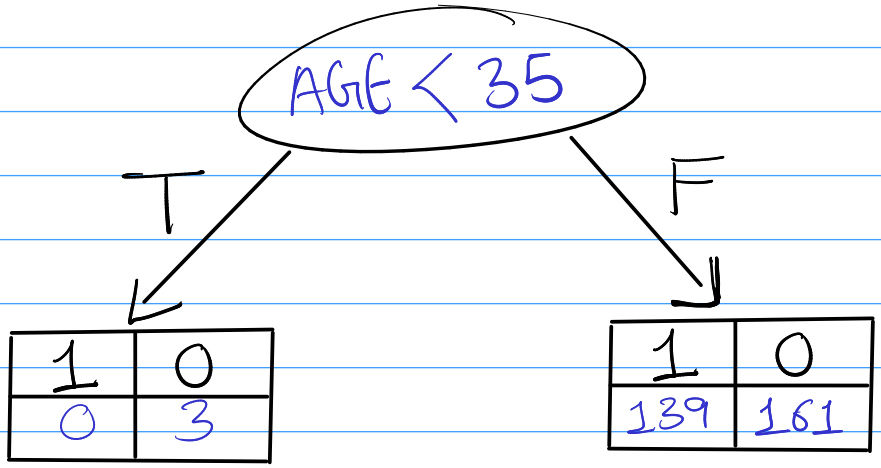
1	0
139	161

P(1) = 0/3
P(0) = 3/3
GINI = 0

P(1) = 139/300
P(0) = 161/300
GINI = 0.49731

Avg_GINI = (3/303) * 0 + (300/303) * 0.49731
= 0.49238

4)



P(1) = 0/3
P(0) = 3/3
GINI = 0

P(1) = 139/300
P(0) = 161/300
GINI = 0.49731

Avg_GINI = (3/303) * 0 + (300/303) * 0.49731
= 0.49238

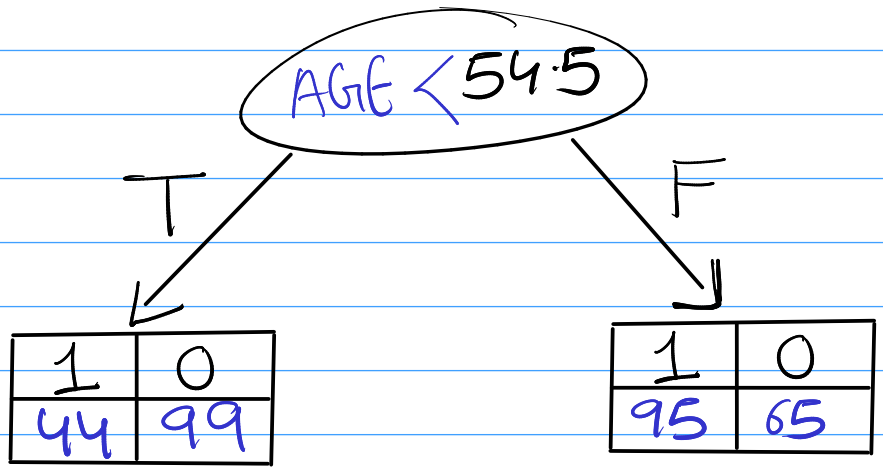
Similarly if we do this 302 times, for each pair we get Avg_GINI. We need to find the minimum of all the Avg_GINI's.

$$\text{OPTIMAL GINI} = \min(\text{Avg\_GINI})$$

Doing this by hand will be too tedious, so after implementing the code on JUPYTER NOTEBOOK, I got the following:

OPTIMUM GINI = 0.45581
OPTIMUM age = 54.5

The tree looks like this:



P(1) = 44/143
P(0) = 99/143
GINI = 0.42603

P(1) = 95/160
P(0) = 65/160
GINI = 0.48242

AVG GINI = (143/303) * 0.42603 + (160/303) * 0.48242

= 0.45581

Q2) Consider two features, slope and heart disease to create a decision tre with Information gain.

Ans) (i) First we need to calculate the entropy of heart disease

$$E(\text{heart Disease}) = E(139, 164)$$
$$= E(0.458, 0.541)$$
$$= -(0.458) \cdot \log_2(0.458) - (0.541) \cdot \log_2(0.541)$$
$$= 0.995$$

(ii) Entropy using the frequency table of features.

		Heart Disease		
		1	0	
Slope	1	36	106	142
	2	91	49	140
	3	12	9	21
				303

$$E(\text{Heart Disease}, \text{slope}) = \sum_{i=1}^{3} P(s_i) * E(s_i) \;\Big|\; s_i \in \text{Slope}$$

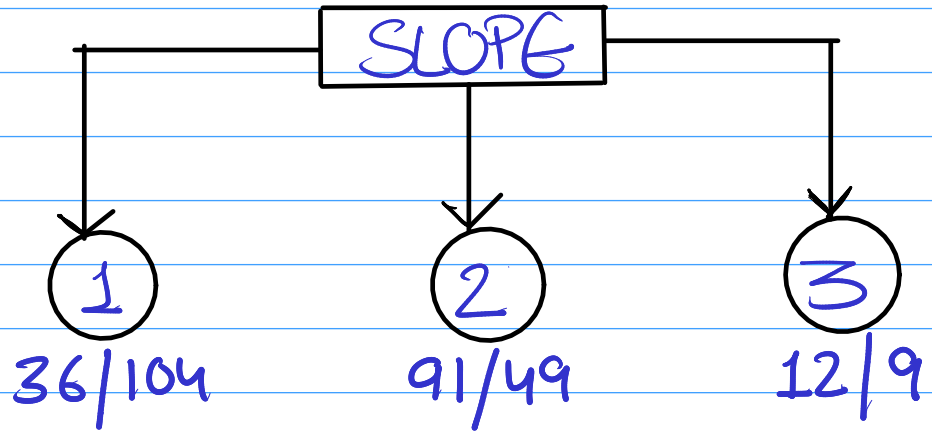
$$\Rightarrow P(1) * E(\text{slope}=1) + P(2) * E(\text{slope}=2) + P(3) * E(\text{slope}=3)$$
$$= 0.468 * E(36, 106) + 0.462 * E(91, 49) + 0.069 * E(12, 9)$$
$$= (0.468 * 0.816) + (0.462 * 0.934) + (0.069 * 0.985)$$
$$= \underline{\underline{0.8826}}$$

(iii) Now we need to find the gain.

$$\text{Gain}(\text{Heart Disease}, \text{slope}) = E(\text{Heart Disease}) - E(\text{Heart Disease}, \text{slope})$$
$$= 0.995 - 0.8826$$
$$= \underline{\underline{0.1124}}$$

Ideally we would many features, then we would calculate the GAIN for all the features and then find the MAX.GAIN and set it as the ROOT node of the decesion tree.

In our case we have only the slope as a feature ∴ we set it to root.



There are no more features that we can use to further branch the nodes. So this will be our final tree.