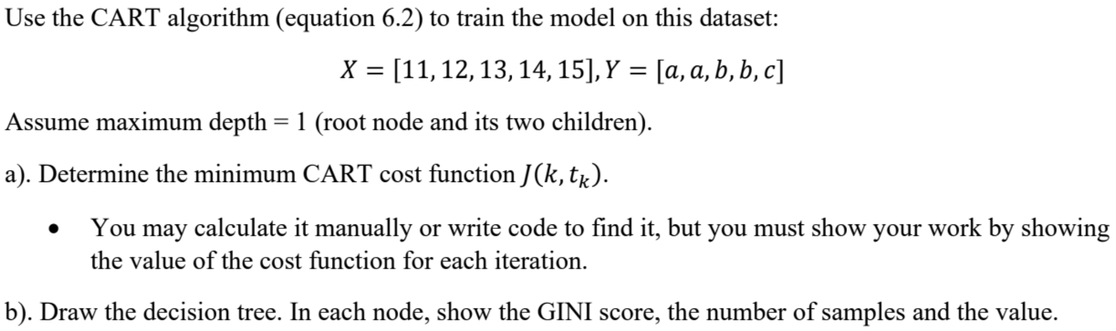
Homework - 7
COEN 240-Machine Learning

Pujitha Kallu
ID : W1653660
pkallu@scu.edu



Use the CART algorithm (equation 6.2) to train the model on this dataset:

$$X = [11, 12, 13, 14, 15], Y = [a, a, b, b, c]$$

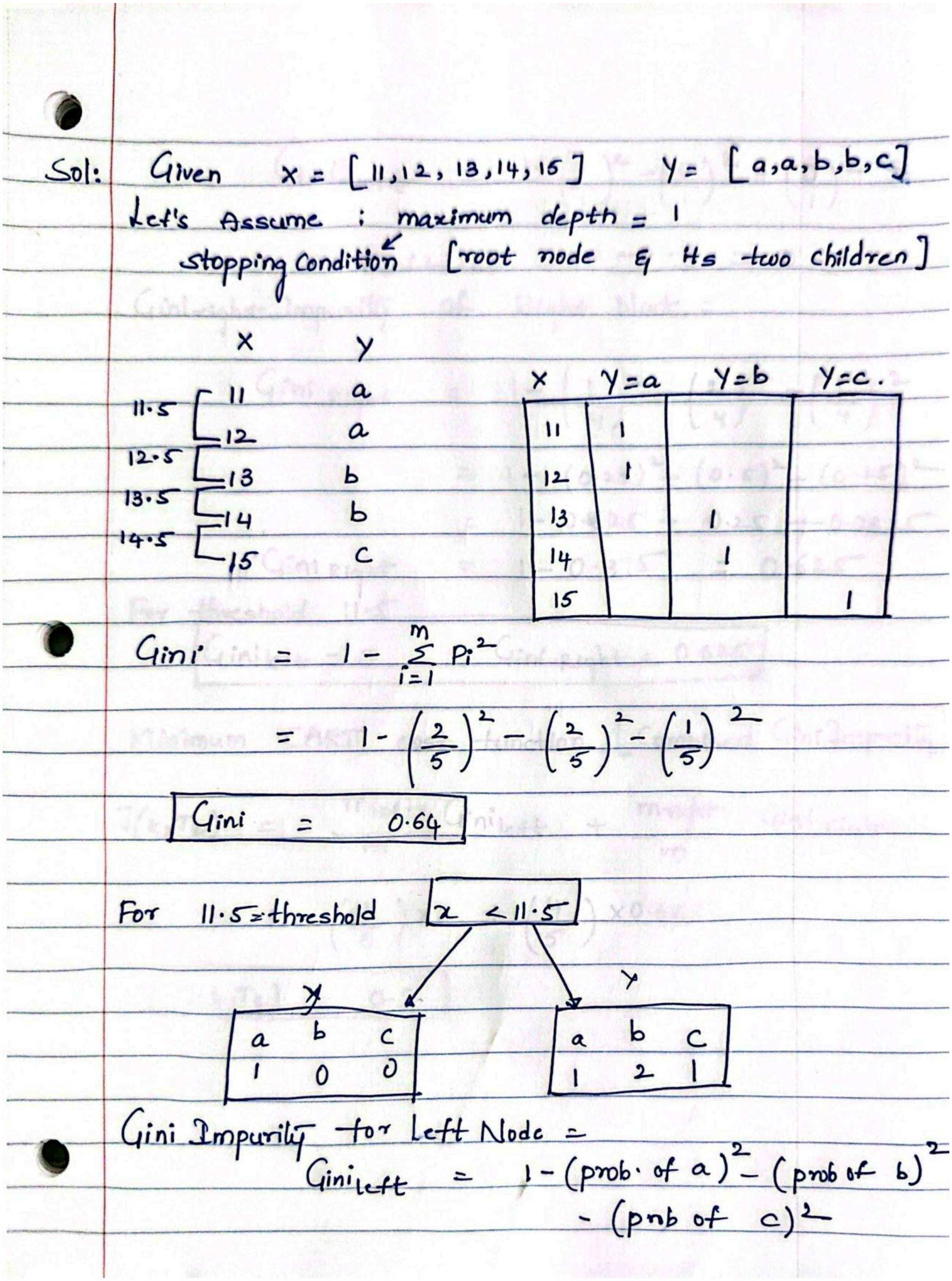
Assume maximum depth = 1 (root node and its two children).

a). Determine the minimum CART cost function $J(k, t_k)$.

- You may calculate it manually or write code to find it, but you must show your work by showing the value of the cost function for each iteration.

b). Draw the decision tree. In each node, show the GINI score, the number of samples and the value.

Sol:



Sol: Given $x = [11, 12, 13, 14, 15]$ $y = [a, a, b, b, c]$

Let's Assume : maximum depth = 1

stopping condition [root node & its two children]

	x	y
11.5	11	a
	12	a
12.5	13	b
13.5	14	b
14.5	15	c

x	y=a	y=b	y=c.
11	1		
12	1		
13		1	
14		1	
15			1

$$\text{Gini} = 1 - \sum_{i=1}^{m} P_i^2$$

$$= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2$$

$$\boxed{\text{Gini} = 0.64}$$

For 11.5 = threshold $\boxed{x < 11.5}$

y

a	b	c
1	0	0

y

a	b	c
1	2	1

Gini Impurity for Left Node =

$$\text{Gini}_{left} = 1 - (\text{prob. of } a)^2 - (\text{prob of } b)^2 - (\text{prob of } c)^2$$

$$\text{Gini}_{\text{left}} = 1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2-\left(\frac{0}{1}\right)^2$$

$$\text{Gini left} = 1-1-0-0 = 0$$

Giniright Impurity of Right Node =

$$\text{Gini}_{\text{Right}} = 1-\left(\frac{1}{4}\right)^2-\left(\frac{2}{4}\right)^2-\left(\frac{1}{4}\right)^2$$

$$= 1-(0.25)^2-(0.5)^2-(0.25)^2$$

$$= 1-0.0625-0.25-0.0625$$

$$\text{Gini}_{\text{Right}} = 1-0.375 = 0.625$$

For threshold 11.5

$$\boxed{\text{Gini}_{\text{left}} = 0 \qquad \text{Gini}_{\text{Right}} = 0.625}$$

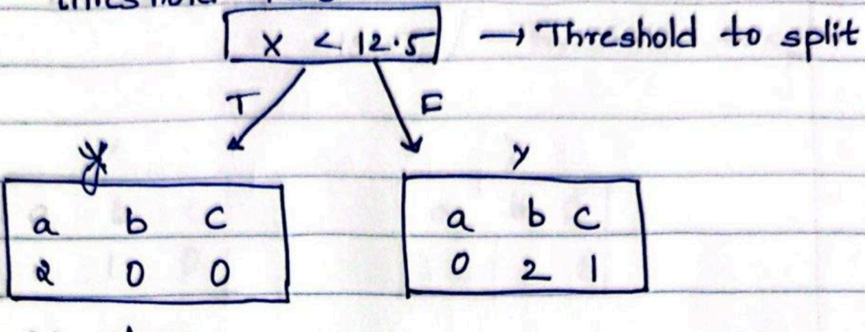
Minimum CART cost function [Combined Gini Impurity]

$$J(k,T_k) = \frac{m_{\text{left}}}{m}\,\text{Gini}_{\text{left}} + \frac{m_{\text{right}}}{m}\,\text{Gini}_{\text{right}}$$

$$= \left(\frac{1}{5}\right)\times 0 + \left(\frac{4}{5}\right)\times 0.625$$

$$\boxed{J(k,T_k) = 0.5}$$

For threshold 12.5



a	b	c
2	0	0

a	b	c
0	2	1

Same as above

$$\text{Gini}_{\text{Left}} = 1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2-\left(\frac{0}{2}\right)^2$$

$$= 1-1-0-0$$

$$\text{Gini}_{\text{Left}} = 0$$

$$\text{Gini}_{\text{Right}} = 1-\left(\frac{0}{3}\right)^2-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2$$

$$= 1-0-0.446-0.11$$

$$= 1-0.556$$

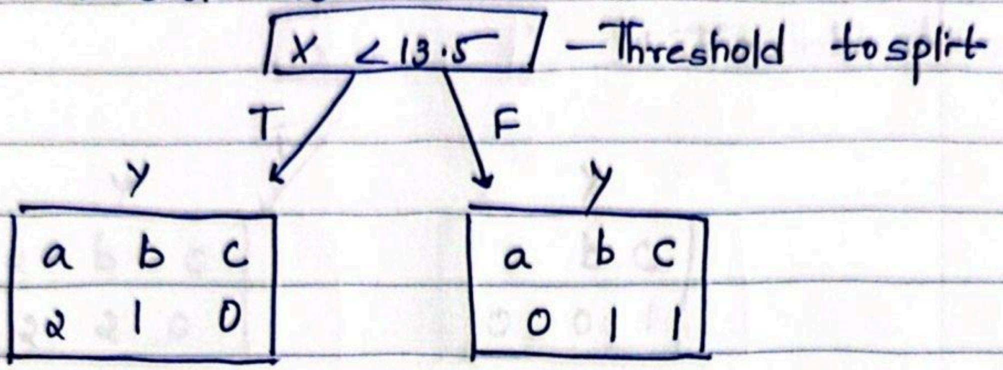
$$\text{Gini}_{\text{Right}} = 0.444$$

Minimum CART cost $J(k,t_k) = \left(\frac{2}{5}\right)\times 0 + \left(\frac{3}{5}\right)\times 0$

$$J(k,t_k) = 0 + 0.2664 .$$

$$\boxed{J(k,t_k) = 0.2664}$$

For threshold 13.5



a	b	c
2	1	0

a	b	c
0	1	1

$$Gini_{Left} = 1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2-\left(\frac{0}{3}\right)^2$$

$$= 1 - 0.444 - 0.11 - 0$$

$$Gini_{Left} = 0.444$$

$$Gini_{Right} = 1-\left(\frac{0}{2}\right)^2-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2$$

$$= 1 - 0 - 0.25 - 0.25$$

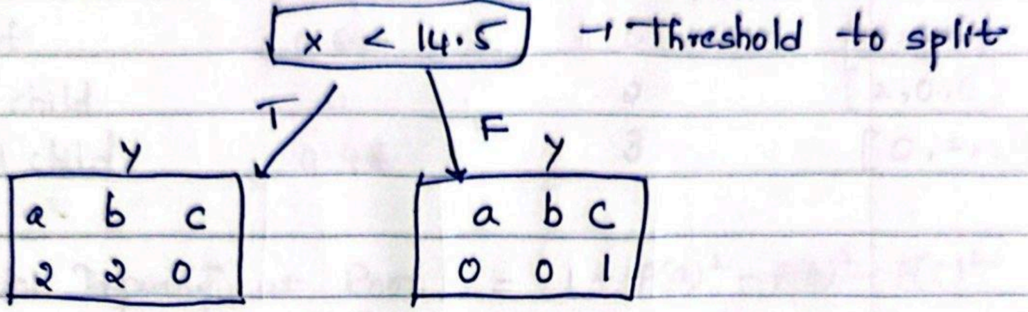
$$Gini_{Right} = 0.5$$

Minimum CART cost $J(k,t_k) = \left(\frac{3}{5}\right)\times 0.444 + \left(\frac{2}{5}\right)\times 0.5$

$$= 0.2664 + 0.2$$

$$\boxed{J(k,t_k) = 0.4664}.$$

For threshold 14.5



$$Gini_{Left} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 - \left(\frac{0}{4}\right)^2$$

$$= 1 - 0.25 - 0.25 - 0$$

$$Gini_{Left} = 0.5$$

$$Gini_{Right} = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2$$

$$Gini_{Right} = 0$$

Minimum CART cost $J(k, t_k) = \left(\frac{4}{5}\right) \times 0.5 + \left(\frac{1}{5}\right) \times 0$

$$= 0.4 + 0$$

$$\boxed{J(k, t_k) = 0.4}$$

∴ Of All the minimum CART cost is at threshold $\boxed{12.5 \text{ \& } J(k, t_k) \text{ is } 0.2664}$

	Gini Impurity score.	# of samples	Value.
Root	0.64	5	[2,2,1]
Lef child	0	2	[2,0,0]
Right child.	0.46	3	[0,2,1]

Gini Impurity at Root [Before Splitting] $= 1 - P(a)^2 - P(b)^2 - P(c)^2$

$= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2$

$= 1 - 0.16 - 0.16 - 0.04$

$= 1 - 0.36$

<u>Gini Impurity at Root $= 0.64$</u>

Gini Impurity at Leftchid $= 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{0}{2}\right)^2$

$= 1 - 1$

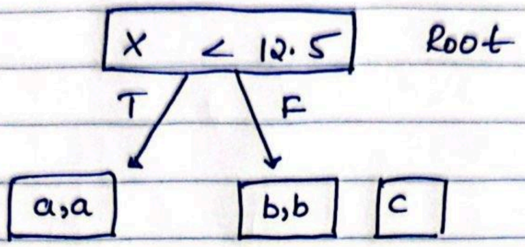
<u>Gini Impurity at Leftchild $= 0$.</u>

Gini Impurity at Right child $= 1 - \left[\frac{0}{3}\right)^2 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right.$

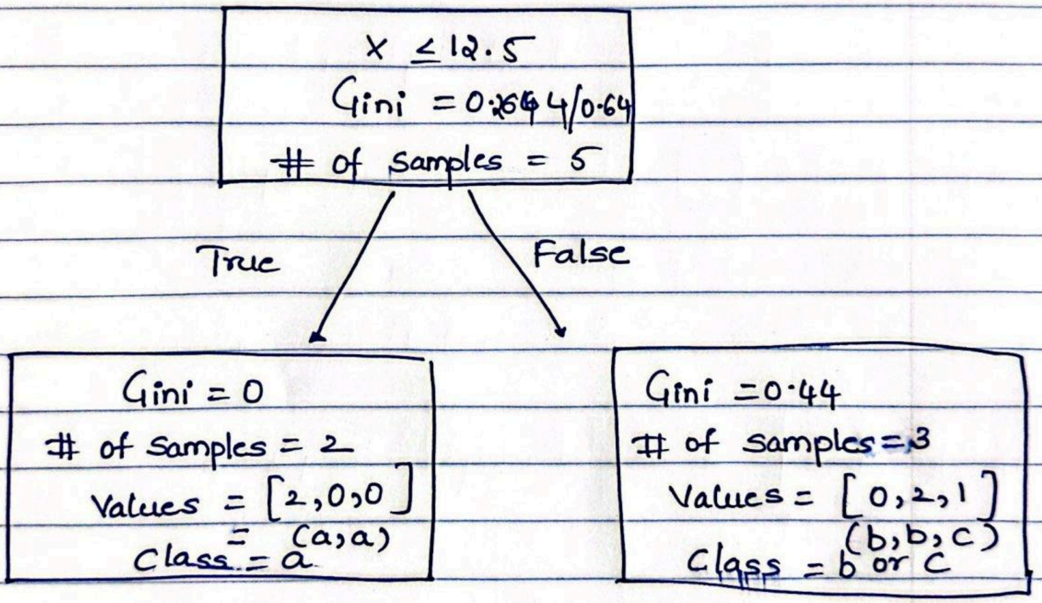
$= 1 - 0 - 0.44 - 0.1$

<u>Gini Impurity at Right Child. $= 0.46$</u>

b) Decision tree:



Decision tree:



References:

1. Class notes: example cart algorithm:
2. https://www.linkedin.com/pulse/decision-tree-cart-algorithms-mathematics-all-behind-algorithm-patel/
3. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/
4. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/