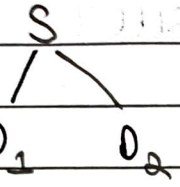
Date: / /

MA5710 - Assignment 03 (Devansh - BE19B002)

Case Study 03 : Supervised learning - Classification

Question One



classification on main training set can be based on either age or salary.

Data set :

Age	salary	Class	Index
30	65	G	1
23	15	B	2
40	75	G	3
55	40	B	4
85	100	G	5
45	60	G	6

(i) Age and salary as attributes.

Re-sorting (Age)

23	2
30	1
40	3
45	6
55	5
58	4

Re-sorting (salary)

15	2
40	4
60	6
65	1
75	3
100	5

(ii) Gini Index (GD) = $\frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$

~~Gini (GD) = Gini (D1) / Gini~~

$Gini(A) = 1 - \sum_{j=1}^{N} p_j^2$ N classes

p_j = frequency of a class

here N=2

(I) Age as attribute

a) age ≤ 23 : $|D_1| = 1$, $|D_2| = 5$

$\therefore Gini(D) = \frac{1}{6} Gini(D_1) + \frac{5}{6}\left(1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right)\right)$

$= \frac{4}{15}$

$= 0.26667$

b) age ≤ 30 : $|D_1| = 2$, $|D_2| = 4$

$Gini(D) = \frac{2}{6}\left(1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right)\right) + \frac{4}{6}\left(1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)\right)$

$= \frac{1}{6} + \frac{1}{4}$

$= 5/12 = 0.41666$

c) Age ≤ 40 : $|D_1| = 3$, $|D_2| = 3$

$Gini(D) = \frac{1}{2}\left(1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)\right) + \frac{1}{2}\left(1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)\right)$

$= \frac{4}{9}$

$= 0.4444$

d) Age ≤ 45 : $|D_1| = 4$, $|D_2| = 2$

$Gini(D) = \frac{2}{3}\left(1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)\right) + \frac{1}{3}\left(1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right)\right)$

$= \frac{1}{4} + \frac{1}{6}$

$= 0.41666$

e) Age ≤ 55 : $|D_1| = 6$, $|D_2| = 0$

$Gini(D) = \frac{6}{6}\left(1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right)\right) + 0$

$= \frac{4}{9} = 0.444$

(II) Salary as attribute

a) salary ≤ 15 : $|D_1| = 1$, $|D_2| = 5$

$$Gini(D) = \frac{1}{6}(0) + \frac{5}{6}\left(1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right)\right)$$

$$= 0.2666$$

b) salary ≤ 40 : $|D_1| = 2$, $|D_2| = 4$

$$Gini(D) = \frac{2}{6}(1 - 1) + \frac{4}{6}(1 - 1)$$

$$= 0$$

c) salary ≤ 60 : $|D_1| = 3$, $|D_2| = 3$

$$Gini(D) = \frac{1}{2}\left(1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right)\right) + \frac{1}{2}(1 - 1)$$

$$= \frac{2}{9} = 0.222$$

d) salary ≤ 65 : $|D_1| = 4$, $|D_2| = 2$

$$Gini(D) = \frac{2}{3}\left(1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right)\right) + 0$$

$$= \frac{1}{3} = 0.333$$

e) salary ≤ 75 : $|D_1| = 5$, $|D_2| = 1$

$$Gini(D) = \frac{5}{6}\left(1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right)\right) + 0$$

$$= \frac{2}{5} = 0.4$$

f) salary ≤ 100 : $|D_1| = 6$, $|D_2| = 0$

$$Gini(D) = 1\left(1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right)\right)$$

$$= \frac{4}{9} = 0.444.$$

Date: / /

Best split among age and salary is salary
- with salary $\leq \left(\frac{40+60}{2}\right)$ which has a Gini index of 0.

salary ≤ 80

(ii) Root partition keeping Age ≤ 35

→ Age ≤ 35 : D = 2 samples.

Age	Salary	Class	Index
23	15	B	2
30	65	G	1

Gini (salary ≤ 15) = $\frac{1}{2}\left(1-(1)^2\right)+\frac{1}{2}(1-(1)^2)$

= 0

∴ **Best split = salary ≤ 15.**

→ Age > 35

Age	Salary	Class	Index
40	75	G	3
48	60	G	6
55	100	G	8
55	40	B	4

→

Salary	Index
40	4
60	6
75	3
100	5

(i) Gini (salary ≤ 40) = $\frac{1}{4}\left(1-(1)^2\right)+\frac{3}{4}\left(1-\left(\frac{1}{3}\right)^2\right)$

= 0

Date: / /

(ii) Gini(salary ≤ 60) = $\frac{2}{4}\left(1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right) + \frac{2}{4}\left(1-(1)^2\right)$

$= \frac{1}{4} = 0.25$

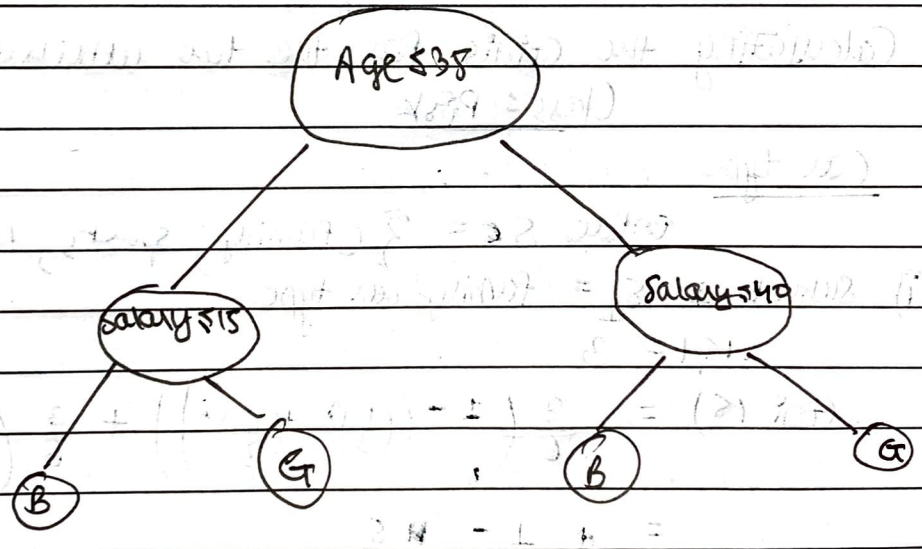
(iii) Gini(salary ≤ 75) = $\frac{3}{4}\left(1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)\right) + \frac{1}{4}\left(1-(1)^2\right)$

$= 1/3 = 0.333$

(iv) Gini(salary ≤ 90) = $\frac{4}{4}\left(1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right) + 0$

$= 3/8 = 0.375$

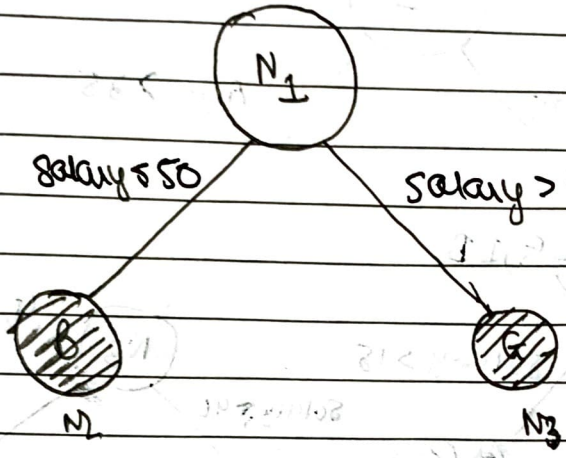
∴ Best split = salary ≤ 60 (at split point = 40)



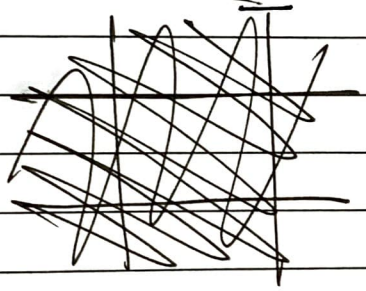
Date: / /

Q3) Accuracy of classifier at every node using histogram.

Ideal Classification Tree



Histogram for N_1 before updating:



	B	G	
L	0	0	→ N_1
R	2	4	

After classification (salary ≤ 50)

	B	G
L	2	0
R	0	4

N_2 :

	B	G
L	–	–
R	2	0

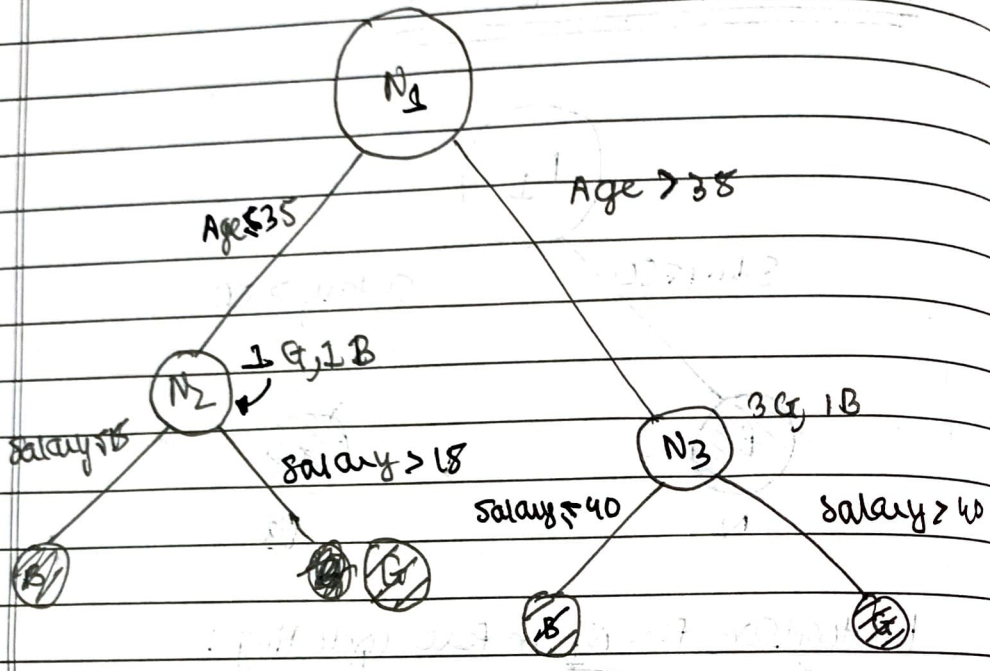
N_3 :

	B	G
L	–	–
R	0	4

Date: / /

Classification Tree using root node

Age ≤ 35



Class histograms

(I) (N_1) after classification:

		B	G
N_2 ←	L	1	1
N_3 ←	R	1	3

Accuracy to get N_2 = $\frac{\text{number of B's in } N_2 \times 100}{\text{expected no of Bs}}$

= $\frac{1}{2} \times 100 = 50\%$

Accuracy to get N_3 = $\frac{\text{number of G's in } N_3 \times 100}{\text{expected number of G}}$

= $\frac{3}{4} \times 100$

= 75%

Accuracy of Nodes N_2, N_3 are 100% since they give the perfect split i.e one class type on the left and other on the right.

When classifying using root node as N_1, we do not get the perfect split as expected using the SLIQ model. The accuracy is calculated for both L and R (i.e how Accurately does it identify B in Left Node N_2 and G in right node N_3.)