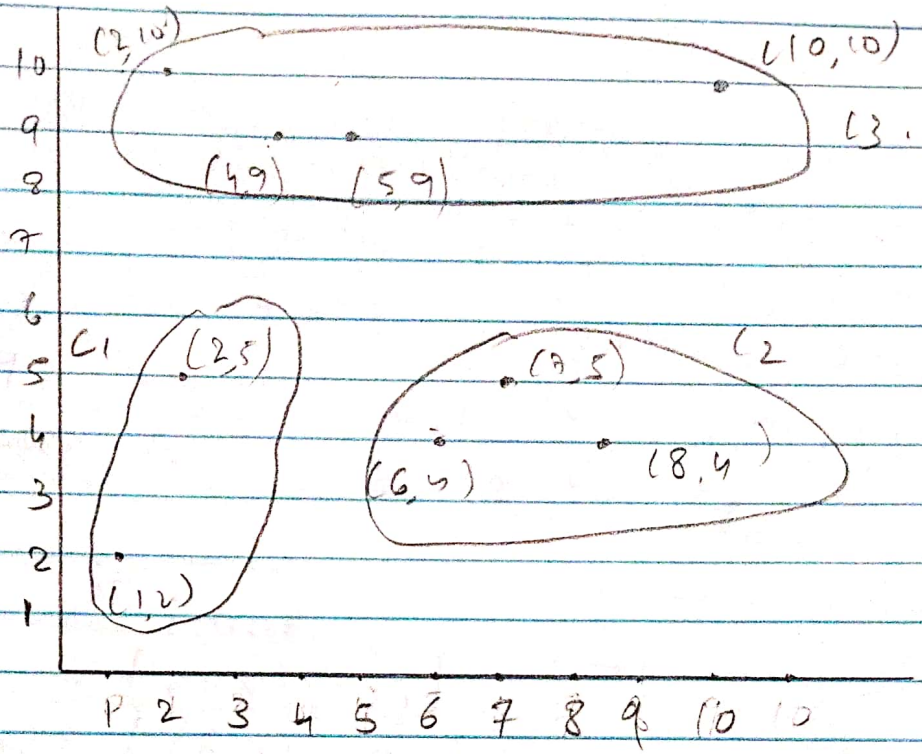
Assignment-4.

Q-A.

1.



2.

Centroids	Points.								
	(2,10)	(2,5)	(8,4)	(5,9)	(7,5)	(6,4)	(1,2)	(4,9)	(10,10)
c_1 (2,5)	5	0	6.08	5	5	4.12	3.16	4.47	9.43
c_2 (5,8)	3.61	4.24	5	1	3.61	4.12	7.21	1.41	5.39
c_3 (4,9)	2.24	4.47	6.40	1	5	5.39	7.62	0	6.08
membership	c_3	c_1	c_2	c_3	c_2	c_2	c_1	c_3	c_3

→ After 1st iteration

$c_1 = \{(2,5), (1,2)\}$

$c_2 = \{(8,4), (7,5), (6,4)\}$

$c_3 = \{(2,10), (5,9), (4,9), (10,10)\}$

New centroids.

$c_1 = \frac{2+1}{2}, \frac{5+2}{2} = (1.5, 3.5)$

$c_2 = \frac{8+7+6}{3}, \frac{4+5+4}{3} = (7, 4.33)$

$c_3 = \frac{2+5+4+10}{4}, \frac{10+9+9+10}{4} = (5.25, 9.5)$.

3. Use cluster created after 1st iteration.

Centroid	Points.								
	(2,10)	(2,5)	(8,4)	(5,9)	(7,5)	(6,4)	(1,2)	(4,9)	(10,10)
(1.5, 3.5)	6.52	1.58	6.52	6.52	5.70	4.53	1.58	6.04	10.7
(7, 4.33)	7.56	5.04	1.05	5.08	0.67	1.05	6.44	5.55	6.41
(5.25, 9.5)	3.29	5.55	6.15	0.56	4.83	5.55	8.62	1.35	4.78
Membership	c_3	c_1	c_2	c_3	c_2	c_2	c_1	c_3	c_3

Since Membership remains unchange.
Final clusters are.

$C_1 = \{(2,5), (1,2)\}$

$C_2 = \{(8,4), (7,5), (6,4)\}$

$C_3 = \{(2,10), (5,9), (4,9), (10,10)\}$

Centroids $(1.5, 3.5), (7, 4.33), (5.25, 9.5)$

B.

	P_1	P_2	P_3	P_4	P_5
P_1	1.00	0.10	0.41	0.55	0.35
P_2	0.10	1.00	0.64	0.47	0.98
P_3	0.41	0.64	1.00	0.44	0.85
P_4	0.55	0.47	0.44	1.00	0.76
P_5	0.35	0.98	0.85	0.76	1.00

Single link hierarchical clustering.

→ From the given matrix we can see that P_2 & P_5 have max. similarity.
Hence merge P_2 & P_5.

	P_1	$P_2 \cup P_5$	P_3	P_4
P_1	1.00	0.35	0.41	0.55
$P_2 \cup P_5$	0.35	1.00	0.85	0.76
P_3	0.41	0.85	1.00	0.44
P_4	0.55	0.76	0.44	1.00

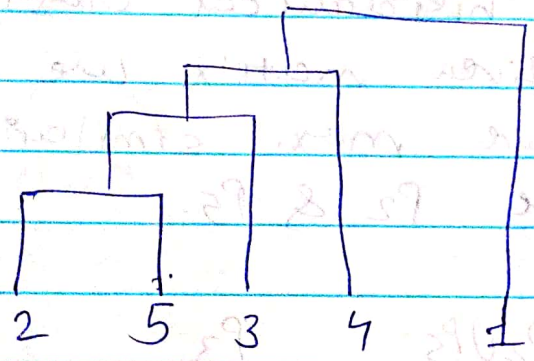
— Now, $P_2 \cup P_5$ & P_3 have max. similarity.
Merge $P_2 \cup P_5$ & P_3.

	P_1	$(P_2 \cup P_5) \cup P_3$	P_4
P_1	1.00	0.41	0.55
$(P_2 \cup P_5) \cup P_3$	0.41	1.00	0.76
P_4	0.55	0.76	1.00

Now $(P_2 \cup P_5 \cup P_3)$ & P_4 has max. Similarity.

	P_1	$P_2 \cup P_5 \cup P_3 \cup P_4$
P_1	1.00	0.55
$P_2 P_5 P_3 P_4$	0.55	1.00.

Final Dendogram will be like



* Complete link hierarchical clustering.

→ From the matrix, P_2 & P_5 has the maximum Similarity. Now merging P_2 & P_5.
Here, similarity matrix will be update with least values.

	P_1	$P_2 \cup P_5$	P_3	P_4
P_1	1.00	0.1	0.41	0.55
$P_2 \cup P_5$	0.1	1.00	0.64	0.47
P_3	0.41	0.64	1.00	0.44
P_4	0.55	0.47	0.44	1.00.

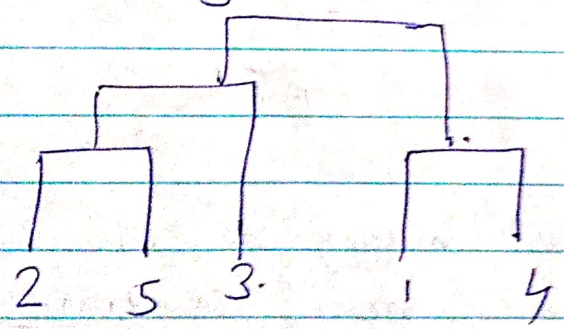
Now, $P_2 \cup P_5$ & P_3 have the max Similarity.

	P_1	$P_2 \cup P_5 \cup P_3$	P_4
P_1	1.00	0.1	0.55
$P_2 \cup P_5 \cup P_3$	0.1	1.00	0.44
P_4	0.55	0.44	1.00

→ Now, P_1 & P_4 has the max Similarity

	$P_1 \cup P_4$	$P_2 \cup P_5 \cup P_3$
$P_1 \cup P_4$	1.00	0.1
$P_2 \cup P_5 \cup P_3$	0.1	1.00

Final Dendogram will be as follows.



c.

Points = {6, 12, 18, 24, ~~30, 42~~, 25, 28, 30, 42, 48}

a. $P_1 = \{5, 7.5\}$

$P_2 = \{15, 25\}$

~~For~~ P_1

	C1 (5)	C2 (7.5)	Cluster
6	1	1.5	C1
12	7	4.5	C2
18	13	10.5	C2
24	19	16.5	C2
25	20	17.5	C2
28	23	20.5	C2
30	25	22.5	C2
42	37	34.5	C2
48	43	40.5	C2

→ Only 1 point belong to C1 & other belongs to C2. Hence new centroids are as ~~follows~~.

$C_1 = 6$

$C_2 = \frac{12+18+24+25+30+28+42+48}{8}$

$= 28.375 \approx 28.38$

Total squared Error.

$(6-6)^2 + (28.38-12)^2 + (28.38-18)^2 + (28.38-24)^2 + (28.38-25)^2 + (28.38-28)^2 + (28.38-30)^2 + (28.38-42)^2 + (28.38-48)^2$

$= 0 + 268.30 + 107.74 + 19.18 + 11.42 + 0.144 + 2.62 + 185.50 + 384.94$

$= SSE = 1 + 20.25 + 110.25 + 272.25 + 306.25 + 420.25 + 506.25 + 1190.25 + 1640.25 = 4467$

Point 2

$P_2 = \{15, 25\}$

	c1 (15)	c2 (25)	cluster
6	9	19	C1
12	3	13	C1
18	3	7	C1
24	9	1	C2
25	10	0	C2
28	13	3	C2
30	15	5	C2
42	27	17	C2
48	33	23	C2.

$C1 = \{6, 12, 18\}$

$C2 = \{24, 25, 28, 30, 42, 48\}$

Update centroid

$C1 = \frac{6 + 12 + 18}{3} = 12$

$C2 = \frac{24 + 25 + 28 + 30 + 42 + 48}{8} = 32.83$

Total square Error.

$(12-6)^2 + (12-12)^2 + (18-12)^2 + (32.83-24)^2 + (32.83-25)^2 + (32.83-28)^2 + (32.83-30)^2 + (42-32.83)^2 + (48-32.83)^2$

$= SSE = 81 + 9 + 9 + 0 + 1 + 9 + 25 + 289 + 529$

$= 99 + 853 = \boxed{952}$

b.

1. for Point $P_1 = \{7.5\}$.

we have calculated new centroids

$C_1 = \{6\}$ $C_2 = 28.38$.

	C1 [6]	C2 [28.38]	cluster
6	0	22.38	C1
12	6	16.38	C1
18	12	10.38	C2
24	18	4.38	C2
25	19	3.38	C2
29	22	0.38	C2
30	24	1.62	C2
42	36	13.62	C2
48	42	19.62	C2

new cluster $C_1 = \{6, 12\}$

$C_2 = \{18, 24, 25, 30, 42, 48\}$.

i. (15, 25)

from set 1 = c_1 =

c_1

New centroid = $6+12/2 = 9$

$c_2 = 30.714$

	$c_1(9)$	$c_2(30.714)$	cluster
6	3	24.714	c_1
12	6	18.714	c_1
18	9	12.714	c_1
24	15	6.714	c_2
25	16	5.714	c_2
28	18	2.714	c_2
30	21	0.714	c_2
42	33	11.286	c_2
48	39	17.286	c_2

New centroids $c_1 = 12$

$c_2 = 32.8$

	$c_1(12)$	$c_2(32.8)$	clusters
6	6	26.8	c_1
12	0	20.8	c_1
18	6	14.8	c_1
24	12	8.8	c_2
25	13	7.8	c_2
28	18	2.8	c_2
30	16	4.8	c_2
42	20	9.2	c_2
48	36	15.2	c_2

Here, they Converged

which is stable solution for Point P1.

→ Now for point P2 (15, 25).

$C_1 = 12$

$C_2 = 32.83$

	C1 (12)	C2 (32.83)	Cluster
6	6	26.83	C1
12	0	20.83	C1
18	6	14.83	C1
24	12	8.83	C2
25	13	7.83	C2
30	8	2.83	C2
42	30	9.17	C2
48	35	15.17	C2

Here After iteration 2 K-means gives same result as iteration which gives stable solution.

c. cluster produced by MIN.

→ points.

6 12 18 24 25 28 30 42 48.

→ Take min distance 1.

6 12 18 (24 25) 28 30 42 48.

→ take min distance 2.

6 12 18 (24 25) (28 30) 42 48.

→ Take min distance 3.

6 12 18 ((24 25) (28 30)) 42 48

Now take min distance 6.

6 (12 (18 ((24 25) (28 30)))) (42 48)

(6 (12 (18 (24 25) (28 30)))) (42 48)

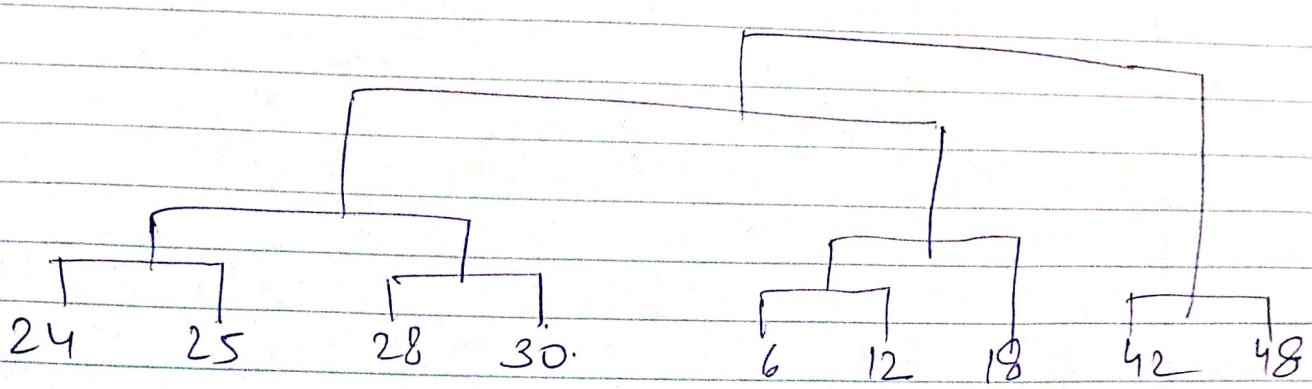
C1 C2

→ Finally 2 clusters are formed

C1 ⇒ {6, 12, 18, 24, 25, 28, 30}

C2 ⇒ {42, 48}.

Dendo gram



d.) MIN, Produces the most natural clustering in given situation

→ MIN.

(e) k-means depend upon the initial center points. If no points could be added to initial center point, then the cluster will be empty with just centroid. k-means is not good option when data points are not separated. It breaks a large cluster as it minimizes squared error and finds new center point.

Part II : classification.

D. Decision Tree.

a ID3 Algorithm.

Class $y \in \{+1, -1\}$.

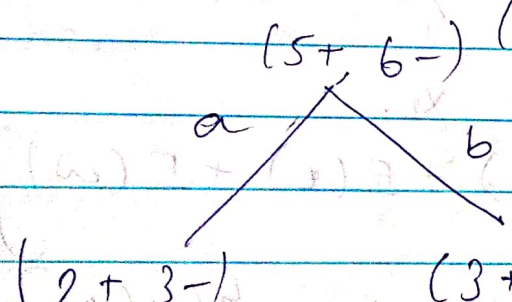
$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

Entropy y =

$$\frac{-5}{11}\log\left(\frac{5}{11}\right) - \frac{6}{11}\log\left(\frac{6}{11}\right)$$

$$= 0.993$$

⇒ Now For X_1,



$$E(a) = -2/5 \log(2/5) - 3/5 \log(3/5)$$

$$= 0.97$$

$$E(b) = \frac{-3}{6}\log(3/6) - 3/6 \log(3/6)$$

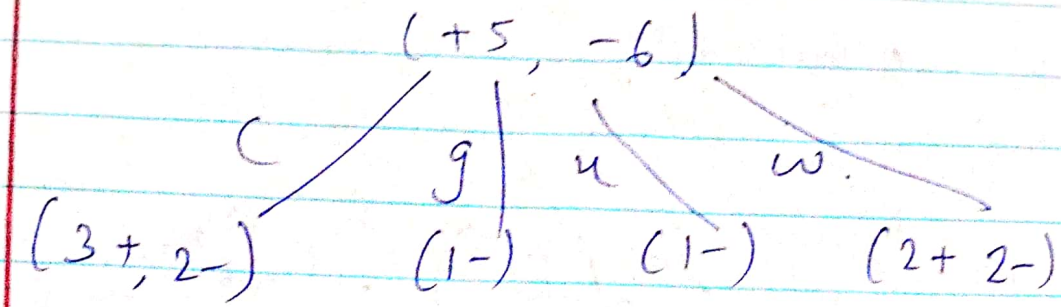
$$= 1$$

Info Gain

$$E(y) - [E(a) + E(b)]$$

$$= 0.993 - \left(\frac{5}{11}(0.97) + \frac{6}{11}(1)\right) = \boxed{0.0663}$$

⇒ Now for X_2.



$E(C) = -3/5 \log(3/5) - 2/5 \log(2/5)$

$= 0.97$

$E(g) = -1 \log(1) = 0$

$E(u) = -1 \log(1) = 0.$

$E(w) = -\frac{2}{4} \log(2/4) - 2/4 \log(2/4) = 1.$

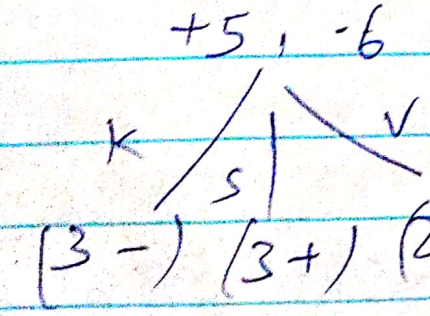
Information Gain ~~for~~ X_2.

$E(Y) - [E(C) + E(g) + E(u) + E(w)]$

$= 0.993 - \left[\frac{5}{11}(0.97) + 0 + 0 + 4/11\,(1)\right]$

$= 0.188.$

⇒ Now for X_3.



$E(K) = 0.$

$E(S) = 0$

$E(V) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right)$

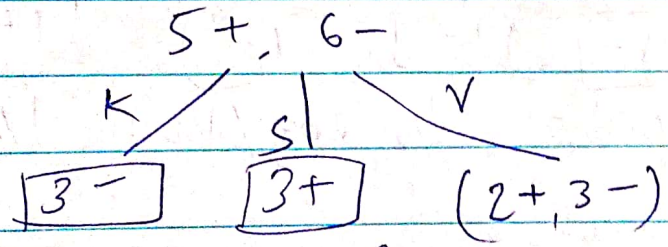
$= 0.970$

— Info gain for X_3.

$0.993 - \left(\frac{5}{11}(0.97)\right)$

$= 0.552.$

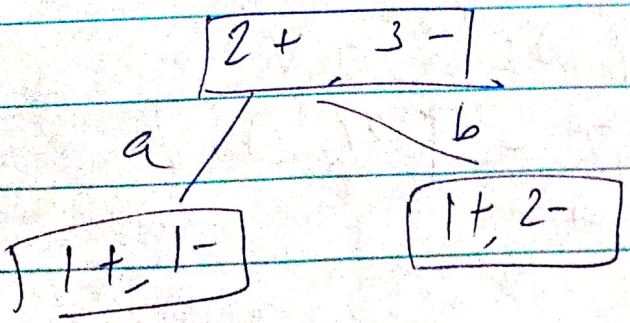
→ Here, X_3 has the highest I.G. among X_1, X_2, X_3. Hence we'll split with X_3.



— Here, K & S are leaf nodes.

→ Now $E(V) = 0.970$.

We split V by X_1.



$E(a) = -1/2 \log(1/2) - 1/2 \log(1/2) = 1$

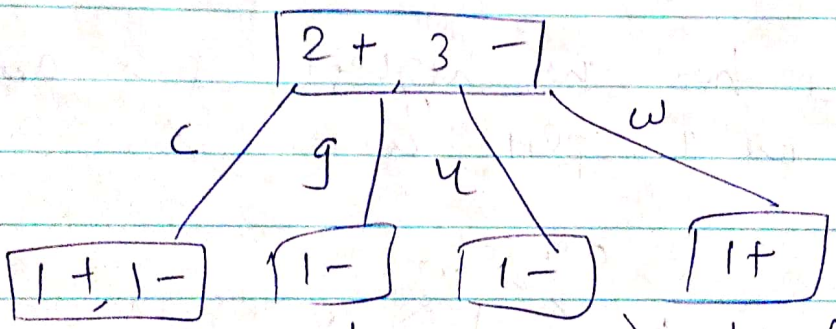
$E(b) = -1/3 \log(1/3) - 2/3 \log(2/3) = 0.917$

I.G. $(X_1 | X_3 = V)$.

$0.917 - \left[\frac{2}{5}(1) + \frac{3}{5}(0.917) \right]$

$= 0.0198$

⇒ Spliting of X_2 when $X_3 = V$.



→ $E(c) = -1/2 \log(1/2) - 1/2 \log(1/2) = 1$

$E(g) = 0$

$E(u) = 0$

$E(w) = 0$.

$I.G = (X_2 | X_3 = V)$

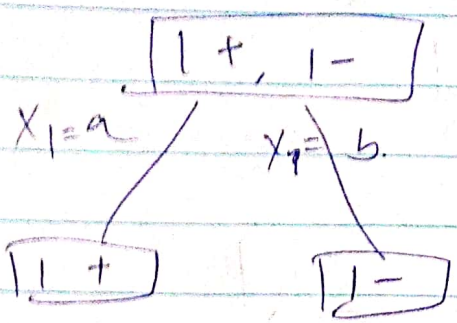
$= 0.917 - [2/5 (1)]$

$= 0.57$

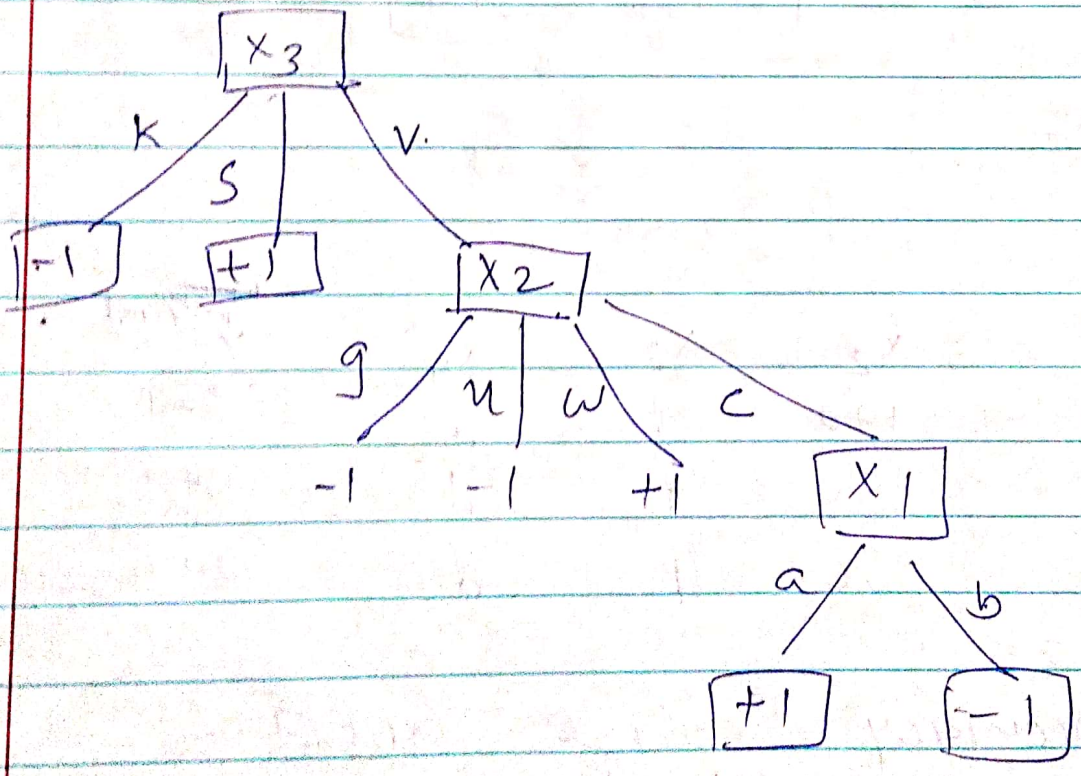
→ I.G. for $X_2 > X_1$. So we split X_2.

Node g, u, w are leaf nodes so we split c of $X_1 = a, b$

$X_2 = c$ $X_3 = v$.

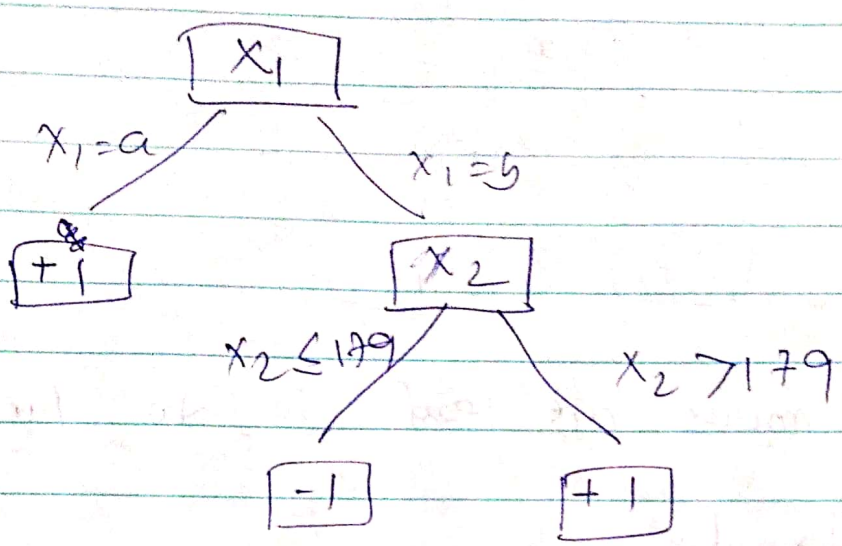


→ All nodes are leaf. So No further split



b. Decision Tree with 2 attrb. 3 leaf class nodes 100 -/. accuracy.

(i)



(ii)

X_1	X_2	X_3	X_4	Actual Y.	predicted o/p.
b	170	f	d	-1	-1
a	150	f	d	+1	+1
b	60	f	d	+1	-1.

Accuracy = 2/3 = 66.67 /.