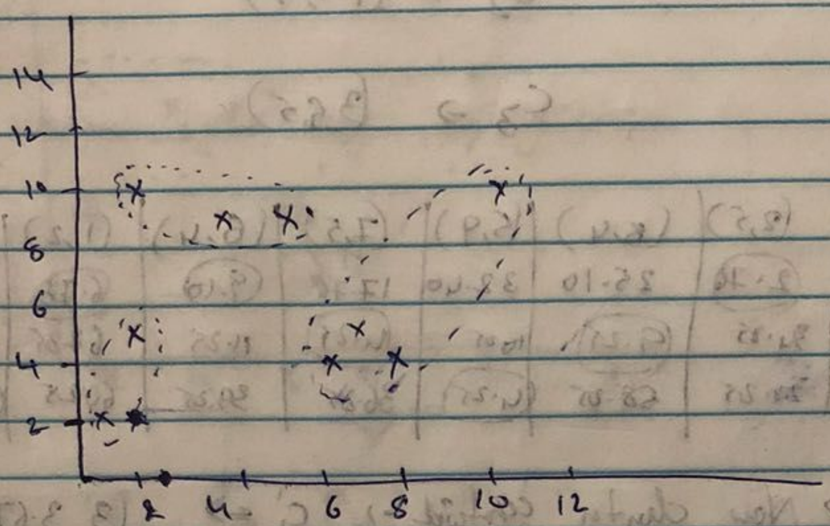
CS 6350 Big Data Analytics & Management

Assign-4

Sashidhar D
sxd173730

Part I: Clustering: Points: (2,10), (2,5), (8,4), (5,9), (7,5), (6,4), (1,2), (4,9), (10,10)

(i)



Plot data to see appropriate Clusters.

(ii) Beginning with (2,5), (5,8), (4,9) as initial cluster centeroids.

	(2,5)	(5,8)	(4,9)
(2,10)	$\sqrt{25}$	$\sqrt{13}$	$\sqrt{5}$
(2,5)	$\sqrt{0}$	$\sqrt{18}$	$\sqrt{20}$
(8,4)	$\sqrt{37}$	$\sqrt{25}$	$\sqrt{41}$
(5,9)	$\sqrt{25}$	$\sqrt{1}$	$\sqrt{1}$
(7,5)	$\sqrt{25}$	$\sqrt{13}$	$\sqrt{25}$
(6,4)	$\sqrt{17}$	$\sqrt{17}$	$\sqrt{29}$
(1,2)	$\sqrt{10}$	$\sqrt{52}$	$\sqrt{58}$
(4,9)	$\sqrt{20}$	$\sqrt{2}$	$\sqrt{0}$
(10,10)	$\sqrt{89}$	$\sqrt{29}$	$\sqrt{37}$

$\Rightarrow$ $C_1 \rightarrow$ (2,5), (6,4), (1,2)

$C_2 \rightarrow$ (8,4), (5,9), (7,5), (10,10)

$C_3 \rightarrow$ (2,10), (4,9)

⇒ New cluster centroids :-

$C_1 \Rightarrow (9/3, 11/3) = (3, 11/3)$

$C_2 \rightarrow (7.5, 7)$

$C_3 \rightarrow (3, 9.5)$

	(2,10)	(2,5)	(8,4)	(5,9)	(7,5)	(6,4)	(1,2)	(4,9)	(10,10)
(3,3.67) C_1	41.06	2.76	25.10	32.40	17.76	9.10	6.78	29.40	89.60
(7.5,7) C_2	39.25	34.25	9.25	10.25	4.25	11.25	67.25	16.25	15.25
(3,9.5) C_3	1.25	21.25	58.25	4.25	36.25	39.25	60.25	1.25	49.25

New clusters:-

$C_1' \rightarrow$ (2,5), (6,4), (1,2)

$C_2' \rightarrow$ (8,4), (7,5), (10,10)

$C_3' \rightarrow$ (2,10), (5,9), (4,9)

⇒ New cluster centroid → $C_1' \Rightarrow (3, 3.67)$

$C_2' \rightarrow (8.33, 6.33)$

$C_3' \rightarrow (3.67, 9.33)$

⇒

	(2,10)	(2,5)	(8,4)	(5,9)	(7,5)	(6,4)	(1,2)	(4,9)	(10,10)
C_1'	41.06	2.76	25.10	32.40	17.76	9.10	6.78	29.4	84.06
C_2'	53.33	41.03	6.53	18.21	3.53	10.85	72.47	25.87	16.25
C_3'	3.24	21.5	47.13	1.87	29.8	33.54	60.85	.217	48.52

New clusters :- $C_1'' = (2,5), (6,4), (1,2)$

$C_2'' = (8,4), (7,5), (10,10)$

$C_3'' = (2,10), (5,9), (4,9)$

New centroids :-

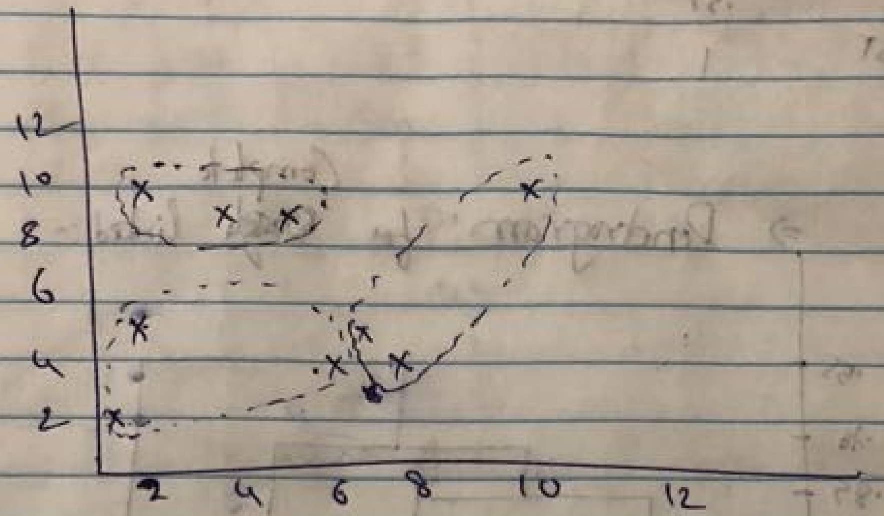
⇒ $C_1'' \rightarrow (3, 3.67)$

$C_2'' \rightarrow (8.33, 6.33)$

$C_3'' \Rightarrow (3.67, 9.33)$

∴ Converged with C_1'', C_2'', C_3'' as final clusters

⇒ Find cluster:



B)

	P_1	P_2	P_3	P_4	P_5
P_1	1				
P_2	.10	1			
P_3	.41	.64	1		
P_4	.55	.47	.44	1	
P_5	.35	.98	.85	.76	1

Complete Link hierarchical clustering:-

⇒

	P_1	P_{25}	P_3	P_4
P_1	1			
P_{25}	.35	1		
P_3	.41	.85	1	
P_4	.55	.76	.44	1

$\max(P_{25}, P_1)$
$\Rightarrow \max((P_1, P_2), (P_1, P_5))$

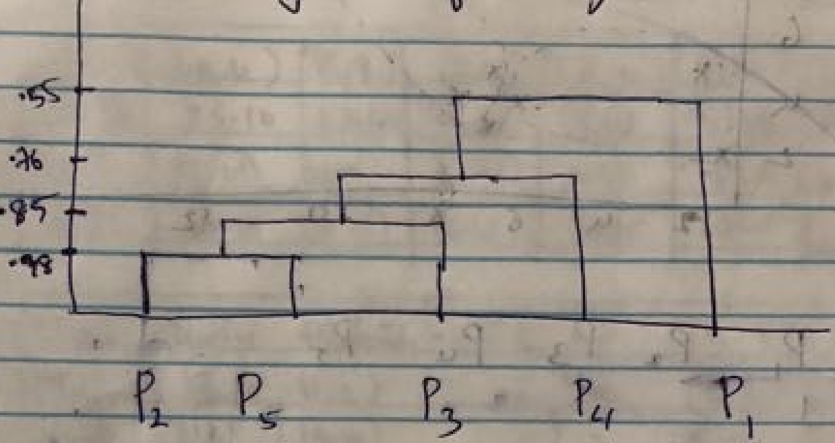
⇒

	P_1	P_{253}	P_4
P_1	1		
P_{235}	.41	1	
P_4	.55	.76	1

$\max(P_{235}, P_1)$
$\Rightarrow \max((P_{25}, P_1), (P_3, P_1))$

	P_1	P_{2345}
P_1	1	.55
P_{2345}	.55	1

⇒ Dendrogram for ~~Single~~ Complete linked :-



Single ~~Complete~~ linked hierarchical clustering :-

	P_1	P_2	P_3	P_4	P_5
P_1	1	.10	.41	.55	.35
P_2		1	.64	.47	.98
P_3			1	.44	.85
P_4				1	.76
P_5					1

⇒

	P_1	P_{25}	P_3	P_4
P_1	1	.10	.41	.55
P_{25}		1	.64	.47
P_3			1	.44
P_4				1

$\min(P_{25}, P_1)$

$\Rightarrow \min((P_2, P_1), (P_5, P_1))$

$= \min(.10, .35)$

$\min(P_{25}, P_4)$

$= \min((P_2, P_4), (P_5, P_4))$

$\min(P_{25}, P_3)$

$= \min((P_2, P_3), (P_5, P_3))$

⇒

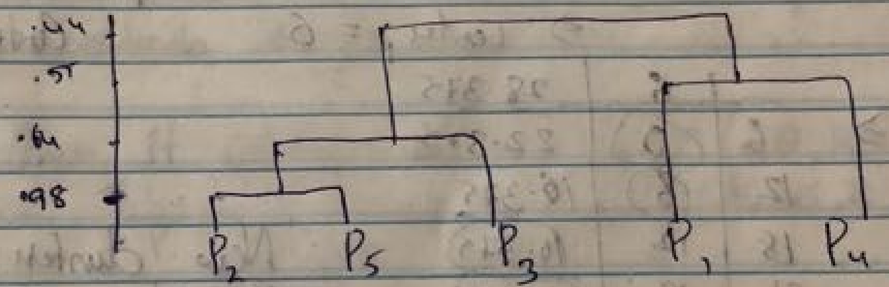
	P_1	P_{235}	P_4
P_1	1		
P_{235}	.70	1	
P_4	.55	.44	1

~~match~~

⇒

	P_{14}	P_{235}
P_{14}	1	.44
P_{235}	.44	1

⇒ ~~Complete~~ dendrogram fo single linked:



Q(c) One-D data points: {6, 12, 18, 24, 25, 28, 30, 32, 48}.

(a) (i) {5, 7.5} → assume 2 cases of centroid of clusters

⇒ $C_1 \to (5) \to 6$

$C_2 \to (7.5) \to 12, 18, 24, 25, 28, 30, 42, 48$

Individual TSE ⇒ $C_1 = 1$

$C_2 = (7.5-12)^2 + (7.5-18)^2 \cdots + (7.5-48)^2$

$= 4466$

(Total) TSE ⇒ 4467

② $\{15, 25\}$

$C_1 \Rightarrow 15 \rightarrow 6, 12, 18$

$C_2 \Rightarrow 25 \rightarrow 24, 30, 28, 42, 48$

$TSE_{C_1} = \left(\sum (15 - c_i)^2\right) +$

$TSE_{C_2} = \left(\sum 25 - c_{2i}\right)^2$

$TSE = 952$

③ (b) For Set ① $\Rightarrow$ $C_1 \rightarrow 6$ | $C_2 \rightarrow 7.5$ $12, 18, 24, 28, 25, 30, 42, 48$

$\Rightarrow$ $Center_1 = 6$ $Center_2 = 28.375$

	6	28.375
6	0	22.375
12	6	16.375
18	12	10.375
24	18	4.375
25	19	3.375
28	22	.375
30	24	1.625
42	36	13.625
48	42	19.625

New clusters:-

(b) $C_1' \rightarrow 6, 12$

$Center_1' = 9$

28.375 $C_2' \rightarrow 18, 24, 25, 30, 42, 48$

$Center_2' = 30.714.$

	9	30.714
6	3	24.714
12	6	18.714
18	9	12.714
24	15	6.714
25	16	5.714
28	18	2.714
30	21	.714
42	33	11.286
48	39	17.286

New cluster:

$9 \rightarrow C_1'' \rightarrow 6, 18, 12$

$30.714 \rightarrow C_2'' \rightarrow 24, 25, 28, 30, 42, 48$

$Center_1'' = 12$

$Center_2'' = 32.8$

	12	32.8
6	6	26.8
12	0	20.8
18	6	26.8
24	12	8.8
25	13	7.8
30	18	2.8
28	16	4.8
42	20	9.2
48	36	15.2

New cluster:-

$C_1''' \to (6, 12, 18)$

$C_2''' \to (24, 25, 28, 42, 30, 48)$

~~Hence~~

Here, $C_1'' \equiv C_1'''$

$C_2'' \equiv C_2'''$

∴ Hence converged.

Now consider Set (2):- and repeat above procedure with initial centroids at {15, 25}

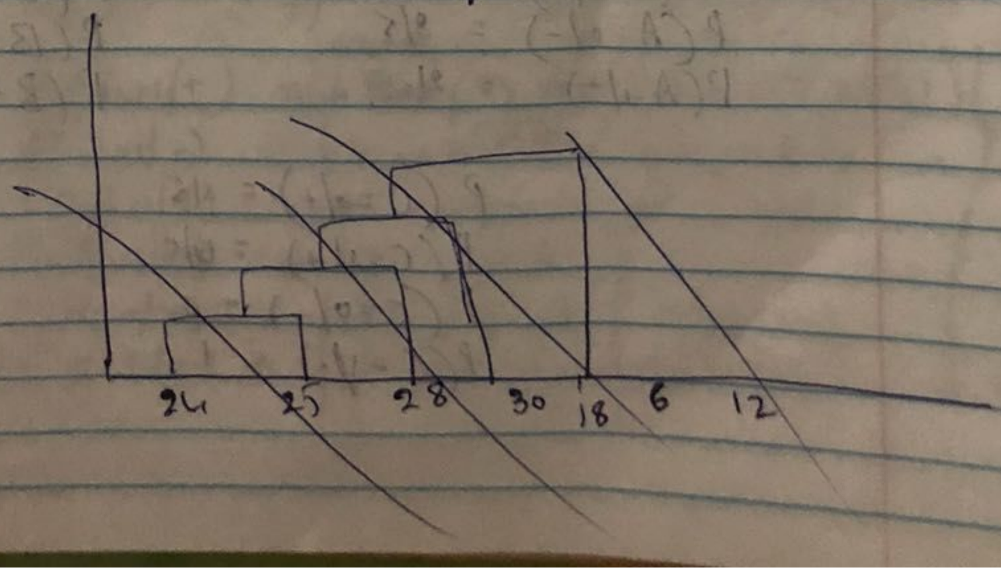
Then, it converges at ⇒ $C_1''' = (6, 12, 18) \to 12$

$C_2''' = (24, 25, 28, 30, 42, 48) \to 32.8$

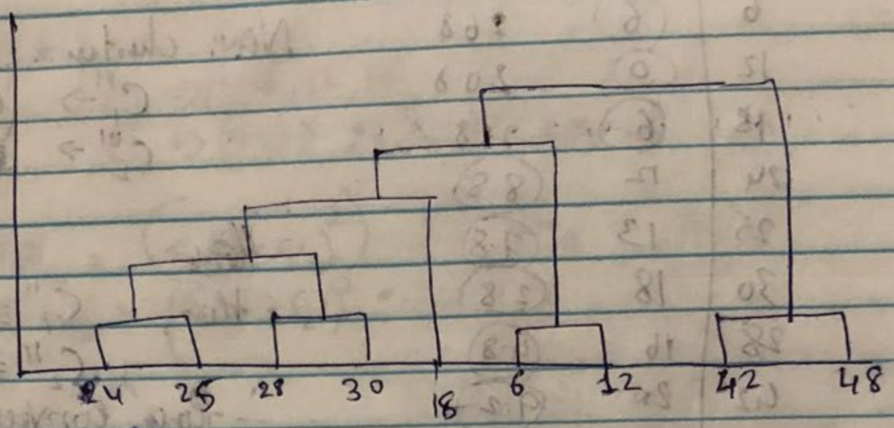
∴ Both are stable solutions since they converged and also have same clustering.

(c) Two clusters produced by MIN (single link clustering)

⇒



MIN
single Linked ⇒



(d) MIN gives most natural clustering than K-means in this situation.

(e) K-means clustering depends on selection of initial centroids.

Part II :- Classification:-

(i) ① $P(A=0|+) = 2/5$
$P(A=1|+) = 3/5$
$P(A=0|-) = 3/5$
$P(A=1|-) = 2/5$

$P(B=0|+) = 4/5$
$P(B=1|+) = 1/5$
$P(B=0|-) = 3/5$
$P(B=1|-) = 2/5$

$P(C=0|+) = 1/5$
$P(C=1|+) = 4/5$
$P(C=0|-) = 0$
$P(C=1|-) = 1$

② Using Naive Bayes Calculate $P(A=1, B=1, C=0)$

$$P(+/A=1,B=1,C=0) = \frac{P(A=1, B=1, C=0/+)\ P(+)}{P(A=1, B=1, C=0)}$$

$$= \frac{P(A=1/+)\ P(B=1/+)\ P(C=0/+)\ P(+)}{K}$$

$$= (3/5)(1/5)(1/5)(1/2)/K = \frac{3}{250K} \quad (A)$$

$$P(-/A=1,B=1,C=0) = \frac{P(A=1, B=1, C=0/-)\ P(-)}{K}$$

$$= \frac{P(A=1/-)\ P(B=1/-)\ P(C=0/-)\ P(-)}{K}$$

$$= (2/5)(4/5)(0)(1/2)/K = 0 \quad (B)$$

$\therefore$ (A) > (B) ⇒ The class of sample $P(A=1, B=1, C=0)$ is (+)

③ Using m-estimate approach ⇒ $p = 1/2$, $m = 4$

$P(A=1/+)$	apply $\left(\frac{n_c + mp}{n+m}\right)$	$= \frac{3 + (4)(1/2)}{5+4}$		$5/9$
$P(A=1/-)$				$4/9$
$P(B=1/+)$				$3/9$
$P(B=1/-)$				$4/9$
$P(C=0/+)$				$3/9$
$P(C=0/-)$				$2/9$

④ repeat part ③ :-

$P(A=1, B=1, C=0) \Rightarrow$ which class?

$\Rightarrow P(+|A=1, B=1, C=0) = \dfrac{P(+)\left(P(A=1|+)\,P(B=1|+)\,P(C=0|+)\right)}{P(A=1, B=1, C=0)}$

$= (5/9)(3/9)(3/9)(1/2)/K = \dfrac{0.301}{K}$ (A)

$\Rightarrow P(-|A=1, B=1, C=0) = \dfrac{P(-)\left(P(A=1|-)\,P(B=1|-)\,P(C|-)\right)}{K}$

$= (4/9)(4/9)(2/4)(1/2)/K$

$= .219/K$ (B)

$\therefore \Rightarrow$ (A) > (B) The classification class of sample is +

⑤ Considering above two cases, when the conditional probability of one of the class is zero, without smoothing the contribution of all other probabilities is not taken into consideration.

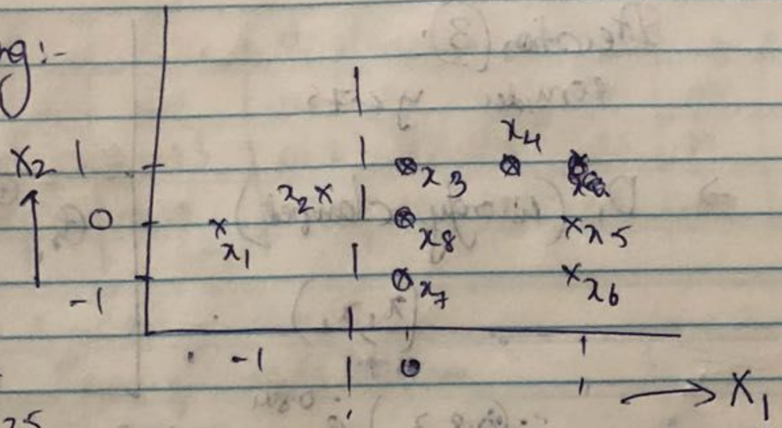
So, then the contribution of other probabilities contribute more to the given class, than the zero prob. value.

Hence, it is better to consider all probabilities by smoothing because it gives minimal probability weight to all the classes.

(iii) Adaboosting:-

Iteration (1)

$\Downarrow$



$w_1 \to w_8$

$\Rightarrow w_i = 1/8$

$= .125$

let $x_1 \geq -0.5 \Rightarrow$ 2 points (x_5, x_6) are misclassified.

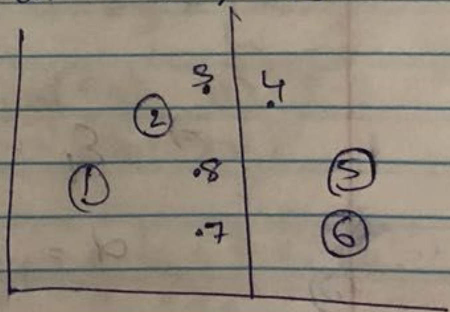
$\Rightarrow \varepsilon_1 = 2 \times 1/8 = 0.25$

$\alpha = \frac{1}{2} \ln \left(\frac{1-0.25}{0.25} \right) = \ln\sqrt{3} = .55$

$Z_1 = 2\sqrt{(.25)(1-0.25)} = .866$

Iteration (2) :- let boundary be at $x \geq .75$

2 points are misclassified (x_5, x_6)



$D_2 = \frac{(1/8)\, e^{0.55}}{.866} = .25$

weights of $x_1, x_2, x_3, x_4, x_7, x_8$ are correctly classified

$\Rightarrow D_2 = \frac{(1/8)\, e^{-0.55}}{.866} = .083$

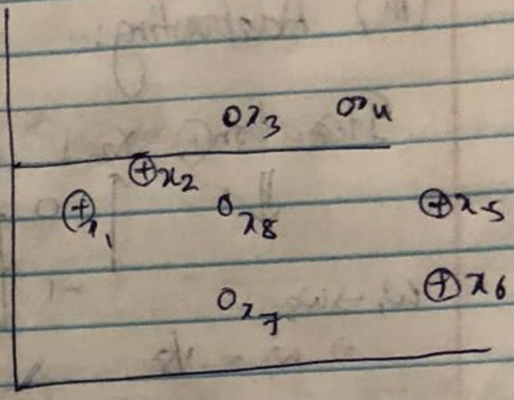
$\varepsilon_2 = 0.083 \times 2 = .167$

$\alpha = \frac{1}{2} \ln \left(\frac{1-\varepsilon_2}{\varepsilon_2} \right) = .804$

$Z_2 = .746$

Iteration (3):

consider $y < .75$



$\Rightarrow$ D_1 (wrongly classified) (x_1, x_2)

$$= \frac{(0.83)\, e^{-0.804}}{.746} = .249$$

D_2 (correctly classified) $\Rightarrow (x_3, x_4, x_8, x_7)$

$$= \frac{(0.083)\, e^{-.804}}{.746} = .050$$

D_3 (correctly classified) $= (x_5, x_6)$

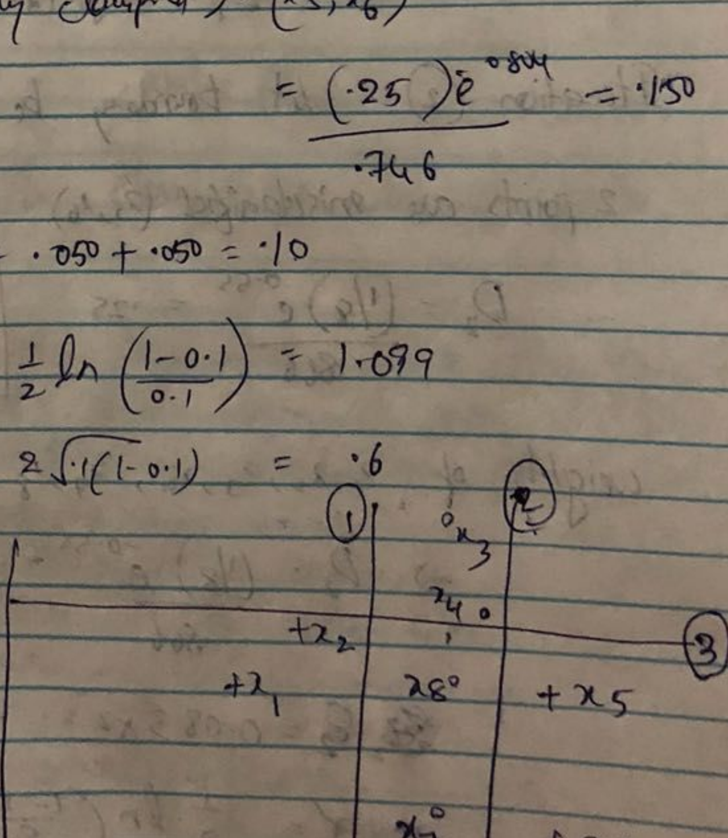
$$= \frac{(.25)\, e^{0.804}}{.746} = .150$$

$\Rightarrow \quad \varepsilon_t = .050 + .050 = .10$

$$\alpha = \frac{1}{2} \ln\left(\frac{1-0.1}{0.1}\right) = 1.099$$

$$Z_3 = 2\sqrt{.1(1-0.1)} = .6$$

(d) Decision Stumps:



$\Rightarrow$

t	ε_t	α_t	Z_t	$w_t(1)$	w_2	w_3	w_4	w_5	w_6	w_7	w_8
1	0.25	.55	.866	.125	.125	.125	.125	.125	.125	.125	.125
2	.167	.804	.746	.083	.083	.083	.083	.25	.25	.083	.083
3	.10	1.099	.6	.249	.249	.050	.050	.150	.050	.050	.050

(2) Training Error of adaboost :-

Adaboost outperform a single decision stump because when single decision stump is used, the training error is ~~[illegible]~~ mostly greater than zero and also has high variance which is not the same using adaboost.

Q(5) $X_1 \in \{a,b\}$ $X_2 \in \{c,g,u,w\}$ $X_3 \in \{k,sv\}$

parent entropy $\Rightarrow -\left(\Sigma\left(\frac{ny_i}{ny_{1i}+ny_{2i}}\right)\log\left(\frac{ny_i}{ny_{1i}+ny_{2i}}\right)\right)$

$$= -\left(\frac{6}{11}\log\frac{6}{11} + \frac{5}{11}\log\frac{5}{11}\right) = .994$$

(1) Now consider sorting on basis of X_1

$$\Rightarrow \text{entropy}(S_a) = -\left(\frac{3}{5}\log 3/5 + 2/5 \log 2/5\right)$$

$$= 0.97$$

$$S_b = -\left(3/6 \log 3/6 + 3/6 \log 3/6\right)$$

$$= 1$$

$$IG = .994 - \left(\left(\frac{5}{11}\right)(.97) + \left(\frac{6}{11}\right)1\right) = .0076$$

② Sort on basis of X_2:

$$S_C = -\left(\frac{3}{5}\log 3/5 + \frac{2}{5}\log 2/5\right)$$

$$= 0.97$$

$$S_{00} = -\left(\frac{2}{4}\log 2/4 + \frac{2}{4}\log 2/4\right)$$

$$= -\left(\log 1/4\right) = \log 4.$$

$$S_g = -\left(1\cdot\log 1\right) = 0, \quad S_4 = 0.$$

$$IG = .994 - \left((.97)\left(\frac{5}{11}\right) + (1.386)\left(\frac{4}{11}\right)\right)$$

$$= .994 - \left(\frac{(4.85) + (5.545)}{11}\right)$$

$$= .601$$

③ Sort on basis of X_3:

$$S_K = -\left(\frac{3}{3}\log 3\right) = 0$$

$$S_V = -\left(\frac{2}{5}\log 2/5 + 3/5\log 3/5\right) = \frac{-1}{5}(\log 36)$$

$$= 3.583$$

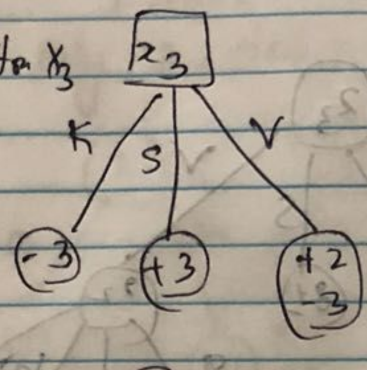
$$S_S = 0$$

$$I_G = .994 - \left(\left(\frac{5}{11}\right)(3.583)\right) = 1.6288$$

$$I_G = .634$$

IG of (3) is highest

$\Rightarrow$ split on x_3



At Node (3) $\Rightarrow$ entropy $= -\frac{2}{5}\log\frac{2}{5} + -\frac{3}{5}\log 3/5$

(parent now)

$= .97$

based on x_1

$S_a = -\frac{1}{2}\log 1/2 - \left(\log 1/2\right)\frac{1}{2} = 1$

$S_b = -\frac{2}{3}\log 2/3 - \frac{1}{3}\log 1/3 = .918$

$IG = .97 - \left(\frac{2}{5}(1) + \frac{3}{5}(.918)\right)$

$= 0.192$

based on x_2

$S_c = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log 1/2 = 1$

$S_g = -\frac{1}{1}\log 1/1 = 0$

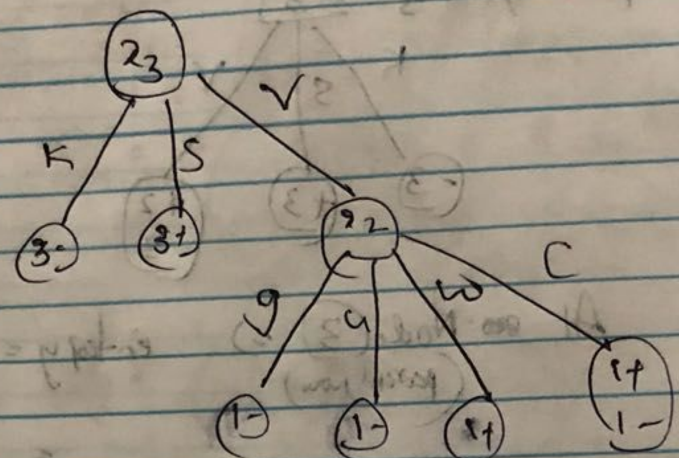
$S_u = 0$

$S_w = 0$

$IG = .97 - \left(\frac{2}{5}\times 1 + \frac{1}{5}\times 0 + 0 + 0\right) = .57$

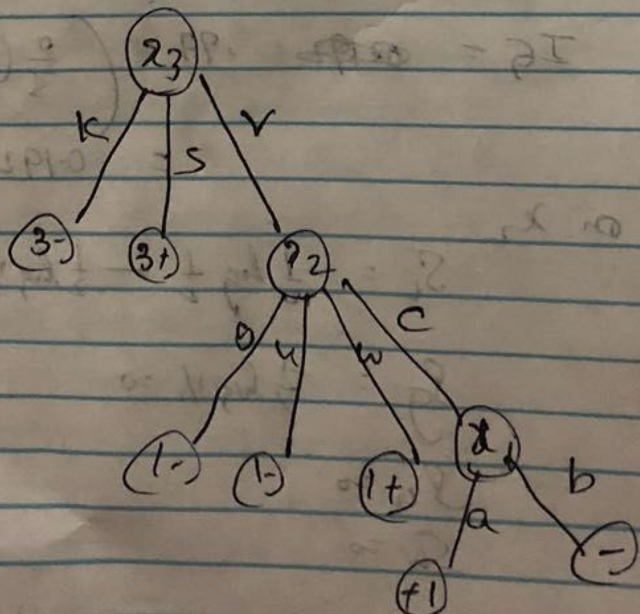
Build stump based on x_2 ($IG_{x_1} < IG_{x_2}$)

$\Rightarrow$



$x_2 \rightarrow$ new parent
but no need to calculate IG because
we are left with x_1 only

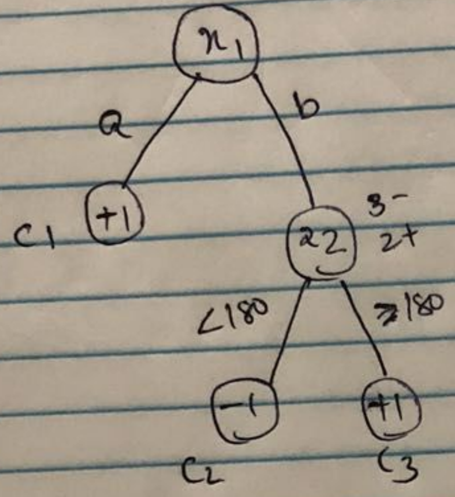
$\Rightarrow$ Find Decision Tree:-



(b) For given data,

Decision Tree:-

(i) only 2 attribute node and 3 leaf nodes with 100% accuracy.



$c_i \rightarrow$ class nodes

$x_i \rightarrow$ attribute nodes.

(ii) Calculating accuracy of test sample using above model:-

$\boxed{\text{From model}}$

$\Rightarrow$ $x_1 = b$, $x_2 = 170$ expected value $\rightarrow -1$

Obtained value $\Rightarrow$ $Y = -1$ ✓

$x_1 = a$, $x_2 = 150$ expected value $\rightarrow +1$

Obtained $\Rightarrow Y = +1$ ✓

$x_1 = b$, $x_2 = 60$ expected value $\Rightarrow +1$

Obtained $\Rightarrow Y = -1$ X

Accuracy $= 2/3 = .667$

$\Rightarrow 66.7\%$