(ii) Maxi...

Let's see a few ex...

...ice Questions

...1.1 : Use the following data and group them using K-means clustering algorithm. Show calculation of centroids.

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Iteration 1 :

Let's randomly choose two centroids.

Centroid 1 = (170, 56) and Centroid 2 = (182, 72).

Now, let's calculate the distance of the data points from the chosen centroids and complete the data points table.

The data point is assigned to the cluster based on the closest centroid.

A sample distance calculation is as following for the first data point.

Distance from Centroid 1 (170, 56) for (185, 72) =

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$d = \sqrt{(170 - 185)^2 + (56 - 72)^2}$$

$$d = 21.93$$

Height	Weight	Distance from Centroid 1	Distance from Centroid 2	Assigned Cluster
185	72	21.93	3.00	2
170	56	0.00	20.00	1
168	60	4.47	18.44	1
179	68	15.00	5.00	2
182	72	20.00	0.00	2
188	77	27.66	7.81	2
180	71	18.03	2.24	2
180	70	17.20	2.83	2
183	84	30.87	12.04	2
180	88	33.53	16.12	2
180	67	14.87	5.39	2
177	76	21.19	6.40	2

Let's re-calculate the centroids for the next iteration.

For Centroid 1 calculation, you have two data points that fall in cluster 1. They are (170, 56) and (168, 60).

Hence, Centroid 1 is the mean of the data points which is $\left(\frac{170 + 168}{2}, \frac{56 + 60}{2}\right)$ = (169, 58).

For Centroid 2 calculation, take the remaining 10 data points that fall in cluster 2.

Hence, Centroid 2 is the mean of these 10 data points which is

$$\left(\frac{185 + 179 + 182 + 188 + 180 + 180 + 183 + 180 + 180 + 177}{10}, \frac{72 + 68 + 72 + 77 + 71 + 70 + 84 + 88 + 67 + 76}{10}\right)$$

= (181.4, 74.5)

Now, you have both the centroids ready for the next and final iteration.

Iteration 2 :

Centroid 1 = (169, 58) and Centroid 2 = (181.4, 74.5).

Now, let's calculate the distance of the data points from the centroids and complete the data points table. The data point is assigned to the cluster based on the closest centroid.

Height	Weight	Distance from Centroid 1	Distance from Centroid 2	Assigned Cluster
185	72	21.26	4.38	2
170	56	2.24	21.73	1
168	60	2.24	19.74	1
179	68	14.14	6.93	2
182	72	19.10	2.57	2
188	77	26.87	7.06	2
180	71	17.03	3.77	2
180	70	16.28	4.71	2
183	84	29.53	9.63	2
180	88	31.95	13.57	2
180	67	14.21	7.63	2
177	76	19.70	4.65	2

You stop here because the number of iterations that you decided are completed and also you see that the clusters assigned in the iteration 1 for the data points did not change.

Hence, you got the two clusters as shown in Fig. P. 5.1.1(a).

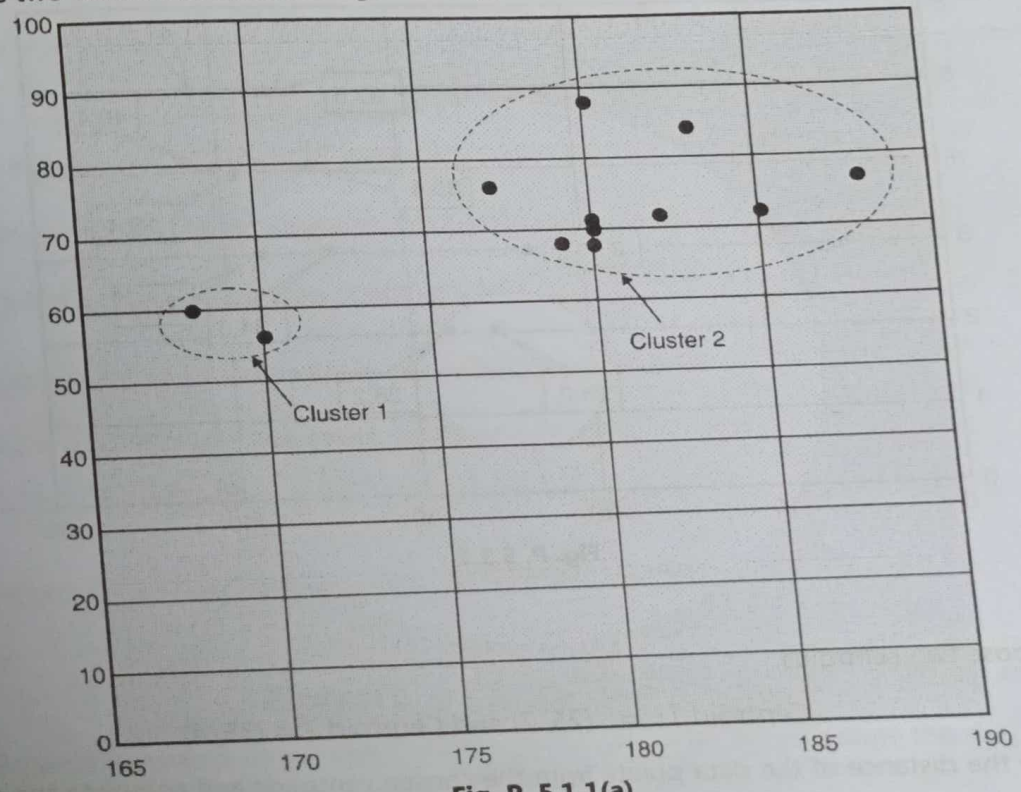


Fig. P. 5.1.1(a)

P1: (3, 7)
P2: (4, 6)
P3: (5, 5)
P4: (6, 4)
P5: (7, 3)
P6: (6, 2)
P7: (7, 2)
P8: (8, 4)
P9: (3, 3)
P10: (2, 6)
P11: (3, 5)
P12: (2, 4)

minPts = 4 and epsilon (ε) = 1.9

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1	0											
P2	1.41	0										
P3	2.83	1.41	0									
P4	4.24	2.83	1.41	0								
P5	5.66	4.24	2.83	1.41	0							
P6	5.83	4.47	3.16	2.00	1.41	0						
P7	6.40	5.00	3.61	2.24	1.00	1.00	0					
P8	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P9	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P10	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P11	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P12	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.41	0

DBSCAN Clustering Algorithm Solved Example – 1

minPts = 4 and epsilon (ε) = 1.9

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1	0											
P2	1.41	0										
P3	2.83	1.41	0									
P4	4.24	2.83	1.41	0								
P5	5.66	4.24	2.83	1.41	0							
P6	5.83	4.47	3.16	2.00	1.41	0						
P7	6.40	5.00	3.61	2.24	1.00	1.00	0					
P8	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P9	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P10	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P11	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P12	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.41	0

P1: P2, P10
P2: P1, P3, P11
P3: P2, P4
P4: P3, P5
P5: P4, P6, P7, P8
P6: P5, P7
P7: P5, P6
P8: P5
P9: P12
P10: P1, P11
P11: P2, P10, P12
P12: P9, P11

Like, Share and Subscribe to Mahesh Huddar Visit: vtupulse.com

P1 is in someone's core point set. then its going to be border

P2 is core itself

P3 is present in someone's core. hence border

P9 is not in anyone's core. hence not a border

P1: P2, P10

P2: P1, P3, P11

P3: P2, P4

P4: P3, P5

P5: P4, P6, P7, P8

P6: P5, P7

P7: P5, P6

P8: P5

P9: P12

P10: P1, P11

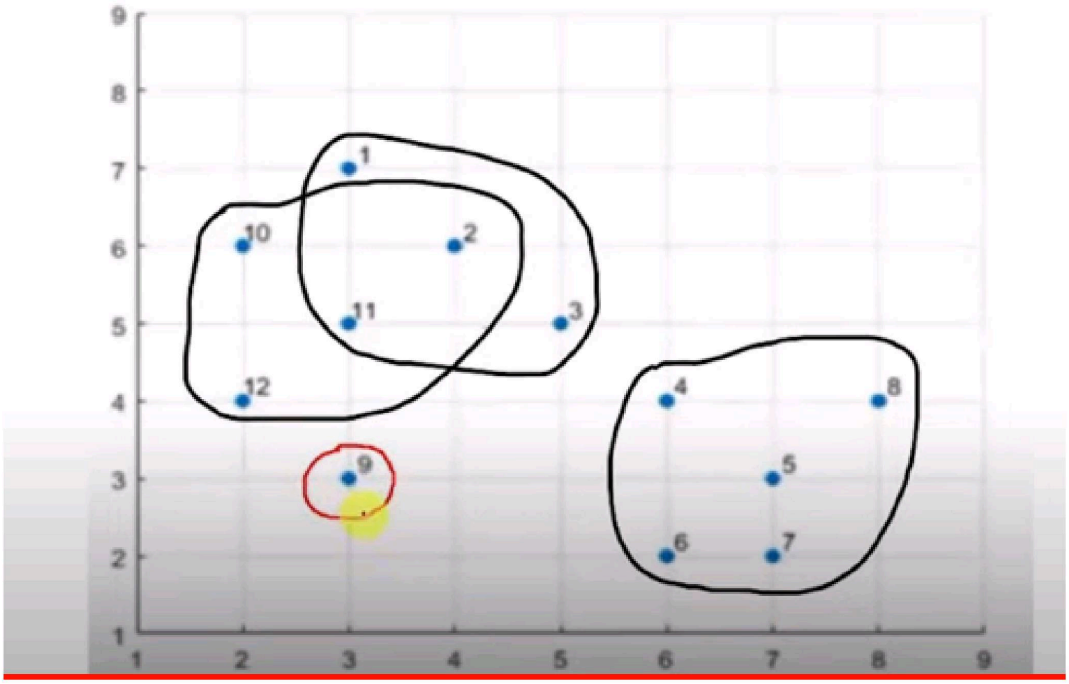
P11: P2, P10, P12

P12: P9, P11

minPts = 4 and epsilon (ε) = 1.9

Point	Status	
P1	Noise	Border
P2	Core	
P3	Noise	Border
P4	Noise	Border
P5	Core	
P6	Noise	Border
P7	Noise	Border
P8	Noise	Border
P9	Noise	
P10	Noise	Border
P11	Core	
P12	Noise	Border

minPts = 4 and epsilon (ε) = 1.9



Ex. 4.3.1 : Given the following data, calculate hyperplane. Also, classify (0.6, 0.9) based on the calculated hyperplane.

A1	A2	y	α
0.38	0.47	+	65.52
0.49	0.61	-	65.52
0.92	0.41	-	0
0.74	0.89	-	0
0.18	0.58	+	0
0.41	0.35	+	0
0.93	0.81	-	0
0.21	0.1	+	0

Soln. :

As you see, the value of α is non-zero for only first two training examples. Hence, the first two training examples are support vectors.

Let's calculate the value of w and b as required to calculate the SVM hyperplanes.

Let $w = (w_1, w_2)$

$$w = \sum_{n=1}^{N} a_n y_n x_n$$

$w_1 = a_1 * y_1 * A1x_1 + a_2 * y_2 * A1x_2 = 65.52 * 1 * 0.38 + 65.52 * -1 * 0.49 = -7.2$

$w_2 = a_1 * y_1 * A2x_1 + a_2 * y_2 * A2x_2 = 65.52 * 1 * 0.47 + 65.52 * -1 * 0.61 = -9.2$

The parameter b can be calculated for each support vector as follows.

$b_1 = 1 - w * x_1 = 1 - (-7.2) * 0.38 - (-9.2) * 0.47 = 8.06$

$b_2 = 1 - w * x_2 = 1 - (-7.2) * 0.47 - (-9.2) * 0.61 = 10$

Averaging b_1, b_2 you get b = 9.03

Hence, the hyperplane line is defined as $-7.2 x_1 - 9.2 x_2 + 9.03 = 0$

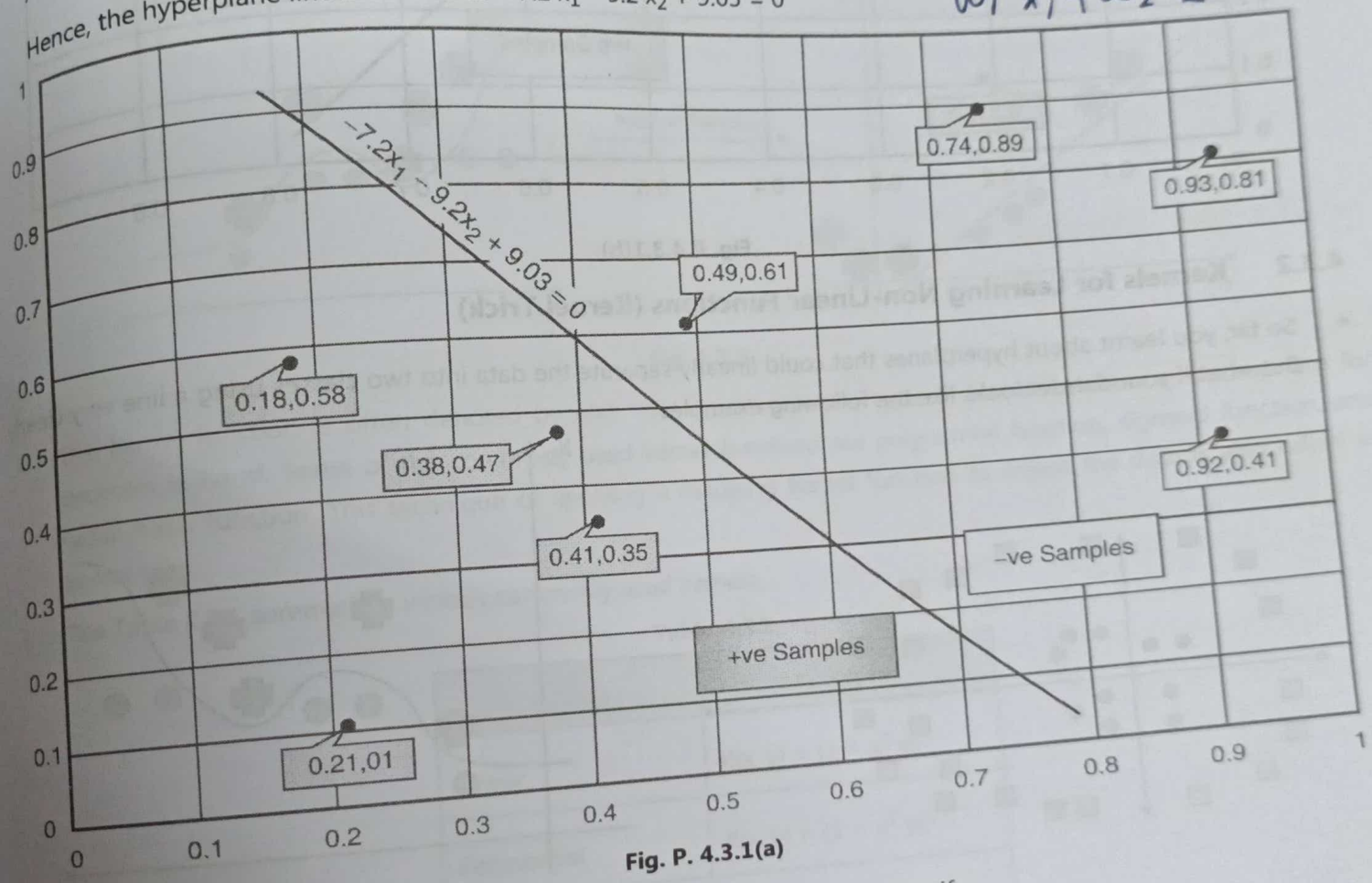


Fig. P. 4.3.1(a)

Now, suppose that there is a new data point (0.6, 0.9) that you want to classify.

Putting the values in the equation $-7.2 x_1 - 9.2 x_2 + 9.03$ you get $-7.2 * 0.6 - 9.2 * 0.9 + 9.03 = -3.57$.

Hence, this is classified as negative sample.

Agglomerative clustering

Q. Consider 1D Data pts. 18, 22, 25, 27, 42, 43.

→ Distance Matrix:

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	(1)	0

Circle the min cell.

Merge 42 & 43. cell (42,43) =

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	(2)	0	15
42,43	24	20	17	15	0

Single Linkage: (42,43) → 18 ∵ Min (42→18) & (43→18)

Complete Linkage: max dist.

Merge 25, 27.

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	(3)	0	15
42,43	24	20	15	0

Merge (25,27) & 22

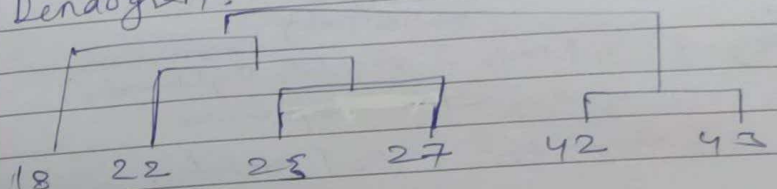
	(18)	(22, (25,27))	(42,43)
(18)	0	4	27
(22, (25,27))	(4)	0	15
(42,43)	24	15	0

	18, (22, (25,27))	(42,43)
18, (22, (25,27))	0	15
(42,43)	15	0

Merge both

	22, 18, (25,27), (42,43)
22, 18, (25,27), (42,43)	0

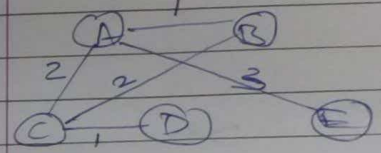
Dendogram:



Divisive method

→ uses MST

	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	4	1	0	
E	3	3	5	3	0



(A,B) (C,D) (A,C)
1 1 2

(B,C) (A,D) (E,A) (E,B) (E,D)
2 2 3 3 3

loop ×. skip

(V-1) edges Reached

∴ stop

ABCDE Connected

E/ABCD. (disconnect max Dist)

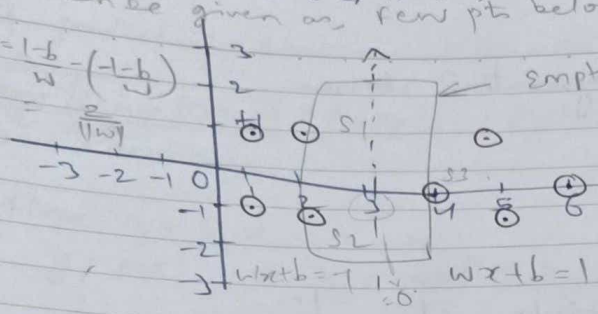
E/AB/CD

E/A/B/C/D.

Q. SVM Hyperplane.

(1,1) (2,1) (1,-1) (2,-1) (4,0) (5,1) (5,-1) (6,0)

(it can be given as, new pts belong to +ve class, others -ve)



1. Plot em
2. Box em
3. SV ↓ S1 S2 S3 (2,1) (2,-1) (4,0) → SV + Bias
4. Eqns → α_i
5. $\bar{w} = \begin{pmatrix} w \\ b \end{pmatrix} \sum \alpha_i S_i$

Hyperplane eq? = $\begin{pmatrix} 1 \\ 0 \end{pmatrix} x - 3$

$S1$ vector $= \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $S2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, $S3 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$

Bias Augment ⇒

$\bar{S1} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ $S2 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$ $S3 = \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}$ } add 1 last.

$\alpha_1 \bar{S_1}\bar{S_1} + \alpha_2 \bar{S_1}\bar{S_2} + \alpha_3 \bar{S_1}\bar{S_3} = -1$ (S1, Constant)

$\alpha_1 \bar{S_1}\bar{S_2} + \alpha_2 \bar{S_2}\bar{S_2} + \alpha_3 \bar{S_3}\bar{S_2} = -1$ (S2)

$\alpha_1 \bar{S_1}\bar{S_3} + \alpha_2 \bar{S_2}\bar{S_3} + \alpha_3 \bar{S_3}\bar{S_3} = +1$ (S3)

assume one side as +ve, other as -ve

$\therefore S1, S2 = +ve$, $S3 = -ve$ or

$S1, S2 = -ve$, $S3 = +ve$

$$\therefore \alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = -1$$

$\therefore \alpha(2\times2 + 1\times1 + 1\times1) + \alpha_2(2\times2 - 1\times1 + 1\times1) + \alpha_3(2\times4 + 0\times1 + 1\times1) = -1$

$\therefore \alpha(6) + \alpha_2(4) + \alpha_3(5) = -1$ — (1)

Similarly, $4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1$

$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = 1$

$\therefore \alpha_1 = -3.25$, $\alpha_2 = -3.25$, $\alpha_3 = 3.5$.

$y = mx + c$

$0 = m(1) - 3$

$m = 3$

$$\bar{w} = \sum \alpha_i \bar{S_i} = -3.25\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} - 3.25\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + 3.5\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} =$$

$$\begin{bmatrix} -3.25\times2 - 3.25\times2 + 3.5\times4 \\ -3.25\times1 + 3.25\times1 + 3.5\times0 \\ -3.25\times1 - 3.25\times1 + 3.5\times1 \end{bmatrix} = \begin{pmatrix} 2 \\ 0 \\ -3 \end{pmatrix} \begin{matrix} w \\ \\ b \end{matrix}$$

∴ Answer = 33.47

Q. Multiple LinReg.

y	x_1	x_2

x_1x_1 x_2x_2 x_1y x_2y x_1x_2

→ $y = B_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

$$B_1 = \frac{\Sigma x_2^2 \, \Sigma x_1 y - \Sigma x_1 x_2 \, \Sigma x_2 y}{\Sigma x_1^2 \, \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$B_2 = \frac{\Sigma x_1^2 \, \Sigma x_2 y - \Sigma x_1 x_2 \, \Sigma x_1 y}{\Sigma x_1^2 \, \Sigma x_2^2 - (\Sigma x_1 x_2)^2}$$

$$\Sigma x_1^2 = \Sigma x_1 \cdot x_1 - \frac{\Sigma x_1 \cdot \Sigma x_1}{N}$$

$$B_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

Quad Poly, $n=2$ $n>4$ ... over fit

1. Linear Reg.

$y = mx + c + e$.

$m = \dfrac{n\overline{xy} - \bar{x}\bar{y}}{n\overline{x^2} - \bar{x}^2}$

$c = \bar{y} - m\bar{x}$

x	y	Error	x^2	xy
Sum =				
Avg $= \bar{x}$	$\bar{y}$	(Mean)	$\overline{x^2}$	$\overline{xy}$

$X =$

When x given, find y

only 1 independent variable x.

Linearly Distributed Data.

or

$y = a + bx$

$$Cov(x,y) = \frac{1}{n-1} \sum (x_i - \bar{x}_i)(y_i - \bar{y}_i)$$

$$Correl(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\, Var(y)}}$$

$$Var(x) = \frac{1}{n-1} \sum (x - \bar{x}_i)^2$$

$$b = \frac{Cov(x,y)}{Var(x)}$$

$$a = \bar{y} - b\bar{x}$$

...on linear SVM –

Q.

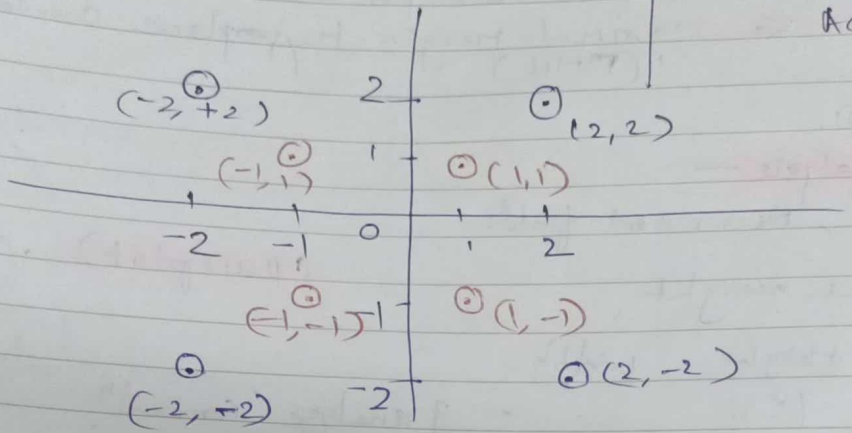
+vely labeled data pts –

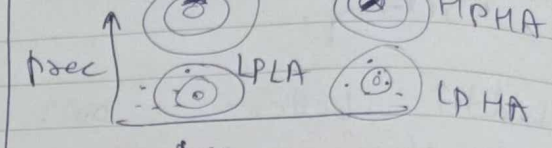
$\begin{pmatrix}2\\2\end{pmatrix}$ $\begin{pmatrix}2\\-2\end{pmatrix}$ $\begin{pmatrix}-2\\-2\end{pmatrix}$ $\begin{pmatrix}-2\\2\end{pmatrix}$

-vely labeled –

$\begin{pmatrix}1\\1\end{pmatrix}$ $\begin{pmatrix}1\\-1\end{pmatrix}$ $\begin{pmatrix}-1\\-1\end{pmatrix}$ $\begin{pmatrix}-1\\1\end{pmatrix}$

$\phi\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases} \begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & \text{if } \sqrt{x_1^2+x_2^2} > 2 \\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & \text{otherwise} \end{cases}$





Converting them from I feature space to another :

① for +vely labeled Data pts :

$\sqrt{2^2+2^2} > 2$ true $\therefore \begin{pmatrix}4-2+|2-2|\\4-2+|2-2|\end{pmatrix} = \begin{pmatrix}2\\2\end{pmatrix}$

$\sqrt{2+(-2)^2} > 2$ true $\therefore \begin{pmatrix}4+2+|2+2|\\4-2+|2+2|\end{pmatrix} = \begin{pmatrix}10\\6\end{pmatrix}$

$\sqrt{(-2)+(-2)^2} > 2$ true $\therefore$ 6, 6

$\sqrt{(-2)^2+(2)^2} > 2$ true $\therefore$ $\frac{6}{10}$

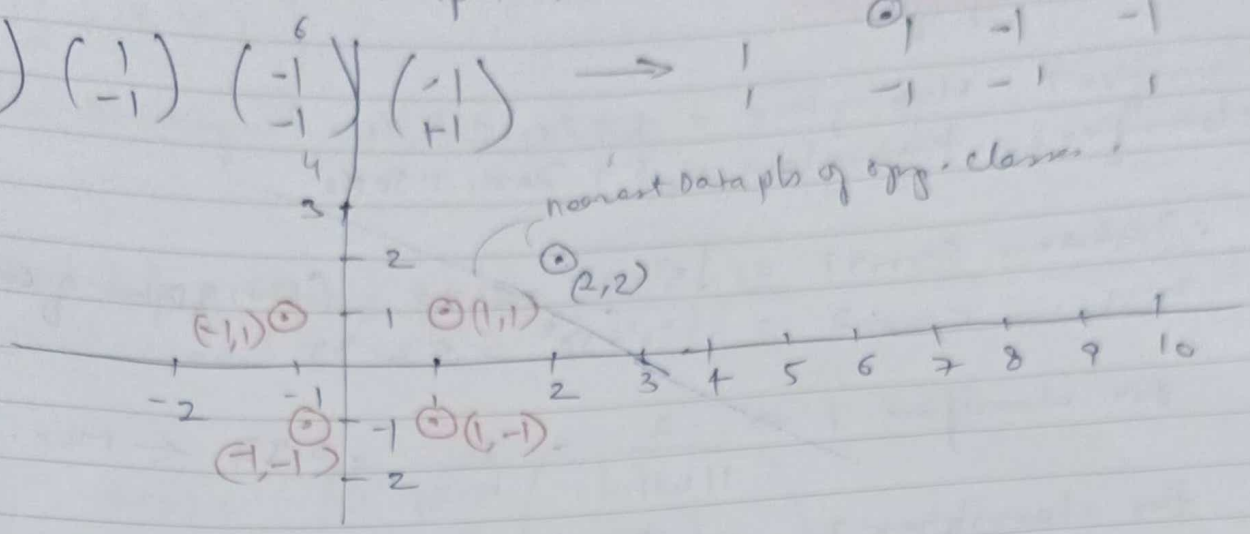
10 -

(6,10)

For -vely labeled data pts :-

(6,6) (10,6)

$\therefore \begin{pmatrix}1\\1\end{pmatrix} \begin{pmatrix}1\\-1\end{pmatrix} \begin{pmatrix}-1\\-1\end{pmatrix} \begin{pmatrix}-1\\+1\end{pmatrix} \rightarrow \begin{matrix}1 & -1 & -1 & 1\\ 1 & -1 & -1 & 1\end{matrix}$



$\therefore S_1 = \begin{matrix}1\\1\end{matrix}$ $\quad S_2 = \begin{matrix}2\\2\end{matrix}$

Add Bias: $\tilde{S}_1 = \begin{matrix}1\\1\\1\end{matrix}$ $\quad \tilde{S}_2 = \begin{matrix}2\\2\\1\end{matrix}$ ... wrt S_1 ... wrt S_2

$\alpha_1 \tilde{S}_1 \tilde{S}_1 + \alpha_2 \tilde{S}_1 \tilde{S}_2 = -1$

$\alpha_1 \tilde{S}_2 \tilde{S}_1 + \alpha_2 \tilde{S}_2 \tilde{S}_2 = +1$

$\therefore \quad 3\alpha_1 + 5\alpha_2 = -1$

$5\alpha_1 + 9\alpha_2 = 1$

$\alpha_1 = -7, \quad \alpha_2 = 4$

$\therefore \tilde{w} = \sum \alpha S =$

$\alpha_1 \tilde{S}_1 + \alpha_2 \tilde{S}_2$

$= -7\begin{pmatrix}1\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix} = \begin{pmatrix}1\\1\\-3\end{pmatrix}$

$w = \begin{matrix}1\\1\end{matrix}$ $\quad b = \text{bias} = -3$

Q. Multivar Reg.

x_1	x_2	y_1	y_2

$Y = [y_1 \; y_2]$ $\quad X = \begin{bmatrix}1 & x_1 & x_2\\1 & & \end{bmatrix}$

$\theta = [X^T X]^{-1} \cdot X^T \cdot Y$

$= \begin{matrix}-41 & 67\\ 59 & 27\\ 2.9 & -2.3\end{matrix}$

$\therefore \hat{y}_1 = -41 + 59x_1 + 2.9x_2$

$\hat{y}_2 = 67 + 27x_1 - 2.3x_2$

Margin ↑ : better control over Generalization error.
↓ : overfitting hojayegi.

Select Best Hyperplane

Eg. classifier 1 Hyperplane = $5 + 2x_1 + 5x_2$
classifier 2 Hyperplane = $5 + 20x_1 + 50x_2$

∴ Distance error 1 = $\sqrt{5^2+2^2} = 5.39$ (sum of square of coeff.)
Distance error 2 = $\sqrt{20^2+50^2} = 53.85$

for classifier 1 $\Rightarrow \frac{2}{\|w\|} = \frac{2}{5.39} = 0.37$ ← Max.

for classifier 2 $= \frac{2}{\|w\|} = \frac{2}{53.85} = 0.037$

∴ Hyperplane 1 is Best

age { youth, Middle Aged, Senior }

(y,m) (m,s) (y,s)

Q. id-3

	age	income	student	creditRating	buys Computer
1	youth	high	no	fair	no
2	youth	h	n	excellent	n
3	Middle	h	n	f	yes
4	Senior	medium	n	f	y
5	Senior	low	yes	f	y
6	Senior	L	y	e	n
7	middle	l	y	e	y
8	youth	m	n	f	n
9	youth	L	y	f	y
10	Senior	m	y	f	y
11	youth	m	y	e	y
12	Middle	m	n	e	y
13	Middle	h	y	f	y
14	Senior	m	n	e	n

n(yes) = 9

n(no) = 5

age	y	n	T	
youth	2	3	5	$-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.971$
middleaged	4	0	4	0
senior	3	2	5	$-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$

Total

Splitinfo $= -\frac{5}{14}\log\frac{5}{14} - \frac{4}{14}\log\frac{4}{14} - \frac{5}{14}\log\frac{5}{14} = 1.5774$

$\therefore$ If $= 0.971 \times \frac{5}{14} + 0.971 \times \frac{5}{14} = 0.6934$

Gain = 0.9403 − 0.6934 = 0.2469

Gain Ratio = $\frac{0.2469}{1.5774} = 0.1559$

income	y	n	T	
high	2	2	4	1
medium	4	2	6	$-\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.9183$
low	3	1	4	$-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$

Split-Info $= -\frac{4}{14}\log\frac{4}{14} - \frac{6}{14}\log\frac{6}{14} - \frac{4}{14}\log\frac{4}{14} = 1.5566$

If $= 1 * \frac{4}{14} + 0.9183 * \frac{6}{14} + 0.8113 * \frac{4}{14} = 0.911$

Gain = 0.9403 − 0.911 = 0.0293

Gain Ratio $= \frac{0.0293}{1.5566} = 0.0188$

Student				
yes	6	1	7	$-\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7} = 0.5917$
no	3	4	7	$-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.9852$
	y	n	T	

Split Info $= -\left(\frac{7}{14}\log\frac{7}{14}\right)\times 2 = 1$

$IG = 0.5917 * \frac{7}{14} + 0.9852 * \frac{7}{14} = 0.7884$

Gain Ratio $= \frac{0.1519}{1} = 0.1519$

Gain $= 0.9403 - 0.7884 = 0.1519$

Credit Rating				
fair	6	2	8	$-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.8113$
Excellent	3	3	6	
	y	n	T	

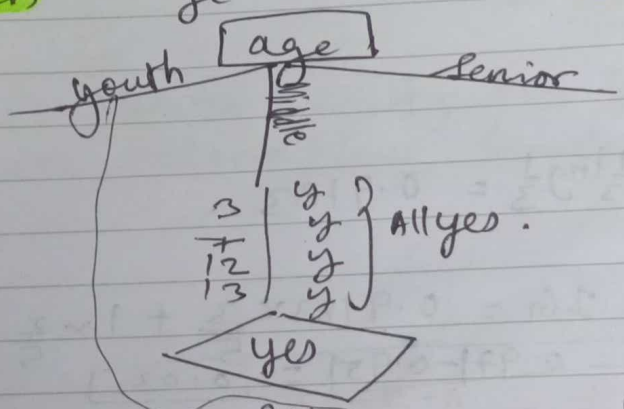
$IG = 0.8113 * \frac{8}{14} + 1 * \frac{6}{14} = 0.8922$

Gain $= 0.9403 - 0.8922 = 0.0481$

Gain Ratio $= \frac{0.0481}{0.9852} = 0.0488$

Split Info $= -\frac{8}{14}\log\frac{8}{14} - \frac{6}{14}\log\frac{6}{14} = 0.9852$

Max Gain .. age ... root ... Max Gain Ratio .. Age



1	n
2	n
8	y
9	y
11	y

n(yes) = 3
n(no) = 2
↓ Gain $= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$

3	y
7	y
12	y
13	y

4	y
5	y
6	n
10	y
14	n

n(yes) = 3
n(no) = 2
↓ Gain $= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$

income				
H	0	2	2	0
L	1	0	1	0
M	1	1	2	1
	y	n	T	

$IG = 1 * \frac{2}{5} = 0.4$

Gain $= 0.971 - 0.4 = 0.571$

Gain Ratio $= \frac{0.571}{1.5219} = 0.3754$

Split info $= -\frac{2}{5}\log\frac{2}{5} - \frac{1}{5}\log\frac{2}{5} - \frac{2}{5}\log\frac{2}{5} = 1.5219$

Student

y	2	0	2	0
n	0	3	3	0
	y	n	T	

$IG = 0$

Gain = 0.971 − 0 = 0.971

SplitInfo = $-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.9709$

Gain Ratio = $\frac{0.971}{0.9709} = 1$

Credit Rating

fair	1	2	3	$\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9183$
Excellent	1	1	2	1
	y	n	T	

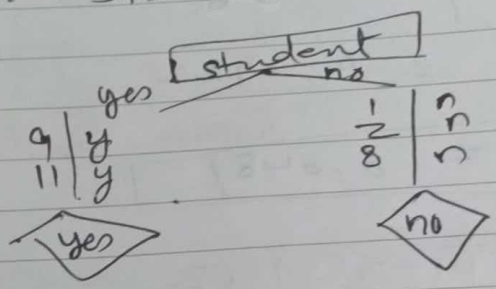
$IG = 1 - \frac{2}{5} + \frac{3}{5} \times 0.9183 = 0.7443$

Gain = 0.971 − 0.7443 = 0.2267

Split info = $-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.9709$

Gain Ratio = $\frac{0.2267}{0.9709} = 0.2335$

Max Gain.. Student.



Max Gain Ratio.. Student.

for Senior.

income

M	2	1	3	$-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9183$
L	1	1	2	1
H	0	0	0	0
	y	n	T	

$IG = 0.9183 \times \frac{3}{5} + 1 \times \frac{2}{5} = 0.951$

Gain = 0.971 − 0.951 = 0.02

Split Info = $-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.9709$

Gain Ratio = $\frac{0.02}{0.9709} = 0.0205$

Student

y	2	1	3	$-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9183$
n	1	1	2	1
	y	n	T	

$IG = 0.9183 \times \frac{3}{5} + 1 \times \frac{2}{5} = 0.02$

Split Ratio = $-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.9709$

Gain Ratio = $\frac{0.02}{0.9709} = 0.0205$

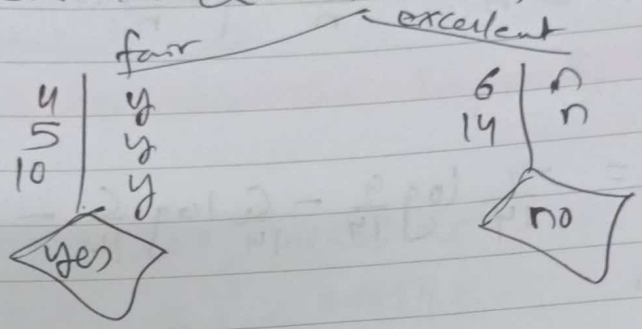
Split Info $= -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.9709$

Gain Ratio $= \frac{0.971}{0.9709} = 1$

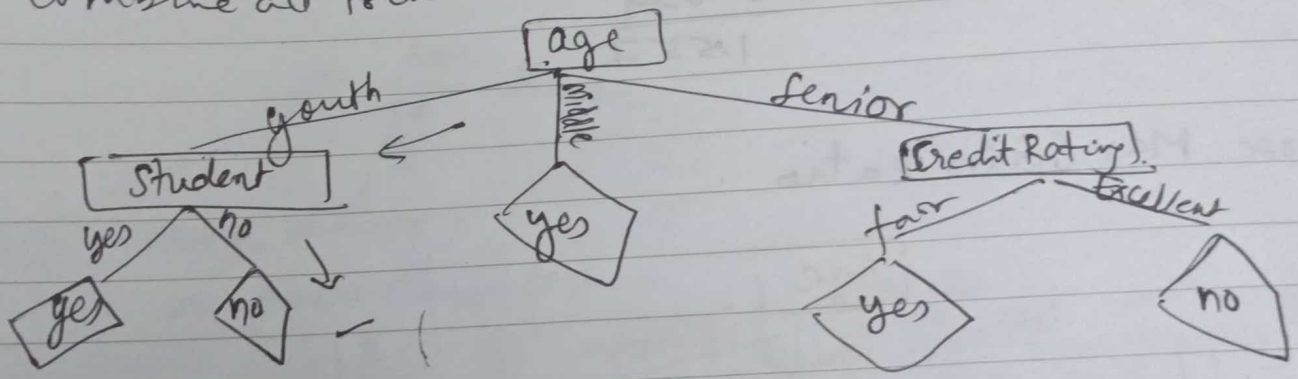
CreditRating						
fair	3	0	3	0	IG = 0	
Excellent	0	2	2	0	Gain = 0.971 − 0 = 0.971	
	y	n	T			

Max Gain ... CreditRating

Max Gain Ratio CreditRating



Combine all trees



Use this final tree to classify the record:

age = youth, income = low, Student = no, CreditRating = Excellent

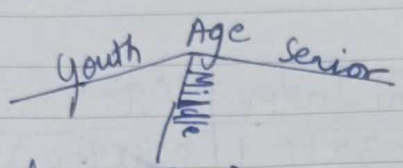
no

...same dataset.

n(yes) = 9, n(no) = 5.

$gini = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$

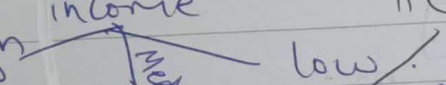
$gini_i = \frac{D_1}{\#D} gini(D_1) + \frac{D_2}{\#D} gini(D_2)$.



(youth, middle, senior)

Consider all pairs (subsets).

youth, middle	$\frac{9}{14}\left(1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2\right)+\frac{5}{14}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)=0.45$ (youth, middle) (senior)
youth, senior	$\frac{10}{14}\left(1-\left(\frac{5}{10}\right)^2-\left(\frac{5}{10}\right)^2\right)+\frac{4}{14}\left(1-\left(\frac{4}{4}\right)^2-\left(\frac{0}{4}\right)^2\right)=0.3571$
middle, senior	$\frac{9}{14}\left(1-\left(\frac{7}{9}\right)^2-\left(\frac{2}{9}\right)^2\right)+\frac{5}{14}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)=0.3936$.



#(h+m) yes. #(L) yes.

(h, m)	10	$\frac{10}{14}\left(1-\left(\frac{2+4}{10}\right)^2-\left(\frac{2+2}{10}\right)^2\right)+\frac{4}{14}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)=0.45$
(m, l)	10	$\frac{10}{14}\left(1-\left(\frac{3+4}{10}\right)^2-\left(\frac{1+2}{10}\right)^2\right)+\frac{4}{14}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)=0.4428$
(h, l)	8	$\frac{8}{14}\left(1-\left(\frac{3+2}{8}\right)^2-\left(\frac{1+2}{8}\right)^2\right)+\frac{6}{14}\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right)=0.4583$

→ has Difficulty when #Classes ↑.
↔ Equalized part: favoured + purity in both parts
all yes / all no

Income < yes / no

(y, n) ⇒ $\frac{7}{14}\left(1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2\right)+\frac{7}{14}\left(1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right)$.

y+n = 14

$= 0.3673$

CreditRating < fair / excellent

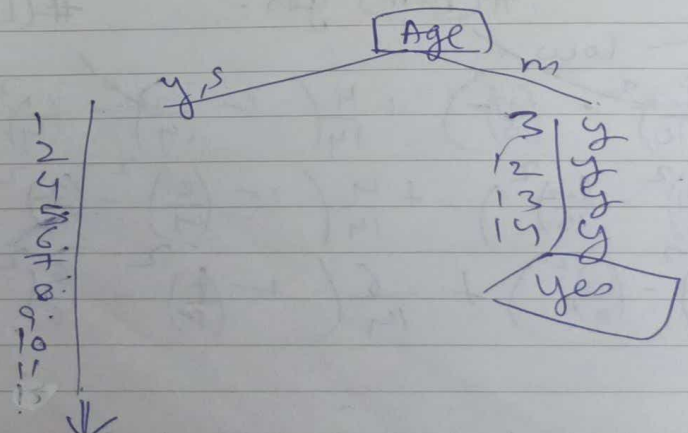
(f, e)

f+e = 14

$\frac{8}{14}\left(1-\left(\frac{6}{8}\right)^2-\left(\frac{2}{8}\right)^2\right)+\frac{6}{14}\left(1-\left(\frac{3}{6}\right)^2-\left(\frac{3}{6}\right)^2\right) = 0.4285$

Attribute	Split	Gini Index	Δgini = .
age	(y, s) + m	[0.3571]	0.459 − 0.3571 = 0.1019
income	(m, h) + l	0.4428	0.459 − 0.4428 = 0.0162
Student	Binary	0.3673	0.459 − 0.3673 = 0.0917
CreditRating	Binary	0.4285	0.459 − 0.4285 = 0.0305

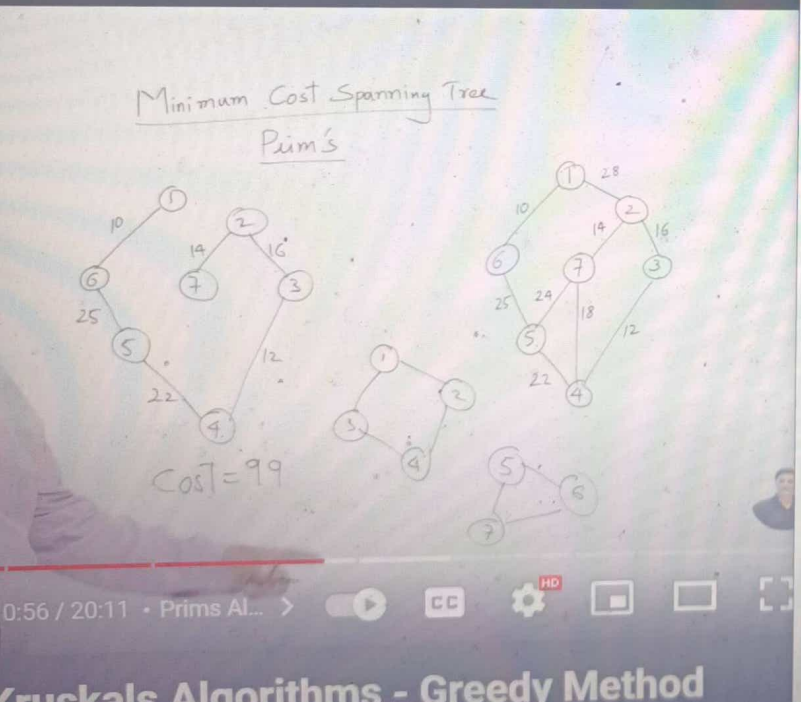
min. Gini Index
↑ Red: in Impurity



again process.

m/watch?v=4ZlRH0eK-qQ

kruskal's algorithm



0:56 / 20:11 • Prims Al...

Kruskals Algorithms - Greedy Method

*Untitled - Notepad

File Edit Format View Help

prim

next min weight edge must be a neighbor so that the graph is connected at every step

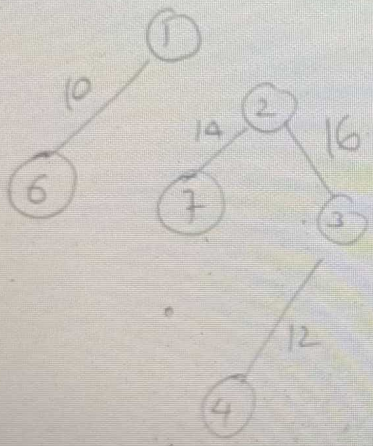
kruskal

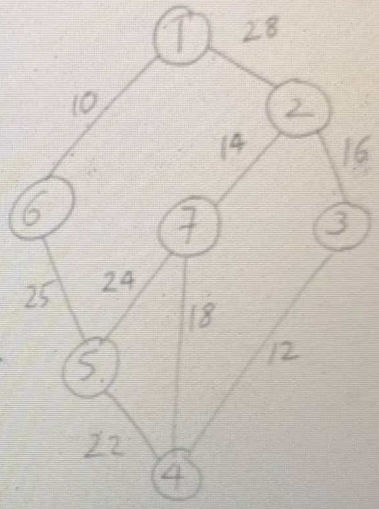
select most min only. its ok if its not adj.

Ln 7, Col 45 100% Windows (CRLF) UTF-8

Minimum Cost Spanning Tree

Kruskal's





Ex. 4.4.1 : Calculate the radial distance of point A having value of 6 with respect to point B and C having value of 8 and 12 respectively.

Soln. :

$$\text{Assume } \gamma = 1$$

$$\text{Distance of point A with respect to B} = e^{-\gamma(x-y)^2} = e^{-1(6-8)^2} = e^{-4} = 0.018$$

$$\text{Distance of point A with respect to C} = e^{-\gamma(x-y)^2} = e^{-1(6-12)^2} = e^{-36} = 0$$

So, point B has higher influence over point A than point C.

Practice Questions

Ex. 2.1.1 : For the following data set, find the linear regression line. Predict the value of Y if X = 10.

X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
0	1				
1	3				
2	2				
3	5				
4	7				
5	8				
6	8				
7	9				
8	10				
9	12				
		$\bar{X} =$	$\bar{Y} =$	$\Sigma =$	$\Sigma =$

Soln. : Calculate the mean of X and Y and then calculate the other values as required.

X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
0	1	– 4.5	– 5.5	24.75	20.25
1	3	– 3.5	– 3.5	12.25	12.25
2	2	– 2.5	– 4.5	11.25	6.25
3	5	– 1.5	– 1.5	2.25	2.25
4	7	– 0.5	0.5	– 0.25	0.25
5	8	0.5	1.5	0.75	0.25
6	8	1.5	1.5	2.25	2.25
7	9	2.5	2.5	6.25	6.25
8	10	3.5	3.5	12.25	12.25
9	12	4.5	5.5	24.75	20.25
$\bar{X}$ (Mean of X) = 4.5	**$\bar{Y}$ (Mean of Y) = 6.5**			**Total = 96.5**	**Total = 82.5**

Hence,
$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{96.5}{82.5} = 1.1696$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_0 = 6.5 - 1.1696 \times 4.5 = 1.2368$$

Hence, the regression line is $Y = 1.2368 + 1.1696X$

A sample plot of the regression line is as shown in Fig. P. 2.1.1.

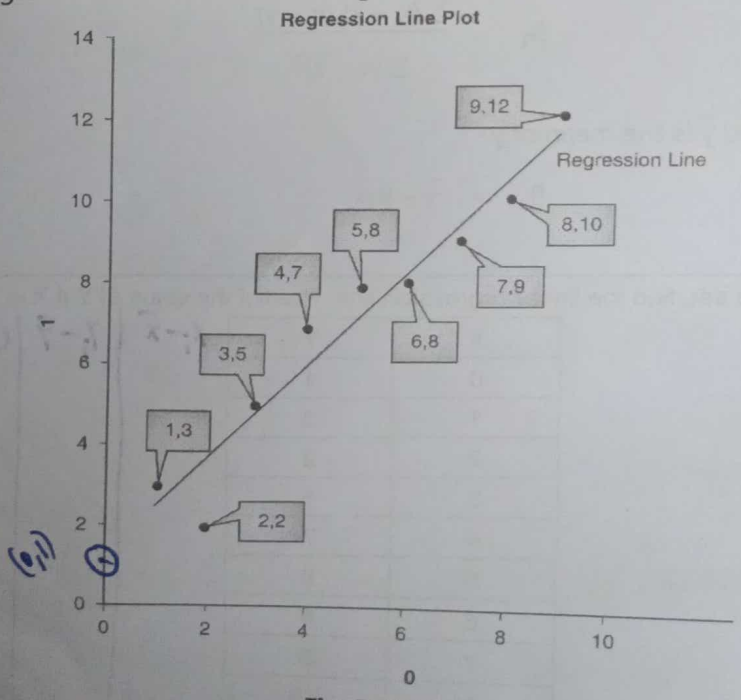


Fig. P. 2.1.1

Ex. 2.4.1 : Calculate the entropy of the following distribution.

Gender	Count
Male	9
Female	5

Soln. :

- $p(\text{Male}) = \frac{9}{14}$
- $p(\text{Female}) = \frac{5}{14}$

So, for calculating the entropy of the distribution you would use the following formula.

$$H_x = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$\text{Entropy} = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right)$$

$$\text{Entropy} = -(-0.53 - 0.40)$$

$$\text{Entropy} = 0.93$$

x. 2.4.3 : Calculate the entropy of the following distribution.

Fruit Colour	Taste	Count
Yellow	Sweet	10
Red	Sweet	5
Green	Sour	15
Orange	Sour	5

35

Soln. :

Note here that there are two input variables - colour and taste.

Assuming that the target of entropy calculation is on the basis of fruit *colour*. In this scenario, the *taste* column is then used as weight for entropy calculation.

- $p(\text{Yellow}) = \frac{10}{35}$
- $p(\text{Red}) = \frac{5}{35}$
- $p(\text{Green}) = \frac{15}{35}$
- $p(\text{Orange}) = \frac{5}{35}$
- $p(\text{Sweet}) = \frac{15}{35}$
- $p(\text{Sour}) = \frac{20}{35}$

Entropy = Sweet $*^{E}$(Yellow) + Sweet $+^{E}$(Red) + Sour $*^{E}$(Green) + Sour $*^{E}$(Orange)

$$= \left[\frac{15}{35} \times \frac{10}{35} \log \frac{10}{35} + \frac{15}{35} \times \frac{5}{35} \log \frac{5}{35}\right] + \ldots$$

$$= 0.92.$$

Practice Questions

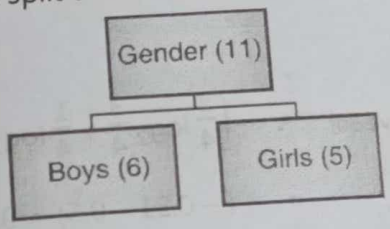
Ex. 2.4.4 : Which attribute would you choose to split the following dataset into a decision tree?

Gender	Marks
Girl	65
Girl	46
Boy	56
Boy	43
Boy	53
Boy	49
Girl	42
Boy	84
Boy	44
Girl	42
Girl	40

Soln. :

There are totally 11 datapoints. There are 5 girls and 6 boys. Also, there are 7 students with marks below 50 and 4 students with marks above 50. Now, you have to find out which of these two attributes (gender or marks) is better for splitting the dataset into a decision tree.

Let's take scenario 1 where you decide to split the tree based on gender.



Calculate entropy of the parent node (Gender).

$$p(\text{Boys}) = \frac{6}{11}$$

$$p(\text{Girls}) = \frac{5}{11}$$

$$\text{Entropy}_{\text{(Parent Node)}} = -\left(\frac{6}{11}\log_2\frac{6}{11} + \frac{5}{11}\log_2\frac{5}{11}\right)$$

$$\text{Entropy}_{\text{(Parent Node)}} = -(-0.48 - 0.52)$$

$$\text{Entropy}_{\text{(Parent Node)}} = 1$$

After split,

$$\text{Entropy}_{\text{(Boys)}} = -\left(\frac{6}{6}\log_2\frac{6}{6}\right) = 0$$

$$\text{Entropy}_{\text{(Girls)}} = -\left(\frac{5}{5}\log_2\frac{5}{5}\right) = 0$$

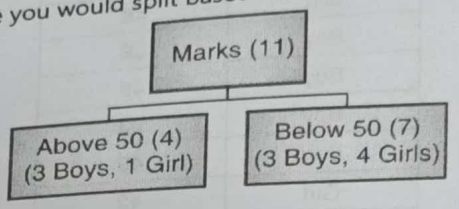
$$\text{Weighted entropy of child nodes} = \frac{6}{11} \times \text{Entropy (Boys)} + \frac{5}{11} \times \text{Entropy (Girls)} = 0$$

Hence,

$$\text{Information Gain, IG}_{\text{(Gender)}} = \text{Parent Entropy} - \text{Child Entropy} = 1 - 0 = 1$$

Information Gain of 1 indicates that after split, pure sets of boys and girls are formed and splitting on gend makes sense.

Now, let's take scenario two where you would split based on marks scored (less than 50, 50 or above it).



After split,

Above 50

- $p(\text{Boys}) = \frac{3}{4}$
- $p(\text{Girls}) = \frac{1}{4}$

Below 50

- $p(\text{Boys}) = \frac{3}{7}$
- $p(\text{Girls}) = \frac{4}{7}$

$$\text{Entropy}_{\text{(Above 50)}} = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$= -(-0.31 - 0.5) = 0.81$$

$$\text{Entropy}_{\text{(Below 50)}} = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right)$$

$$= -(-0.52 - 0.46) = 0.98$$

Weighted entropy of child nodes

$$= \frac{4}{11} \times \text{Entropy (Above 50)} + \frac{7}{11} \times \text{Entropy (Below 50)}$$

$$= \frac{4}{11} \times 0.81 + \frac{7}{11} \times 0.98$$

$$= 0.29 + 0.62 = 0.91$$

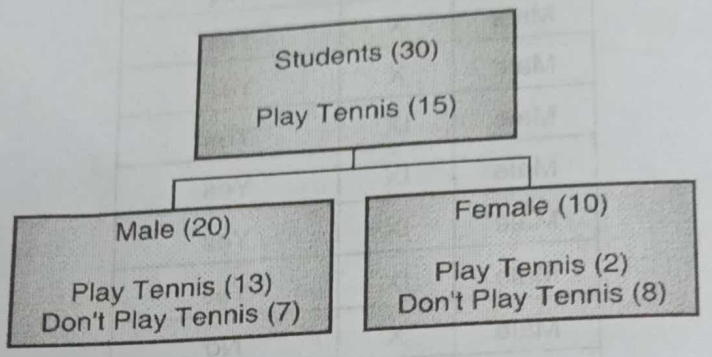
Hence,

$$\text{Information Gain, IG}_{\text{(Marks)}} = \text{Parent Entropy} - \text{Child Entropy}$$

$$= 1 - 0.91 = 0.09$$

As you see, the information gain on splitting the dataset by gender is higher than splitting by marks. Hence, suggested to split the given dataset on gender.

Scenario 1 : Split on Gender



Gini for sub – nodes = $(P_i)^2$

Gini $_{(Male \mid Tennis)}$ = Square of probability of playing tennis + Square of probability of not playing tennis

Gini $_{(Male \mid Tennis)} = \left(\frac{13}{20}\right)^2 + \left(\frac{7}{20}\right)^2 = 0.42 + 0.12 = 0.54$

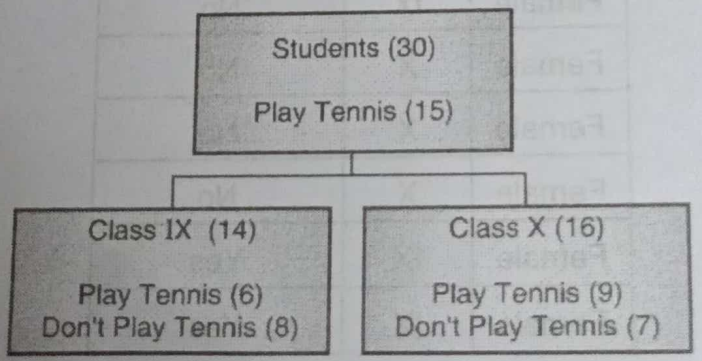
Gini $_{(Female \mid Tennis)}$ = Square of probability of playing tennis + Square of probability of not playing tennis

Gini $_{(Female \mid Tennis)} = \left(\frac{2}{10}\right)^2 + \left(\frac{8}{10}\right)^2 = 0.04 + 0.64 = 0.68$

Gini Index $_{(Gender)}$ = Weight of sub-node $_{(Male)}$ × Gini$_{(Male \mid Tennis)}$ + Weight of sub-node $_{(Female)}$ × Gini$_{(Fema}$

Gini Index $_{(Gender)} = \frac{20}{30} \times 0.54 + \frac{10}{30} \times 0.68 = 0.36 + 0.23 = 0.59$

Scenario 2 - Split on Class



Date: / /

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix}$$

$$C = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$C - \lambda I = 0$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - 121 = 0$$

$$(14)(23) - (14 + 23\lambda) + \lambda^2 - 121 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\therefore \lambda_1 = 30.3849$$
$$\lambda_2 = 6.615$$

For $\lambda_1 = 30.3849$

$$(C - \lambda I)(x) = 0$$

$$\begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$(14-\lambda_1)x_1 - 11x_2 = 0$$
$$-11x_1 + (23-\lambda_1)x_2 = 0$$

$$\therefore \frac{x_1}{11} = \frac{x_2}{14-\lambda_1} = e \quad \text{(let)}$$

$e_1^T = [0.5574 \quad -0.8303]$

$e_2^T = [0.8303 \quad 0.5574]$

$P_{11} \quad P_{12} \quad P_{13} \quad P_{14}$

PC1

$P_{11} = e_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix} = -4.3052$

$P_{12} = e_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} = 3.7363$

$P_{13} = 5.6930$

$P_{14} = -5.1240$

PC1 =	-4.3052	3.7363	5.6930	-5.12
	P_{11}	P_{12}	P_{13}	P_{14}

2.083

$$\begin{bmatrix} 2.73 - 0.0506 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = 0$$

$a_{21} = 0.79$

$a_{22} \quad 0.61$

...nents are given as

... $a_{12} x_2$ | $z_2 = a_{21} x_1 + a_{22} x_2$

PAGE:
DATE:

$e_1 = \begin{pmatrix} 4-8 \\ 11-8.5 \end{pmatrix}$

$8-8$
$4-8.5$

$13-8$
$5-8.5$

$7-8$
$14-8.5$

PC

Total

predicted

Actual	akiec	bcc	df
akiec	1,1 ✓	1,2	1,3
bcc	2,1	2,2 ✓	2,3 ✓
df	3,1	3,2	3,3

TP	1,1
FP	(1,2) (1,3)
TN	(2,2)(2,3) (3,2)(3,3)
FN	(2,1)(3,1)

TP	2,2
FP	(2,1)(2,3)
TN	(1,1)(1,3) (3,1)(3,3)
FN	(1,2)(3,2)

TP	3,3
FP	(3,1)(3,2)
TN	(1,1)(1,2) (2,1)(2,2)
FN	(1,3)(2,3)

Actual \ predicted	+	−
+	TP	FN
−	FP	TN

TP (1,1)

FP (1,2)(1,3)

FN (2,1)(3,1)

TN (2,2)(2,3) (3,2)(3,3)