Machine Learning.

Unit :-2

Date
Page

* Types of descriptive statistics :-

1) distribution 2) central tendency 3) variability

* Measures of central tendency :-
  - Mean
  - Median
  - Mode

Mean :-

Dataset : 15, 3, 12, 0, 24, 3.

sum = 15 + 3 + 12 + 0 + 24 + 3

= 57

Total no. of responses N = 6.

∴ Mean = 57/6 = **9.5**

* Measures of variability :-
  - Range
  - Standard deviation
  - Variance.

Range :- Ordered dataset : 0, 3, 3, 12, 15, 24

Range = 24 - 0

= 24.

Standard deviation :- [S or SD]

It is the avg. amount of variability in the dataset.

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$ ← $n-1$

Ex:. Given:.

Rawdata: 15, 3, 12, 0, 24, 3

Sdⁿ:

Rawdata	deviation from Mean	Squared deviation
15	15 − 9.5 = 5.5	30.25
3	3 − 9.5 = −6.5	42.25
12	12 − 9.5 = 2.5	6.25
0	0 − 9.5 = −9.5	90.25
24	24 − 9.5 = 14.5	210.25
3	3 − 9.5 = −6.4	42.25
M = 9.5		421.5

Mean.

→ Mean $= \frac{15+3+12+0+24+3}{6} = 9.5$

→ S.D OR S. $= \sqrt{\frac{\text{sum of squared deviation}}{N-1}}$

$= \sqrt{\frac{421.5}{5}}$

$= \sqrt{84.3}$

∴ S. = 9.18

Variance : Square of standard deviation

⇒ S = 9.18

∴ $S^2 = 84.3$ ← Variance of given dataset

* Types of descriptive statistics :-

1) Univariate
→ focuses on a one variable at a time

2) Bivariate
→ simultaneously study frequency & variability of two variables.

3) Multivariate
→ same as bivariate but with more than two variable

* Mode :-

→ The value that has higher frequency.

For Ungrouped data :-

Match no.	1	2	3	4	5	6	7	8	9	10.
No. of wickets	2	1	1	3	2	3	2	2	4	1

2 wickets were taken by the bowler frequently.
∴ mode is 2

For grouped data :-

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h.$$

l = lower limit of the model class
h = size of class interval
f_1 = frequency of modal class
f_0 = frequency of class preceding the model class.
f_2 = freq. of class succeeding the model class.

Ex:- In a class of 30 students marks obtained by students in mathematics out of 50 is tabulated as below. Calculate the mode of given data.

Marks obtained	Number of students.	
10-20	5	$\leftarrow f_0$
$l \rightarrow$ 20-30	12	$\leftarrow f_1$
30-40	8	$\leftarrow f_2$
40-50	5.	

$h = 30 - 20$
$= 10$

→ Maximum class freq. = 12
∴ corresponding model class = 20-30

$$\therefore \text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

∴ l = lower limit of modal class = 20
∴ h = size of class interval = 30-20
= 10
∴ f_1 = freq. of modal class = f 12
∴ f_0 = freq. of the class preceding the modal class = 5
∴ f_2 = freq. of the class succeeding the modal class = 8

$$\therefore \text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 20 + \left(\frac{12 - 5}{24 - 5 - 8}\right) \times 10$$

$$= \boxed{26.364}$$

* Medium :

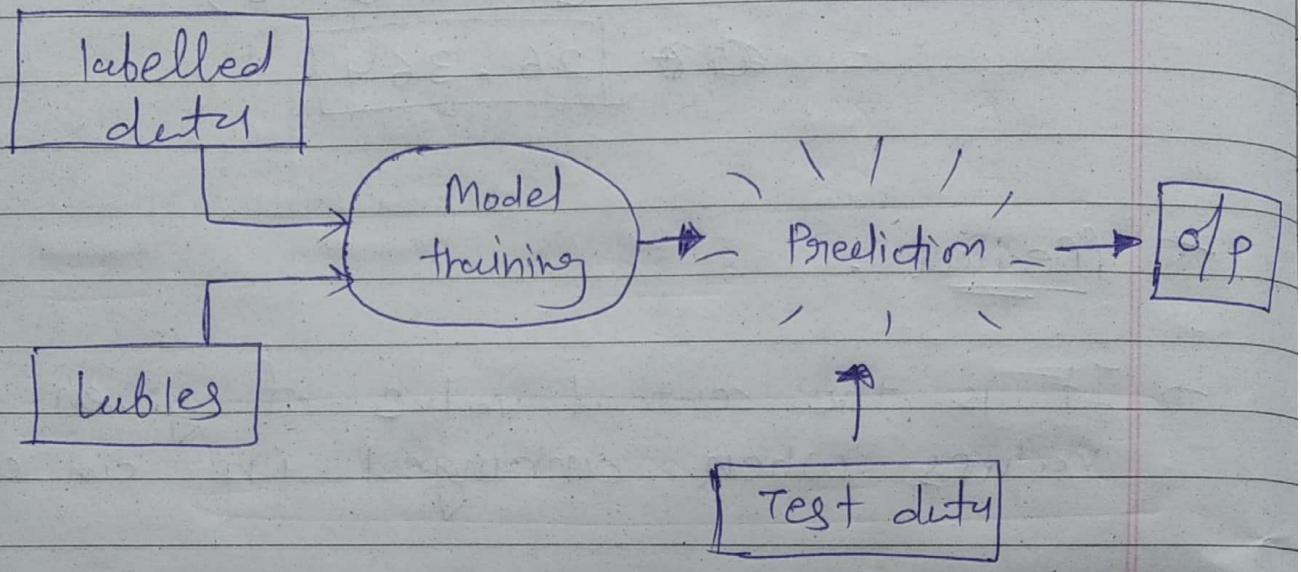
→ It is the central value of given set of values when arrangeed in an order.

* Formula :

Mode = 3 Medium − 2 Mean.

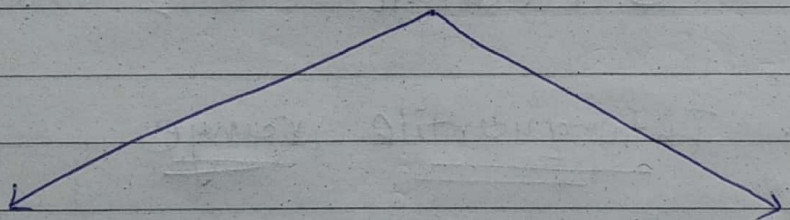
Supervised learning

* Diagram :-



* Types of supervised ML Algorithms :-

Knowledge purpose.



Regression

- Linear Regression
- Bayesian linear regression
- Regression Trees
- Non-linear regression
- Polynomial regression.

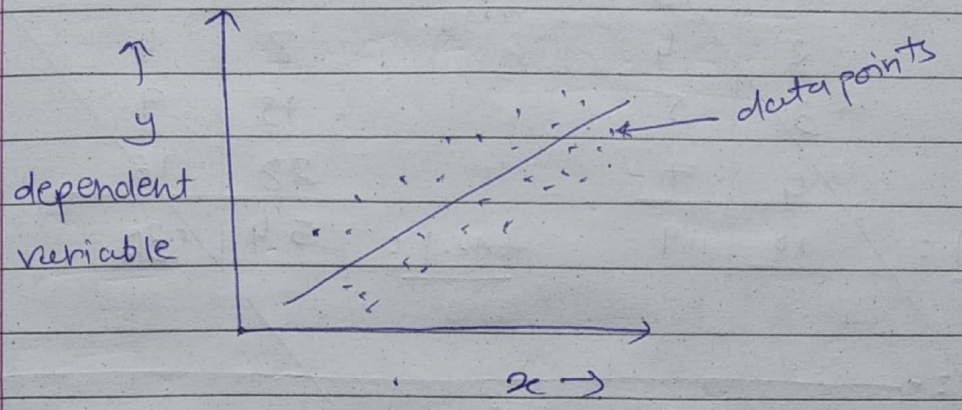
Classification

- Random forest
- Decision Trees
- Logistic regression
- Support vector machine
- Naive Bayes Classifier
- K- Nearest Neighbours.

Date
Page

* Linear Regression :-

→ Dependent variable is continuous in nature.



1) Simple linear regression :

→ one dependent and only one independent variable exists.

$$y = \alpha_0 + \alpha_1 X_1$$

↑ dependent variable (to be found)

α_1, X_1 ← independent variable, co-efficient of regression

2) Multiple linear regression :

→ more than one independent variable and only one dependent variable exists.

$$y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_m X_m$$

α_i = Regression co-efficient
X_i = Independent variable
y = Dependent variable.

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

Date
Page

* Linear Regression

OLS method
Ordinary least square method

Ex:

Number of TV Ads (x)	Number of cars sold (y)
1	14
3	24
2	18
1	17
3	27
10	100

→ $\bar{x} = \frac{10}{5} = 2$ | $\bar{y} = \frac{100}{5} = 20$

→ Regression eqn : $\boxed{y = ax + b}$

1* $a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ slope

2* $b = \bar{y} - a\bar{x}$ intercept

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	14	-1	-6	6	1
3	24	1	4	4	1
2	18	0	-2	0	0
1	17	-1	-3	3	1
3	27	1	7	7	1
10	100			20	4

$$\therefore a = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$= \frac{20}{4}$$

$$\therefore \boxed{a = 5}$$

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

$= 20 - (5)(2)$

$$\therefore \boxed{b = 10}$$

$\therefore$ Estimated regression eqn $\Rightarrow \hat{y} = aX + b$

$$\hat{y} = 5X + 10$$

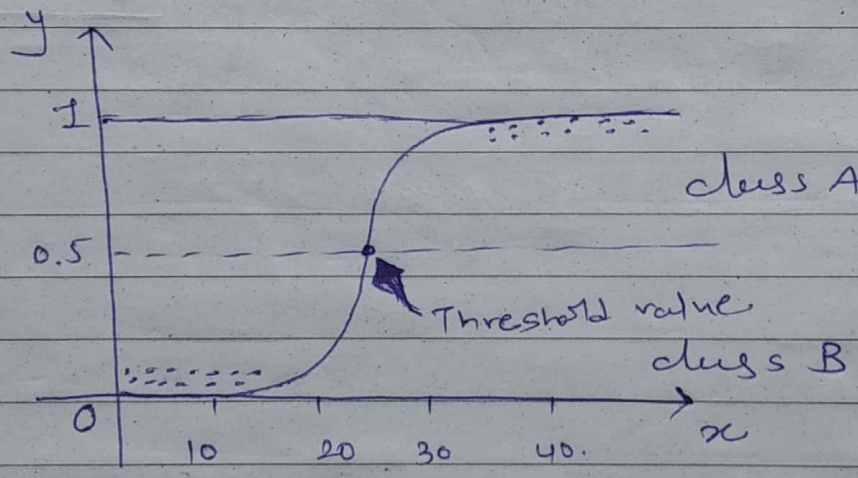
* Logistic Regression :-

→ It gives probabilistic values which lies between 0 and 1.

→ Used to solve classification problems.

→ sigmoid function $p = \frac{e}{1 + e^{-y}}$; $e = 2.718$

sigmoid function is simply trying to convert the independent variable (x) into expression of probability that ranges b/w 0 and 1 w.r.t. the dependent variable.



→ Dataset should be free of missing values.

→ Data points may vary between 30 - 50. for one class. :- 60 - 100 datapoints. Should be there.

→ Eqⁿ : $\log\left[\frac{p}{1-p}\right] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_m X_m$

* Types of logistic regression :-

Binomial : two possible types of dependent variable.
→ Yes or No, 0 or 1, Pass or Fail.

Multinomial : 3 or more possible unordered types of dependent variable.
→ cat, dog, sheep.

Ordinal : 3 or more possible ordered types of dependent variable.
→ low, medium, high.

* K neurest neighbour classification :-

** Ex:- $x = (\text{Maths} = 6, \text{CS} = 8)$, $K = 3$

	Maths	CS	Result
1)	4	3	F
2)	6	7	P
3)	7	8	P
4)	5	5	F
5)	8	8	P.

$$d = \sqrt{|x_{0_1} - x_{A1}|^2 + |x_{0_2} - x_{A2}|^2}$$

observed value (maths: 6, CS: 8)

Actual value

(1) $\sqrt{|6-4|^2 + |8-3|^2} = \sqrt{4+25} = 5.38$

(2) $\sqrt{|6-6|^2 + |8-7|^2} = \sqrt{1} = 1$

(3) $\sqrt{|6-7|^2 + |8-8|^2} = \sqrt{1+0} = 1$

(4) $\sqrt{|6-5|^2 + |8-5|^2} = \sqrt{1+9} = 3.16$

(5) $\sqrt{|6-8|^2 + |8-8|^2} = \sqrt{4} = 2$

→ Here, 2nd, 3rd and 5th data points/neighbours are neurest.

→ 2nd → Pass
3rd → Pass
5th → Pass.

∴ 3 of them are pass. ~~and~~ ~~score of~~.
∴ Probability of pass is greater than fail.

→ Therefore, we can declare ~~x~~ as pass.

* Conditional Probability :-

→ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

ex:- → $P(B) = \frac{30}{100} = 0.3$

$P(A \cap B) = \frac{20}{100} = 0.2$

∴ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$= \frac{0.2}{0.3} = 0.67$

* Baye's Theorem :-

→ $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

likelihood, Prior, hypothesis or data, evidence, Posterior, Marginal

* Naive Bayes Classifier :-

Given :- Fruits = { yellow, Sweet, Long }

Attributes.

Fruit	Yellow	Sweet	Long	Total
Orange	350	450	0	650
Banana	400	300	350	400
Others	50	100	50	150
Total	800	850	400	1200

Sol :- → $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$.

→ Probability of yellow given that my fruit is orange,

$$P(\text{Yellow}/\text{Orange}) = \frac{P(\text{Orange}|\text{Yellow}) \cdot P(\text{Yellow})}{P(\text{Orange})}.$$

orange + yellow = 350

$$= \frac{\frac{350}{800} \times \frac{800}{1200}}{\frac{650}{1200}}$$

total yellow ← 800 ← total yellow fruits

1200 ← total fruits

$$= 0.538$$

→ Probability of sweet given that my fruit is Orange,

$$P(\text{Sweet}/\text{Orange}) = \frac{P(\text{Orange}/\text{Sweet}) \cdot P(\text{Sweet})}{P(\text{Orange})}$$

$$= \frac{\frac{450}{850} \times \frac{850}{1200}}{\frac{650}{1200}}$$

$$= 0.692$$

→ Probability of Long given that my fruit is Orange,

$$P(\text{Long}/\text{Orange}) = \frac{P(\text{Orange}/\text{Long}) \cdot P(\text{Long})}{P(\text{Orange})}$$

$$= \frac{\frac{0}{400} \times \frac{0}{400}}{650/1200}$$

$$= 0.$$

→ Now, Probability of Fruit given that orange is my fruit,

$$P(\text{Fruit}/\text{Orange}) = P(\text{Yellow}/\text{Orange}) \times P(\text{Sweet}/\text{Orange}) \times P(\text{Long}/\text{orange})$$

$$= 0.538 \times 0.692 \times 0$$

$$= 0$$

→ Now, probability of Fruit given that banana is my fruit,

$$P(\text{Fruit} \mid \text{Banana}) = P(\text{Yellow} \mid \text{Banana}) \times P(\text{Sweet} \mid \text{Banana}) \times P(\text{long} \mid \text{Banana})$$

$$\therefore P(\text{Yellow} \mid \text{Banana}) = \frac{P(B \mid Y) \cdot P(Y)}{P(B)}$$

$$= \frac{\frac{400}{800} \times \frac{800}{1200}}{\frac{400}{1200}}$$

$$= 1$$

$$\& \; P(\text{Sweet} \mid \text{Banana}) = \frac{P(B \mid S) \cdot P(S)}{P(B)}$$

$$= \frac{\frac{300}{880} \times \frac{850}{1200}}{\frac{400}{1200}}$$

$$= 0.75$$

$$\& \; P(\text{long} \mid \text{Banana}) = \frac{P(B \mid L) \cdot P(L)}{P(B)}$$

$$= \frac{\frac{350}{400} \times \frac{400}{1200}}{\frac{400}{1200}}$$

$$= 0.875$$

$\therefore$ P(Fruit / Banana) = 1 X 0.75 X 0.875

= 0.656

→ Now, probability of Fruit given that the fruit is from others category,

P(Fruit / others) = P (~~Red~~ Yellow / others) X

P (Sweet / others) X

P (long / others).

= 0.33 X 0.66 X 0.33

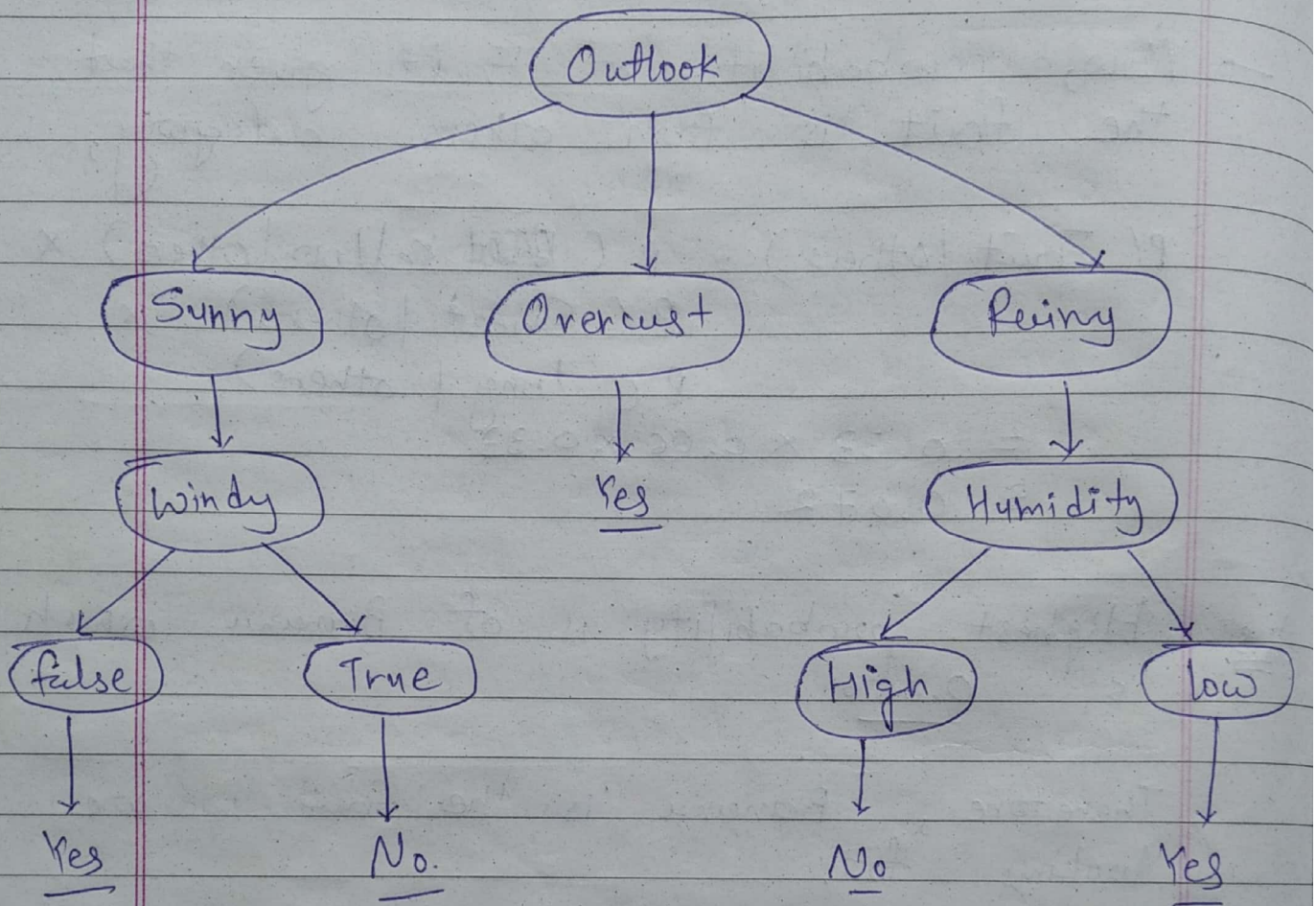
= 0.072

Ans: ⇒ Highest probability is of Banana which is 0.656.

Therefore, Banana is the fruit we are looking for.

* Decision Tree :-

Play Golf



* Entropy :-

$$E(S) = \sum_{i=1}^{c} - P_i \log_2 P_i$$

→ One attribute :-

Play Golf	
Yes	No
9	5

$E(5,9)$ = Entropy $\left(\frac{5}{14}, \frac{9}{14}\right)$

= Entropy $(0.36, 0.64)$

$= -(0.36 . \log_2 0.36) - (0.64 . \log_2 0.64) = \boxed{0.94}$

→ Two attributes :-

		Play Golf: Yes	Play Golf: No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$E(\text{PlayGolf}, \text{Outlook}) = P(\text{Sunny}) \cdot E(3,2) + P(\text{Overcast}) \cdot E(4,0) + P(\text{Rainy}) \cdot E(2,3)$$

$$= \left(\frac{5}{14}\right)\left(-\frac{3}{5}\cdot\log_2\frac{3}{5} - \frac{2}{5}\cdot\log_2\frac{2}{5}\right) + \left(\frac{4}{14}\right)\left(-\frac{4}{4}\cdot\log_2\frac{4}{4} - \frac{0}{0}\cdot\log_2 0\right) + \left(\frac{5}{14}\right)\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\cdot\log_2\frac{3}{5}\right)$$

$$= (0.357)(-0.6\cdot(-0.73) - 0.4\cdot(-1.32)) + (0.285)(0) + (0.357)(-0.4\cdot(-1.32) - 0.6\,(-0.736))$$

$$= 0.693$$

* Example :-

Day	Outlook	Temp	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No.

→ 4 Attributes : Outlook, Temp, Humidity, Wind.

→ Max IG will be considered as root node.

→ 5 No's and 9 Yes's.

(1) → Entropy of entire dataset

→ $E(S) = E(+9, -5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

(S)

$= 0.94$

(2) → Outlook :- Sunny, Overcast, Rain

Yes ↓ No's in Sunny's rows.

→ E(Sunny) = E(2+, 3-)

$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

→ E(Overcast) = E(4+, 0-) = 0

→ E(Rain) = E(3+, 2-)

$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$

→ Gain(S, Outlook)

= E(S) − (5/14) E(Sunny) − (4/14) E(Overcast) − (5/14) E(Rain)

(5 → no. of times Sunny is appearing; 14 → Total total)

$= 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971)$

$= 0.2464$

(3) → Temp :- Hot, Mild, Cool.

→ $E(Hot) = E(2+, 2-)$ → same ⇒ 1

$= -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

→ $E(Mild) = E(4+, 2-)$

$= -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$

→ $E(Cool) = E(3+, 1-)$

$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$

→ Gain (S, Temp)

$= E(S) - \frac{4}{14} E(Hot) - \frac{6}{14} E(Mild) - \frac{4}{14} E(Cool)$

$= 0.94 - \frac{4}{14}(1) - \frac{6}{14}(0.9183) - \frac{4}{14}(0.8113)$

$= 0.0289$

(4) → Humidity :- High, Normal

→ E(High) = E(3+, 4-)

$= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$

→ E(Normal) = E(6+, 1-)

$= -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.5916$

→ Gain(S, Humidity)

$= E(S) - \frac{7}{14}E(High) - \frac{7}{14}E(Normal)$

$= 0.94 - \frac{7}{14}(0.9852) - \frac{7}{14}(0.5916)$

$= 0.1516$

(5) → Wind :- Strong, Weak

→ E(Strong) = E(3+, 3-) = 1

→ E(Weak) = E(6+, 2-) = 0.8113

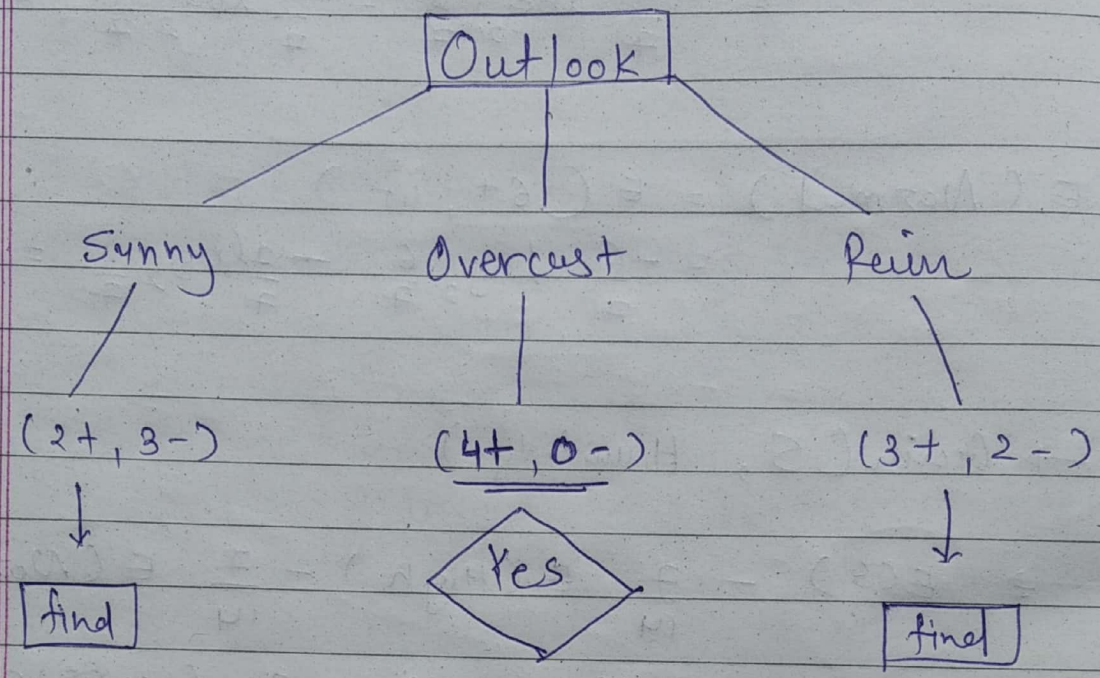
→ Gain(S, Wind)

$= E(S) - \frac{6}{14}E(Strong) - \frac{8}{14}E(Weak)$

$= 0.94 - \frac{6}{14}(1) - \frac{8}{14}(0.8113)$

$= 0.0478.$

→ From the above calculations, Outlook is having maximum entropy.



→ Now, outlook is considered.
Therefore for sunny,

Day	Temp	Humidity	Wind	PlayTennis
1	Hot	High	Weak	No
2	Hot	High	Strong	No
8	Mild	High	Weak	No
9	Cool	Normal	Weak	Yes
11	Mild	Normal	Strong	Yes.

→ We don't need to consider outlook further.

* Confusion Matrix :-

* Consider the confusion matrix given below for a binary classifier predicting the presence of a disease.

The classifier made a total of 150 predictions. out of those 150 cases, the classifier predicted yes 100 times and no 50 times.

In reality, 100 patients in the sample have the disease and 50 patients do not.

	Predicted No	Predicted Yes
Actual No	45 (TN)	5 (FP)
Actual Yes	5 (FN)	95 (TP)

* Accuracy : $\frac{TN + TP}{\text{total predictions}}$ ← total correctness

$$= \frac{45 + 95}{150}$$

$$= 93.33\ \%$$

* Precision : When it predicts yes, how often is it correct ?

* Precision = $\frac{TP}{\text{total predicted yes}} = \frac{95}{100} = 95\%$

* Recall = $\frac{TP}{\text{total Actual Yes}} = \frac{95}{100} = 95\%$

* Example :-

	Predicted No	Predicted Yes
Actual No	2 TN	0 FP
Actual Yes	3 FN	5 TP

→ Accuracy = $\frac{TN + TP}{\text{total predictions}}$

$= \frac{2+5}{10} = \frac{7}{10} \times 100 = 70\%$

→ Precision = $\frac{TP}{\text{total predicted yes } (TP + FP)} = \frac{5}{5} \times 100 = 100\%$

→ Recall = $\frac{TP}{\text{total actual yes } (TP + FN)} = \frac{5}{8} \times 100 = 62.5\%$

→ F1 Score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

* 3 x 3 Confusion matrix :-

		A	B	C	total of actual
		Predicted			total
A	Actual	15	2	3	20.
B		7	15	8	30.
C		2	3	45	50
	total	24	20	56	100

→ Accuracy = $\frac{TP + TN}{\text{total predicted}}$ (Cross axis)

$= \frac{15 + 15 + 45}{100} \times 100$

$= 75\ \%$

→ Precision$_A$ = $\frac{\text{Correctly predicted}}{\text{Total predicted}}$

$= \frac{15}{24} = 0.625.$

Precision$_B$ = $\frac{15}{20} = 0.75$

Precision$_C$ = $\frac{45}{56} = 0.80$

*/ → F-1 score = $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

for(A) $= 2 \times \frac{0.625 \times 0.75}{0.625 + 0.75} = \frac{0.9375}{1.375} = 0.681$

→ Recall = $\frac{\text{Correctly classified}}{\text{Actual}}$

$\text{Recall}_A = \frac{15}{20} = 0.75$

$\text{Recall}_B = \frac{15}{30} = 0.5$

$\text{Recall}_C = \frac{45}{50} = 0.9$

* Cross validation :-

→ It is a statistical method used to estimate the performance (or accuracy) of ml models.

Types of cross validation

Exhaustive	Non-exhaustive
1) Hold out method	1) Leave one out cross validation
2) k-fold cross validation	2) Leave p out cross validation
3) Stratified k-fold cuss validation	

* MSE : Mean Squared Error : $(\text{Actual} - \text{forecast})^2$

ex:

Day	Actual	forecast	Squared error
1	67	70	9
2	50	49	1
3	36	38	4
4	74	76	4
5	84	83	1
6	84	80	16
7	64	67	9
8	34	30	16
9	23	20	9
10	72	75	9
11	62	60	4
12	42	38	16
			98

$$MSE = \frac{1}{n}\Sigma(\text{Act} - \text{Fore})^2$$

$$= \frac{98}{12}$$

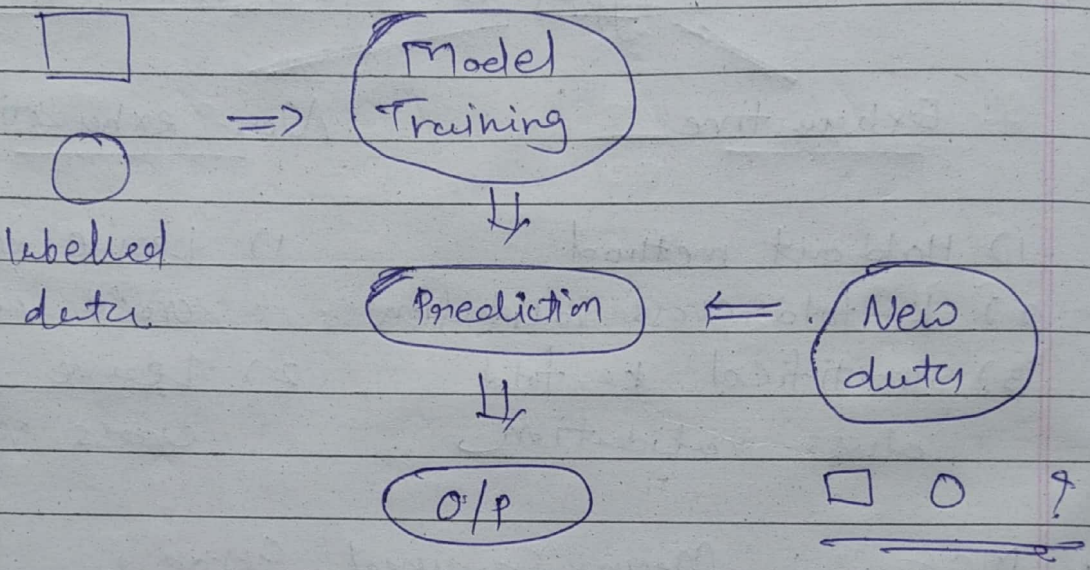
$$= 8.166$$

$$\approx \boxed{8.17}$$

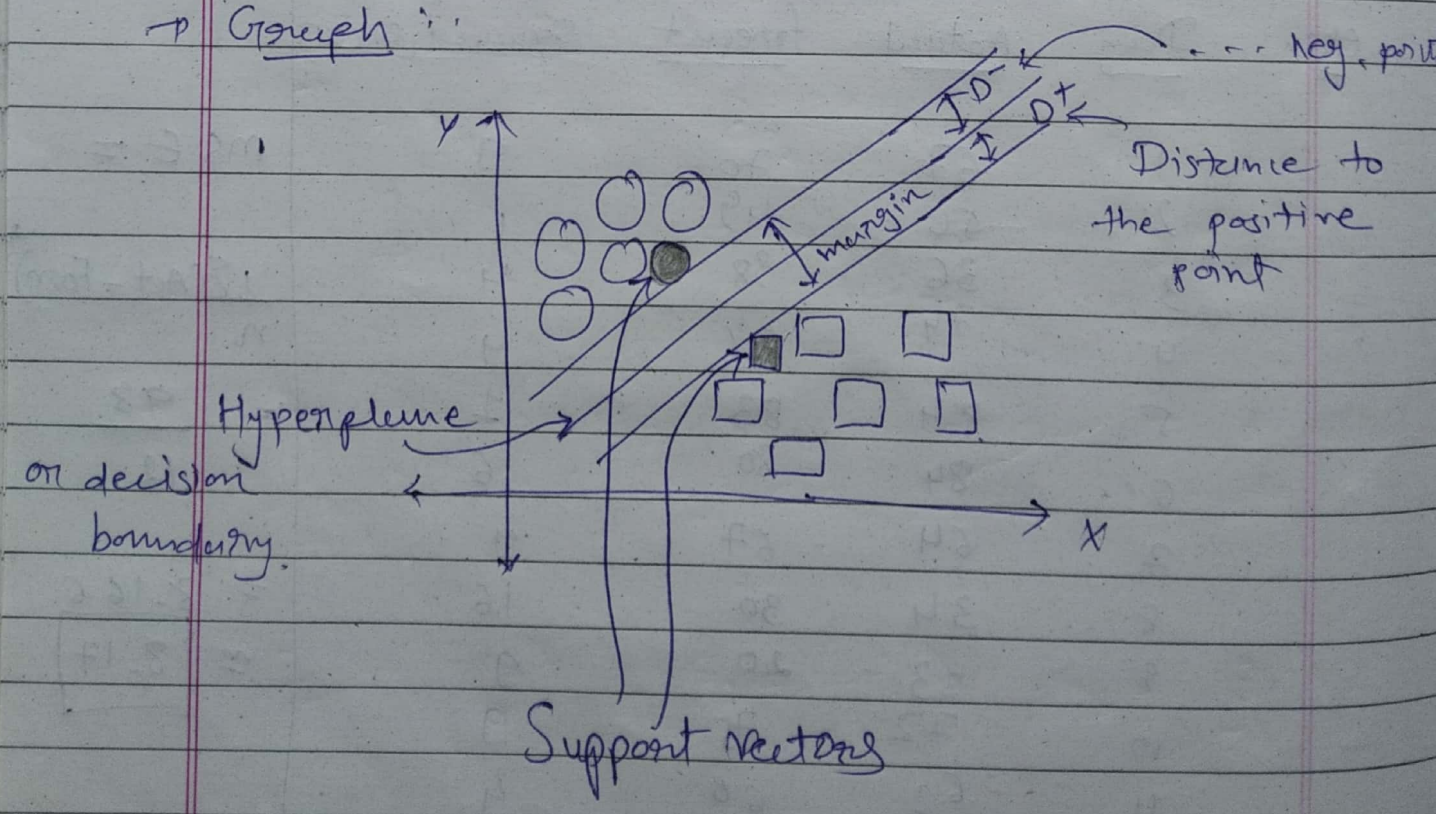
* Support Vector Machine :-

→ Follows Supervised Learning.

→ Diagram :-



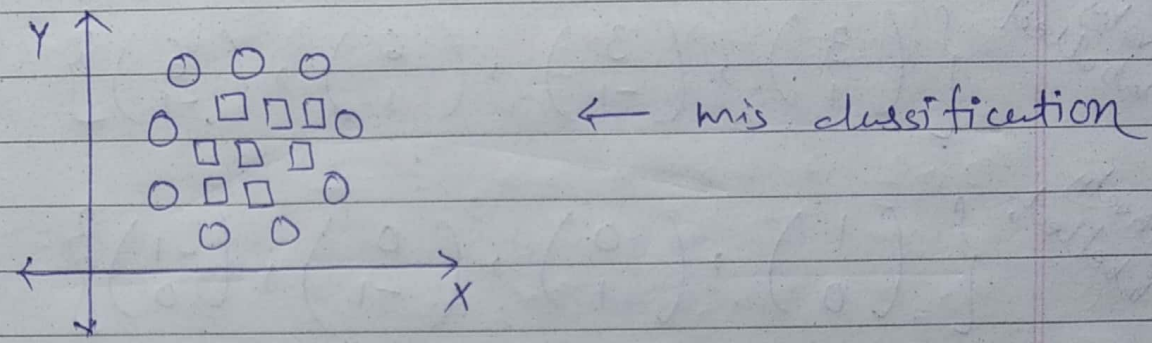
→ Graph :-



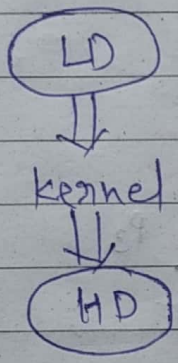
→ Maximum width of hyperplane is chosen to get better result.
or margin

→ Maximal Margin Hyperplane is selected

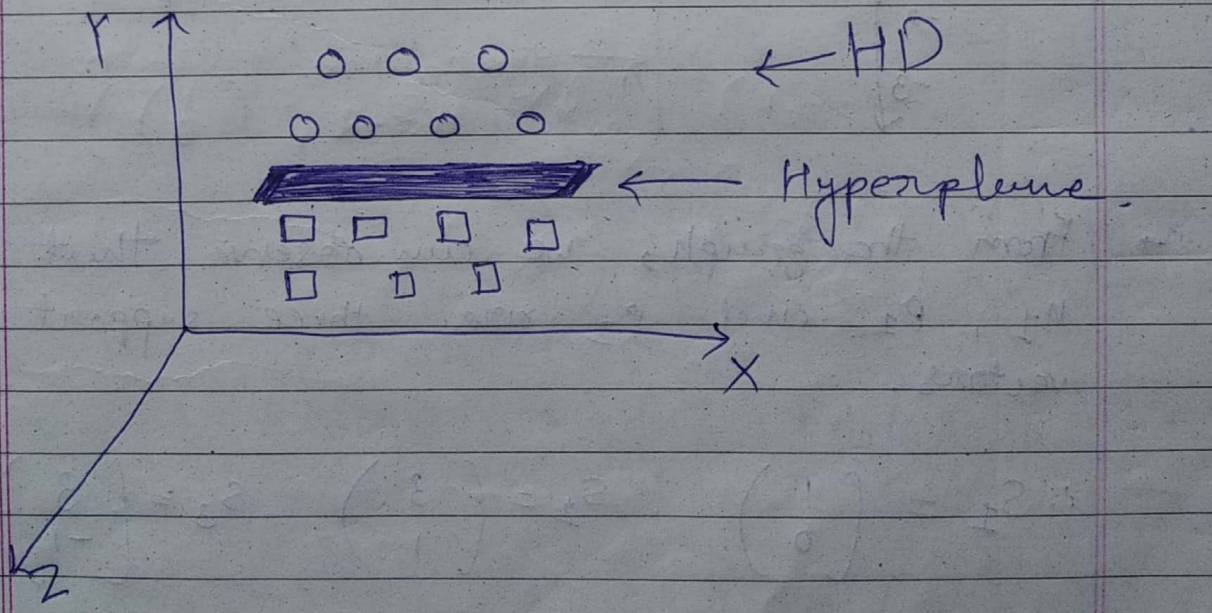
* Non-linear SVM :-



→ To classify this, kernel function is used.



→



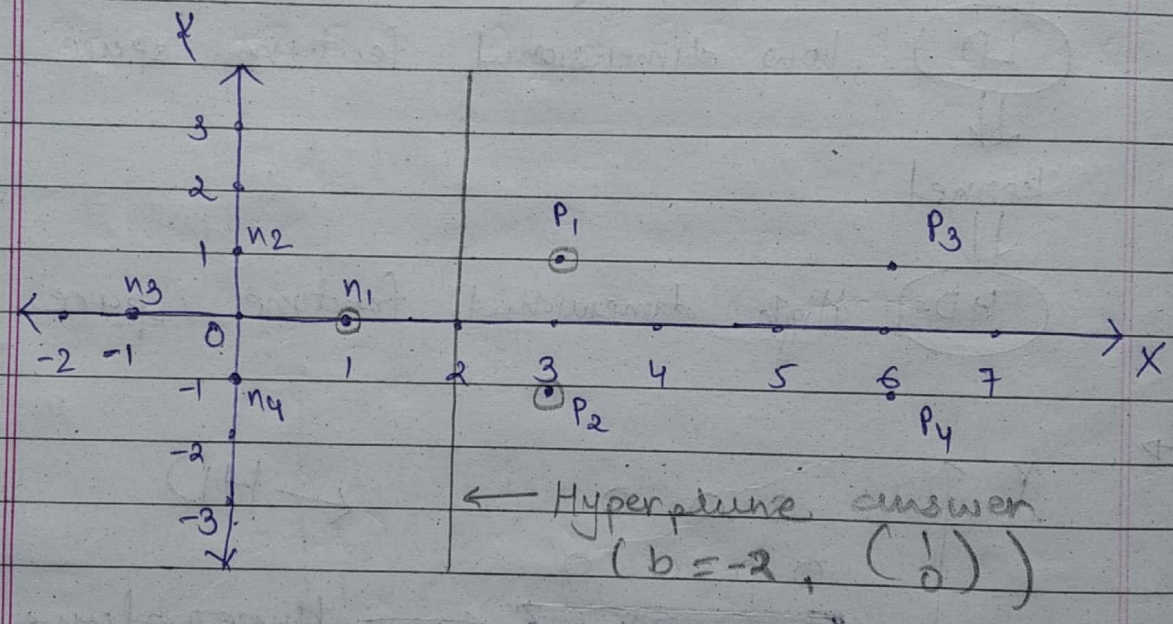
* Linear SVM Examples :-

ex:- Dataset :-

Positively labelled $\left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 6 \\ -1 \end{pmatrix} \right\}$

Negatively labelled $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right\}$

Sol^n :- Plot a graph.



→ From the graph, we can observe that n_1, P_1 and P_2 are three support vectors.

$\therefore S_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $S_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$, $S_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$

→ Now, augment each vector with a '1' as a bias input.

$$S_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow \bar{S}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \leftarrow 1 \text{ is added}$$

$$S_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \Rightarrow \bar{S}_2 = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix} \Rightarrow \bar{S}_3 = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

→ Calculate α,

S_1 is common

$$\alpha_1 \bar{S}_1 \cdot \bar{S}_1 + \alpha_2 \bar{S}_2 \cdot \bar{S}_1 + \alpha_3 \bar{S}_3 \bar{S}_1 = -1$$

($\because S_1$ is present in the other side)

&

$$\alpha_1 \bar{S}_1 \cdot \bar{S}_2 + \alpha_2 \bar{S}_2 \bar{S}_2 + \alpha_3 \bar{S}_3 \bar{S}_2 = +1$$

&

$$\alpha_1 \bar{S}_1 \cdot \bar{S}_3 + \alpha_2 \bar{S}_2 \bar{S}_3 + \alpha_3 \bar{S}_3 \cdot \bar{S}_3 = +1$$

$$\therefore \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = -1$$

&

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = +1$$

&

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = +1$$

$\therefore \alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3+0+1) = -1$

& $\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = 1$

& $\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = 1$

$\therefore 2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$

& $4\alpha_1 + 11\alpha_2 + 9\alpha_3 = 1$

& $4\alpha_1 + 9\alpha_2 + 11\alpha_3 = 1$

$\therefore \alpha_1 = -3.5 \quad , \alpha_2 = 0.75 \quad , \alpha_3 = 0.75$

→ Calculate weight vector :-

$$\bar{w} = \sum_{i=1}^{n} \alpha_i \bar{s}_i$$

$$= (-3.5)\begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\-1\\1\end{pmatrix}$$

$$= \begin{pmatrix}1\\0\\-2\end{pmatrix}$$

it's a bias

→ We can equate the last entry in $\bar{w}$ as the hyperplane offset b and write the separating hyperplane equation :

$y = wx + b$

where, $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$b = -2$

if $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow$ parallel to y axis

$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow$ parallel to x axis

$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow 45^\circ$ with respect to x and y axis.

* Support Vector Machine :-

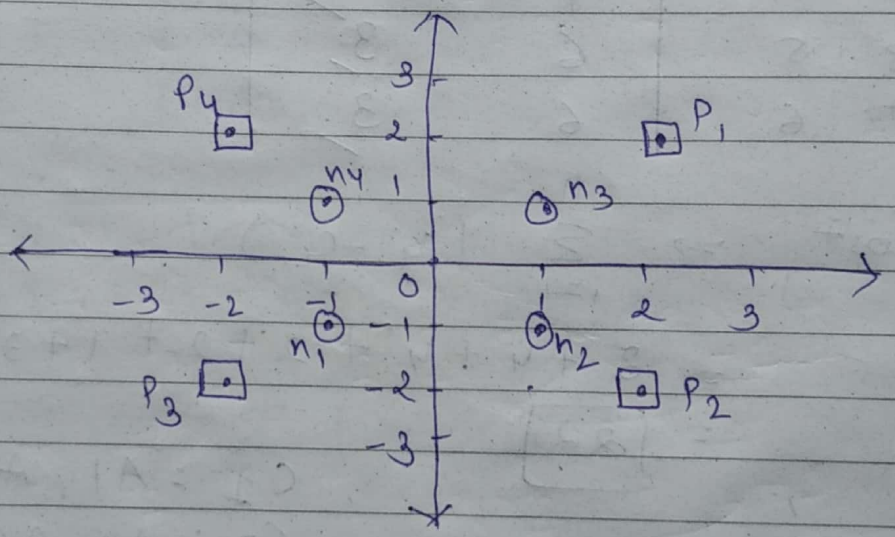
Non-Linear

Ex →

positively labelled : $\left\{ \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \end{pmatrix} \right\}$

Negatively labelled : $\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$

Soln.



→ To convert these points into non-linear plane, this formula is used.

$$\phi_1 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 4 - x_2 + |x_1 - x_2| \\ 4 - x_1 + |x_1 - x_2| \end{pmatrix}; & \text{if } \sqrt{x_1^2 + x_2^2} > 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases}$$

→ For positively labelled,

$$\begin{pmatrix} 2 \\ 2 \end{pmatrix} \Rightarrow \begin{pmatrix} 4-2+0 \\ 4-2+0 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ -2 \end{pmatrix} \Rightarrow \begin{pmatrix} 4-(-2)+|2-(-2)| \\ 4-2+|2-(-2)| \end{pmatrix} = \begin{pmatrix} 10 \\ 6 \end{pmatrix}$$

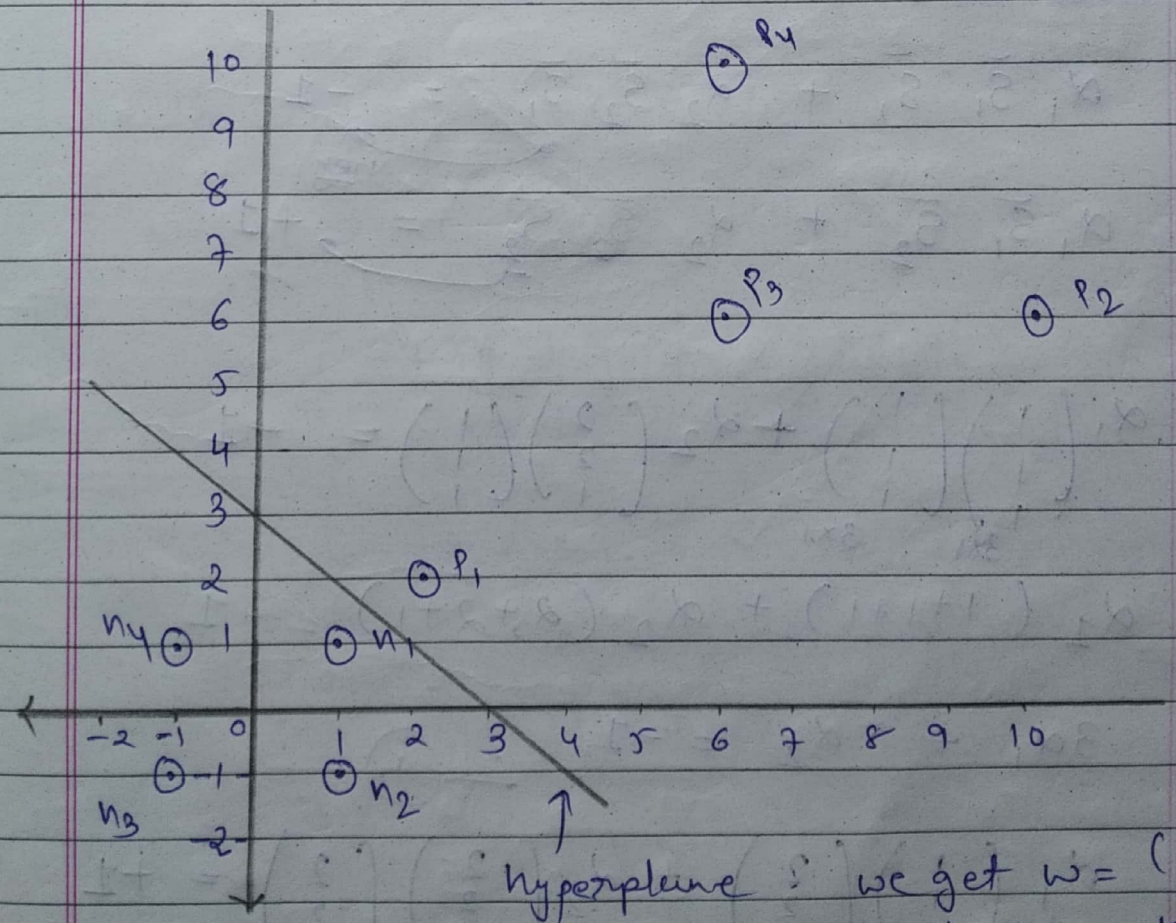
$$\begin{pmatrix} -2 \\ -2 \end{pmatrix} \Rightarrow \begin{pmatrix} 4-(-2)+|-2-(-2)| \\ 4-(-2)+|-2-(-2)| \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$$

$$\begin{pmatrix} -2 \\ 2 \end{pmatrix} \Rightarrow \begin{pmatrix} 4-2+|-2-2| \\ 4+2+|-2-2| \end{pmatrix} = \begin{pmatrix} 6 \\ 10 \end{pmatrix}$$

→ For negatively labelled,

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix} \Rightarrow \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ 1 \end{pmatrix} \Rightarrow \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$



Hyperplane : we get $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ which means that it will make 45° angle with x and y axis.

→ Here, n_1 and p_1 are nearest to each other belongs from two different labelles.

$$\therefore \left\{ S1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad S2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}$$

→ Add 1 as a bias input.

$$\tilde{S1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \tilde{S2} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$$

→ find α,

$$\alpha_1 \tilde{S_1} \tilde{S_1} + \alpha_2 \tilde{S_2} \tilde{S_1} = -1$$

neg

$$\alpha_1 \tilde{S_1} \tilde{S_2} + \alpha_2 \tilde{S_2} \tilde{S_2} = +1$$

$$\therefore \alpha_1 \underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}_{3\times1} \underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}_{3\times1} + \alpha_2 \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -1$$

$$\therefore \alpha_1 (1+1+1) + \alpha_2 (2+2+1) = -1$$

$$\therefore 3\alpha_1 + 5\alpha_2 = -1 \quad \text{---(1)}$$

and $$\alpha_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} = +1$$

$\therefore a_1(2+2+1) + a_2(4+4+1) = +1$

$\therefore 5a_1 + 9a_2 = 1$ — (2)

~~$\therefore -5a_1 + 9a_2 = -1$~~

$\therefore$ (1) $\times 5$ - (2) $\times 3$.

$\therefore 15a_1 + 25a_2 = -5$

$= 15a_1 + 27a_2 = -3$

$-2a_2 = -8$

$\therefore \boxed{a_2 = 4}$

$\therefore 5a_1 + 9(4) = 1$

$\therefore 5a_1 = -35 \Rightarrow \boxed{a_1 = -7}$

→ Calculate weight vector :-

$$\vec{w} = \sum_{i=1}^{n} a_i \vec{s_i}$$

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

$$= \begin{pmatrix}-7\\-7\\-7\end{pmatrix} + \begin{pmatrix}8\\8\\4\end{pmatrix}$$

$$= \begin{pmatrix}1\\1\\-3\end{pmatrix}$$

bias input = b.

$\therefore y = wx + b$

$w = \begin{pmatrix}1\\1\end{pmatrix}$

* K-means clustering :-

→ Suppose that the data mining task is to cluster points into three clusters,

where points are;

A1 (2,10), A2 (2,5), A3 (8,4).
B1 (5,8), B2 (7,5), B3 (6,4).
C1 (1,2), C2 (4,9).

→ The distance function is Euclidean distance.

Soln :- A1, B1, C1 are the centre of each cluster, (if not given, we can select).

∴ Initial Centroids : A1, B1, C1.

	Datapoints x_1 y_1	Distance to Centroid x_2 y_2: (2, 10)	x_2 y_2: (5, 8)	x_2 y_2: (1, 2)	Cluster	New Cluster
A1	2 10	0	3.61	8.06	1	
A2	2 5	5	4.24	3.16	(3)	
A3	8 4	8.49	5	7.28	2	
B1	5 8	3.61	0	7.21	2	
B2	7 5	7.07	3.61	6.71	2	
B3	6 4	7.21	4.12	5.39	2	
C1	1 2	8.06	7.21	0	(3)	
C2	4 9	2.24	1.41	7.62	2	
cluster no.		(1)	(2)	(3)		

→ Now, Calculate Euclidean's distance using formula,

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

→ Now, smallest cluster will be assigned in each row.

ex:- for A1 : 0, 3.61, 8.06

0 is the smallest ⇒ cluster (1) will be assigned.

for A2 : 5, 4.24, 3.16 ⇒ 3.16 is smallest

∴ cluster (3) will be assigned

like wise for all Datapoints.

→ After assigning all the clusters, we need to calculate new centroids for each one.

→ for cluster no. (3) :

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

→ for cluster no. (2) :

$$\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$$

$$\Rightarrow (6, 6)$$

→ for cluster no. (1) :

$$\left(\frac{2}{1} + \frac{10}{1}\right) \Rightarrow (2, 10).$$

∴ Current Centroids : A1 (2, 10)
B1 (6, 6)
C1 (1.5, 3.5)

→ Now, using current centroids, we need to perform the same operations.

old clusters

	Datapoints x_1	y_1	Distance to Centroid 2	10	6	6	1.5	3.5	Cluster	New cluster
A1	2	10	0		5.66		6.52		1	(1)
A2	2	5	5		4.12		1.58		3	[3]
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.7		2	2
B2	7	5	7.07		1.41		5.7		2	2
B3	6	4	7.21		2		4.53		2	2
C1	1	2	8.06		6.4		1.58		3	[3]
C2	4	9	2.24		3.61		6.04		~~1~~ 2	(1)

→ Here, cluster is what we found in our first round and new cluster is the one having smallest distance.
∴ Assign numbers to new clusters.

→ C2 is moving from cluster 2 to 1
∴ We need to again calculate the new centroids.

→ for cluster no (3),

(1.5, 3.5)

for cluster no (2),

$$\left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) \Rightarrow (6.5, 5.25)$$

for cluster (1),

$$\left(\frac{2+4}{2}, \frac{10+9}{2}\right) \Rightarrow (3, 9.5)$$

→ Current centroids, A1 (3, 9.5)
B1 (6.5, 5.25)
C1 (1.5, 3.5)

Datapoints		Distance to centroid			Cluster	New Cluster
		3 9.5	6.5 5.25	1.5 3.5		
A1	2 10	1.12	6.54	6.52	1	1
A2	2 5	4.61	4.51	1.58	3	3
A3	8 4	7.43	1.95	6.52	2	2
B1	5 8	2.5	3.13	5.7	2	1
B2	7 5	6.62	0.56	5.7	2	2
B3	6 4	6.26	1.35	4.53	2	2
C1	1 2	7.76	6.39	1.58	3	3
C2	4 9	1.12	4.51	6.04	1	1

↑ New clusters of last round.

→ Again, cluster 2 is moved to 1.

∴ Calculate new centroids.

A1 (3.67, 9)
B1 (7, 4.33)
C1 (1.5, 3.5)

	Datapoints	Distance to Centroid			Cluster	New Cluster
		3.67 9	7 4.33	1.5 3.5		
A1	2 10	1.94	7.56	6.52	1	1
A2	2 5	4.33	5.04	1.58	3	3
A3	8 4	6.62	1.05	6.52	2	2
B1	5 8	1.67	4.18	5.7	1	1
B2	7 5	5.21	0.67	5.7	2	2
B3	6 4	5.52	1.05	4.53	2	2
C1	1 2	7.49	6.44	1.58	3	3
C2	4 9	0.33	5.55	6.04	1	1

→ Here, All the clusters and new clusters are same.

∴ (1)st cluster : A1, B1, C2

(2)nd cluster : A3, B2, B3

(3)rd cluster : A2, C1.

* Nearest Neighbour Clustering :-

ex:- Dataset is of points (A, B, C, D, E)

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

Column wise

distance given

threshold = 2

Soln →

$K_1 = \{A, B, C, D\}$, $K_2 = \{E\}$

A ↑ initially added

$\leq t$

→ For threshold = 1 :

$K_1 = \{A, B\}$ ($\because dist(A,B) \leq t$)

A ↑ initially added

$K_2 = \{C, D\}$ ($\because dist(C,D) \leq t$)

$K_3 = \{E\}$

ex:

A1 = (2,10)	A4 = (5,8)	A7 = (1,2)	
A2 = (2,5)	A5 = (7,5)	A8 = (4,9)	
A3 = (8,4)	A6 = (6,4)	t = 4	

Soln:

A1 is placed in a cluster by itself.

$\therefore K_1 = \{A1\}$

A2 → $d(A1, A2) = \sqrt{25} = 5 > t$

$\therefore K_2 = \{A2\}$

A3 → $d(A1, A3) = \sqrt{36+36} = 8.48 > t$

$d(A2, A3) = \sqrt{37} > t$ ⇓ form a new cluster

$\therefore K_3 = \{A3\}$

A4 → $d(A1, A4) = \sqrt{13} = < t$

$d(A2, A4) = \sqrt{18}$

$d(A3, A4) = \sqrt{25}$

⇓ Add into already initiated cluster having nearest point or smallest distance

$\therefore K_1 = \{A1, A4\}$

A5 → $d(A1, A5) = \sqrt{50}$

$d(A2, A5) = \sqrt{25}$

$d(A3, A5) = \sqrt{2} \leq t$

$d(A4, A5) = \sqrt{13}$

$\therefore K_3 = \{A_3, A_5\}$

$A_6 \rightarrow d(A_5, A_6) = \sqrt{2} < t$

$d(A_3, A_6) = \sqrt{2} < t$

$\therefore K_3 = \{A_3, A_5, A_6\}$

$A_7 \rightarrow d(A_2, A_7) = \sqrt{10} < t$

$\therefore K_2 = \{A_2, A_7\}$

$A_8 \rightarrow d(A_8, A_4) = \sqrt{2} < t$

$\therefore K_1 = \{A_1, A_4, A_8\}$

if $> t \Rightarrow$ form a new cluster
else if $\leq t$
$\Rightarrow$ Add into an already assigned cluster (nearest distance)

$\therefore$ Clusters :-

$K_1 = \{A_1, A_4, A_8\}$

$K_2 = \{A_2, A_7\}$

$K_3 = \{A_3, A_5, A_6\}$

* Clusters Using Single and Complete Link Clustering :-

Agglomerative Clustering ::

→ Consider the following set of 6 one dimensional data points.

18, 22, 25, 42, 27, 43.

→ Apply Agglomerative hierarchical clustering algorithm to build the hierarchical clustering dendogram.

→ Merge the clusters using Min distance and update the proximity matrix accordingly

→ 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) ← min distance
43	25	21	18	16	1	0

Single Link

Date
Page

Sol^{n} :- Min distance among all is 1.
which is b/w 42 and 43.
∴ Merge 42 and 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	min of (17,15)
27	9	5	2	0	15	
42,43	24	20	17	15	0	

→ Now, 2 is min dist. b/w 25 and 27.
∴ merge both of them.

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

→ 3 is min dist. ∴ merge 22, ~~(0,)~~ 25,27
b/w 22 and 25,27

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

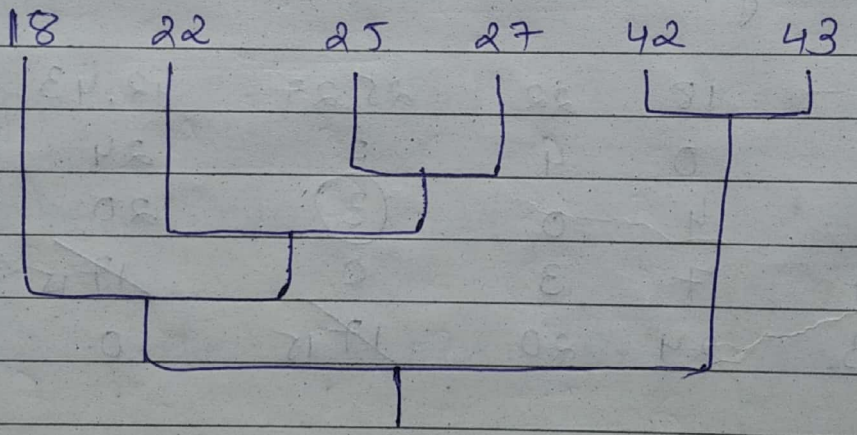
→ 4 is min dist b/w 18, and 22,25,27.

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

→ min = 15 ∴ 18,22,25,27 + 42,43.

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

∴ dendogram :-



* Complete link :

distance mat:

	1	2	3	4	5
1	0	4	7	9	(1)
2	4	0	3	5	3
3	7	3	0	2	6
4	9	5	2	0	8
5	1	3	6	8	0

Complete Link

Date
Page

Sol^{n}:- min dist = 1. ∴ merge 1 and 5. b/w 1 & 5. and TAKE MAX of two distance except (0).

	1,5	2	3	4
1,5	0	4	7	9
2	4	0	3	5
3	7	3	0	(2)
4	9	5	2	0

→ min is 2 b/w 3 and 4
∴ merge 3 and 4.

	1,5	2	3,4
1,5	0	9	7
2	9	0	3
3,4	7	3	0

	1,5	2	3,4
1,5	0	(4)	9
2	4	0	5
3,4	9	5	0

max

→ min = 4 b/w 1,5 and 2 ∴ merge them.

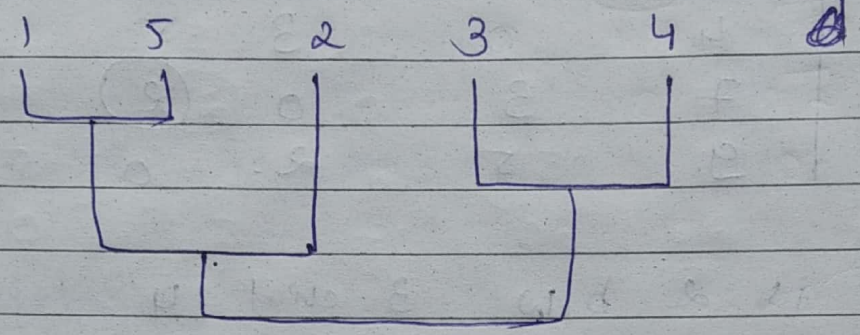
	1,5,2	3,4
1,5,2	0	9
3,4	(9)	0

∴ min = 9 b/w 1,5,2 & 3,4.

	1, 5, 2, 3, 4
1, 5, 2, 3, 4	0

∴ 1 — 2 — 3 — 4 — 5 — dendrogram

∴



* Average Link :-

	A	B	C	D	E
A	0	(1)	(2)	(2)	(3) ← Avg
B	1	0	(2)	(5)	(4)
C	2	2	0	5	4
D	2	5	3	0	6
E	3	4	6	6	0

So, ① →

	A, B	C	D	E
A, B	0	(2)	(3.5)	(3.5)
C	2	0	5	4
D	3.5	5	0	6
E	3.5	4	6	0

(2)

	A, B, C	D	E	
A, B, C	0	4.25	3.75	min
D	4.25	0	6	
E	3.75	6.	0.	

Avg

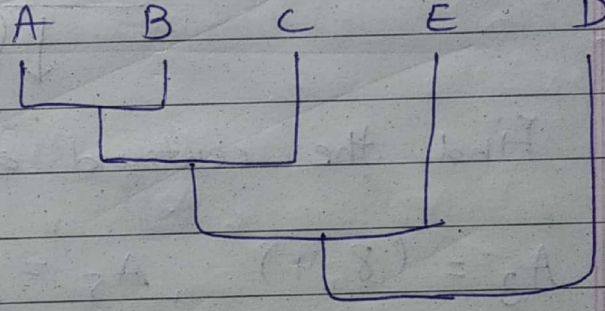
(3)

	A, B, C, E	D.	
A, B, C, E	0	5.125	min.
D	5.125	0.	

(4)

	A, B, C, E, D
A, B, C, E, D	0.

dendrogram



* Centroid Link :-

A1 = (2, 10)
A2 = (2, 5)
A3 = (8, 4)
A4 = (5, 8)
A5 = (7, 5)
A6 = (6, 4)
A7 = (1, 2)
A8 = (4, 9)

→ dist. matrix:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

→ {A1} {A2} {A3, A5, A6} {A4, A8} {A7}.

Dt 52

↓ (B) ↓ (C)

→ Find the centroid of these two clusters

$A_3 = (8,4)$, $A_5 = (7,5)$, $A_6 = (6,4)$

$\therefore$ Centroid $= \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right)$

$\therefore$ B = (7, 4.33)

$A_4 = (5,8)$ $A_8 = (4,9)$

$\therefore$ Centroid $= \left(\frac{5+4}{2}, \frac{8+9}{2}\right) = (4.5, 8.5)$

$\therefore$ C = (4.5, 8.5)

→ $d(A1, B) = \sqrt{(7-2)^2 + (4.33-10)^2}$

$D_t \leq 3$

A1 : (2, 10) $= 7.55$

B : (7, 4.33)

$d(A1, C) = \sqrt{(4.5-2)^2 + (8.5-10)^2}$

A1 : (2, 10) $= 2.91 \leq 3$

C : (4.5, 8.5)

∴ $\{A_1, A_4, A_8\}$ $\{A_2\}$ $\{A_3, A_5, A_6\}$ $\{A_7\}$

dist $= \sqrt{10}$

$= 3.16 > 3$

→ Centroid of A1, A4, A8

$D_t \leq 4$ $= \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right)$

D = (3.66, 9)

~~cancel (centroid of A2, A7~~

$d(A2, D) \Rightarrow D_t \leq 4)$

∴ → $\{A1, A4, A8\}$ $\{A2, A7\}$ $\{A3, A5, A6\}$

Ans

D E B

$d(D, B) = 5.74$ E = (1.5, 3.5) (centroid)

$d(D, E) = 5.9$

* DBSCAN :-

(Density based Clustering method)

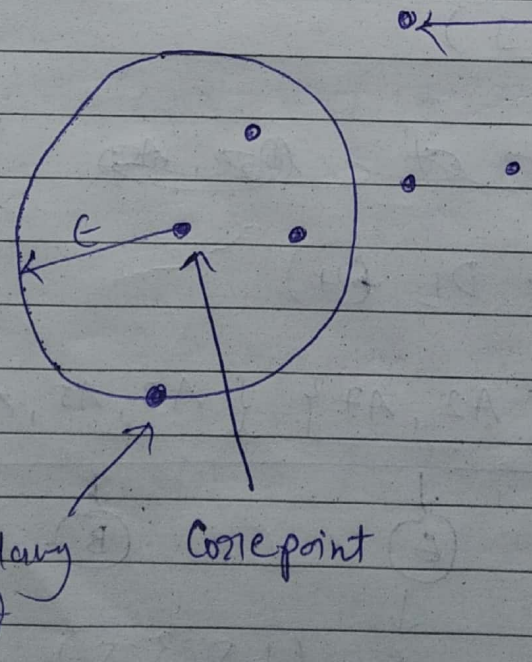
→ ϵ , MinPts ($\epsilon > 0$)

↑ Radius ↑ Neighbourhood points within ϵ.

→ Core object : got atleast minPts objects or points within ϵ (including itself)

→ Boundary points : An object or a point which is not a core point but it is in neighbourhood of the core point within radius ϵ.

→ Noise/outlier : Neither a core point nor a boundary point



Ex:- $\epsilon = 3.5$, MinPts = 3

A1 = (5, 7)	A5 = (3, 7)
A2 = (8, 4)	A6 = (6, 7)
A3 = (3, 3)	A7 = (6, 1)
A4 = (4, 4)	A8 = (5, 5).

→ dist. matrix:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0							
A2	4.24	0						
A3	4.47	5.1	0					
A4	3.16	4	1.41	0				
A5	2	5.83	4	3.16	0			
A6	1	3.61	5	3.61	3	0		
A7	6.08	3.61	3.61	3.61	6.71	6	0	
A8	2	3.16	2.83	1.41	2.83	2.24	4.12	0

Sol:- (1) → Points : Their Neighbours within 3.5

A1	A4, A5, A6, A8
A2	A8
A3	A4, A8
A4	A5, A8, A1, A3
A5	A6, A8, A1, A4
A6	A8, A1, A5
A7	None
A8	A1, A3, A4, A5, A6, A2

Check Row wise & Column wise

(2) → Checking whether each point is core point or not:

If a point has minimum / atleast MinPts points (including itself) in the radius ϵ, it is declared as core point.

~~S1, S3, S4, S5,~~

∴ A1, A3, A4, A5, A6, A8 are corepoints.
A3 ← (∵ including itself)

and A2, A7 are Noise.

(3) → Now, checking whether A2 and A7 can be qualified as a boundary point ?

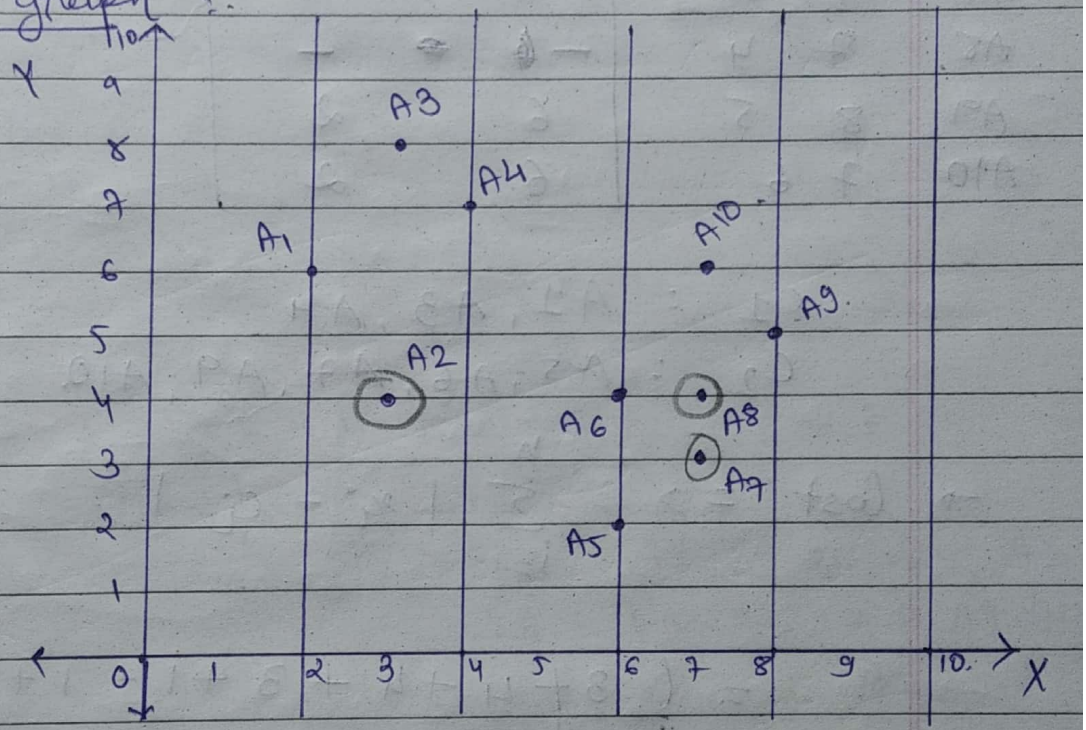
→ Here, A2 is in the vicinity of A8
∴ It can be considered as a boundary point.
while & A7 is not the neighbour of any point (Core point).

∴ A2 → Boundary Point
A7 → Noise

* Medoids Link :-

A1	= (2,6)	A6	= (6,4)
A2	= (3,4)	A7	= (7,3)
A3	= (3,8)	A8	= (7,4)
A4	= (4,7)	A9	= (8,5)
A5	= (6,2)	A10	= (7,6).

① → Plot a graph :-



→ Select two random (within two diff. cluster observing from the graph) objects.

$C_1 = (3,4)$
$C_2 = (7,4)$.

② → Now, create a table of distance from all objects to C_1 and C_2.

(Manhattan distance $|x_1-x_2|+|y_1-y_2|$ is used)

$C_1 = (3,4)$
$C_2 = (7,4)$

	Datapoints	distance from C_1	distance from C_2	Cluster no.	Cluster
A1	2 6	3	7	min →	C_1
A2	3 4	—	—	—	
A3	3 8	4	8		C_1
A4	4 7	4	6		C_1
A5	6 2	5	3		C_2
A6	6 4	3	1		C_2
A7	7 3	5	1		C_2
A8	7 4	—	—	—	
A9	8 5	6	2		C_2
A10	7 6	6	2		C_2

C_1 : A1, A3, A4

C_2 : A5, A6, A7, A9, A10

→ Cost => $\sum_{i=1}^{n} |x_i - c_i|$

$= (3+4+4+3+1+1+2+2)$

$= \boxed{20}$

→ Again we need to select one more point as C_3.

(Choose such a point which is near to either C_1 or C_2.)

Here, let $C_3 = (7,3)$.

$C_1 = (3,4)$
$C_3 = (7,3)$.

	Datapoints		distance from C_1	distance from C_3		cluster
A1	2	6	3	8	→ min	C_1
A2	3	4	—	—		—
A3	3	8	4	9		C_1
A4	4	7	4	7		C_1
A5	6	2	5	2		C_2
A6	6	4	3	2		C_2
A7	7	3	—	—		—
A8	7	4	4	1		C_2
A9	8	5	6	3		C_2
A10	7	6	6	3		C_2

$\therefore$ Cost $\Rightarrow \sum_{i=1}^{4} |x_i - c_i)$ . .

$= 3 + 4 + 4 + 2 + 2 + 1 + 3 + 3$

$= \boxed{22}$

C_1 : A1, A3, A4, A2
C_3 : A5, A6, A8, A9, A10, A7

→ Here, cost increases.

$\therefore$ Previously selected C1 and C2 were best.
$\therefore$ C1 and C2 are center of clusters.

Clusters
$C_1 = \{(2,6), (3,8), (4,7), \underbrace{(3,4)}_{C_1 \text{ itself}}\}$

$C_2 = \{(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)\}$.

* Association Rules :-

IMP Apriori Algorithm :-

→ Here are a dozen sales transactions.

→ The objective is to use this transaction data to find affinities b/w products, that is, which products sell together often.

→ The support level will be set at 33 percent. The confidence level will be set at 50 percent.

→ Rule $X \rightarrow Y$:

$$\text{Support} = \frac{\text{freq}(x, y)}{N}$$

$$\text{Confidence} = \frac{\text{freq}(x, y)}{\text{freq}(x)}$$

Example : Min support = 50 %.
Threshold confidence = 70 %.

Transaction ID	Items
T_1	1 3 4
T_2	2 3 5
T_3	1 2 3 5
T_4	2 5

Confidence = $\frac{S(A \cup B)}{S(A)}$ | support = $\frac{\text{freq}}{\text{total transaction}}$

Date
Page

Soln :

min support = 50%
confidence = 70%

(1)

item set	support	
1	2/4 = 50%	✓
2	3/4 = 75%	✓
3	3/4 = 75%	✓
4	1/4 = 25%	
5	3/4 = 75%	✓

∴ {1, 2, 3, 5}.

(2)

itemset	support	
{1, 2}	1/4 = 25%	
{2, 3}	2/4 = 50%	✓
{3, 5}	2/4 = 50%	✓
{5, 1}	1/4 = 25%	
{1, 3}	2/4 = 50%	✓
{2, 5}	3/4 = 75%	✓

∴ {1, 2, 3, 5}.

(3)

itemset	support	
{1, 2, 3}	1/4 = 25%	
{1, 2, 5}	1/4 = 25%	
{2, 3, 5}	2/4 = 50%	✓
{1, 3, 5}	1/4 = 25%	

∴ {2, 3, 5}.

(4) Find Association Rule :-

∴ Confidence = $\frac{S(A \cup B)}{S(A)}$.

For, $(2 \wedge 3) \rightarrow 5 = \frac{S((2 \wedge 3) \cup 5)}{S(2 \wedge 3)}$ ← Support of 2, 3 only.

(Support of 2, 3, 5.)

$= \frac{2}{2} = 100\%$

{2, 3, 5}

Rules	support	Confidence
(2 ^ 3) → 5	2	2/2 = 100 %
(2 ^ 5) → 3	2	2/3 = 66 %
(3 ^ 5) → 2	2	2/2 = 100 %
5 → (2 ^ 3)	2	2/3 = 66 %
3 → (2 ^ 5)	2	2/3 = 66 %
2 → (3 ^ 5)	2	2/3 = 66 %

divide by support of

→ Confidence threshold is 70 %.

∴ (2 ^ 3) → 5 and (3 ^ 5) → 2 are applicable rules.

– By Rushik Rathod