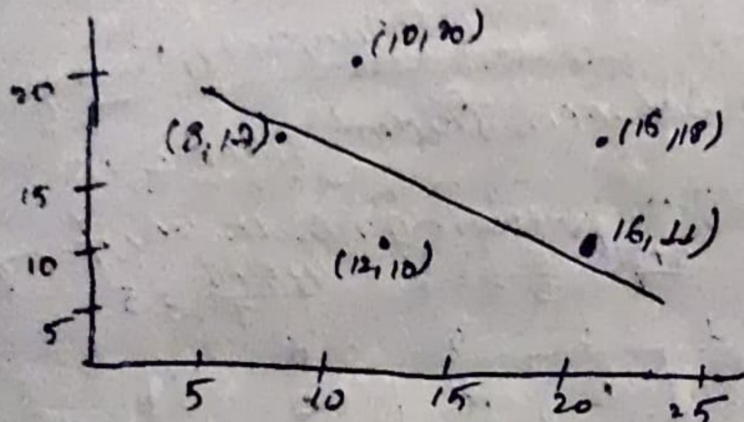
Pearson Correlation

Lecture-4 (Monday)
15th Jan, 2024,

Example: 2

Science	16	15	12	10	8
Geometry	11	18	10	20	17



Spearman's rank coefficient (r_s)

Q. Whether the values pearson correlation values are differ from spearman's rank coefficient.

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

D= difference both the ranks.

Ex:2 → (different answer)

	5	4	3	2	1
S	16	15	12	10	7
G	11	18	10	20	17
	2	4	1	5	3

Set rank

D-2   2-4
= -2

Ans: (-0.3)

Soln:

S	G	r_s	r_g	d	d^2
16	11	5	2	3	9
15	18	4	4	0	0
12	10	3	1	2	4
10	20	2	5	-3	9
8	17	1	3	-2	4
					26

$$r_s = \frac{1-6\times 26}{5(5^2-1)} =$$

Any value that is ordered we can use Spearman's rank coefficient correlation.

Other one which is distributed we can use Pearson's correlation.

Lecture-5

16th Jan, 2024, Tuesday

In the study of relation between student grades in Mathematics and science, the following results were found for some students.

$\frac{2+3}{2} = \frac{5}{2} = 2.5$ $\frac{4+5+6}{3} = 5$

	4	2	1	3	5	6
Mathematics	D	B	A	B	D	D
Science	C	C	B	A	C	F
	3	4	2	1	5	6

$\frac{3+4+5}{3} = 4$

find the spearman correlation coefficient.

$$\boxed{r_s = 1 - \frac{6\Sigma D^2}{n(n^2-1)}}$$

⇒ Soln:

M	S	r_m	r_s	d	d^2
D	C	5	4	1	1
B	C	2.5	4	1.5	2.25
A	B	1	2	-1	1
B	A	2.5	1	1.5	2.25
D	C	5	4	1	1
D	F	5	6	-1	1
					8.5

$D^2 = 1^2 + (1.5)^2 + 1^2 + (1.5)^2 + 1^2 + 1^2$
$= 1 + 2.25 + 1 + 2.25 + 1 + 1$
$D^2 = 8.5$

$r_s = 1 - \frac{6 \times 8.5}{6 \times 35} = \underline{0.757} \, //$

Q.

$\frac{1+2+3}{3} \quad \frac{4+5}{2} =$

Matematics	D	B	A	B	D	D
	3	4.5	6	4.5	2	2
R_M	5	2.5	1	2.5	5	5

Science	C	C	B	A	C	F
R_S	4	4	2	1	4	6
	3	3	5	6	3	1

$\Rightarrow$

M	S	r_M	r_S	d_i	d_i^2
D	C				
B	C				
A	B				
B	A				
D	C				
D	F				

Dimensionality Reduction:

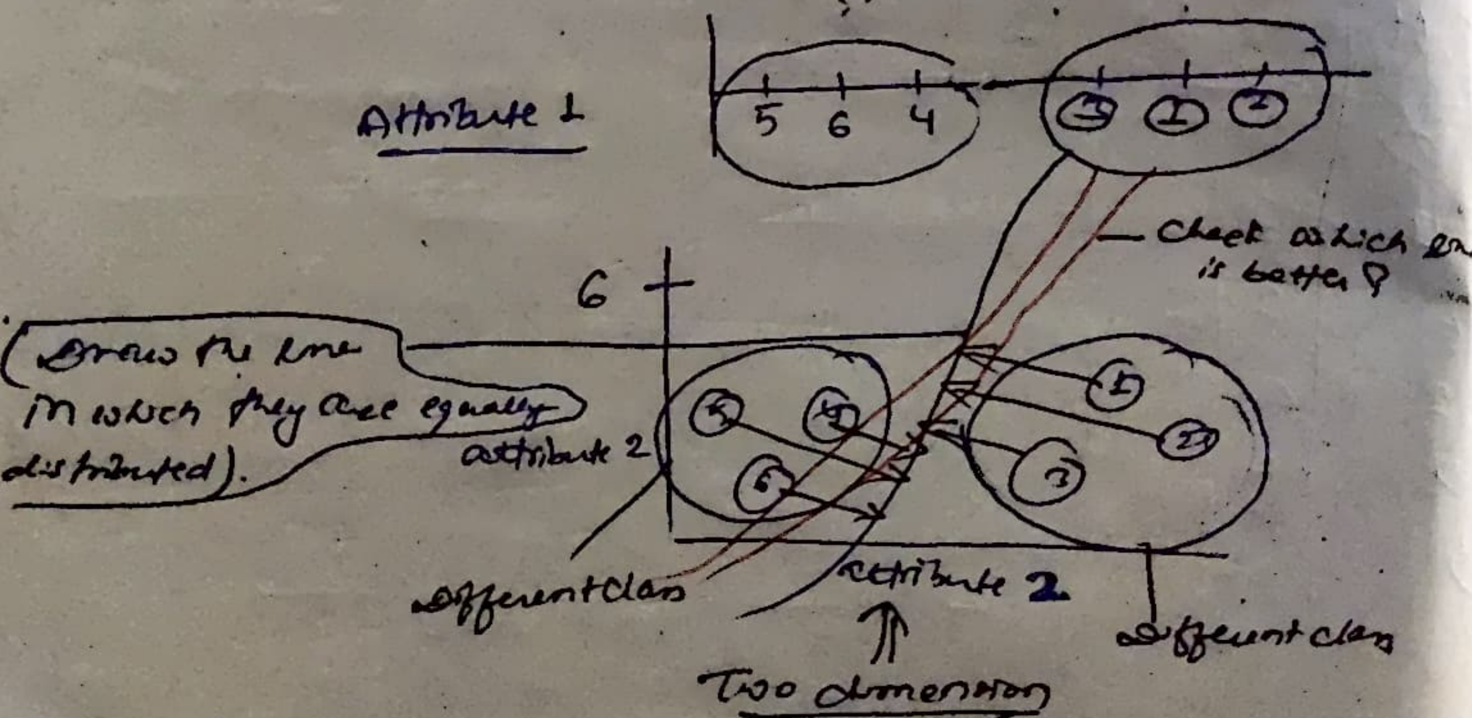
Advantage:

1) Understanding and developing
2) Reduce overfitting.
3) Fail with testing data; is called over fitting.

lecture-6

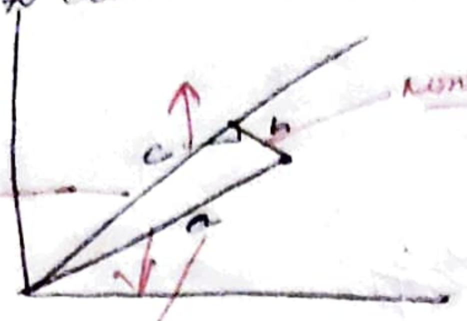
18th Jan, 2024

	object 1	object 2	obj 3	obj 4	obj 5	obj 6
attribute 1	10	11	8	3	1	2
attribute 2	6	4	5	3	2.8	1



∴ only for 3 dimension, we can only visualize.
4 dimension will be difficult.

→ Have to draw a line such that it is equally distributed (above)



Minimizing 'b' distances

Maximizing 'c' value distances

where, a is constant

$a^2 = b^2 + c^2$

(a^2 doesnot change; only b and c changes).

(We have to check,

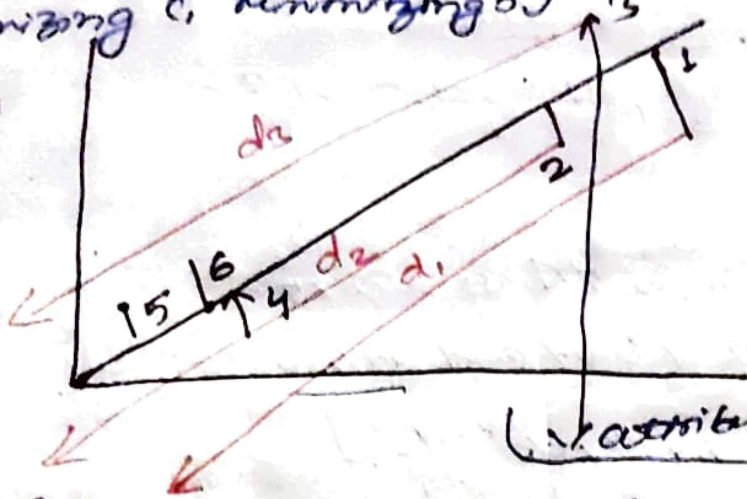
c → increases,

b → decreases

Example:

Maximizing c, minimizing b by objective

(5.8, 3.6) mean value



$d_1\ d_2\ d_3\ d_4\ d_5\ d_6$.

∴ attribute value of mean 1? = 5.8 (from first ex).

mean value of attribute 2 = 3.63

∴ mean value of attribute 1 = 5.8

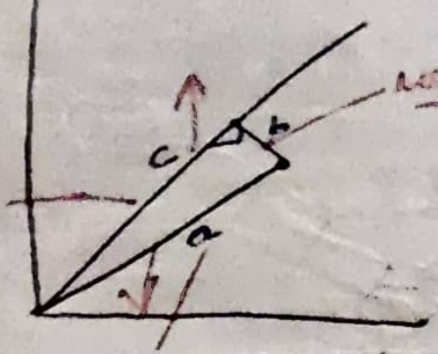
mean value of attribute 2 = 3.63

∴ Square sum for the first line drawn:

$(SS = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2)$

~~# Next~~

→ Have to draw a line such that it is equally distributed (above)



Maximizing b's distances

$a^2 = b^2 + c^2$

(a^2 does not change; only b and c changes)

(we have to check,

c → increases

b → decreases

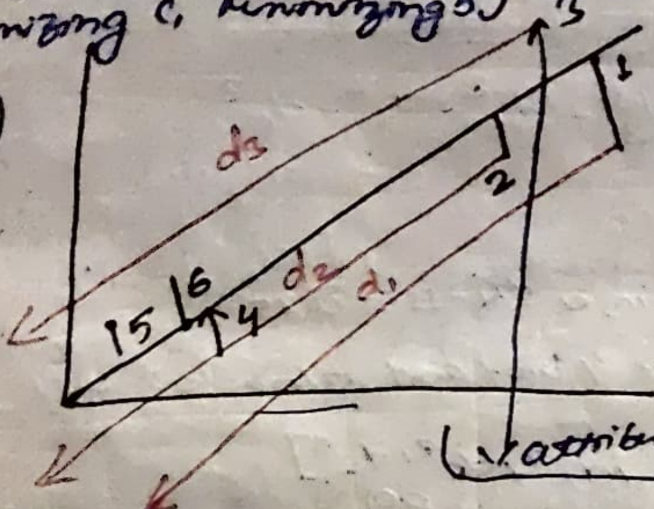
Maximizing 'c' value distances

where, a is constant

Example:

Maximizing c, minimizing b } objective 3

(5.8, 3.6) mean value



$d_1\ d_2\ d_3\ d_4\ d_5\ d_6$

attribute value of mean 1? = [illegible] (from the ex)

mean value of attribute 2 = 3.6

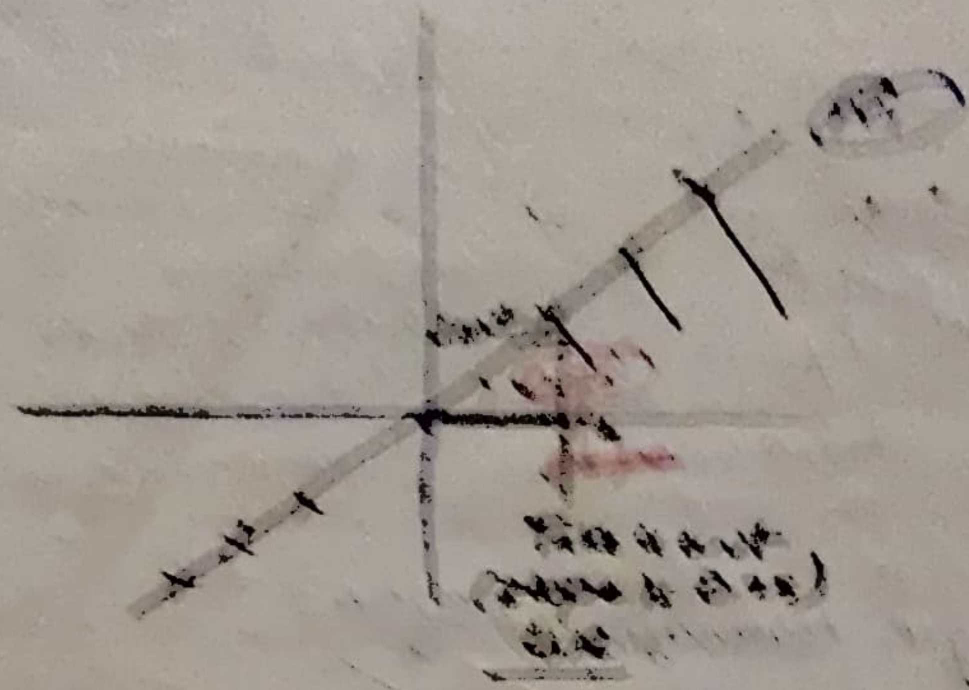
∴ mean value of attribute 1 = 5.8

mean value of attribute 2 = 3.6

∴ Square sum for the first line drawn:

$(SS = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2)$

~~#~~

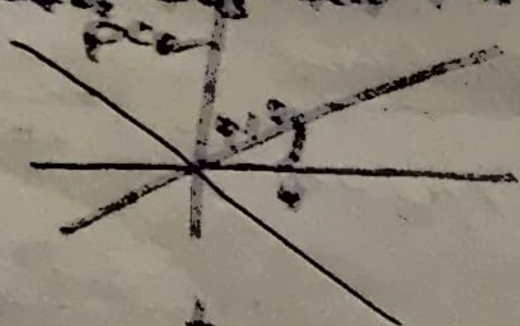


$SS = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$

Max SS

- Slope line that maximize the SS = [illegible]
- [illegible]

[illegible]
[$SS = d_1^2 + d_2^2 + d_3^2$ [illegible]]

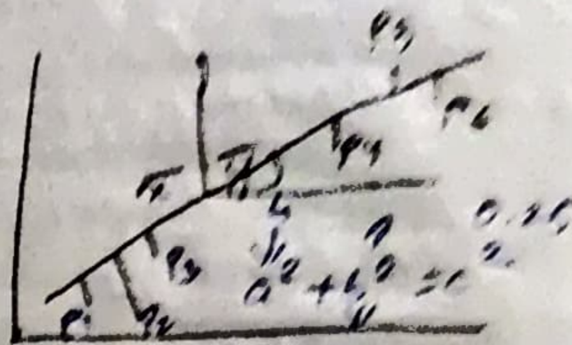


$\frac{4.12}{4.12}$ $\frac{4}{4.12}$, $\frac{1}{4.12}$

$\Rightarrow 1$ 0.92 0.242

Lecture 7 [PCA]

23 Jan, 2024 (Tuesday).



	x	y
P_1	4	11
P_2	8	4
P_3	13	5
P_4	7	14

$\bar{x} = 8$, $\bar{y} = 8.5$

Calculate;
cov (x, x)

$$S = \begin{bmatrix} (x,x) & (x,y) \\ (y,x) & (y,y) \end{bmatrix}$$

$$\text{cov}(x,x) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})$$

$$\Rightarrow \frac{1}{3} \{ (4-8) + (8-8) + (13-8) + (7-8) \} \times (\ldots)$$

$$= \frac{1}{3} \{ (-4) + (0) + (5) - (-1) \}$$

$$\text{cov}(y,x) = \text{cov}(x,y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{cov}(y,y) = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

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

Find eigen values of covariance matrix.

$$0 = \det(S - \lambda I)$$

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

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

(using this eigen values, find eigen vectors)

Find Eigen vectors using eigen values;

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda_1 I)X$$

$$= \begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} (14-\lambda_1)u_1 - 11u_2 \\ -11u_1 + (23-\lambda_1)u_2 \end{bmatrix} = \begin{matrix} 0 \\ 0 \end{matrix}$$

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda} = t$$

t = any real number.

Find out the vector;

$u_1 = 11t$, $\quad u_2 = (14-\lambda_1)t$

$t = 1$

$u_1 = 11 \qquad u_2 = (14-\lambda_1)$

$$u_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix}$$

$$\|u_1\| = 19.73$$

$$e_1 = \begin{bmatrix} 24/43.011 \\ (3x - 23)/43.011 \end{bmatrix}$$

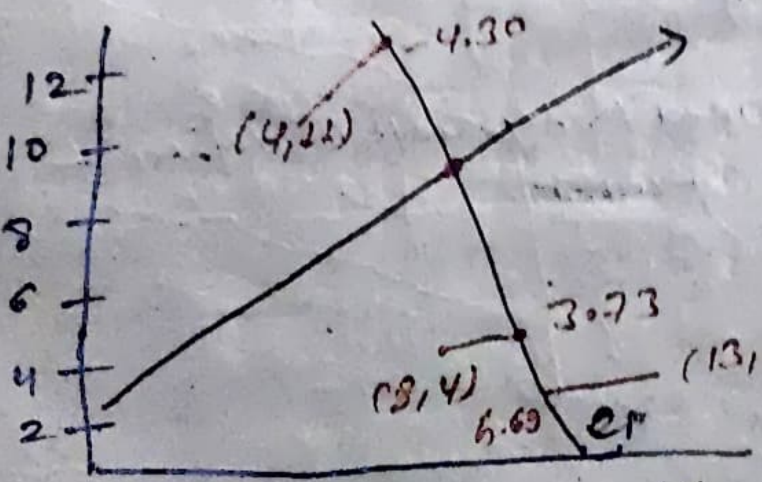
$$= \begin{bmatrix} 0.5575 \\ -0.8303 \end{bmatrix}$$

$$e_1^T \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}$$

$$= 0.5574(x - \bar{x}) - 0.8303(y - \bar{y})$$

[principle component using above formula].

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \end{bmatrix}$$



[illegible]

[illegible]

→ [illegible]

([illegible] separable [illegible] separable
([illegible])

→ Example : ([illegible])

⇒ Easy to separate

⇒ need to separate in different dimension.

2D → 3D

[illegible]

Example :

Trying to make guarmone fnction 2D to 3D.

$f(x, y)$

(x, z)

$\phi(x) = (x^2 + \sqrt{2}\,xy + y^2)$ (∴ converting 2D points to 3D)

↓

mapping into a function.

Which kernal we have to, so that we get the right classification.

Trying to looking the dimension into multi dimension and then

KPCA :

⇒ KERNEL PCA: kernel PCA uses a kernel function to project dataset into a higher dimensional feature space where it is linearly separable.

⇒ mapping data points in multi dimension and then trying to reduce.

⇒ KCPA : is a technique used in ML for non-linear dimensionality reduction.

Covariance

Kernel PCA : (Kernel Trick):- PCA on which we applying kernel trick and then map into.

↳ a simple transformation func which we going to use.

Kernel function is important

Lecture [illegible]

[illegible], Monday

Kernel PCA

Methods

Observed & features

$x \longrightarrow \phi(x)$

observed space high-dimensional space — feature space

KPCA:

$x \rightarrow \mathcal{H}$

$x \rightarrow \mathcal{H} \rightarrow y$

Embedding space:

look data into different space from that space we trying to look into linear space.

Q Calculating kernel of a matrix:

$$A = \begin{bmatrix} 1 & 1 & 2 & 1 \\ 3 & 1 & 4 & 4 \\ 4 & -4 & 0 & 8 \end{bmatrix} \in \mathbb{R}^{3\times 4}$$

Find $\ker(A) = \{x \in \mathbb{R}^4 \mid Ax = 0\}$

converting the data into a ~~[illegible]~~ system of linear Equations}.

$$\Rightarrow \ker(A) = \begin{bmatrix} 1 & 1 & 2 & 1 \\ 3 & 1 & 4 & 4 \\ 4 & -4 & 0 & 0 \end{bmatrix} \begin{matrix} \\ R_2 - 3R_1 \leftarrow R_2 \\ R_3 - 4R_1 \leftarrow R_3 \end{matrix}$$

as many no. of rows into a zero.

$$\Rightarrow \begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & -2 & -2 & 1 \\ 0 & -8 & -8 & 4 \end{bmatrix} R_3 - 4R_2 \leftarrow R_3$$

$$\Rightarrow \begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & -2 & -2 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = 0$$

$$\Rightarrow \begin{pmatrix} 1 & 1 & 2 & 1 \\ 0 & -2 & -2 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = 0$$

$$\Rightarrow -2x_2 - 2\overset{\alpha}{x_3} + \overset{\beta}{x_4} = 0$$

Let say,

$x_3 = \alpha$

$x_4 = \beta$.

$$x_2 = -\alpha + \tfrac{1}{2}\beta$$

Then,

$$x_1 + x_2 + 2x_3 + x_4 = 0.$$

$$x_1 + x_2 + 2\alpha + \beta = 0$$

$x_1 \to$

$$x_1 + (-\alpha + \tfrac{1}{2}\beta) + 2\alpha + \beta = 0$$

$$x_1 = -\alpha - \tfrac{3}{2}\beta$$

$$\left\{ \begin{pmatrix} -\alpha - \frac{3}{2}\beta \\ -\alpha + \frac{1}{2}\beta \\ \alpha \\ \beta \end{pmatrix} \middle| \ \alpha, \beta \in \mathbb{R} \right\}$$

~~Bases~~ Basis for calculating kernal (A).

$$\alpha \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} -3/2 \\ 1/2 \\ 0 \\ 1 \end{pmatrix} \ \Big| \ \alpha, \beta \in \mathbb{R}$$

$\therefore$ For this Example;

kernal is two dimensional

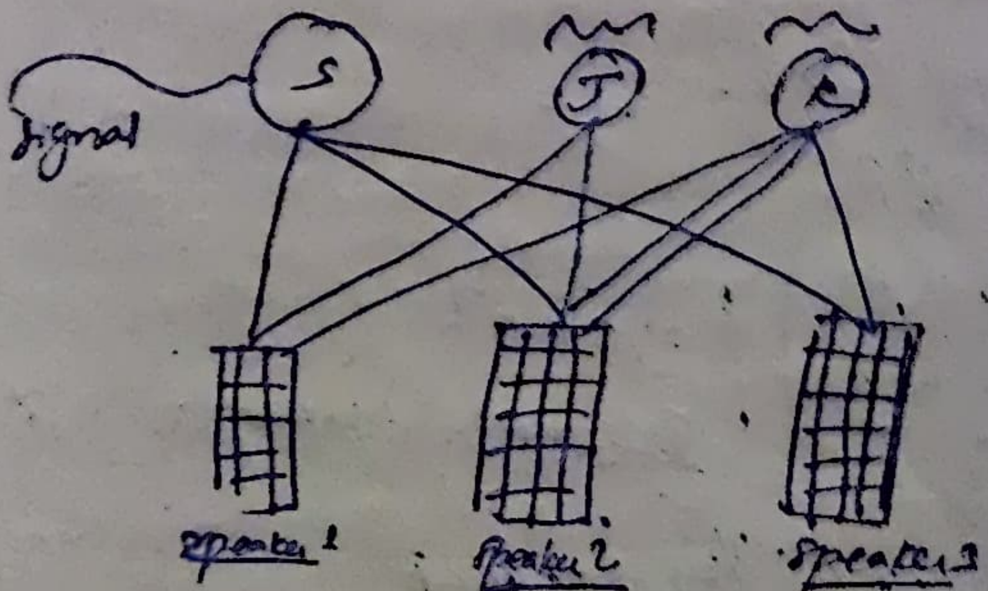
ICA { Independent Component Analysis (ICA) }

Lecture-10

30th Jan 2024, Tuesday

ICA :- Independent Component Analysis

Example:



Make them into original signal. What is the signal generated by each of them. Filter out the original signal produced by S, J, R.

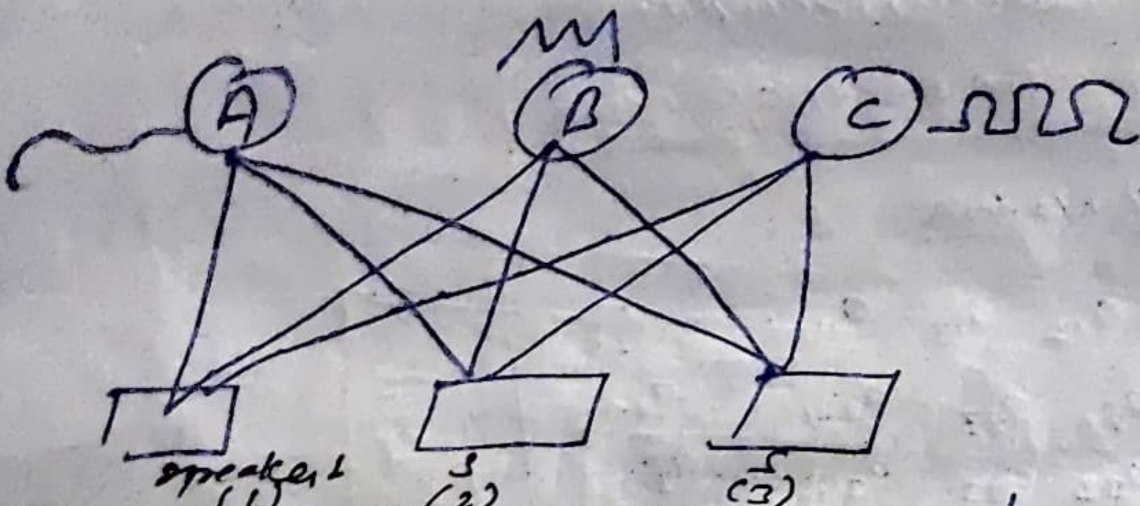
Application: Noise cancellation

lecture-13

Independent Component Analysis (ICA)

Objective : different location different signal want to make original signal.

Linear combination of all signals captured by speaker 2. but we would like to identify the original signal generated by A, B, and C.



Trying to get independent component from signal.

ICA also called as. Blind signal separation

Used for

↳ noise cancellation

↳ Audio processing.

↳ Medical data.

↳ coding

The simple "cocktail party" problem.

[illegible] (Lecture 11)

Independent Component Analysis (ICA)

[illegible] statistically independent non-gaussian data. [illegible] finding underlying [illegible] time [illegible] data. [illegible] components from multi-dimensional [illegible] ~~[illegible]~~

$\boxed{\text{[illegible] and ICA}}$

[illegible]

[illegible]

[illegible] mathematical approach

$\boxed{\text{Mixing matrix}}$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots & a_{3n} \\ a_{41} & a_{42} & a_{43} & a_{44} & \cdots & a_{4n} \\ \vdots & & & & & \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ \vdots \\ s_n \end{bmatrix}$$

$x = As$,

Also written as; $\therefore A$ is a mixing matrix.

$$x = \sum_{j=1}^{n} a_j s_j$$

Assumptions of ICA: The components s_i are statistically independent. Eg: $P(A \cap B) = P(A) \cdot P(B)$.

Limitations of ICA:

1) Cannot determine the variances of independent components as s, A are unknown

2) Unit variance: $E[s_i^2] = 1$.

3) Ambiguity of the sign.

~~[illegible]~~

1) Can not determine the order of independent components.
↳ i.e. are continuous.
↳ freely change the order of the terms.

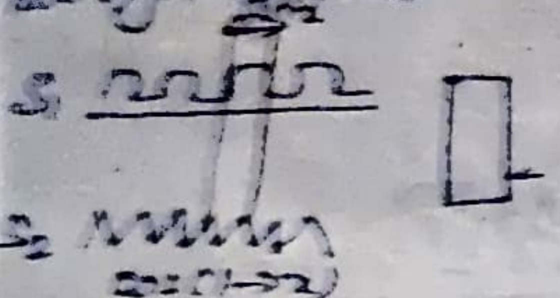
ICA process:

1) centering data (remove mean)
2) whitening process (sphere data)
whitening + non-gaussianity → independent.

LECTURE-12

~~[illegible]~~

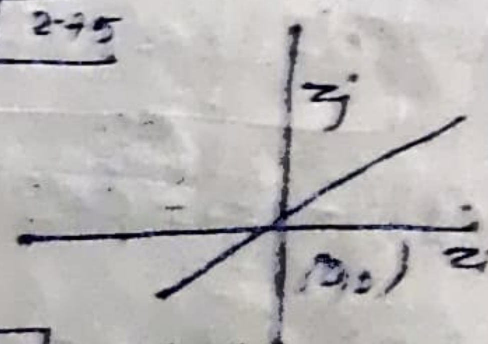
Two signals need to analyze independent component of signal.



$x_i = (A) s_i$
↓
mixture matrix

x_i	y_i
2	3
7	2
4	1
5	5
4.5	2.75

z_i	z_j
-2.5	0.25
2.5	-0.25
0.5	-1.75
0.5	2.25



$\boxed{\text{mixture matrix - independent}}$ is to find its determination of a determined to get the

centering, whitening order).

$\boxed{A = W^{-1}}$

whitening:

$x_i \leftarrow E D^{-1/2} E x^T \frac{1}{2}$ ⟶ ICA - learning

SVD → single value decomposition

$\boxed{W \leftarrow g^{-1} E W}$

$w_j \leftarrow w_j / \| w_j \|$

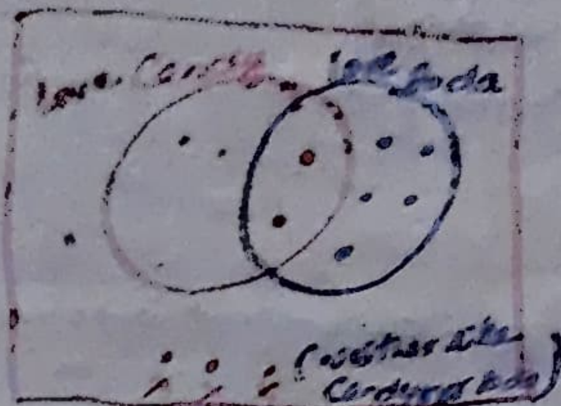
W → is used to calculate x_i signal.

Fast ICA

lecture 2:

[illegible], Tuesday

Conditional Probability



		Candy		
		LC	NC	
Soda	LS	2	5	$\frac{2+5}{14}$ $\frac{7}{14}$
	NS	4	3	$\frac{7}{14}$
		$\frac{6}{14}$	$\frac{8}{14}$	

$P(LC \,\&\, S) = \frac{2}{14} = 0.14$

$P(LC \,\&\, NS) = \frac{4}{14} = 0.29$
(only LC)

$P(NC \,\&\, LS) = \frac{5}{14} = 0.36$
(only LS)

$P(NC \,\&\, NS) = \frac{3}{14} = 0.22$

$\underbrace{P(LC \& LS/LS)}$ → ~~[illegible]~~ evident that a person love soda.

⇓

<u>This condition is Conditional Probability</u>

→ (Figuring out whether love candy as well love soda)

$\Rightarrow P(LC \& LS) = \frac{2}{14}$

$P(LS) = \frac{7}{14}$

$$= \frac{\frac{2}{14}}{\frac{7}{14}}$$

$$= \frac{2}{7} = 0.29$$

$\Rightarrow P(NC \text{ and } LS/LS) = \frac{5}{7}$

$$\boxed{P(NC \& LS/LS) = \frac{P(NC \& LS)}{P(LS)}}$$ ⇒ used to calculate <u>conditional probability</u>.

The Bayes Theorem

$$P(NC \,\&\, LS/LS) = \frac{P(NC \,\&\, LS)}{P(LS)}$$

or written

$$P(NC/LS) = \frac{P(NC \,\&\, LS)}{P(LS)}$$

$$P(NC \,\&\, LS/LS) = \frac{P(NC \,\&\, LS)}{P(LS)}$$

$$P(NC \,\&\, LS/NC) = \frac{P(NC \,\&\, LS)}{P(NC)}$$

to find out

$$P(NC \,\&\, LS/LS) \times P(LS) = \frac{P(NC \,\&\, LS)}{\cancel{P(LS)}} \times P(LS)$$

to find out

$$P(NC \,\&\, LS/NC) \times P(NC) = \frac{P(NC \,\&\, LS)}{P(NC)} \times P(NC)$$

to find out

$$P(NC \,\&\, LS/LS) \times P(LS) = P(NC \,\&\, LS/NC) \times P(NC)$$

A = does not love candy (NC)

B = love soda (LS)

$P(A \,\&\, B/B) = P($

$$P(A \& B/A) = \frac{P(A \& B/B) \times P(B)}{P(A)}$$

Lecture-14

12th Feb, ~~...~~ 2024, Monday

Bayesian Classifier

↳ Mainly used to take individual attributes.
↳ no attribute will affect another attribute

Dataset

$$X = \{A_1, A_2, A_3 \ldots\ldots A_n\}$$

$$x = \{x_1, x_2, x_3 \ldots\ldots x_n\}$$

where x_1 is present in A_1.
x_2 is present in A_2.
⋮
S.on.

we have,

Class(c), $c = \{c_1, c_2, \ldots c_m\}$

Given tuples, x,

The classifier will predict that x belongs to the class having highest posterior probability.

$$P(C_i/x) > P(C_j/x) \text{ for } i \le j \le m, j \ne i$$

$$P(A \& B/A) = \frac{P(A \& B/B) \times P(B)}{P(A)}$$

Lecture-14

12th Feb, ~~07~~ 2024, Monday

Bayesian Classifier

↳ Mainly used to take individual attributes.
↳ no attribute will affect another attribute.

Dataset

$$X = \{A_1, A_2, A_3 \ldots\ldots A_n\}$$

$$x = \{x_1, x_2, x_3 \ldots\ldots x_n\}$$

where x_1 is present in A_1.

x_2 is present in A_2.

⋮

so on.

we have,

Class(C), $C = \{C_1, C_2, \ldots\ldots C_m\}$

Given tuples, x,

The classifier will predict that x belongs to the class having highest posterior probability.

$$P(C_i/x) > P(C_j/x) \text{ for } i \leq j \leq m,\ j \neq i.$$

Maximize $P(C_i/x)$

$$P(C_i/x) = \frac{P(x/C_i)\,P(C_i)}{P(x)}$$

posterior

prior probability of x.

Baye's theorem

~~If it is not~~

If it is given;

$$\cancel{P(C_i) = \frac{|C_{i,D}|}{|D|}}$$

$$P(C_i) = \frac{|C_{i,D}|}{|D|}$$

where D is dataset

~~prior probability is ...~~

• prior probability of x is common for all the class.

$$P(X/C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \ldots P(x_n|C_i)$$

Maximize $P(C_i|x)$

$$P(C_i|x) = \frac{P(x|C_i)\,P(C_i)}{P(x)}$$

posterior

prior probability of x.

Baye's theorem

~~If it is not~~

If it is given;

$$\cancel{P(C_i) = \frac{|C_{i,D}|}{|D|}} \qquad P(C_i) = \frac{|C_{i,D}|}{|D|}$$

where, D is dataset.

~~prior probability is defined as~~

• Prior probability of x is common for all the class.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \ldots P(x_n|C_i)$$

RID: Record Id:

Example

RID	age	Income	(Y or N) Student	Credit Rating	Class label (buys laptop)
1	Youth	High	N	F	N
2	Youth	High	N	E	N
3	middle aged	High	N	F	Y
4	S	Medium	N	F	Y
5	S	L	Y	F	Y
6	S	L	Y	E	N
7	MA	L.	Y	E	Y
8	Youth	Medium	N	F	N
9	Youth	L	Y	F	Y
10	S	Medium	Y	F	Y
11	Youth	Medium	Y	E	Y
12	MA	Medium	N	E	Y
12	MA	Me~~dium~~	Y	F	Y
14	S	Medium	N	E	S

X = (age = Youth,
~~Student~~ Income = Medium,
Student = Yes
Student rating = fair)

(4 ~~..~~ variables are categorial variables).
(4 values)

~~P(buy~~ <u>Prior probability of the class</u>

$P(\text{buys computer} / \text{yes}) = \frac{9}{14} = 0.643$

$P(\text{buys computer} / \text{no}) = \frac{5}{14} = 0.357$

Compute $P(x/C_i)$ for $i = 1, 2$

$P(\text{age} = \text{youth} / \text{buy computer} = \text{yes}) = \frac{2}{9}$

$P(\text{age} = \text{youth} / \text{buy computer} = \text{no}) = \frac{3}{5}$

$P(\text{income} = \text{Medium} / \text{buy computer} = \text{yes}) = \frac{4}{9}$

$P(\text{income} = \text{Medium} / \text{buy computer} = \text{no}) = \frac{2}{5}$

~~Pl~~ Calculate, student = yes
Credit = fair

$P(\text{student} = \text{yes} / \text{buy computer} = \text{yes}) = \frac{6}{9}$

$P(\text{student} = \text{yes} / \text{buy computer} = \text{no}) = \frac{1}{5}$

$P(\text{Credit Rating} = \text{Fair} / \text{buy computer} = \text{yes}) = \frac{6}{9}$

$P(\text{Credit Rating} = \text{Fair} / \text{buy computer} = \text{no}) = \frac{2}{5}$

$\text{Yes} = \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} = P(x/C_i)_{i=1,2}$

$= 0.044 \times 0.643 (9/14)\ P(C_i) [P(\text{class})$

$\boxed{= 0.028}$

$NO = \left(\frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5}\right) P(\text{no class})$

$= \frac{5}{14} \times 0.357$

$\boxed{\cancel{0.0357}}$

$\boxed{= 0.007.}$

Final probability for Yes $= 0.028$ (High)

No $=$ ~~0.0357~~ 0.007

~~If the values~~

If we have continous variables, Then,

$$P(x_k | c_i) = g(x_k, \mu c_i, \sigma c_i)$$

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\boxed{\mu \rightarrow \text{Mean} \quad \sigma \rightarrow \text{S.D}}$

calculate probability of Yes class

calculate probability of no class.

(Gaussian function of variable) $\rightarrow$

mean for attribute
variance

$$\text{Posterior (male)} = \frac{P(M) * P(H/m) * P(w/M) * P(Fs/M)}{\text{Evidence}}$$

Gaussian $\quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$H/M = \frac{1}{\sqrt{2\pi * (0.035033)}} e^{-\frac{(6-5.811)^2}{2*(0.035033)^2}}$$

$= 1.5789.$

Lecture-15

13th Feb, 2024, Tuesday

(Problems) (Mail is spammed/normal)

(c) Datasets..

when you attributes are;

Q In which dataset i.e. normal or spam the above msg will come.

Dataset

(Normal)	(spam) ~~Spam~~
Dear - 8	Dear - 2
Friend - 5	Friend - 1
~~lug~~ lunch - 3	lunch - 0
Money - 1	Money - 4
(17)	(7)

tuple

① X = Dear Friend ⇒ Normal

② X = lunch Money Money Money Money ⇒ Normal

whether it goes to Normal or Spam?

⇒

17+7 datasets

A_1	A_2	A_3	A_4	N/S
Dear	Friend	lunch	Money	

(A_1 ... A_4 – Attributes)

Q.1 : $P(Dear|N) = 8/17$ $P(Friend|N) = 5/17 = 0.29$

Sol: $P(Dear|S) = 3/7$ $P(Friend|S) =$

Q.2 : $P(lunch|N) = 0.17$ $P(Money|N) = 0.06$

Sol: $P(lunch|S) = 0.00$ $P(Money|S) = 0.57$

⇒ ~~Probability~~ ⇒ P ① P(Dear & Friend) = P(Dear friend)
= NORMAL.

② P(lunch | N) = 0.18 P(Money | normal) = 0.06
P(lunch | S) = 0.00 P(Money | Spam) = 0.57
normal dataset so, NORMAL (Spam is greater).

• (parameter adding : to make right classification called alpha parameter "), [$\alpha = 1$]

③ last add +1 : (Tuples are increased by 17+4 and 7+4)

Normal	Spam
Dear = 3+1	Dear = 2+1
Friend = 5+1	Friend = 1+1
lunch = 3+1	lunch = 0+1
Money = 1+1 (21)	Money = 4+1 (11)

P(lunch | N) = 0.14 P(Money | N) = ~~0.09~~ 0.09
P(Money | S) = 0.45
P(lunch | S) = 0.09

∴ It will go in spam dataset.

Rule based classifier

SPAM

Parameter we add to become right normalization
↳ α parameter.

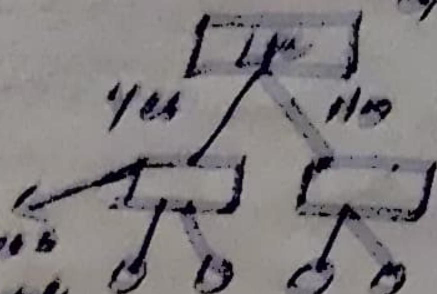
Decision tree classifier

Supervised Learning

Machine Learning

Decision Tree & Classification Tree.

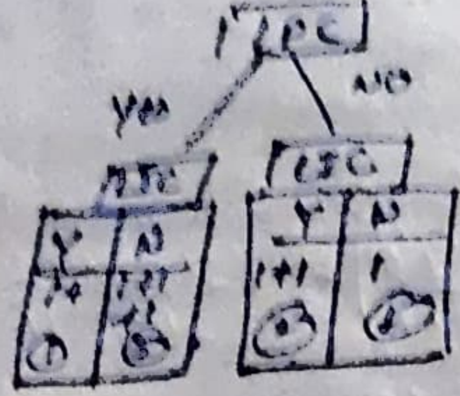
It's tree use for classification.



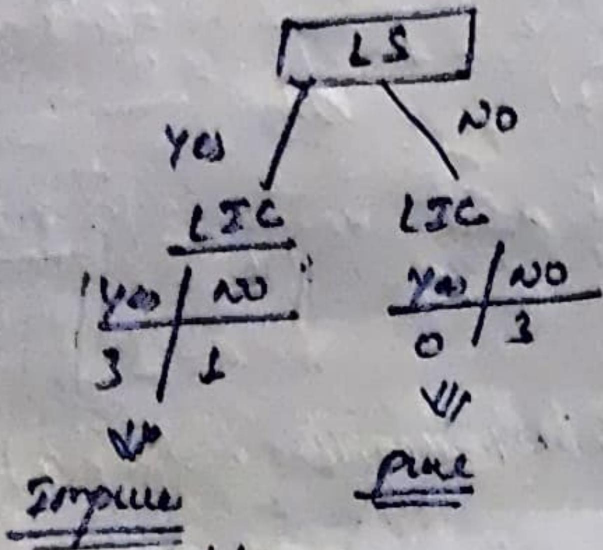
[illegible]

Age are in sorted order.

(LP) Loves Popcorn	(LS) Loves Soda	Age	Class label [illegible] (LJC)
Yes	Yes	7	No
Yes	No	12	No
No	Yes	25	Yes
Yes	Yes	35	Yes
Yes	Yes	37	Yes
Yes	No	50	No
No	No	13	No



(purity / Impurity of node)



3 Imp parameters
- Entropy
- Info gain
- Gini Index *

(1) LPC

GINI INDEX → find out impurity of a node

GI of left node

$$GI = 1 - \sum_{i=1} (P)^2$$

$$= 1 - (\text{prob of yes})^2 - (\text{prob of NO})^2$$

Left node $= \cancel{1 - \left(\frac{1}{1+3}\right)} = 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{3+1}\right)^2 = \boxed{0.375}$ → 4 people

Right node $= 1 - \left(\frac{2}{1+2}\right)^2 - \left(\frac{1}{1+2}\right)^2 = \boxed{0.444}$ → 3 people

Total GINI impurity = weighted avg of GINI Impurities for the leave.

GI = weighted average of LPC

$$= \left(\frac{4}{4+3}\right) \times 0.375 + \left(\frac{3}{4+3}\right) 0.444 = \boxed{0.405}$$

GI LPC = 0.405

(7) What is

GI is?

GI of left node $= 1 - (\text{Prob Yes})^2 - (\text{Prob. No})^2$

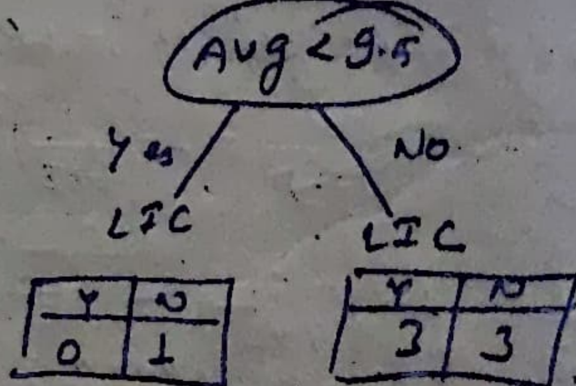
$$= 1 - \left(\frac{3}{3+1}\right)^2 - \left(\frac{1}{3+1}\right)^2 = 1 - \frac{9}{16} - \frac{1}{16} = \frac{16-10}{16} = \frac{6}{16} = 0.375$$

Right node $= 1 - \left(\frac{0}{0+3}\right)^2 - \left(\frac{3}{0+3}\right)^2 = 1 - \frac{1}{1} = 0$

$$\Rightarrow \left(\frac{4}{4+3}\right)\frac{6}{16} \Rightarrow \frac{4}{7} \times \frac{6}{16} = \frac{3}{14} = \boxed{0.214}$$

GI age? (which has min value will be root node)?

Avg	Gini Index
9.5 (7, 12)	→ 0.429
15 (12, 18)	→ 0.343
26.5 (18, 35)	→ 0.476
36.5 (35, 38)	→ 0.476
44 (38, 50)	→ 0.343 min
66.5 (50, 83)	→ 0.429

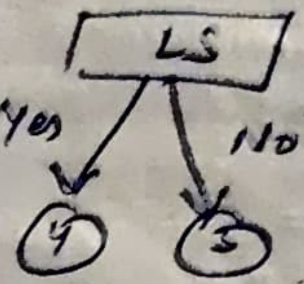


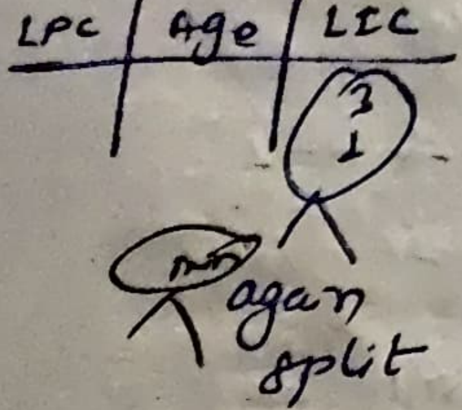
$\Rightarrow \left(1 - \frac{1}{1}\right)^2 = 0 \Rightarrow 1-1-1 \Rightarrow -1$

a1 ⇒ age < 15 = 0.343 since minimum value

$\boxed{LS = 0.214}$ — it will become the root node

LPC = 0.405





(19th Feb, 2024, Monday) lecture-17

> ## Entropy

> **Info gain**: finding info given tuple ∈ which class.

$$\text{Info}(D) = \sum_{i=1}^{m} p_i \log_2(p_i)$$

> info gain depends on entropy.
> m: possible value dataset to cal.
  i.e. possible values.

for A = 1

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{info}(D_j)$$

$$\text{GAIN}(A) = \underbrace{\text{Infropy}(D)}_{\text{entire DS}} - \underbrace{\text{Infropy}_A(D)}_{\text{one attributes.}}$$

Q Calculate info. gain of age.

$$\text{Info}(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) + \frac{-5}{14}\log\left(\frac{5}{14}\right)$$
$$= 0.940$$

$Info_{age}(D) = $ [illegible]

[illegible]

Gain(age) = Info(D) - Info_age(D)

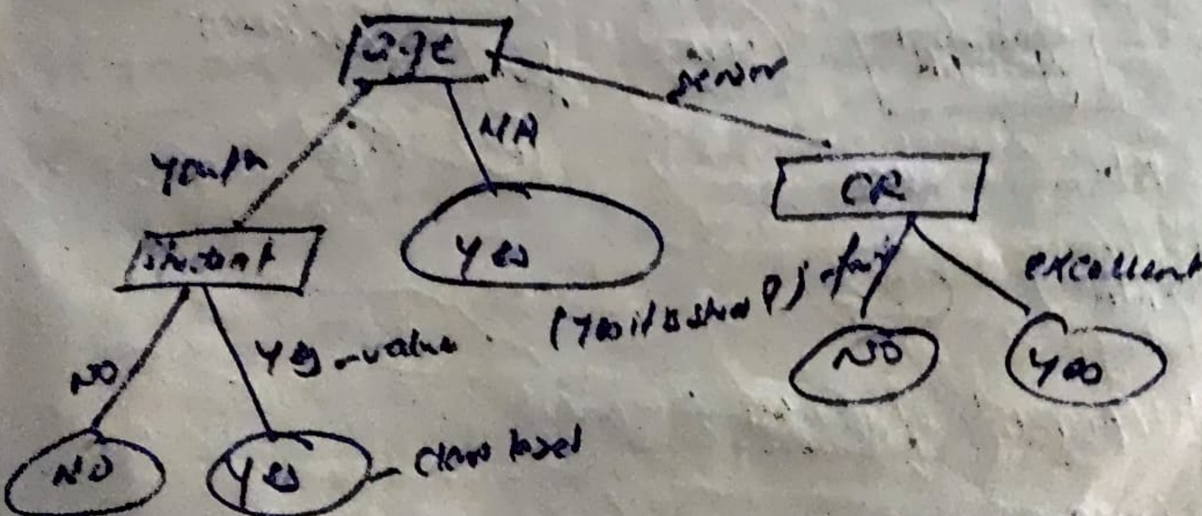
= 0.694

Entropy - lowest one

Gain - highest value (entire dataset)

Root Node: one has highest info gain.

- Info gain
- Gini Index



Rule based classification

If condition then conclusion.

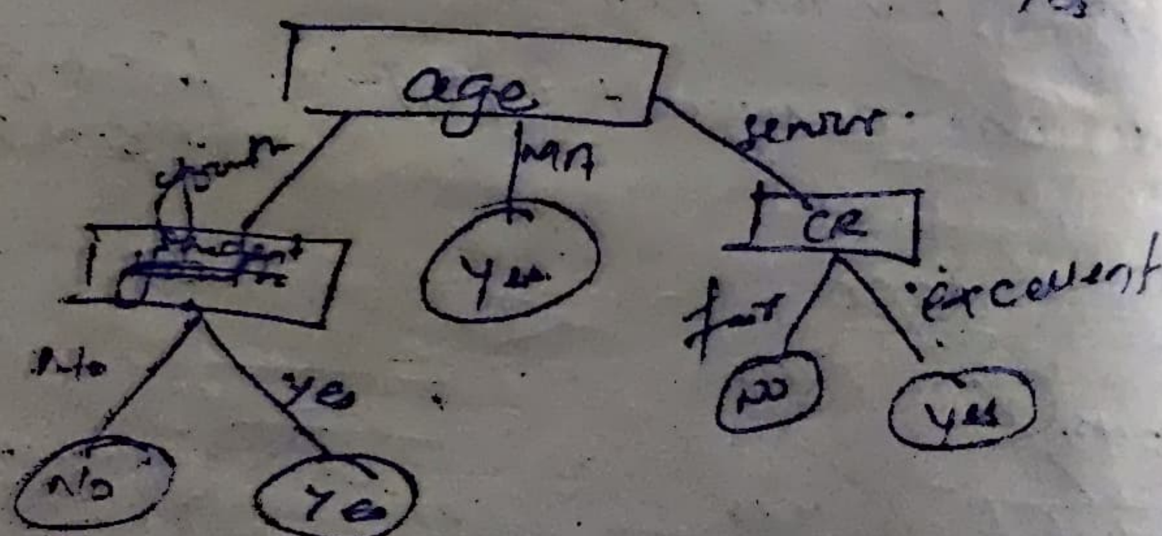
or

Precondition — Rule antecedent

Rule consequent.

R_1: If age = youth and student = yes
then BC = yes.

R_1: $(age = youth) \wedge (student = yes) \Rightarrow BC = yes$



Quality of Rules:
1. Coverage
2. Accuracy

(1) $\text{Coverage}(R) = \dfrac{n_{covers}}{|D|}$

(2) $\text{Accuracy}(R) = \dfrac{n_{coverage}}{n}$ $\quad \dfrac{n_{correct}}{n_{covers}}$

Rule-Based Classifier

20th Feb, 2024, Tuesday lecture-18

Firing: Any tuple which satisfies two rules is called firing
or more than two rules

Two rules for Rule-Based Classifier

1) Mutually Exclusive: No two rules are ~~being~~ trigged by same record.

2) Exhaustive rule: Any rule which triggers all ~~rule~~ records are called exhaustive rule.

Cross-validation:

learning
testing → 2 important

dividing dataset for test.

✓
✓
✓
Testing

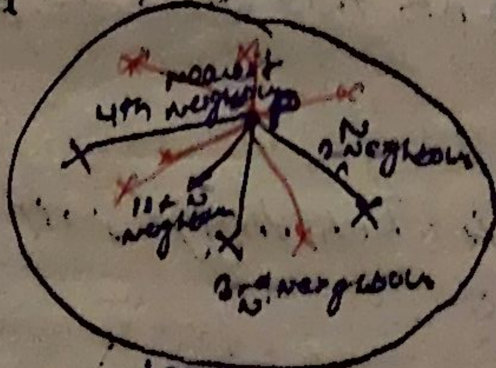
use my 2 blocks for Testing model → LEARNING

	Yes	No	measure of correct...
Bayesian classifier	5	4	
Decision Tree	6	3	
Rule-Based Classifier	5	4	
	6	3	
4-fold	6	3	
	6	3	

Four-fold cross validation:

(Because we have 4 blocks; 3 is used for training ~~testing~~ and 1 is for testing)

k-Nearest Neighbour classifier



Matrix is a measure, where we can calculate neighbours or we know own.

3 properties of matrix! (10 marks)

1) $d(x,y) \geq 0$
2) $d(x,y) = d(y,x)$
3) $d(x,z) \leq d(x,y) + d(y,z)$

- want to classify P nearest neighbour k. ⇒

$\Rightarrow$ $k=10$ (10 points are there)

$p=6$ (majority of Red × are more)

According to paper.

$y = 0.1 + 0.78x$

y-intercept, slope

So since the slope not zero $\Rightarrow$ we will get a point in the size dimension to find the intersection point with size axis. depends on this value we will check if our value is good or bad depend on the lower value or may depends on the function we were

$d(x,y) \geq 0$

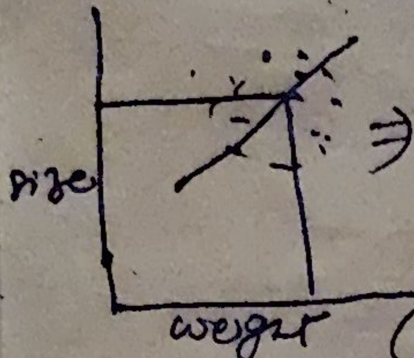
$d(x,y) = d(y,x)$

$d(x,z) \leq d(x,y) + d(y,z) + d($

$d($

Linear regression

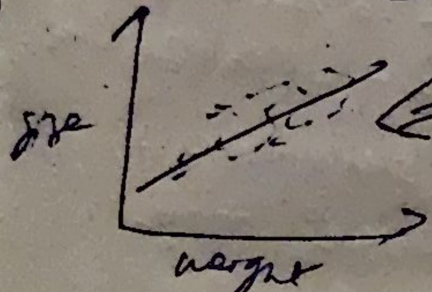
lecture-19
22nd Feb, 2024,
Tuesday



~~(1) Use least~~

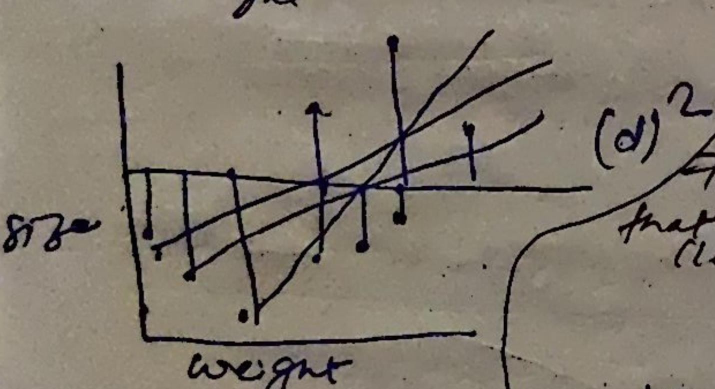
⇒ Used to predict value for variable depending on other.

* we should understand 3 - important values
(1) least square to fit a line
(2) calculate R^2
(3) calculate a P value for R^2.



⇐ How To draw This line

difference make it square (d^2).
for every line I'll check the d^2 values and draw the corresponding point in the graph ⇒ we will get a graph as polynomial features.



⇒ The Best value of the line that achieve the ~~max~~ min value (least square ~~fitted~~ factor / fitted)

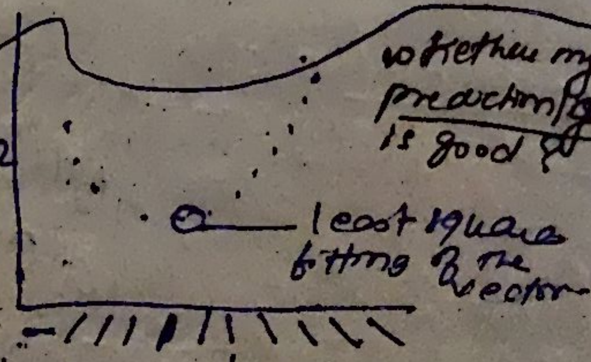
line equation of right line which is fitted is least square fitted d^2

$$\boxed{y = (0.1) + (0.78)x}$$

y-intercept — slope

(if slope ≠ 0 then,
→ with respect to ~~we can predict some~~ weight then we can give some weight to the size.

→ If w is some 'x' then

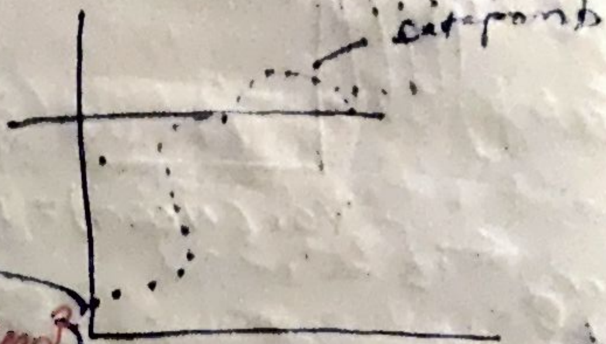


• starting line
• calculate $(d)^2$ value.

$\boxed{R^2}$ — weight is affecting the size, we're getting relationship
value

R^2 measured in pr

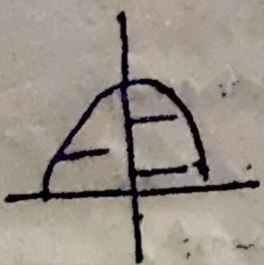
↳ percentage of variation explained by the relationship between two variables.

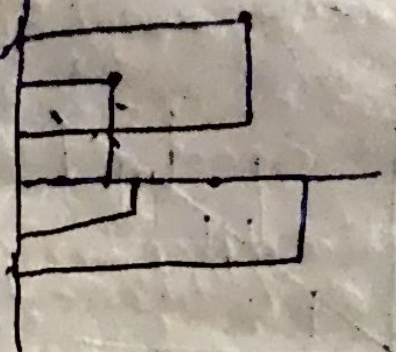


$$\boxed{SS(Mean) = (data - Mean)^2}$$

variation around the mean $= \dfrac{(data - Mean)^2}{n}$

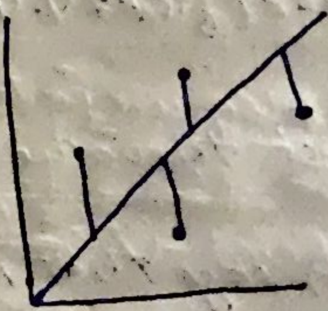
with respective to data points.





$(SS/fit) = (data - line)^2$

variation of fit $= \dfrac{(data - line)^2}{n}$



Variation of (ST) $= \dfrac{SS}{n}$ → sum of square

(ST) ↓ something

↓

Average sum of square.

$$\boxed{R^2 = \dfrac{Var(Mean) - Var(fit)}{Var(Mean)}}$$ — 1

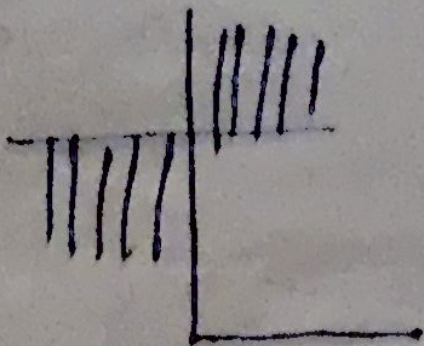
Mean is calculated with respect to size

↓

~~How much variation of size of weight came into account.~~

Here R^2 ⇒ tell us how much variation in size can be explained by taking weight into account. (predicting only 60% of variance) perfectly → 100%

// Mean is calculated by below figure:



Var (Mean) = 11.1

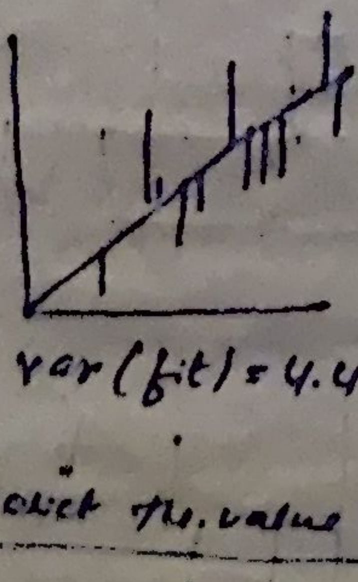
$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$= 0.6$

$\boxed{= 60\%}$

(This is 60% of production of variation when we take weight into account)

//



var (fit) = 4.4

↳ Making weight as independent

↳ predict the value of size when weight is given?

↳ we can say that:

weight explains 60% of the variation in size.

↳ we can also make R^2, simply;

$$R^2 = \frac{100-40}{100} = 60\%$$

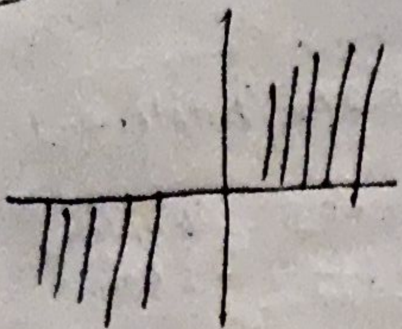
coming same

$$\frac{SS^2(\text{mean}) - SS(\text{fit})}{SS(\text{Mean})}$$

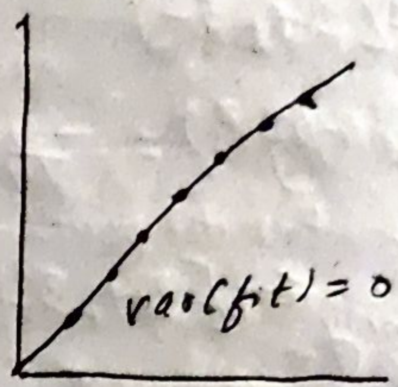
(can also be written like this).

$$R^2 = \frac{Var(\text{Mean}) - Var(\text{fit})}{Var(\text{Mean})}$$

Example:



11.1
var(mean)

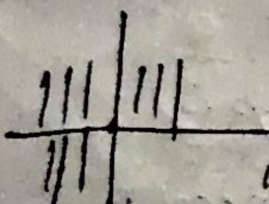


$R^2 = 1 = 100\%$

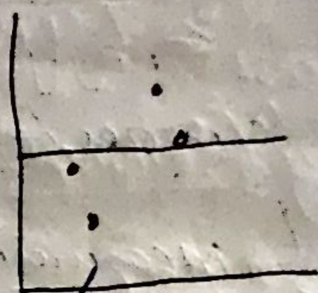
→ weight explains 100% variation in the size.

(since data points are aligned together)

→ Now parameter is weight which explains 100% size.



$R^2 = 0$



→ weight will not explain about the size.

(fit line is same as mean line)

→ when calculating

(Collapsing weight on x-axis).

Collapsing weight on y-axis

R^2 = will tell you how other variables are effected.

Linear Regression

(1) Quantifying the relationship in the data
(this is R^2)

$y = 0.1 + 0.7x$ (simple eqn)

$y = 0.1 + 0.7x + 0.8z$ (complex eqn)

{even if it is compressed}

1) R^2 needs to large value.

large means two variables are highly, strongly correlated, good relationship.

(2) Determines how reliable that relationship is

→ Measured in terms of 'p' value. This need to be ~~strong~~ small. i.e. P-value should be small.

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Attr = Outlook

Values(Outlook) = Sunny, Overcast, Rain

$S[9+, 5-]$

$Entropy(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.9$

$S_{Sunny} \leftarrow [2+, 3-]$

$Entropy(S_{Sunny}) = -\frac{2}{14}\log_2\left(\frac{2}{14}\right) - \frac{3}{14}\log_2\left(\frac{3}{14}\right)$

$\Rightarrow 0.971$

$S_{overcast} \leftarrow [4+, 0-]$

$Entropy(S_{overcast}) = -\frac{4}{14}\log_2\left(\frac{4}{14}\right) - \left(\frac{0}{14}\log_2\left(\frac{0}{14}\right)\right.$

$\Rightarrow 0$

$S_{Rain} \leftarrow [3+, 2-]$

$Entropy(S_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \Rightarrow 0.971$

$Gain(S, outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \left|\frac{S_v}{S}\right| Entrop$

After Midterm

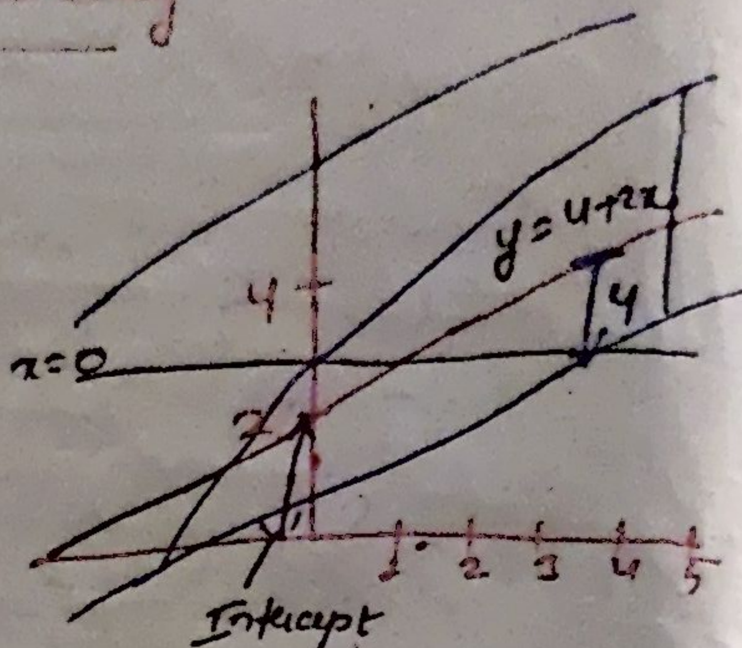
Lecture 20

4th March, 2024, Monday

$y = a + bx$

$y = 4 + 2x$

What is the purpose of intercept?



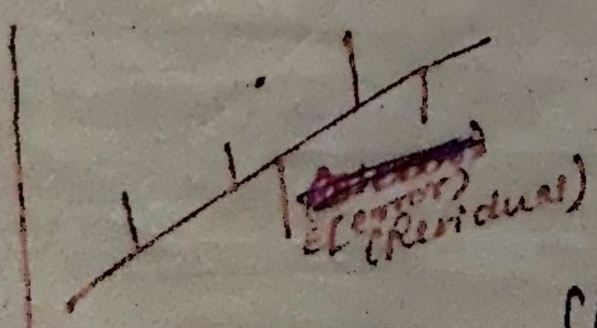
Changing the equation

$$y = 3 + 2x$$

Also, $y = -2 + 2x$ nothing but a line parallel to $y = 4 + 2x$.

$y = 4 + 3x$, every change of 1 unit x, y is changing 3 units.

Data is a scatter data



$y = a + bx + \varepsilon$ → two parameters to be considered

y → Scored marks

x → height

(how is the height equal to the scored marks?)

Does the intercept affect the distribution of the value?

=) Does not affect much in the relationship.

Main one which are affected.

Income & consumption : (indirectly proportional)

but if the family size is not given and we have to ~~derive~~ derive the relation, consumption/per week.

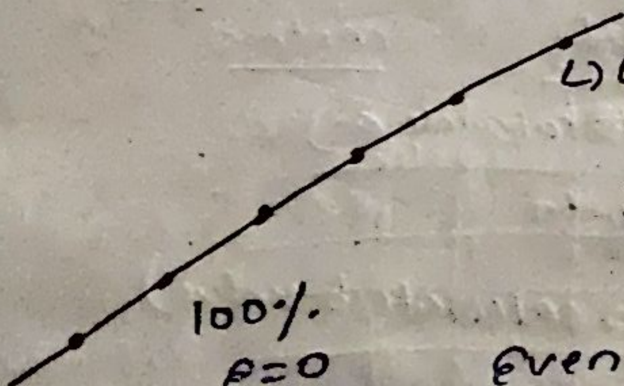
R^2

for R^2,

$$\frac{\text{least squared distance}}{\frac{\text{variance (mean)}}{\text{variance (fit)}}}$$

What does this equation tells?

↳ consumption = 49.13 + 0.85 (income) + e

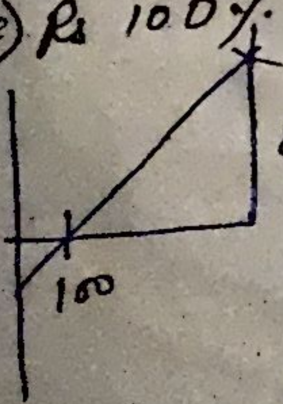


100%.
e=0
y=x

Income is in weeks.
Consumption is in weeks.

Even if income is zero, there is some consumption, hence, we need to make assumption in such way so that we get real consumption rate.

⇒ Rs 100% - every week



85

100

imp.

β_0 β_1 ε

beta beta error

(In some cases, Intercept is also important).

Least Square : use to reduces residual.

lecture-21

5th March, 2024, Tuesday)

$$y = \underline{a} + \underline{bx} + \varepsilon$$

a - intercept

b - slope

ε - calculation → square sum method.

a,b calculation → a - intercept → Radient decent method.

b - intercept

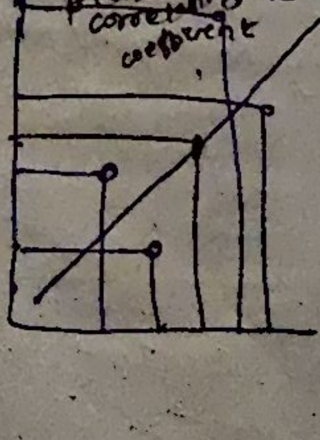
$a = \bar{y} - b\bar{x}$ (Inorder to calculate a value.)

standard deviation

$b = r, \frac{Sy}{Sx}$ (Inorder to calculate b value)

Pearson correlation coefficient

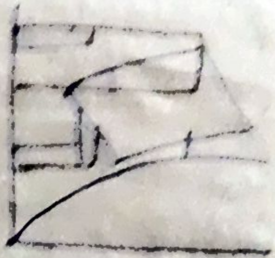
→ we can write)



$\boxed{\text{Slope}(x)\ \beta_1} = r\frac{Sy}{Sx}$

$\text{Slope}(yz) = r\frac{Sy}{Sz}$

$$\frac{SS(mean) - SS(fit)}{SS(mean)}$$



$\hat{y}$

$y = y_{intercept} + \text{slope}(x) + \text{slope}(z) + \epsilon$

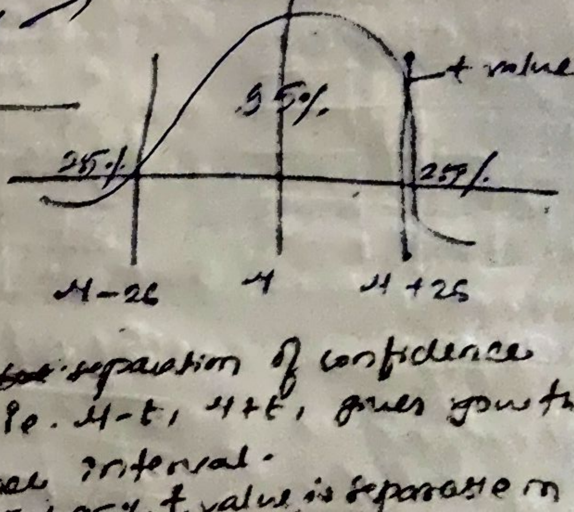
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n + \epsilon$

slope of $\beta_n = \gamma \frac{\delta y}{\delta n}$ | $\frac{\text{slope}(x)}{\beta_1} = \gamma \frac{\delta y}{\delta x}$

$\text{slope}(z) = \gamma \frac{\delta y}{\delta z}$

goodness of equations can be set by → $\boxed{R^2 \text{ value}}$

Matrix representation of multiple regression



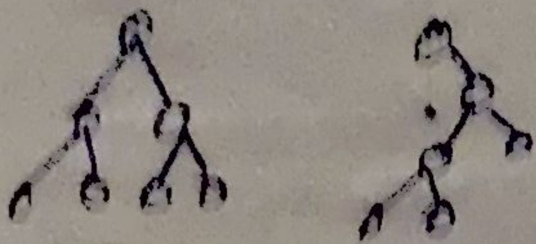
t-value : declares the ~~...~~ separation of confidence interval. i.e. μ − t, μ + t, gives you the confidence interval.
that 95% t value is separate in above diagram.

μ-range : one which is closest to μ mean.

confidence interval [μ − 2σ, μ + 2σ]

// lecture 09 (11th March, 2024)

Random forests



1) Bootstrap datasets :

A	B	C	D	class
1				
2				
3				
4				

ABCD.

B.D.S

A	B	C	D	class
2				
4				
3				
4				

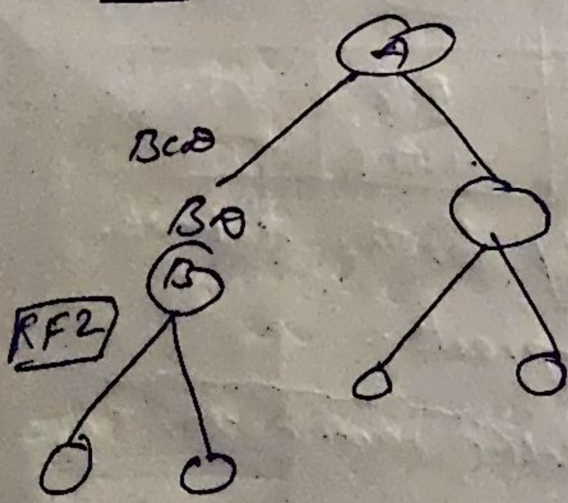
B	D	class
2		
4		
3		
4		

Create Boot strap Dataset

1	
1	
2	
3	

- Original dataset contain 4 Tuples
- Even Boot strap Dataset contain 4 Tuples.

ABCD



Note:

→ Take every time you sample.
→ sample ⇒ Bootstrap Dataset
→ Try to do more random.

Test Sample

A	B	C	D	class
x_1	x_2	x_3	x_4	?

(By voting; we conclude which class does it belong)

R.F1 —
which is better?

tuple belongs to yes classes

Yes	No
1+1	1 — Error has to be minimized
5	

Bagging : Constructing or Bootstrapping;
take the voting,
test the sample

Out-of-Bag Dataset :- All the tuples which will not come into Bootstrap; That datasets is called Out-of-Bag dataset.
(dataset)

↳ About $\frac{1}{3}$rd of data will not enter into Bootstrap datasets.

lecture 23 (12th March, 2024) Tuesday

Out-of-Bag-sample

A-B-C-D – Yes

→ Sample contains class label

set of Random forest: RF1 – yes, RF2 – yes, RF3 – NO

Maximum vote

That sample is classified.

correctly or not that will tell you out-of-Bag-error

$\frac{2}{10}$ → error

0.2

$\underbrace{\quad\quad\quad\quad}_{\text{Accuracy of dataset.}}$

Out-of-Bag-Error

- A-B-C-D (can select 3 or 4)

→ which has less error consider to be good Random forest.

Steps / Functionality:

- Create a bootstrap (simply replicas of no of counts)
- Select few no. of attributes. (Suppose 2 in this case).
- From out of Bag datasets it is not included in Bootstrap dataset.
- Verify with Random forest.
- verify according to the class.
- calculate Error values.

If datasets (original) has some missing values:

~~Good Blood Circulation~~

Chest pain	Good Blood circulation	Blocked	Weight	~~Height~~ Heart Diagnosis
No	No	No	125	No
yes	yes	yes	180	yes
yes	yes	No	210	No
yes	yes	? No	? 167.5	No

We need to construct similarity for this datasets.
Similarity means create random forests.
At every node (If two tuples ends in same node) then similarity is 1.

(10RF)

Suppose;

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		8
4	1	1	8	

(3 and 4 are ending in same leaf node)

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weighted frequency of 'yes' = $\frac{1}{3}$ × the weight for yes

$yes = \frac{1}{3} = 0.1 = \frac{0.1}{1} = 0.1$

$no = \frac{2}{3} = 0.1 + 0.1 + 0.8$

$\frac{2}{3} * \frac{0.1 + 0.8}{0.1 + 0.1 + 0.8} = 0.03$

$\Rightarrow 125 \times 0.1 + 180 \times 0.1 + 210 \times 0.8$

$\Rightarrow 198.5$

If yes Tuple;

yes | yes | — | 190 | yes

lecture 27

Adaboost

1. Order
2. Stamp
3. weak learners

Three important steps made by Adaboost:

1) Adaboost combines a lot of weak learners to make the classification.

2) Some stamps get more steps in the classification than others.
say

3) Each stamp is made by the previous stamp mistakes into accure.

~~Also~~

Chest pain	Blocked Arteries	Patient weight	Heart Disease	SW (sample weight)	New sample weight (NSW)	New weight
yes	yes	205	yes	1/8	0.05	0.07
no	yes	180	yes	1/8	0.05	0.07
yes	no	210	yes	1/8	0.05	0.07
yes	yes	167	yes	classification of wrongly classified 1/8	0.05	0.48
no	yes	156	no	1/8	0.05	0.07
no	yes	125	no	1/8	0.05	0.07
yes	no	168	no	1/8	0.05	0.07
yes	yes	172	no	1/8	0.05	0.07
					0.68	1.00

new weight

Questions

we can say:

After Gini index we find the result

weight > 176

yes

C	NC
3	0

no

C	NC
4	1

← first stamp

Hence, above tuple is wrongly classified, we need to m[...] the weight of the particular ~~tuple~~ tuple.

We can give the sample weight.

Here, the total error will be between 0 and 1.

[0 1]

If error = 0, perfect stump.

If error = 1, it has error.

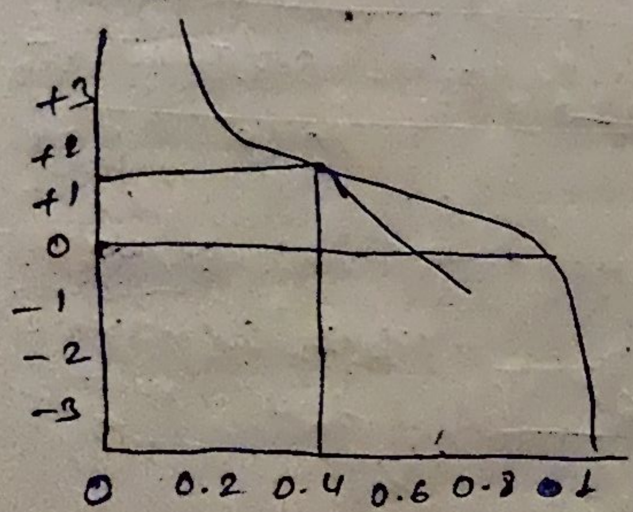
→ The total error to be determined.

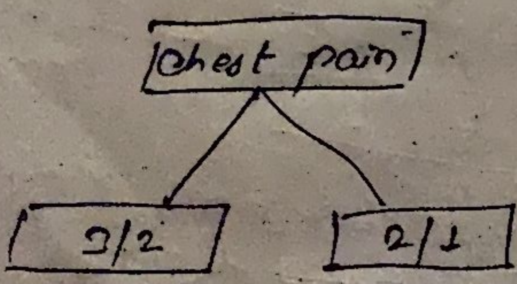
Amount of say this stump has the final classification with the formula.

$$\text{Amount of say} = \frac{1}{2}\log\left(\frac{1-\text{total error}}{\text{total error}}\right)$$

$$= \frac{1}{2}\log\left(\frac{1-1/8}{1/8}\right)$$

$$= 0.97$$





$$\text{Total error} = \frac{1}{8} + \frac{1}{8} + \frac{1}{8}$$

from formula,

Amount of say $= \frac{1}{2}\log(2/3) = 0.42$

first stamp is there need to find another stamp.

new sample weight $=$ sample weight $\times e^{\text{(amount of say)}}$

$= \frac{1}{8} \times e^{0.97}$

$= \frac{1}{8} \times 2.64$

$= 0.33$

(in case of error because we want to increase the wt for the incorrect variables to amount of say)

new sample weight $=$ sample wt $\times e^{-\text{amount of say}}$

$= \frac{1}{8} \times e^{-0.42}$

$=$

Weighted Gini index

↳ to calculate the next stamp

lecture-25

21st March, 2024

Unsupervised learning

Two important classification :

↳ partition based

↳ Hierarchical based → depending on aerial group that is created linkage between them.

where, we use partition and Hierarchical based method.

Leader's Clustering Algorithm

↳ predefined value (ϵ)

10 points in dataset

1 point which is taking in 1 Cluster

1

2

3

4

5

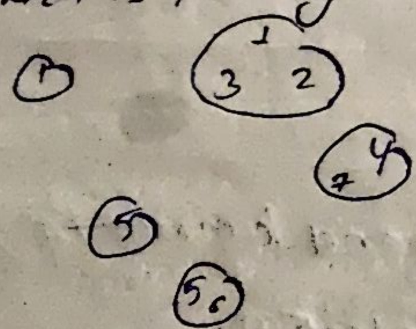
6

7

8

9

10



$d(1,2) < \epsilon$

$d(5,6) < \epsilon$

$d(4,7) < \epsilon$

like that we populate things.

Adv : Traversing will be done only once.

whether if we change the order ~~Exte~~ ex: start from point 10, then this will be different. (Cluster)

Imp point

- order of point should be distributed.
- If we change the distance, their cluster will be change.

Dis adv:

↳ People don't use this algorithm because it is order dependent.

k-means

Disadv

↳ k value, what to be selected as k value is difficult

↳ overlapping of points.

Adv (Convergence: No change in the cluster)

↳ It may not converge every time.

↳ This will converge very fast

I = very small
k = small
n = Big

Generally, $I << n$
$k << n$

k Means ++

difference between k means and k means ++

↳ Initial start point selection.

lecture-26

Maryagam

HCM

1. SL
2. CL
3. AL
4. CL

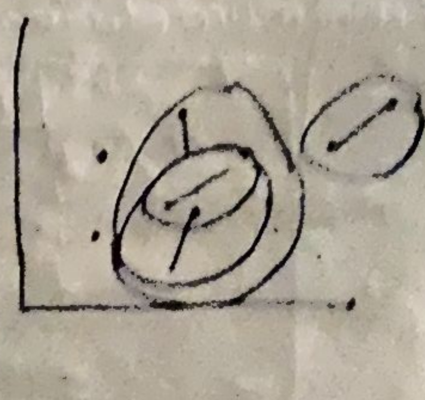
} Commonly agglomerating
↳ called Bottom-up approach.

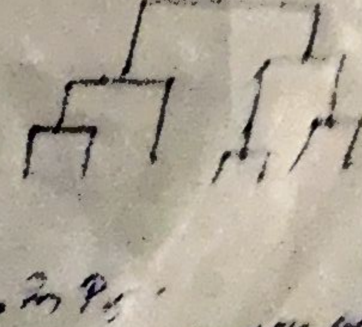
divisive method ⇒ Top down method

P_1 P_4

Suppose there is 10 points;
Initially every point is cluster;

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	42	0				
P_3	36	72	0			
P_4	50	41	33	0		
P_5	51	36	22	12	0	
P_6	12	22	27	13	51	0





Single link will find the minimum
Here taking first point, (12) which lies in P_5.
(we will compute the distance between each pair and find the minimum
distance and combine them as a new cluster.

	P_1	P_3	P_4P_2	P_5	P_6
P_1	0				
P_3	2	0			
P_4P_2	3	30	0		
P_5	20	5	11	0	
P_6	15	15	12	12	0

and we do the same process until we combine all pair.

4) to find the new value for each combine pair; we will find the min point between them.

$P_1 (P_4, P_5)$

$\min [d(P_1, P_4), d(P_1, P_5)]$

$\min [d(P_2 P_4), d(P_2, P_5)$

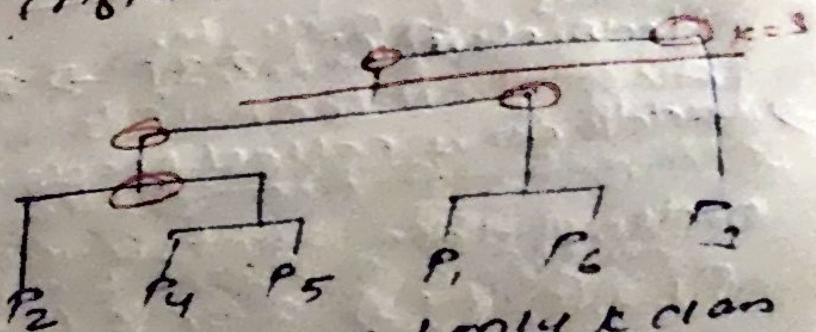
~~Step~~

(The maximum distance will be P_1 and P_6.)

New Table;

	$P_1 P_6$	$P_2 P_4 P_5$	P_3
$P_1 P_6$			
$P_2 P_4 P_5$	0		
P_3			

(After that we will draw the graph)



	$P_1 P_6 P_2 P_4 P_5$	P_3
$P_1 P_6 P_2 P_4 P_5$		
P_3		

(when we want only k class from our data we will draw a line. This is cut the graph in k point → we get k class)

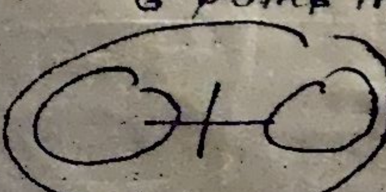
Hierarchical clustering single link method

II Complete link clustering Method / complete leakage

SL Min C
CL Max C
Average
Centroid — 3 methods. methods have its own dendrogram.

$$\frac{d(P_1, P_2) + d(P_1, P_6)}{2}$$

6 point in cluster



(each methods are having their dedogram. These all are bottom up approach).

lecture-27
4th April, 2023

~~k-Medoids~~

k-Medoids

Medoids of {1, 2, 3} ⇒ 2 is medoid

⇒ Middle no is called Medoids

order of data points is important to find Medoids.

{1, 2, 3, 4} ⇒ 2 or 3 is medoid.

for clustering,
The centroid point is not a real point.
whereas for medoid the mid point is considered.

		x	y	$d(P_i, m_1)$	$d(P_i, m_2)$	Cost (min distance)
M_1 —	P_1	2	3	0	2	0
M_2 —	P_2	4	3	2	0	0
M_1	P_3	2	1	2	4	2
	P_4	4	1	4	1	1
	P_5	3	3	1	1	1
	P_6	6	1	6	4	4

	P_1 M_1	M_2 (3,3)	$m_1 m_2$
P_1 (2,2)	0	2	0
P_2 (3,3)	2	0	0
P_3 (4,4)	4	2	2
P_4 (2,1)	1	3	1
P_5 (3,1)	2	2	2
P_6 (4,1)	3	3	3
			8

$k=2$

PAM → partition around medoids

M_1	M_2
P_1 P_4 P_2 P_5	P_2 P_6

	M_1	M_2 (4,4)	$m_1 M_2$
P_1 (2,2)	0	4	0
P_2 (3,3)	2	2	2
P_3 (4,4)	4	0	0
P_4 (2,1)	1	5	1
P_5 (3,1)	2	4	2
P_6 (4,1)	3	3	3
			8

BRICH Clustering algorithm → based on BTree structure Balance Reduce Iterative Clustering.

$\boxed{R, D}$

Radius of a cluster $= \dfrac{d_1 + d_2 + d_3 + d_4}{4}$

Avg of all the points.

Diameter of the cluster

Avg sum of highest distance

$= 2 \times R$.

Sum of all distances ↑ 4 points

→ all the points

$$\frac{\sum d(\ )}{n(n-1)} \qquad \frac{{}^4C_2}{2\,C_2}$$

↑

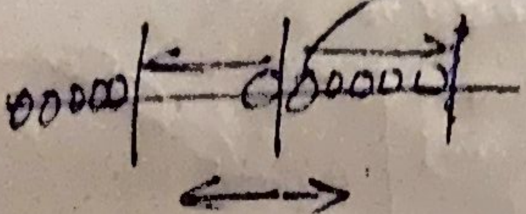
diameter

lecture 28
8th April, 2024, Monday..

Cross validation

Support vector classifier
Support vector Machine

→ nearest point from the threshold

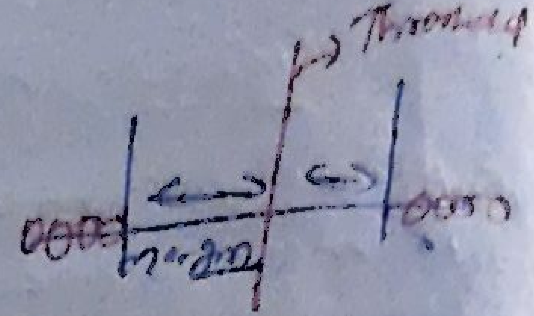


→ need to find maximum margin classifier

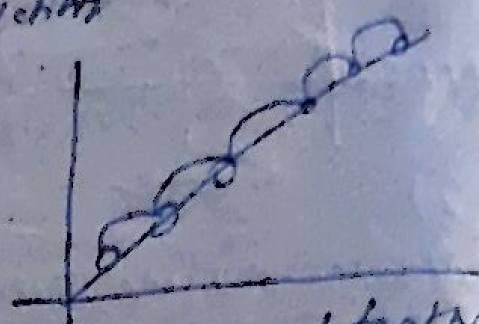
Two classes are there and classified by Threshold value.
↳ can change the threshold value which satisfy the data
edge-points of the.
↳ maximal marginal classifier

3. Support vector classifier (memin) → for multinomial classification

Maximal margin classifier



Penalty Term → using cross validation



$\frac{d^2}{e}$ $\sqrt{d}$

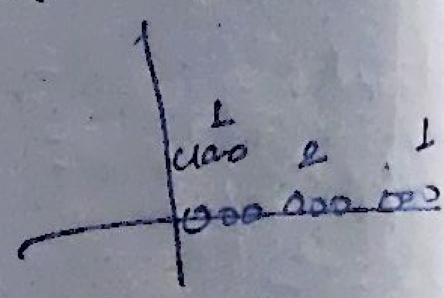
$\frac{1}{}$

after that find [SVC] and applying cross validation. (transform the point to other dim and find the ~~distance~~ which is the distance between two point i.e. original and next point)

if we have data in one specific dimension distributed in way we can't apply ~~this~~ SVM to classify ⇒ so we increase the dimension of the data.

So, we can separate the point in higher dimension so, the relation will be so clearly.

we can this way when we can't have soft margin.



// Lecture: 23

9th April, 2024, Thursday

$P_1(3,1)$

$P_2(3,-1)$

$P_3(6,1)$

$P_4(6,-1)$

$P_5(1,0)$

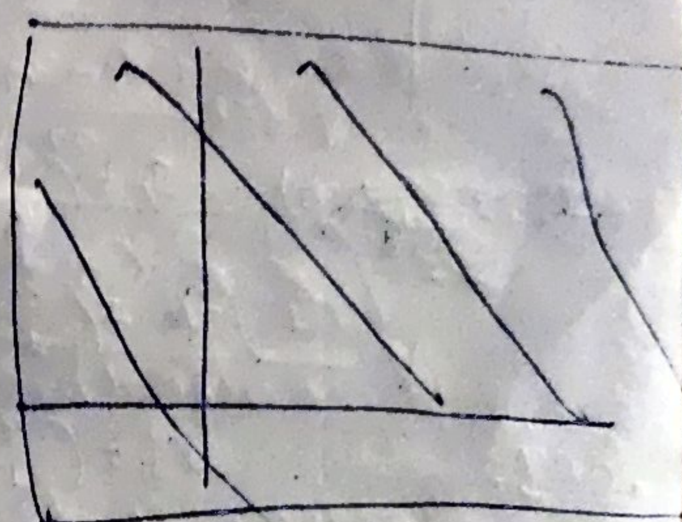
$P_6(0,1)$

$P_7(0,-1)$

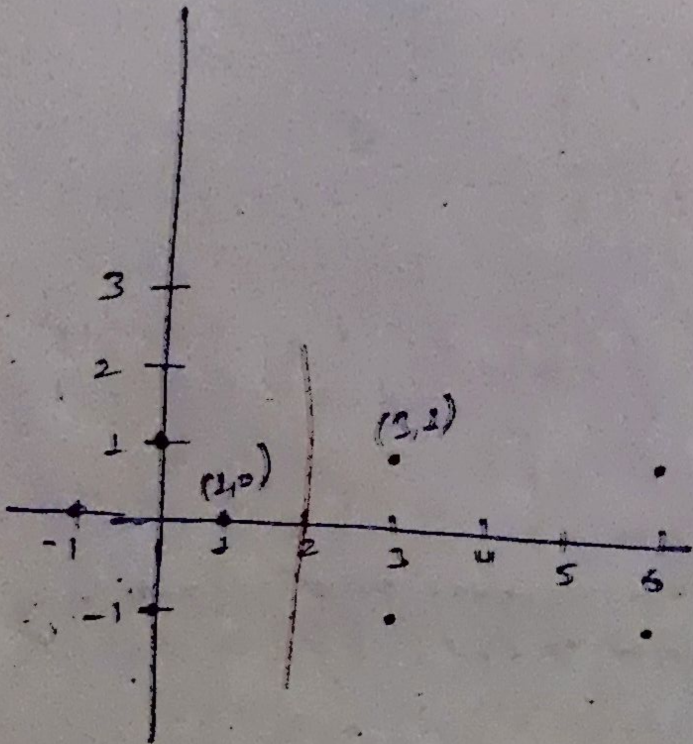
$P_8(-1,0)$

Find out the hyperplane explaination for the data using SVM (Support Vector Machine)

=) Solution:



J.7.0.



$s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \tilde{s}_1$

$s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ $\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \tilde{s}_2$

$s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ $\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \tilde{s}_3$

$\alpha_1 \tilde{s}_1 \tilde{s}_1 + \alpha_2 \tilde{s}_2 \tilde{s}_1 + \alpha_3 \tilde{s}_3 \tilde{s}_1 = -1$

$\alpha_1 \tilde{s}_1 \tilde{s}_2 + \alpha_2 \tilde{s}_2 \tilde{s}_2 + \alpha_3 \tilde{s}_3 \tilde{s}_2 = +1$

$\alpha_1 \tilde{s}_1 \tilde{s}_3 + \alpha_2 \tilde{s}_2 \tilde{s}_3 + \alpha_3 \tilde{s}_3 \tilde{s}_3 = +1$

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

$2x_1 + 4x_2 + 4x_3 = -1$

$4x_1 + 11x_2 + 3x_3 = +1$

$4x_1 + 3x_2 + 11x_3 = +1$

$x_1 = -3.5$

$x_2 = 0.75$

$x_3 = 0.75$

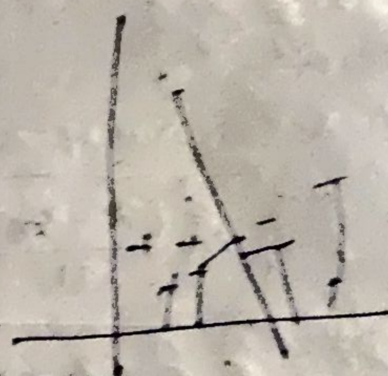
$$-3.5\begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75\begin{pmatrix}3\\-1\\1\end{pmatrix}$$

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

$y = wx + b$

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

$b = -2$



Lecture-30

15th April, 2024, Monday

$\boxed{\text{BIRCH}}$

(Balanced iterative Reducing and clustering using Hierarchies)

• **CF** – clustering Feature (define entity of a this cluster)

⟨n, LS, SS⟩ $\boxed{CF}$

Radius } of cluster
Diameter }

centroid = x_0

$$\boxed{x_0 = \sum_{i=1}^{n} \frac{x_i}{n}}$$

Radius :- average sum of distances from centroid from all other points

$$R = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_0)^2}{n}}$$

$$D = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2}{n(n-1)}}$$

• **CFT** :-

Two important factors:

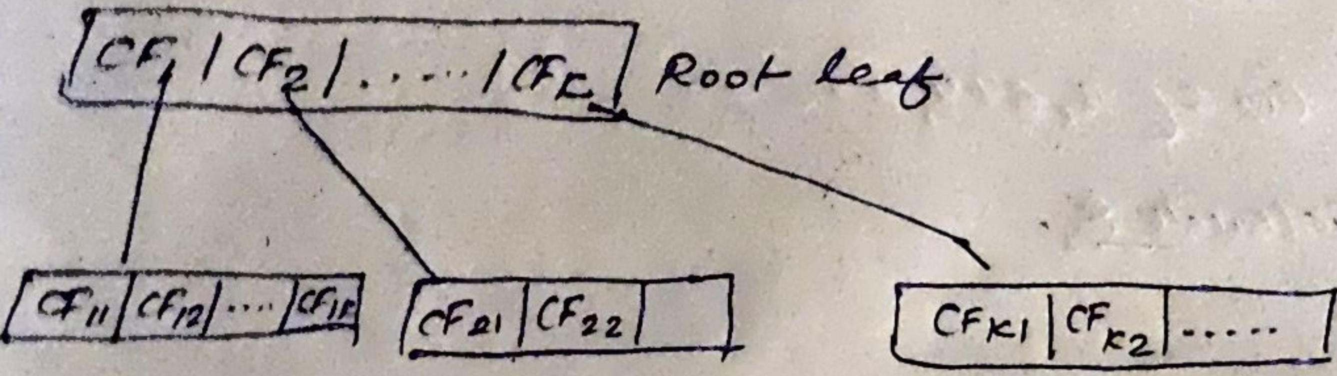
↳ (B) Blocking Factor :- max no of children for a given leaf node.

↳ (T) Threshold

B → Maximum number of children per non leaf node

T → Specifies the max diameter of sub cluster stored at the leaf nodes of the tree.

we have,



for ex1, Two point.

(2, 3)

(4, 1)

$CF_1 =$ ~~<...>~~ $\langle 1, 5, 13 \rangle$

$CF_1 =$ ~~<...>~~ only summary available in the main memory.

Calculate the radius with formula;

$$\boxed{R = \sqrt{\frac{SS - \frac{LS^2}{N}}{N}}}$$

$SS \Rightarrow 25$

keeping in formula;

$$\cancel{\sqrt{\frac{25 - 25}{1}}} \Rightarrow \sqrt{\frac{(13 - 25)}{\sqrt{-12}}}$$

centroid.

points

Radius $= 0$ (since there is only one point i.e. 0)

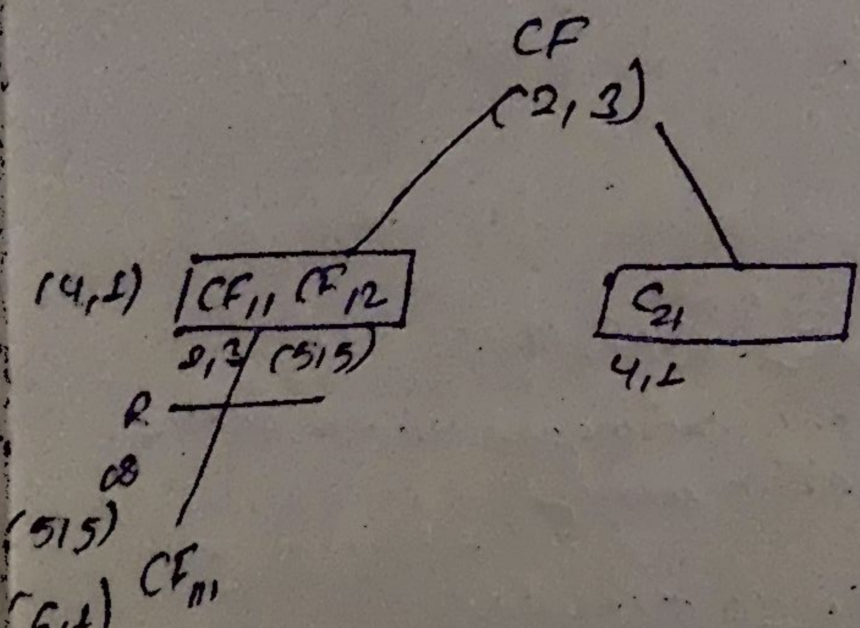
When we are going to split particular node;

After adding,

(4, 1)

$$CF_2 = (2, 6+4, 2^2+3^2+4^2+1^2)$$

Calculate R,



If two points are failed, then third point came split again.

Main advantage: Scalability.

only defined using Clustering Feature (CF).

→ Threshold is user defined.

→ Reduce threshold, cluster increase.

$\boxed{T>>}$

(less no. of clusters)

$\boxed{T<<}$

(More no. of clusters)

Iteration = (i)

i = 1 (one iteration)

t = 3

O O O

These cluster is used to CF.

<u>How they are created?</u>

<u>Example</u>:

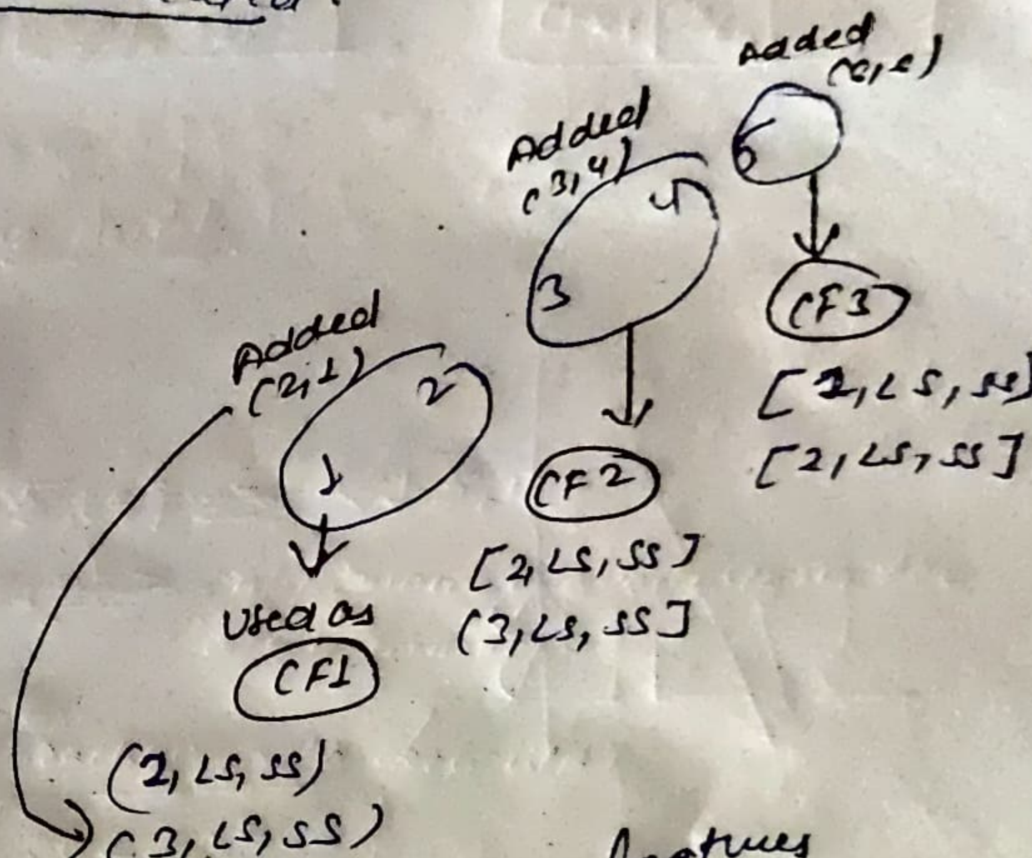
(1,1)

(2,2)

(3,3)

(4,4)

(5,5)



<u>Note</u>: features

- Based on linear sum and square sum, we can calculat
- Used for large amount of data.