POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

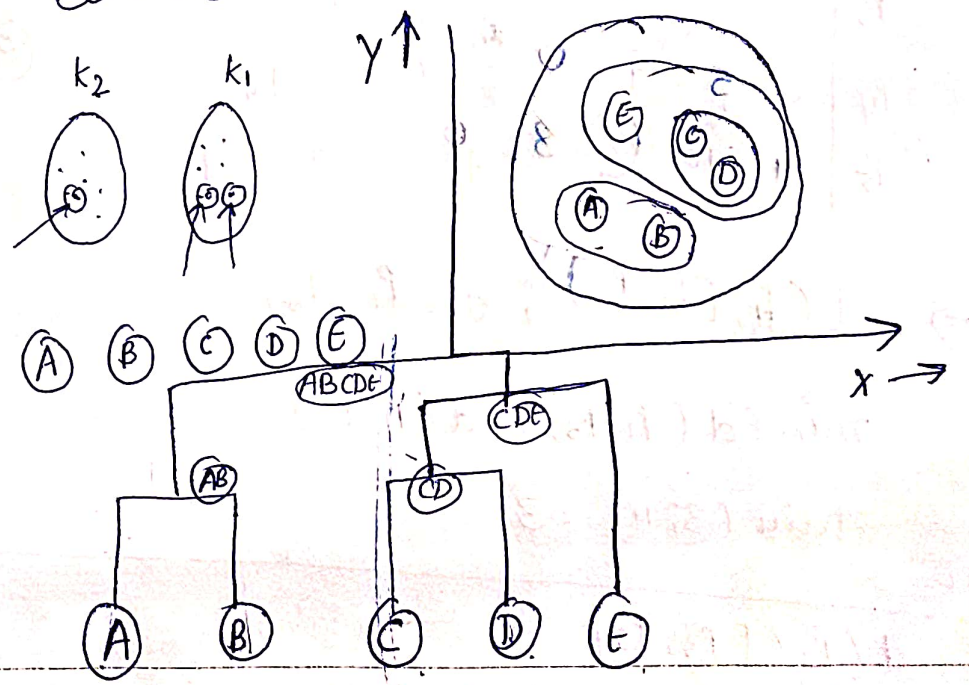
PAGE NO

Regression is a statistical that helps us understand and predict the relationship between variables

Hierarchical clustering

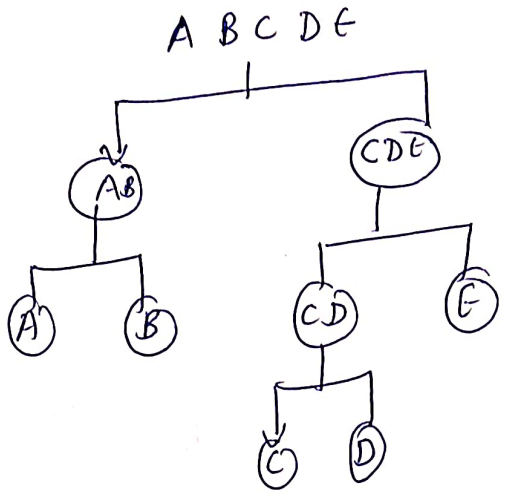
→ Agglomerative [Dendrogram]

→ Divise



Aggm bottom to Top.

→ Divise Algo:- Top to bottom.



Agglomerative clustering:-

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

⇒

	P_1	P_2	$[P_3\ P_5]$	P_4
P_1	0			
P_2	9	0		
$[P_3\ P_5]$	3	7	0	
P_4	6	5	8	0

⇒ ~~$d([P_1, [P_3\ P_5]), d(P_1, P_5)$~~

$\min(d(P_1, P_3), d(P_1, P_5))$

$\min(3, 11) = 3$

$d(P_2, [P_3, P_5])$

⇒ $\min(d(P_2, P_3), d(P_2, P_5))$

⇒ $\min(7, 10)$ ⇒ 7

$d(P_4[P_1\ P_3\ P_5)]$

$\Rightarrow \min(P_4,P_1), d(P_4,P_3), d(P_4,P_5]$

$\Rightarrow \min(6, 9, 8)$

$\Rightarrow \underline{6}$ ✓

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

	$[P_1\ P_3\ P_5]$	P_2	P_4
$[P_1\ P_3\ P_5]$	0		
P_2	7	0	
P_4	6	5	0

	$[P_1\ P_3\ P_5]$	$[P_2\ P_4]$
$[P_1\ P_3\ P_5]$	0	
$[P_2\ P_4]$	6	0

$[P_1\ P_2\ P_3)\ [P_4\ P_5]$

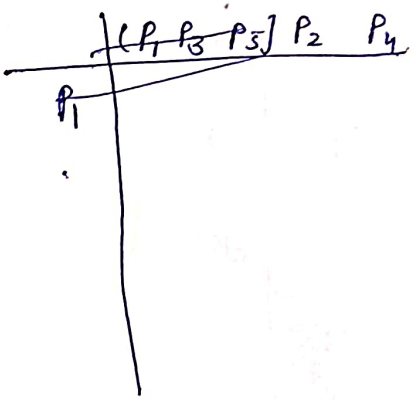
9 [

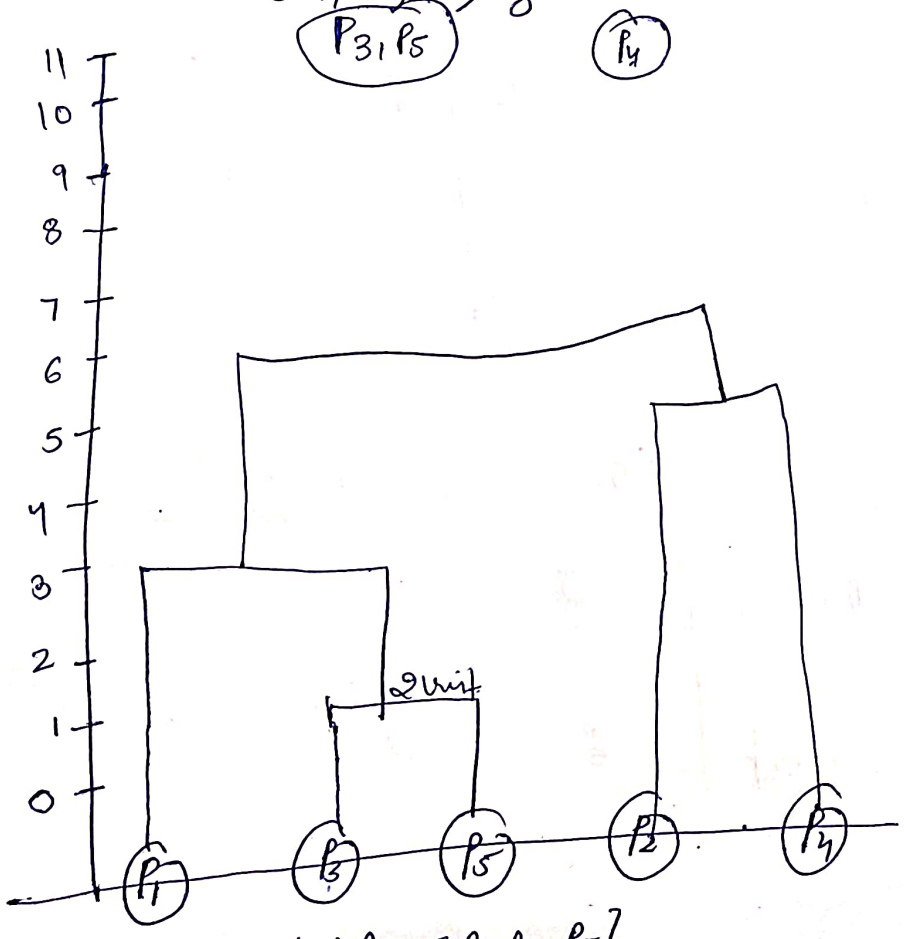
PAGE NO. .....

$\Rightarrow d[P_4, (P_3, P_5)]$

$\Rightarrow \min d(P_4, P_3), d(P_4, P_5)$

$\Rightarrow \min(9, 8) \Rightarrow 8$





$d[P_2, [P_1\ P_3\ P_5]$

$\min(d(P_2\ P_1), d(P_2, P_3), d(P_2, P_5))$

$\min(9, 7, 10) = 7$

	$P_1\ P_3\ P_5$	$(P_2\ P_4]$
$[P_1\ P_3\ P_5]$	0	
$(P_2\ P_4]$	⑥	0

$d[P_1\ P_3\ P_5][P_2\ P_4]$

$\Rightarrow \min d(P_2, P_1), d[P_3, P_3], d[P_2, P_5], d[P_4, P_1]$
$d[P_4, P_3), d[P_4, P_5)$

$\Rightarrow \min(9, 7, 10, 6, 9, 8)$

$\Rightarrow$ ⑥

PAGE NO. .....

	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0	3.16	2.82	5.09	6.32	2.23
S_2		0	3.16	4.47	7.07	5
S_3			0	7.07	8.94	3
S_4				0	3.16	7.28
S_5					0	8.06
S_6						0

$\{S_1, S_2, S_3\}$
$\{S_4, S_5\}$
$\{S_6\}$
3.17
3.17

Divise Algorithm:- (a, b, c, d, e)

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

$C_i = \{a, b, c, d, e\}$

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

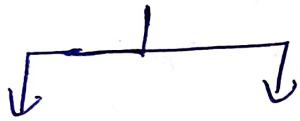
probabilistic clustering :-

Basic concept of probabilistic model-Based clustering.

mixture models for cluser analysis.

⇒ Gaussian mixture models.

⇒ the expectation-maximization Emalgo.



⇒ analysis of mixture model method.

what is probabilistic model :-

⇒ model the data from a generative procen.

⇒ Assume the data are generated by a mixture of underlying probability distribution.

⇒ Attempt to optimize the fit between the observed data and some mathmetical model using a probabilistic approach.

Probabilistic model-based clustering :-

⇒ each cluster can be represented mathematically by a parametric probability distribution (eg. a Gaussian or poisson distribution)

⇒ cluster : Data points (or objects) that most likely belongs to the same distribution.

⇒ clustering parameter estimation so that they will have a maximum likihood value to the model by a mixture of k component distribution (e.g k cluster)

Application :- image segmentation, document clustering.

Two type
mixture model

⇒ A very continous probability distribution is Gaussian distribution it is called as a bell curve. the eq. of one-dimension

$$P(X) = \frac{1}{\sqrt{2\pi\sigma \exp\left[-(x-\mu)^2/2\sigma^2\right]}}$$

σ → standrd μ → mean

Expectation-maximization algo.

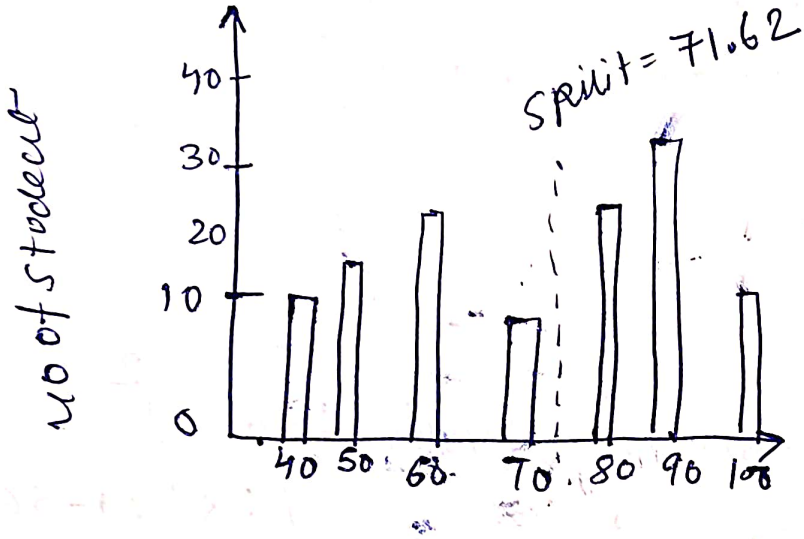
A general techniques to find maximum likelihood estimation in mixture models.

Em algo. for Gaussian mixture model.

Example:-

PAGE NO. .....

Histogram of exam score:-



score	num	$5 \times num$
40	10	
45	14	
50	19	
55	28	
60	22	
65	19	
70	16	
75	12	
80	30	
85	34	
90	24	
95	19	
100	12	
	Σf_i	$\Sigma x_i f_i$

total no student = 259

Step-① spirit the histogram:-

inial spirit point = $40 \times 10 + 45 \times 14 + 50 \times 19 + 55 \times 28 + 60 \times 22 + 65 \times 19 + 70 \times 16 \times 75 \times 12 + 80 \times 30 + 85 \times 34 + 90 \times 24 + 95 \times 19 + 100 \times 12$

total mean :- $\dfrac{}{10+14+19+28+22+19+16+12+30+34+24+19+12}$

$= 71.62$

itially, score less than 71.62 is treated as the first component.
score greater than 71.62 is treated as the second component.

Step-II initialization:-

First component

Score	num
40	10
45	14
50	19
55	28
60	22
65	19
70	16

$$\text{mean} = \frac{40\times10+45\times14+50\times19+55\times28+60\times22+65\times19+70\times16}{10+14+19+28+22+19+16}$$

$$m_1 = 56.21$$

$$\text{variance} = \sigma^2 = \frac{(x_i-\bar{x})^2}{n-1} = \frac{10\times(40-56.21)^2+14\times(45-56.21)^2+\cdots\cdots+16\times(70-56.21)^2}{(10+14+19+28+22+19+16)-1}$$

$$V_1 = 78.64$$

II component:- initialization:-

$$\text{mean} = \frac{75\times12+80\times30+85\times34+90\times24+95\times19+100\times12}{12+30+34+24+19+12}$$

$$m_2 = 86.68$$

$$\text{variance} = \frac{12\times(75-86.68)^2+30\times(80-86.68)^2+\cdots\cdots+12\times(100-86.68)^2}{(12+30+34+24+19+12)-1}$$

$$V_2 = 52.16$$

Step:-3 iteration (compute new weight)

this is the eq. to compute the new wight.

$$W_{k,new} = \frac{1}{N} \sum_{n=1}^{N} \frac{w_k \cdot G(x_n | m_k, v_k)}{w_k \cdot G(x_n | m_k, v_k)}$$

$N \rightarrow$ No. of students

$w_k \rightarrow$ weight (either w_1 or w_2) $= 0.5$

m_k — $m_1 = 56.21$ & $m_2 = 86.68$

v_k :- variance $v_1 = 78.64$

$v_2 = 52.16$

$x_n \rightarrow$ exam score (40, 45, ---100) from each of the student

259.

$k \rightarrow$ cluster value $= 2$

$$G(x | m_1, v_1) = \frac{1}{(2\pi)^{1/2} |v_1|^{1/2}} \exp\left[-\frac{1}{2}[x-m_1]^T v_1^{-1} [x-m_1]\right]$$

$$= \frac{1}{\text{(let solve numerator First)}}$$

$$= \frac{1}{(2\times 3.14\times 78.64)^{1/2}} \exp\left[-\frac{[X-56.21)\times[X-56.21}{2\times(78.64)}\right.$$

$X \rightarrow$ substitute!- 40, 45, 50, 55, 60, 65, 70, 75
80, 85, 90, 95, 100

$G(x_1|v_1) =$ -0.008, 0.002, 0.035, 0.045, 0.041, 0.028,
0.013, 0.005, 0.001, 0, 0, 0, 0

as II Component:-

$$\frac{1}{(2\times 3.14\times 52.16)^{1/2}} \exp\left[-\frac{(X-86.68)\times(X-86.68)}{2\times[52.16]}\right]$$

$X \rightarrow$ 40, ~~50~~ 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
$\downarrow$
-0 0 0 0 0 0.001, 0.004, 0.015, 0.036,
0.054, 0.05, 0.028, 0.0

$$= \left[\frac{0.5\times 0.008}{0.5\times(0.008+0)} \times 10 + \frac{0.5\times 0.002}{0.5\times(0.002+0)} \times 14 + \frac{0.5\times 0.035}{0.5\times(0.035+0)} \times 19\right.$$

$$+ \frac{0.5\times 0.045}{0.5\times(0.045+0)} \times 28 + \frac{0.5\times 0.041}{0.5(0.041+0)} \times 22$$

$$\left. + \frac{0.5\times 0.028}{0.5(0.028+0.001)} \times 19 + \text{-------} \right] \times \frac{1}{259}$$

$$= 0.494$$

II Substitution:-

$$\left[\frac{0.5\times 0}{0.5\times(0.008+0)}\right]\times 10+\frac{0.5\times 0}{0.5\times(0.002+0}\Big]\times 14+\frac{0.5\times 0}{0.5\times(0.035+0}\times 19$$

$$+\frac{0.5\times 0}{0.5\times(0.045+0}\Big)\times 28+\frac{0.5\times 0}{0.5\times(0.041\times 0}\Big)\times 22$$

$$\cdots\cdots\cdots\cdots\frac{0.5\times 0.01}{0.5\times(1+0.01)}\times 12\Big]\times\frac{1}{259}$$

$$= 0.506$$

$w_1 = 0.5$

$w_2 = 0.5$

$w_1 = 0.494$

$w_2 = 0.506$

$w_1 = 0.494$, $m_1 = 56.45$, $v_1 = 85.95$

$w_2 = 0.506$, $m_2 = 86.46$, $v_2 = 57.64$

Em algorithm:- em algo is a problem in which the data d is a set of instances generated by probability distribution that is mixure of k district normal distribution. the case where k=2 and where the ~~distance~~ instances are the point x axis. Each instance is generated using a two-step proces.

First one, k normal distribution is seted at random. Second a single random instance x_i is generated according to this seted attributed.

where each of k-normal distribution has a same variance σ^2

the learning task is ouput H

$$= \{ u_1 \cdots u_k \}$$

the mean of each of the k distribution.

it is easily calculated max. likelihood hypothesis for the mean of a single normal distribution observed data instance $x_1, x_2 \cdots x_n$

n → training instance

$$u_{ML} = \underset{u}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - u)^2$$

in case the sum of sequred error is minimized by sample mean.

DETAILED LECTURE NOTES

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

$$\mu_{ML} = \frac{1}{M}\sum_{i=1}^{m} x_i$$

Step-I :- Calculate the expected value $E[z_i]$ of each hidden variable z_{ij}, assuming the current hypothesis $h = (\mu_1, \mu_2)$

Step 2 :- Calculate a new maximum likelihood $h = (\mu_1, \mu_2)$ assuming the value taken on by each hidden z_{ij}.

$$E[z_{ij}] = \frac{P(x = x_i) \mid \mu = \mu_j]}{\sum_{n=1}^{2} P(x = x_i) \mid \mu = \mu_n]}$$

Association rule mining :-

~~ARM also called as market~~

Association rule mining is one of the ways to find pattern in data.

- feature (dimension) which occur together
- feature (ee) which are correlated.

Association rule in any dataset where feature take only two values 0/1.

- market Basket analysis is a popular application of association rule mining.

set of item is transaction is called market Basket.

Example:- For the following given transaction Data-set Generate rule using

Support = 50% Confidence = 75%.

measure of effectiveness of the Rule:-

- support
- Confidence
- lift
- other Affinity, leverage

DETAILED LECTURE NOTES

Campus: ..... Course: ..... Class/Section: ..... Date: .....
Name of Faculty: ..... Name of Subject: ..... Code: .....

Transaction item

ID	Purchased
1	Bread, Cheese, Egg, Juice
2	Bread, ~~milk~~, Cheese, Juice
3	Bread, milk, Yogurt
4	Bread, Juice, milk
5	Cheese, Juice, milk

Support = $\frac{\text{no. of times item}}{\text{total no. of transaction}}$

Frequent item set:-

items	Frequency	Support
Bread	→ 4	$\frac{4}{5}$ = 80%
Cheese	→ 3	$\frac{3}{5}$ = 60%
Egg	→ 1	$\frac{1}{5}$ = 20%
Juice	→ 4	$\frac{4}{5}$ = 80%
milk	→ 3	$\frac{3}{5}$ = 60%
Yogurt	→ 1	$\frac{1}{5}$ = 20%

minimum support 50%

we will Remove these two then support is less than 50%.

make 2 - item candidate set and wrote their fequany.

item pair	Fequany	support
(Bread, cheese)	→ 2	→ 2/5 = 40%
(Bread, Juice)	→ 3	→ 3/5 = 60%
(Bread, milk)	→ 2	→ 2/5 = 40%
(cheese, Juice)	→ 3	→ 3/5 = 60%
(cheese, milk)	→ 1	→ 1/5 = 20%
(Juice, milk)	→ 2	→ 2/5 = 40%

There support ≥ 50%

For Rules make
- (Bread, Juice) — (I)
- (cheese, Juice) — (II)

(I) (Bread, Juice)

(Bread → Juice) (Juice → Bread)

Confidence :-

$(A \to B) = \frac{\text{support of } A \cup B}{\text{suppor}(A)}$

(1) Bread → Juice.

$\frac{S(B \cup J)}{S(B)} = \frac{3 \times 5}{5 \cdot 4} = 75\%$

(2) Juice → Bread

$\frac{S(J \cup B)}{S(J)} = \frac{3}{5} \cdot \frac{5}{4} = 75\%$

Both rule are good enough to be implement.

(2) Cheese → Juice)
Juice → cheese.

$\frac{S(C \cup J)}{S(C)} =$

$\frac{S(J \cup C)}{S(J)} =$ $3/5 \cdot \frac{5}{3} = 100\%$

$\frac{3 \cdot 5}{5 \cdot 4} = 75\%$

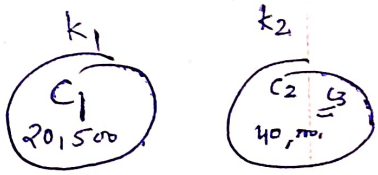
All rule are good.

K-Mean clustering :-

Sr NO	age	amount
C1	20	500
C2	40	1000
C3	30	800
C4	18	300
C5	28	1200
C6	35	1400
C7	45	1800

Euclidean Distance :-

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



$$C_3 = \sqrt{(30-20)^2 + (800-500)^2}$$

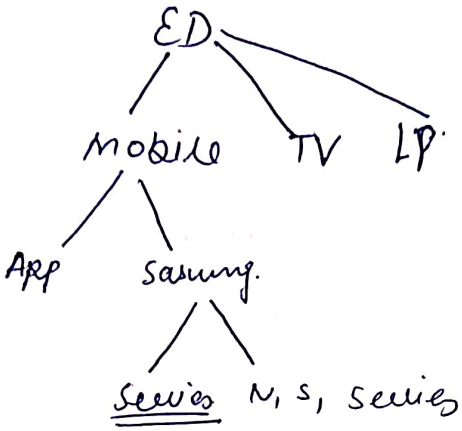
$$\sqrt{(10)^2 + (300)^2} = \underline{3000}$$

$$C_3 = \sqrt{(30-40)^2 + (800-1000)^2} = \underline{200}$$

K-mean clustering :-

Central, Hirarchy.

Tree divided.



Step-1 Take mean value.

Step 2 Find nearest no. of mean and put in cluster.

Step 3 :- Report one and two until we get same mean.

K = { 2, 3, 4, 10, 11, 12, 20, 25, 30 }

K = 2,

$M_1 = 4$ $M_2 = 12$

K_1 K_2

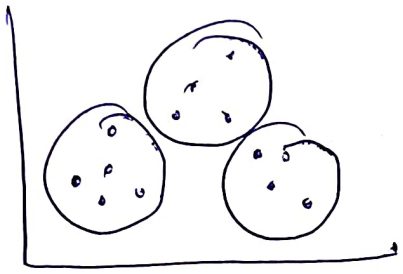
$K_1 = \{2, 3, 4\} \Rightarrow M_1 = 9/3 = 3$ (average)

$K_2 = \{10, 11, 12, 20, 25, 30\}$

$M_2 = \frac{108}{6} = 18$

Numerical:-

- k-mean clustering is an unsupervise iterative clustering technique
- it partitians the given data set into k predefined distuict cluster.
- A cluster is defined as a collection of data points cuttain similarties



O(n k t)

- n → no. of instance
  k → no. of cluster.
  t → no. of iteration

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,3)$ $A_4(5,8)$
$A_5(7,5)$, $A_6(6,4)$, $A_7(1,2)$, $A_8(4,9)$

intial cluster $A(2,10)$, $A_4(5,8)$ and $A_7(1,2)$

$$P(a,b) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

iteration 01:- $A_1(2,10)$, $C_1(2,10)$

$(2-2) + (10-10)$
$= 0$

$A_1(2,10)$ and $C_2(5,8)$
$(5-2| + |8-10| = 3+2 = 5$

PAGE NO.

$M_1 = 3$, $M_2 = 18$

$k_1 = \{2, 3, 4, 10\}$

$k_2 = \{11, 12, 20, 25, 30\}$

$M_1 = 4.75 (5)$, $M_2 = 19.6 = (20)$

$k_1 = \{2, 3, 4, 10, 11, 12\}$

$k_2 = \{20, 25, 30\}$

$m_1 = 7$, $M_2 = 25$

$k_1 = \{2, 3, 4, 10, 11, 12\}$

$k_2 = \{20, 25, 30\}$

$m_1 = 7$, $M_2 = 25$

Thus we are getting same mean we have to stop.

New cluster will be

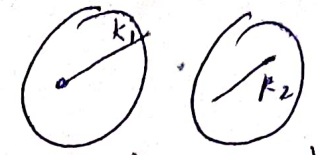
$k_1 = \{2, 3, 4, 10, 11, 12\}$

$k_2 = \{20, 25, 30\}$

PAGE NO.

Height weight.

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



(185, 72) (170, 56)

$k_1 \rightarrow \sqrt{(168-185)^2+(60-72)^2}$

$= 26.82$

$k_2 \rightarrow \sqrt{(168-170)^2+(60-56)^2}$

$= 4.48$

New centraid:-

$k_2 = \left(\frac{170+168}{2}\right), \left(\frac{60+56}{2}\right) = (169, 58)$

(185, 72) (169, 58)

$k_1 \rightarrow \{1, 4, 5, 6, 7, 9, 10, 11, 12\}$

$k_2 \rightarrow \{2, 3\}$

$k_1 = \sqrt{(179-185)^2 + (168-72)^2}$

$\Rightarrow \sqrt{(6.32)}$

$k_2 = \sqrt{(179-169)^2+(168-58)^2}$

$= 14.14$

PAGE NO .....

$A_1(2,10)$, and $C_3(1,2)$ =

$|1-2| + |2-10|$

$1 + 8 = 9$

Given Points	Distance (2,10) clus 01	Distance (5,8) clus 02	Distance (1,2) clus 03	Point belongs to clus-
A1 (2,10)	0	5	9	C1
A2 (2,5)	5	6	4	C3
A3 (8,4)	12	7	9	C2
A4 (5,8)	5	0	10	C2
A5 (7,5)	10	5	9	C2
A6 (6,4)	10	5	7	C2
A7 (1,2)	9	10	0	C3
A8 (4,9)	3	2	10	C2

DETAILED LECTURE NOTES

PAGE NO.....

suppose = 50%.
confidence = 75%.

Transaction ID	items purchased
1	Bread, cheese, egg, juice
2	Bread, cheese, juice.
3	Bread, milk, Yogurt.
4	Bread, juice, milk
5	cheese, juice, milk

Frequent itemset.

support → $\frac{\text{Bread}}{4}$

items	Frequency.	
Bread	4	4/5 x 100 = 80%.
cheese	3	3/5 x 100 = 60%.
egg	1	1/5 = 20%.
juice	4	4/5 = 80%.
milk	3	3/5 = 60%.
Yogurt	1	1/5 = 20%.

Item pairs	Frequency	support
(Bread, cheese)	2	2/5 = 40%
(Bread, juice)	3	3/5 = 60%
(Bread, milk)	2	2/5 = 40%
(cheese, juice)	3	3/5 = 60%
(cheese, milk)	1	1/5 = 20%
(juice, milk)	2	2/5 = 40%

For Rules → (Bread, juice)
→ (cheese, juice)

(i) (Bread, juice)

(Bread → juice) (juice → Bread

Confidence (A → B) = $\frac{\text{support}(A \cup B)}{\text{support } A}$

1) Bread → juice = $\frac{S(B \cup j)}{S(B)} = \frac{3}{5} \cdot \frac{5}{4} = \frac{3}{4}$ = 75%

(2) juice → Bread = $\frac{3 \cdot 5}{5 \cdot 4}$ = 75%

(ii) cheese → juice

$3/5 \cdot 5/3 = 100\%$

juice → cheese

$3/5 \cdot 5/4 = 75\%$

A. Frequent pattern set (L) is built which will contain all element whose frequny is greater than or equal to the

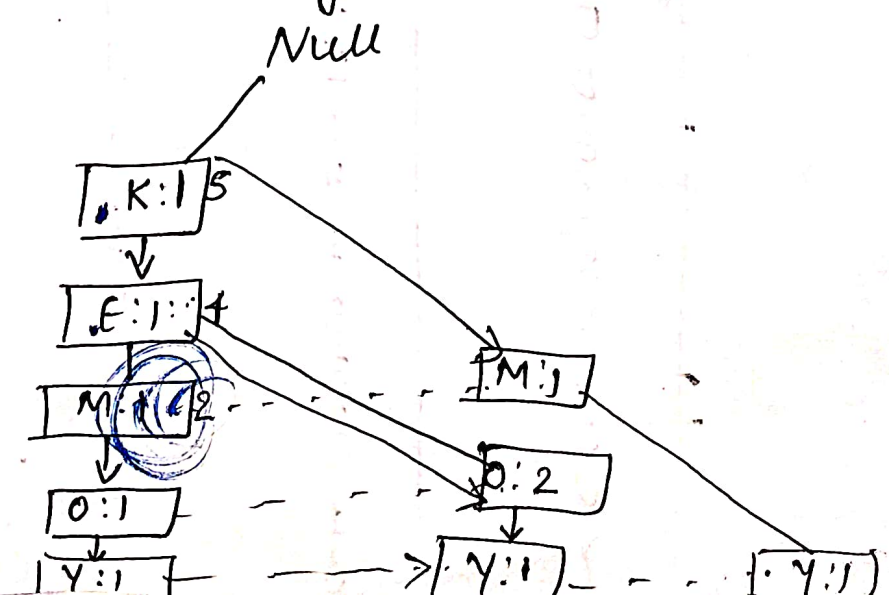
These elements are stored in desending order of their respectives.

$L = \{K:5, E:4, M:3, O:3, Y:3\}$

Frequnt pattern set $L = \{K:5, E:4, M:3, O:3, Y:3\}$

Transaction ID	items	ordered-itemset
T1	{E, K, M, N, O, Y}	~~{K, E, M, O, Y}~~ {K, E, M, O, Y}
T2	{D, E, K, N, O, Y}	~~{K, E, O, Y}~~ {K, E, O, Y}
T3	{A, E, K, M}	~~{A, E, K, M}~~ {K, E, M}
T4	{C, K, M, U, Y}	~~{C, K, M, U, Y}~~ {K, M, Y}
T5	{C, E, I, K, O, O}	~~{C, E, I, K, O, O}~~ {K, E, O}

a) inserting the set {K, E, O, Y}



DETAILED LECTURE NOTES

PAGE NO.....

Frequent pattern (FP growth):-

Transaction ID	items
T1	{E, K, M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O, O}

the above-given data is a hypothetical dataset of transaction with each letter representing an item

- Let the minimum support is 3.

item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

DETAILED LECTURE NOTES

PAGE NO.....

Y	{K, E, M, O : 1}, {K, E, O : 1} [K, M : 1]
O	{ K, E, M : 1], { K, E : 2]
M	{ K E : 2], { K : 1]
E	{ K : 4]
K	

Y	{K, E, M, O : 1}, {K, E, O : 1} { K, M : 1]	Condition fever {K : 3]
O	{[K, E, M : 1], { K, E : 2]	{ K, E : 3]
M	{ K, E : 2], { K = 1]	{ K : 3)
E	{ K = 4]	{ K : 4]
K		

Y	{<K, Y : 3 >
O	{< K, O : 3 >, < E, O : 3 >, < E, K, O : 3 >
M	{< K, M : 3 >]
E	{< E, K : 3 >]
K	

DETAILED LECTURE NOTES

PAGE NO.....

Association rule Mining :-

min support

Transcation ID	items
T1	HOTDOGS, Buns, ketchup
T2	HOTDogs, Buns
T3	HOTDogs, coke, chips
T4	chips, cake
T5	chips, ketchup
T6	HOTDogs, cake, chips

2. Transcation. Threshold = 33.34 %.

Q.2

T-ID	itemset
T=1000	M, O, N, K, E, Y
T=1001	D, O, N, K E, Y
T=1002	M, A, K, E
T=1003	M, U, C, K y
T=1004	C, O, O, K E

Min Support 60%

Confidence = 80%

=

Support = $\frac{60}{100} \times 5 = 3$

itemset	Supp. count
M	3
O	4
K	5
E	4
Y	3

MO	1
MK	3
ME	2
MY	2
OK	3
O, E	3
O, Y	2
K, E	4
K, Y	3
E, Y	2

→

Now again C_2 in min support

M, K	3
O, K	3
O, E	3
K, Y	3
K, E	4

↓

MOK	1
MKE	2
MKY	2
OKE	3
OK, Y	2

OKE = 3

O∧K = E
O∧E = K
} marked Basket

O∧K=E	3	100
O∧K=K	3	100

→

1 { Bread, Butter, Milk }
2 { Bread, Butter }
3) { Beer, cookies, Diapers }
4) { milk, Diapers, Bread, Butter }
5) { Beer, Diapers }