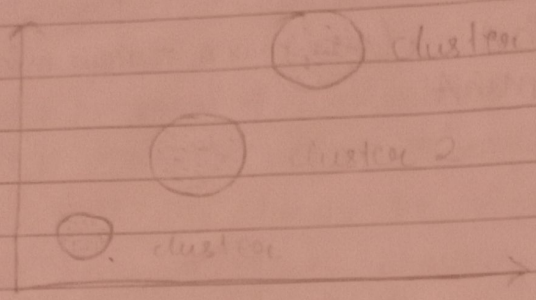
Clustering : It is a method where we find similarity between the objects and similar objects are grouped into one forming a cluster.
Object in one cluster is dissimilar than the object in second cluster.



Types of clustering

classification of clustering algorithm

Hierarchical	Partitional	Categorical	Density-based
(1) Agglomerative	(1) K-means - Euclidean		(1) DBSCAN
(2) Divisive	(2) K-medoid - Manhattan		(2) OPTICS.

K-mean clustering algorithm

Eg: Use K-means algorithm to create two clusters for a given data set as below
dataset { 2, 4, 10, 12, 3, 20, 30, 11, 25 }

Method 1

Step 1: Assume means
$m_1 = 3$, $m_2 = 4$

Step 2: Create the cluster from the means and recalculate the means.

$K = 2$

$K_1 = \{2, 3\}$

$K_2 = \{4, 10, 12, 20, 30, 11, 25\}$

$m_1 = 2.5$

$m_2 = 16$

Step 3: Reassign the clusters and recalculate the means.

$K_1 = \{2, 4, 3\}$

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

$m_1 = 3$

$m_2 = 18$.

Step 4: Reassign the clusters and recalculate the means

$K_1 = \{2, 4, 10, 3\}$

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

$m_1 = 4.75$

$m_2 = 19.6$

Step 5: Reassign the clusters and recalculate the means.

$K_1 = \{2, 4, 10, 12, 3, 11\}$

$K_2 = \{20, 30, 25\}$

$m_1 = 7$

$m_2 = \cancel{21.5}\ 25$

Step 6:

$K_1 = \{2, 4, 10, 12, 3, 11\}$

$K_2 = \{20, 30, 25\}$
$m_1 = 7$
$m_2 = 25$

Step 7: As we have got same means in Step 5 and 6, the given dataset is classified into two clusters as
$K_1 = \{2, 4, 10, 12, 3, 11\}$
$K_2 = \{20, 30, 25\}$
with means as 7 and 25 respectively

Method 2

Step 1:
$K_1 = \{2, 10, 3, 30, 25\}$
$K_2 = \{4, 12, 20, 11\}$
$m_1 = 14$
$m_2 = 11.75$

Step 2: Reassign the clusters and recalculate means.

$K_1 = \{20, 30, 25\}$
$K_2 = \{2, 10, 3, 4, 12, 11\}$
$m_1 = 25$
$m_2 = 7.$

Step 3: Reassign the clusters and recalculate means

$K_1 = \{20, 30, 25\}$
$K_2 = \{2, 10, 3, 4, 12, 11\}$
$m_1 = 25$
$m_2 = 7$

Step 4: As we have got same means in step 2 and 3, the given dataset is classified into two clusters as

$K_1 = \{20, 30, 25\}$

$K_2 = \{2, 10, 3, 4, 12, 11\}$

with means 25 and 7 respectively.

c Use k-means algo to create two clusters for given set of data.

$\{10, 4, 2, 12, 3, 20, 30, 11, 25, 31\}$

→ step 1:

$K_1 = \{10, 2, 3, 30, 25\}$

$K_2 = \{4, 12, 20, 11, 31\}$

$m_1 = 14$

$m_2 = 15.6$

Step 2: Reassign the clusters and recalculate means.

$K_1 = \{10, 4, 2, 12, 3, 11\}$

$K_2 = \{20, 30, 25, 31\}$

$m_1 = 7$

$m_2 = 26.5$

Step 3: Reassign the clusters and recalculate means.

$K_1 = \{10, 4, 2, 12, 3, 11\}$

$K_2 = \{20, 30, 25, 31\}$

$m_1 = 7$

$m_2 = 26.5$

Suppose we have 4 objects and each object has 2 attributes shown in table below. Our goal is to group these objects into two clusters.

object	att(A)	att(B)
med(A)	1	1
med(B)	2	1
med(C)	4	3
med(D)	5	4

$K = 2$

→ Iteration 0

Step 1: Assume the centroids

assume anything from table but both should be from same object

$C_1 = (1,1)$, $C_2 = (2,1)$

Step 2: Calculation of objects distance matrix

calculate distance using Euclidean distance formula

$$D^0 = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 0 & 1 & 3.6 & 5 \\ 1 & 0 & 2.8 & 4.2 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} C_1 = (1,1) \rightarrow \text{group1} \\ C_2 = (2,1) \rightarrow \text{group2} \end{array}$$

Step 3: Grouping of clusters

$$G^0 = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} \text{Group 1 } C_1 = (1,1) \\ \text{Group 2 } C_2 = (2,1) \end{array}$$

While creating groups for cluster, for smallest value we put 1 in group matrix and all other values will be 0 (Have to check vertically).

Iteration 1:

Step 4: Calculation of centroids (4n

$C_1 = (1,1)$

$$C_2 = \left(\left(\frac{2+4+5}{3}\right), \left(\frac{1+3+4}{3}\right)\right) = (3.67, 2.67).$$

In these we calculate no. of 1 in group matrix horizontally. If there is only one 1 then the representation of it is taken as centroid. If there is more than one 1 in the group matrix then the mean of x-cordinate and y-coordinate is taken respectively.

Step 5: Calculation of objects distance matrix.

$$D' = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 0 & 1 & 3.6 & 5.4 \\ 3.05 & 2.76 & 0.56 & 1.97 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{l} \text{Group 1} \quad C_1 = (1,1) \\ \text{Group 2} \quad C_2 = (3.6, 2.6). \end{array}$$

Step 6: Group of clusters

$$G' = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{l} \text{Group 1} \quad C_1 = (1,1) \\ \text{Group 2} \quad C_2 = (3.6, 2.6). \end{array}$$

As G_0 is not same as G', we will be performing iteration 2.

Iteration 2:

Step 7: Calculation of centroids.

$$C_1 = \left(\left(\frac{1+2}{2}\right), \left(\frac{1+1}{2}\right)\right) = (1.5, 1)$$

$$C_2 = \left(\left(\frac{4+5}{2}\right), \left(\frac{3+4}{2}\right)\right) = (4.5, 3.5)$$

Step 8 : Calculation of objects distance matrix.

$$D^2 = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 3.2 & 4.6 \\ 4.3 & 3.5 & 0.7 & 0.7 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{ll} \text{Group 1} & C_1 = (1.5, 1) \\ \text{Group 2} & C_2 = (4.5, 3.5) \end{array}$$

Step 9 : Grouping of clusters.

$$G^2 = \begin{array}{c} \begin{array}{cccc} (1,1) & (2,1) & (4,3) & (5,4) \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array} \begin{array}{ll} \text{Group 1} & C_1 = (1.5, 1) \\ \text{Group 2} & C_2 = (4.5, 3.5). \end{array}$$

As G^2 matches with G^1, we have to stop, and given example is grouped as follows.

Group 1 $K_1 = \{A, B\}$
Group 2 $K_2 = \{C, D\}$

Q.

A_1	2	10
A_2	2	5
A_3	8	4
B_1	5	8
B_2	7	5
B_3	6	4

C_1	1	2
C_2	4	9

Take $k = 3$

→ Iteration 0:

Step 1: Assume the centroids

$C_1 = (2, 10)$ $C_2 = (5, 8)$ $C_3 = (1, 2)$.

Step 2: Calculation of objects distance matrix.

$$D^0 = \begin{array}{c} \begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{array} \\ \begin{bmatrix} 0 & 5 & 8.4 & 3.6 & 7.0 & 7.2 & 8.0 & 2.2 \\ 3.6 & 4.2 & 5 & 0 & 3.6 & 4.1 & 7.2 & 1.4 \\ 8.0 & 3.1 & 7.2 & 7.2 & 6.7 & 5.3 & 0 & 7.6 \end{bmatrix} \\ \begin{array}{cccccccc} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{array} \end{array} \begin{array}{l} C_1 = (2,10) \\ C_2 = (5,8) \\ C_3 = (1,2). \end{array}$$

Step 3: Grouping of clusters

$$G^0 = \begin{array}{c} \begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \\ \begin{array}{cccccccc} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{array} \end{array} \begin{array}{l} C_1 = (2,10) \\ C_2 = (5,8) \\ C_3 = (1,2) \end{array}$$

~~Step~~ Iteration 1:

Step 4: Calculation of centroids.

$C_1 = (2, 10)$

$$C_2 = \left(\left(\frac{8+5+7+6+4}{5} \right), \left(\frac{4+8+5+4+9}{5} \right) \right)$$

$= (6, 6)$.

$$C_3 = \left(\left(\frac{2+1}{2}\right), \left(\frac{5+2}{2}\right)\right)$$

$$= (1.5, 3.5).$$

Step 5: Calculation of object distance matrix.

$$D' = \begin{array}{c} \begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{array} \\ \begin{bmatrix} 0 & 5 & 8.4 & 3.6 & 7.07 & 7.21 & 8.06 & 2.23 \\ 5.6 & 4.1 & 2.8 & 2.2 & 1.4 & 2 & 6.4 & 3.6 \\ 6.5 & 1.5 & 6.5 & 5.7 & 5.7 & 4.5 & 1.5 & \cancel{6.04}\ 6.5 \end{bmatrix} \end{array} \begin{array}{l} C_1 = (2,10) \\ C_2 = (6,6) \\ C_3 = (1.5, 3.5 \end{array}$$

Step 6: Grouping of clusters.

$$G' = \begin{array}{c} \begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{l} C_1 = (2,10) \\ C_2 = (6,6) \\ C_3 = (1.5, 3.5) \end{array}$$

Iteration 2:

Step 7: Calculation of centroids

$$C_1 = \left(\left(\frac{2+4}{2}\right), \left(\frac{10+9}{2}\right)\right)$$

$$= (3, 9.5)$$

$$C_2 = \left(\left(\frac{8+5+7+6}{4}\right), \left(\frac{4+8+5+4}{4}\right)\right)$$

$$= (6.5, 5.25)$$

$$C_3 = (1.5, 3.5).$$

Step 8: Calculation of object distance matrix.

$$D^2 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 1.1 & 4.6 & 7.43 & 2.5 & 6.02 & 6.2 & 7.76 & 1.1 \\ 6.5 & 4.5 & 1.95 & 3.1 & 0.5 & 1.3 & 6.3 & 4.5 \\ 6.5 & 1.5 & 6.5 & 5.7 & 5.7 & 4.5 & 1.5 & 6.02 \end{bmatrix} \\ \begin{matrix} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{matrix} \end{array} \begin{matrix} C_1 = (3, 9.5) \\ C_2 = (6.5, 5.25) \\ C_3 = (1.5, 3.5) \end{matrix}$$

Step 9: Grouping of clusters

$$G^2 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{matrix} C_1 = (3, 9.5) \\ C_2 = (6.5, 5.25) \\ C_3 = (1.5, 3.5) \end{matrix}$$

Iteration 3:

Step 10: Calculation of centroids

$$C_1 = \left(\left(\frac{2+8+5+4}{9} \right), \left(\frac{10+4+8+9}{4} \right) \right)$$

$$= (4.75, 7.75)$$

$$C_2 = \left(\left(\frac{7+6}{2} \right), \left(\frac{5+4}{2} \right) \right)$$

$$= (6.5, 4.5)$$

$$C_3 = (1.5, 3.5)$$

Step 11: Calculation of object distance matrix.

$$D^3 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 9.5 & 3.8 & 24.6 & 0.3 & 3.5 & 3.9 & 6.8 & 1.4 \\ 7.1 & 4.5 & 1.5 & 3.8 & 0.7 & 0.7 & 6.4 & 5.4 \\ 6.5 & 1.5 & 6.5 & 5.7 & 5.7 & 4.5 & 1.5 & \cancel{36.5}\ 8.04 \end{bmatrix} \\ \begin{matrix} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{matrix} \end{array} \quad \begin{matrix} C_1 = (4.75, \\ C_2 = (6.5, \\ C_3 = (1.5, 3 \end{matrix}$$

Step 12: Grouping of clusters.

$$G^3 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \\ \begin{matrix} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{matrix} \end{array} \quad \begin{matrix} C_1 = (4.75, \\ C_2 = (6.5, 4 \\ C_3 = (1.5, 3 \end{matrix}$$

As G_3 is same as G_2 we have to stop and given example is grouped as follows

Group 1 = $\{A_1, B_1, C_2\}$
Group 2 = $\{A_3, B_2, B_3\}$
Group 3 = $\{A_2, C_1\}$

K-medoid clustering algorithm

In K-medoid, we use centrally located point (reference point) known as medoid. It is also known as PAM (Partioning Around medoids)
One dimensional.

Q. Given the dataset D as
$D = \{1, 2, 6, 7, 8, 10, 15, 17, 20\}$
Apply K-medoid algorithm to form three clusters

K-medoid algo:

(1) Select initial K-medoids.
(2) Assign all points to closest medoid.
(3) See if any other point can act as better medoid. Calculate total swapping cost.
(4) if (swapping_cost > 0)
    stop
else
    repeat step 3 and step 4.

→ Step 1: Calculation of Adjacency matrix.
(optional step)

0	1	2	6	7	8	10	15	17	20
1	0	1	5	6	7	9	14	16	19
2	1	0	4	5	6	8	13	15	18
6	5	4	0	1	2	4	9	11	14
7	6	5	1	0	1	3	8	10	13
8	7	6	2	1	0	2	7	9	12
10	9	8	4	3	2	0	5	7	10
15	14	13	9	8	7	5	0	2	5
17	16	15	11	10	9	7	2	0	3
20	19	18	14	13	12	10	5	3	0

Step 2: Assumption of initial medoids
$K_1 = 6 : 1, 2$
$K_2 = 7 :$
$K_3 = 8 : 10, 15, 17, 20.$

Step 3: Check whether 8 can be replaced by 15 or not?
$K_1 = 6 : 1, 2.$

$K_2 = 7 : 8, 10$
$K_3 = 15 : 17, 20$

Step 4 : Calculation of TSC
Total swapping cost = cost(K_1) + cost(K_2) + cost(K_3)

cost of (K_1) = cost(1) + cost(2)

cost(1) = distance of an element from its new medoid − distance of element to its old medoid
= 5 − 5 = 0.

cost(2) = 4 − 4 = 0.
cost(K_1) = 0 + 0 = 0.
cost (K_2) = cost(8) + cost(10)

cost(8) = 1 − 0 = 1

cost(10) = 3 − 2 = 1

cost(K_2) = 1 + 1 = 2.

cost(K_3) = cost(17) + cost(20)

cost(17) = 2 − 9 = −7

cost(20) = 5 − 12 = −7

cost(K_3) = −7 + (−7) = −14.

Page No.:
Date:

$TSC = cost(K_1) + cost(K_2) + cost(K_3)$

$= [(5-5)+(4-4)] + [(1-0)+(3-2)] + [(2-9)+(5-12)]$

$= -12.$

Step 5 : As $TSC \not> 0$
8 can be replaced by 15
$\therefore$ $K_1 = 6 : 1, 2$
$K_2 = 7 : 8, 10$
$K_3 = 15 : 17, 20.$

Step 6 : Now check whether 7 can be replaced by 10 ~~nor~~ not?

$K_1 = 6 : 1, 2, 7$
$K_2 = 10 : 8$
$K_3 = 15 : 17, 20.$

Step 7 : Calculation of TSC

$TSC = cost(K_1) + cost(K_2) + cost(K_3)$
$= [(5-5)+(4-4)+(1-0)] + [(2-1)] + [(2-2)+(5-5)]$

$= (1 + 1 + 0)$
$= 2.$

Step 8 : As $TSC > 0$
We have to stop
7 cannot be replaced by 10
So we get final answers as
$K_1 = 6 : 1, 2$
$K_2 = 7 : 8, 10$
$K_3 = 15 : 17, 20.$

Q. $D = \{2, 4, 10, 12, 3, 20, 30, 11, 25\}$
$K = 3$.

→ Step 1 : Assumption of initial medoids

$K_1 = 10 : 2, 3, 4$
$K_2 = 12 : 11$
$K_3 = 25 : 20, 30$

Step 2: Check whether 10 can be replaced by 4 or not?

$K_1 = 4 : 2, 3$
$K_2 = 12 : 10, 11$
$K_3 = 25 : 20, 30$

Step 3 : Calculation of TSC

$TSC = cost(K_1) + cost(K_2) + cost(K_3)$
$= [(2-8)+(1-7)] + [(2-0)+(1-1)] + [(5-5)+(5-5)]$
$= -6 - 6 + 2$
$= -10$.

Step 4: As $TSC \ngtr 0$
10 can be replaced by 4
$\therefore K_1 = 4 : 2, 3$
$K_2 = 12 : 10, 11$
$K_3 = 25 : 20, 30$

Step 5 : Now check whether 30 can be replaced by 25?

$K_1 = 4 : 2, 3$
$K_2 = 12 : 10, 11, 20$.
$K_3 = 30 : 25$

step 6: calculation of TSC

$$TSC = cost(K_1) + cost(K_2) + cost(K_3)$$
$$= [(2-2)+(1-1)] + [(2-2)+(1-1)+(8-5)] + [(5-0)]$$
$$= 0+0+0+0+3+5$$
$$= 8.$$

Step 7: As TSC > 0.
we have to stop
30 can be replaced by 25
So we get final answers as
$K_1 = 4$: 2, 3
$K_2 = 12$: 10, 11, 20
$K_3 = 30$: 25

For multidimensional

Q. Coordinates of 10 objects are given below. Apply k-medoid algorithm where no. of clusters are 2.

Objects	x	y
1	1	4
2	5	1
3	5	2
4	5	4
5	10	4
6	25	4
7	25	6
8	25	7
9	25	8
10	29	7

→ Step 1: Assume initial medoids
Here objects 1 and 5 are selected as initial medoids.

Step 2:

objects	Dissimilarity from obj 1	Dissimilarity from obj 5	minimal Dissimilarity	closest representative object.
1	0	9	0	1
2	3	8	7	1
3	6	7	6	1
4	4	5	4	1
5	9	0	0	5
6	24	15	15	5
7	26	17	17	5
8	27	18	18	5
9	28	19	19	5
10	31	22	22	5
		Avg cost	10.8	

(1,4)
(5,1)
(5,2)
(5,4)
(10,4)
(25,4)
(25,6)
(25,7)
(25,8)
(29,7)

Step 3: Assume medoids.
Here object 4 and object 8 are selected as medoids

Step 4:

objects	Dissimarility from obj 4 (5,4)	Dissimilarity from obj 8 (25,7)	minimal dissimilarity	closest representative obj
1	4	27	4	4
2	3	26	3	4
3	2	25	2	4
4	0	23	0	4
5	5	18	5	4
6	20	3	3	8
7	22	1	1	8
8	23	0	0	8
9	24	1	1	8
10	27	4	4	8

~~Avg~~ Avg cost = 2.3

Avg cost = Avg of minimal dissimilarity

TSC = new cost − old cost
= ~~10.8~~ 2.3 − 10.8
= −8.5

As TSC is less than 0, we have got new medoids as 4 and 8.

Step 5: Now we are choosing next medoids as 5 and 8.

Step 6:

Object	Dissimilarity from obj 5 (10,4)	Dissimilarity from obj 8 (25,7)	minimal dissimilarity	Closest representative object
1	9	27	9	5
2	8	26	8	5
3	7	25	7	5
4	5	23	5	5
5	0	18	0	5
6	15	3	3	8
7	17	1	1	8
8	18	0	0	8
9	19	1	1	8
10	22	4	4	8

Avg cost = 3.8

TSC = new cost − old cost
= 3.8 − 2.3
= 1.5

Step 7: As total swapping cost > 0
we have to stop.
∴ Final medoids will be 4 and 8.
$K_1 = 4$: 1, 2, 3, 5
$K_2 = 8$: 6, 7, 9, 10

Q.

1	8	7
2	3	7
3	4	9
4	4	9
5	8	5
6	5	8
7	7	3
8	8	4
9	7	5
10	4	5

K = 2

→ Step 1: Assume initial medoids
Here objects 3 and 7 are selected as initial medoi

step 2:

	objects	Dissimilarity from obj 3 (4, 9)	Dissimilarity from obj 7 (7, 3)	minimal dissimilarity	closest repr object
1	(8, 7)	6	5	5	7
2	(3, 7)	3	8	3	3
3	(4, 9)	0	9	0	3
4	(4, 9)	0	9	0	3
5	(8, 5)	8	3	3	7
6	(5, 8)	2	7	2	3
7	(7, 3)	9	0	0	7
8	(8, 4)	9	2	2	7
9	(7, 5)	7	2	2	7
10	(4, 5)	4	5	4	3
			Avg cost	21	

Step 3: Assume medoids

Here object 5 and 10 are selected as medoids.

object	dissimilarity from obj 5 (8,5)	dissimilarity from obj 10 (4,5)	minimal dissimilarity	closest representative object
(8,7)	2	6	2	5
(3,7)	7	3	3	10
(4,9)	8	4	4	10
(4,9)	8	4	4	10
(8,5)	0	4	0	5
(5,8)	6	4	4	10
(7,3)	3	5	3	5
(8,4)	1	5	1	5
(7,5)	1	3	1	5
(4,5)	4	0	0	10

Avg cost = 22

TSC = new cost - old cost

= 22 - 21

= 1

Step 5:

As TSC ≥ 0

we have to stop

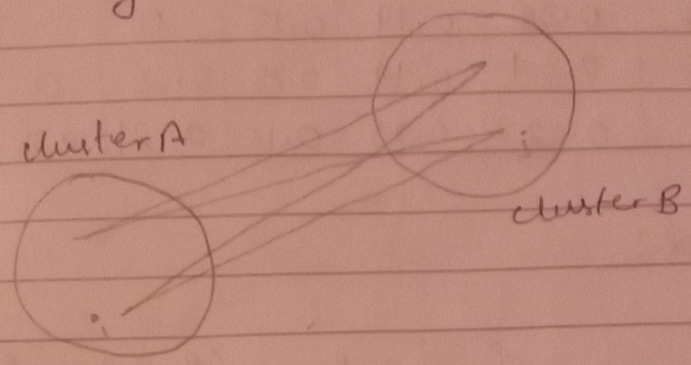
∴ Final medoids will be 3 and 7.

K_1 = 3 : 2, 4, 6, 10

K_2 = 7 : 1, 5, 8, 9

member of one cluster to any member of another cluster

(iii) Average linkage



$$d(A, B) = \text{mean}\{d(i, j)\}$$

- It is also known as mean method.
- It finds out the mean between any member of one cluster to the any member of another cluster

Q. Assume data base D is given as follows. Apply single link, complete link and average link technique to find clusters in D. Use Euclidean distance measure.

D =

	X	Y
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.41
P_6	0.45	0.30

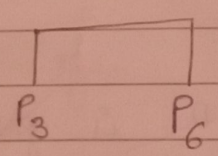
single linkage

→ Step 1: Find distance matrix

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.23				
P_2	0.23	0				
P_3	0.21	0.14	0			
P_4	0.36	0.19	0.15	0		
P_5	0.34	0.14	0.28	0.28	0	
P_6	0.23	0.25	0.11	0.21	0.38	0

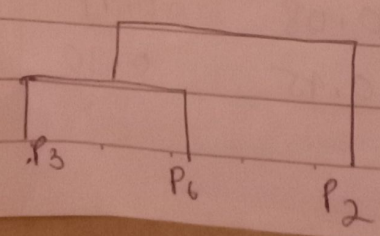
Step 2:

	P_1	P_2	(P_3, P_6)	P_4	P_5
P_1	0				
P_2	0.23	0			
$(P_3 P_6)$	0.21	0.14	0		
P_4	0.34	0.19	0.15	0	
P_5	0.23	0.14	0.28	0.28	0



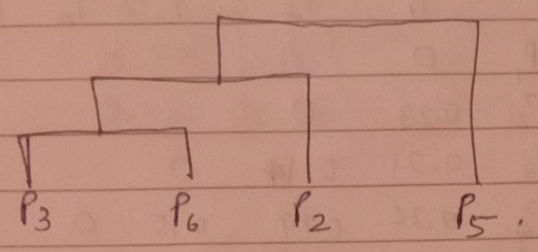
Step 3:

	P_1	$\{P_2, (P_3 P_6)\}$	P_4	P_5
P_1	0			
$\{P_2 (P_3 P_6)\}$	0.21	0		
P_4	0.36	0.15	0	
P_5	0.34	0.14	0.28	0



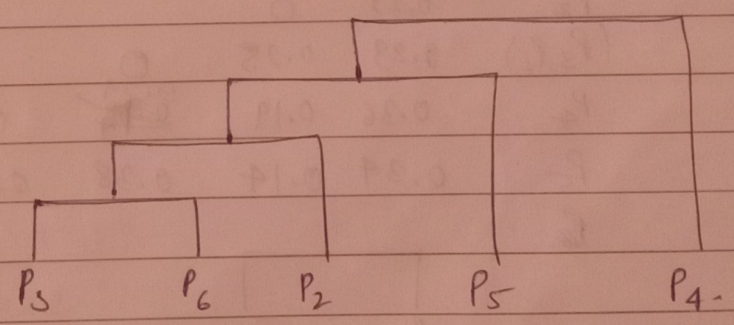
Step 4:

	P_1	$[P_5\{P_2(P_3, P_6)\}]$	P_4.
P_1	0		
$[P_5\{P_2(P_3, P_6)\}]$	0.21	0	
P_4	0.36	0.15	0



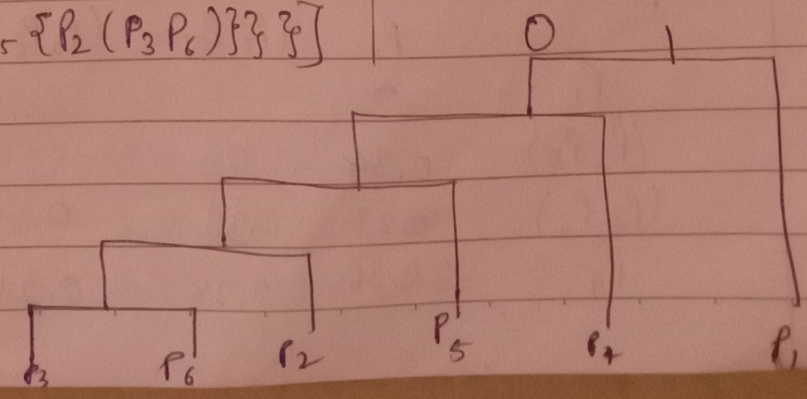
Step 5:

	P_1	$[P_4[P_5\{P_2(P_3, P_6)\}]]$
P_1	0	
$[P_4[P_5\{P_2(P_3, P_6)\}]]$	0.21	0



Step 6

	$[P_1\{P_4\{P_5\{P_2(P_3 P_6)\}\}\}]$
$[P_1\{P_4\{P_5\{P_2(P_3 P_6)\}\}\}]$	0



Note: while merging 2 clusters ~~minimum distance (for all 3 linkages)~~ & while placing new values in merge matrix see which algo. is to be used

Complete linkage

Step 1: Find distance matrix.

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.21	0.14	0			
P_4	0.36	0.19	0.15	0		
P_5	0.34	0.14	0.28	0.28	0	
P_6	0.23	0.25	0.11	0.21	0.38	0

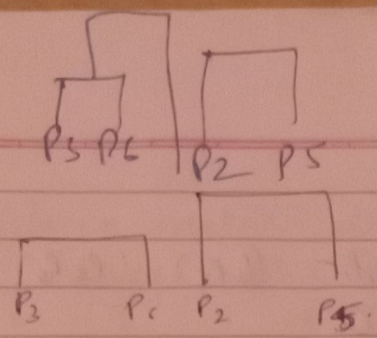
Step 2:

	P_1	P_2	$(P_3 P_6)$	P_4	P_5
P_1	0				
P_2	0.23	0			
(P_3, P_6)	0.23	0.25	0		
P_4	0.36	0.19	0.21	0	
P_5	0.34	0.14	0.38	0.28	0

P_3 P_6

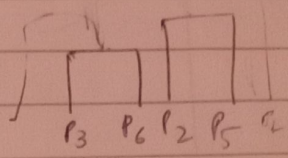
Step 3:

	P_1	$(P_2 P_5)$	$(P_3 P_6)$	P_4
P_1	0			
$(P_2 P_5)$	0.34	0		
$(P_3 P_6)$	0.23	0.38	0	
P_4	0.36	0.28	0.21	0



Step 4:

	P_1	$(P_2\ P_5)$	$[P_4(P_3\ P_6)]$
P_1	0		
$(P_2\ P_5)$	0.34	0	
$[P_4(P_3\ P_6)]$	0.36	0.38	0



Step 5:

	$[P_1(P_2\ P_5)]$	$[P_4(P_3\ P_6)]$
$[P_1(P_2\ P_5)]$	0	
$[P_4(P_3\ P_6)]$	0.38	0

Step 6:

	$\{P_1(P_2\ P_5)[P_4(P_3\ P_6)]\}$
$\{P_1(P_2\ P_5)[P_4(P_3\ P_6)]\}$	0

Average linkage:

Step 2:

	P_1	P_2	$(P_3\ P_6)$	P_4	P_5
P_1	0				
P_2	0.23	0			
(P_3, P_6)	0.22	0.19	0		
P_4	0.36	0.19	0.18	0	
P_5	0.34	0.14	0.33	0.28	0

Step 3:

	P_1	$(P_2 P_5)$	$(P_3 P_6)$	P_4
P_1	0			
$(P_2 P_5)$	0.28	0		
$(P_3 P_6)$	0.22	0.26	0	
P_4	0.36	0.23	0.18	0

P_3 P_6 P_2 P_5

Step 4:

	P_1	$(P_2 P_5)$	$[P_4(P_3 P_6)]$
P_1	0		
$(P_2 P_5)$	0.28	0	
$[P_4(P_3 P_6)]$	0.29	0.24	0

Step 5:

	P_1	$[P_2 P_5 [P_4 (P_3 P_6)]]$
P_1	0	
$[P_2 P_5 [P_4 (P_3 P_6)]]$	0.28	0

Step 6:

	$\{P_1 \{P_2 P_5 [P_4 (P_3 P_6)]\}\}$
$\{P_1 \{P_2 P_5 [P_4 (P_3 P_6)]\}\}$	0

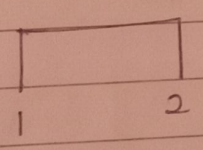
Q. For a given distance matrix, draw a single link, complete link, average link dendogram

	1	2	3	4	5
1	0				
2	2	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

→ Single link

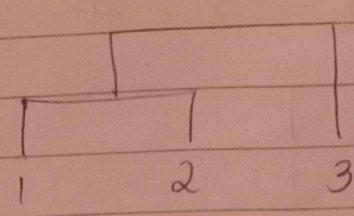
Step 1:

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



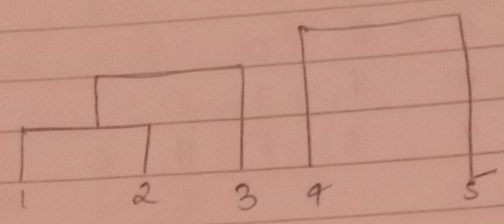
Step 2:

	[3(1,2)]	4	5
[3(1,2)]	0		
4	7	0	
5	5	4	0



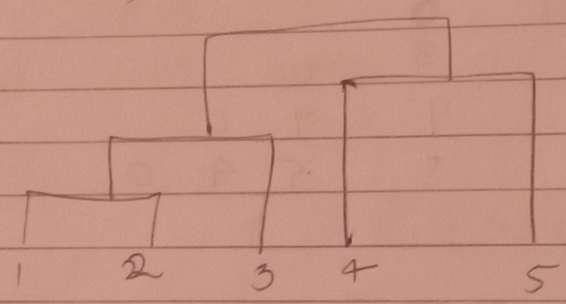
step 3:

	[3(1,2)]	(4,5)
[3(1,2)]	0	
(4,5)	5	0



step 4:

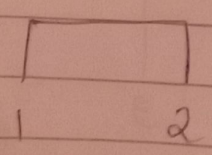
	{(4,5),[3(1,2)]}
{(4,5),[3(1,2)]}	0



Complete link

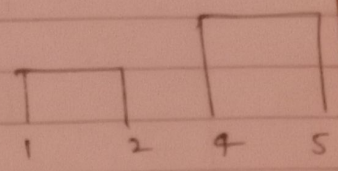
step 1:

	(1,2)	3	4	5
(1,2)	0			
3	6	0		
4	10	7	0	
5	9	5	4	0



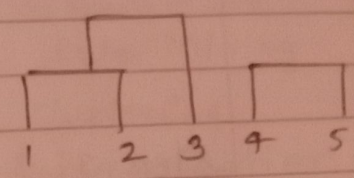
Step 2:

	(1,2)	3	(4,5)
(1,2)	0		
3	6	0	
(4,5)	10	7	0



Step 3:

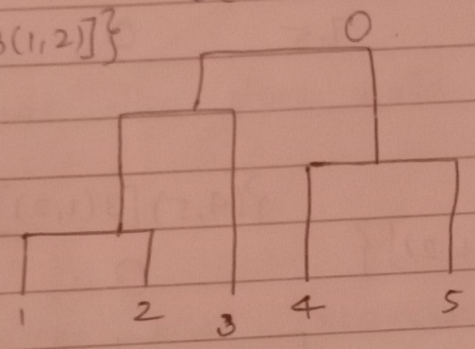
	[3(1,2)]	(4,5)
[3(1,2)]	0	
(4,5)	10	0



Step 4:

{(4,5)[3(1,2)]}

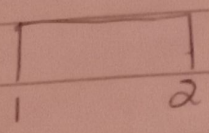
{(4,5)[3(1,2)]} 0



Average link.

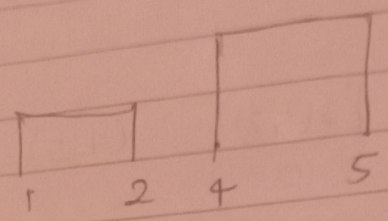
Step 1:

	(1,2)	3	4	5
(1,2)	0			
3	4.5	0		
4	9.5	7	0	
5	8.5	5	4	0



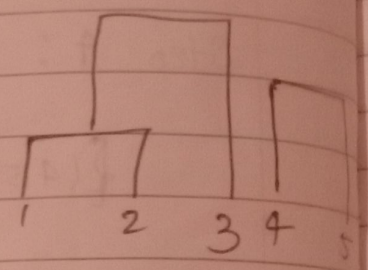
Step 2:

	(1,2)	3	(4,5)
(1,2)	0		
3	4.5	0	
(4,5)	9	6	0



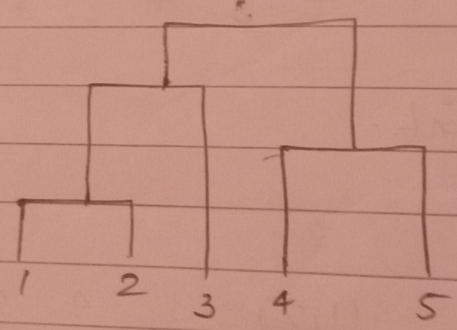
Step 3:

	[3(1,2)]	(4,5)
[3,(1,2)]	0	
(4,5)	7.5	0



Step 4:

	{(4,5)[3(1,2)]}
{(4,5)[3(1,2)]}	0



classification involve 2 stage ie Training stage and Testing stage

Training stage
- It generates the rules by considering/applying classification algorithm on training data (past records).

Testing
- In testing phase, rules are used to predict new incoming data.

IMP Naive ~~based~~ Baye's classifier

eg:

Car no.	colour	Type	Origin	Stolen
1	Red	sports	Domestic	Yes
2	Red	sports	Domestic	No
3	Red	sports	Domestic	Yes
4	Yellow	sports	Domestic	No
5	Yellow	sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	sports	Imported	Yes

Apply Naive Baye's classifier algo to classify unseen sample (x) = ⟨red, SUV, Domestic⟩

→ Step 1: Find the probability of final class (car stolen)

$P(Yes) = \frac{5}{10} = 0.5$ $P(No) = \frac{5}{10} = 0.5$

Step 2: Find the probability of color attribute

$P(red / Yes) = 3/5$

$P(red / No) = 2/5$

Step 3: Find the probability of type attribute

$P(SUV / Yes) = 1/5$

$P(SUV / No) = 3/5$

Step 4: Find the probability of origin attribute

$P(Domestic / Yes) = 2/5$

$P(Domestic / No) = 3/5$

Step 5: Calculate probability of unseen sample

$P(X / Yes) \cdot P(Yes) = P(Red / Yes) \cdot P(SUV / Yes) \cdot P(Domestic / Yes) \cdot P(Yes)$

$= \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{10} = 0.024$

$P(X / No) \cdot P(No) = P(Red / No) \cdot P(SUV / No) \cdot P(Domestic / No) \cdot P(No)$

$= \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{5}{10} = 0.072$

As Probability of No is greater than probability of Yes i.e $0.072 > 0.024$,

$\therefore$ Given unseen sample is classified as No.

car sto

Q.

Name	Hair	Height	weight	Dublin	Result
	Blonde	avg	light	No	Sunburned
	Blonde	tall	avg	Yes	None
	Brown	Short	avg	Yes	None
	Blonde	Short	avg	No	Sunburned
	Red	avg	Heavy	No	Sunburned
	Brown	tall	Heavy	No	None
	Brown	avg	Heavy	No	None
	Brown	short	Light	Yes	None

$X = \langle$ Brown, tall, avg, No $\rangle$

→ Step 1: Find the probability of final class (Result)

$P(\text{Sunburned}) = \frac{3}{8}$ $\quad$ $P(\text{None}) = \frac{5}{8}$

Step 2: Find the probability of hair attribute

$P(\text{Sunburned}) = 0/3$

$P(\text{Brown} \mid \text{None}) = 4/5$

Step 3: Find the probability of height attribute.

$P(\text{tall} \mid \text{Sunburned}) = 0/3$

$P(\text{tall} \mid \text{None}) = 2/5$

Step 4: Find the probability of weight attribute

$P(\text{avg} \mid \text{sunburned}) = 1/3$

$P(\text{avg} \mid \text{None}) = 2/5$

Step 5: Find the probability of dublin attribute

$P(\text{No} \mid \text{Sunburned}) = 3/3$

$P(\text{No} \mid \text{None}) = 2/5$

Step 6: Calculate Probability of UnSeen Sample

$P(x/\text{Sunburned}) \cdot P(\text{Sunburned}) = P(\text{Brown}/\text{Sunburned}) \cdot P(\text{tall}/\text{Sunburned}) \cdot P(\text{avg}/\text{sunburned}) \cdot P(\text{No}/\text{Sunburned}) \cdot P$

$= \frac{0}{3} \times \frac{0}{3} \cdot \frac{1}{3} \cdot \frac{3}{3} \cdot \frac{3}{8}$

$= 0$

$P(x/\text{None})\, P(\text{None}) = P(\text{Brown}/\text{None}) \cdot P(\text{tall}/\text{None}) \cdot P(\text{avg}/\text{None}) \cdot P(\text{No}/\text{None}) \cdot P(\text{No}$

$= \frac{4}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{5}{8}$

$= 0.032$

Since the probability of None is greater than Sunburned i.e $0 < 0.032$.

$\therefore$ Given unseen sample is classified as None

Q.

Trans id	Income	Credit	Decision
1	V. High	Excellent	Authorize
2	High	Good	Authorize
3	medium	Excellent	Authorize
4	High	Good	Authorize
5	V. High	Good	Authorize
6	Medium	Excellent	Authorize
7	High	Bad	reg. ID
8	Medium	Bad	reg ID
9	High	Bad	reject
10	Low	Bad	call police

$X = \langle 15, \text{medium}, \text{good} \rangle$

Step 1: Find the probability of final class (Decision)

$P(\text{Authorize}) = \frac{6}{10}$, $P(\text{reg. ID}) = \frac{2}{10}$, $P(\text{reject}) = \frac{1}{10}$

$P(\text{call police}) = \frac{1}{10}$

Step 2: Find probability of income attribute.

$P(\text{medium} / \text{authorize}) = \frac{2}{6}$

$P(\text{medium} / \text{reg. ID}) = \frac{1}{2}$

$P(\text{medium} / \text{reject}) = \frac{0}{1}$

$P(\text{medium} / \text{call police}) = \frac{0}{1}$

Step 3: Find probability of credit attribute

$P(\text{good} / \text{authorize}) = \frac{3}{6}$

$P(\text{good} / \text{reg. ID}) = \frac{0}{2}$

$P(\text{good} / \text{reject}) = \frac{0}{1}$

$P(\text{good} / \text{call police}) = \frac{0}{1}$

Step 4: Calculate probability of unseen sample

$P(x/\text{authorize}) \cdot P(\text{authorize}) = \frac{2}{6} \cdot \frac{3}{6} \cdot \frac{6}{10} = \frac{1}{10}$

$P(x/\text{reg. ID}) \cdot P(\text{reg. ID}) = \frac{1}{2} \cdot \frac{0}{2} \cdot \frac{2}{10} = 0$

$P(x/\text{reject})\, P(\text{reject}) = \frac{0}{1} \cdot \frac{0}{1} \cdot \frac{1}{10} = 0$

$P(x/\text{call police})\, P(\text{call police}) = \frac{0}{1} \cdot \frac{0}{1} \cdot \frac{1}{10} = 0.$

As Probability of Authorize is greater than probability of reg. ID, reject, call police i.e
$0.1 > 0$
∴ Given unseen sample is classified as Authorize

Q

ID	name	Gender	Height	Class
1	A	F	1.6 m	short
2	B	M	2 m	tall
3	C	F	1.9 m	medium
4	D	M	2.1 m	tall
5	G	F	1.7 m	short
6	F	M	1.85 m	medium
7	q	F	1.6 m	short
8	H	M	1.7 m	short
9	I	M	2.2 m	tall

$X = \langle \text{Adam}, M, 1.95\,m \rangle$

→ Step 1: Find probability for final class (class)

$P(\text{Short}) = \frac{4}{9}$ $P(\text{Tall}) = \frac{3}{9}$ $P(\text{Medium}) = \frac{2}{9}$

Step 2:
As height attribute is in integer datatype, it must be represented in range
∴ I have assumed range as
(0.0 - 1.60) (1.61 - 1.70) (1.71 - 1.80) (1.81 - 1.90)
(1.91 - 2.0) (2.01 - 2.1) (2.1 - infinity)

Step 3: Find probability of Gender attribute

$P(M / \text{Short}) = \frac{1}{4}$

$P(M / \text{tall}) = \frac{3}{3}$

$P(M / \text{medium}) = \frac{1}{2}$

Step 4: Find probability of height attribute

$P(1.91 - 2.0 / \text{short}) = \frac{0}{4}$

$P(1.91 - 2.0 / \text{tall}) = \frac{1}{3}$

$P(1.91 - 2.0 / \text{medium}) = \frac{0}{2}$

Step 5: Calculate Probability for unseen sample

$P(x / \text{short}) \cdot P(\text{short}) = P(M / \text{short}) \cdot P(1.91 - 2.0 / \text{short}) P(\text{Short})$

$= \frac{1}{4} \cdot \frac{0}{4} \cdot \frac{4}{9} = 0$

$P(x|tall)\ P(tall) = P(M|tall) \cdot P(1.91-2.0|tall) \cdot P(tall)$

$= \frac{3}{3} \cdot \frac{1}{3} \cdot \frac{3}{9} = \frac{1}{9}$

$P(x|medium) \cdot P(medium) = P(M|medium) \cdot P(1.91-2.0|medium)\ P(medium)$

$= \frac{1}{2} \cdot \frac{0}{2} \cdot \frac{2}{9} = 0$

As probability of tall is greater than probability of short and medium i.e $1/9 > 0$

∴ Given unseen sample is classified as tall.

Sampling

Population is the collection of elements having some or other features common. The subset of population is known as sample.

The process of selecting samples is known as sampling.

Two methods for sampling are

1. Sampling with replacement
2. Sampling without replacement

Sampling with replacement

In this, data can be picked up again and again whereas in sampling without replacement, data will not be repeated again.

Ensemble models

(1) Bootstrap

Apriori Algorithm

Q. Consider the database given below. Apply Apriori Algorithm with minimum support count as 2 and generate association rules with its confidence.

Tid	List of items
T_1	1, 2, 5
T_2	2, 4
T_3	2, 3
T_4	1, 2, 4
T_5	1, 3
T_6	2, 3
T_7	1, 3
T_8	1, 2, 3, 5
T_9	1, 2, 3

→ Step 1: Scan the database D and find the support count. The candidate list is {1, 2, 3, 4, 5}

C_1 =

Item set	support count
1	6
2	7
3	6
4	2
5	2

Step 2: Compare candidate support count with minimum support count i.e 2 and discard all items having support < 2

↳ pruning method

L_1 =

Item set	support count
1	6
2	7
3	6
4	2
5	2

Step 3 : Generate C_2 from L_1
↳ Joining method

C_2 =

Item set	Support count
1, 2	4
1, 3	4
1, 4	1
1, 5	2
2, 3	4
2, 4	2
2, 5	2
3, 4	0
3, 5	1
4, 5	0

Step 4 : Compare candidate support count with minimum support count i.e 2 & discard all items having support < 2

L_2 =

Item set	support count
1, 2	4
1, 3	4
1, 5	2
2, 3	4
2, 4	2
2, 5	2

Step 5: Generate C_3 from L_2

C_3 = Item set	Support count
1, 2, 3	2
1, 2, 5	2
1, 3, 5	1
2, 3, 4	0
2, 3, 5	1
2, 4, 5	0

Step 6:

L_3 = Item set	Support count
1, 2, 3	2
1, 2, 5	2

Step 7: Find association rules
Assuming confidence as 75%

Association rule	Support	Confidence	% confidence
$1 \Rightarrow 2 \wedge 3$	2	2/6	33.3%
$2 \Rightarrow 1 \wedge 3$	2	2/7	28.5%
$3 \Rightarrow 1 \wedge 2$	2	2/6	33.3%
$1, 2 \Rightarrow 3$	2	2/4	50%
$1, 3 \Rightarrow 2$	2	2/4	50%
$2, 3 \Rightarrow 1$	2	2/4	50%
$1 \Rightarrow 2 \wedge 5$	2	2/6	33.3%
$2 \Rightarrow 1 \wedge 5$	2	2/7	28.5%
$5 \Rightarrow 1 \wedge 2$	2	2/2	100%

$1,2 \Rightarrow 5$	2	2/4	50%
$1,5 \Rightarrow 2$	2	2/2	100%
$2,5 \Rightarrow 1$	2	2/2	100%

As minimum threshold for confidence 75%, following rules will be selected as final answer

$5 \Rightarrow 1,2$, $1,5 \Rightarrow 2$, $2,5 \Rightarrow 1$

Q. Find frequent item sets for following transactions with minimum support has 50% and confidence as 50%

Tid	Items
T_1	a, b, c
T_2	a, c
T_3	a, d
T_4	b, e, f

→ Step 1: As support is 50%, it is recalculated as support count $= \frac{50}{100} \times 4$

$= 2$

Step 2: Scan database D and find the support count

candidate list is {a, b, c, d, e, f}

$C_1 =$

Item set	support count
a	3
b	2
c	2
d	2

e	1
f	1

Step 3: Compare candidate support count with minimum support count i.e 2 and discard all items having support < 2

L_1 =

Item set	Support count
a	3
b	2
c	2

Step 4: Generate C_2 from L_1

C_2 =

Item set	Support count
ab	1
ac	2
bc	1

Step 5: Compare candidate support count with minimum support count i.e 2 and discard all items having support < 2

L_2 =

Item set	Support count
ac	2

Step 6: Find association rules

Association rule	Support	confidence	% confidence
$a \Rightarrow c$	2	2/3	66.6%
$c \Rightarrow a$	2	2/2	100%

As minimum threshold for confidence is 50%, following rules will be selected as final answer

$a \Rightarrow c$, $c \Rightarrow a$

FP Tree

Tid	Items bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

min support = 3

→ Step 1: Find min support for each item

C_1 =

Item set	support
a	3
b	3
c	4
d	1
e	1
f	4
g	1
h	1
i	1
j	1
k	1
l	2
m	3
n	1
o	2
p	3

Step 2: Compare candidate support count with minimum support count i.e 2 and discard all items having support < 2

L_1 = Item	min support
a	3
b	3
c	4
f	4
m	3
p	3

Step 3: Arrange frequent item sets according to their descending order

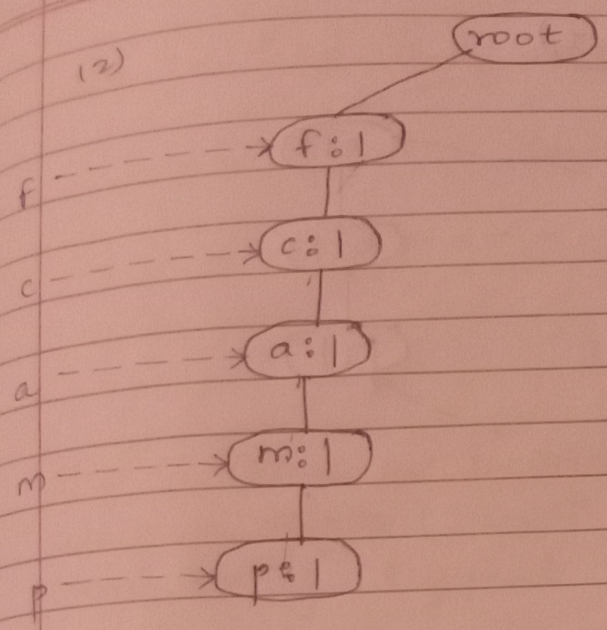
Tid	Items bought	Order for itemset
1	f, a, c, d, g, i, m, p	f, c, a, m, p
2	a, b, c, f, l, m, o	f, c, a, b, m
3	b, f, h, j, o	f, b
4	b, c, k, s, p	c, b, p
5	a, f, c, e, l, m, n	f, c, a, m, f

Step 4: Creation of FP tree

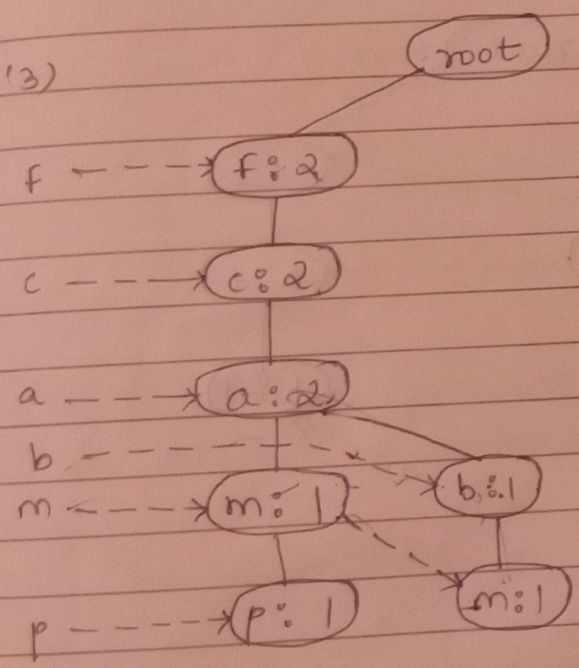
(1) Create root.

(root)

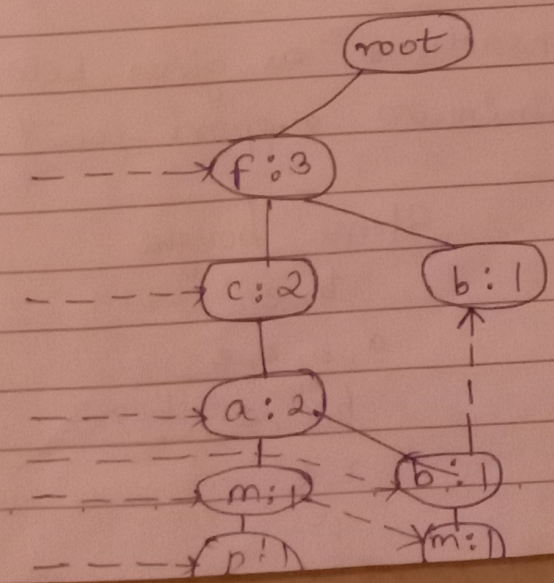
(2)

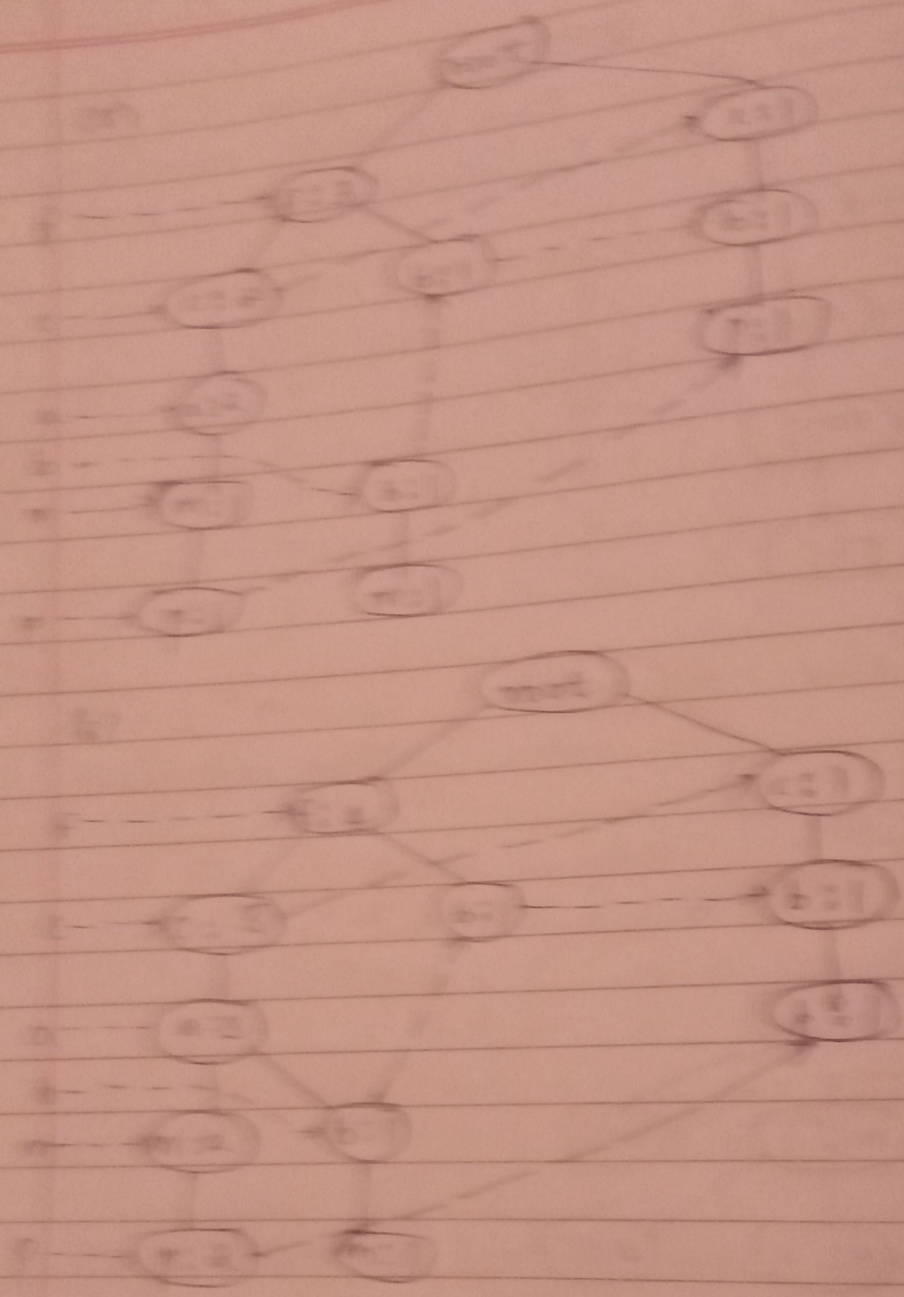

root
f
f:1
c
c:1
a
a:1
m
m:1
p
p:1

(3)


root
f
f:2
c
c:2
a
a:2
b
b:1
m
m:1
p
p:1
m:1

(4)


root
f:3
c:2
b:1
a:2
m:1
b:1
p:1
m:1



6. Consider the transactions as given below and show it was being minimum support count as 2.

Tid	Items bought
1	b, e
2	a, b, c, e
3	b, d, e
4	a, b, d
5	A

Step 1: Find min support for each item

C_1 = Items	Support
a	3
b	3
c	3
e	3

Step 2: Compare candidate support count with minimum support count i.e 2 and discard all items having support < 2.

L_1 = Items	support
b	3
e	3
a	3
c	3

Step 3: Arrange frequent item sets according to their descending order.

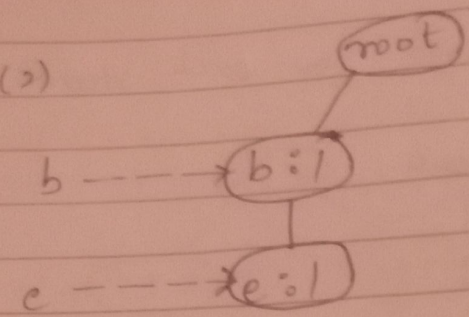
Tid	order for itemset
1	b, e
2	b, e, a, c
3	b, e, c
4	a, c
5	a

Step 4: Creation of FP tree.

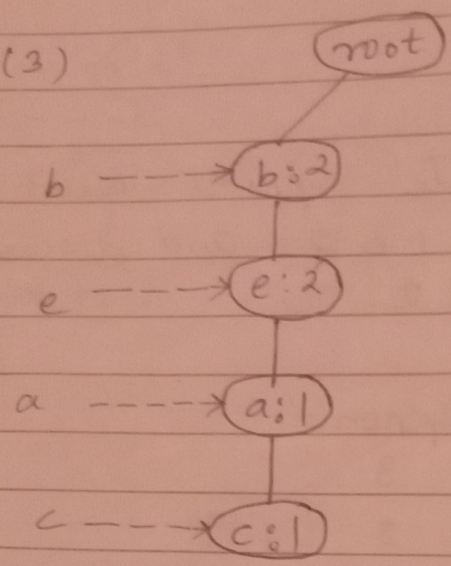
(1) Creation of root node

root

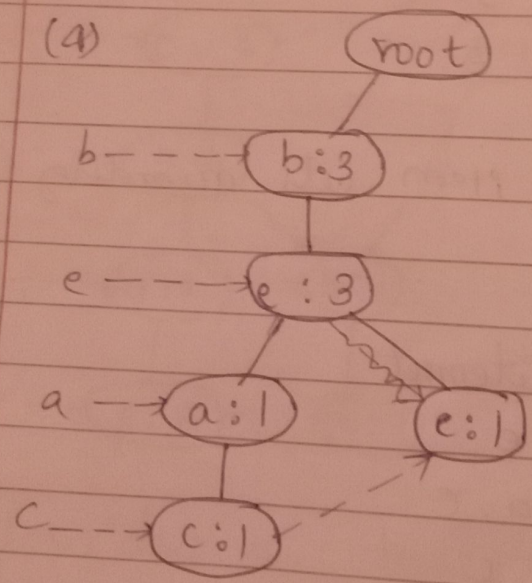
(2)



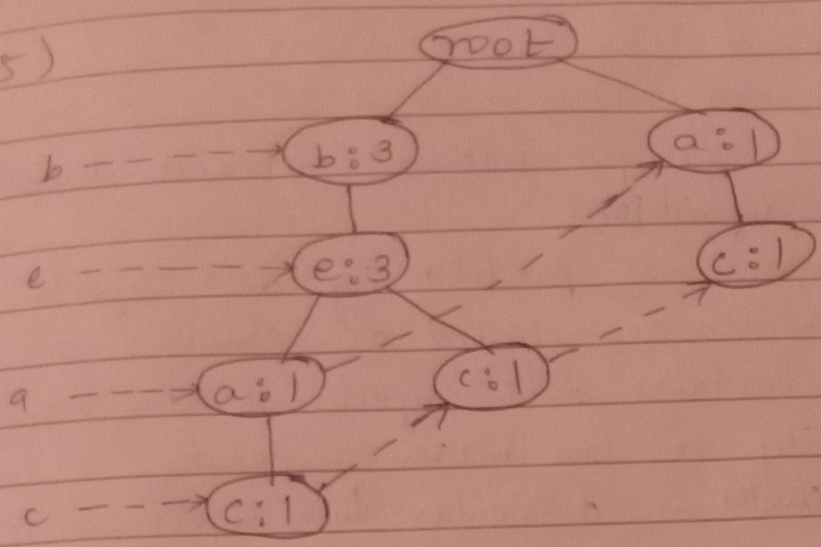
(3)



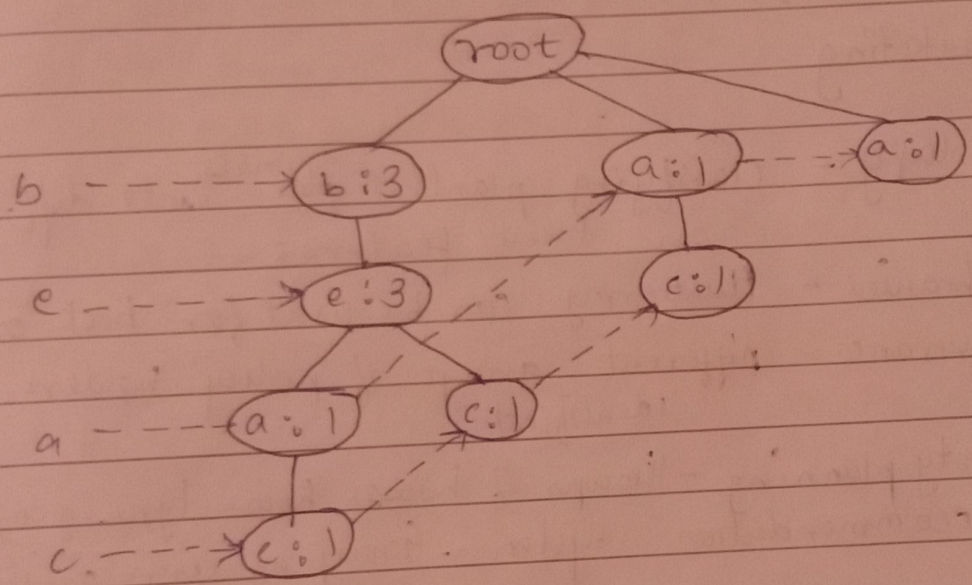
(4)



(5)



(6)



Mining Frequent Itemset w/o candidate generation

① Horizontal Data format (HDF)