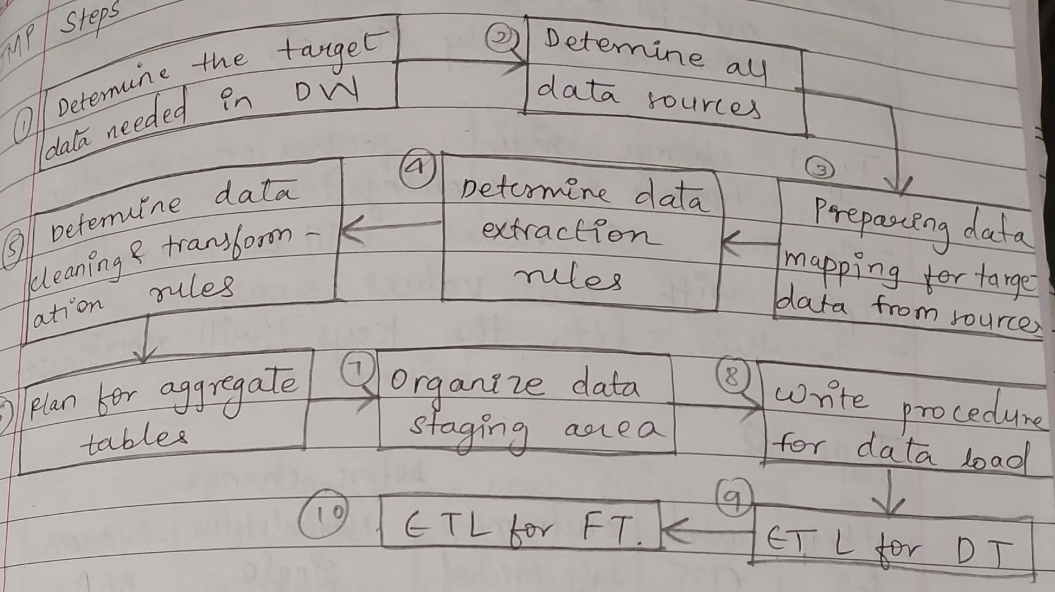
5/8/23

IMP Steps in ETL



Updates to DT:

(1) Type 1 :- Overwrite DT (correction of errors)

(2) Type 2 :- Add a new dimension record (Preservation of history).

(3) Type 3 :- Create new fields in dimension records

Type 1

cust_key	cust_id	custname	martialstatus	address
S01	C125	July Michael	single	AAA.

Before change.

custkey	custid	custname	martialstatus	Address
S01	C125	July Michel	single	AAA.

After change

incremental load
cust_id = C125
custname - July Michel

- Type1 change mainly focuses on correction of errors (errors in spellings).
- In this, old values will be discarded and replace with new values.
- In this, both the keys will remain same.

Type 2

before change

custkey	custid	custname	Martial status	Address
S01	C125	July Michel	single	AAA

custkey	custid	custname	Martial status	Address	Effective date
S01	C125	Jyly Michel	single	AAA	—
S02	C125	July Michel	Married		13th June 19

after change.

incremental load
custid = C125
martial status = married
Effective date = 13th June 1998

- Type 2 change, history is preserved by adding new record and field to the table.
- In this, surrogate key will be changed but primary key will remain same.

Drawback : It increases storage space

Type 3

Before change.

custkey	custid	custname	M.S	Address
501	C125	July Michel	Single	AAA

after change

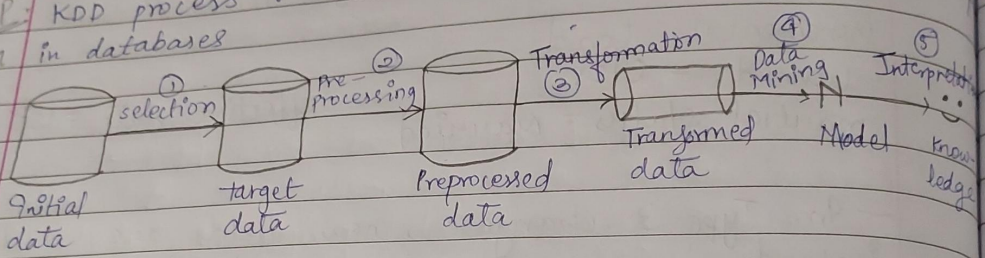
custkey	custid	custname	old M.S	New M.S	Effective date
501	C125	July Michel	Single	Married	13th June 1998

incremental load
custid = C125
martial status = Married
effective date = 13th June 1998.

- In Type 3 change, new fields are added instead of records. thus storage space is less as compared to type 2.
- History is preserved and both the keys will remain same

Data Mining : It is a process of extracting or mining knowledge from data warehouses. It is also termed as deductive learning exploratary data analysis.

IMP. 10M KDD process : It stands for Knowledge Discovery in databases



KDD process involves 5 steps :

(1) Selection
- In this step, data is selected from heterogeneous sources (internal, external, archived, production).

(2) Preprocessing
- In this, data is preprocessed by applying cleaning methods.

(3) Transformation
- In this, data is transformed into standardized format by checking datatypes or fieldlengths.

(4) Data Mining
- We apply various algorithms for extracting data to generate desired results.

(5) Intrepretation / Visualization / Evaluation.
- In this, various visualization and GUI

strategies are used. Visualization refers to the visual representation of data.

Visualization methods :-

(1) Graphical - Bar chart, line graph, pie chart
(2) Geometric - Boxplot, scatter diagram
(3) Icon-based - Icons, figures, menu
(4) Pixel-based - Coloured pixel
(5) Hiearchial - Tree structure
(6) Hybrid - combination of two or more methods

Issues in DM

(1) Human Interaction
- In DW, data input is often not clear and therefore domain as well technical experts are needed to use DM algorithms.

(2) Overfitting
- This problem arises when model is developed by considering current states. Such model is not capable for analyzing future states.

(3) Outliers
- Outliers are the values which do not fit nicely in given dataset (valid values).

(4) Intrepretation / visualization of result
- As many experts are involved in DM, data can be represented / visualized differently

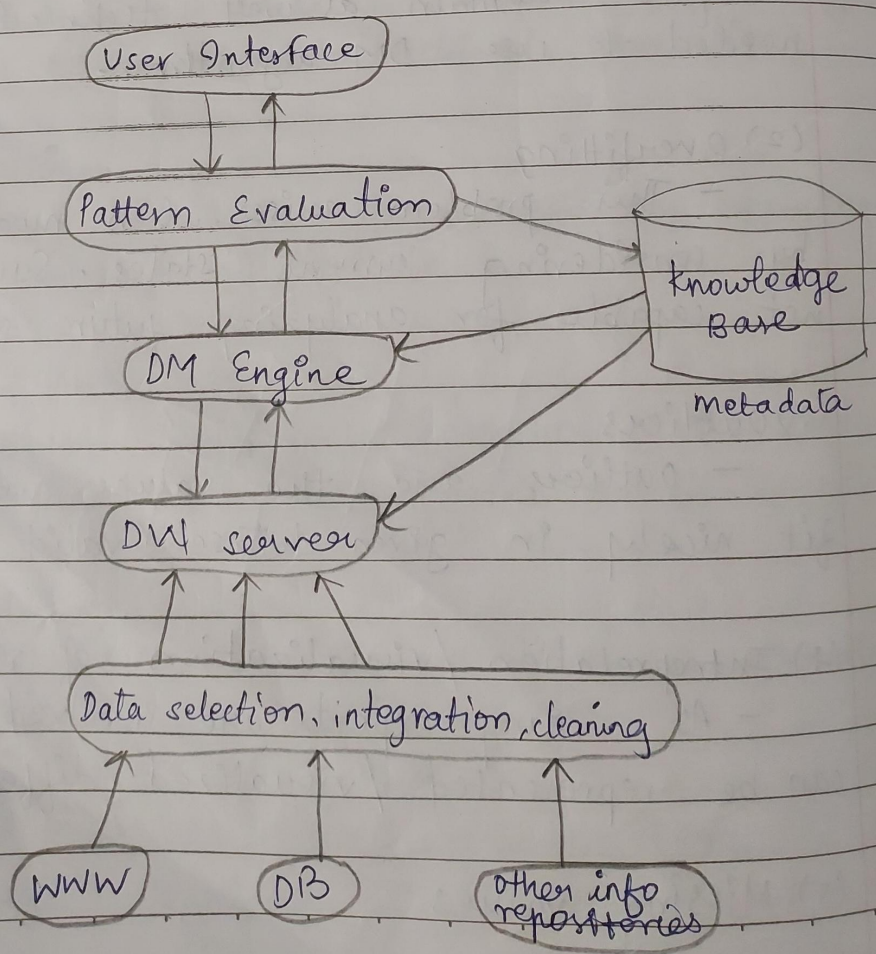
(5) Missing values

- Applying default values to missing data may create wrong data in DW which is not proper for analysis.

(6) Irrelevant / Noisy data
— Values which are invalid / wrong termed as irrelevant values

(7) Changing Data
— DM must be assumed to be static, however most DB is assumed data to be non-static. So this changing data creates problem in DW as data is needed for analysis.

DM architecture



In DM, architecture data is collected from heterogeneous sources (WWW, DB). These sources are cleansed and converted in form of subjects and stored in DW server.
Users are using GUI strategies for communicating with DW.
Pattern evaluation component is used to evaluate patterns based on user's interest.
DM engine is responsible for applying various algo, patterns.
Knowledge Base is used as meta data to DM architecture.

Types of Attributes

Attribute describes the features or characteristic of object

1. Nominal Attribute.
- It relates to the name of person, places and events
Ex: Hair color (Black, Brown, White)
- We do not have ranking for the values of nominal attribute.

2. Ordinal Attribute
- We do have ranking for the values of ordinal attributes.
Ex: Designation of Employee

3. Binary Attribute
- It will take only 2 values (0 or 1, True or False, Yes or No, Positive or Negative).

- It is also referred as Boolean attribute.

(a) Symmetric Binary Attribute
- If both states are equally valuable, it is symmetric binary attribute
Ex: Gender of person

(b) Asymmetric Binary Attribute
- Where one state is more important than other state it becomes a symmetric binary attribute
Ex:- Test result of disease

4. Discrete VS continuous Attribute
- A discrete attribute has finite or countable values
Ex: Grades of students
- Attribute which is not discrete is known as continuous
Ex: set of natural no.

5. Numeric Attribute
- It is represented with the help of numbers

(a) Interval scaled
- measured on scale of equal size of units (+, - possible, x, ÷ not possible)
Ex: temperature

(b) Ratio scaled
- Represented with the help of multiple of another value
Ex: age

Measuring Similarity and Dissimilarity – for grouping purpose

1. Data Matrix VS Dissimilarity matrix

- Data Matrix is represented with the help of table whereas dissimilarity matrix is represented by finding out differences between objects.

(a) data matrix

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & & x_{np} \end{bmatrix}$$

$n \times p$

rows x columns

(b) dissimilarity matrix

object j

object i	1	2	3	4
1	d(1,1)			
2	d(2,1)	d(2,2)		
3	d(3,1)	d(3,2)	d(3,3)	
4	d(4,1)	d(4,2)	d(4,3)	d(4,4)

Distance (4 − 4) = 0.
~~take~~ distance is always +ve

$$sim(i, j) = 1 - d(i, j)$$

2. Proximity measures for nominal attributes

- For creating similarity and disimilarity for nominal attribute, 0's and 1's are used.
- When both the objects are matched we put 0 and when there is dissimilarity we put value as 1.

object Identifier	Test 1 (nominal)	Test 2 (ordinal)	Test 3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	68
4	code A	excellent	35

Dissimilarity / similarity matrix

	A	B	C	A
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
A	0	1	1	0

~~Similarity matrix~~

	A	B	C	A
A	1	0		
B	0	1		
C	0	0		
A	1	0		

3. Proximal measures for ordinal attributes

- In this, we are representing values with the help of range between 0 to 1

dissimilarity matrix

	(1) e	(0.5) F	(0) g	(1) e
E (1)	0	0.5	1	0
F (0.5)	0.5	0	0.5	0.5
g (0)	1	0.5	0	1
E (1)	0	0.5	1	0

e = 1
f = 0.5
g = 0

From above matrix we can conclude that

object (4,1) is most similar and object (3,1) and (4,3) are most dissimilar

4. Bonuity measures for binary attributes

		object j		
		1	0	sum
object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

where $p = q + r + s + t$

$q = 1, 1$ $\quad s = 0, 1$

$r = 1, 0$ $\quad t = 0, 0$

SBA

dissimilarity

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

ABA

$$d(i,j) = \frac{r+s}{q+r+s}$$

Eg:

name	Gender	fever	cough	T_1	T_2	T_3	T_4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Meny	F	Y	N	P	N	P	N.

→ Assume $Y = 1$, $N = 0$, $P = 1$, $N = 0$.

① $d(\text{Jack}, \text{Jim}) = \frac{r+s}{q+r+s} = \frac{1+1}{1+1+1} = \frac{2}{3} = 0.67$

② $d(\text{Jim}, \text{Meny}) = \frac{r+s}{q+r+s} = \frac{1+2}{1+1+2} = \frac{3}{4} = 0.75$

less value → similar
more value → dissimilar

③ $d(\text{Jack}, \text{Mary}) = \frac{r+s}{q+r+s} = \frac{0+1}{2+0+1} = \frac{1}{3} = 0.33$

From above we can conclude that Jack and Mary has diseases similar in nature and ~~Jack~~ and ~~mary~~ has diseases dissimilar in nature.

14/05/23 3. Proximity measure for numeric attribute

(1) Euclidean measure distance

Numeric attribute is represented by following three form

(1) Euclidean distance (when $h=2$)

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \ldots\ldots + (x_{ip}-x_{jp})^2}$$

(2) Manhattan distance (when $h=1$)

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \ldots\ldots + |x_{ip}-x_{jp}|$$

(3) Minkowski distance

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \ldots\ldots + |x_{ip}-x_{jp}|^h}$$

$x_1 = (1,2)$, $x_2 = (3,5)$

Euclidean distance $d(i,j) = \sqrt{(1-3)^2 + (2-5)^2} = \sqrt{13} =$

Manhattan distance $d(i,j) = |1-3| + |2-5| =$

~~Min~~

Minkowski distance is generalisation of Euclidean distance and Manhattan distance.

6. Cosine similarity

It is used to find out similarity between two documents and it is given as below

$$sim(x,y) = \frac{\Sigma(x \cdot y)}{||x|| \cdot ||y||}$$

where $||x||$ = Euclidean distance of x

$||y||$ = Euclidean distance of y.

Doc	team	coach	hockey	baseball	score	win	loss	trophy
1	5	0	3	2	2	2	0	0
2	3	0	2	0	0	1	0	1

$x = (5, 0, 3, 2, 2, 2, 0, 0)$

$y = (3, 0, 2, 0, 0, 1, 0, 1)$

consider object x and y

$$sim(x,y) = \frac{\Sigma(xy)}{||x|| \cdot ||y||}$$

$$\Sigma(xy) = [(5\times3) + (0\times0) + (3\times2) + (2\times0) + (2\times0) + (2\times1) + (0\times0) + (0\times1)]$$

$$= 15 + 6 + 2 = 23$$

$$||x|| = \sqrt{(5-0)^2 + (0-0)^2 + (3-0)^2 + (2-0)^2 + (2-0)^2 + (2-0) + (0-0)^2}$$

$$= \sqrt{5^2 + 0^2 + 3^2 + 2^2 + 2^2 + 2^2 + 0^2 + 0^2}$$

$$= \sqrt{46}$$

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

$$= \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 1^2 + 0^2 + 1^2}$$

$$= \sqrt{15}$$

$$\sin(i,j) = \frac{\Sigma(xy)}{\|x\|\cdot\|y\|} = \frac{23}{\sqrt{46}\cdot\sqrt{15}} = 0.$$

Data Pre-processing

- Us

Data Pre-processing gets divided into 4 parts :-

(1) Data Cleaning
(2) Data Transformation
(3) Data Integration
(4) Data Reduction

1. Data Cleaning - Used to handle missing data and noisy data.

(a) Missing data

- It is handle by following 4 ways :

(i) Ignore the tuple

- When values are missing for the classes those tuples are not be the part of analysis and hence can be ignored.

(ii) Fill the missing values manually

- In these we manually check for missing values and replace those missing values with default value

(iii) Use global constant.

- In these missing values are replaced by either unknown class or $-\infty$.

- This method is not feasible has data mining will consider this classes as actual classes and will start classfiying them.

(iv) Use measure central tendency

- In this, missing values are replaced by mean or the median.

(b) Noisy data
- It is the data which is known as unwanted ~~data~~ error or irrelevant data present in DB
- It is handled by the method called 'binning' where we sort the data based on its neighbourhood.
- Binning provides 3 methods :-
(i) Smoothing by means
(ii) Smoothing by medians
(iii) Smoothing by boundaries

Q Ex : Following data set is given in which ~~to~~ noisy data is present. Apply binning concept having bin size = 3.
4, 8, 15, 21, 21, 24, 25, 26, 29, 9, 28, 34.
→ Step 1: Sort the data in ascending order

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

Step 2: Partition into equal bins
bin 1 = 4, 8, 9, 15
bin 2 = 21, 21, 24, 25
bin 3 = 26, 28, 29, 34.

Step 3: Smoothing by means (Find avg & replace all values by avg no)
bin 1 = 9, 9, 9, 9
bin 2 = ~~22.75, 22.75, 22.75, 18~~ 22.75
bin 3 = 29.25, 29.25, 29.25, 29.25

Step 4 : Smoothing by median
bin 1 = 8.5, 8.5, 8.5, 8.5
bin 2 = 22.5, 22.5, 22.5, 22.5
bin 3 = 28.5, 28.5, 28.5, 28.5

Step 5: Smoothing by bin boundaries
bin 1 = 4, 4, 4, 15
bin 2 = 21, 21, 25, 25
bin 3 = 26, 26, 26, 34.

Data integration - Process of managing data from heteregenous sources
It involves following problems
(1) Entity Identification Problem
- As DW is larger in size, identifying the entity is the biggest challenge in data integrity.
- As various attribute are involved in entity identification of same attribute is also the challenge.
for eg:- How stakeholder will differentiate between customer_id and c_id that are of same types
- Solution to Entity Identification problem is to maintain metadata.

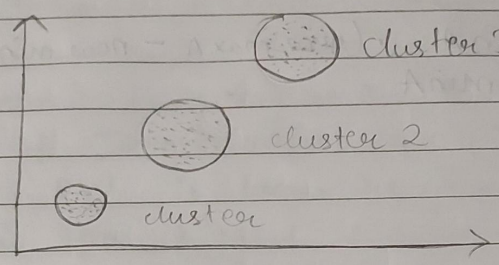
Data Transformation - Process of changing the form or structure of the existing attribute.
- It is separate from data cleaning as it does not correct existing attribute or add new attribute instead it structures the existing attribute for data mining
- It is carried out with the help of normalization
- Normalization is the process of changing the scale of existing attribute such a way that they will fall in range of 0.0 - 1.0.

$$v' = \frac{v - minA}{minA - maxA} (new_maxA - new_minA) + new_minA.$$

For eg:- We have range for salary as 20,000 to 400,000 and it should be normalized and transform into range 0.0 - 1.0.
If we have to find out the new value for 83,500, it should be calculated as follows.

$$V' = \frac{V - \min A}{\max A - \min A} \cdot (\text{new_max}A - \text{new_min}A) + \text{new_min}A$$

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		(1) DBSCAN
(2) Divisive	(2) K-medoid		(2) OPTICS.

Syllabus

V.V. IMP 10M

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.

Q 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.

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

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

$m_1 = 7$

$m_2 = 26.5$

Use K-means algo to create three clusters for given set of data:
{2, 3, 6, 8, 9, 12, 15, 18, 22}

→ Step 1:

$K_1 = \{2, 8, 15\}$
$K_2 = \{3, 9, 18\}$
$K_3 = \{6, 12, 22\}$

$m_1 = 8.3$
$m_2 = 10$
$m_3 = 13.33$

Step 2:

$K_1 = \{2, 8, 3, 9, 6$
$K_2 = \{$ ~~18,~~
$K_3 = \{18, 15, 12$

Suppose we have 4 objects and each object has 2 attributes shown in table below. Our goal is to group there 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 (In

$C_1 = (1,1)$

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

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-coordinate and y-coordinate is taken respectively.

Step 5: Calculation of objects distance matrix.

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

Step 6: Group of clusters

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

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

	1	2
C_1	4	9
C_2		

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

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

~~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.

$$D' = \begin{array}{c} \begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{array} \\ \begin{bmatrix} 0 & 5 & 8.4 & 3.6 & 7.07 & 7.21 & 8.06 & 2.25 \\ 5.6 & 4.1 & 2.8 & 2.2 & 1.4 & 2 & 6.4 & 3.6 \\ 6.5 & 1.5 & 6.5 & 5.7 & 5.7 & 4.5 & 1.5 & \cancel{6.04}\ 6.5 \end{bmatrix} \end{array} \begin{array}{l} C_1 = (2,10) \\ C_2 = (6,6) \\ C_3 = (1.5, 3.5) \end{array}$$

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.

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

$C_1 = (3, 9.5)$
$C_2 = (6.5, 5.25)$
$C_3 = (1.5, 3.5)$

Step 9: Grouping of clusters

$$G' = \begin{array}{c} \\ \\ \\ \end{array}\begin{array}{cccccccc} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$C_1 = (3, 9.5)$
$C_2 = (6.5, 5.25)$
$C_3 = (1.5, 3.5)$

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

$$D^3 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 8.5 & 3.8 & 24.6 & 0.3 & 3.5 & 3.9 & 6.8 & 1.4 \\ 7.1 & 4.5 & 1.5 & 3.8 & 0.7 & 0.7 & 6.4 & 5.4 \\ 6.5 & 1.5 & 6.5 & 5.7 & 5.7 & 4.5 & 1.5 & \cancel{8.0}\ 36.5 \end{bmatrix} \\ \begin{matrix} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{matrix} \end{array} \quad \begin{matrix} C_1 = (4.75, 7.5) \\ C_2 = (6.5, 4.5) \\ C_3 = (1.5, 3.5) \end{matrix}$$

Step 12: Grouping of clusters.

$$G^3 = \begin{array}{c} \begin{matrix} (2,10) & (2,5) & (8,4) & (5,8) & (7,5) & (6,4) & (1,2) & (4,9) \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \\ \begin{matrix} A_1 & A_2 & A_3 & B_1 & B_2 & B_3 & C_1 & C_2 \end{matrix} \end{array} \quad \begin{matrix} C_1 = (4.75, 7.5) \\ C_2 = (6.5, 4.5) \\ C_3 = (1.5, 3.5) \end{matrix}$$

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).

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 : Calullation 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.

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

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

$= -12.$

Step 5 : As $TSC \neq 0$
$\therefore$ can be replaced by 15

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

Step 6 : Now check whether 7 can be replaced by 10 ~~or~~ 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 answer as

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

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

$K = 3.$