

$$\begin{aligned} WJMK &= \frac{3}{50} = 0.06 \\ &= 6\% \end{aligned}$$

7 Jan 2024
Tuesday

Lecture - 1 :-

Data Analytics / Mining :-

Analysis + Algorithm

↳ which enables us to do analysis

* 3 components associated with it :-

1) Association mining ↳ Apriori algo

2) Classification ↳ then its several variations.

3) Clustering ↳

② frequent pattern (FP) growth algorithm

③ observations :-
statistical add ons,
metric sufficiency,
clustering of ARs
(Association rules)

- ① Decision tree and its variations
- ② Bayesian
- ③ Back propagation
- ④ SVM
- ⑤ Linear regression

① partition based

② hierarchical based

③ density based

① single, average and complete linkage algo.
② GPRCH

① CURE

② DBSCAN

③ BFR

k-mean, PAM, clara, clarans.

Variations of k-mean & PAM, to enable larger data analysis

① Introduction

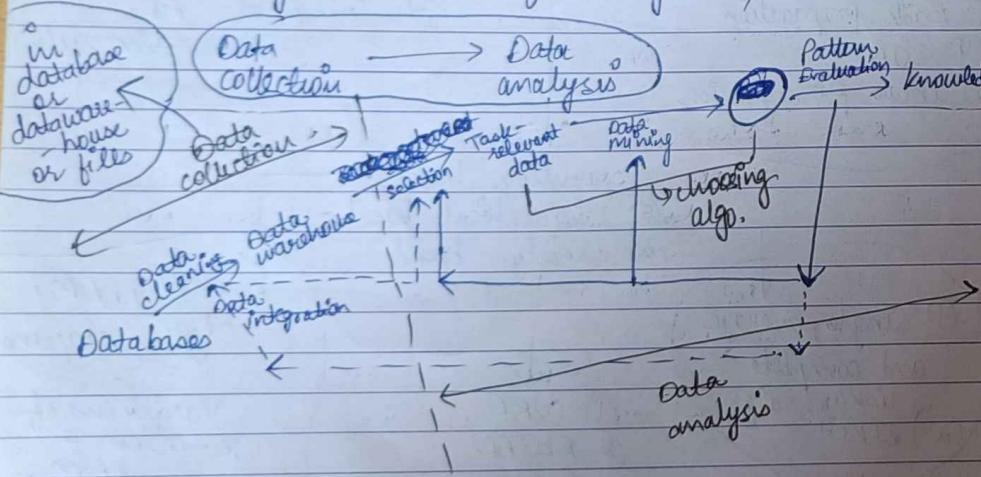
- ↳ ① architecture
- ② DA engineering
- ③ steps
- ④ functionalities
- ⑤ Issues

② we are in era of data. we have data of different types: audio, video, text, numerical etc.

↳ How to handle and analyse such big data ??

Data Warehouse ?

Data analysis involves following steps :-



* Database and warehouse are two different things.

WORK
30/02/2023

WORK
30/02/2023

* Data Mining — KDD process

↳ the core of knowledge discovery process.

→ If ek company ka database bnana h, then what u will ask from company :-

① what applications are req. for your company

so,

database → application oriented data collection.

↳ so wo data to go application mein help nahe.

eg. our college ki applications are :-

library, exp etc.

like college

does not need our siblings details, as they are not useful for that application to be executed.

whereas,

(Data warehouse) is subject-oriented and is always historical data.

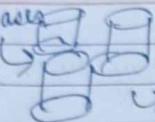
↳ means like we are not in warehouse as we are still in clg and not history, jab clg se chle gye, then we are entered in warehouse.

→ Means data in warehouse is fixed, and will never change like my CGPA after going out from clg.

→ consists data from that date jab se clg bna

→ Database stores current data.

database



why 3 cylinders?

as database changes its formats

with time, like pihli S&L, then kuch
aur, kabhi koi column add kr diya hogar, kabhi
kuch update etc.

warehouse → collection of entire data

jisker kisi saare format
of database ko integrate karna phlega

for that need some interface programming,
to reformat things so that they can talk to
each other.

database of
different format

Analysis

1st thing :- get / extract task-relevant
data out of entire huge data.

Now how to extract this data from big data???

with help of finding queries

which will definitely be a
complex query.

Now, task-relevant data mil gyaa, ab
ML, Deep learning algo use keinge.

Now to use which algo??

which algo to select??

most difficult thing
to understand (?)

WJMK
= 50%
50%

WJMK
= 50%
50%

WJMK
= 50%
50%

→ algo to be used will depend on requirement of
customer and not of developer.

→ when we apply algo

we get patterns

mostly 100s of pattern
nibble hain,
these depends on algo.

then pattern evaluation comes into picture,
to figure out which pattern is good
for customer!

for this matrixes are used,
and then we get 'knowledge'

yehi toh humara main

target hai!

① 1st category of state:-

(*) Association Rule Mining :-

traces your pattern of
purchase and interprets what and
how you are purchasing.

Say:- Bread and butter has strong association.

How they come to know?

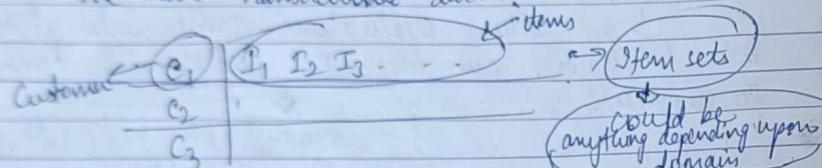
By looking into your
basket and applying maths.

ARMK

Say total 500 people, out of them 400 bought bread, (80%) and out of 400, 360 bought butter, so 90% confidence level based on purchase that bread and butter have strong association.

→ each customer is one transaction for retail shop.

We have Transactional data:-



↳ we do not need personal details of the customer.

→ Once transactional data hai, then association rule mining say, apply also on this data and find patterns (also known as association rules).
like:- $X \rightarrow Y$.

X, Y may be single items or multiple items.

age (< 30) ^ buy (laptop) → buy (earphone)

X e Y

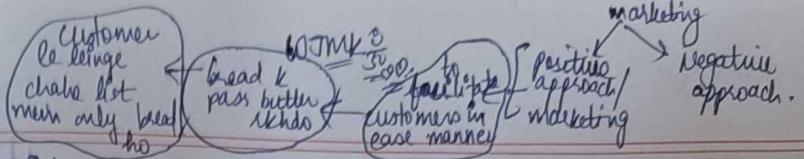
2 dimension rule

↳ age, buy

buy (bread) → buy (butter)

↳ single dimension

So, rule may be single dimension or multiple dimension.



→ bread - butter ka confidence level abhi jyada than laptop - earphone, toh so confidence level varies with items, and yeh chalta hai,

agar confidence level of bread - butter = 90%, then " " " laptop - earphone = 20%, is enough for retailers, as, bread - butter jitna frequently koi buy krega, laptop - earphone nahi to krega.

and when you go to buy laptop, toh wo earphones thi dikhale hain and your mind may change.

Negative marketing approach

↳ butter and bread deno dena store,

say list has both, then bread lekar, butter lena jaoge, toh main items dikhenge and you will start buying, check wo list main nahi thi.

→ AR helps in inventory & also,

or demand to manufacturers

association rules also (AR)

↳ also helps in managing warehouses, helps in not only restricted to retail market, also comes in fraud detection (???)

Usually support and confidence of AR is user defined.

① Most important algo that comes in this category:-

mostly retail market tells what is best and supports and confidence for them, a user analyst can also figure out at background because of knowledge he has.

(a) Types of AR:-

① Boolean Quantitative ARs :-
 You look for absence or presence.
 e.g. bread \rightarrow butter

age (X , "30 to 39") & income (X , "42k...48k")
 \rightarrow buys (X , projection TV)

There are 3 dimensions

② Single/Multi-dimensional ARs.
 ③ Single/multi-level ARs.

Bread Butter
 white... full
 brown cream
 butter

Bread and butter bhi age different types k. hote hain, but not interested in that till now, but might be interested, then that comes under this type of ARs.

\rightarrow s may also be interested to know that how brown bread is associated with bread, butter.

brown bread \rightarrow butter (?)
 Bread \rightarrow full cream butter (?)

so, bread, butter are supersets.

brown bread \rightarrow butter \Rightarrow say (70%, 80%)
 support confidence

\Rightarrow even if (60%, 70%) data, then it is also pattern for retailers.

Support and confidence:-

- ④ A rule must have some minimum user-specified confidence and support.
- ⑤ AR $X \Rightarrow Y$ holds with support T, if T% of transactions in DB that support X also support Y,
 so jaise thi confidence and support decide kaise hain, then unko min. " " " bethe hain, means agar unko niche. " " ", then we will reject.

① \rightarrow set of all items

② \rightarrow transactional database.

AR $A \Rightarrow B$ has support 's' if 's' is % of transactions in D that contain $A \cup B$ (both A and B)

$$s(A \Rightarrow B) = P(A \cup B) \rightarrow \text{kitne logon ne bread-butter khareeda.}$$

AR $A \Rightarrow B$ has confidence 'c' in D if c is % of transactions in D containing A that also contain B.

$$c(A \Rightarrow B) = P(B/A) = \frac{P(A \cup B)}{P(A)}$$

Kitne logon ne bread k. baad butter bhp khareeda.

Example:-

Transaction ID	purchased items
1	{1, 2, 3, 4}
2	{1, 4, 5}
3	{1, 3, 5}
4	{2, 5, 6, 7}

for min s = 50%, min c = 50%
what rules we should have.

A. Try:- 1 → 2

$$s = 1/4$$

$$c = 1/3$$

so not!

1 → 3 ✓

$$s = 2/4 = 50\%$$

$$c = 2/3 = 66.66\% \text{ so chlega } \checkmark$$

(*) Algo for finding (F₁) (frequent itemsets):-

① Apriori

② Sampling

③ Partitioning

④ Hash based technique

⑤ Transaction reduction

etc.

6. Variation of apriori only, to solve some specific problem.

WOMEN'S
TOPS.

so, main algo → apriori hai.

Apriori algo :-

↳ has main 2 properties:-

(1) Apriority property.

(2) Anti-monotone property.

say min s = 50%.

→ so item is said to be frequent, then it is which appears in more than or equal to 50% of transaction.

↳ means it must satisfy min. support.

→ if SABY is frequent itemset, both SA³, SB³ should also be frequent itemset.

Anti-monotone property

↳ if a set does not

pass property, then its every superset will also not pass property.

→ So both properties are complementary to each other and both helps in analysis.

→ dusri property ketti agar chota set fail, big switching that chota, dekho chota hi mat.

① Apriori algo example:-

Database (R)

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

say :- $s = 50\%$
 Step 1 list unique items
 $\rightarrow 1, 2, 3, 4, 5$

Step 2 :- Candidate set formed

(frequency kii h sbii)

Itemset	sup.
1	2
2	3
3	3
4	1
5	3

$L_1 \rightarrow$ frequent itemsets, large items.

Itemset	sup
1	2
2	3
3	3
5	3

Step 3 :- $L_1 \times L_1$ (join kiji)

Itemset
1, 2
1, 3
1, 5
2, 3
2, 5
3, 5

\Rightarrow

Itemset	sup
1, 2	1
1, 3	2
1, 5	1
2, 3	2
2, 5	3
3, 5	2

Step 6

Itemset	sup
1, 3	2
2, 3	2
2, 5	3
3, 5	2

\rightarrow then do $L_2 \times L_2$

W.M.K
 20 Mar.

Step 7

Itemset	sup
2, 3, 5	2

Step 8

Itemset	sup
2, 3, 5	2

only one combination

otherwise $L_4 \dots$ aage jaati p till ek combi
 left, so algo stops
 reh jaaye!

here L_3 pe mukri, so !

$L = L_1 \cup L_2 \cup L_3$ (large itemsets).

* Single items can not generate sufficient rules.

$1 \rightarrow 3$] \rightarrow are different, (billkul different)

$3 \rightarrow 1$

\rightarrow donka confidence will be different.

So, ~~1, 2, 3, 5~~ deletion? ismin also
 may go fav $\{2, 3\} \rightarrow 5$
 on $2 \rightarrow \{2, 5\}$

and many other.

WEEK 9
TERM

Now, so confidence bhi dekhlo
so want min c = 90%.

so jinka $C \geq 90\%$. wo nako, and baaki
discard ker!!!!

→ But this algo would be very-very complex, if transaction
data is very-very large!
What is TC=? of this algo.

→ In each iteration, we are scanning data!

say D = 50 GB

and RAM = 8 GB.

↓
then each time, 7 blocks u have to
retrieve,
so 7 times input-output, swap-in \rightarrow
swap out!

as $D \uparrow$, I/O cost \uparrow ,
that's why variations in apriori algo has been
proposed.

Ex: $L = \{2, 3, 5\}$
 $\{2, 3\}, \{2, 5\}, \{3, 5\}, \{2, 3\}, \{3, 5\}, \{2, 5\}$

↳ AR ans:-

$2, 3, 5 \rightarrow 100\%$

$2, 5, 3 \rightarrow 66\%$

$3, 5, 2 \rightarrow 100\%$

$2, 2, 3 \rightarrow 100\%$

$3, 2, 5 \rightarrow ??$

WEEK 9
TERM

WEEK 9
TERM

28/1/25

TUESDAY

lec-3 :-

- When we talk about database, or database house, the task-relevant data $\xrightarrow{\text{(TRD)}}$ is very important not in files, bcz. files can be simply called in the files.
- and getting TRD is with help of complex queries which definitely must satisfy all constraints, as our analysis requires only specific data.

STEPS of a KDD process :-

(1) Learn the application domain:

— relevant prior knowledge and goals of the application

— databases are actually created on the basis of application, bhut raab reformat krna pdh jaata hai data, clean krke data ko.

(2) Creating a target data set

(3) Data cleaning and preprocessing

(4) Data reduction and transformation

(5) choosing fns of data mining

(6) choosing mining algo(s)

(7) Data mining

(8) Pattern ~~visualiz.~~ evaluation and knowledge presentation.

↳ (9) Use of discovered knowledge.

very critical as it req. logical reasoning, no such hard and fast rule mostly,

as have to know what features are important and what are not.

WJMK
2020

→ As we get so many patterns mostly and
sare zaroori nahi hote, so step (8) very
very important in which we filter patterns
after evaluating them.
↳ 3 things to consider:-

(1) pattern should be simple enough so that
it can be interpreted by you.

(2) Novel i.e. new data

(3) Confidence

statistically | mathematically

(4) Durability

↳ yeh y cheezekin zaroori hain, tabhi uss
pattern ko consider krna.
pattern ka future
main koi use ho,

koi ek bhi
cheez nahi hai,
then pattern
of no use.

→ Preprocessing → most difficult task and intensive task
out of all.

↳ firstly we will integrate data into 1
but for it also we need middleware / interface
programming as data can be in different
format, and we want that they can
interact with each other and integrate them, and
even missing cheezon fill kro yea ignore
kro, in a data.

WJMK
2020

What kind of data, can I mine?

↳ any kind of data

Yes tareeka bdl jayegi, algo
bdl jayegi, preprocessing ka tareeka
bdl jayegi but mining ho skti hai.

Spatial data

↳ locations are stored.

Time-series data

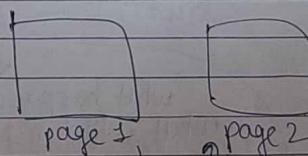
↳ Here time component is very-very
important.

Heterogeneous data

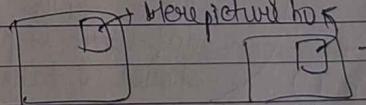
↳ data stored in different locations and
do not share same software or format,
www → semi structured

XML → semi structured

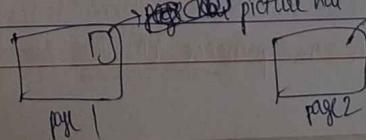
↳ can make it (structured) as well
difference?



changes everything → unstructured



→ trying to structure it



→ semi structured, as some
structured and some not

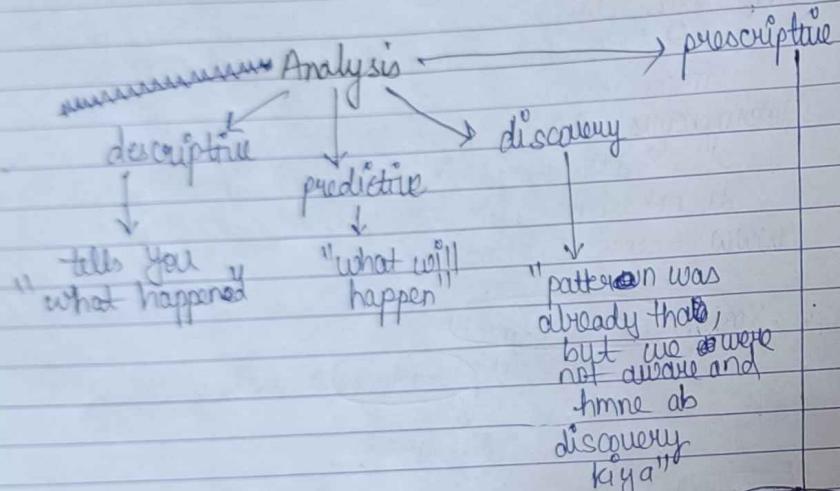
wrong form
If a page has set template, then structured or not.

Data Mining

↳ confluence of multiple disciplines

① Information Science

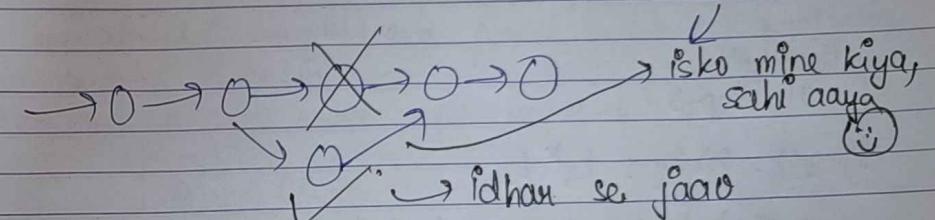
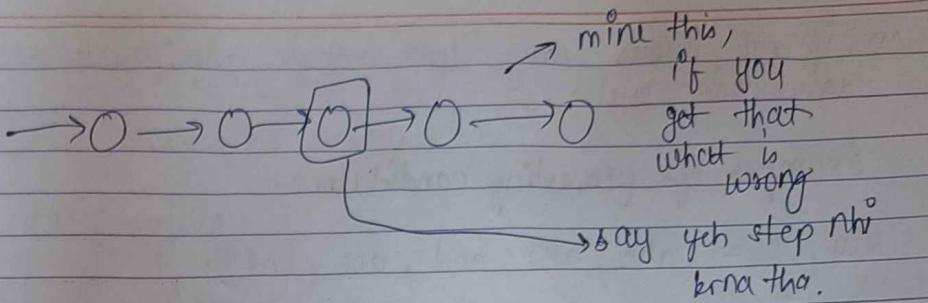
↳ here we can calculate entropy etc.



→ aaj kal mathey prescriptive +
descriptive analysis ki jaati hai.

combinedly known as process mining (PM)

WJMK
S
=



② process mining aaj kal, fyada, "system log" par ho raha hai,

Transaction Reduction :-

↳ we stop where candidate set can not be generated again

↳ it says that in subsequent stages of apne aur, if any transaction becomes insignificant, then we can ignore it, in all further steps. (that transaction)

↳ means jithi bhi items is main hain, that are no longer significant, means they are not appearing in large item set.
↳ how we can ignore them? what are advantages?
what will be impact?

↳ support % formula will change, so support both facayega, toh agar ignore nahi karega, then we may miss out some prominent transactions or item sets. so ignore karna zaruri.

so, if not ignore, may lose some significant association rules.

Examples of generating candidates:-

say, $L_3 = \{abc, abd, acd, ace, bcd\}$
↑
large item set at step-3

① self-joining: $L_3 \times L_3$

→ abcd from abc and abd
→ acde from acd and ace.

② pruning

→ acde is removed bcz ade is not in L_3 .

so, $C_4 = \{abcd\}$

so no need to scan database, evaluate support . . . etc.

so pruning is done before scanning database and calculate support.

ARs from FIs:-

for each FI l , generate all non-empty subsets of l , and for each non-empty subset s of l , output

then the rule:-

$[s \Rightarrow (l-s)]$ if $\frac{\text{support-count}(l)}{\text{support-count}(s)} \geq \text{min-conf}$
minimum confidence.

WORKING

Variations of Apriori :-

(1) Transaction reduction
↳ done.

(2) Sampling

random transactions of the original database are selected (sampled) and placed in a much smaller sampled database.
→ our task is not reduce input/output (I/O).

① → Apriori → AR's
→ large I/O.

so instead take samples-

→ apply apriori on each sample
→ calculate AR's

But these might

not represent all patterns,
so ek sample lo, apriori lgao, AR nikalo, then duera
sample lo, apply apriori, then AR milenge usko pehle
wale main add-on kro, → ~~not LI~~

② → apriori → PL → potentially large items.

then we talk about NBC fn (Negative borderline function)

FI = $P \cup_{NB}^{PL}$
↓
frequent item sets.

returns itemsets that are

not in PL but has all of their
subsets in PL

WORKS
when finding PL from sampled database.

Usually, min support threshold is covered.
How much can be covered?
depends on application domain

$$I = \{a, b, c, d\}$$

$$PL = \{\{a\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}, \{c, d\}\} \rightarrow \text{after step 2.}$$

$$C_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}, \{c, d\}\}$$

Since a and c are frequent, so $\{a, c\}$ also.
a and d → d →
so,

$$NFC = \{ac, ad\} \text{ till now.}$$

a, ac, ad, cd, a, c, d are frequent,
thus, acd also frequent.

so, $NFC = \{ac, ad, acd\}$
(ed) frequent hai but nahi hogा in NFC , as we
already PL mein hai, after step-2.

(3) Partitioning

Instead of sampling, do 'n'
partitions:- D_1, D_2, \dots, D_n .
Improve performance

Q. How to partition? what will size of each partition?
↳ each partition size must be less or equivalent to RAM
↳ so that each partition can be executed at once.

$$D_1 \rightarrow \text{apriori} \rightarrow FI_1$$

$$D_2 \rightarrow \text{apriori} \rightarrow FI_2$$

$$\vdots \quad \vdots \quad \vdots$$

$$D_n \rightarrow \text{apriori} \rightarrow FI_n$$

$$L = FI = FI_1 \cup FI_2 \cup \dots \cup FI_n$$

$$L = L_1 \cup L_2 \cup \dots \cup L_n$$

$$L = L^1 \cup L^2 \cup \dots \cup L^n$$

then generate rules.

then what difference it makes?

↳ as still doing same thing only,
accessing same and whole data.

will there be same time complexity as of apriori
original?

No! it will be less.

say

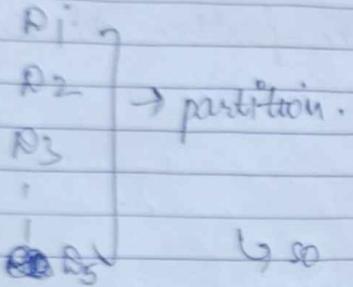
$$(D) \rightarrow \text{apriori} \quad (2 \text{ GB} = \text{RAM})$$

$$10 \text{ GB} \rightarrow \text{Iteration of apriori}$$

→ for each D_i , 5 input/output are req.
min

say 5 iteration

then min $5 \times 5 (= 25)$ I/O req.

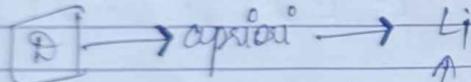


Now each iteration will take \pm 1 I/O.
min

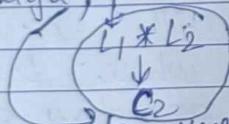
↳ so overall 5 I/O.

(4) Hash based

↳ it uses hash function to generate LIs.



then
hash
use
kiya,



to prune C_2 .

so instead of scanning whole data base for L_2 ,
we will do pruning first with help of L_1 .

fp → frequent pattern approach

fp growth
approach.

↳ In this just one iteration of database is req.

H/W → utility mining

↳ when we consider quantity of items in transaction database.

WJMK

5

WJMK

5

WJMK

5

eg: how many bread and butter customer has taken.

4/2/25

Tuesday

lec-4 :-

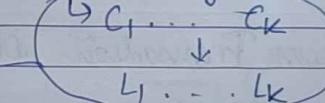
FP growth algorithm

frequent pattern

↳ which is actually tree data structure based algorithm.

in apriori → large I/O.

↳ FP growth algo was revised,
candidate generation



FP growth says that 'no candidate generation req'.
so this part will be skipped in FP,
and that time of scanning database will get reduced.

FP

↳ divide and conquer methodology : decompose mining tasks into smaller ones.

→ requires 2 scans of transaction DB
→ 2 phase algo

↳ Phase I: - construct FP tree

↳ Phase II: - seeing. FP tree

FP-tree Construction:-

↳ stem prefix tree

↳ Most frequent will always be preferred.)

WJML 30
20M

WJML 30
20M

- FP Header table
- dependent on ordering of items
- sort items in decreasing order of support count
so highest frequency will ↑ preferred.
- Non FIs are ignored
- each Tr. is viewed as a list of FIs in descending order of support count.

Item	frequency	support
1	5	→ so → order → 1
2	3	3
3	4	2

Construct FP tree from Transaction DB,

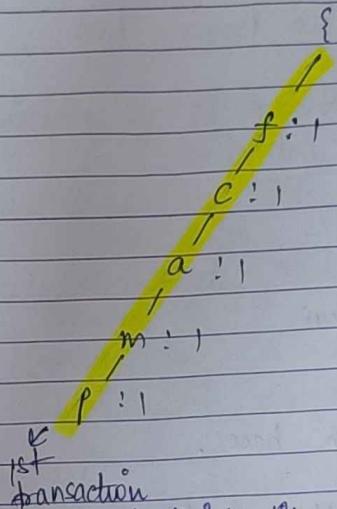
→ jinki frequency unkaise mazgi order mein
lo, no difference.

TID	Items	(Ordered) frequent items
100	f, a, c, d, g, i, m, p	f, c, g, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m	f, c, g, m, p

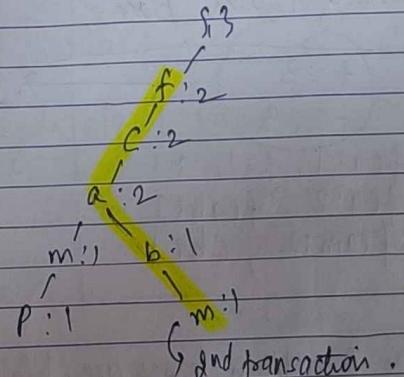
as min. support = 0.5

Header Table	
Item	frequency head
f	4
c	4
b	3
m	3
p	2

Now, we'll make tree :- null item as root.



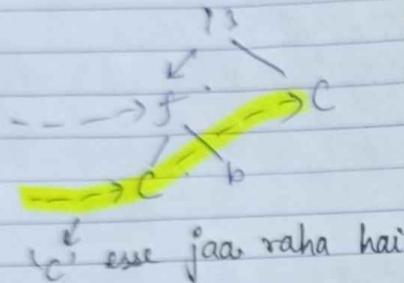
(Just think like tree, maintaining frequency as well)
2nd transaction :- f, c, a, b, m



so on, complete the tree !!!.

WJMK
=
=

→ dotted arrows tell us how transactions are participating



* Major steps to mine FP tree :-
↳ conditional pattern base.

* From FP tree to Conditional Pattern Base.

conditional Pattern Base.	
c	f : 3
a	fc : 3
b	fca:1, f:1, c:1
m	fca:2, fcab:1
p	fcam:2, cb:1

These 2 pattern bases are highly frequent.
confidence

$f \Rightarrow c$	3/4
$c \Rightarrow f$	1

can derive easily from pattern base

WJMK
=
=

→ So ab jo 'min confidence' condition di hogi uss thisaab se dekh leinge, what to ignore and what not to (:

Principles of FP growth:-

↳ pattern growth property.

Let α be frequent in DB, B be α 's conditional pattern base and β be itemset in B, then $\alpha \cup \beta$ is also frequent itemset in DB iff β is frequent in B, o/w just prune the things.

"abcdef" is FP \Leftrightarrow "abcde" is FP and "f" is frequent in the set of transactions containing "abcde".

(and you can see, it is very different from apriori).

Hash based algo (variation of apriori) :-

primarily works for k=2

you, we can also use for $k \geq 3$, but mostly not used

means C_2 tak jaana phega.

1stly generate C_1 , then using hash, we calculate C_2 , no need to go to database for scanning again (:

WJMK
 100 $\{(1, 3), (1, 4), (3, 4)\}$
 200 $\{(2, 3), (2, 5), (3, 5)\}$
 300 $\{(1, 2), (1, 3), (1, 5), (2, 3), (2, 5), (3, 5)\}$
 400 $\{(2, 5)\}$

$$H(x, y) = \{(\text{order of } x) * 10 + (\text{order of } y)\} \bmod 7$$

$$(3 * 10 + 5) \bmod 7 = 35 \bmod 7 = 0$$

Hash table.

			1, 3
3, 5		2, 5	
3, 5	2, 3	2, 5	3, 4
1, 4	1, 5	2, 3	1, 2
3	1	2	1, 3
0	2	3	3
count	1	2	6

Bucket No,
yjo hash function se pta chl raha hai.

so count ≥ 2 wale, to.

so bit vector $[\pm 0 \pm 0 1 0 1]$

$L_1 * L_2 = \{ \{1, 2\}, \{1, 3\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 5\} \}$
 No. in the
bucket
with
Item set.
 so,
 $C_2 = \{1, 3\}, \{2, 3\}, \{2, 5\}, \{3, 5\}$
 baaki & ignored.

WJMK
 support (???)
 as min - ~~0.00~~ = 50%
 4 transactions in total

so $\geq 2 \checkmark$ count lelo.

if Item sets are 'words' rather than 'no.', then
hash function may not work.

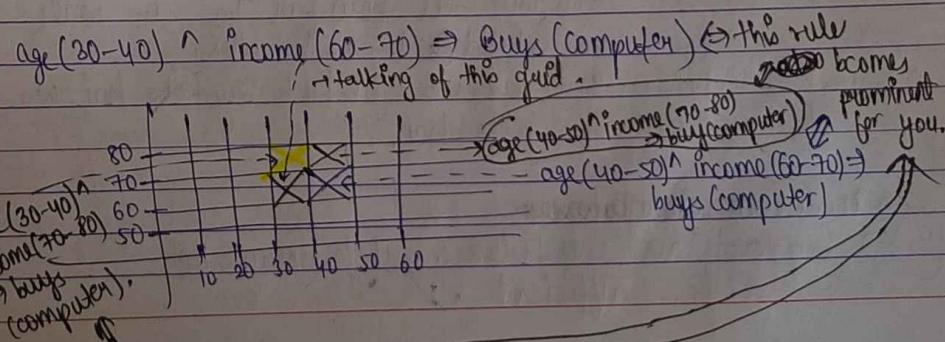
Q. If problem changes, how do figure out to the
hash function ???

*Multiple-Level Association Rules:-

- Items often ~~not~~ form hierarchy.
- Items at lower level are expected to have lower support.
- Rules regarding items at appropriate levels could be quite useful.

Observation:-

(1) clustering of association rules is making life easy, as jyada rules we can not handle, so ek rule bnab but then that rule must contain all other rules.



WTMK
= $\frac{\sum_{i=1}^n \text{sup}_{i,j}}{\sum_{i=1}^n \text{sup}_{i,j}}$

$A \Rightarrow B$

3rd metric can use is referred as lift

→ instead of these 4 rules, make single rule :-

age (30-50) ^ income (60-80) \Rightarrow buys (computer)
↳ best rule.

→ and unhi rules ko cluster kerna which has minimum confidence, so rules jo cluster kr rhe hne must be genuine one.
↳ also known as "long".

② → mining vertical Transactional database (TD)

→ abhi tak jf the bhi phelein hain TD, wo sare horizontal database the.
↳ trying to figure out association b/w items

TD	Item
T ₁₀	a, b
T ₁₁	a, c, d

→ Horizontal DB

Items ID	TD
a	T ₁₀ , T ₁₁
b	T ₁₀
c	T ₁₁
d	T ₁₁

→ trying to figure out association b/w customers.
↳ so that we can handle similar customers
↳ why this?
↳ what advantages in practical?

↳ this will tell how far T₁₀, T₁₁ wale customers are related to each other.

→ Transaction is nothing but customer jab kuch buy keta hai, wo ek transaction bn jaata hai.

③ → ~~support and confidence~~

Support and confidence akele may not be sufficient ;

b/w A and B, correlation

lift (A \Rightarrow B)

\downarrow
 $\left\{ \begin{array}{l} > 1 \rightarrow \text{positively correlated} \\ < 1 \rightarrow \text{negatively } " \end{array} \right.$

yeh chlega $\left\{ \begin{array}{l} = 1 \rightarrow \text{no correlation / Independent} \\ \text{yeh nahi, as no association,} \end{array} \right.$

H/W
↳ one example having lift as well
↳ what is lift.

④ → sometimes in place of 'lift', we may use

$$(chi)^2 = (O - E)^2$$

→ H/W → isse also ek example and write what it is,

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

* Data ~~representation~~ (preprocessing) :-

Data can be text, image, audio, video.

process of converting data into meaningful data.

① why we do this?

↳ because our data is dirty, incomplete, noisy, inconsistent and duplicate hoga.

and we want to clean it.

① steps of data preprocessing :-

- ↳ Data cleaning
- ↳ Data integration
- ↳ Data reduction
- ↳ Data Transformation

Data Cleaning

↳ handles 2 tasks :-

- (1) missing value
- (2) noisy data

Data Integration

↳ merge the data from different sources.

Data reduction

↳ redundancy

Data Transformation

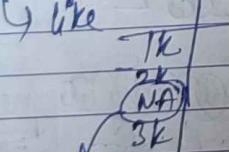
↳ normalise karna data ko.

↳ helps in data modelling.

② Techniques to remove noise :-

① Binning

↳ random error / unwanted data.



Ex:-
10, 2, 19, 18, 20, 18, 25, 28, 22

3 steps :-
① hole han, ②

smoothing by bin means
means
medians
boundary

Types of
Binning

WORME
= 5000

SMOOTHING
by bin numbers

Step 1
Sort data :-

2, 10, 18, 18, 19, 20, 22, 25, 28

then Step 2,
divide data into BIN buckets
say, bucket of 3 size
each

2, 10, 18

3 buckets.

18, 19, 20

22, 25, 28

* say smoothing by bin mean :-

2, 10, 18 → mean = 10

18, 19, 20 → " " 19

22, 25, 28 → " " 25.

so step 0 mean 'k' equal, kdo →

10 10 10

19 19 19

25 25 25

③ If by median :- odd = $\left(\frac{n+1}{2}\right)$ even = $\left(\frac{n}{2}\right)$

10
19
25 } → aayega → 80

10 10 10
19 19 19
25 25 25

WJMK
= 5 cm

say by boundary

min max

$$2, 10, 18 \rightarrow \text{min} = 2 \\ \text{max} = 18$$

replace 10 by min as usko paas hain

2, 2, 18
18, 18, 20
22, 22, 25

hmm \rightarrow chhota hain
dikha yeh kiske paas, as min and max deno
jyada

k paas, toh min ko prefer kroge.
as then big value dominate rabi kroga jab
ever nikaalenge yaa jab tree mein distance

Normalization

kringe ek specific range mein laane
k liye.

2 techniques:-

(1) min - max normalization technique

(2) Z - ~~square~~, normalization "

~~square~~ \rightarrow square

House	Sq. foot	Bedrooms	price (In lakhs)
1	1200	3	50
2	1500	4	80
3	1000	2	40
4	1800	5	80

WJMK
= 5 cm

old value.

$$X' = \frac{X - \text{min}}{\text{max} - \text{min}}$$

new value

for price \Rightarrow

$$\text{min} = 40$$

$$\text{max} = 80$$

$$\text{so for House 1} \rightarrow X' = \frac{50 - 40}{80 - 40}$$

$$= \frac{10}{40} = \frac{1}{4} = 0.25$$

[0, 1] range
mean

aagya :)

$$\text{for House 2} \rightarrow X' = \frac{80 - 40}{80 - 40}$$

$$= \frac{40}{40} = 1$$

$$\text{for House 3} \rightarrow X' = 0$$

$$\text{for House 4} \rightarrow X' = \frac{40 - 40}{80 - 40} = 0$$

for bedroom \Rightarrow

$$\text{min} = 2$$

$$\text{max} = 5$$

$$\text{for House 1} \rightarrow X' = 0.33$$

$$2 \rightarrow X' = 0.6$$

$$3 \rightarrow X' = 0$$

$$4 \rightarrow X' = 1$$

for price \Rightarrow

$$\text{for House 1} \rightarrow X' = 0.25$$

$$2 \rightarrow X' = 0.5$$

$$3 \rightarrow X' = 0$$

$$4 \rightarrow X' = 1$$

$$WJM \stackrel{K}{=} \frac{\sum M_i}{M_m}$$

if values are on same scale then
use Z-square method,

Z-square

also known as O-mean method.

Student	Height (in inches)
1	64
2	70
3	72
4	68
5	76

$$Z' = \frac{x - \text{mean}(\mu)}{\text{Standard deviation}(\sigma)}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$

x_i → Term given in data

\bar{x} → mean

N = total no. of terms

$$\mu = \frac{\sum x_i}{n} \quad (i \text{ from } 1 \text{ to } N)$$

and

$$\sigma = \sqrt{\frac{(64-70)^2 + (70-70)^2 + (72-70)^2 + (68-70)^2 + (76-70)^2}{5}} \\ = 4$$

then,
student

student	Height	x'
1	64	$x' =$
2	70	$x' =$
3	72	$x' =$
4	68	$x' =$
5	76	$x' =$

WJMK
लूप्प

कालत्सु कारेचरित्त यनर मरिष्ट,
नियमुक्ते कलेवग आमरष्ट॥०७॥
WJMK
लूप्प

Home work Solution :-

~~Support~~

[LIFT] :-

- it is a simple correlation measure
- The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B) = P(A) \cdot P(B)$, otherwise both itemsets are dependent and correlated.

$$\text{lift}(A, B) = \frac{P(A \cap B)}{\frac{P(A) \cdot P(B)}{P(A \cup B)}}$$

→ If $\text{lift}(A, B) < 1$, occurrence of A is negatively correlated with occurrence of B, meaning that occurrence of one likely leads to the absence of another one.

→ If $\text{lift}(A, B) > 1$, then A and B are positively correlated, meaning occurrence of one implies occurrence of other.

→ If $\text{lift}(A, B) = 1$, then A, B are independent and there is no correlation between them.

for example:- Total transactions analyzed = 10,000
transactions which include computer games = 6000
transactions which include vedios = 7500
transactions which include both = 4000
say, min-support = 30%
min-confidence = 60%

WJMK
2023

WJMK
2023

Then following association rule can be discovered :-

buys(X, "games") \Rightarrow buys(X, "videos")

$$\text{as, support} = \frac{4000}{10,000} = 40\% \geq 30\%$$

$$\text{confidence} = \frac{4000}{6000} = 66\% \geq 60\%$$

but this is misleading because probability of purchasing videos is 75% $> 66\%$.

Infact games and videos are negatively associated because purchase of one of these items decreases likelihood of purchasing the other.

So here, confidence is deceiving and does not measure real strength (or lack of strength) of correlation and implication b/w A and B.
Hence, alternative required.

As,

$$P(\{\text{game}\}) = 0.6$$

$$P(\{\text{video}\}) = 0.75$$

$$P(\{\text{game, video}\}) = 0.4$$

$$\text{lift}(\text{game} \Rightarrow \text{video}) = \frac{0.4}{0.6 \times 0.75} = 0.89 < 1 \rightarrow (\text{negative correlation})$$

$\boxed{\chi^2}$:-

The second correlation measure is χ^2 measure.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

for example :-

Observed :-

	game	game	Σ_{row}
video	4000	3500	7500
video	2000	500	2500
Σ_{col}	6000	4000	10,000

Expected :-

	game	game	Σ_{row}
video	4500	3000	7500
video	1500	1000	2500
Σ_{col}	6000	4000	10,000

Now,

$$\begin{aligned} \chi^2 &= \frac{(4000 - 4500)^2}{4500} + \frac{(3500 - 3000)^2}{3000} + \frac{(2000 - 1500)^2}{1500} \\ &\quad + \frac{(500 - 1000)^2}{1000} = 555.6 \end{aligned}$$

Now, $\chi^2 > 1$ and observed value of slot (game, video) = 4000, < 4500 (expected), so game and video are negatively correlated.

- ① So, χ^2 measure, test if two categorical variables are independent or correlated.

WJMK
= 3.
=

χ² test checks whether observed values in a dataset are different from expected values; if there was no relationships.

- # High $\chi^2 \rightarrow$ Items are highly correlated.
- # Low $\chi^2 \rightarrow$ ~~Observed~~ means if χ^2 close to 0, observed values match the expected values, meaning no correlation.
- # If ~~observed~~: χ^2 large, > 0 , and observed value is less than expected value \Rightarrow negatively correlated.
- # If χ^2 large, > 0 and Observed $>$ expected \Rightarrow positively correlated.

→ Support tells us how frequently an itemset appears,
→ confidence tells us how strong rule is,

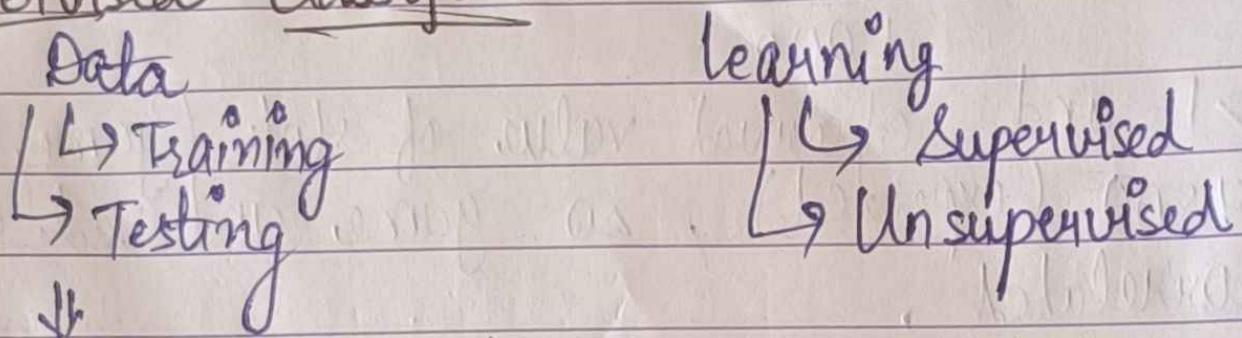
confidence. tells us how strong

25/2/25
Tuesday

Lec-5 :-

Classification :- Decision Tree :-

* Supervised Classification :-



based on data, domain etc., we can decide the ratio.

→ Training data jyada rkha hai to train model in nice manner, but kitne %, depends on total data we've,

agar

If total data = 1 MB \rightarrow then 80% being training mem,
then very less.

But,

If total data = 1 TB, then 80% is more than enough.

Decision Trees:- (DT)

- ↳ classification scheme
- ↳ represents a model of different classes
- ↳ generates tree and set of rules
- ↳ a node w/o children is leaf node,
o/w internal node

each internal node has associated splitting predicate eg: binary predicates,

↳ eg: Age $<= 20$.

profession in { student, teacher }.

3* salary - 10000 > 0

grades	marks
A +	> 90
A	> 80
B +	$8 < 90$
B	:
nothing but "class"	:

* "Sample DT" might be DT of ~~training~~ data, but not necessarily to be DT.

NOMK
= 35/50

NOMK
= 35/50

* Leaf node \rightarrow classes / rule

[* Root + internal node \rightarrow feature / attributes

* edges \rightarrow predicate / cond's

\rightarrow How to decide which node is root? and kis feature ko konsa node dena hain?????

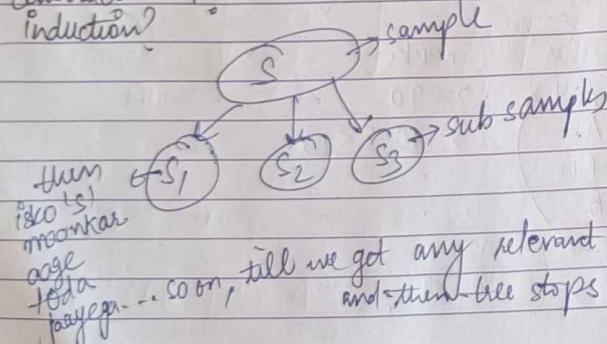
\hookrightarrow sare features whi relate hm DT, sinf wo rkhte hain which has high feature importance, means those which play significant role in decision making

\rightarrow and then we will make their priority list.

\rightarrow This all is part of preprocessing.

* Again distinct values tam hain, then sbki ek-2 branch bna skte hain, o/w do grouping.

How to construct DT?
DT induction?



Entropy:- (Info theory)

is measure of uncertainty associated with a random variable

$$H(Y|X) = \sum_x p(x) H(Y|X=x)$$

Conditional Entropy

* Attribute Selection Measure

\hookrightarrow Information Gain ($C_{ID3}/C_{4.5}$)

\rightarrow Select attribute with highest info gain,

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

where,

p_i is probability that an arbitrary tuple in D belongs to C_i estimated by $\frac{|C_i \cap D|}{|D|}$.

(Expected info(entropy)
needed to classify a
tuple in D)

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

Info needed to classify D after using A split to split D into m partitions

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

\uparrow
Info gained by branching on attribute A.

* Gain for continuous valued attributes !

→ firstly sort data / attributes in ascending order.

Gain Ratio for attribute selection (C4.5) :-

$$\text{Split Info}(A) = \sum_{j=1}^v \frac{|A_j|}{|D|} \log_2 \left(\frac{|A_j|}{|D|} \right)$$

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{splitInfo}(A)}$$

→ attribute with max gain ratio is selected as splitting attribute.

* what are special cases, where we can be in problem in ID3?

temp given	grouped
65	= < 75
70	= < 75
76	> 75
78	> 75
60	= > 75

→ yehi kita TD3,
but can't
group key based
attributes
like your roll no,

so then TD3 fails

and then only gain ratio kaam

aa skti hai, i.e. C4.5.

* Gini Index (CART, IBM Intelligent Miner) :-

$$\text{gini}(D) = 1 - \sum_{j=1}^n p_j^2$$

where p_j is relative frequency of class j in D
and total there are n classes.

WJMK
= GINI

$$\Delta \text{gini}(A) = \text{gini}(A) - \text{gini}(D)$$

* gini index is index of impurity
(so hm usko minimum se minimum lekhne
ki koshish karte hain)

- biased to multivalued attributes.
- has difficulty when # of classes is large
- tends to favor tests that results in equal sized partitions.

Other attribute selection measures:-

→ CHAID, C-SEP, G-statistic, (MDL), CART,
multivariate splits

↓
quite
popular

Which attribute selection measure is best?

based on data, domain and features.

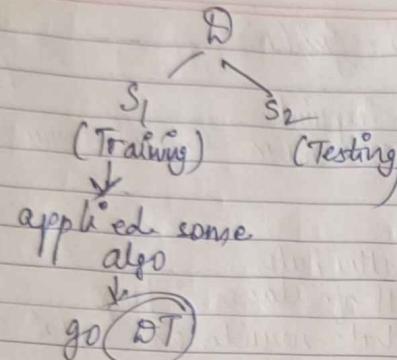
ID3 → when no noise, no key
so on!!!!

Overfitting the Data → when BT is built, many of the branches may reflect anomalies in the training data due to noise or outliers.

→ we may grow the tree just deeply enough to perfectly classify the training data set.

→ This problem is known as "overfitting the data".

NJM&
SOM



- 8 ways to avoid overfitting:-
- (1) during AT construction
Pre-pruning
 - (2) after proper bgya,
fir tree ko scan kro
↓
post-pruning

problems of overfitting :- ?

Underfitting :-

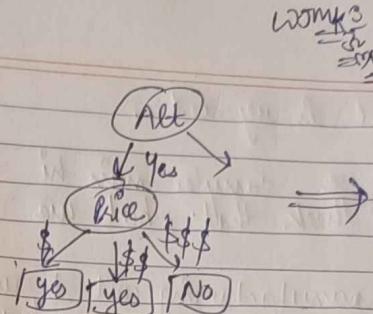
- 1. Usually we don't go for it, as we've to track things so on.
- tree is pruned by halting the tree construction i.e. by deciding not to further split on partition the subset of training samples at a given node.
- Upon halting, a node becomes a leaf node.
- leaf may hold the most frequent class among the subset samples.

Post pruning:-

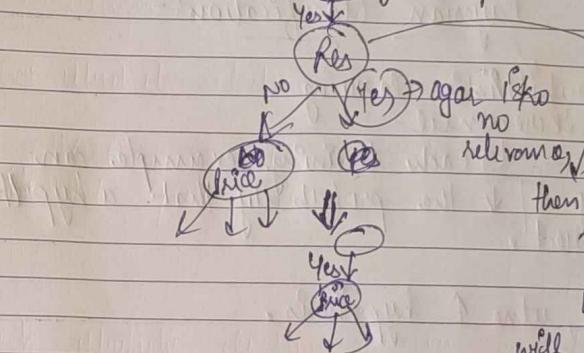
- more easy to implement as have entire tree with us, now scan and see whether we can prune something or not.

Two methods :-

- (1) Subtree replacement replaces subtree with a single leaf node,



- (2) Subtree raising - moves subtree to a higher level in decision tree, subsuming its parent.



then also no relevance, so remove

because 'Yes'

will create problem in testing, may give less accuracy, while may be good at case of training.

Bayesian Classification :-

Bayes Thm

jab shoes/sandals beche jada hain on discount even w/o any defect is based on bayes theorem because companies works on lot, women se kuch samples nikal ke discount p beche, defect dekhe, agar threshold niche data gya

than decided, then whole lot is rejected.

↓
chale bari ki mat se samples mein hi
khraab tha, baaki theek tha, but we
reject (i).
→ issi liye jin shoes ki manufacturing cost ↓, unk
bhk ↑ rate par bckta hain, becz unhone plots
ko reject kya hota hai.
(bhut se)

Bayesian Classification:-

- probabilistic learning
- Incremental

↳ each training example can
incrementally ↑ ↓ the probability that a hypothesis
is correct.

→ fitni ↑ classes, utne ↑ hypothesis.

Estimating probabilities:-

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

Naive Bayesian classification:-

① class conditional Independence
↳ also called simple BC,

↳ This assumption simplifies computations.

From Bayes theorem:-

$$P(C_i|X) = \frac{P(X|C_i) P(C_i)}{P(X)}$$

* $P(X)$ is constant.

only $P(X|C_i) \cdot P(C_i)$ needs to be
maximized.

Naive Assumption:-

$$P(X|C_i) = \prod_{k=1}^n P(x_k|c_i)$$

→ This algo may fail also. (H/w)

(Naive Bayes classification)

What ~~is~~ remedies
to do, so that
Naive Bayes classification
can be applied.

$$P(C_i^o | X) = \frac{P(X | C_i^o) P(C_i^o)}{P(X)}$$

* $P(X)$ is constant.

only $P(X | C_i^o) \cdot P(C_i^o)$ needs to be maximized.

Naive Assumption :-

$$P(X | C_i^o) = \prod P(x_k | C_i^o) \text{ over } k=1 \text{ to } n.$$

→ This algo may fail also. (H/W)

(Naive Bayes classification) → what ~~is~~ remedies to do so that Naive Bayes classification can be applied.

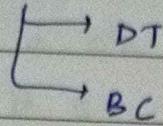
85/3/2025

Tuesday

Lec-6

:-

Classification



→ BP

→ SVM

→ LR

→ KNN

} Test (next week)

Clustering

Boosting algorithms

Feature selection

Data clustering : Unsupervised learning.



↳ Class label not known

Partition based clustering

→ K-means

→ PAM

→ K-Means++

→ Clara

→ Clarans

} Variations
in K-mean as K-mean is
not good for large set of
data.

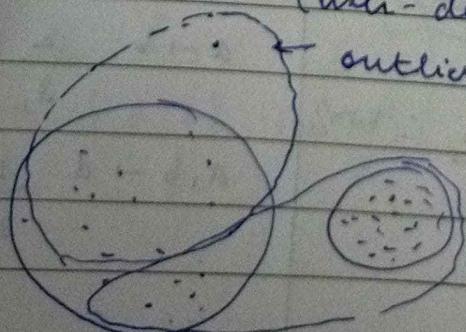
- ↳ might not give proper clusters
- ↳ lacks accuracy & efficiency
- ↳ Sparse matrix

Problem: Outliers are also part of a cluster

k: number of clusters

(user-defined parameter)

outlier



Due to presence of
outlier means
converge to a
value impacted
largely by one outlier

Binary variables:

a	11
b	10
c	01
d	00

Both objects for a particular instance have 1.

Symmetric:

$$d(i, j) = \frac{b+c}{a+b+c+d} = \frac{\underbrace{10+01}_{\text{1 for } x_1, 0 \text{ for } x_2}}{\underbrace{11+01+10+00}_{11+01+10+00}}$$

Asymmetric: $d(i, j) = \frac{b+c}{a+b+c}$

How to know if variable is sym/asym.

If $\begin{cases} 10 \rightarrow 1 \\ 2 \rightarrow 1 \rightarrow 0 \end{cases} \Rightarrow$ symmetric binary variable.

If $10 \neq 1$ then asymmetric binary variable.

Eg: No.: $N \rightarrow 0$

$Y_{12}: Y \mid \text{Presence: } 1 \rightarrow 1$

no instance
for 10.

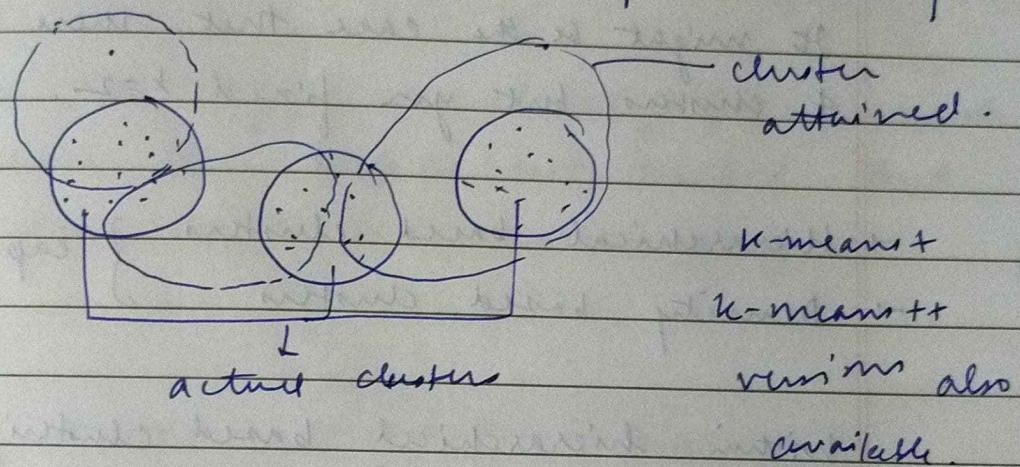
$$d(jack, Mary) = \frac{10+01}{11+01+10} = \frac{0+1}{2+1} = \frac{1}{3} = 0.33$$

The points which were initially part of cluster 1 can now be part of cluster 2.

Large data \Rightarrow Large noise

\Rightarrow Mean affected largely

\therefore The actual clusters may not be identified which impacts accuracy.

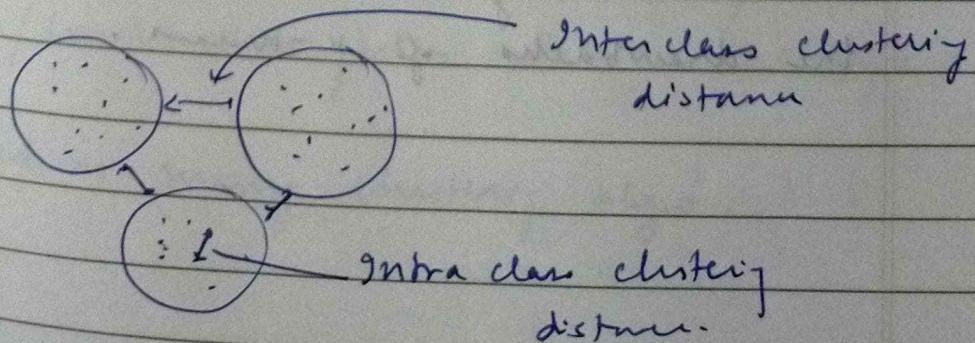


To apply k-means on a large data, remove noise from the data.

\Rightarrow can be used in Text mining, etc.

Data clustering: grouped based on similarity measure (could be distance). Can be on multi-dimensional data.

OPTIMISATION PROBLEM



Data types

- Interval based
- Ratio based
- Binary
- Ordinal
- Nominal



Real data has
to be transformed
into a data of
one of these types
before applying
clustering

Data from kaggle is already pre-processed
& hence this step is not needed. But
raw data must be pre-processed first

Data matrix : attributes in columns
entities in rows.

$$\text{Eg. } \begin{matrix} A_1 & A_2 & \dots & A_n \\ \left[\begin{array}{cccc} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{array} \right] \end{matrix} \quad \text{Object by variable.}$$

Dissimilarity metric:

Object by object : Distance b/w all objects
& all other objects.

Upper/lower triangular matrix is enough.

Used in many clustering algs.

Clustering requires 2 optimisations at the same time:

- (1) Inter class distance, ^{of clusters} should be maximised
- (2) Minimise intra class ^{data} _{distance of clusters}

Problem with K-means : Fixing k

It might be the case that there are actually 4 clusters but you fixed $k=2$.

- Hierarchical based clusters } capable of doing
- Density based clusters

Within hierarchical based clustering

we have :

single link

heavy link

complete link

Birch

Density based : → CURE

→ DBSCAN

→

3 researchers
→

BFR : Primarily uses K-means & overcomes the drawbacks of K-means.

Smaller unit \rightarrow larger variable range

To select basketball players based on height, we want clear distances in the differences due to closeness in height.

\therefore Lower down the unit -

Give more weightage to height.

\therefore Convert Euclidean into ~~weight~~ weighted distance -

$$w(h_1 - h_2)$$

↓

weight assigned to assign higher weight to height.

Suppose 3 attributes: A₁, A₂, A₃

assign higher weight to the one with highest importance

$$w_1 > w_2 > w_3$$

→ Mean absolute deviation better than standard deviation due to absolute values.

Assigning weights to Minkowski distance would give weighted distance:

$$\sqrt[m]{w_1|x_1 - \bar{x}|^m + \dots}$$

Not necessary to assign weight to all attributes

कालिंग, कारेचरिन्त रुपरे ग्रिंडा,
निवासुंदर कालिंग अस्ट्रिंडा।
WJMK & WOKS
३००१. ३००२.

HomeWork :- Nominal, Ordinal, Ratio Data

(1) Nominal Data :- (categorical - Unordered)

→ Nominal data represents categories or labels without any intrinsic order or ranking.

Key characteristics :-
(1) Purely qualitative
(2) No meaningful numerical value
(3) Can not be ordered or ranked
(4) mode is the only central tendency that can be used.

Examples :- (1) gender :- male, female, other

(2) Blood type :- A, B, AB, O

(3) marital status :- single, married, divorced

(4) Country name :- India, USA, Japan

(2) Ordinal Data :- (Categorical - Ordered)

→ Ordinal data represents categories that have a natural order, but the intervals between them are not necessarily equal or known.

Key characteristics :-
(1) Qualitative but with order.
(2) Ranks can be compared (eg. higher or lower), but not quantified.
(3) Median and mode can be used.

Examples :-

1) Educational level : Higher school < Bachelor < master or PhD

WJMK = INOM.

(2) Customer Satisfaction :- Very Unsatisfied < Unsatisfied < Neutral < Satisfied & very satisfied

(3) Economic class:-

low income < middle income < high income

(3) Ratio Data :- (Quantitative absolute zero)

Ratio data is numerical data with a true zero point, meaning that zero indicates a complete absence of the quantity.

Key characteristics:-

(1) Quantitative and continuous

(2) Has absolute zero (e.g. 0kg = no weight)

(3) All mathematical operations (addition, multiplication etc.) are valid

(4) mean, median, mode and standard deviation can be used.

Example:-

(1) Weight : 0kg, 50kg

(2) Height : 0cm, 180cm

(3) Age : 0 years, 25 years.

(4) Income : ₹0, ₹10,000

8/4/25
Tuesday

[UC-7] :-

• Nominal variables:-

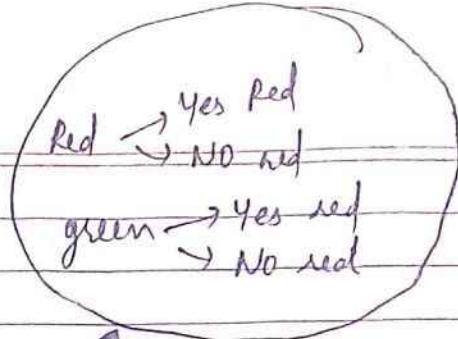
→ generalization of binary variable in that it can take more than 2 states e.g. red, yellow, blue, green.

Method-1 :-

m : - no. of matches

p : - total no. of variables

$$d(i, j) = \frac{p-m}{p}$$

Method-2 :-

use a large no. of binary variables.

↓
creating new binary variable for each of M nominal variable.Ordinal Variables :-

An ordinal variable can be discrete

or continuous, where order (rank) is important.

→ can be treated as interval-scaled;

→ replacing x_{if} by their rank

$$r_{if} \in \{1, \dots, M_f\}$$

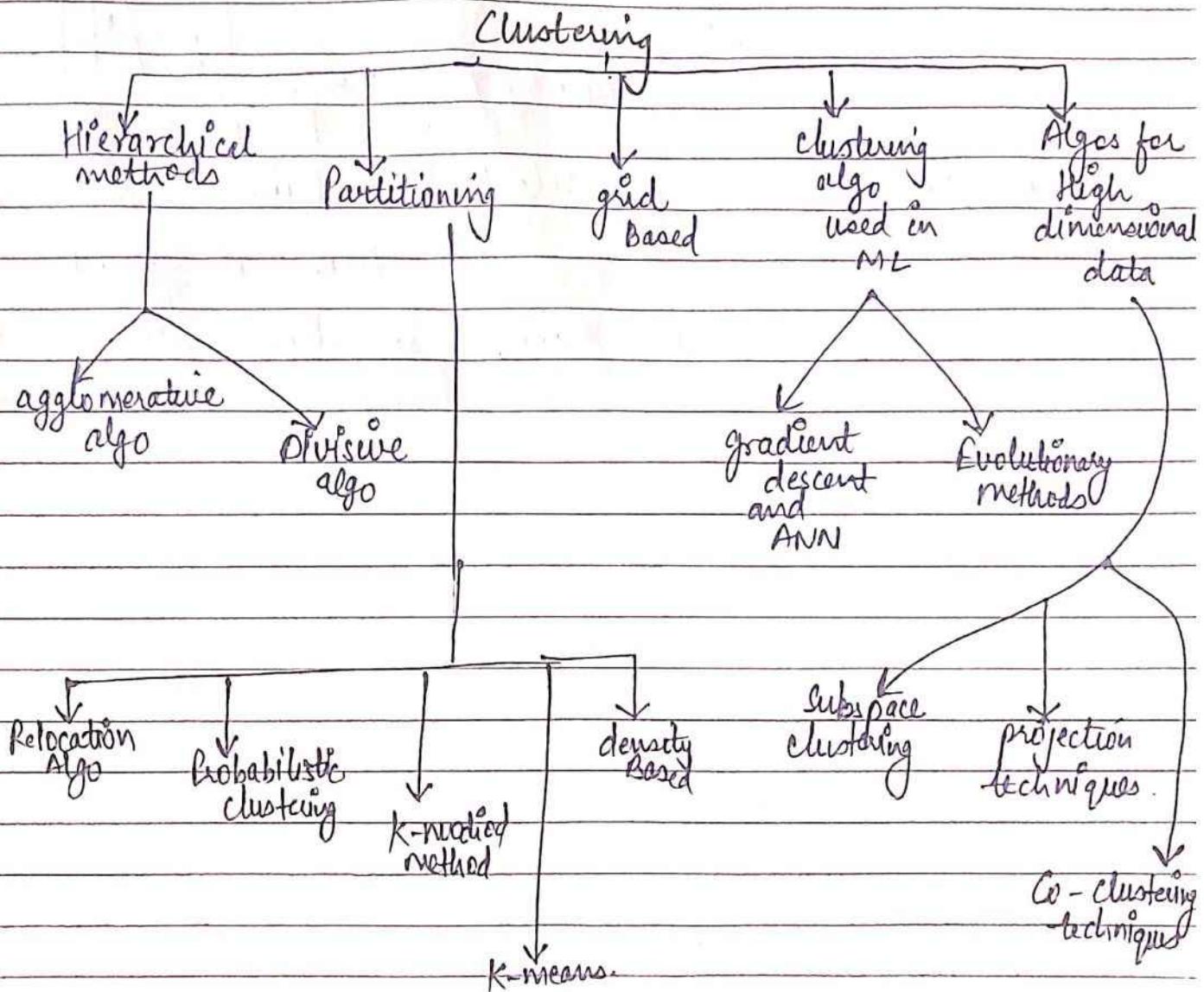
→ map range of each variable onto $[0, 1]$ by replacing i^{th} object in f^{th} variableRatio-Scaled variables:-

↳ A positive measurement on a non linear scale, approximately at exponential scale, such as:-

$$Ae^{Bt} \text{ or } Ae^{-Bt}$$

↳ Sometimes we want mixed variables, mix of ^{range} two or all 3 variables,

* Types of clustering algorithm:-



* Distance b/w clusters:-

If $|p - p'|$ is distance b/w 2 pts / objects.

$m_i^o \rightarrow$ mean of cluster C_i^o

$n_i^o \rightarrow$ no. of objects in C_i^o

$$d_{min}(C_i^o, C_j^o) = \min_{p \in C_i^o, p' \in C_j^o} |p - p'|$$

$$d_{max}(C_i^o, C_j^o) = \max_{p \in C_i^o, p' \in C_j^o} |p - p'|$$

$$d_{mean}(C_i^o, C_j^o) = |m_i - m_j|$$

$$d_{avg}(C_i^o, C_j^o) = \left(\frac{1}{n_i \cdot n_j} \right) \left(\sum_{p \in C_i^o} \sum_{p' \in C_j^o} |p - p'| \right)$$

K-means :-
no. of clusters

k pts. select kme. b/w

select jthmth data cluster points
k-means, iterations until
kam "honge".

Disadvantages :-

- (1) K should be known in prior
- (2) outliers

so for k-means, data should be noise
free, only then k-means \Rightarrow se satith
output aayega.

CLARA :- (clustering large applications)

PAM :- Partition around medoids

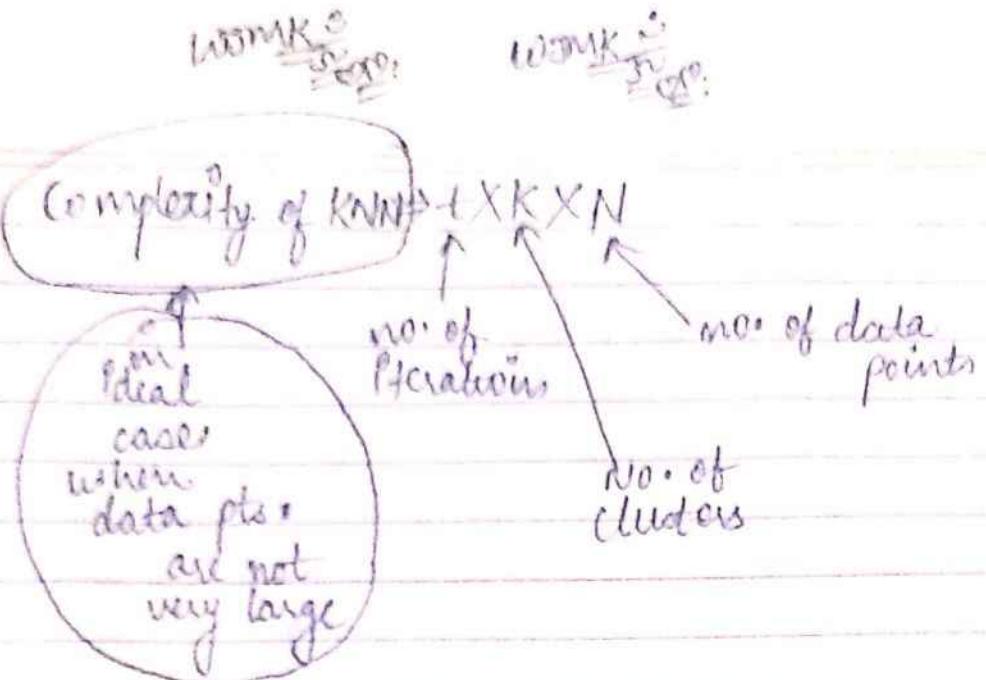
solves some drawbacks of k-means
 \rightarrow 4 cost fn's
 \hookrightarrow to see ptth is medoid or
non-medoid

How \rightarrow How PAM works

& 2 variations of k-means and PAM

CLARA CLARANS

\hookrightarrow "randomized" CLARA



Q. Why to consider neighbouring pts. not far.
 A. \uparrow better clustering \uparrow (sparse data)

Hierarchical clustering:-

↳ clusters are created in levels, actually creating sets of clusters at each level.

↳ AGNES

(Agglomerative AGNES)

↓

Bottom-up
Approach

DIANA.

(Dissimilarity Analysis)

↓
Top-Down Approach

(1 2 3 4 5)

AGNES

grouping

kind of data



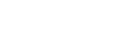
Divisive → dividing

data on

basis of

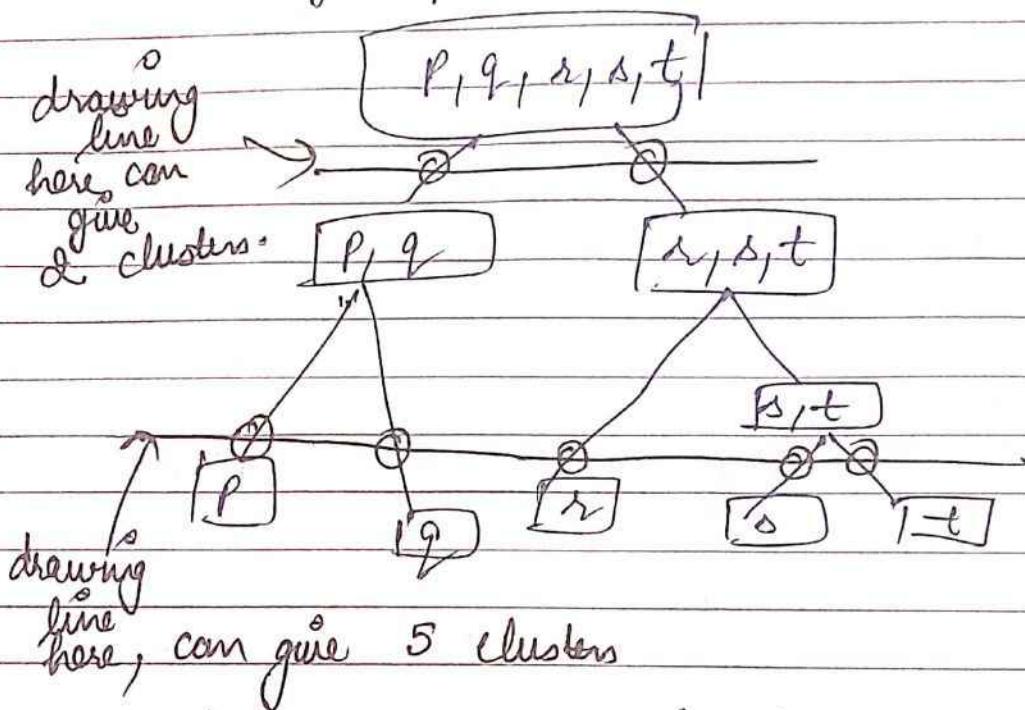
similarity

criterion.



WJM K^o
5/2020

* clusters always depend on our requirement



→ like we draw special type of graphs
"dendograms".

Eg: (1) BIRCH

(2) MST

~~H/W Answer:-~~

* How does PAM works?

Algorithm : k-medoids PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input :

k: no. of clusters

D: a data set containing n objects

Output:

A set of K clusters

Method:

- (1) arbitrarily choose k objects in \mathbb{D} as the initial representative objects or seeds;
- (2) repeat
- (3) assign each remaining object to the cluster with the nearest representative object;
- (4) randomly select a non-representative object, O_{random} ;
- (5) compute the total cost, S , of swapping representative object, O_j , with O_{random} ;
- (6) if $S < 0$ then swap O_j with O_{random} to form the new set of k representative objects;
- (7) until no change;

EXAMPLE OF PAM:-

$$\mathbb{D} = \{1, 3, 4, 10, 11, 12, 25\}$$

$$k=2$$

Step 1 :- Initial medoids $\rightarrow 1$ and 25

Step 2 :- Assign each point to the nearest medoid.

WDMK 20%

Point	Dist to 1	Dist to 25	Assigned to
1	0	24	1
3	2	22	1
4	3	21	1
10	9	15	1
11	10	14	1
12	11	13	1
25	24	0	25

cluster 1 (mediod 1) :- {1, 3, 4, 10, 11, 12}

cluster 2 (Mediod 25) :- {25}

$$\text{Total cost} = 0 + 2 + 3 + 9 + 10 + 11 + 0 = 35$$

Step 3:- Try Swapping

Try swapping $25 \leftrightarrow 3 \rightarrow$ new mediods: 1 and 3

Point	Dist to 1	Dist to 3	Assigned to
1	0	2	1
3	2	0	3
4	3	1	3
10	9	7	3
11	10	8	3
12	11	9	3
25	24	22	3

Cluster 1: {1} \rightarrow cost = 0

Cluster 2: - {3, 4, 10, 11, 12, 25}

$$\text{cost} = 0 + 1 + 7 + 8 + 9 + 22 = 47$$

worse than $\uparrow 35$
so, reject.

work = 30%

Try Swapping $25 \leftrightarrow 11 \rightarrow$ medoids: 1 and 11

Point	Dist to 1	Dist to 11	Assigned to
1	0	10	1
3	2	8	1
4	3	7	1
10	9	1	11
11	10	0	11
12	11	1	11
25	24	14	11

$$\text{Cluster 1: } \{1, 3, 4\} \rightarrow \text{Cost} = 0 + 2 + 3 = 5$$

$$\text{Cluster 2: } \{10, 11, 12, 25\} \rightarrow \text{Cost} = 1 + 0 + 1 + 14 = 16$$

$$\text{Total cost} = 5 + 16 = 21 \rightarrow \text{Better} \rightarrow \text{Accept}.$$

Try swapping $1 \leftrightarrow 3 \rightarrow$ medoids: 3 and 11

Point	Dist to 3	Dist to 11	Assigned to
1	2	10	3
3	0	8	3
4	1	7	3
10	7	1	11
11	8	0	11
12	9	1	11
25	22	14	11

$$\text{Cluster 1: } \{1, 3, 4\} \rightarrow \text{Cost} = 2 + 0 + 1 = 3$$

$$\text{Cluster 2: } \{10, 11, 12, 25\} \rightarrow \text{Cost} = 1 + 0 + 1 + 14 = 16$$

$$\text{Total cost} = 19 \rightarrow \text{Better} \rightarrow \text{Accept}.$$

Try other swaps \rightarrow No improvement \rightarrow Stop

ANAL RESULT:-

Final mediods: 3 and 11
 Clusters:-

$$\text{Cluster 1} : \{1, 3, 4\}$$

$$\text{Cluster 2} : \{10, 11, 12, 25\}$$

Final total cost = 19.

15 April 2025
 Tuesday

i	x	y	
x_1	2	6	
x_2	3	4	
x_3	3	8	
x_4	4	7	
x_5	6	2	
x_6	6	4	
x_7	7	3	
x_8	7	4	
x_9	8	5	
x_{10}	7	6	

[lec-8] :-
 [PAM] :-

manhattan distance of
 (x, y) from C_1

$c_1 \leftarrow$ manhattan distance of
 (x, y) from C_2

$c_2 < c_1$
 so in cluster C_2

Step 1:- select random points (mediods)

$$C_1 : (3, 4)$$

$$C_2 : (7, 4)$$

Step 2:- find manhattan distance from both C_1 and C_2 .

Step 3:- Assign to cluster jisse distance km.

WORK
W^{ED}

\vec{x}	cluster
x_1	C ₁
x_2	C ₁
x_3	C ₁
x_4	C ₁
x_5	C ₂
x_6	C ₂
x_7	C ₂
x_8	C ₂
x_9	C ₂
x_{10}	C ₂

+ by finding distance ^{going} and
adding ^{return} median time

Step 4:- find clustering cost

\rightarrow 1st cluster has 4 points
 \rightarrow 2nd " " 6 points

cluster 1:- { (2, 6), (3, 4), (3, 8), (4, 7) }

(order is
~~not~~ very important)

$$\text{Cluster 2} = \{(6, 2), (6, 4), (7, 3), (7, 4), (8, 5), \\ \rightarrow |3-2| + |4-6|\} \quad (7, 6)\}$$

$$\begin{aligned}
 \text{Cost} = & \left\{ \text{cost}((3, 4), (2, 6)) + \cancel{\text{cost}((3, 4), (2, 7))} + \right. \\
 & \quad \text{cost}((3, 4), (3, 8)) + \text{cost}((3, 4), (4, 7)) \cancel{+} \\
 & \quad \text{cost}((7, 4), (6, 2)) + \text{cost}((7, 4), (6, 4)) + \\
 & \bullet \quad \text{cost}((7, 4), (7, 3)) + \text{cost}((7, 4), (8, 5)) + \\
 & \quad \left. \text{cost}((7, 4), (7, 6)) \right\} \\
 = & 3 + 4 + 4 + \cancel{10} + \cancel{10} + \cancel{10} + 3 + 1 + 1 + 2 + 2 \\
 = & 20
 \end{aligned}$$

WOMKE
340%

Step 5:- new random mediod
 $C_3 = (7, 3)$.

x_i	c_1	c_3
x_1	3	8
x_2	0	5
x_3	4	9
x_4	4	7
x_5	5	2
x_6	3	2
x_7	5	0
x_8	4	1
x_9	6	3
x_{10}	6	3

$$\begin{aligned} \text{cost} &= 3+4+4+2+9+1+ \\ &= 3+3 \end{aligned}$$

we want optimal cluster.

Now,

$$\begin{aligned} S &= \text{current cost} - \text{previous cost} \\ &= 22 - 20 \\ &= 2 > 0 \end{aligned}$$

↑
 $\text{so } (7, 3) \text{ is not}$

optimal mediod as it mein computational
 cost / time jyada hoga.

so, optimal mediods $\rightarrow (3, 4) \& (7, 4) \checkmark$
 $(7, 3) \times$

Backpropagation:-

Q. Assume that the neurons ~~are~~ have "sigmoid activation" function, perform a forward pass and a backward pass on the network. Assume that the actual output of y is (0.5) and learning rate is (1) .

→ In ANN

we can go only forward, but if "only" that can correct weights / perception itself is called backpropagation, and then again forward pass, then such cycle called chain, '1 cycle' is known as '1 epoch'

Perform another forward pass.

$$x_2 = 0.3$$

$$w_{13} = 0.1$$

$$w_{14} = 0.4$$

$$w_{24} = 0.6$$

$$w_{23} = 0.8$$

$$w_{35} = 0.3$$

$$w_{45} = 0.9$$

~~a₄~~ $a_4 = (w_{14} * x_1) + (w_{24} * x_2)$
 $= (0.4 * 0.33) + (0.6 * 0.9)$
 $= 0.14 + 0.54$
 $= 0.68$

so,
 $y_4 = \frac{1}{1 + e^{-0.68}} = \frac{1}{1 + 0.5066} = 0.6637$
 $= \frac{1}{1.5066} = \cancel{0.6637}$
~~0.6637~~

Similarly for y_5 :-

~~a₅~~ $a_5 = (y_3 * w_{35}) + (y_4 * w_{45})$
 $= (w_{35} * y_3) + (w_{45} * y_4)$
 $= (0.3 * 0.68) + (0.9 * 0.6637)$
 $= 0.801.$

so,
 $y_5 = \frac{1}{1 + e^{-0.801}}$
 $= \frac{1}{(1+0.4488)} = \frac{1}{1.4488}$
 $= 0.69.$

↑
N/w output.

Error = $y_{target} - y_5$
 $\Rightarrow (0.5) - (0.69)$

here subtract like error nikalao, as we are not given ki error kaise nikalna hai.

WORK
TOP

→ Usually 'mean square error' waja likha hoga, tab hm woh use kenge.

Ans:

* Now for Backpropagation :-

Each weight is changed by :-

$$\Delta w_{ji} = \eta (s_j) o_i$$

\uparrow
learning rate

← change in weight.

$$s_j = o_j (1 - o_j) (t_j - o_j) \quad (\text{if } j \text{ is output unit})$$

$$s_j = o_j (1 - o_j) \sum s_k w_{kj} \quad (\text{if } j \text{ is hidden unit})$$

Where,

η = learning rate

t_j = correct output for unit j

s_j = error measured.

* Backward pass :-

for o_5 ,

$$s_5 = y_5 (1 - y_5) (y_{\text{target}} - y_5)$$

$$= 0.69 (1 - 0.69) (0.5 - 0.69)$$
$$= -0.0406$$

Now, for hidden unit 3 -

$$s_3 = y_3(1-y_3)(w_{35} * s_5)$$

$$= 0.68 * (1-0.68) * (0.3 * (-0.0406))$$

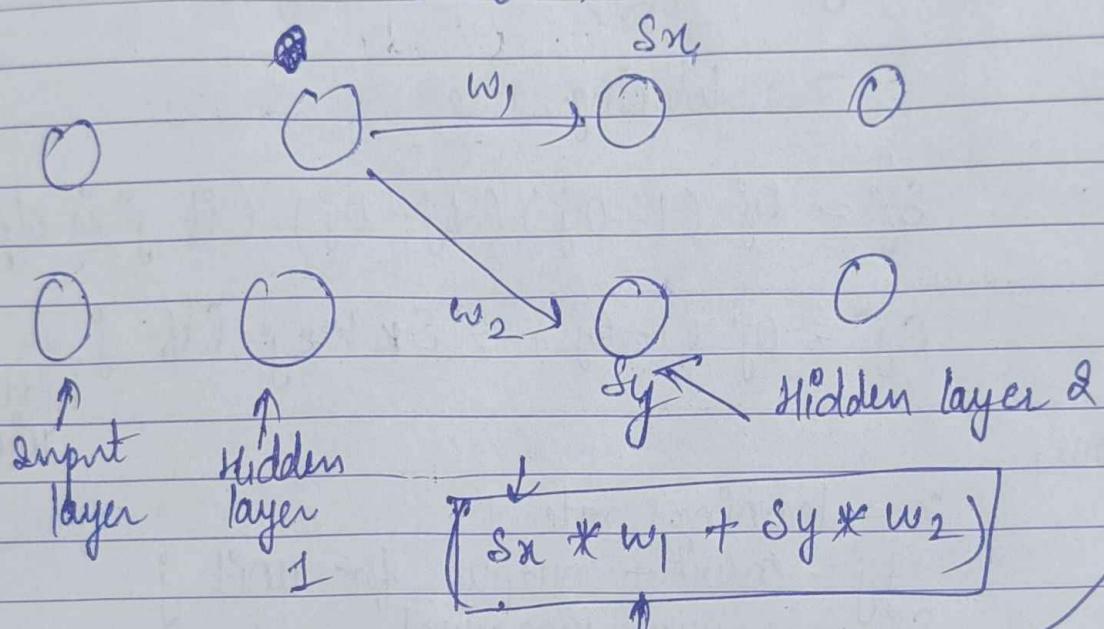
$$= -0.00265$$



$$s_4 = y_4(1-y_4)(w_{45} * s_5)$$

$$= (0.6637)(1-0.6637)(0.9 * (-0.0406))$$

$$= -0.0082$$



Isi peye, 1' ukha in formula.

Now these above are 8 terms.

change in weights:-



WORK
-058

$$\begin{aligned}\Delta w_{45} &= \eta s_5 y_4 \\ &= 1 * (-0.0406) * (0.6637) \\ &= -0.0269\end{aligned}$$

Now, ~~w₄₅ = 0.9~~

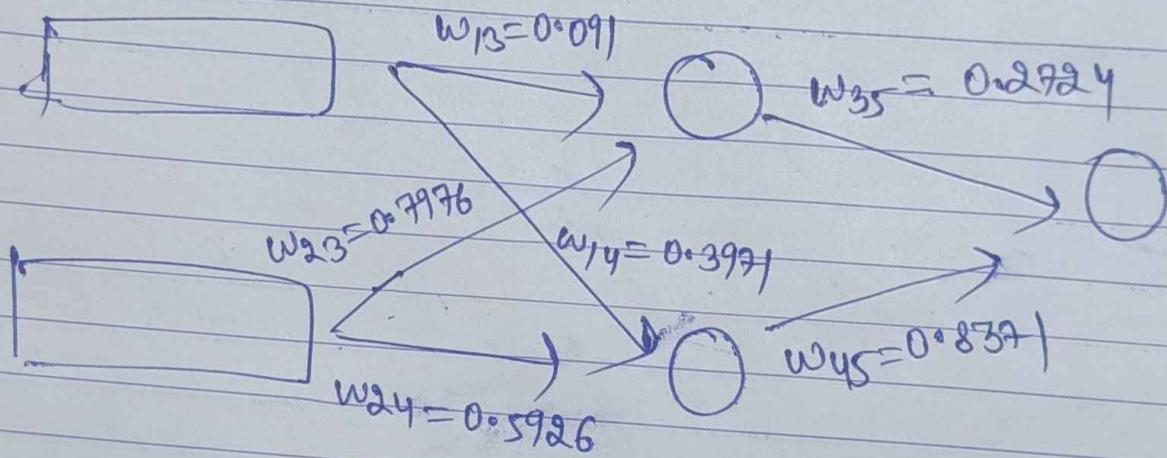
$$\begin{aligned}w_{45} &= \Delta w_{45} + w_{45}(\text{old}) \\ &= -0.0269 + 0.9 \\ &= 0.8731\end{aligned}$$

similarly,

$$\Delta w_{14} = -0.00287$$

$$\begin{aligned}\text{New } w_{14} &= 0.4 + (-0.00287) \\ &= 0.3971.\end{aligned}$$

when we will update all weights:-



Now, again we'll do the forward pass to see the change! -

WORKING

For this :-

$$a_j^o = \sum_i (w_{ij}^o * x_i^o)$$

$$y_i^o = \text{sigmoid}(a_j^o)$$

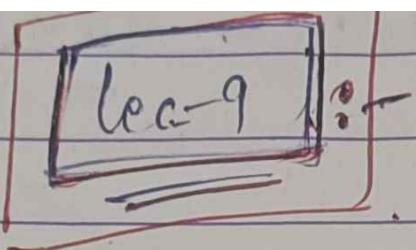
$$y_3 = 0.6797$$

$$y_4 = 0.6620,$$

$$y_5 = 0.6820, \rightarrow \text{final output}$$

In \downarrow one pass we can see that error has reduced a little.

~~29/4/25~~
Tuesday



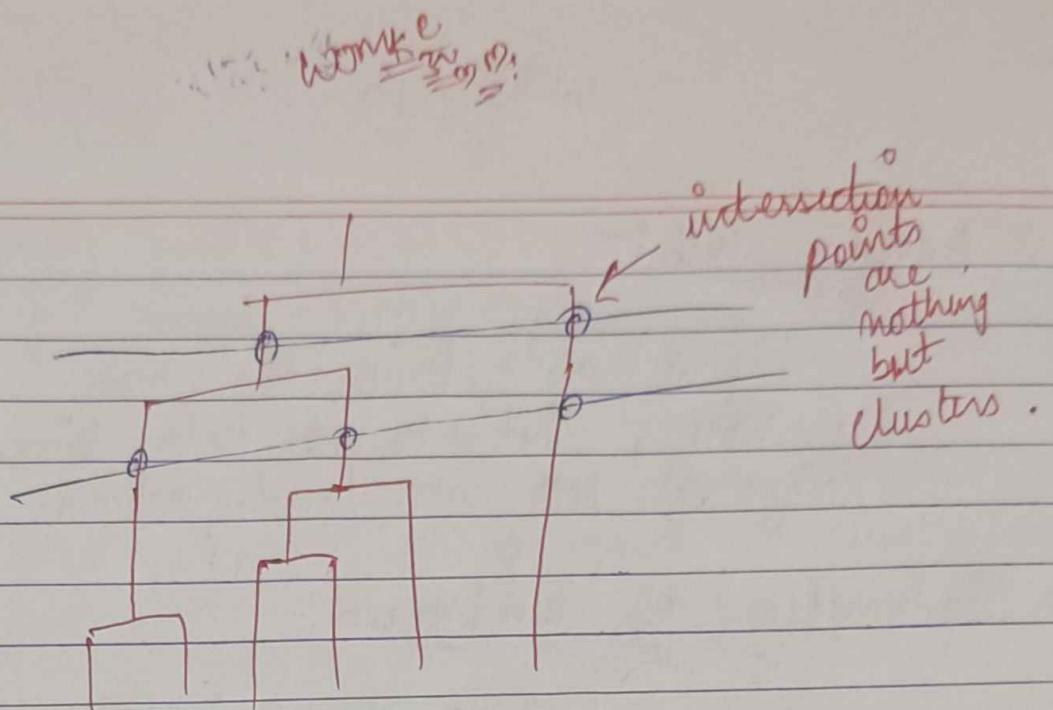
* Hierarchical Algorithms :-

all such
algos make
use of
special type
of graphs

known as Dendograms

- ① single link
- ② MST single link
- ③ complete link
- ④ Average link

actually enables you to get
cluster out of it,

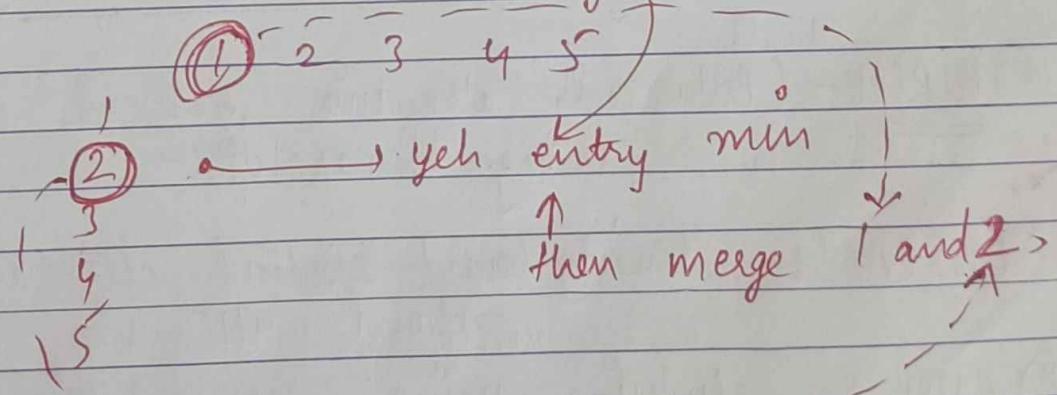


* Single link :- very fast wst computations

$D(i, j) \rightarrow$ dissimilarity matrix

usually hm lower Δ_L yaa upper Δ_U
lete ham bs, but you can work with complete
matrix as well.

→ har vaar minimum dekho say



* complete link :-

here hm ~~not~~ consider kreiinge har
vaar,
but merge kte huye max length.
pehle min le rahi the
~~numbers ka~~ dono rows yaa
columns ka.

* Average link :-

↳ now ab bhi hm min. dakhinge har rasta, but merge karte time avg lengths of both rows and columns entries.

* Disadvantage of dendograms :-

↳ becomes messy when very large dataset, as no. of lines will be very high and it will become very difficult to identify the clusters.

clustering large Data :-

↳ Algorithms :- ① BIRCH

② DBSCAN \rightarrow density base.
③ CURE. \rightarrow clustering algos.

④ BIRCH :- (Balanced Iterative reduced clustering using hierarchy) :-

⑤ DBSCAN :- density based spatial clustering applications with noise

⑥ CURE :- clustering using representatives.

⑦ BIRCH :-

↳ only one scan required, just very highly efficient.

① CF (clustering feature)

↳ used for entire process of BIRCH algorithm

$N \rightarrow$ no. of points in cluster

$LS \rightarrow$ sum of 'i' in cluster

$SS \rightarrow$ sum of sq. of points in cluster

↳ we make tree, known as CF tree.

$B =$ max. no. of CF in non-leaf node

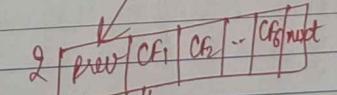
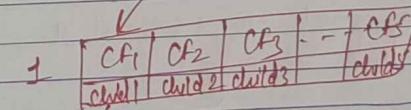
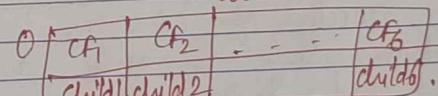
$L =$ max. no. of CF in leaf mode

$T =$ max. radius of subcluster

$O(CF_1), I(CF_1), \Delta(CF_1) \rightarrow$ all 3 are different.

as, $n(CF) = (N, LS, SS)$.

root



$$\Rightarrow \Delta(CF_1) = (7, 144, 545)$$

these 2 we just assumed and mostly tm yet to change.

and

here →

$$IC(CA) = 2(CA_1 + CA_2 + \dots + CA_6)$$

and,

$$OCCA_1 = IC(CF_1 + CF_2 + \dots + CF_5)$$

say New data pt. (N_2) aaya, and it is close to $2(CF_2)$, then it must go there, update kaise usko, then
 ① like same, CF_{root} B must be updated,
 and ②

and say $2(CF_2)$ mein N_2 can't be accommodated, then shifting kriji pdlegi, kuch Data pt. in $2(CF_1)$, kuch in $2(CF_3)$, some may go till $2(CF_6)$. so some times shifting may be req'd, known as maintenance of CF-tree.
 → But beauty of CF-tree, b-tht calculation is easy as only add'n and multiplication and only one scan is required.

* **CURE**!:- → use many points to represent a cluster instead of only one.

Effectiveness of CURE
 ↗ one scan, $TC=O(n)$, sensitive to user-defined parameters (sample size etc.)
 ↗ does not handle categorical attributes

WJMK
STUDY

cf additive property.

* Density based methods:-

WJMK
STUDY

①

DBSCAN :- 2 inputs required from user

core point???

no point

of this cluster

bnd skta hai,

having pts. > MinPts

within ϵ distance.

MinPts

E.

for

each pt in

cluster there

must be

another

pt in it less

than than

this distance

away

core point
for each pt in cluster there must be another pt in it less than than this distance away

core point

1st level of subcluster

$$N_{Eps}(p) = \{q \in D \mid \text{dist}(p, q) \leq Eps\}$$

E ($Epsilon$)

A pt. p is directly density reachable from a pt. q w.r.t ϵ minpts if
 1) q belongs to $N_{Eps}(p)$
 2) p is a core pt.

$$|N_{Eps}(p)| \geq \text{MinPts}$$

(Pinski definition slides mein get hai) \rightarrow yeh copy wali sahi hai

K-Means

* A pt. p is density reachable from a pt. q wrt Eps , Minpts , if there is a chain of pts. p_1, p_2, \dots, p_n , $p_1 = q$, $p_n = p$

s.t. p_{i+1} is directly density reachable from p_i .

* Density-connected ???

→ this algo can generate clusters with very-
different shapes (and any)

→ and noise agar hai toh uska bilkul
different cluster bnega, thus handles
noise very effectively.

[DBSCAN]

→ process of making clusters tak tak chilega jab
tak sare points exhaust na ho
jaayein.

H/w MST single link :-

H/w. BFR :-

Beauty is it uses K-means and can
handle large data.

also uses CF.