7/4/23
Sunday

Module-4
Association Rule Analysis (ARA)

Possibilities of buying two items together

It is also known as market Basket Analysis

It is used to find relationship/association b/w large set of data items.

$X \longrightarrow Y$

Support and confidence are used to find association b/w items in a data set.

Support

The support for $x \rightarrow y$ is the probability of both x & y appearing together, i.e., $P(X \cup Y)$

$$\text{Support}(A \cup B) = P(A \cup B)$$

Confidence

The confidence of $X \longrightarrow Y$ is the conditional probability of Y appearing given that X exists. It is written as $P(Y/X)$ and read as POF Y given X.

$$\text{confidence}(A \Rightarrow B) = P(B/A) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

Find the support value of each itemset

Transaction ID	Items purchased
1	Bread, cheese, Egg, Juice
2	Bread, cheese, juice
3	Bread, Milk, yogurd.
4	Bread, Juice, Milk.
5	cheese, juice, Milk.

Support(item) = Freaquency of item / Total no: of transaction

Total no: of transaction = 5

		Support(%)
support(Bread) =	4/5 =	80%.
support(cheese) =	3/5 =	60%
support(Egg) =	1/5 =	20%
support(Juice) =	4/5 =	80%
support(Milk) =	3/5 =	60%
support(yogurt) =	1/5	20%.

Methods to discover Association rules

1. Basic idea behind the method to discover association rule.

step1: Find all freaquent itemset

step2: Generate strong association rules from the freaquent itemset

Frequent itemset

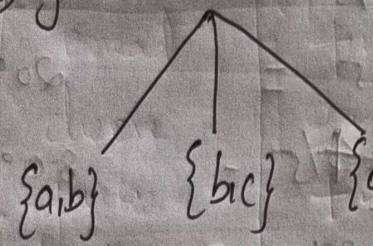
An itemset $X \subseteq A$ is said to be a frequent itemset in T with respect to σ, if

$$support(X) \geq \sigma$$

Downward closure property: Any subset of a frequent set is a frequent set.

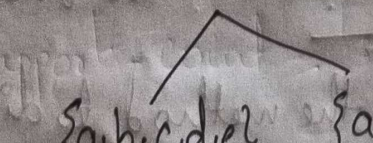
upward closure property: Any superset of an infrequent set is an infrequent set.

eg: $\{a, b, c\}$ → frequent.



{a} {b} {c}

→ If $\{a, b, c\}$ is infrequent then all



the superset of $\{a, b, c\}$ is infrequent.

APRIORI algorithm

It is also known as level-wise algorithm

Step 1: discover all frequent (single) items that have support above the minimum support count

Step 2: use the set of frequent items to generate the association rules that have high enough confidence level.

The candidate generation process and the pruning process are the most important parts of the algorithm.

Candidate set generation method

Given represented as C_k.

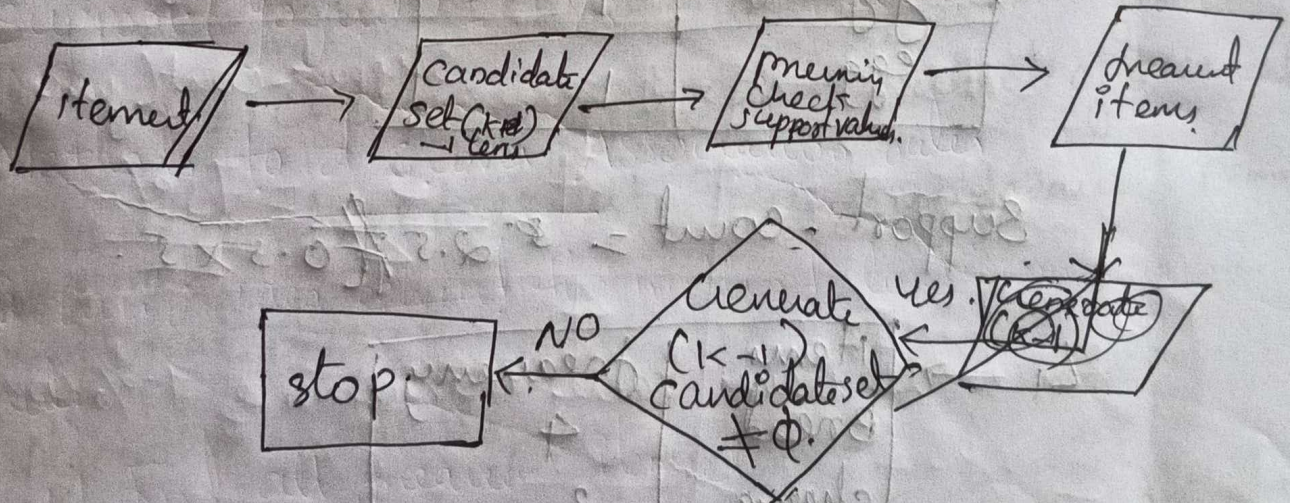
Given L_{k-1} the set of all frequent $(k-1)$-itemset, we want:

eg: {1, 2, 3} {1, 2, 4}

C = {1, 2, 3, 4}.

Pruning

It is used to eliminate infrequent items.



Find association rule from the following data with minimum support of 50% & minimum confidence of 75%

Transaction ID	items.
100	Bread, cheese, egg, juice
200	Bread, cheese, juice
300	Bread, Milk, yogurt.
400	Bread, juice, Milk
500.	cheese, juice, Milk.

step 1: Find L1

The items that are frequent (≥ 50% of support)

Item	frequency
Bread	4 ✓
cheese	3 ✓
juice	4 ✓
Milk	3 ✓
yogurt	1
egg	1

Support - count = 0.5 × 5 = 2.5

$L_1 \rightarrow$

Item	frequency
Bread	4
cheese	3
juice	4
Milk.	3

step 2: Find C_2 → candidate set of two items.

Items	frequency
{Bread, cheese}	2
{Bread, juice}	3
{Bread, Milk}	2
{cheese, juice}	3
{cheese, Milk}	1
{juice, Milk}	2

3. Find L_2. (≥ 2.5).

$L_2 \rightarrow$

Items	frequency
{Bread, juice}	3
{cheese, juice}	3

4. Find C_3.

Items. frequency.

It is not possible to create a candidate set C_3 {Bread, cheese, juice} because all subset of 3 frequent Subset must be frequent. There is no frequent items of the form {Bread, cheese}.

Algorithm will stop

Generate 4 association rules from the input data.

Bread → Juice

Juice → Bread

Cheese → Juice

Juice → Bread, cheese.

Bread → Juice → confidence = support_count(Bread, Juice) / support_count(Bread)

$= \frac{3}{4} = 75\%$.

Juice → Bread

confidence = $\frac{3}{4} = 75\%$.

Cheese → juice

confidence = support_count(cheese, juice) / support_count(cheese)

$= \frac{3}{3} = 100\%$

juice → cheese

confidence = support_count(juice, cheese) / support_count(juice)

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

Since all items have minimum of 75% confidence.

2. Find AR from the following data using Apriori algorithm with minimum support count measure & 2 & confidence 70%.

TID.	List of item_IDs.
T100	I1, I2, I5.
T200	I2, I4.
T300	I2, I3.
T400	I1, I2, I4.
T500	I1, I3
T600	I2, I3.
T700.	I1, I3.
T800	I1, I2, I3, I5
T900	I1, I2, I3.

1. Find L_1.

Items	frequency.
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2.

Support_count >= 2.

$L_1 \rightarrow$

Items	frequency.
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

2. Find C_2.

Items	frequency.
$\{I_1, I_2\}$	4.
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2.
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2.
$\{I_2, I_5\}$	2.
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

3. Find L_2. ≥ 2.

Items	frequency.
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2.
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2.
$\{I_2, I_5\}$	2.

$\{I_1, I_2, I_3\}$ ✓

$\{I_1, I_3, I_5\}$ ✗

$\{I_1, I_2, I_5\}$ ✓

$\{I_1, I_2, I_4\}$ ✗

$\{I_1, I_3, I_4\}$ ✗

$\{I_1, I_4, I_5\}$ ✗

$\{I_2, I_4, I_5\}$ ✗

4. Find C_3

Items	frequency.
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2

5. Find L_3

Items	frequency.
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2

6. Algorithm stop.

Association rule.

~~$\{I_1, I_2, I_3\}$~~:

$\{I_1, I_2, I_3\}$.

$\{I_1, I_2\} \rightarrow \{I_3\}$

$\{I_1, I_3\} \rightarrow I_2$

$\{I_2, I_3\} \rightarrow I_1$

$I_1 \longrightarrow \{I_2, I_3\}$.

$I_2 \longrightarrow \{I_1, I_3\}$.

$I_3 \rightarrow \{I_1, I_2\}$.

$I_1 \longrightarrow \{I_2, I_3\}$.

$$\text{Confidence} = \frac{\text{support\_count}(I_1 \& I_2, I_3\})}{\text{support\_count}(I_1)}.$$

$= \frac{2}{6} = 33.3\%$.

Confidence.

$I_2 \rightarrow \{I_1, I_3\}$

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_3\})}{\text{support\_count}(I_2)}$$

$$= \frac{2}{7} = 28.2\%.$$

$I_3 \rightarrow \{I_1, I_2\}$

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_3)}{\text{support\_count}(I_3)}$$

$$= \frac{2}{6} \quad 33.3\%.$$

$\{I_1, I_2\} \rightarrow \{I_3\}$

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_3)}{\text{support\_count}(I_1, I_2)}$$

$$= \frac{2}{4} = 50\%.$$

$\{I_1, I_3\} \rightarrow I_2$.

$$\text{Confidence} = \frac{\text{support\_count}(I, I_2, I_3)}{\text{support\_count}(I, I_3)}$$

$= \frac{2}{4}$ 50%

$\{I_2, I_3\} \rightarrow I_1$

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_3)}{\text{support\_count}(I_2, I_3)}$$

$= \frac{2}{4}$ 50%.

$\{I_1, I_2, I_5\}$.

$I_1 \rightarrow \{I_2, I_5\}$

$I_2 \rightarrow \{I_1, I_5\}$

$I_5 \rightarrow \{I_1, I_2\}$.

$\{I_2, I_5\} \rightarrow I_1$

$\{I_1, I_5\} \rightarrow I_2$

$\{I_1, I_2\} \rightarrow I_5$.

$\{I_1 \rightarrow \{I_2, I_5\}$.

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_1)}$$

$= \frac{2}{6}$ 33.3%

$I_2 \rightarrow \{I_1, I_5\}$.

$$\text{confidence} = \frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_2)}$$

$I_5 \rightarrow \{I_1, I_2\}$

Confidence = $\frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_5)}$

$= \frac{2}{2} = 1 \quad 100$

$\{I_1, I_5\} \rightarrow I_1$

Confidence = $\frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_1, I_5)}$

$= \frac{2}{2} = 1 \quad 100$

$\{I_1, I_5\} \rightarrow I_2$

Confidence = $\frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_1, I_5)}$

$= \frac{2}{2} \quad 100$

$\{I_1, I_2\} \rightarrow I_5$

Confidence = $\frac{\text{support\_count}(I_1, I_2, I_5)}{\text{support\_count}(I_1, I_2)}$

$= \frac{2}{4} \quad 50\%$

Final o/p

$I_5 \rightarrow \{I_1, I_2\}$

$\{I_2, I_5\} \rightarrow I_1$

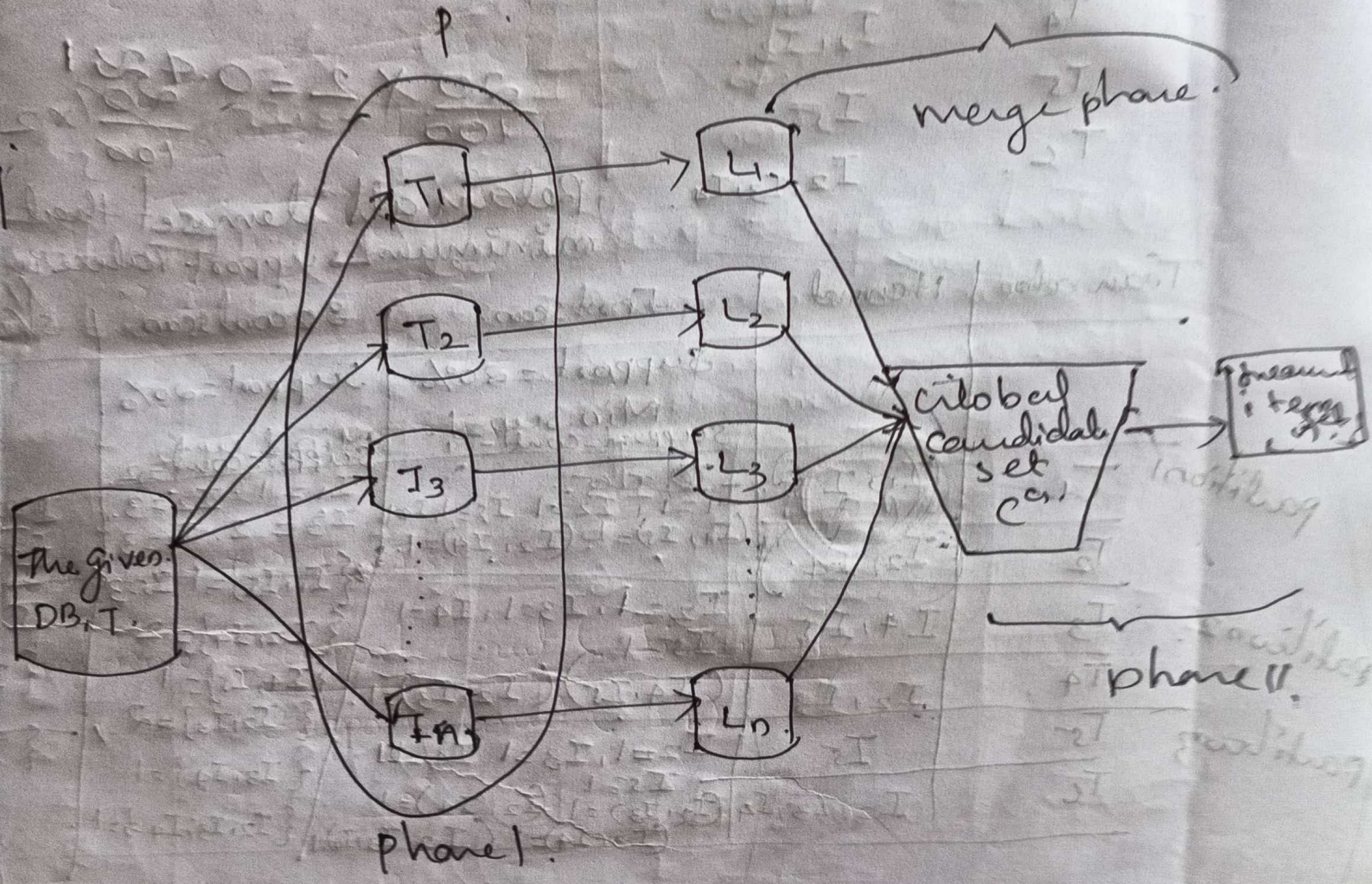
$\{I_1, I_5\} \rightarrow I_2$

Partition Algorithm

Partition algorithm scans the database only twice.

First scan: generate a set of all potentially large itemset.

Second scan: setup counters for each potentially large itemset and compute their actual supports



Transaction: Given
Find frequently itemset from the following data with support value of 20% using partition algorithm of size 3

Transaction	Itemset
T_1	I_1, I_5
T_2	I_2, I_4
T_3	I_4, I_5
T_4	I_2, I_3
T_5	I_5
T_6	I_2, I_3, I_4

$$\text{Support\_count} = \frac{\text{support}}{100} \times \text{no. of transactions}$$

$$\frac{20}{100} \times 2 = 0.4 \approx 1 \qquad \frac{20}{100} \times 2 = 0.4 \approx 1$$

Retain all itemset that have minimum support value in first scan.

	Transaction	Itemset	First scan: Support = 20%. Min. sup = 1. support_count = 1	Second scan: Support = 20%. Min. support_count = 2.	Shortlisted
partition 1	T_1	I_1, I_5	$I_1=1, I_2=1, I_5=1, I_4=1$	$I_1=1, I_2=3$	$I_2=3$
	T_2	I_2, I_4	$(I_1, I_5)=1, (I_2, I_4)=1$	$I_3=2, I_4=3$	$I_3=2$
partition 2	T_3	I_4, I_5	$I_2=1, I_3=1, I_4=1$	$\{I_1, I_5\}=1, I_5=3$	$I_4=3$
	T_4	I_2, I_3	$I_5=1$	$\{I_2, I_4\}=2$	$I_5=3$
partition 3	T_5	I_5	$(I_4, I_5)(I_2, I_3)=1$	$\{I_4, I_5\}=1$	$\{I_2, I_4\}=2$
	T_6	I_2, I_3, I_4	$I_2=1, I_3=1, I_4=1$	$\{I_2, I_3\}=2$	$\{I_2, I_3\}=2$
			$I_5=1$	$\{I_3, I_4\}=1$	
			$(I_2, I_3)=1, (I_2, I_4)=1$	$\{I_2, I_3, I_4\}=1$	$\{I_2, I_3$
			$(I_3, I_4)=1, (I_2, I_3, I_4)=1$		

→ Local frequent itemset

→ Here, all itemsets are satisfying the condition. All items are retains during first scan. There itemset are local frequent itemset

final o/p frequent itemset

$\{I_2, I_4\}$ — 2

$\{I_2, I_3\}$ — 2

$I_2 \rightarrow I_4$

$$\text{confidence} = \frac{\text{support\_count}(I_2, I_4)}{\text{support\_count}(I_2)}$$

$$= \frac{2}{1} = 2$$

confidence = support_count.

$\frac{20}{100} \times$

$\frac{20}{100} \times 2$

$\frac{1}{5} \times 2 = 0.4 \approx 1$

Fp Growth Algorithm → Fp Tree

Find AR using the following data using Fp growth algorithm with min support count = 2 & confidence 80%

Transaction	Itemset
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3

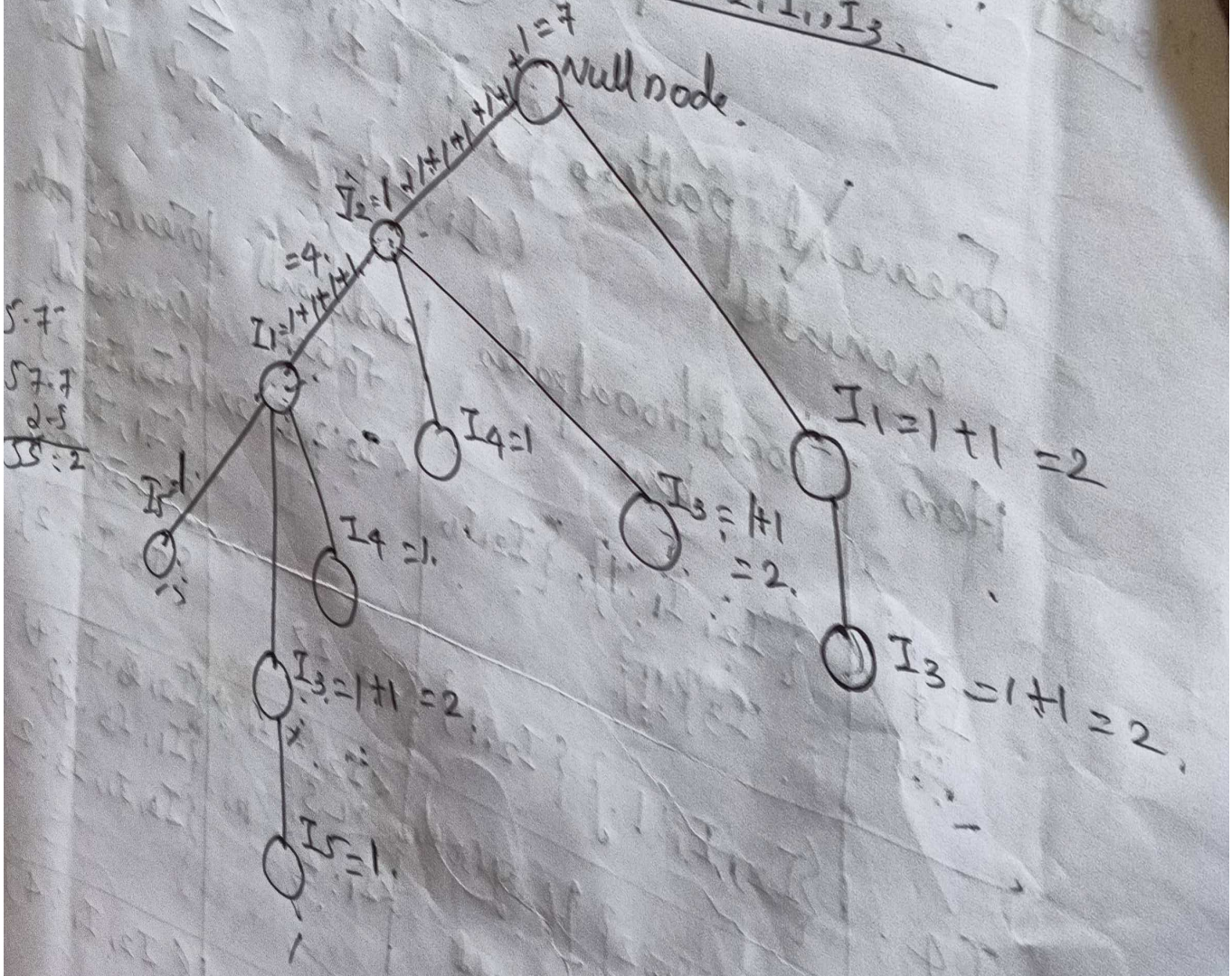
1) Find frequent pattern set (L): It contains all items whose frequency is greater than or equal to minimum support value and these items are arranged in descending order.

Item	Frequency
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

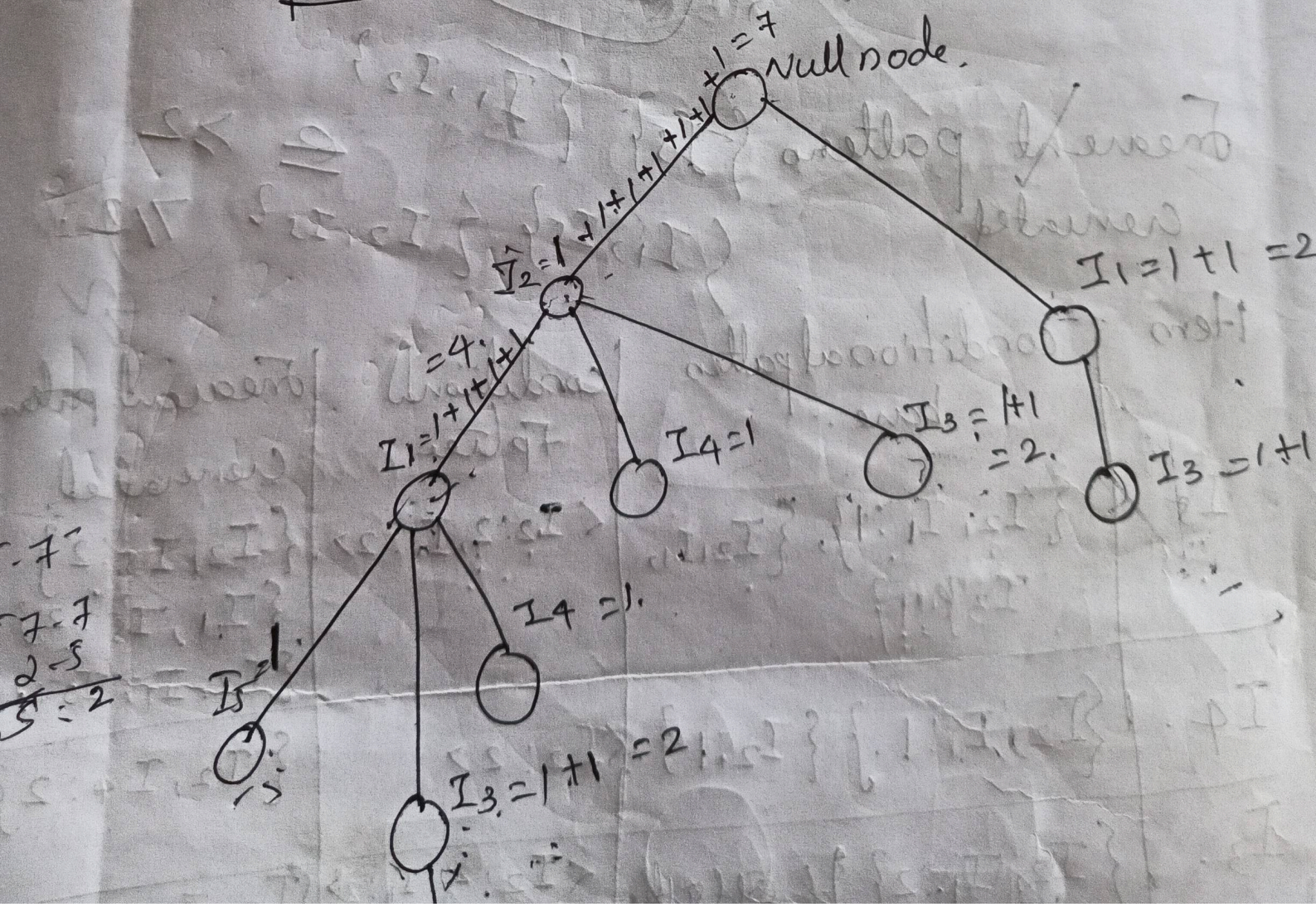
Item	Frequency
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

2) Find ordered itemset

Transaction	Item_set	Ordered Itemset
T_{100}	I_1, I_2, I_5	I_2, I_1, I_5
T_{200}	I_2, I_4	I_2, I_4
T_{300}	I_2, I_3	I_2, I_3
T_{400}	I_1, I_2, I_4	I_2, I_1, I_4
T_{500}	I_1, I_3	I_1, I_3
T_{600}	I_2, I_3	I_2, I_3
T_{700}	I_1, I_3	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5	I_2, I_1, I_3, I_5
T_{900}	I_1, I_2, I_3	I_2, I_1, I_3



Transaction	Item_set	ordered itemset
T100	I_1, I_2, I_5	I_2, I_1, I_5
T200	I_2, I_4	I_2, I_4
T300	I_2, I_3	I_2, I_3
T400	I_1, I_2, I_4	I_2, I_1, I_4
T500	I_1, I_3	I_1, I_3
T600	I_2, I_3	I_2, I_3
T700	I_1, I_3	I_1, I_3
T800	I_1, I_2, I_3, I_5	I_2, I_1, I_3, I_5
T900	I_1, I_2, I_3	I_2, I_1, I_3



Item	conditional pattern base	Conditional fp tree
I_5	$\{I_2, I_1 : 1\}, \{I_2, I_1, I_3 : 1\}$	$\langle I_1 = 2, I_1 : 2 \rangle$
I_4	$\{I_2, I_1 : 1\}, \{I_2, I_1, I_4 : 1\}$	$\langle I_2 I_2, I_4 = 2 \rangle$
I_3	$\{I_1, I_3 : 2\}, \{I_2, I_3 : 2\}$ $\{I_2, I_1, I_3 : 1\}$	
	$I_2, I_5 : 2 \rangle L_5$	
	$I_2, I : 2$	
	$I_1, : 2 \rangle R_1$	
I_1		

123
Sunday Today

$\{I_5\}$ $\{I_1, I_2\}$

Frequent pattern generated. $\{I_5\}$ $\{I_1, I_2\}$ $\{I_1, I_5\}$ $\{I_2, I_5\}$ $\{I_1, I_2, I_5\}$ ≥ 2 ?

Item	conditional pattern base	conditional fp tree	Frequent pattern generated
I_5	$\{I_2, I_1 : 1\}, \{I_2, I_1, I_3 : 1\}$	$\langle I_2 : 2, I_1 : 2 \rangle$	$\{I_2, I_5 : 2\}$ $\{I_1, I_5 : 2\}$ $\{I_2, I_1, I_5 : 2\}$
I_4	$\{I_2, I_1 : 1\}$ $\{I_2 : 1\}$	$\langle I_2 : 2 \rangle$	$\{I_2, I_4 : 2\}$
I_3	$\{I_2, I_1 : 2\}$ $\{I_2 : 2\}$ $\{I_1 : 2\}$ Right sibling	$\langle I_2 : 4, I_1 : 2 \rangle$ L_5 $I_1 : 2 \rangle R_5$	$\{I_2, I_3, I_3 : 4\}$ $\{I_1, I_3 : 4\}$ $\{I_2, I_1, I_3 : 2\}$
I_1	$\{I_2 : 4\}$	$\langle I_2 : 4 \rangle$	$\{I_2, I_1 : 4\}$

Association rule

I_5 & I_3 are highest frequent pattern

2/5/23 Tuesday

Dynamic itemset counting algorithm (DIC)

Itemsets are marked in 4 different ways.

□ solid box : confirmed frequent itemset.

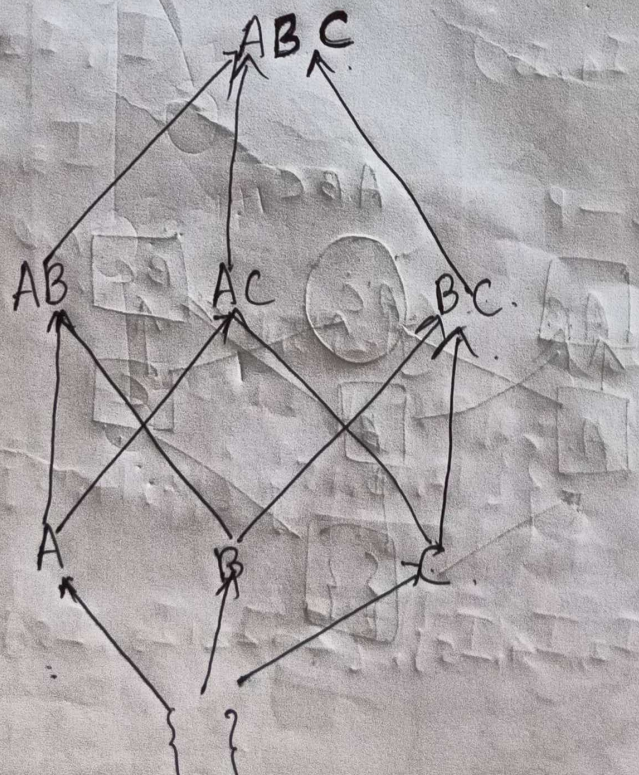
○ solid circle : confirmed infrequent itemset

: dashed box : suspected frequent itemset

: dashed circle : suspected infrequent itemset

min supp = 25% ($S = 5 \times 0.25 \times 4 = 1$) : M = 2

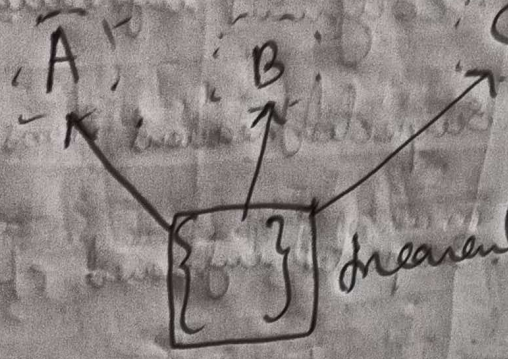
TID	A	B	C	
T_1	1	1	0	AB
T_2	1	0	0	A
T_3	0	1	1	BC
T_4	0	0	0	-



lattice

ABC.

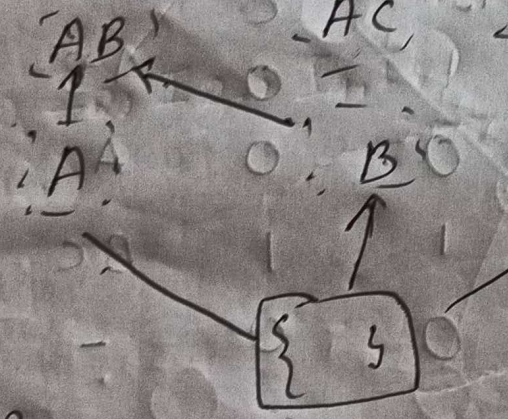
AB AC BC.



contu
A = 0
B = 0
C = 0

incrément

ABC.

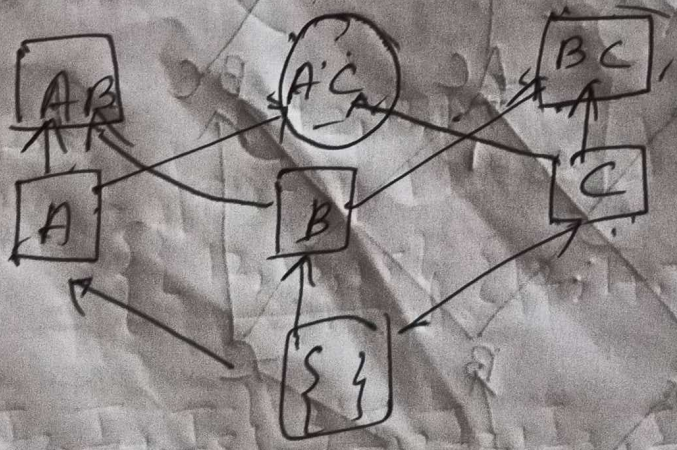


contu
A = 2
B = 1
C

mise à zéro
(All tous)

A = 2
B = 2
C = 1.

ABC



3/5/23 Wednesday

Pincer search Algorithm

Bottom up procedure.

BFS method.

computations

Pass 1:

Item	Count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

$C_1 = \{\{I_1\}, \{I_2\}, \{I_3\}, \{I_4\}, \{I_5\}\}$

MFCS $= \{I_1, I_2, I_3, I_4, I_5\}$.

MFS $= \{\}$.

$\{I_1, I_2, I_3, I_4, I_5\} - 0$

LC =

$L_1 = \{\{I_1\}, \{I_2\}, \{I_3\}, \{I_4\}, \{I_5\}\}$

$S_1 = 9$. infrequent item

Itemset	Sup-count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

count (MFCs) = 0

$L_2 = \{I_1, I_2\}\ \{I_1, I_3\}; \{I_1, I_5\}\ \{I_2, I_3\}$

$\{I_2, I_4\}\ \{I_2, I_5$

$S_2 = \{\{I_1, I_4$

Delete $\{I_1, I_4\}$ from MFCs one at at time.

$\{I_1, I_2, I_3, I_4, I_5\}$.

(I_2, I_3, I_4, I_5) (I_1, I_2, I_3, I_5).

(I_2, I_3, I_4, I_5) (I_1, I_2, I_3, I_5)

~~(I_3, I_4)~~ — P.

I_3, I_4 — absent

no change.

(I_2, I_4, I_5) (I_2, I_3, I_5)

New MFCs = $\{\{I_2, I_3, I_5\}\ \{I_2, I_4, I_5\}\ \{I_1, I_2, I_3, I_5\}$

$= \{\{I_2, I_4, I_5\}, \{I_1, I_2, I_3, I_5\}\}$

Delete $\{I_3, I_5\}$ from each item in MFCS one at a time

I_1, I_2, I_3, I_5 — (I_3, I_5) — P.

I_2, I_4, I_5 — (I_3, I_5) → absent → No change

$\underline{I_3}$

(I_1, I_2, I_3)

$\underline{I_5}$

(I_1, I_2, I_3)

New MFCS.

$\{\{I_1, I_2, I_5\}, \{I_1, I_2, I_3\}, \{I_2, I_4, I_5\}$

Delete (I_1, I_4, I_5) }.

$\{I_2, I_4, I_5\}$

$\underline{I_2}$

$\{I_4, I_5\}$.

$\underline{I_5}$

$\{I_2, I_5\}$

$I_2, I_5 \subseteq \{I_2, I_4, I_5\}$.

final MFS. $\{I_1, I_2, I_5\}, \{I_1, I_2, I_3\}\ \{I_2, I_4, I_5\}$

Final o/p.

Frequent pattern → $\{I_1, I_2, I_5\}$

$\{I_1, I_2, I_3\}$.

4/5/23
Thursday

Clustering

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. (clusters)

Measures of dissimilarity

Numeric data

Measures of distance between data points.

Euclidean distance $= \|\bar{x} - \bar{y}\|_2 = \sqrt{(x_1-y_1)^2 + \cdots + (x_n-y_n)^2}$

Squared euclidean distance $= \|\bar{x} - \bar{y}\|_2^2 = (x_1-y_1)^2 + \cdots + (x_n-y_n)^2$

City Block / Manhattan distance $= \|\bar{x} - \bar{y}\|_1 = |x_1-y_1| + \cdots + |x_n-y_n|$

Maximum distance $= \|\bar{x} - \bar{y}\|_\infty = \max\{|x_1-y_1|, \ldots |x_n-y_n|\}$

$X = (2, 2)$

$Y = (1, 1)$

1) Euclidean distance: $\sqrt{(x_1-y_1)^2 + \cdots}$

$\sqrt{(2-1)^2 + (2-1)^2} = \sqrt{1^2+1^2} = \sqrt{2}$

2) Squared

3) CB/Manh $= |x_1-y_1| + |x_2-y_2|$

$= |2-1| + |2-1|$

$= 1+1 = 2$

Maxdist = max { |x1-y1|, |x2-y2| }

= max { |2-1|, |2-1| = max {1,1} = 1

x = (2,1,3)

y = (2,0,1)

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

Minkowski distance.

Non-numeric data

hemenshetin distance.

Clustering methods

Hierarchial's partition

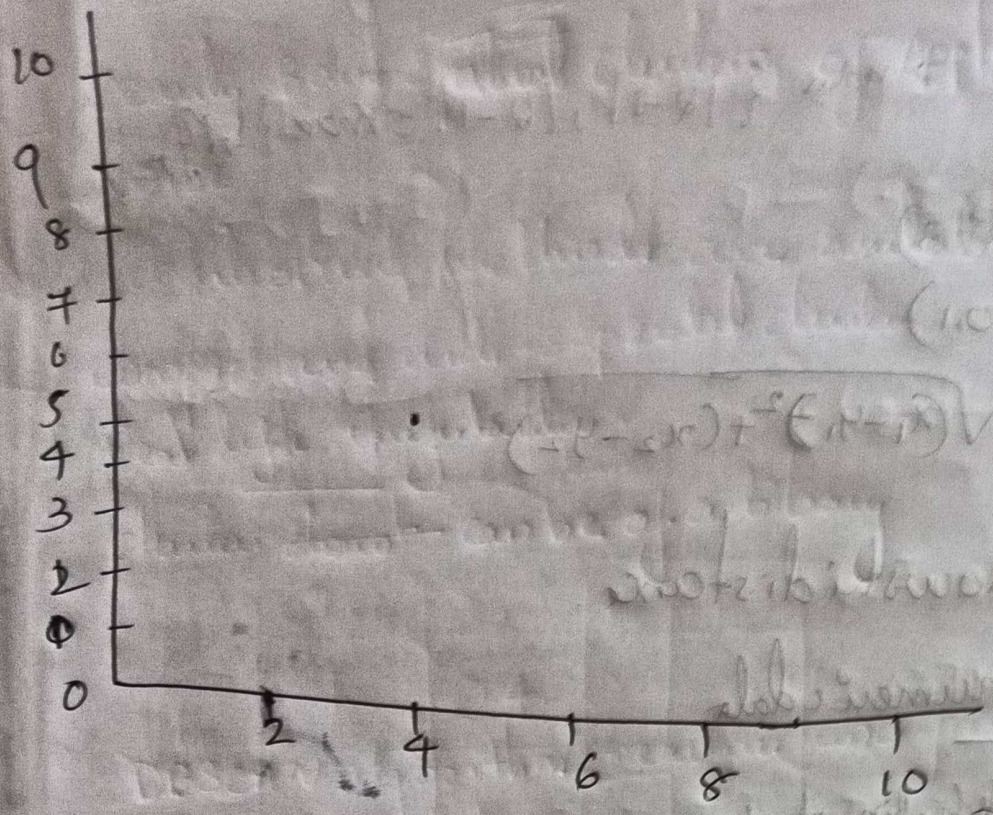
PAM - partition Around Medoid Algorithm

	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5

$|x_1-y_1|+|x_2-y_2|$

$|8-7|+|8-4|$

$2+3=3$



$C_1(4,5)$, $C_2(8,5)$.

Calculating cost.

	X	Y	Dissimilarity from C_1	Diss from C_2
C_2	8	7	6	2
C_1	3	7	3	1
	4	9	4	8
			6	2
			4	6
			5	3
			5	1
			34	1

Manhattan

$(8,7)$ from $(4,5)$

$D = |8-4| + |7-5| =$

$4 + 2 = 6$

$(8,7)$ from $(8,5)$

$C_1 \rightarrow 2+2+4 = 8$

$C_2 \rightarrow 3+3+1+1+4$

$= 12$

Total = 20

new medoid (8, 4)

DBSCAN

Density Based spatial clustering of Applications with Noise.

Partitioning methods are suitable only for compact and well-seperated clusters

Real life irregularities

clusters ~~dont~~ can be of arbitrary.

DBSCAN algorithm requires two parameters.

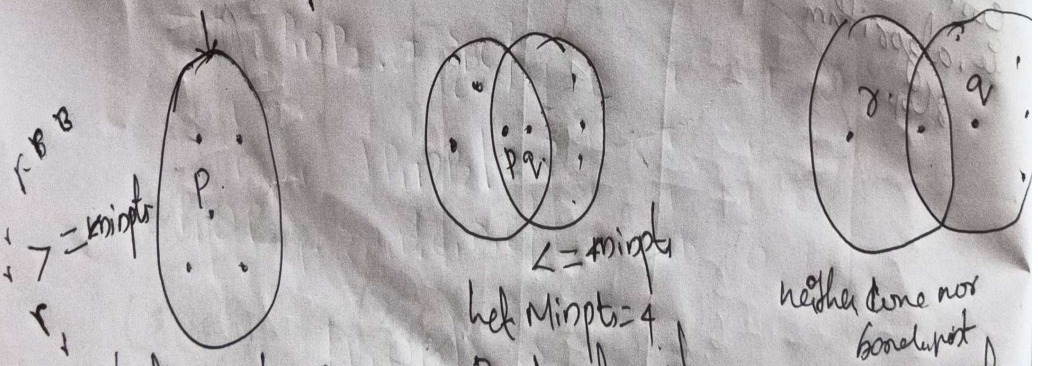
1. eps(ϵ): it defines
2. MinPts

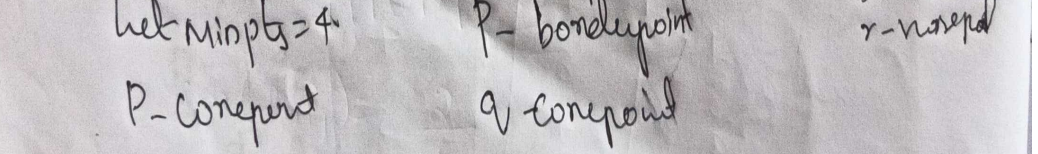
3 types of datapoints.

core point: A point is a core point if it has more than MinPts points within ϵ of it.

Borderpoint: A

noise point

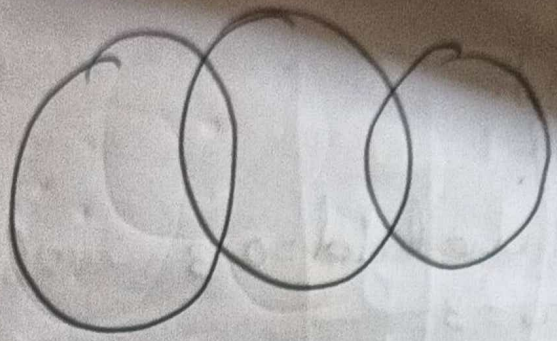




let Minpts = 4
P - corepoint

let Minpts = 4
P - borderpoint
q corepoint

neither core nor borderpoint
r - noisepoint



DBScan
Palm
modern

Density based - DBScan
Partition - palm

• Categorical clustering - Rock

Robust clustering using links.
Rock belongs to the class of Agglomerative hierarchical clustering Algorithm.
Rock works for categorical attributes

Data → draw random simple → cluster with links.

Neighbours.

$$Sim(P_i, P_j) \geq \theta$$

Jaccard coefficient for $sim(T_1, T_2)$,

$$Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \rightarrow \text{coefficient.}$$

$link(P_i, P_j)$ no: of common neighbours.

Good new means

$$g(C_i, C_j) = \frac{link(C_i, C_j)}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

$$f(\theta) = \frac{1-\theta}{1+\theta}$$

$P_1 = \{A, B, C, D\}$.

$P_2 = \{E, B, C\}$.

$P_3 = \{D, E, B\}$.

$P_4 = \{E, C, F\}$.

Similiarity threshold = 0.3

No. of clusters = 3.

1) Similiarity table.

	P_1	P_2	P_3	P_4
P_1	1	0.4	0.4	0.17
P_2		1	0.5	0.5
P_3			1	0.2
P_4				1

$$sim(P_i, P_j) = \frac{|P_i \cap P_j|}{|P_i \cup P_j|}$$

P_1, P_2 $\quad P_1 \cap P_2 = \{B, C\}$.

$$|P_1 \cap P_2| = 2$$

$P_1 \cup P_2 = \{A, B, C, D, E\}$.

$$|P_1 \cup P_2| = 5$$

$$sim(P_i, P_j) = \frac{2}{5}$$

) Adjacency table

	P_1	P_2	P_3	P_4
P_1	1	1	1	0
P_2		1	1	1
P_3			1	0
P_4				1

$Sim(P_1, P_2) = 0.4$

$0.4 > 0.3$

$Sim(P_1, P_4) = 0.17$

$0.17 < 0.3$

$= 0.$

if $Sim(P_i, P_j) \geq$ threshold $= 1.$

" " $\leq$ " $= 0.$

$nP_1 = 1$

$nP_2 = 1$

3. No: of links / common neighbours.

	P_1	P_2	P_3	P_4
P_1	-	3	3	1
P_2		+	3	2
P_3			+	1
P_4				-

No: of links (P_i, P_j)

= Admat x Adjmat

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \end{bmatrix}$$

4) Goodness Measure.

$$g(c_i, c_j) = \text{link}[C_i, C_j]$$

$$g(P_1, P_2) = \frac{\text{link}[P_1, P_2]}{(n_1+n_2)^{1+2f(\theta)} - n_1^{1+2f(\theta)} - n_2^{1+2f(\theta)}}$$

$$f(\theta) = \frac{1-\theta}{1+\theta} = \frac{1-0.3}{1+0.3} = 0.57$$

$$= \frac{3}{2^{1+2\times0.57} - 1^{1+2\times0.57} - 1^{1+2\times0.57}}$$

$$= 1.35$$

Pair	CM
P_1, P_2	1.35
P_1, P_3	1.35 — max value.
P_1, P_4	0.45
P_2, P_3	1.35
P_2, P_4	0.90
P_3, P_4	0.45

highest equal value for 3 pairs.

cluster P_1, P_2

new link table. old link table → single link.

	P_1, P_2	P_3	P_4
P_1, P_2	-	6	3
P_3		-	1
P_4			-

$(P_1, P_2), P_3 \rightarrow (P_1, P_2)\ (P_2, P_3)$

link $[(P_1, P_2), P_3] =$ link $[P_1, P_3]$ + link $[P_2, P_3]$.

$= 3 + 3 = 6$

n_i = no: of points in $[P_1, P_2]$

$= 2.$

n_j = no: of points in $[P_3] = 1$