Association Rules & Item Sets

Frequent pattern :-

Eg : (itemsets, subsequences or substructures) patterns appear frequently in a dataset

Ex Set of Items such as milk and bread appear frequently in a data set. (together)

> Freq. Itemset = milk & bread

Subsequence :-

Buying first pc, then digital camera then memory card } occurs frequently in shopping history DB.

↓

called (frequent) sequential pattern.

Substructure :

different structural forms such as subgraphs, subtrees, or sublattices which may be combined with itemsets or subsequences.

→ if substructure occured frequently called (frequent) structured pattern.

MBA (Market Basket Analysis algm)

→ Analyze customer buying habbit

→ discover correlation relationship between business txn. records

→ Helps for Business decision making.

→ Rep. frequently associated or purchased together

→ It rep'td in the form of association Rule.

Ex Customer

↳ buys Computer

↳ also tend to buy Antivirus at the sametime

Computer → Antivirus_slw

- Support = 2% (Computer & Antivirus slw purchased together out of all txns of the entire shop)
- Confidence = 60% (60% of customer purchased computer also bought the slw).

→ It needs to satisfy minimum support threshold & minimum confidence threshold

→ set by users / domain experts

Frequent Itemsets, Closed Itemsets & Association Rules

Freq. Itemset

Itemset whose support is greater than user specified minimum support.

closed freq. Itemset

if none of its immediate supersets has the same support as that of the Itemset

Maximal freq. Itemset

if none of its immediate supersets is frequent

Downward closure property of frequent patterns

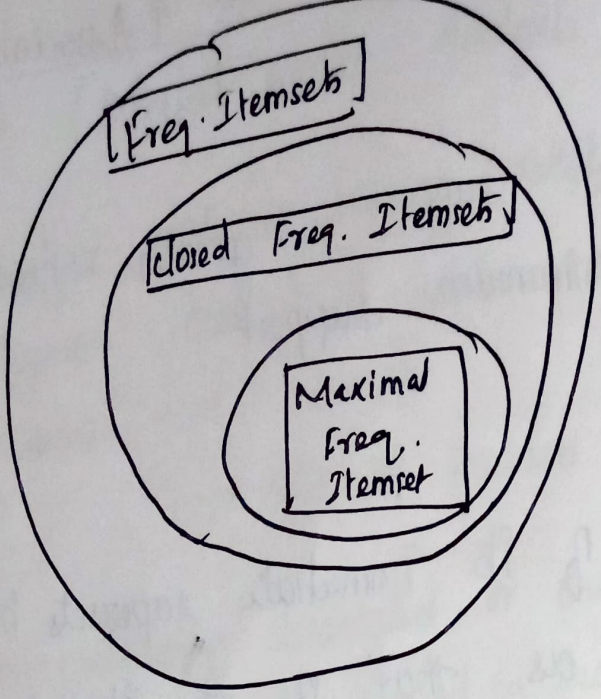
All subset of any frequent Itemset must also be freq.

Ex

if Milk, Bread, Butter is a freq. Itemset then the following Itemsets are freq.

- Milk
- Bread
- Butter
- Milk, Bread
- Milk, Butter
- Bread, Butter

} smaller subset of freq. pattern Itemset



Need of closed & Maximal Itemset

It useful when huge amount of data used in association rule mining

If length of freq. itemset is 'k' then.

by downward closure property

→ All of it 2^k subsets are also freq.

Disadv of Maximal freq Itemsets

Even all its subsets are frequent,
we don't know their support

For mining rules, support info is very ∴

→ Therefore closed freq itemset is preferred

TID	List of Items
T1	A, B, C, D
T2	A, B, C, D
T3	A, B, C
T4	B, C, D
T5	C, D

Min sup. count = 3

Find freq, closed, Maximal itemset

Item	Count
A	3
B	4
C	5
D	4

All items A, B, C & D are frequent because their sup count is greater than or equal to min. sup count

↳ sup count ≥ min sup count

2 itemset as combination

Item	Count
A, B	3
A, C	3
A, D	2
B, C	4
B, D	3
C, D	4

All 2 items are freq. except AD

∵ min sup count = 3

A(3) → AB(3), AC(3), AD(2)

↳ sup. count — Immediate superset

A (count) is not > its immediate superset → Yes

condn not satisfied

↓ close

A is not closed

In A's immediate superset,

Itemset are present with min sup. count 3

A is not maximal.

[If immediate superset is not frequent (not satisfy min support count) → called maximal]

B (4) → AB (3), BC (4), BD (3)

B (count) is not greater than its immediate superset

B is not closed

B's Immediate superset, Itemset are present with min sup. count 3

B is not maximal.

C (5) → AC (3), BC (4), CD (2)

C(count) > immed. superset (count)

∴ C is closed

C Not maximal

D (4) → AD (2), BD (3), CD (4)

D(count) > im. superset (count)

D is not closed

Im. superset count > min sup count (3)

→ No

→ D is not maximal

3 Itemset Generation

Item	count
A B C	3
A B D	2
A C D	2
B C D	3

AB (3) → ABC (3), ABD (2), ACD (2), BCD (3)

AB (count) > Is (count)

AB is not closed

Is(count) > min sup. count (3)

No → AB not maximal

AC(3) → ABC(3), ACD(2)

↳ Not closed

↳ not maximal.

AD(2) → ABD(2), ADC(2)

↳ not frequent ∵ sup count < min sup count.

BC(4) → ABC(3), BCD(3)

↳ closed

↳ not maximal

BD(3) → ABD(2), BCD(3)

↳ not closed

↳ not maximal

CD(4) → ACD(2), BCD(3).

↳ closed

↳ Not maximal

4 Itemset Generation

item	count
A, B, C, D	2

min $\frac{3}{=}$

ABC(3) → ABCD(2).

↳ closed

↳ itemset not present with min sup count (3)

↳ ~~Not~~ maximal

ABD (2) → Not freq.

ACD (2) → Not freq.

BCD (3) → ABCD (2)

↳ closed

↳ Maximal.

Assignment

TID	List of items
T1	B, J, P
T2	B, P
T3	B, M, P
T4	E, B
T5	E, M

Min sup. count = 40 % ⇒ 2 //

Find frequent, closed, max. Itemset

→ All freq except J

Item	count
B	4
E	2
J	1 → < min sup. count
M	2
P	3
	12

B (4) → BE (1), BM (1), BP (3) → closed, Not max

E (2) → BE (1), EM (1), EP (0) → closed, max

M (2) → BM (1), EM (1), MP (1) → closed, max

P (3) → BP (3), EP (0), MP (1) → not closed, not max.

∵ Itemset empty so BP is closed & maximal.

Sup & conf blw 0% and 100% rather than 0 & 1.0.

Itemset set of items

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

$$\text{Confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)}$$

Freq. itemset

itemset occurs frequently $\Rightarrow \geq$ minsup. count

Generate strong association rules from freq. itemset

rules must satisfy min support and min. confidence.

Ex

freq itemset of length 100

$\{a_1, a_2, \ldots a_{100}\}$ contain $\binom{100}{1} = 100$

freq. 1 itemset: $\{a_1\}\ \{a_2\} \ldots \{a_{100}\}$

$\binom{100}{2}$ freq. 2 itemset: $\{a_1, a_2\}\ \{a_1, a_3\} \ldots \{a_{99}, a_{100}\}$

Tot. no. of freq itemset contains

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}$$

This is too huge number of itemsets for any computer to compute or store.

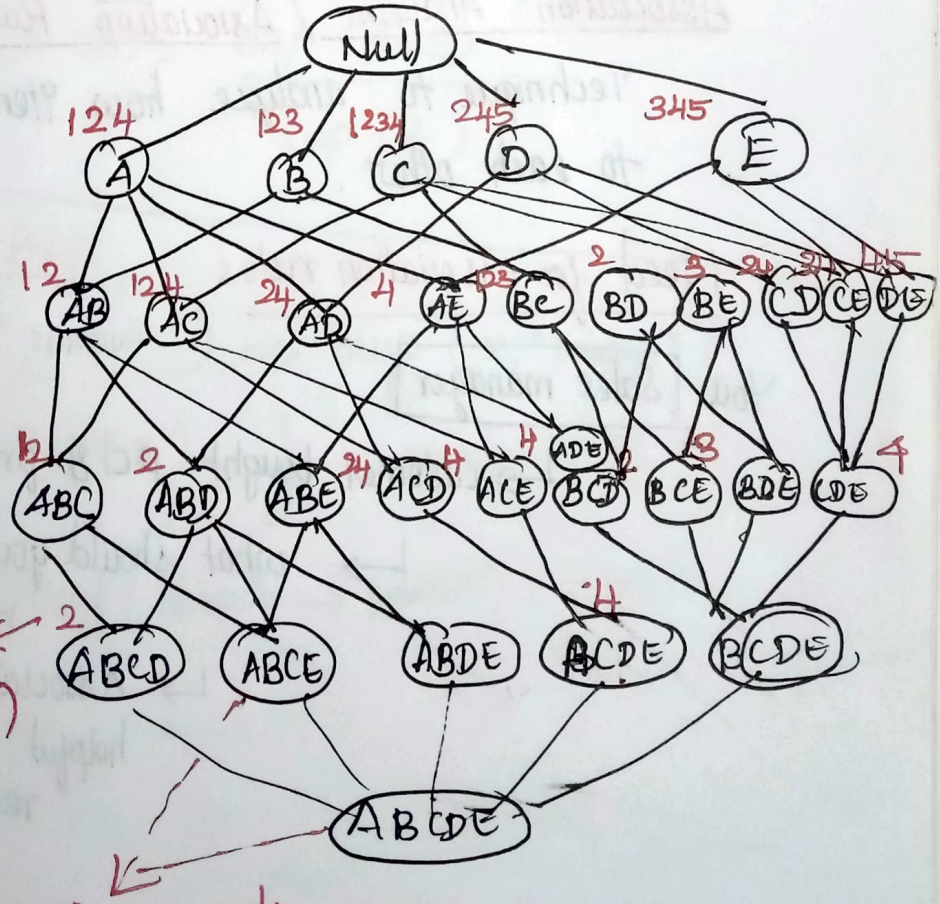
⇓ overcome this difficulty

closed freq. Itemset & maximal freq itemset introduced.

⇓

Maximal ~~Closed~~ freq Itemset	closed ~~Maximal~~ freq Itemset
if none of its immediate supersets is frequent.	If none of its immediate supersets has the same support as the itemset.

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



min sup = 2

closed, not max

C, D, E, ~~AC, BC, CE, DE~~

closed & Max

CE, DE, ABC, ACD

closed = 9
max = 4

Apriori Algorithm : Finding frequent Itemsets by confined Candidate Generation

→ It uses frequent Itemsets to generate association rules.

Itemset :- set of items together

Ex { Milk, Bread, Egg }

→ k Itemset
→ Itemset contains k items

freq:
↳ satisfy min TH value for sup & conf.

Association Analysis / Association Rules

Technique to analyze how items are associated to each other.

Need for association rules

You [Sales manager]
↳ customer bought PC & printer recently
↳ what should you recommend to her next?
↳ Association rules are helpful in making your recommendation

Association rules

determine which items are frequently purchased together within the same txn.

$A \Rightarrow B$

↳ implies / determines

some value of itemset A

↳ determines the value of itemset B

under condition in which min sup & conf are met

$$\text{sup}(A) = \frac{\text{No. of txn in which A appears}}{\text{Total no. of Txns}}$$

$$\text{conf}(A \rightarrow B) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)}$$

Customer Buys

(X "computer") ⇒ buys (X, "printer")

$$\text{sup}(x) = \frac{1}{2} \quad \cancel{= 50\%} \quad 1\ \%$$

$$\text{conf}(\text{computer} \rightarrow \text{printer}) = 50\ \%$$

Support count (σ):

Freq of occurrence of itemset

Ex $\sigma(\{Bread, Milk, Sugar\}) = 2$.

T.ID	Items bought
I1	Bread, Milk.
I2	Bread, Egg, Sugar
I3	Milk, Sugar, Jam.
I4	Bread, Milk, Sugar, Jam.
I5	Bread, Milk, Sugar, Eggs

Apriori algm
Finding freq. Itemsets by confined Candidate Generation

2 step process

1. Join step

generate (k+1) itemset from k itemsets by joining each item with itself.

2. prune step

→ scans the count of each item in the DB
→ If the candidate item doesn't meet min sup.
→ infrequent → remove it
→ step to reduce the size of the candidate itemsets

Ex

TID	Item set
T100	I1, I2, I5
T200	I2, 4
T300	2, 3
T400	1, 2, 4
T500	1, 3
T600	2, 3
T700	1, 3
T800	1, 2, 3, 5
T900	1, 2, 3

Step 1

Generate 1 °itemset [C1]

algm count occurrance of each item.

Itemset	sup count
I1	6
I2	7
I3	6
I4	2
I5	2

Step 2

min sup count = 2 //

Pruning step: No need of pruning step (No need to del. items has min sup count ≥ 2) ∵ all items had atleast min sup count

After pruning also same sup count maintained along with Items

Is	Sup. Count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Step (3) Joining - To discover the set of freq 2-itemsets, L_2 the algm uses the join $L_1 \bowtie L_1$ to generate a candidate set of 2 itemsets C_2

Step (4)

L_1	L_1
I_1	I_1
I_2	I_2
I_3	I_3
I_4	I_4
I_5	I_5

→ Single set converted into a set using join operation

C_2

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)

Step ④ freq 2 itemsets L_2 determined

pruning → Remove itemset which has < minsupcount

min sup count = 2

L_2

Itemset	Sup. count
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

Step ⑤ Generate 3 itemsets candidate C_3 using join step

$C_3 = L_2 \bowtie L_2$

L_2	$\bowtie$	L_2
I_1, I_2		I_1, I_2
I_1, I_3		I_1, I_3
I_1, I_5		I_1, I_5
I_2, I_3		I_2, I_3
I_2, I_4		I_2, I_4
I_2, I_5		I_2, I_5

⇒

Itemset	Sup count
I_1, I_2, I_3	2
I_1, I_2, I_5	2
I_1, I_2, I_4	1
I_1, I_3, I_5	1
I_2, I_3, I_4	0
I_2, I_3, I_5	1
I_2, I_4, I_5	0

Correlation Analysis classification

- Study of degree of linear relationship between 2 or more variables.

⇒ Relation b/w variables

Ex A, B

X, Y, Z.

Types of correlation

variables

- positive → X, Y ⇒ Both ↑sed (or) ↓sed.
  - changes equal. Ex Yield, fertilizers
- Negative → X, Y ⇒ opposite changes
  - X ↑se ⇒ Y ↓se (or) X ↓se ⇒ Y ↑se
  - Ex: Item purchase based on prize (less)
- Simple → X, Y → using only 2 variable, do correlation simply
- partial → X, Y, Z, a, b → more variables gn but we calculate only 2 var.
- Multiple → X, Y, Z, a, b → — do — Finding correlation b/w 2 but Pts based on all other variables

4 methods under Simple linear correlation

01 — Scatter diagram.

Best 02 — karl pearson correlation coefficient

03 — Spearman's Rank correlation coeff.

04 — Correlation coeff. by concurrent method.

Ex karl pearson correlation coefficient (r)

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - (\sum x)^2} - \sqrt{N\sum y^2 - (\sum y)^2}}$$

X : 1 3 5 7 9 → Ise order

Y : 2 4 6 8 10 → Ise order

} positive correlation

Calculate coefficient of correlation b/w x & y

x	y	xy	x^2	y^2
1	2	2	1	4
3	4	12	9	16
5	6	30	25	36
7	8	56	49	64
9	10	90	81	100
Σ = 25	30	190	165	220

N = 5 → Total no. of values

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} - \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5 \times 190 - 25 \times 80}{\sqrt{5(165) - 625} - \sqrt{5(220) - (900)}}$$

$(\Sigma x)^2 = 625$

$(\Sigma y)^2 = 900$

$$= \frac{950 - 750}{\sqrt{825 - 625} - \sqrt{1100 - 900}}$$

$$= \frac{200}{\sqrt{200} - \sqrt{200}}$$

$2 \Rightarrow \sqrt{2} \times \sqrt{2}$

$$= \frac{200}{200}$$

$$\boxed{r = 1}$$

Correlation Analysis classification : Basic learning / mining tasks

used to measure the relationship b/w 2 variables

var $x\uparrow \rightarrow y\uparrow$ $\Rightarrow$ +ve correlation

var $x\uparrow \rightarrow y\downarrow$ $\Rightarrow$ -ve correlation

$$r_{A,B} = \frac{\Sigma (A-A')(B-B')}{(n-1)\sigma A \sigma B}$$

gamma

Karle pearson correlation coefficient

$A'B' \Rightarrow$ Mean of A & B

$\sigma_A, \sigma_B \Rightarrow$ std. deviation of A & B

n = no. of tuples in DB.

$r \rightarrow$ 3 values $(0, -1, +1)$

$r \rightarrow +1 \rightarrow$ perfect positive correlation

$r \rightarrow 0 \rightarrow$ No correlation (no dependence) b/w 2 var.

↳ A doesn't affect B

$r \rightarrow -1 \rightarrow$ perfect -ve correlation

B u A

Ex

A	B
20	8
12	34
9	4
Σ 41	46

- $A' = \frac{20+12+9}{3} = 13.66$

(Mean of A)

$B' = \frac{8+34+4}{3} = 15.33$

(Mean of B)

$\sigma_A = \sqrt{\frac{\Sigma (A-A')^2}{n-1}}$

/

sigma

$= \sqrt{\frac{(20-13.66)^2 + (12-13.66)^2 + (9-13.66)^2}{2}}$

$= \sqrt{\frac{40.20 + 2.75 + 21.72}{2}} = \sqrt{\frac{64.68}{2}}$

$= \sqrt{32.34}$

$= 5.68$ // Ans

$\sigma_B = \sqrt{\frac{\Sigma (B-B')^2}{n-1}}$

$= \sqrt{\frac{(8-15.33)^2 + (34-15.33)^2 + (4+15.33)^2}{2}}$

$= 16.28$ // Ans.

$$r_{A,B} = \frac{\Sigma[(A-A')(B-B')]}{(n-1)\sigma A \sigma B}$$

$$= \frac{(20-13.66)(8-15.33) + (12-13.66)(34-15.33) + (9-13.66)(4-15.33)}{2 \times 5.68 \times 16.28}$$

$$= -1.xxx$$

$$\approx -1 \longrightarrow \text{approximately equal to } -1$$

ie. Negative Correlation

↳ one value ↑se, other value ↓se

↳ Inverse proportion