(Unsupervised Learning)

Association Rule Mining

Examples
- Market Basket Analysis
- Recommendation Systems

Sale
- Web log analysis

Frequent pattern → pattern occurs frequently in a dataset.

Generate association rules from frequent patterns.

Dataset

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke.

$\frac{n}{N}$

Support Count (σ)
→ Frequency of occurrence of an itemset

→ e.g in above example {Bread, Milk, Diaper} occurs two times

Support
→ Fraction of transactions that contain an itemset {Milk, Bread, Diaper}

Formula: $\frac{n}{N} = \frac{2}{5}$

support will tell frequent item

Association Rule :

1) $X \rightarrow Y$ (support)
2) $Y \rightarrow X$ (confidence)

X and Y are itemsets.

{Milk, Diapper} → {Beer}

For Association we need Rule Evaluation Metrics

Support (s) : $\frac{n}{N}$

Confidence : how often items in Y appear in transactions that contain X

$A \rightarrow B \Rightarrow P(B/A) = \frac{P(A \cup B)}{P(A)}$ or $\frac{\sigma(A,B)}{\sigma(A)}$

if we want to find support and confidence for this case

{Milk, Diaper} ⇒ Beer

$s = \frac{2}{5}$ As it occurs 2 times

$c = \frac{2/5}{3/5} = \frac{2}{3}$

$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})}$

Find all rules

support ≥ minsup threashold

confidence ≥ minconf threshold

Brute-force approach

{A, B, C}

S becau

{A} → {B, C}, {B} → {A, C}, {C} → {A, B}

we'll
ll set {A, B} → {C}, {B, C} → {A}, {A, C} → {B}
and
ne
hich

$2^3 - 2 = 6$

Apriori Algorithm

All subsets must be frequent

Anti-monotone property

$(Y \subset X)\ s(X) \geq s(Y)$

Pruning

The superset of an infrequent itemset are infrequent

two operations Join → prune

C_k : Candidate itemset of size k

L_k : Frequent itemset of size k

L_1 = {Frequent items};

for ($k=1$; $L_k != \emptyset$; $k++$)

do begin

Join step: join L_k with itself to produce C_{k+1}

PRUNE step: discard $(k+1)$-itemsets from C_{k+1} that contain non-frequent k-itemsets as subsets.

C_{k+1} = candidates generated from L_k.

for each transaction t in database do increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min-support.

end

return $\cup_k L_k$;

union of all those

Two items must be common for joining

Step 1:

self-joining L_k

Joining ($C_k = L_{k-1} \bowtie L_{k-1}$)

The join, $L_{k-1} \bowtie L_{k-1}$, is performed only if their first (k-2) items are in common.

Example of Candidate-generation

$L_3 = \{abc, abd, acd, ace, bcd\}$

Self-joining: $L_3 * L_3$

~~abcd, acde, two items~~

abcd, ~~ab~~ → abc, abd

acde → acd, ace

Pruning:

abcd

↓

abc ✓, abd ✓, bcd ✓

all subsets are present in L_3

acde ✗

↓

acd ✓, ade ✗ not include

$C_4 = \{abcd\}$

Apriori Example

Transaction ID	Items
T1	{I1, I3, I4, I6}
T2	{I2, I3, I5}
T3	{I1, I2, I3, I6}
T4	{I1, I5, I6}

Iteration 1

C_1

Itemset	Count
{I1}	3
{I2}	2
{I3}	3
{I4}	1
{I5}	3
{I6}	2

L_1

Itemset	Count
{I1}	3
{I2}	2
{I3}	3
{I5}	3
{I6}	2

0.5

$\frac{0.5}{100} * 4$

Min support threshold 0.5

= 50 %.

$= \frac{50}{100} \times 4 = 2$

$\frac{50}{100} \times 4$

$= 2$

$L_1 \bowtie L_1$

k = 2 element comma

Iteration 2

$L_1 \times L_1$ (k-2) 2-2 = 0

Itemset	Item	4-2=2
{I1, I2}	{I1}	
	{I2}	
	{I5}	
	{I5}	
	{I6}	

{I1, I2}

b

C2

{I1, I2}	1 X
{I1, I3}	2
{I1, I5}	2
{I1, I6}	2
{I2, I3}	2
{I2, I5}	2
{I2, I6}	0 X
{I3, I5}	2
{I3, I6}	1 X
{I5, I6}	1 X

L_2

{I1, I3}	2
{I1, I5}	2
{I1, I6}	2
{I2, I3}	2
{I2, I5}	2
{I3, I5}	2

1 element should be Comm°

$L_2 \times L_2$

$\{I_1, I_3\}$	$\{I_1, I_3\}$
$\{I_1, I_5\}$	$\{I_1, I_5\}$
$\{I_1, I_6\}$	$\{I_1, I_6\}$
$\rightarrow \{I_2, I_3\}$	$\{I_2, I_3\}$
$\{I_2, I_5\}$	$\{I_2, I_5\}$
$\{I_3, I_5\}$	$\{I_3, I_5\}$

~~$\{I_1, I_3, I_5\}$~~
~~$\{I_1, I_3\}$~~
$\{I_1, I_3, I_5\}$
$\{I_1, I_3, I_6\}$
~~$\{I_2, I_3, I_5\}$~~
$\{I_1, I_5, I_6\}$
$\{I_2, I_3, I_6\}$

	~~4~~
	4
	~~3~~

$\downarrow$

C_3

✓ $\rightarrow$	$\{I_1, I_3, I_5\}$	$\{I_1, I_3\}, \{I_1, I_5\}, \{I_3, I_5\}$ ✓
$\rightarrow$ X	$\{I_1, I_3, I_6\}$	$\{I_1, I_3\}, \{I_3, I_6\}, \{I_1, I_6\}$ X
$\rightarrow$	$\{I_1, I_5, I_6\}$	$\{I_1, I_5\}, \{I_5, I_6\}, \{I_1, I_6\}$ X
✓ $\rightarrow$	$\{I_2, I_3, I_5\}$	$\{I_2, I_3\}, \{I_3, I_5\}, \{I_2, I_5\}$ ✓

~~C_3~~

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

C_3

$\{I_1, I_2, I_3\}$	1
$\{I_2, I_3, I_5\}$	2 ✓

L_3

$\{I_2, I_3, I_5\}$	2

two sets need to join

C_4 is empty

Frequent Item set $\{I_2, I_3, I_5\}$

L_1	
$\{I_1\}$	3
$\{I_2\}$	2
$\{I_3\}$	3
$\{I_5\}$	3
$\{I_6\}$	2

L_2	
$\{I_1, I_3\}$	2
$\{I_1, I_5\}$	2
$\{I_1, I_6\}$	2
$\{I_2, I_3\}$	2
$\{I_2, I_3\}$	2
$\{I_3, I_5\}$	2

L_3
$\{I_2, I_3, I_5\}$
$\{I_2, I_3, I_5\}$
$\{I_2, I_3\} \rightarrow \{I_5\}$

for L_3

$\{I_2, I_3\} \rightarrow \{I_5\}$	$2/2$
$\{I_3, I_5\} \rightarrow \{I_2\}$	$2/2$
$\{I_2, I_5\} \rightarrow \{I_3\}$	$2/2$
$\{I_2\} \rightarrow \{I_3, I_5\}$	$2/5$
$\{I_3\} \rightarrow \{I_2, I_5\}$	$2/3$
$\{I_5\} \rightarrow \{I_2, I_3\}$	

$$\frac{P(I_2 \cup I_3 \cup I_5)}{P(I_2 \cup I_3)}$$

60

Example 2.

Transaction ID	Items
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Transaction ID	Items
T100	{E, K, M, N, O, Y}
T200	{D, E, K, N, O, Y}
T300	{A, E, K, M}
T400	{C, K, M, U, Y}
T500	{C, E, I, K, O, O}

Q

Itemset	Count
{A}	1
{C}	2
{D}	1
{E}	4
{I}	1
{K}	5
{M}	3
{N}	2
{O}	3
{U}	1
{Y}	3

Min Support
60%.
$= \frac{60}{100} \times 5$
$= 3$.

L_1

Itemset	Count
{G}	4
{K}	5
{M}	3
{O}	3
{Y}	3

L_1 Itemset	L_1 Item
{E}	{E}
{K}	{K}
{M}	{M}
{O}	{O}
{Y}	{Y}

{E, K}

C_2

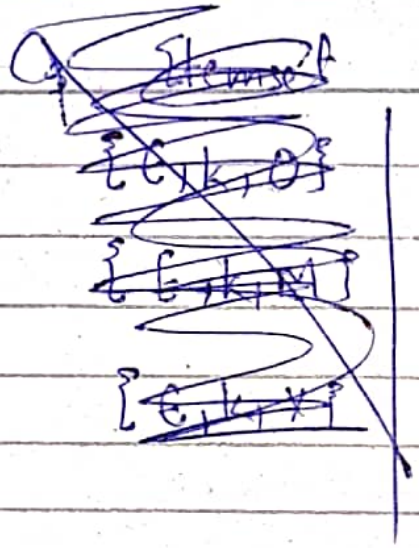
Itemset	
{E, K}	4
{E, M}	2
{E, O}	3
{E, Y}	2
{K, M}	3
{K, O}	3
{K, Y}	3
{M, O}	1
{M, Y}	2
{O, Y}	2

L_2

	Itemset	Count
{E, K} 4	{E, O}	3
	{K, M}	3
	{K, O}	3
	{K, Y}	3
	~~{E, K}~~	1

$L_2 \times L_2$

{E, K}	{E, K}
{E, O}	{E, O}
{K, M}	{K, M}
{K, O}	{K, O}
{K, Y}	{K, Y}



C_2

Itemset
{C, K, O}
{K, M, O}
{K, M, Y}
{K, O, X}

Itemset	
{C K, O}	{C, K}, {K, O}, {C, O}
{K, M, O}	{K, M}, {K, O}, {M, O}
{K, M, Y}	{K, M}, {M, Y}, {K, Y}
{K, O, Y}	{K, O}, {O, Y}, {K, Y}

C_3

{C, K, O}	3

C_4 is empty

L_3

{C, K, O}	3

L_4 is empty

L1 L2 L3

Itemset	Count
{E}	4
{K}	5
{M}	3
{O}	3
{Y}	3

L2

Itemset	Count
{E,K}	4
{E,O}	3
{K,M}	3
{K,O}	3
{K,Y}	3

L3

Itemset	Count
{E,K,O}	3

{E,K} → {O}	3/4	= 75 %
{K,O} → {E}	3/3	= 100 %
{E,O} → {K}	3/3	= 100 %
{O} → {E,K}	3/3	= 100 %
{E} → {K,O}	3/4	= 75 %
{K} → {E,O}	3/5	= 60 %

FP Growth Algorithm

Frequent Pattern

→ tree

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, ~~I3, I5~~
T800	I1, I2, I3, I5
T900	I1, I2, I3

We have to scan database once

Itemset	Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

order →

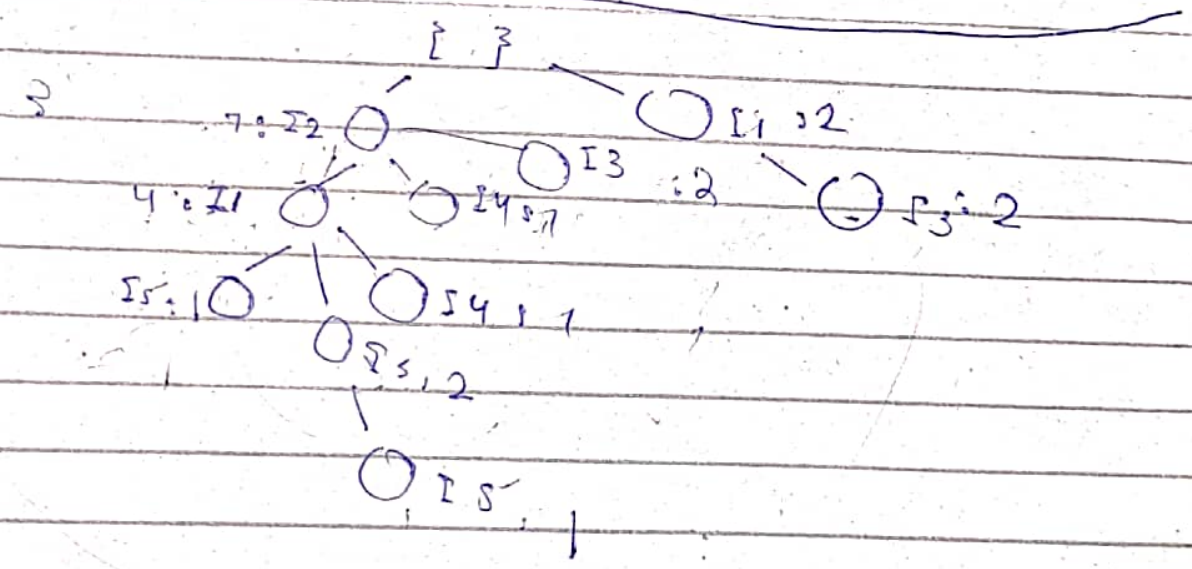
Itemset	Count
{I2}	7
{I1}	6
{I3}	6
{I4}	2
{I5}	2

Construct FP tree

arrange the table having more count

Arranged Table

Itemset	Count
T100	I2, I1, I5
T200	I2, I4
T300	I2, I3
T400	I2, I1, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I2, I1, I3, I5
T900	I2, I1, I3

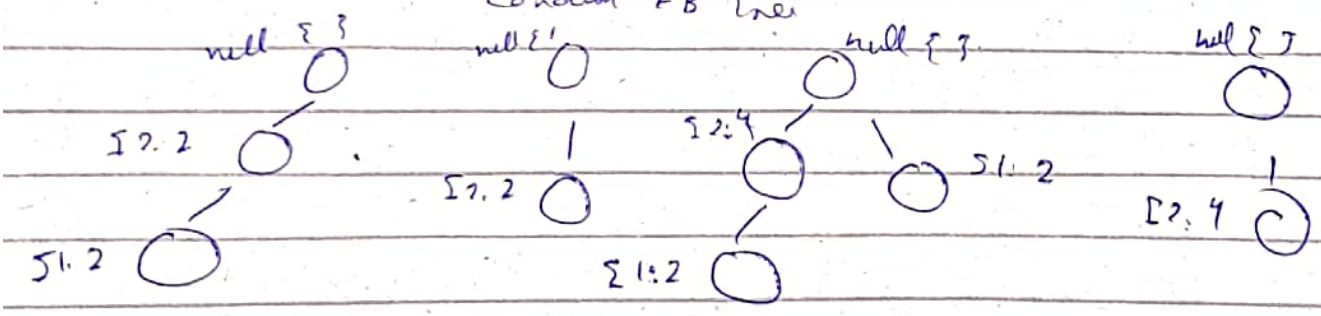


Item ID	Count
{I5}	2
{I4}	2
{I3}	6
{I1}	6
{I2}	7

Item ID	Conditional Pattern Base
{I5}	{I2, I1 : 1}, {I2, I1, I3 : 1}
{I4}	{I2, I1 : 1}, {I2 : 1}
{I3}	{I2, I1 : 2}, {I2 : 2}, {I1 : 2}
{I1}	{I2 : 4}

Item ID	Conditional FP Tree
{I5}	<I2 : 2, I1 : 2>
{I4}	<I2 : 2>
{I3}	<I2 : 4, I1 : 2> <I1 : 2>
{I1}	<I2 : 4>

Condition FP Tree



Item ID	Frequent Pattern set
{I5}	{{I2, I5: 2}, {I1, I8: 2}, {I2, I1, I5: 2}}
{I4}	{I2, I4: 2}
{I3}	{{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}}
{I1}	{I2, I1: 4}
{I2}	

Frequent Pattern

{{I1}, {I2}, {I3}, {I4}, {I5}, {I2, I5}, {I1, I5}, {I2, I4}, {I2, I3}, {I1, I3}, {I2, I1}, {I2, I1, I5}}

Example 3

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

min sup = 60%

= 60%

= 60/100 × 5

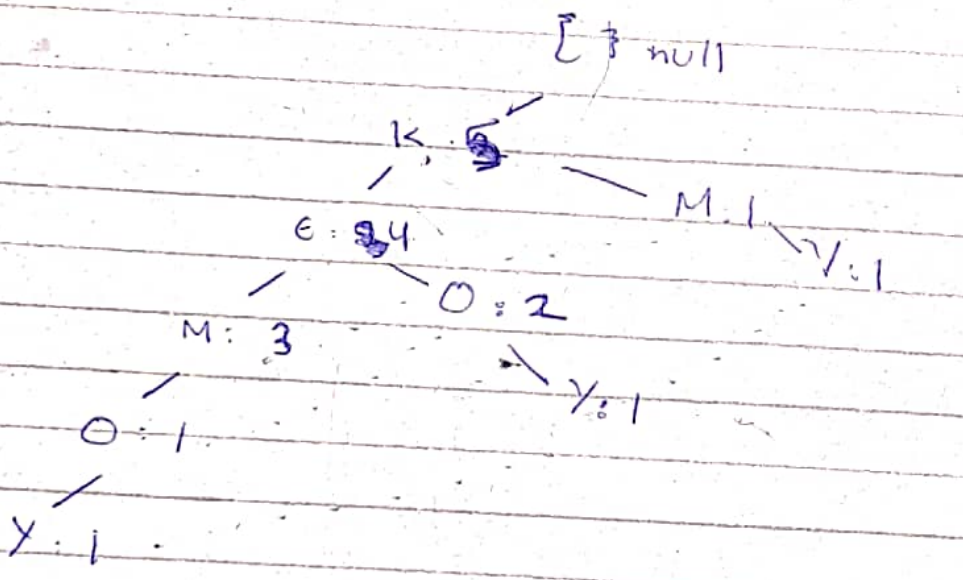
= 3

Itemset	Count
{M}	3
{O}	3
{N}	2
{K}	5
{E}	4
{Y}	3
{D}	1
{A}	1
{U}	1
{C}	2
{I}	1

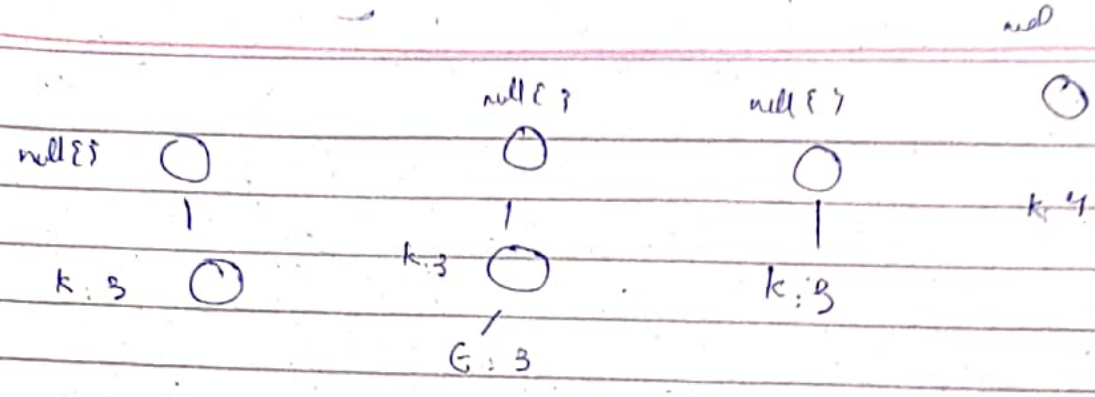
Itemset	Count
{M}	3
{O}	3
{K}	5
{E}	4
{Y}	3

Itemset	Count
{K}	5
{E}	4
{M}	3
{O}	3
{Y}	3

Transaction ID	Items
T100	{K, E, M, O, Y}
T200	{K, E, O, Y}
T300	{K, E, M}
T400	{K, M, Y}
T500	{K, E, O}



Item ID	Conditional Pattern Base
{Y}	{{K, E, M, O : 1}, {K, E, O : 1}, {K, M : 1}}
{O}	{{K, E, M : 1}, {K, E : 2}}
{M}	{{K, E : 3}, {K : 1}}
{E}	{{K : 4}}
{K}	



Item ID	Pattern Base Conditional FP Tree	Conditional FP Tree
{Y}	{{k,E,M,O,Y:1}, {k,G,O:1}, {k,M:1}}	{k:3}
{O}	{{k,E,M:1}, {k,E:2}}	{k,E:3}
{M}	{k,E:2}, {k:1}	{k:3}
{E}	{k:4}}	{k:4}
{k}		

Frequent Pattern

{kY:3}
{kO:3, EO:3, kEO:3}
{kM:3}
{kE:4}

Frequent: {Y} {O} {M}, {E}, {k},
{kY}, {kO}, {EO}, {kEO}, {kM},
{kE}