09-09-2020

Naman Garg
B032
BTech CS B.

DATA MINING ~~ASSIGNMENT~~ 2

Apriori Algorithm : Support = 60% = $\frac{60}{100} \times 5 = 3$

Q1 (i) Count number & support for each item C_1

C_1	Support Count
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

FREQUENT ITEMSET supcount >= 3

L_1	Support Count
E	4
K	5
M	3
O	3
Y	3

(ii) SELF JOIN: $L_1 \rightarrow C_2$ $L_1 \bowtie L_1$

$L_1 = \{E, K, M, O, Y\}$

C_2	Support count
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

Frequent item set L_2
Support Count >=3

L_2	Support Count
E, K	4
E, O	3
K, M	3
K, O	3
K, Y	3

(iii) Self join $L_2 \rightarrow C_3$ using the prune technique i.e. we must check the fact that all non-empty subsets of a frequent itemsets must be a frequent itemset as well.

$L_2 = \{\Sigma K, \Sigma O, KM, KO, KY\} \bowtie \{\Sigma K, \Sigma O, KM, KO, KY\}$
$= \{\Sigma KO, \Sigma KM, \Sigma KY, KMO, KMY\}$.

$\Sigma KM \Rightarrow$ as $\{\Sigma, K\}$ and $\{K, M\}$ are frequent item sets but $\{\Sigma, M\}$ is not here hence, we remove it.

$\Sigma KY \Rightarrow \{\Sigma Y\}$ is not frequent item set, we remove it.

all subsets of:
$\{\Sigma KO\}$, $\{KMY\}$ and $\{KMO\}$ are included as frequent subsets ∴ we keep them.

thus, we have, :

C_3	Support Count
Σ, K, O	3
K, M, O	1
K, M, Y	2

∴ $L_3 \Rightarrow$

L_3	S.Count
Σ, K, O	3

there is only 1 value in L_3 ∴ we cannot make an C_4.
∴ No more associations are possible.

ASSOCIATION RULES

1. $[K, O] \rightarrow \Sigma = 3/3 = 100\%$.
2. $[\Sigma, K] \rightarrow O = 3/4 = 75\%$.
3. $[\Sigma, O] \rightarrow K = 3/3 = 100\%$.
4. $\Sigma \rightarrow [K, O] = 3/4 = 75\%$.
5. $O \rightarrow [\Sigma, K] = 3/3 = 100\%$.
6. $K \rightarrow [\Sigma, O] = 3/5 = 60\%$.

we want confidance $\geq 80\%$.
∴ we discard 2, 4, 6 ∧ we keep 1, 3, 5 formulas they have confidance ≥ 80

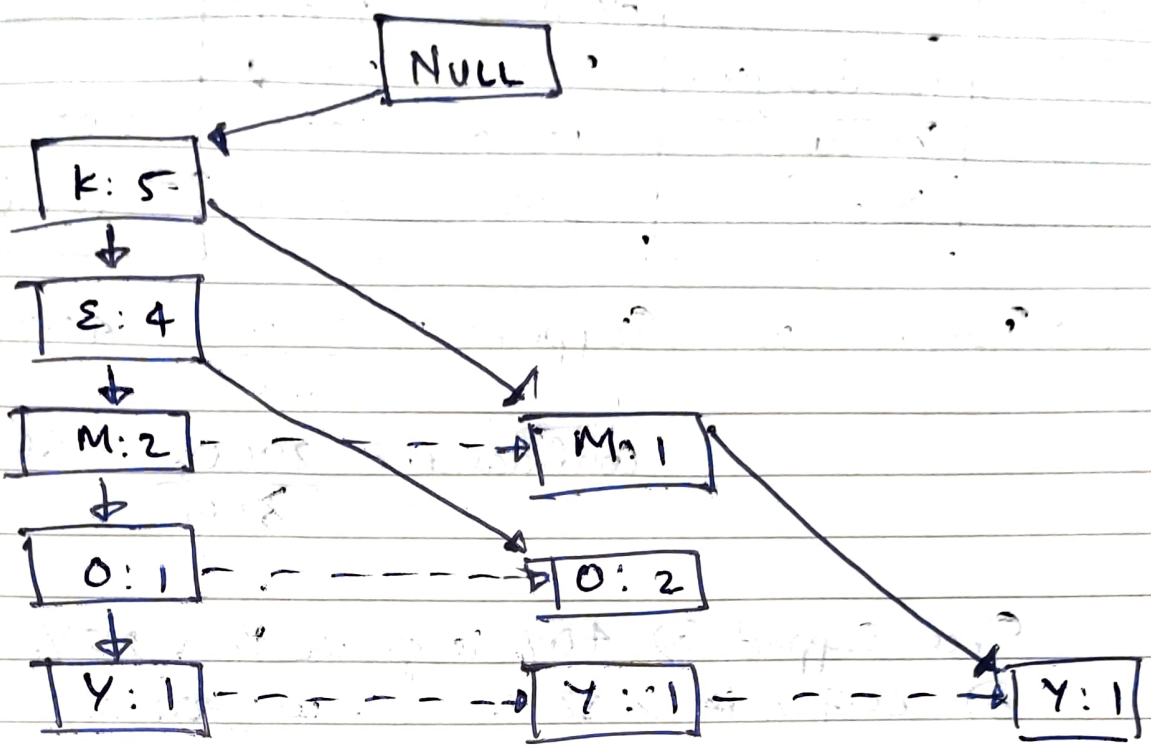
Using FP Growth : Min Support = 60% $\frac{60}{100} \times 5 = 3$

Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

∴ Frequent Pattern set L = {K:5, E:4, M:3, O:3, Y:3}

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

FP Tree



Items	Frequent Pattern Generated
Y	{<K, Y: 3>}
O	{<K, O: 3>, <E, O: 3>, <E, K, O: 3>}
M	{<K, M: 3>}
E	{<E, K: 3>}
K	

Association Rules: Conf

1. [E, K] → O = 3/4 = 75%.
2. [K, O] → E = 3/3 = 100%.
3. [E, O] → K = 3/3 = 100%.
4. E → [K, O] = 3/4 = 75%.
5. K → [E, O] = 3/5 = 60%.
6. O → [E, K] = 3/3 = 100%.

Rules 1, 4, 5 are discarded because we want confidance ≥ 80%. & rules 2, 3, 6 are selected.

Q2

	Hotdogs	Not hotdogs	Σ row
hamburgers	2000	500	2500
Not hamburgers	1000	1500	2500
Σ col.	3000	2000	5000

(i) For the rule, Support = $\frac{2000}{5000} = 40\%$.

Confidance = $\frac{2000}{5000} = 66.67\%$.

Since Support of 40% is greater than 25% & confidance of 66.7% > 50%. ∴ the association rule is STRONG

(ii) Corr (hotdog, hamburger) = $\frac{P(\{\text{hotdog, hamburger}\})}{P(\{\text{hotdog}\})\,P(\{\text{hamburger}\})}$

$= \frac{0.4}{(0.5 \times 0.6)} = 1.33 > 1$

∴ Since the corr value > 1 the purchase of hotdogs is NOT idependent of the purchase of hamburgers. Then is a POSITIVE correlation b/w them.