Data-240 sec 12 Data Mining / Analysis

HW-4

Prayag Nikul Purani (01741673 7)

Question 1 (frequent item set & Association rules)

T10 — 1, 3, 4

T20 — 2, 3, 5

T30 — 1, 2, 3, 5

T40 — 2, 5

min_sup = 40%

min_conf = 70%

a)

Scan 1 ⇒

items	sup	min_sup %
1	2	50%
2	3	75%
3	3	75%
~~4~~	~~1~~	~~25%~~
5	3	75%

Pruning 1 ⇒ {1}, {2}, {3}, {5}

frequent 1-itemset (L_1) ⇒ sup > 40%

{1}, {2}, {3}, {5}

DOMS | Page No.
Date / /

Scan 2 ⇒

{1, 2}	1	25%
{1, 3}	2	50%
{1, 5}	1	25%
{2, 3}	2	50%
{2, 5}	3	75%
{3, 5}	2	50%

Pruning 2 ⇒ {1, 3}, {2, 3}, {2, 5}, {3, 5}

Frequent 2-itemset (L_2) sup ≥ 40%

Scan 3 ⇒

{1 2 3}	1	25%
{1 2 5}	1	25%
{2 3 5}	2	50%

Frequent 3-itemset (L_3) sup ≥ 40%

{2 3 5}

b) Association rules

from {2 3 5}

we will generate all the combination of association rule and then will pruning them based on the > min-conf = 70%.

DOMS | Page No.
Date / /

	sup(X∪Y)	sup(X)	conf
{2 3} → {5}	2	2	100%
{2 5} → {3}	2	3	66.67%
{3 5} → {2}	2	2	100%
{2} → {3}	2	3	66.67%
{3} → {2}	2	3	66.67%
{2} → {5}	3	3	100%
{5} → {2}	3	3	100%
{3} → {5}	2	3	66.67%
{5} → {3}	2	3	66.67%

final strong association rules

{23} → {5} conf = 100%

{35} → {2} conf = 100%

{2} → {5} conf = 100%

{5} → {2} conf = 100%

Question 2 (Association rule mining)

a) {hog dogs} ⟶ {hamburgers}

$$sup(\text{hot dog} \cup \text{hamburgers}) = \frac{2000}{5000} = 0.4$$

40%.

⟹ since $sup(\text{hot dog} \cup \text{hamburger}) > 25\%$.

$$conf(\text{hog dogs} \rightarrow \text{hamburgers})$$

$$= \frac{sup(\text{hot dog} \cup \text{hamburger})}{sup(\text{hot dog})}$$

$$= \frac{\frac{2000}{5000}}{\frac{3000}{5000}} = \frac{2000}{3000} = 66.67\%$$

since confidence (66.67%) > 50%.

So, hot dogs ⟹ hamburger is a strong association rule as both

sup (40%) > min_sup (25%)

conf (66.67%) > min_conf (50%)

meet the threshold.

DOMS | Page No.
Date / /

b)

To, check for if the purchases of hot dogs & hamburgers are independent, we compare expected & observed value

$$\text{Expected count} = \frac{3000 \times 2500}{5000} = 1500$$

⇒ Since the observed value for both is 2000 is higher than expected value (1500) so purchases of hot dogs & hamburger are independent.

Correlation → 2000 > 1500, there is a +ve correlation between hot dog and hamburgers.

c)

$$\text{Lift} = \frac{P(HD \cup H)}{P(HD) \times P(H)}$$

$$= \frac{0.4}{0.6 \times 0.5} = 1.33$$

1.33 > 1 so there is a +ve association.

2) Correlation

Observed joint frequency of HDf H is higher than expected

so +ve correlation b/t hotdogs & hamburgers

3) All confidence (min conf)

= min(P(hotdog), P(hamburger))
= min(0.6, 0.5)
= 0.5

as all confidence (0.5) is lower than confidence (0.667) showing lower bound of support across both items.

4) Max confidence

= max(P(hotdog), P(hamburger))
= max(0.6, 0.5)
= 0.6

Max conf higher than All conf but still lower than actual conf. This measure shows the potential max sup. for either of items

DOMS | Page No.
Date / /

5) Cosine measure

$$= \sqrt{P(A|B) \times P(B|A)}$$

$$= \sqrt{\frac{P(A \cup B)}{P(B)} \times \frac{P(B \cup A)}{P(A}}$$

$$= \sqrt{\frac{P(HO \cup H)}{P(HO)} \times \frac{P(HO \cup H)}{P(H)}}$$

$$= \sqrt{\frac{2000}{2500} \times \frac{2000}{3000}} = 0.73$$

0.73 indicates a strong relationship b/t HO & H by balancing both conditional probabilities based toward either item.

All Conf ⇒ conservative lower bound
useful when stability across both items is desired.

Max conf ⇒ highest likelihood → overestimation

Cosine measure ⇒ balances both item set ⇒
no bias without favouring either item.

DOMS | Page No.
Date / /

Lift ⇒ a direct measure of strength, by comparing joint probability with independence.

Correlation ⇒ Offers directional insights if relationship is +ve / -ve.

Question 3 Sequential pattern mining with prefixSpan

S_1 $\langle (ab)ca \rangle$
S_1^1 $\langle (ab)bc \rangle$
S_3^2 $\langle bcd \rangle$
S_4^3 $\langle b(ab) \rangle$

min_sup = 3

a) Length 1 sequential patterns

$\langle a \rangle$	3
$\langle b \rangle$	4
$\langle c \rangle$	4
~~$\langle d \rangle$~~	~~1~~

so length 1 pattern
$\Rightarrow \langle a \rangle, \langle b \rangle, \langle c \rangle$

b) Projection of database $\langle a \rangle$

S_1 $\langle (\_b)ca \rangle$
S_1^1 $\langle (\_b)bc \rangle$
S_4^2 $\langle (\_b) \rangle$

Projection of database $\langle b \rangle$

S_1 $\langle ca \rangle$
S_1^1 $\langle bc \rangle$
S_3^2 $\langle cd \rangle$
S_4^3 $\langle (ab) \rangle$

Projection of <c>

S_1 <a>
S_2 < >
S_3 <d>

c) projection of <a>

S_1 <(_b)c a>
S_2 <(_b) bc>
S_3 <(_b)>
S_4

<a a> — 1
<a b> — 1
<a c> — 2
<(a b)> — 3

min_sup = 3

projs of <b>

S_1 < c a>
S_2 < b c>
S_3 < c d>
S_4 <(ab)>

<b a> — 2
<b b> — 2
<b c> — 2
<b d> — 1

projection of <c>

S_1 <a>
S_2 < >
S_3 <d>

<c a> — 1
<c d> — 1

So the only frequent sequence of length - 2 is <(a b)>

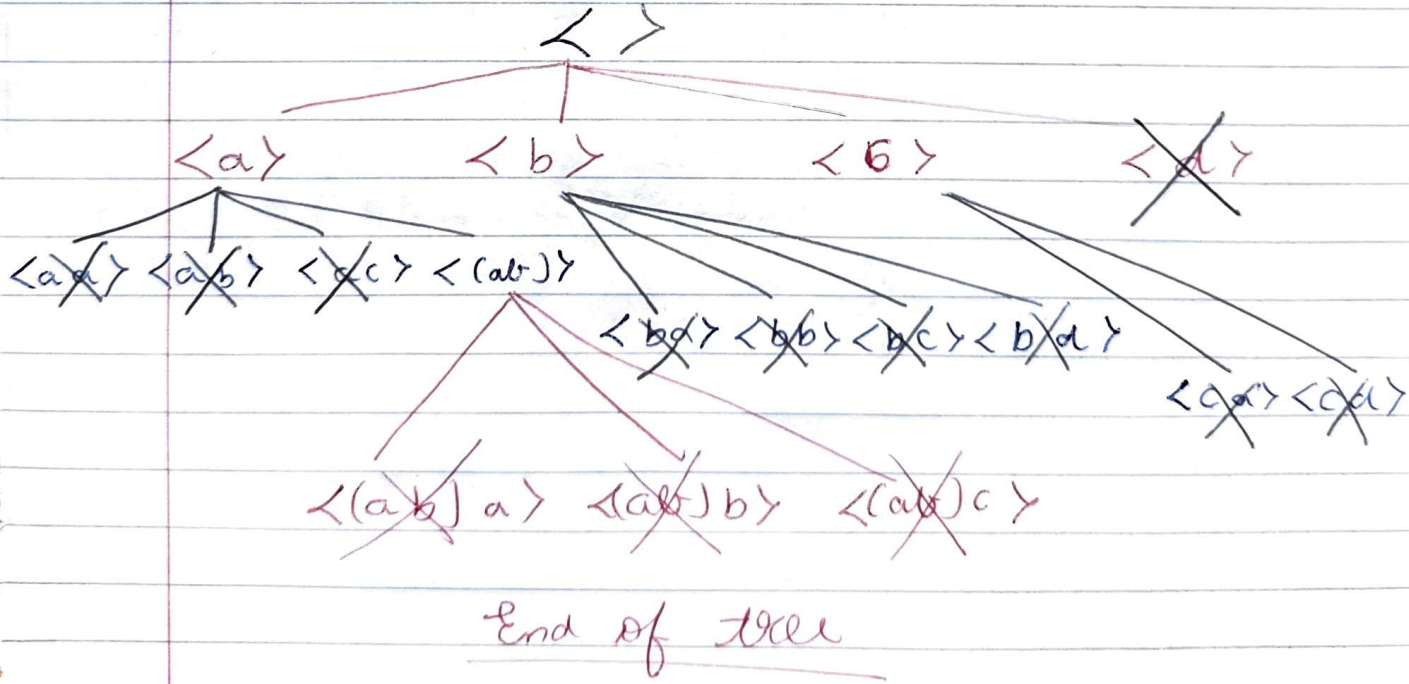
d) Projection of <(a b)>.

S_1 <c a>
S_2 <b c>
S_3 <>

<(ab) a> — 1
<(ab) b> — 1
<(ab) c> — 2

as min_sup = 3 and we will not have any more frequents to proceed with.

Depth-first search



End of tree