Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 6 (Descriptive Analytics - part 1)

Q1. Find all association rules in the following database:
(minsup = 2, minconf = 70%)

TID	Items
1	a, b, c
2	b, c, d, e
4	c, d
3	a, b, d
5	a, b, c

Ans:

Firstly, find all itemsets with support >=2;

C_1

set	supp
{a}	3
{b}	4
{c}	4
{d}	3
~~{e}~~	~~1~~

C_2

set	supp
{a,b}	3
{a,c}	2
~~{a,d}~~	~~1~~
{b,c}	3
{b,d}	2
{c,d}	2

C_3

set	supp
{a,b,c}	2
~~{b,c,d}~~	~~1~~

F = {{a}, {b}, {c}, {d}, {a,b} , {a,c}, {b,c}, {b,d}, {c,d}, {a,b,c}}

Then, split all frequent itemsets in all possible ways

BACS3013 Data Science

set	supp
{}	5
{a}	3
{b}	4
{c}	4
{d}	3
{a,b}	3
{a,c}	2
{b,c}	3
{b,d}	2
{c,d}	2
{a,b,c}	2

(Trivial rules: X => {} 100%)

ac => b	100%
a => b	100%
b => a	75%
b => c	75%
c => b	75%
{} => b	80%
{} => c	80%

BACS3013 Data Science

Q2. For the confidence of association rules, a monotonicity principle can be stated between rules that are based on the same itemset (X=>Y is based on X∪Y). Explain how can it be exploited?

Ans:

Let $I = X \cup Y = X' \cup Y'$, and let $X' \subseteq X$

Then: $sup(X => Y) = sup(X' => Y')$

$conf(X => Y) \geq conf(X' => Y')$

Example:

$conf(\{a,b\} => \{c\}) \geq conf(\{a\} => \{b,c\})$

$\geq conf(\{\} => \{a,b,c\})$

$sup(X => Y) = sup(X \cup Y) = sup(I)$

$= sup(X' \cup Y') = sup(X' => Y')$

$conf(X => Y) = sup(X \cup Y) / sup(X)$

$\geq sup(X \cup Y) / sup(X')$ $(X' \subseteq X)$

$= sup(X' \cup Y') / sup(X)$ $(X \cup Y = X' \cup Y')$

$= conf(X' => Y')$

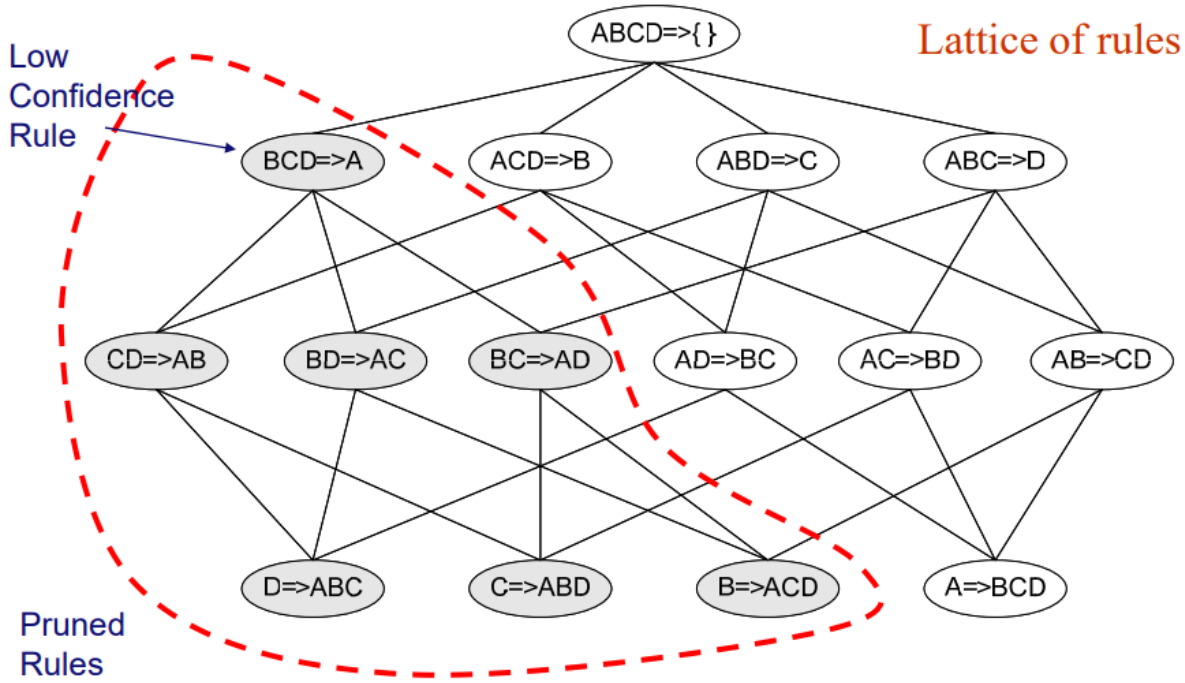
Note:

When splitting I into $I = X \cup Y$, first consider singleton Y

Only consider X => Y if:

for all $y \in Y$, $X \cup \{y\} => Y \setminus \{y\}$ was confident

BACS3013 Data Science



Q3. A transaction database is shown below:

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

BACS3013 Data Science

a. Why is every subset of a frequent set frequent?
b. Given frequent set {A,B,E}, what are the possible association rules?
c. Given frequent set {A,B,E}, what association rules have minsup=2 and mincof =50%?

Ans:

a. Suppose {A,B} is frequent. Since each occurrence of A, B includes both A and B, then bnoth A and B must also be frequent.

b.

- A => B, E
- A, B => E
- A, E => B
- B => A, E
- B, E => A
- E => A, B
- __ => A,B,E (empty rule), or true => A,B,E

c.

A, B => E : conf=2/4 = 50%

A, E => B : conf=2/2 = 100%

B, E => A : conf=2/2 = 100%

E => A, B : conf=2/2 = 100%

Don't qualify

A =>B, E : conf=2/6 =33%< 50%

B => A, E : conf=2/7 = 28% < 50%

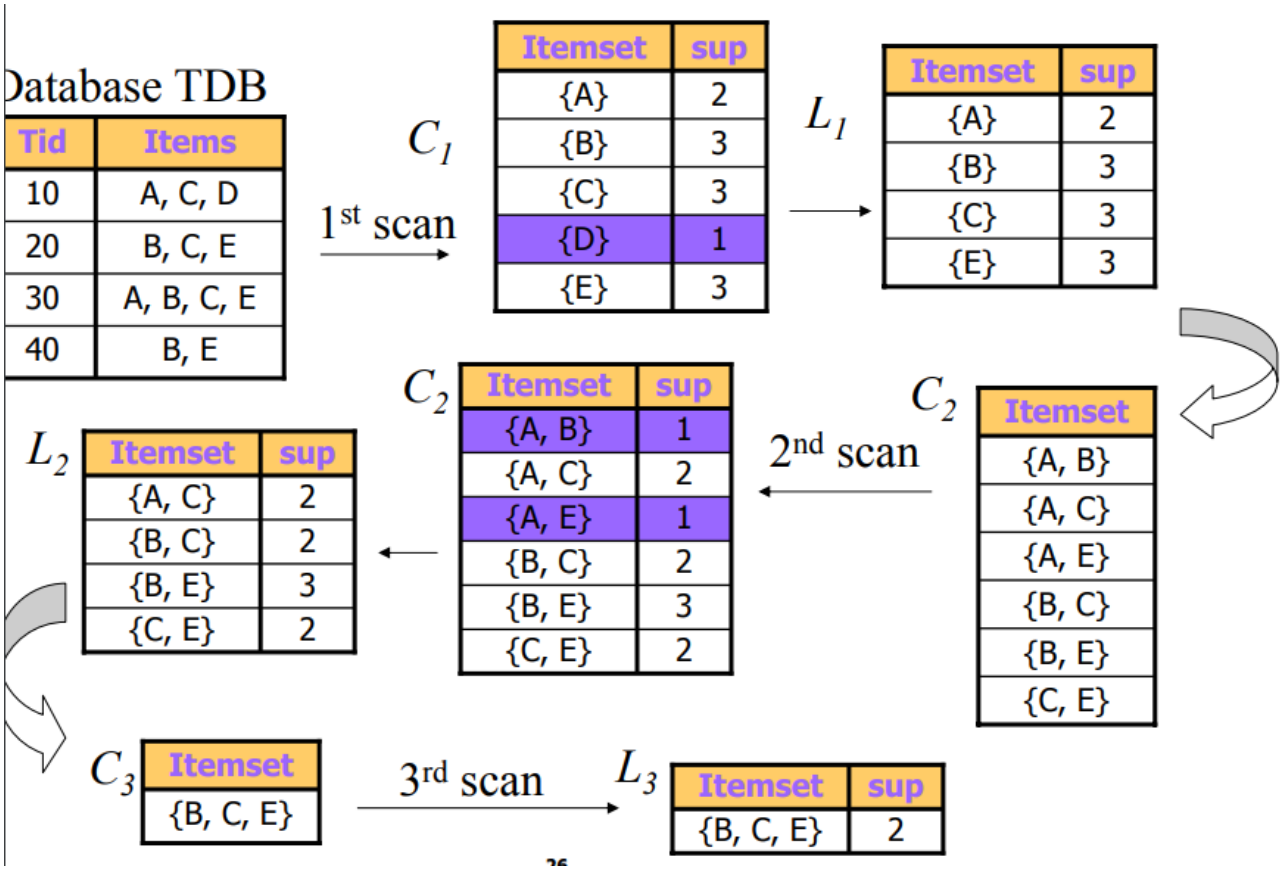
__ => A,B,E : conf: 2/9 = 22% < 50%

Q4. By using Apriori algorithm, mine the association rules for the following transaction:

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

Ans:

BACS3013 Data Science



Q5. The following is an example of customer purchase transaction data set.

CID	TID	Date	Items Purchased
1	1	01/01/2001	10,20
1	2	01/02/2001	10,30,50,70
1	3	01/03/2001	10,20,30,40
2	4	01/03/2001	20,30
2	5	01/04/2001	20,40,70
3	6	01/04/2001	10,30,60,70

3	7	01/05/2001	10,50,70
4	8	01/05/2001	10,20,30
4	9	01/06/2001	20,40,60
5	10	01/11/2001	10,20,30,60

Note: CID = Customer ID and TID = Transactions ID

a) Calculate the support, confidence and lift of the following association rule. Indicate if the items in the association rule are independent of each other or have negative or positive impacts on each other.
{10} -> {50,70}

b) The following is the list of large two item sets. Show the steps to apply the Apriori property to generate and prune the candidates for large three itemsets. Describe how the Apriori property is used is in the steps. Give the final list of candidate large three item sets.
{10,20} {10,30} {20,30} {20,40}

c) Does customer 1 support the sequence <{20} {50,70} {10}>? Justify your answer.

d) Calculate the support of <{10}, {30}>.

e) Based on the types of association rules discussed in class, identify which type(s) of rules {10}-> {50,70} is.

Ans:

a)
Support = Support ({10,50,70}) = 2/10 = 20%
Confidence = Support ({10,50,70})/ Support({10}) = 0.2/0.7 = 2/7 = 29%
Lift = Confidence/Support({50,70}) = 2/7/0.2 = 10/7 = 1.43 > 1

Since lift is larger than 1, it's a positive rule.

b)
{10,20} {10,30} {20,30} {20,40}
*****O: describe how the apriori property is used to decide which 2 large item sets are joined together and to determine which 3 item set should be pruned.**

Join: {10,20,30} {20,30,40}
Prune: {10,20,30} ({20,30,40} is pruned)
Final list: {10,20,30}

d)

The sequence of customer 1 is:
<{10,20} {10,30,50,70} {10,20,30,40}>

Since {20}⊆{10,20}, {50,70}⊆{10,30,50,70}, and {10}⊆{10,20,30,40},
<{20} {50,70} {10}> is contained in the sequence of customer 1. Therefore, customer 1 supports sequence <{20} {50,70} {10}>.

e)
Only customer 1 supports the sequence <{10} {30}> and there are 5 customers, therefore,

Support = 1/5 = 20%

f)
The association rule {10} -> {50,70} is a single-level, single-dimensional and Boolean association rule.