Assignment 5

Question 1: Consider a dataset for frequent set mining as in the following table where we have 6 binary features and each row represents a transaction.

	$I1$	$I2$	$I3$	$I4$	$I5$	$I6$
$T1$	0	0	1	0	1	0
$T2$	0	1	1	1	0	1
$T3$	1	0	0	0	1	0
$T4$	1	1	1	0	0	0
$T5$	0	0	0	1	0	0
$T6$	1	0	0	1	0	1
$T7$	0	0	1	1	1	1
$T8$	1	0	1	0	1	0
$T9$	1	0	0	1	0	0
$T10$	0	1	1	0	0	1

1. Illustrate the first three levels of the Apriori algorithm (set sizes 1, 2 and 3) for support threshold of 3 transactions, by identifying candidate sets and calculating their support. What are the maximal frequent sets discovered in the first 3 levels?
   For the first scan, C1 is

Transactions	Items
T1	{I3, I5}
T2	{I2, I3, I4, I6}
T3	{I1, I5}
T4	{I1, I2, I3}
T5	{I4}
T6	{I1, I4, I6}
T7	{I3, I4, I5, I6}
T8	{I1, I3, I5}
T9	{I1, I4}
T10	{I2, I3, I6}

Therefore L1 is

Items	I1	I2	I3	I4	I5	I6
Count	5	3	6	5	4	4

For the second scan, C2 is

Items	Count
{I1, I2}	1
{I1, I3}	2
{I1, I4}	2
{I1, I5}	2
{I1, I6}	1
{I2, I3}	3
{I2, I4}	1
{I2, I5}	0
{I2, I6}	2
{I3, I4}	2
{I3, I5}	3
{I3, I6}	3
{I4, I5}	1
{I4, I6}	3
{I5, I6}	1

Therefore L2 is

Items	{I2, I3}	{I3, I5}	{I3, I6}	{I4, I6}
Count	3	3	3	3

For the third scan, C3 is

Transactions	Items
{I2, I3, I4}	1
{I2, I3, I5}	0
{I2, I3, I6}	2
{I3, I5, I6}	1

Note that no count is 3 or more; therefore there are no items in L3.

2. Pick one of the maximal sets and check if any of its subsets are association rules with frequency at least 0.3 and confidence at least 0.6. Explain your answer and show all work.

   Pick the first row from L2, or $\{I2, I3\}$. Its count is 3. The count of $I2$ and $I3$ is 3 and 6, respectively. Then the association rule for $I2 \to I3$ is

$$\text{confidence} = \frac{\text{count of I2 and I3}}{\text{count of I2}} = \frac{3}{3} = 1$$

   and for $I3 \to I2$ is

$$\text{confidence} = \frac{\text{count of I2 and I3}}{\text{count of I3}} = \frac{3}{6} = 0.5$$

   Since only the confidence of $I2$ is greater than the minimum confidence of 0.6, only $I2 \to I3$ is an association rule.

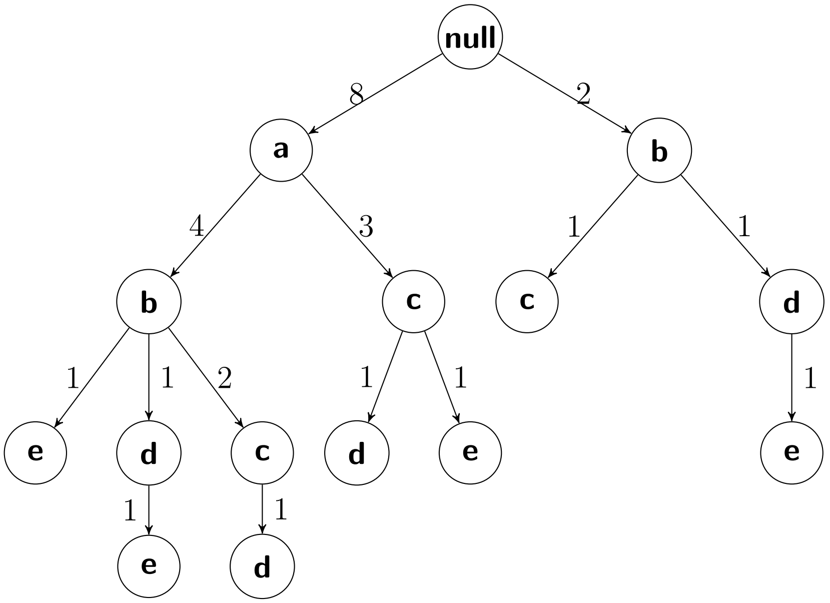
Question 2: Given the following transaction database, let the min_support = 2.

TID	Items
1	{a,b,e}
2	{a,b,c,d}
3	{a,c,d}
4	{a,c,e}
5	{b,c,f}
6	{a}
7	{a,b,c}
8	{b,d,e}
9	{a,c}
10	{a,b,d,e}

1. Construct FP-tree from the transaction database.
   Counts:

a	b	c	d	e	f
8	6	6	4	4	1

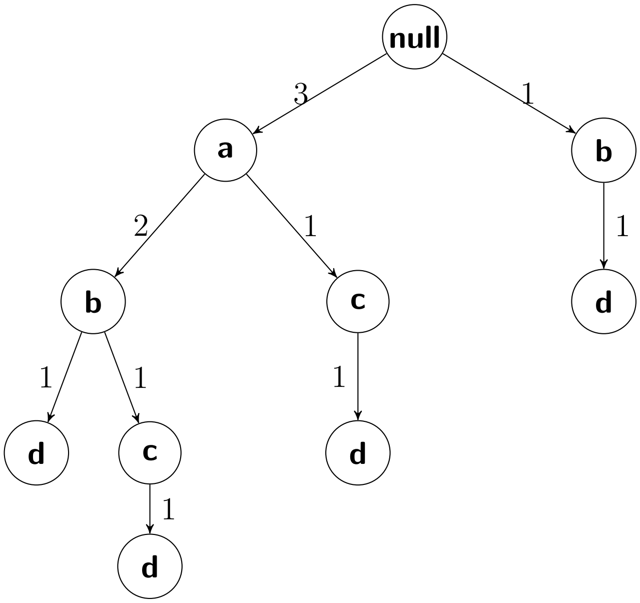
Remove f since min_support = 2.



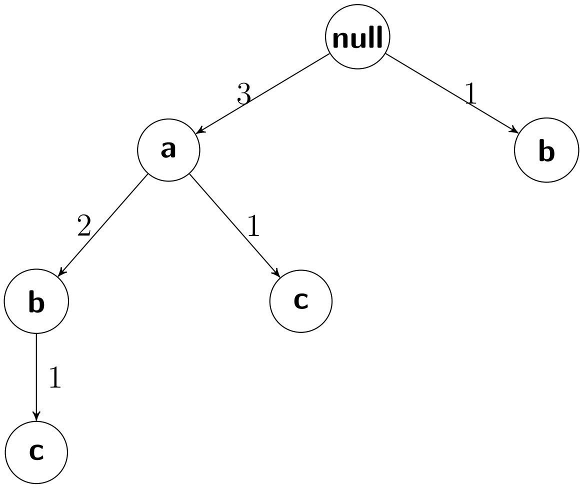
2. Show d's conditional pattern base (projected database), d's conditional FP-tree and find frequent patterns based on d's conditional FP-tree.
   d's Conditional Pattern Base:

ab	ac	abc	b
1	1	1	1

d's Conditional FP-tree:

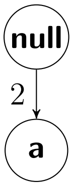


Remove d from the FP-tree and then look at Prefix trees.



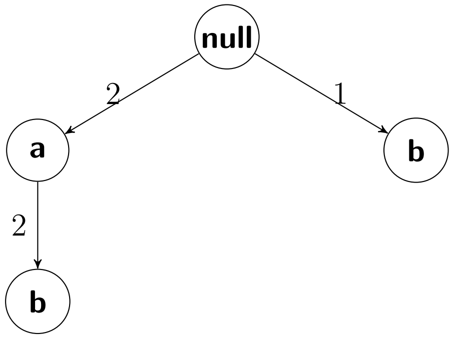
Found: ad, bd, cd since $a = 3 \geq 2$, $b = 2 \geq 2$ and $c = 2 \geq 2$.

For "ad":

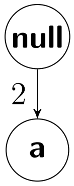


None found because a is in highest order and ad is found.

For “bd”:

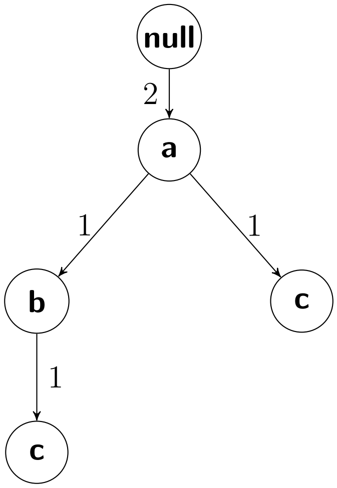


to

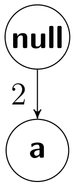


Found: abd since $b = 2 \geq 2$.

For “cd”:



to



Found acd since $a = 2 \geq 2$.

All frequent item set for d’s condition FP-tree:

ad, bd, cd, abd, acd

Question 3: In the GSP algorithm, suppose the length-3 frequent pattern set L_3 is

Row 1: $\langle\{2\}\ \{3\}\ \{4\}\rangle$

Row 2: $\langle\{2\ 5\}\ \{3\}\rangle$

Row 3: $\langle\{3\}\ \{4\}\ \{5\}\rangle$

Row 4: $\langle\{1\}\ \{2\}\ \{3\}\rangle$

Row 5: $\langle\{1\}\ \{2\ 5\}\rangle$

Row 6: $\langle\{1\}\ \{5\}\ \{3\}\rangle$

Row 7: $\langle\{5\}\ \{3\ 4\}\rangle$

Generate length-4 candidates set C_4 and frequent pattern set L_4. Show all work by writing down the details of the joins and prune steps.

For the first step, for each row set, drop the first element and compare it to the other row sets whose last element is dropped.

For Row 1, look for $\langle\{3\}\ \{4\}\rangle$:

~~$\langle\{2\}\ \{3\}\ \{4\}\rangle$~~
$\langle\{2\ 5\}\ \{3\}\rangle$
✓ $\langle\{3\}\ \{4\}\ \{5\}\rangle$
$\langle\{1\}\ \{2\}\ \{3\}\rangle$
$\langle\{1\}\ \{2\ 5\}\rangle$
$\langle\{1\}\ \{5\}\ \{3\}\rangle$
$\langle\{5\}\ \{3\ 4\}\rangle$

Set 3.

For Row 2, look for $\langle\{5\}\ \{3\}\rangle$:

$\langle\{2\}\ \{3\}\ \{4\}\rangle$
~~$\langle\{2\ 5\}\ \{3\}\rangle$~~
$\langle\{3\}\ \{4\}\ \{5\}\rangle$
$\langle\{1\}\ \{2\}\ \{3\}\rangle$
$\langle\{1\}\ \{2\ 5\}\rangle$
$\langle\{1\}\ \{5\}\ \{3\}\rangle$
✓ $\langle\{5\}\ \{3\ 4\}\rangle$

Set 7.

For Row 3, look for $\langle\{4\}\ \{5\}\rangle$:

$\langle\{2\}\ \{3\}\ \{4\}\rangle$
$\langle\{2\ 5\}\ \{3\}\rangle$
~~$\langle\{3\}\ \{4\}\ \{5\}\rangle$~~
$\langle\{1\}\ \{2\}\ \{3\}\rangle$
$\langle\{1\}\ \{2\ 5\}\rangle$
$\langle\{1\}\ \{5\}\ \{3\}\rangle$
$\langle\{5\}\ \{3\ 4\}\rangle$

None Found.

For Row 4, look for ⟨{2} {3}⟩:

✓ ⟨{2} {3} {4}⟩
⟨{2 5} {3}⟩
⟨{3} {4} {5}⟩
~~⟨{1} {2} {3}⟩~~
⟨{1} {2 5}⟩
⟨{1} {5} {3}⟩
⟨{5} {3 4}⟩

Set 1.
For Row 5, look for ⟨{2 5}⟩:

⟨{2} {3} {4}⟩
✓ ⟨{2 5} {3}⟩
⟨{3} {4} {5}⟩
⟨{1} {2} {3}⟩
~~⟨{1} {2 5}⟩~~
⟨{1} {5} {3}⟩
⟨{5} {3 4}⟩

Set 2.
For Row 6, look for ⟨{5} {3}⟩:

⟨{2} {3} {4}⟩
⟨{2 5} {3}⟩
⟨{3} {4} {5}⟩
⟨{1} {2} {3}⟩
⟨{1} {2 5}⟩
~~⟨{1} {5} {3}⟩~~
✓ ⟨{5} {3 4}⟩

Set 7.
For Row 7, look for ⟨{3 4}⟩:

⟨{2} {3} {4}⟩
⟨{2 5} {3}⟩
✓ ⟨{3} {4} {5}⟩
⟨{1} {2} {3}⟩
⟨{1} {2 5}⟩
⟨{1} {5} {3}⟩
~~⟨{5} {3 4}⟩~~

Set 3.

Therefore the length-4 candidates set $C4$ is

Row 1: $\langle\{2\}\ \{3\}\ \{4\}\ \{5\}\rangle$
Row 2: $\langle\{2\ 5\}\ \{3\ 4\}\rangle$
Row 3: $\langle\{1\}\ \{2\}\ \{3\}\ \{4\}\rangle$
Row 4: $\langle\{1\}\ \{2\ 5\}\ \{3\}\rangle$
Row 5: $\langle\{1\}\ \{5\}\ \{3\}\ \{4\}\rangle$
Row 6: $\langle\{5\}\ \{3\ 4\}\ \{5\}\rangle$

To prune, check if subsequences of above candidates are in L3. If all of a certain candidates' subsequences exist in L3, put it in L4, the length-4 frequent pattern set.
For Row 1: the subsequences are

$$\langle\{3\}\ \{4\}\ \{5\}\rangle,\ \langle\{2\}\ \{4\}\ \{5\}\rangle,\ \langle\{2\}\ \{3\}\ \{5\}\rangle,\ \langle\{2\}\ \{3\}\ \{4\}\rangle$$

of which only the first and last exist in L3.
For Row 2: the subsequences are

$$\langle\{5\}\ \{3\ 4\}\rangle,\ \langle\{2\}\ \{3\ 4\}\rangle,\ \langle\{2\ 5\}\ \{4\}\rangle,\ \langle\{2\ 5\}\ \{3\}\rangle$$

of which only the first, second, and fourth exist in L3.
For Row 3, the subsequences are

$$\langle\{2\}\ \{3\}\ \{4\}\rangle,\ \langle\{1\}\ \{3\}\ \{4\}\rangle,\ \langle\{1\}\ \{2\}\ \{4\}\rangle,\ \langle\{1\}\ \{2\}\ \{3\}\rangle$$

of which only the first and last exist in L3.
For Row 4, the subsequences are

$$\langle\{2\ 5\}\ \{3\}\rangle, \langle\{1\}\ \{5\}\ \{3\}\rangle,\ \langle\{1\}\ \{2\}\ \{3\}\rangle,\ \langle\{1\}\ \{2\ 5\}\rangle$$

of which all four subsequences exist in L3. Remove this row from L4.
For Row 5, the subsequences are

$$\langle\{5\}\ \{3\}\ \{4\}\rangle,\ \langle\{1\}\ \{3\}\ \{4\}\rangle,\ \langle\{1\}\ \{5\}\ \{4\}\rangle,\ \langle\{1\}\ \{5\}\ \{3\}\rangle$$

of which only the first and last exist in L3.
For Row 6, the subsequences are

$$\langle\{3\ 4\}\ \{5\}\rangle,\ \langle\{5\}\ \{4\}\ \{5\}\rangle,\ \langle\{5\}\ \{3\}\ \{5\}\rangle,\ \langle\{5\}\ \{3\ 4\}\rangle$$

of which only the first and last exist in L3.

Therefore,

C4
$\langle\{2\}\ \{3\}\ \{4\}\ \{5\}\rangle$
$\langle\{2\ 5\}\ \{3\ 4\}\rangle$
$\langle\{1\}\ \{2\}\ \{3\}\ \{4\}\rangle$
$\langle\{1\}\ \{5\}\ \{3\}\ \{4\}\rangle$
$\langle\{5\}\ \{3\ 4\}\ \{5\}\rangle$

L4
$\langle\{1\}\ \{2\ 5\}\ \{3\}\rangle$

Question 4: For the following two time series:

$$X = [39\ 44\ 43\ 39\ 46\ 38\ 39\ 43]$$
$$Y = [37\ 44\ 41\ 44\ 39\ 39\ 39\ 40]$$

Calculate the DTW distance between X and Y and point out the optimal warping path. (The local cost function is defined as the absolute difference of the two values, e.g., $c(x_1, y_1) = d(39, 37) = 2$)

Cost Matrix:

40	1	4	3	1	6	2	1	3
39	0	5	4	0	7	1	0	4
39	0	5	4	0	7	1	0	4
39	0	5	4	0	7	1	0	4
44	5	0	1	5	2	6	5	1
41	2	3	2	2	5	3	2	2
44	5	0	1	5	2	6	5	1
37	2	7	6	2	9	1	2	6
	39	**44**	**43**	**39**	**46**	**38**	**39**	**43**

Then after finding the optimal lowest costs, the cost matrix becomes:

40	15	18	20	6	11	13	9	11
39	14	19	17	5	12	13	8	12
39	14	15	13	5	12	9	8	12
39	14	10	9	5	12	8	8	12
44	14	5	5	9	7	13	18	19
41	9	5	4	5	10	13	15	17
44	7	2	3	8	10	16	21	22
37	2	9	15	17	26	27	29	35
	39	**44**	**43**	**39**	**46**	**38**	**39**	**43**

The optimal warping path is shown below:

40	15	18	20	6	11	13	9	11
39	14	19	17	5	12	13	8	12
39	14	15	13	5	12	9	8	12
39	14	10	9	5	12	8	8	12
44	14	5	5	9	7	13	18	19
41	9	5	4	5	10	13	15	17
44	7	2	3	8	10	16	21	22
37	2	9	15	17	26	27	29	35
	39	**44**	**43**	**39**	**46**	**38**	**39**	**43**

In total, the DTW distance between X and Y is:

$$DTW(X, Y) = 11$$