CS 412: Introduction to Data Mining
ASSIGNMENT 2

Submitted by: Sugandha
Net ID: sugandh2

Answer 1:
Base Cell1: (a1; a2; a3; a4; a5; ... ; a20)
Base Cell 2: (a1; b2; a3; b4; a5; ... ; b20)
Base Cell 3: (c1; a2; c3; a4; c5; ... ; a20)

Ans 1.1 Number of dimensions = 20
Therefore, number of non-empty cuboids = $\mathbf{2^{20}}$

Ans 1.2 Non-empty aggregate closed cells present in the cube:

- Cells of type (a1; *; a3; *; a5; *; ……; a19; *)
- Cells of type (*; a2; *; a4; ……; a20)
- 1 apex cell

Therefore, total number of distinct (non-empty) aggregate closed cells = **3**

Ans 1.3

- Number of *count 3 cells* = 1 (apex cell – overlap twice)
- Number of count 2 cells:

- To get overlapping cells from base cells 1 and 2, we need to replace all even entries in both by *. We are left with 2^{10} overlapping cells
- There is no overlap between base cells 2 and 3
- To get overlapping cells from base cells 1 and 3, we need to replace all odd entries in both cells by a *.
  Again, we are left with 2^{10} overlapping cells. Subtracting 1 for the case of apex cell,

Therefore, no. *of count 2 cells* = 2 x (2^{10} – 1) [overlap - once]

- Base cells = 3

Number of non-empty aggregate cells = Total cells – overlap twice cells *2 – overlap once cells *1 – base cells

= $3 * 2^{20} – 1*2 - (2^{10}-1)*2 – 3$

Ans 1.4 Number of aggregate cells with count >=2 =once overlapping + twice overlapping cells
= **$2 * (2^{10}-1) + 1$**

Answer 2:

Ans 2.1 True
Since Multi way array aggregation starts computations from base cuboids and proceeds towards more general cuboids, it cannot support ice berg pruning; because ice berg pruning requires descendent cells to be rejected if parents do not satisfy minimum support criteria, but here descendent cells are computed before parent cells.

Ans 2.2 False
Star cubing uses Depth First traversal, not breadth first.

Ans 2.3 False

High dimensional OLAP methodology handles high dimensional data efficiently by pre computing certain cubes, so that all computations do not have to be performed online at the time of query processing.

Ans 2.4 True
Confidence intervals are given as a quality measure in high dimensional OLAP in sampling cubes
Ans 2.5 False
Uncorrelated or weakly correlated attributes must be included in query expansion because these have the least effect on the query's semantics.

Ans 2.6 False
Prediction cube, not ranking cube stores prediction models in multidimensional data space and supports prediction in OLAP manner. Ranking cube is used for top-k query computation.

Answer 3:

T#	Items
1	b, d, f, g, l
2	f, g, h, l, m, n
3	b, f, h, k, m
4	a, f, h, j, m
5	d, f, g, j, m

Minimum support = 0.4
Therefore, minimum support count = 0.4 * 5 = 2

Ans 3.1 Apriori Algorithm:

Iteration 1:
C_1:

Itemset	Sup. Count
{a}	1
{b}	2
{d}	2
{f}	5
{g}	3
{h}	3
{j}	2
{k}	1
{l}	2
{m}	4
{n}	1

Trimming items with count< min. sup. Count

L_1:

Itemset	Sup. Count
{b}	**2**
{d}	**2**
{f}	**5**
{g}	**3**
{h}	**3**
{j}	**2**
{l}	**2**
{m}	**4**

Iteration 2:

C_2:

Itemset	Sup. Count
{b,d}	1
{b,f}	2
{b,g}	1
{b,h}	1
{b,j}	0
{b,l}	1
{b,m}	1
{d,f}	2
{d,g}	2
{d,h}	0
{d,j}	1
{d,l}	1
{d,m}	1
{f,g}	3
{f,h}	3
{f,j}	2
{f,l}	2
{f,m}	4
{g,h}	1
{g,j}	1
{g,l}	2
{g,m}	2
{h,j}	1
{h,l}	1
{h,m}	3
{j,l}	0
{j,m}	2
{l,m}	1

L_2:

Itemset	Sup. Count
{b,f}	**2**
{d,f}	**2**
{d,g}	**2**
{f,g}	**3**
{f,h}	**3**
{f,j}	**2**

{f,l}	**2**
{f,m}	**4**
{g,l}	**2**
{g,m}	**2**
{h,m}	**3**
{j,m}	**2**

Iteration 3:

C_3:

Itemset	Sup. Count
{d,f,g}	2
{f,g,l}	2
{f,g,m}	2
{f,h,m}	3

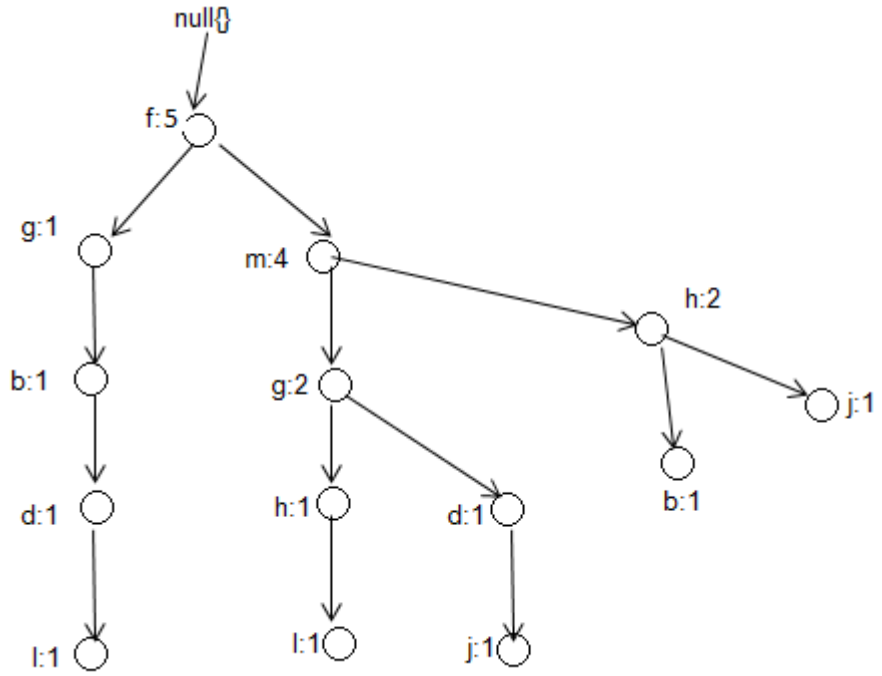
L_3:

Itemset	Sup. Count
{d,f,g}	**2**
{f,g,l}	**2**
{f,g,m}	**2**
{f,h,m}	**3**
{f,j,m}	**2**

Ans 3.2 Transactions (with sup_count>=min_sup_count) arranged by decreasing support count:

Itemset	Sup. Count
{f}	5
{m}	4
{g}	3
{h}	3
{b}	2
{d}	2
{j}	2
{l}	2

FP Tree:
(Threads and table have not been included in tree construction here because they are used for storage and retrieval purpose.)



FP-tree Mining

Item	Conditional Pattern Base	Conditional FP tree	Frequent Patterns Generated
l	<f,g,b,d:1> , <f,m,g,h:1>	<f:2, g:2>	<f,g,l:2> , <f,l:2> , <g,l:2>
j	<f,m,g,d:1> , <f,m,h:1>	<f,m:2>	<f,m,j:2> , <f,j:2> , <m,j:2>
d	<f,g,b:1> , <f,m,g:1>	<f:2,g:2>	<f,g,d:2> , <f,d:2> , <g,d:2>
b	<f,g:1> , <f,m,h:1>	<f:2>	<f,b:2>
h	<f,m,g:1> , <f,m:2>	<f:3,m:3>	<f,h:3> , <f,m,h:3> , <m,h:3>
g	<f:1> , <f,m:2>	<f:3,m:2>	<f,g:3> , <m,g:2> , <f,m,g:2>
m	<f:4>	<f:4>	<f,m:4>

Ans 3.3 Following are the closed frequent patterns: (Eliminating subsets with same support count from above table)

1. <f,g,l>
2. <f,g,d>
3. <f,m,h>
4. <f,m,g>
5. <f,j>
6. <f,b>
7. <f,g>
8. <f,m>
9. <f> (since no superset of f has support count 5)

Ans 3.4 Max frequent patterns (Eliminating subsets from closed pattern list):

1. <f,g,l>
2. <f,g,d>
3. <f,m,h>

4. <f,m,g>
5. <f,j,m>
6. <f,b>

Ans 3.5 Min_confidence = 0.6 = 60%

Frequent itemset	Subsets	Association rules	Confidence
{f,g,l}	{f} , {g} , {l} , {f,g} , {f,l} , {g,l}	f => {g,l}	2/5 = 40%
		g => {f,l}	2/3 = 66.6%
		l => {f,g}	2/2 = 100%
		{f,g} => l	2/3 = 66.6%
		{f,l} => {g}	2/2 = 100%
		{g,l} => {f}	2/2 = 100%
{f,l}	{f} , {l}	f => l	2/5 = 40%
		l => f	2/2 = 100%
{g,l}	{g} , {l}	f => g	2/5 = 40%
		g => l	2/3 = 66.6%
{f,m,j}	{f} , {m} , {j} , {f,m} , {f,j} , {m,j}	f => {m,j}	2/5 =40%
		m => {f,j}	2/4 = 50%
		j => {f,m}	2/2 = 100%
		{f,m} => {j}	2/4 = 50%
		{f,j} => m	2/2 = 100%
		{m,j} => f	2/2 = 100%
{m,j}	{m} , {j}	m => j	2/4 = 50%
		j => m	2/2 = 100%
{f,j}	{f} , {j}	f => j	2/5 = 40%
		j => f	2/2 = 100%
{f,g,d}	{f} , {g} , {d} , {f,g} , {f,d} , {g,d}	f => {g,d}	2/5 = 40%
		g => {f,d}	2/3 = 66.6%
		d => {f,g}	2/2 = 100%
		{f,g} => d	2/3 = 66.6%
		{f,d} => g	2/2 = 100%
		{g,d} => f	2/2 = 100%
{f,d }	{f} , {d}	f => d	2/5 = 40%
		d =>f	2/2 = 100%
{g,d}	{g} , {d}	g => d	2/3 = 66.6%
		d => g	2/2 = 100%
{f,b}	{f} , {b}	f => b	2/5 = 40%
		b =>f	2/2 = 100%
{f,h}	{f} , {h}	f => h	3/5 = 60%
		h => f	3/3 = 100%
{f,m,h}	{f} , {m} , {h} ,	f => {m,h}	3/5 = 60%

	{f,m} , {f,h} , {m,h}	m => {f,h}	3/4 = 75%
		h => {f,m}	3/3 = 100%
		{f,m} => h	3/4 = 75%
		{f,h} => m	3/3 = 100%
		{m,h} => f	3/3 = 100%
{m,h}	{m} , {h}	m => h	3/4 = 75%
		h => m	3/3 = 100%
{f,g}	{f} , {g}	f => g	3/5 = 60%
		g => m	3/3 = 100%
{m,g}	{m} , {g}	m => g	2/4 = 50%
		g => m	2/3 = 66.6%
{f,m,g}	{f} , {m} , {g} , {f,m} , {m,g} , {f,g}	f => {m,g}	2/5 = 40%
		m => {f,g}	2/4 = 50%
		g => {f,m}	2/3 = 66.6%
		{f,m} => g	2/4 = 50%
		{m,g} => f	2/2 = 100%
		{f,g} => m	2/3 = 66.6%
{f,m}	{f} , {m}	f => m	4/5 = 80%
		m => f	4/4 = 100%

Association rules satisfying the minimum confidence requirement are:

1. g => {f,l}
2. l => {f,g}
3. {f,g} => l
4. {f,l} => {g}
5. {g,l} => {f}
6. l=> f
7. g => l
8. j => {f,m}
9. {f,j} => m
10. {m,j} => f
11. j => m
12. j => f
13. g => {f,d}
14. d => {f,g}
15. {f,g} => d
16. {f,d} => g
17. {g,d} => f
18. d =>f
19. g => d
20. d => g
21. b =>f
22. f => h
23. h => f
24. f => {m,h}
25. m => {f,h}
26. h => {f,m}

27. {f,m} => h
28. {f,h} => m
29. {m,h} => f
30. m => h
31. h => m
32. f => g
33. g => m
34. g => m
35. g => {f,m}
36. {m,g} => f
37. {f,g} => m
38. f => m
39. m => f