**Assignment 9**

**Pranjal Naik**

**1.1**

**1**

Support is how often a rule is applicable for the provided data set.
Confidence is how frequently an item in Y appear in transactions that contain X.

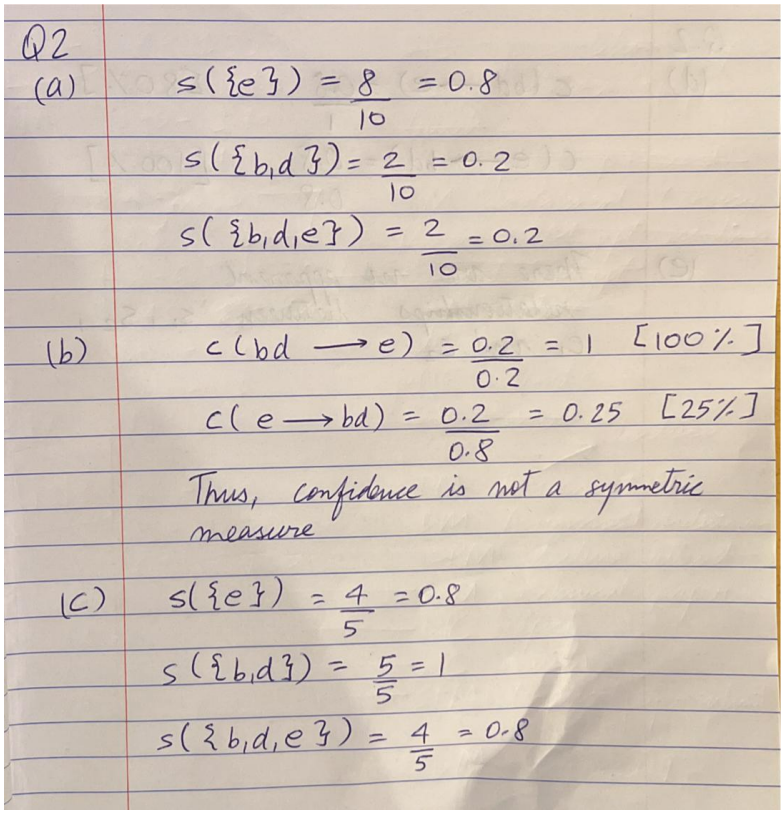
A) A rule that has high support and high confidence
Example of an Association Rule: Cereal → Milk
As there is nothing new to learn from this rule, considering these items are far too common. Thus it is **not interesting**.

B) A rule that has reasonably high support but low confidence.
Example of an Association Rule: Milk → Tuna
Not all transactions that contain milk also contain tuna. Such low-confidence rule tends to be **not interesting**.

C) A rule that has low support and low confidence.
Example of an Association Rule: Batteries → Toothpaste
These items are infrequent, also they do not have any common use cases. Such low-confidence rule are **not interesting**.

D) A rule that has low support and high confidence.
Example of an Association Rule: Vodka → Caviar
As these complement the taste, these items are usually ordered together and occur relatively frequent together. This rules seems **interesting**.

**2**



Q.2

(d) $c(bd \rightarrow e) = \frac{0.8}{1}$ [80%.]

$c(e \rightarrow bd) = \frac{0.8}{0.8}$ [100%.]

(e) There are no apparent relationships between $S_1$, $S_2$, $C_1$ and $C_2$

**9(a)**

Q.9

(a) Minimum Support (minsup) = 0.3 = 30%.

Support Calculations for all levels -
[Support = s]

Level 1 -
Support (null) = 1

Level 2 -
s(A) = 0.5
s(B) = 0.7
s(C) = 0.5
s(D) = 0.9
s(E) = 0.6

Level 3 -

| | |
|---|---|
| s(AB) = 0.3 | s(BD) = 0.6 |
| s(AC) = 0.2 | s(BE) = 0.4 |
| s(AD) = 0.4 | s(CD) = 0.4 |
| s(AE) = 0.4 | s(CE) = 0.2 |
| s(BC) = 0.3 | s(DE) = 0.6 |

{AC} and {CE} have support below threshold, these nodes will be marked as 'I'

Level 4 -

$s(ABD) = 0.2$

$s(ABE) = 0.2$

$s(ADE) = 0.4$

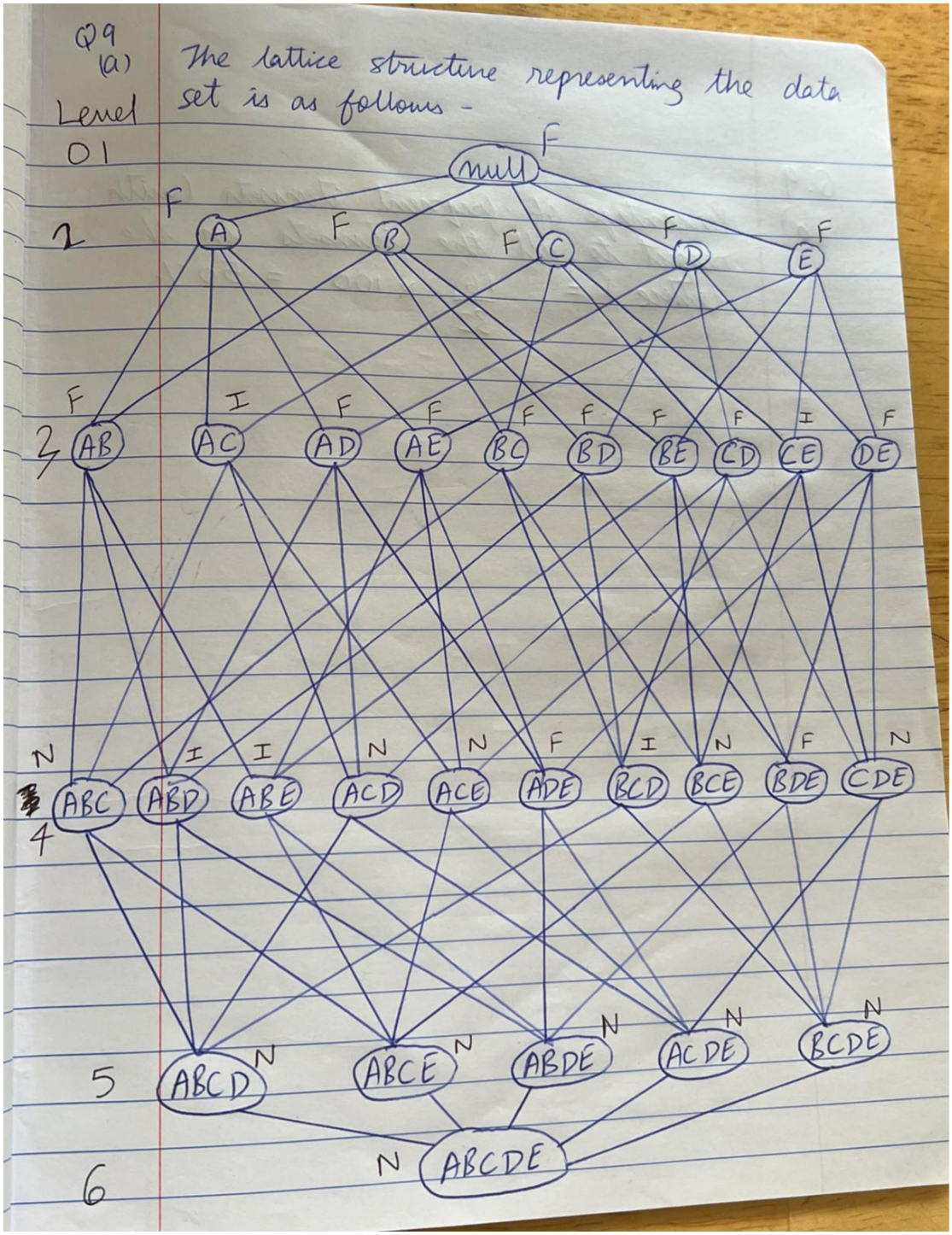
$s(BCD) = 0.2$

$s(BDE) = 0.3$

{ABD}, {ABE} and {BCD} have support below threshold, these nodes will be marked as 'I'.
All nodes with {AC} and {CE} as contributor, will be marked as 'N'

Level 5 and 6 -
Every contributor to every node does not have all nodes marked as 'F' and 'I'. Thus, no need to calculate support.

Q9
(a) The lattice structure representing the data set is as follows -




**9(b)**

Q.9
(b) Percentage of frequent itemsets (with respect to all the itemsets in the lattice) = $\frac{16}{32} \times 100 = 50\%$.

**15**

(a) Which data set(s) will produce the most number of frequent itemsets?
Data set **(e)** will produce the most number of frequent itemsets because it has to generate the longest frequent itemset along with its subsets.

(b) Which data set(s) will produce the fewest number of frequent itemsets?
Data set **(d)** will produce the least number of frequent itemsets because it does not produce any frequent itemsets at 10% support threshold.

(c) Which data set(s) will produce the longest frequent itemset?
Data set **(e)**

(d) Which data set(s) will produce frequent itemsets with highest maximum support?
Data set **(b)**

(e) Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%).
Data set **(e)**

**1.2**

**1(a)**

1.2

1

(a)

| tid | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| t1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| t2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| t3 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| t4 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| t5 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| t6 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| t7 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| t8 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

Minimum Support (minsup) = $\frac{3}{8}$

Level 1 -

s(null) = 8

Level 2 -

s(A) = 5

s(B) = 4

s(C) = 5

s(D) = 6

s(E) = 1

s(F) = 4

s(G) = 5

{E} is below threshold

Level 3 -

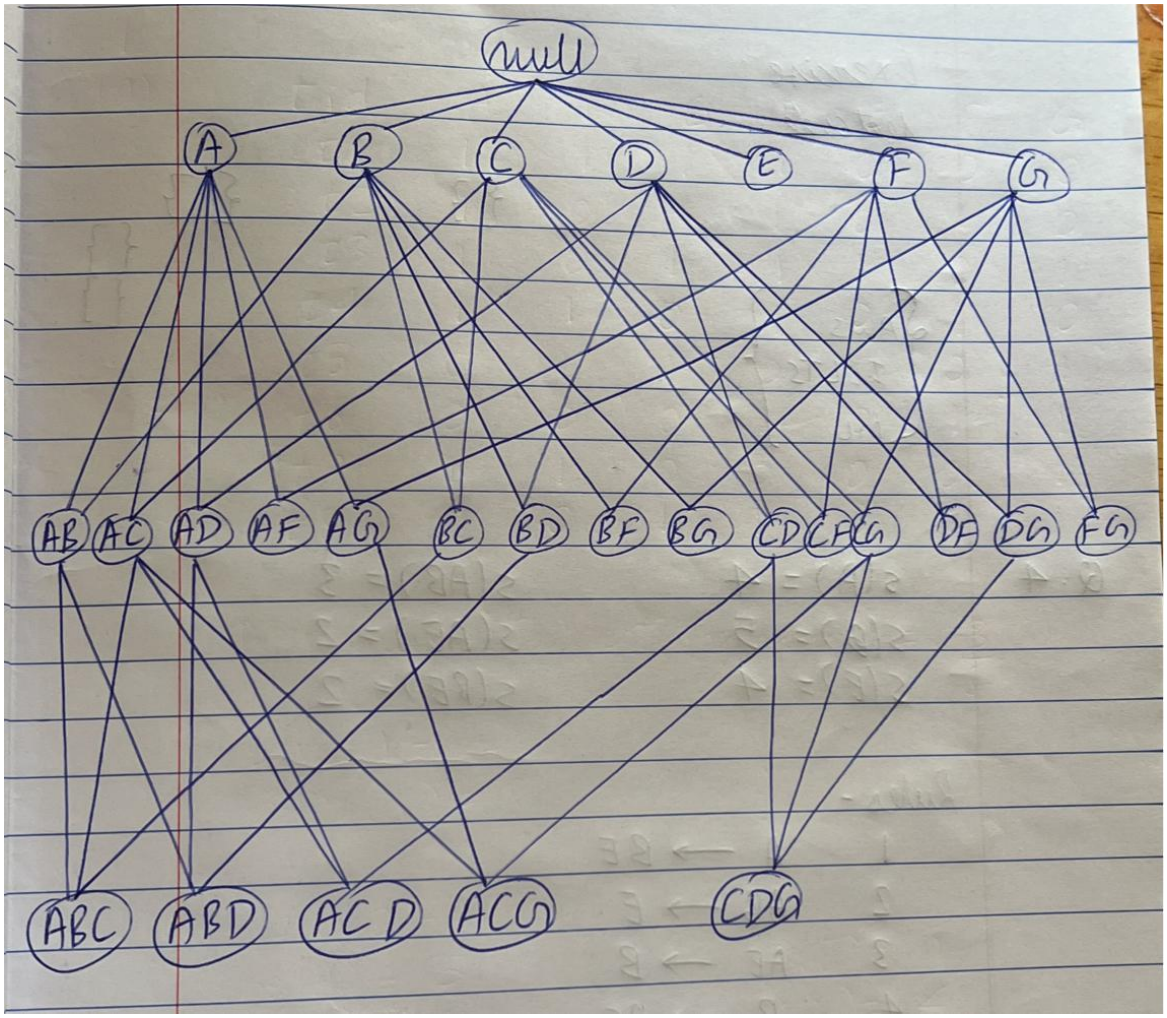
s(AB) = 3
s(AC) = 3
s(AD) = 4
s(AF) = 2
s(AG) = 2
s(BC) = 2
s(BD) = 2
s(BF) = 1
s(BG) = 2
s(CD) = 4
s(CF) = 2
s(CG) = 3
s(DF) = 4
s(DG) = 3
s(FG) = 2

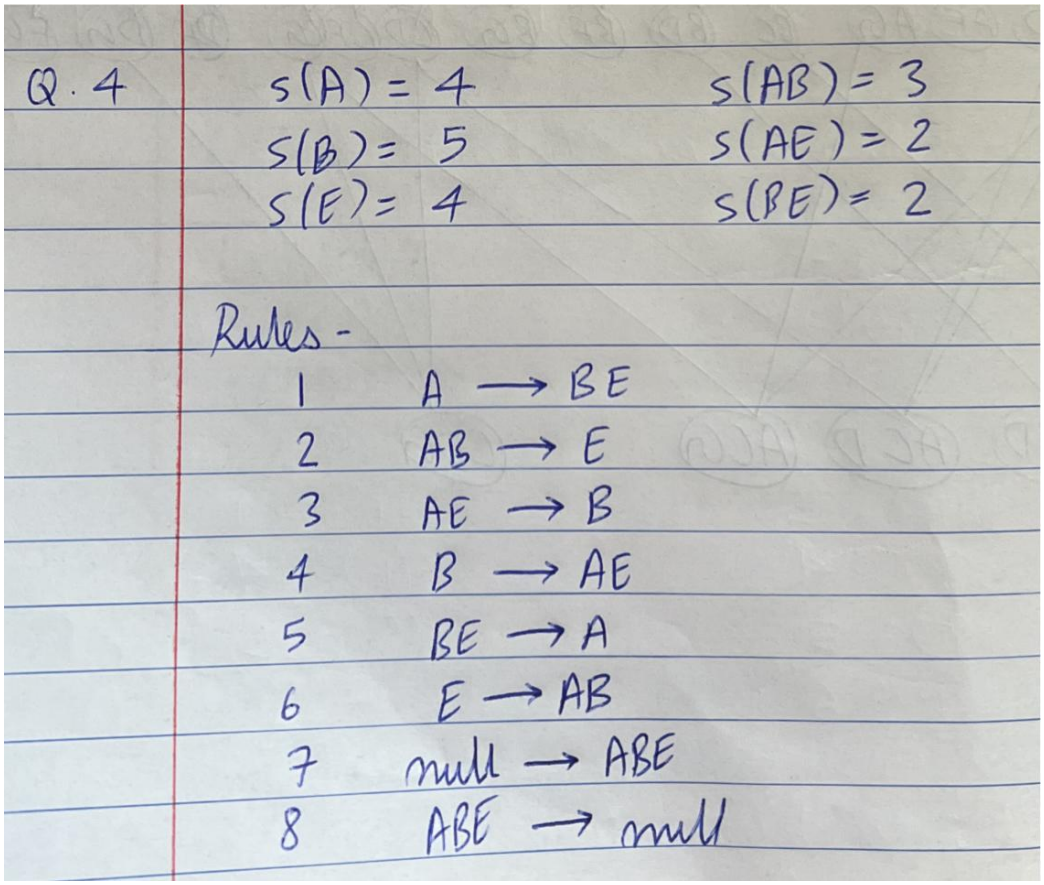
{AF}, {AG}, {BC}, {BD}, {BF}, {BG}, {CF} and {FG} are below threshold

Level 4 -

s(ABC) = 1
s(ABD) = 2
s(ACD) = 3
s(ACG) = 1
s(CDG) = 2

{ABC), {ABD}, {ACG} and {CDG} are below threshold


null
A
B
C
D
E
F
G
AB
AC
AD
AF
AG
BC
BD
BF
BG
CD
CF
CG
DF
DG
FG
ABC
ABD
ACD
ACG
CDG


Q. 4
s(A) = 4
s(B) = 5
s(E) = 4
s(AB) = 3
s(AE) = 2
s(BE) = 2
Rules -
1 A → BE
2 AB → E
3 AE → B
4 B → AE
5 BE → A
6 E → AB
7 null → ABE
8 ABE → null