CSC 422 HW 4

Group 32: Sagnik Nayak, Vinhson Phan, Rhea John

Q1

$y_i(w^T X_i + w_0) - 1 = 0$

<u>Class 1</u>   <u>Class 2</u>

(4,5)   (6,7)

$-1(4w_1 + 5w_2 + w_0) - 1 = 0$

$-4w_1 - 5w_2 - w_0 - 1 = 0 \quad -(i)$

||'ly $\quad 6w_1 + 7w_2 + w_0 - 1 = 0 \quad -(ii)$

from (i) & (ii),

$$\boxed{2w_1 + 2w_2 - 2 = 0}$$

$d(+) + d(-) = \dfrac{2}{\|W\|_{2_2}}$

$X_2 = \dfrac{-W_1}{W_2} X_1 - \dfrac{W_0}{W_2}$

$m = \dfrac{7-5}{6-4} = \underline{\underline{1}}$

$m\,hyper = \underline{\underline{-1}}$

$\Rightarrow \dfrac{-W_1}{W_2} = \underline{1}$

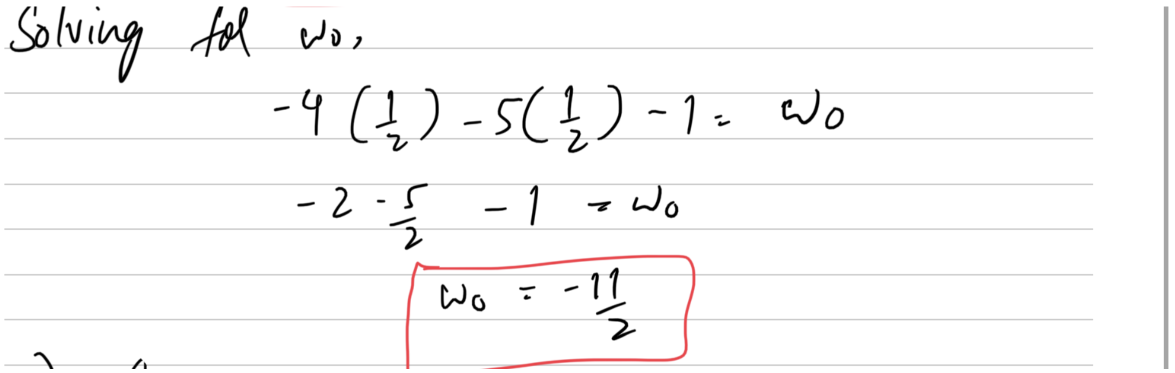
Thus,
Let $w_1 = w_2$

$2(2W_2) + 2W_2 - 2 = 0$

$4W_2 = 2$

$$\boxed{W_2 = \frac{1}{2}}$$

a) i.



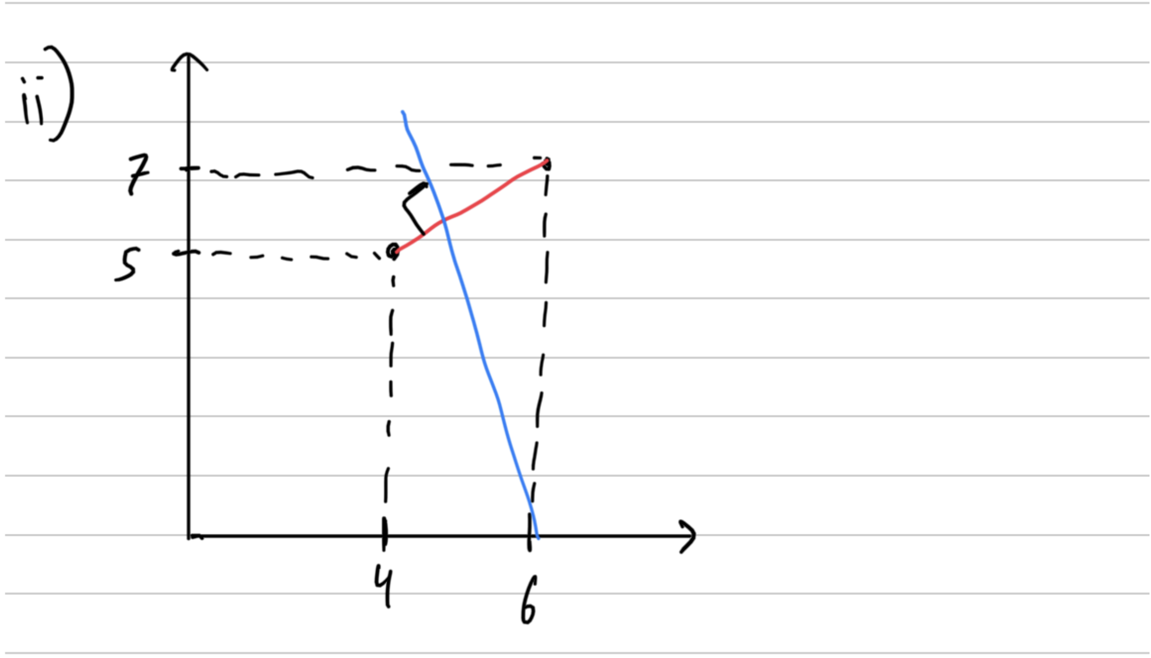
Solving fol w_0,

$$-4\left(\frac{1}{2}\right) - 5\left(\frac{1}{2}\right) - 1 = w_0$$

$$-2 - \frac{5}{2} - 1 = w_0$$

$$\boxed{w_0 = -\frac{11}{2}}$$

ii.

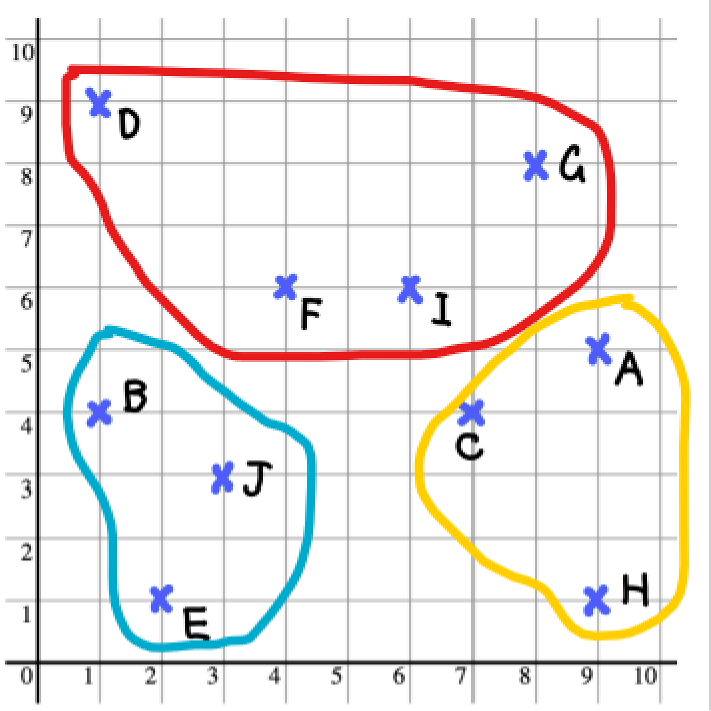


Q2

a)

Round	Points	Cluster	Centroid
1	B,D,F,G,I	1	I(6,6)
	A,E,J,C	2	C(7,4)
	H	3	H(9,1)

2	D,F,G,I,	1	M (4, 6.6)
	B,C,E,J	2	N (5.25, 3.25)
	A,H	3	H (9, 1)
3	D,F,G,I	1	Q (4.75, 7.25)
	B,E,J	2	R (3.25, 3)
	A,C,H	3	S (9, 3)
4	D,F,G,I	1	Q (4.75, 7.25)
	B, E, J	2	U (2, 2.66)
	A,C,H	3	V (8.33, 3.33)



Round One Seed Centroids:
Cluster 1: I (6, 6)
Cluster 2: C (7, 4)
Cluster 3: H (9, 1)

Round Two Centroids:
Cluster 1:
Avg of D, B, F, I, G
Point = M (4, 6.6)

Cluster 2:
Avg of E, J, A, C
Point = N (5.25, 3.25)

Cluster 3: H (9, 1)

Round Three Centroids:
Cluster 1:
Avg of D, F, G, I
Point = Q (4.75, 7.25)

Cluster 2:
Avg of B,C,E,J
Point = R (3.25, 3)

Cluster 3:
Avg of A, H:
Point = S (9, 3)

Round 4 Centroids:
Cluster 1:
Avg of D,F,G,I,
Point = Q (4.75, 7.25)

Cluster 2:
Avg of B,E,J
Point = (2, 2.66)

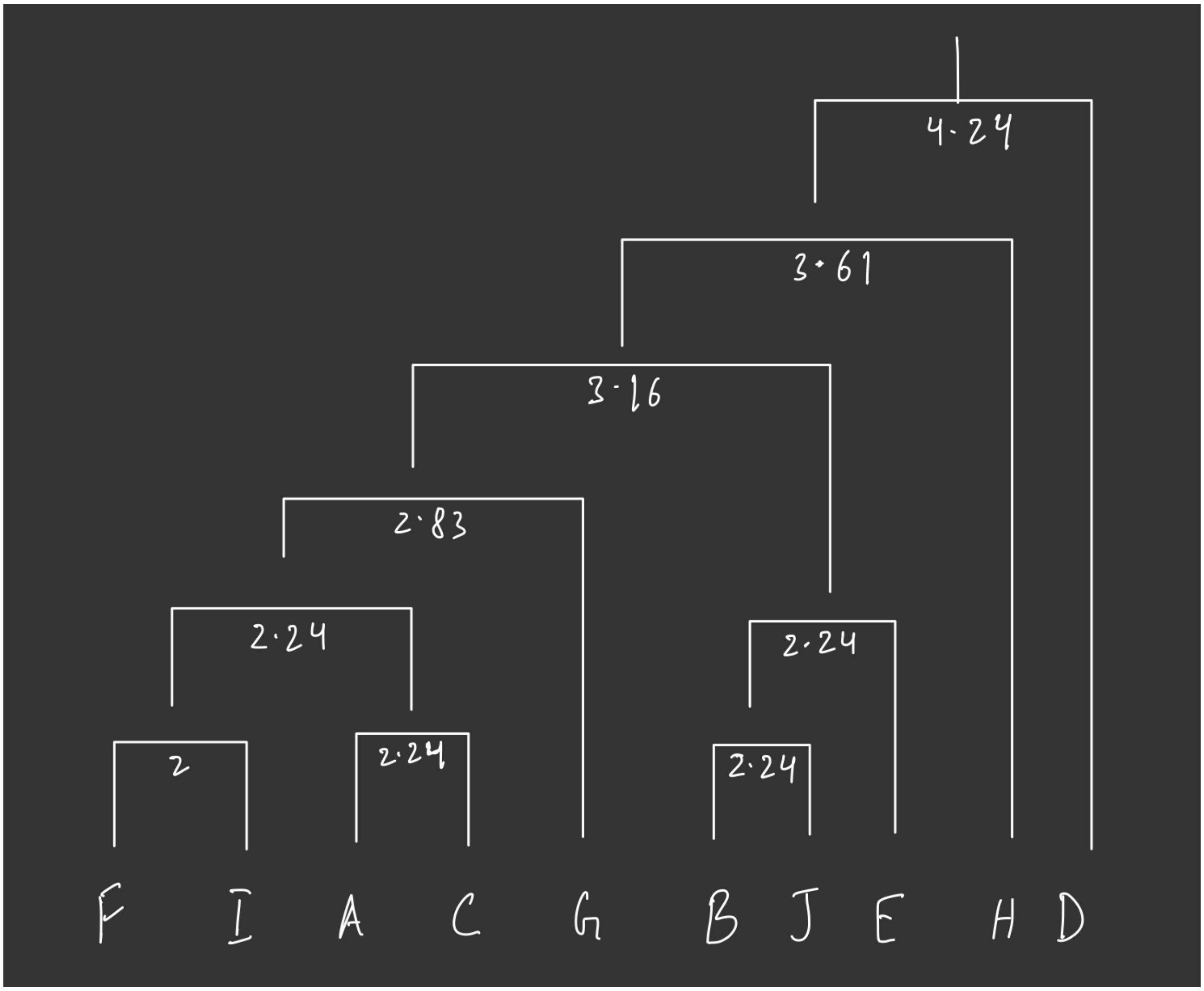
Cluster 3:
Avg of A,C,H
Point = (8.33, 3.33)

b)

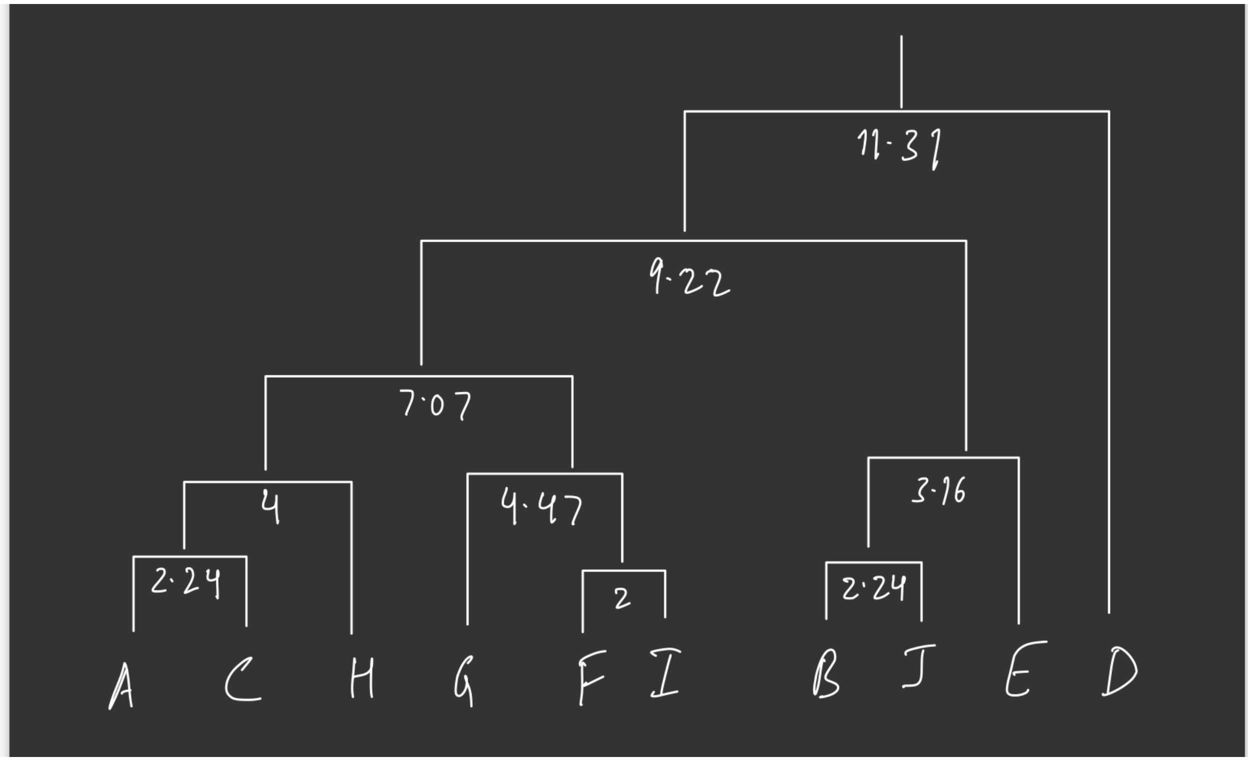
4 rounds

Q3)

a)



b)



c)

Complete Linkage is the better approach to clustering for this dataset due to better splits based on the two clusters. Complete Linkage also had lower SSE compared to Single linkage.

d)

K Means clusters:
Cluster 1: D,F,G,I
Centroid: Q (4.75, 7.25)
SSE = 33.5

Cluster 2: B, E, J
Centroid: U (2, 2.66)
SSE = 6.66

Cluster 3: A,C,H
Centroid: V (8.33, 3.33)
SSE = 11.333

K Means SSE = 33.5 + 6.666 + 11.333 = **51.499**

Single Link clusters:
Cluster 1: F, I, C, A, G, B, J, E
Centroid = (5, 4.625)
SSE = 25.359

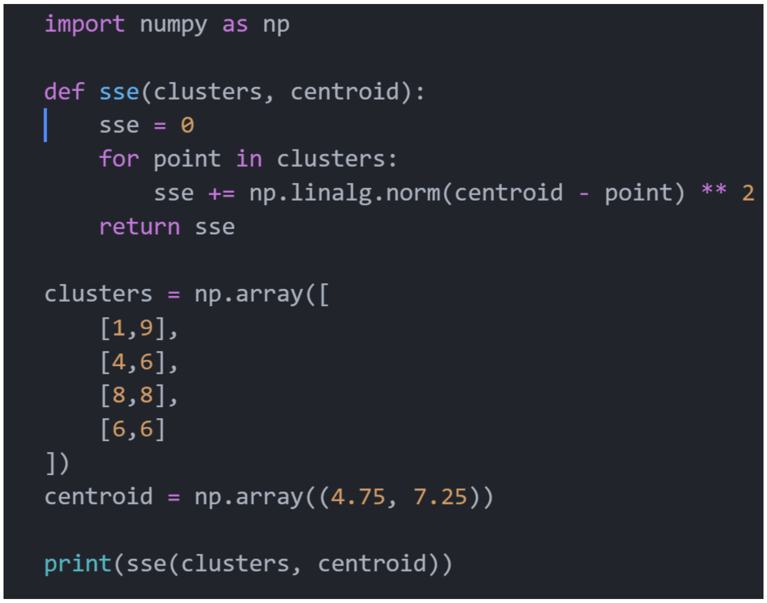
Cluster 2: H
Centroid = H
SSE = 0

Cluster 3: D
Centroid = D
SSE = 0

Single Link SSE = 25.359 + 0 + 0 = **91.875**

Since K Means has a lower SSE compared to Single Link clustering (51.5 < 91.9), K means is the better clustering algorithm for this use case and dataset.

Script:



```
import numpy as np

def sse(clusters, centroid):
    sse = 0
    for point in clusters:
        sse += np.linalg.norm(centroid - point) ** 2
    return sse

clusters = np.array([
    [1,9],
    [4,6],
    [8,8],
    [6,6]
])
centroid = np.array((4.75, 7.25))

print(sse(clusters, centroid))
```

Q4

a)

N items = 6

Given, d items, total itemsets is given by 2^d.
Hence, 2^6 = **64**

b)

Given, d items, the number of association rules that can be extracted is given by:
3^d + 2^(d+1) + 1
I.e. 3^(6) + 2^(6+1) + 1 = **602**

c)

Total transaction = 10
n(support: {Eggs, Cola}) = 2
Hence, support({Eggs, Cola}) = 2/10 = **0.2**

d)

support(Bread -> Butter) = n(Bread U Butter) / n(Transactions) = 3 / 10 = **0.3**
confidence(Bread -> Butter) = n(Bread U Butter) / n(Bread) = 3 / 6 = **0.5**

e)

In the given dataset with 10 transactions, a 0.3 support implies the items must co occur at least three times, which is {Bread, Milk, Cola}.
conf({Bread, Milk} -> {Cola}) = 3 / 4 = 0.75 > 0.6 (confidence threshold)
conf({Cola, Milk} -> {Bread}) = 3 / 4 = 0.75 > 0.6 (confidence threshold)

Hence, the valid association rules are :-

1. **{Bread, Milk} -> {Cola}**
2. **{Milk, Cola} -> {Bread}**

f)

$$\{A,B\} \subseteq \{A,B,D\} \subseteq \{A,B,C,D\}$$

$$S(\{A,B\}) \geq S(\{A,B,D\}) \geq S(\{A,B,D,C\})$$

$$0.46 \geq S(\{A,B,D\}) \geq 0.23$$

Support of $\{A\} \rightarrow \{B,D\}$ must be ≤ 0.46 & ≥ 0.23

Q5

a)

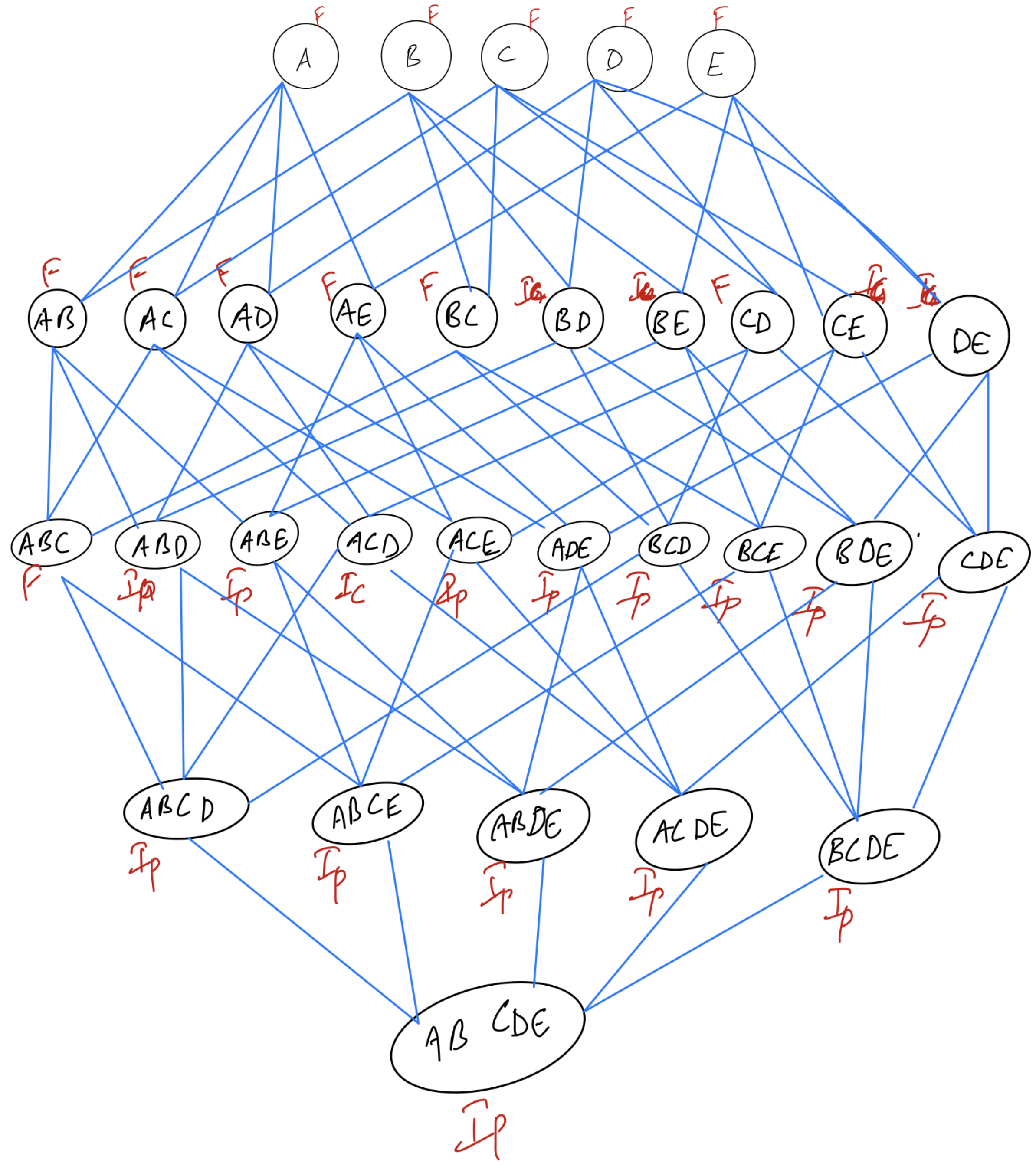
1	Itemset	Count
	A	7
	B	5
	C	5
	D	5
	E	4

2	Itemset	Count	Skip
	A,B	5	
	A,C	4	
	A,D	4	
	A,E	4	
	B,C	3	
	B.D	2	x

	B,E	2	x
	C,D	3	
	C,E	2	x
	D,E	2	x

3	Itemset	Count	Skip
	A,B,C	3	
	A,C,D	2	x

Hence. Frequent Itemset with minimum support = 3 is {A, B, C} b)


A
F
B
F
C
F
D
F
E
F
AB
F
AC
F
AD
F
AE
F
BC
F
BD
Ic
BE
Ic
CD
F
CE
Ic
DE
Ic
ABC
F
ABD
Ip
ABE
Ip
ACD
Ic
ACE
Ip
ADE
Ip
BCD
Ip
BCE
Ip
BDE
Ip
CDE
Ip
ABCD
Ip
ABCE
Ip
ABDE
Ip
ACDE
Ip
BCDE
Ip
AB CDE
Ip