ALDA Homework 4 – 2018 :

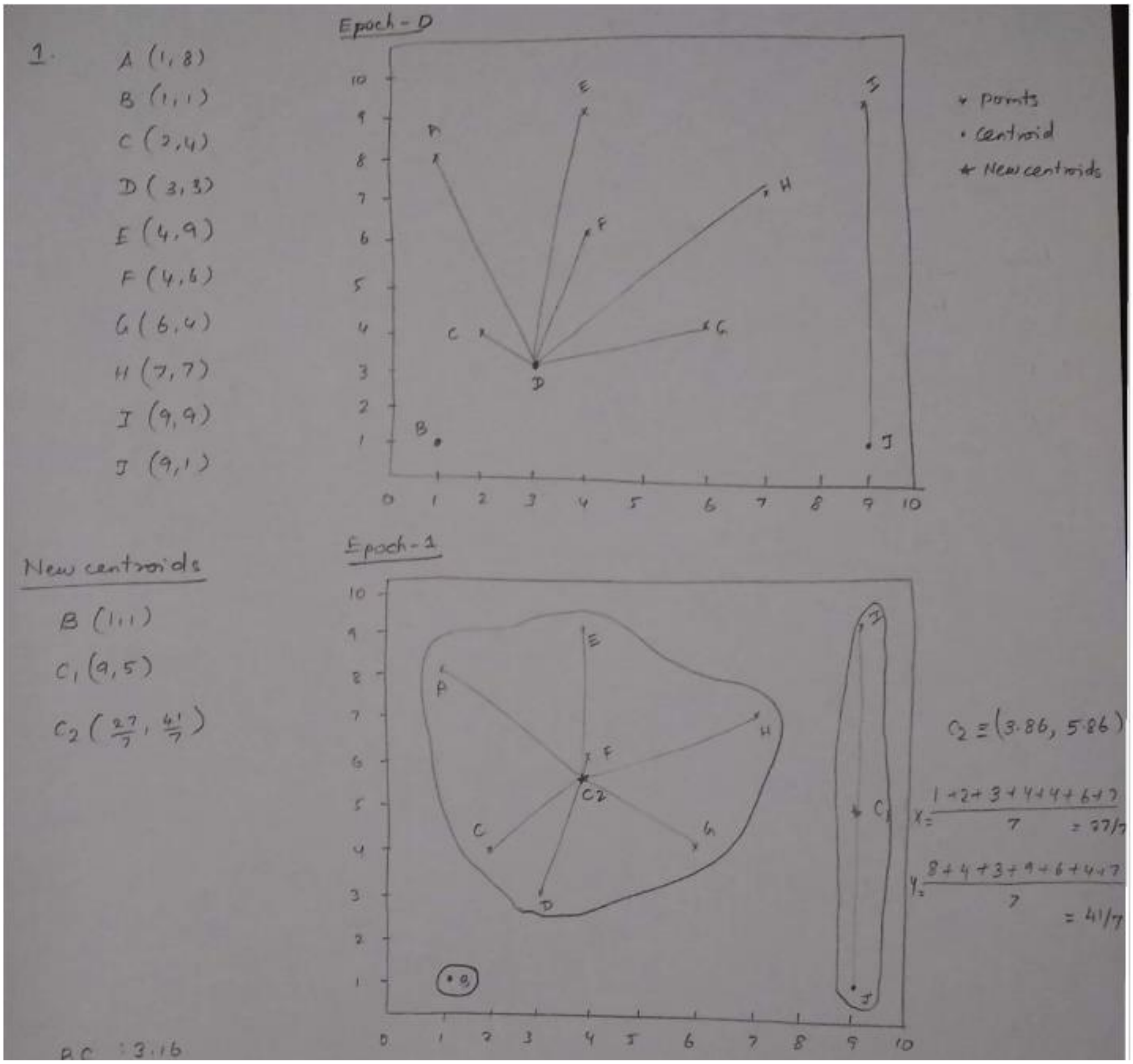
Submitted by :

Srinivasan Balan Unity id : sbalan

Harika Malapaka – Unity Id : hsmalapa

1. (13 points) [K-means Clustering][Xi Yang] Use K-means clustering algorithm with Euclidean Distance to cluster the 10 data points in Figure 1 into 3 clusters. Suppose that the initial seeds are at points: B, D and J. Answer the following questions:

(a) (4 points) Run 1 round of k-means algorithm. What are the coordinates of the new centroids? What are the new clusters? Show your work in Figure 1.



1. A (1,8)
B (1,1)
C (2,4)
D (3,3)
E (4,9)
F (4,6)
G (6,4)
H (7,7)
I (9,9)
J (9,1)

New centroids

B (1,1)
C_1 (9,5)
C_2 ($\frac{27}{7}$, $\frac{41}{7}$)

BC : 3.16

C_2 = (3.86, 5.86)

$$x = \frac{1+2+3+4+4+6+7}{7} = 27/7$$

$$y = \frac{8+4+3+9+6+4+7}{7} = 41/7$$

In the above figure, epoch 0 is the initial coordinates given.

Epoch 1 is the clusters and centroids derived after 1 round of K-means.

After epoch 1, the new coordinates are :

(1,1)

(9,5)

(3.8,5.8)

The clusters are {B}, {A,C,D,E,F,G,H} and {I,J}

(b) (9 points) How many rounds are needed for the K-means clustering algorithm to converge? Draw the result clusters and new centroid at the end of each round (including the first round) in the Figure 2. Indicate the coordinates along side corresponding centroids. Add new graphics if needed; Stop when the algorithm converges and clearly label on the graph where the algorithm converges.

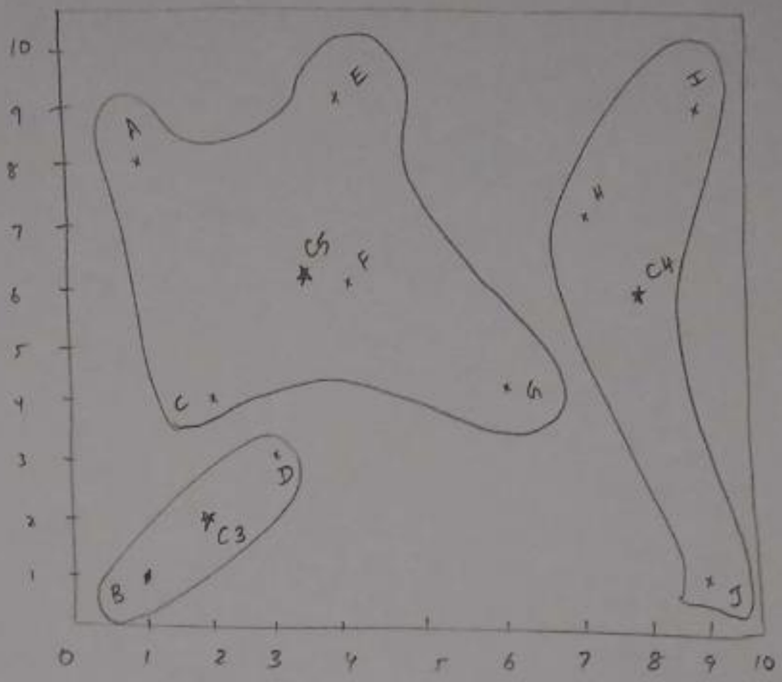
New centroids

c_3 $(2, 2)$

c_4 $(25/3, 17/3)$

c_5 $(17/5, 31/5)$

Epoch - 2



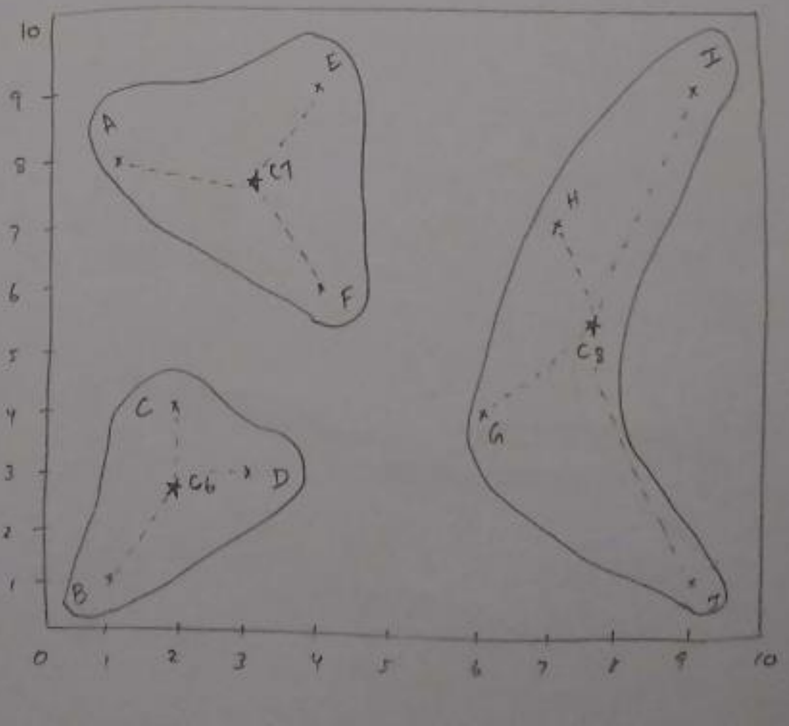
Epoch - 3

New centroids

c_6 $(2, 8/3)$

c_7 $(3, 23/3)$

c_8 $\left(31/4, \frac{21}{4}\right)$

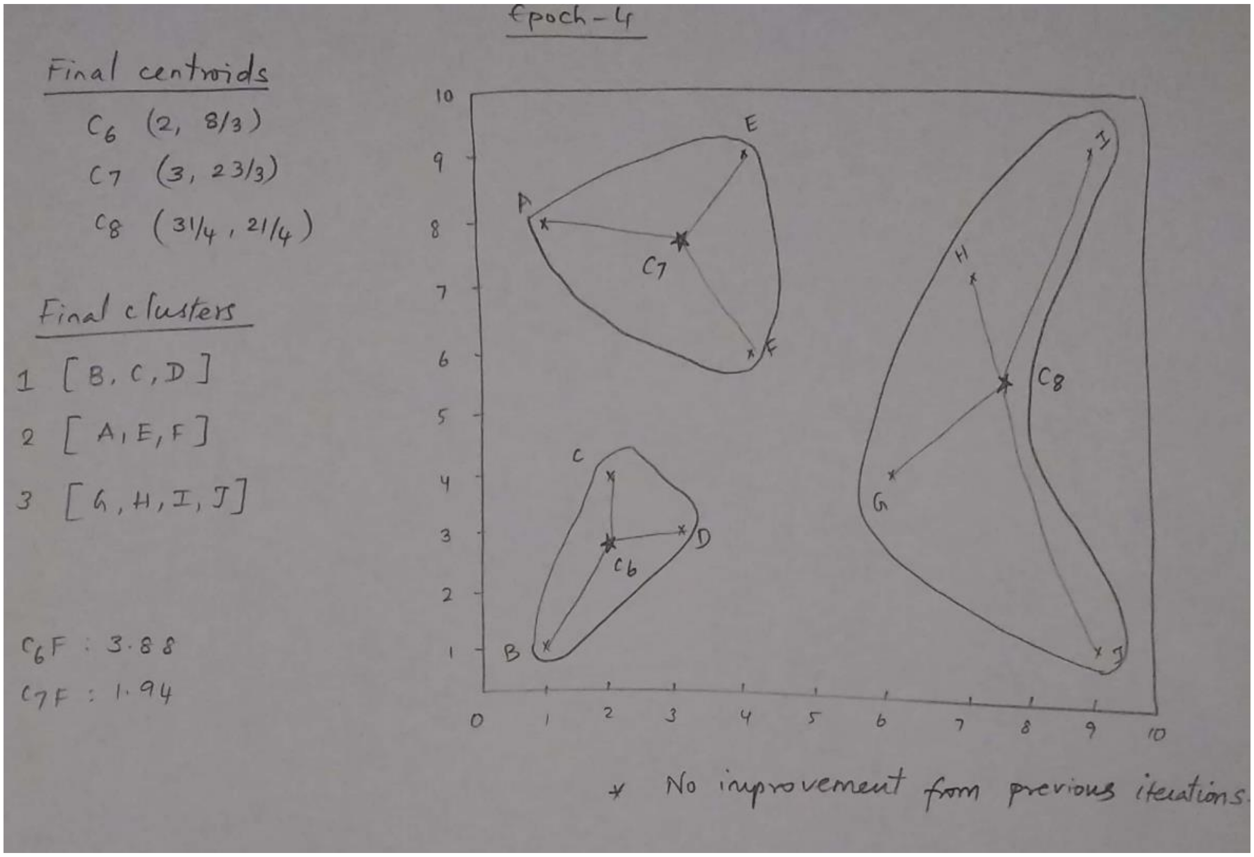


✓ C_3C : 2

C_5C : 2.6

✓ C_4G : 2.86

C_5G : 3.4



Epoch-4

Final centroids

C_6 (2, 8/3)

C_7 (3, 23/3)

C_8 (31/4, 21/4)

Final clusters

1 [B, C, D]

2 [A, E, F]

3 [G, H, I, J]

C_6F : 3.88

C_7F : 1.94

* No improvement from previous iterations.

Therefore, 4 rounds were required to make sure the algorithm converges.

The final centroids are :

(2,2.67),(3,7.67),(7.75,5.25)

The final clusters are :

[B,C,D], [A,E,F],[G,H,I,J]

2. (20 points) [Hierarchical Clustering] [Ruth Okoilu] We will use the same dataset A-J as in Question 1 for Hierarchical Clustering. The Euclidean Distance matrix between each pair of the datapoints are listed in the Table 3 below:

(a) (10 points) Perform single and complete link hierarchical clustering. Show your results by drawing corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged.

2.

a)

	A	B	C	D	E	F	G	H	I	J
A	—	7	4.12	5.39	3.16	3.61	6.4	6.08	8.06	10.63
B		—	3.16	2.83	8.54	5.83	5.83	8.49	11.31	8
C			—	1.41	5.39	2.83	4	5.83	8.6	7.62
D				—	6.08	3.16	3.16	5.66	8.49	6.32
E					—	3	5.39	3.61	5	9.43
F						—	2.83	3.16	5.83	7.07
G							—	3.16	5.83	4.24
H								—	2.83	6.32
I									—	8
J										—

Bottom-up

distance : Smaller the better

similarity : larger the better

Single link

	A	B	E	F	G	H	I	J
C-D	4.12	2.83	5.39	2.83	3.16	5.66	8.49	6.32
A	—	7	3.16	3.61	6.4	6.08	8.06	10.63
B		—	8.54	5.83	5.83	8.49	11.31	8
E			—	3	5.39	3.61	5	9.43
F				—	2.83	3.16	5.83	7.07
G					—	3.16	5.83	4.24
H						—	2.83	6.32
I							—	8
J								—

	A	E	F	G	H	I	J
C-D-B	4.12	5.39	2.83	3.16	5.66	8.49	6.32
A	—	3.16	3.61	6.4	6.08	8.06	10.63
E		—	3	5.39	3.61	5	9.43
F			—	2.83	3.16	5.83	7.07
G				—	3.16	5.83	4.24
H					—	2.83	6.32
I						—	8
J							—

	A	E	G	H	I	J
C - D - B - F	3.61	3.0	2.83	3.16	5.83	6.32
A	—	3.16	6.4	6.08	8.06	10.63
E		—	5.39	3.61	5	9.43
G			—	3.16	5.83	4.24
H				—	2.83	6.32
I					—	8
J						—

	A	E	H	I	J
C - D - B - F - G	3.61	3.0	3.16	5.83	4.24
A	—	3.16	6.08	8.06	10.63
E		—	3.61	5	9.43
H			—	2.83	6.32
I				—	8
J					—

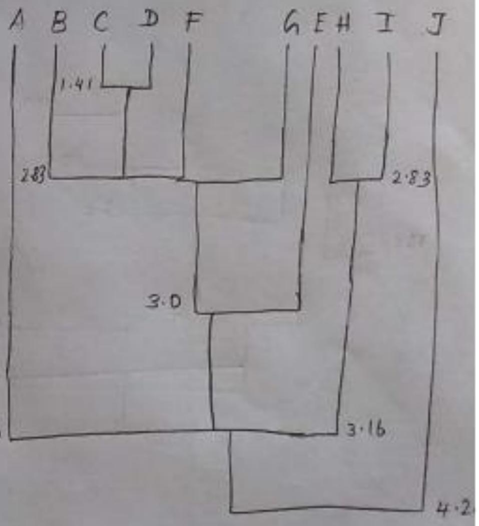
	A	E	H - I	J
C - D - B - F - G	3.61	3.0	3.16	4.24
A	—	3.16	6.08	10.63
E		—	3.61	9.43
J			6.32	—

	A	H-I	J
C-D- B-F-G- E	3.16	3.16	4.24
A	—	6.08	10.63
J		6.32	—

	H-I	J
C-D-B-F-G-E-A	3.16	4.24
J	4.24	—

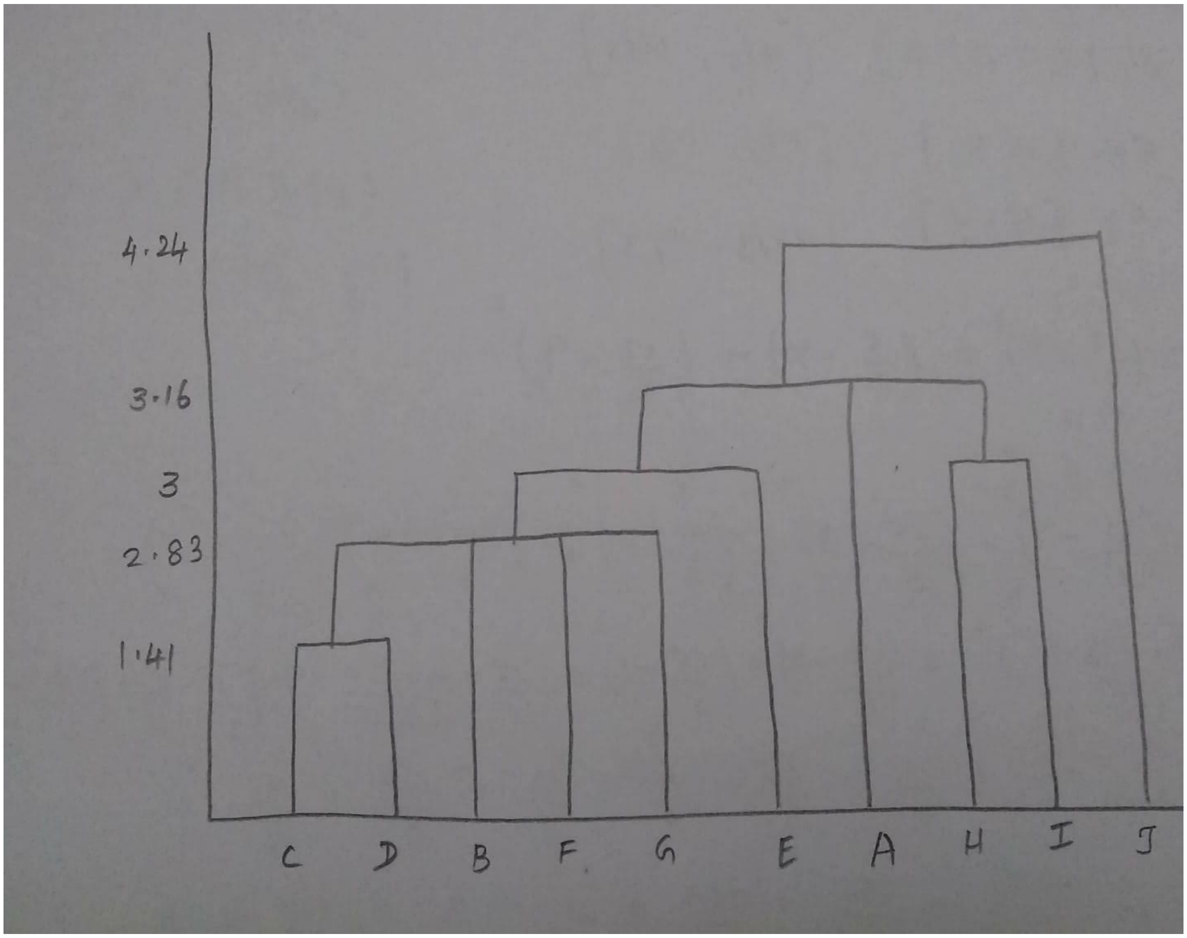
	J
C - D - B - F - G - E - A - H - I	$\boxed{4.24}$

Dendrogram

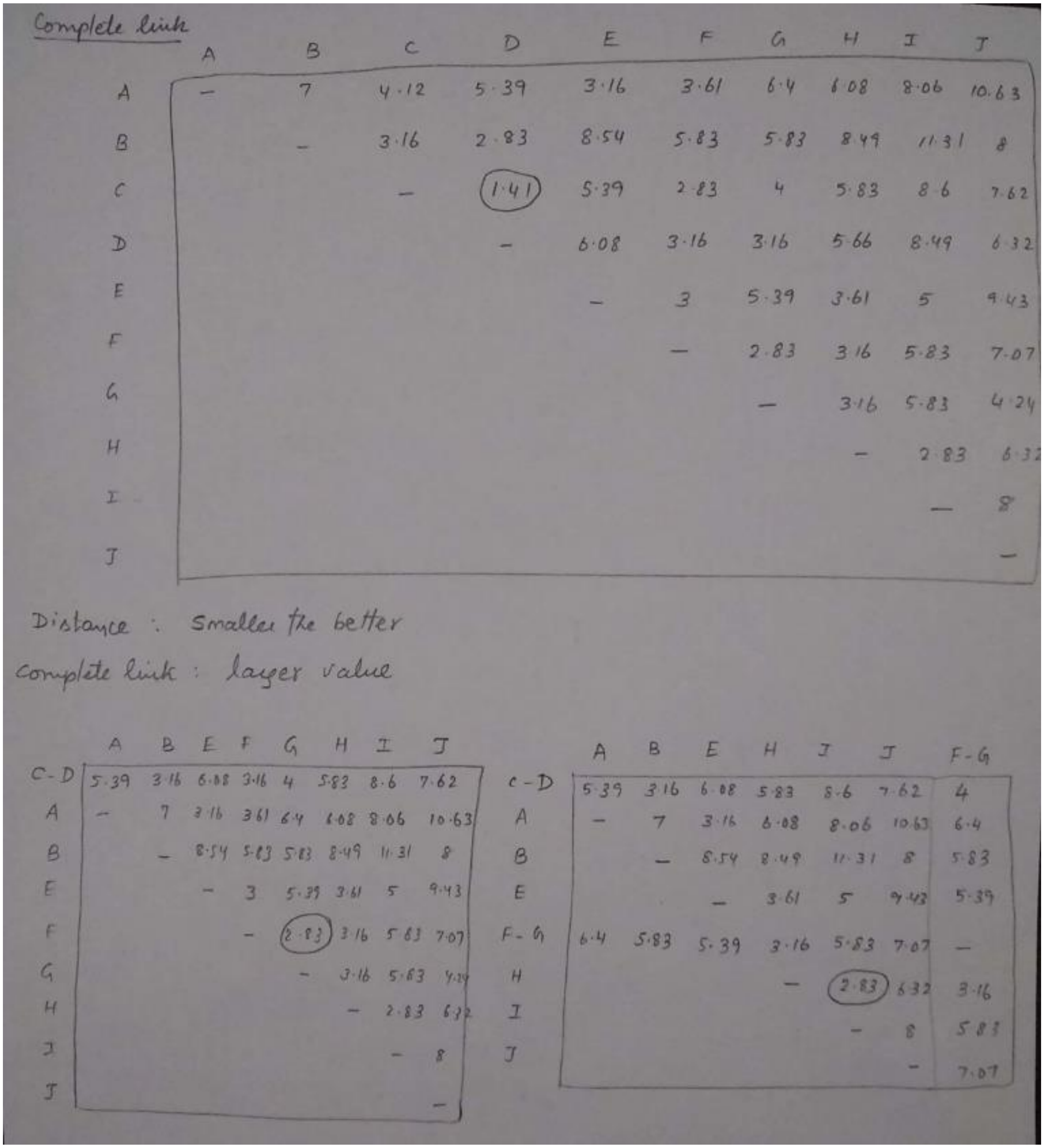


A	1
B	2
C	3
D	4
E	5
F	6
G	7
H	8
I	9
J	10

The Dendogram according to order of height is :



COMPLETE LINK :



Complete link

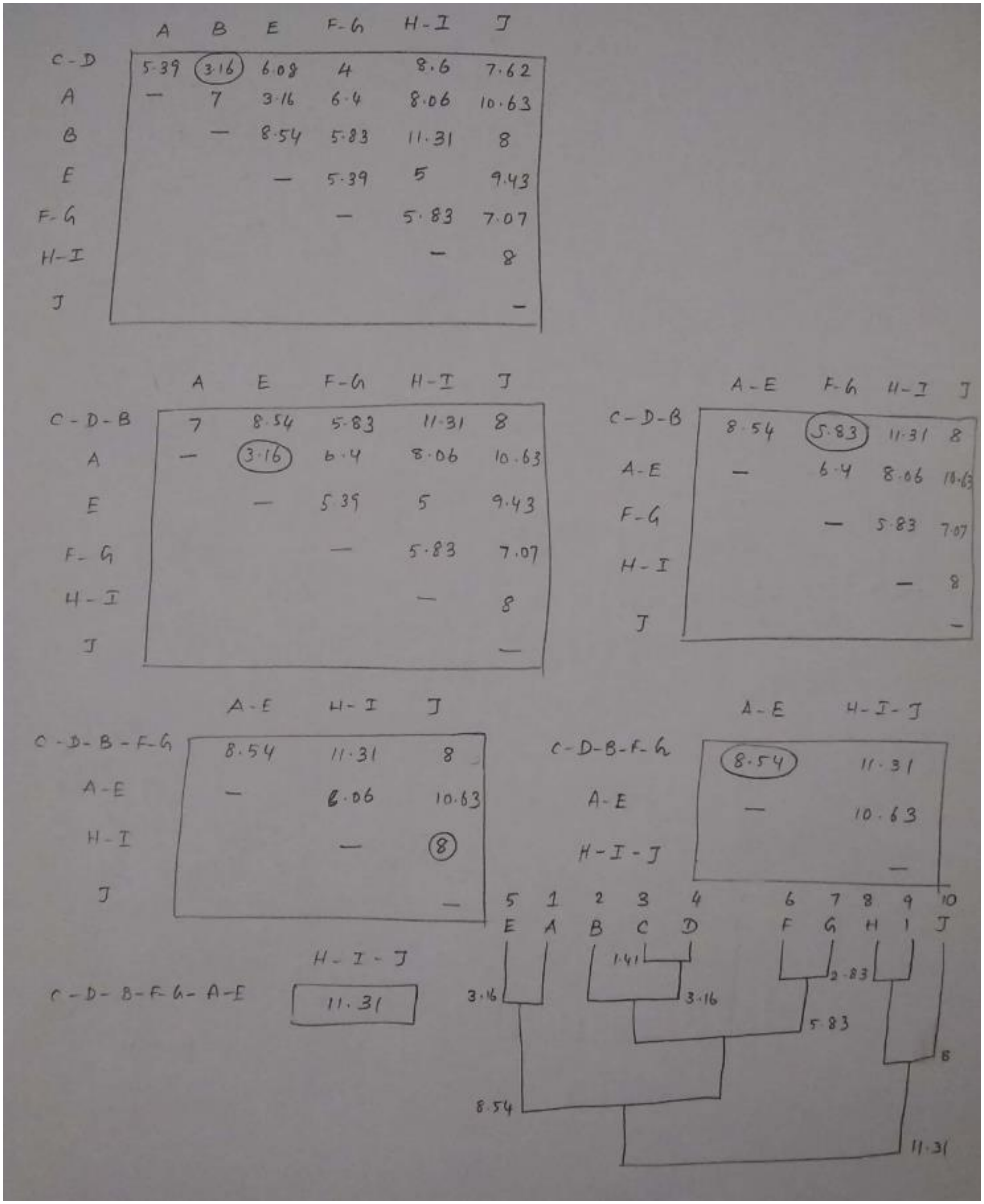
	A	B	C	D	E	F	G	H	I	J
A	–	7	4.12	5.39	3.16	3.61	6.4	6.08	8.06	10.63
B		–	3.16	2.83	8.54	5.83	5.83	8.49	11.31	8
C			–	1.41	5.39	2.83	4	5.83	8.6	7.62
D				–	6.08	3.16	3.16	5.66	8.49	6.32
E					–	3	5.39	3.61	5	9.43
F						–	2.83	3.16	5.83	7.07
G							–	3.16	5.83	4.24
H								–	2.83	6.32
I									–	8
J										–

Distance : smaller the better

complete link : larger value

	A	B	E	F	G	H	I	J
C-D	5.39	3.16	6.08	3.16	4	5.83	8.6	7.62
A	–	7	3.16	3.61	6.4	6.08	8.06	10.63
B		–	8.54	5.83	5.83	8.49	11.31	8
E			–	3	5.39	3.61	5	9.43
F				–	2.83	3.16	5.83	7.07
G					–	3.16	5.83	4.24
H						–	2.83	6.32
I							–	8
J								–

	A	B	E	H	I	J	F-G
C-D	5.39	3.16	6.08	5.83	8.6	7.62	4
A	–	7	3.16	6.08	8.06	10.63	6.4
B		–	8.54	8.49	11.31	8	5.83
E			–	3.61	5	9.43	5.39
F-G	6.4	5.83	5.39	3.16	5.83	7.07	–
H				–	2.83	6.32	3.16
I					–	8	5.83
J						–	7.07



	A	B	E	F-G	H-I	J
C-D	5.39	3.16	6.08	4	8.6	7.62
A	—	7	3.16	6.4	8.06	10.63
B		—	8.54	5.83	11.31	8
E			—	5.39	5	9.43
F-G				—	5.83	7.07
H-I					—	8
J						—

	A	E	F-G	H-I	J
C-D-B	7	8.54	5.83	11.31	8
A	—	3.16	6.4	8.06	10.63
E		—	5.39	5	9.43
F-G			—	5.83	7.07
H-I				—	8
J					—

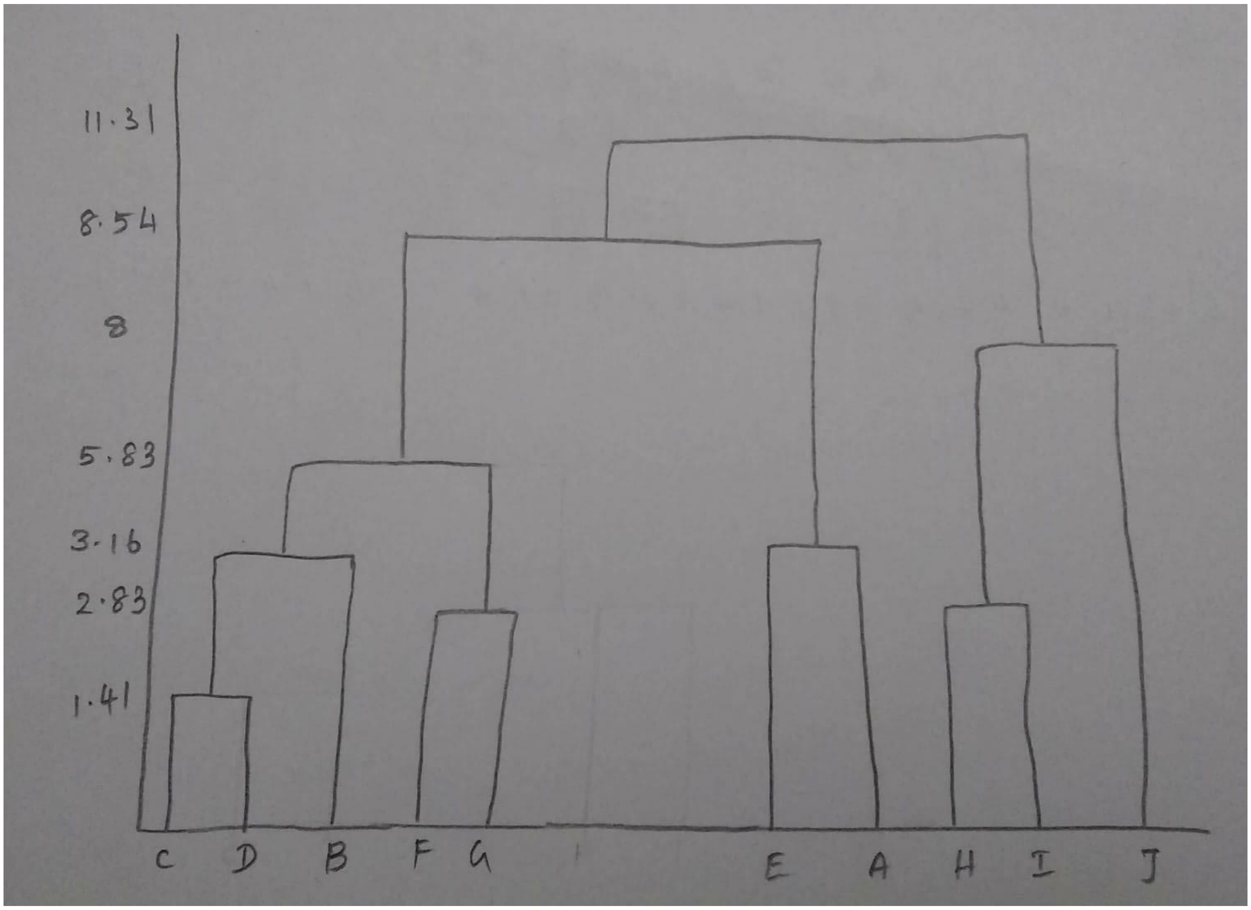
	A-E	F-G	H-I	J
C-D-B	8.54	5.83	11.31	8
A-E	—	6.4	8.06	10.63
F-G		—	5.83	7.07
H-I			—	8
J				—

	A-E	H-I	J
C-D-B-F-G	8.54	11.31	8
A-E	—	8.06	10.63
H-I		—	8
J			—

	A-E	H-I-J
C-D-B-F-G	8.54	11.31
A-E	—	10.63
H-I-J		—

	H-I-J
C-D-B-F-G-A-E	11.31

The dendogram according to order of height is :



(b) (5 points) If we assume there are three clusters, which of the single and complete link hierarchical clustering will give better resulted clusters? Justify your answer.

If we assume 3 clusters:

In single link, the 3 clusters are : {A,B,C,D,E,F,G},{H,I},{J}

In complete link, the 3 clusters are : {A,E}, {B,C,D,F,G}, {H,I,J}

If we observe, in single link, the cluster {J} is all by itself.

This is like an indication that it might be an outlier.

If we actually check the distance from J to every other point, we could observe that J is the farthest point from any other point.

The minimum distance from J to any other point is 4.24 (to G).
Considering it as an outlier, Single link clustering distinguished it as an outlier by not assigning any other points in that cluster.

Since detecting outliers is an important thing, Single link did a better job over Complete link.

Justification 2 :

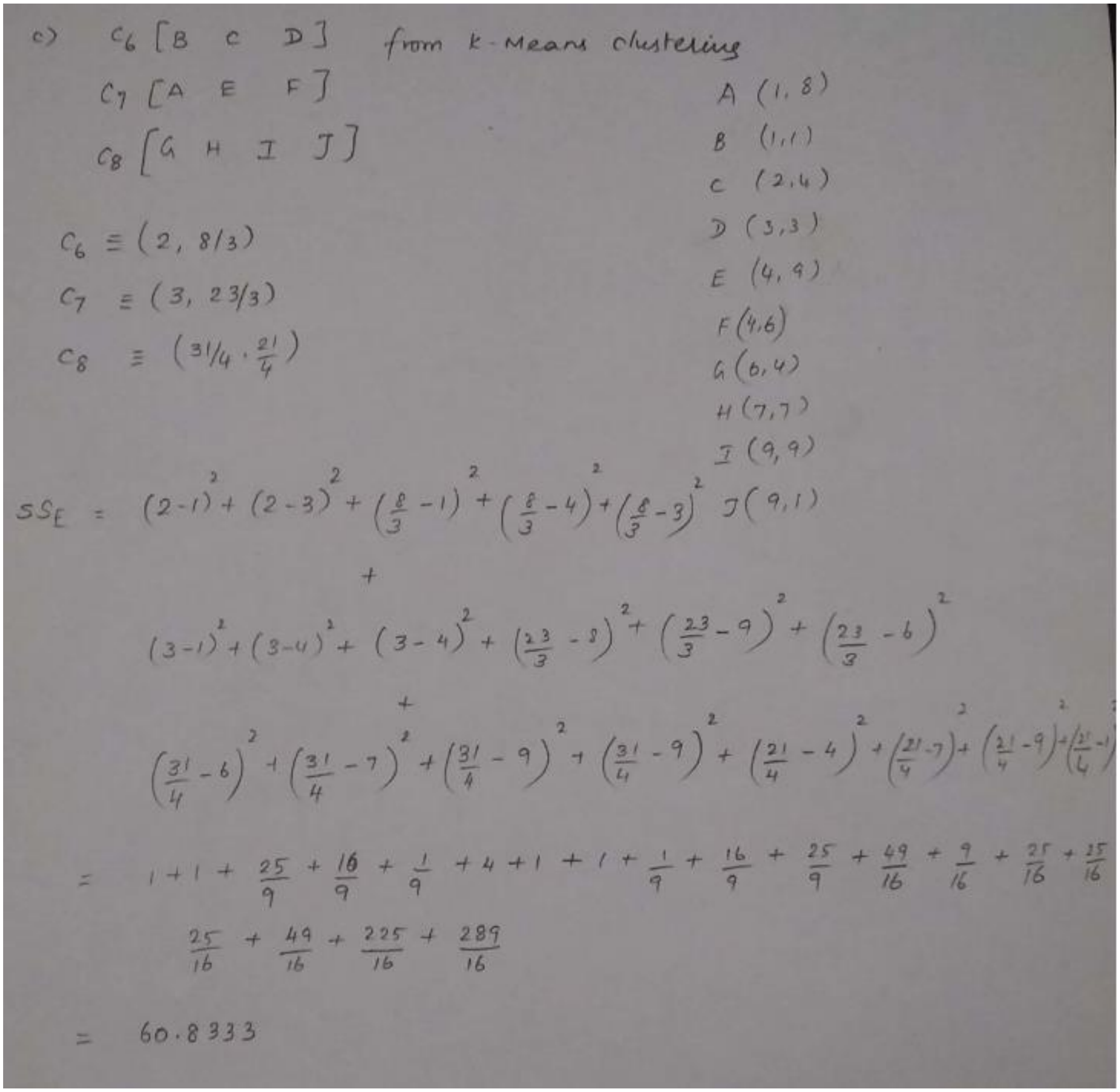
> If we consider the SSE values though, we get Complete link hierarchial clustering as better since it's value is bit less than Single link clustering.

> Single link may produce big and odd-shaped clusters since it considers local minimum values and forms big chains of nodes.

> In complete link, the breaking of the dendogram where we can see clearly 3 clusters is possible. In this view, since breaking clusters is easy, we can say complete link is better than single link.

Thus we can't really say what the best clustering method is , it all depends on the situation.

(c) (5 points) Compare your resulted clusters from 2(b) with the resulted clusters using K-means in Question 1 by calculating their corresponding Sum of Squared Error (SSE). Based on their SSE results, which resulted clusters, 1(b) or 2(b), are better?



c) C_6 [B C D] from k-Means clustering

C_7 [A E F]

C_8 [G H I J]

A	(1, 8)
B	(1,1)
C	(2,4)
D	(3,3)
E	(4, 9)
F	(4,6)
G	(6,4)
H	(7,7)
I	(9,9)
J	(9,1)

$C_6 \equiv (2, 8/3)$

$C_7 \equiv (3, 23/3)$

$C_8 \equiv (31/4, \frac{21}{4})$

$$SS_E = (2-1)^2 + (2-3)^2 + \left(\frac{8}{3}-1\right)^2 + \left(\frac{8}{3}-4\right)^2 + \left(\frac{8}{3}-3\right)^2$$

$$+$$

$$(3-1)^2 + (3-4)^2 + (3-4)^2 + \left(\frac{23}{3}-8\right)^2 + \left(\frac{23}{3}-9\right)^2 + \left(\frac{23}{3}-6\right)^2$$

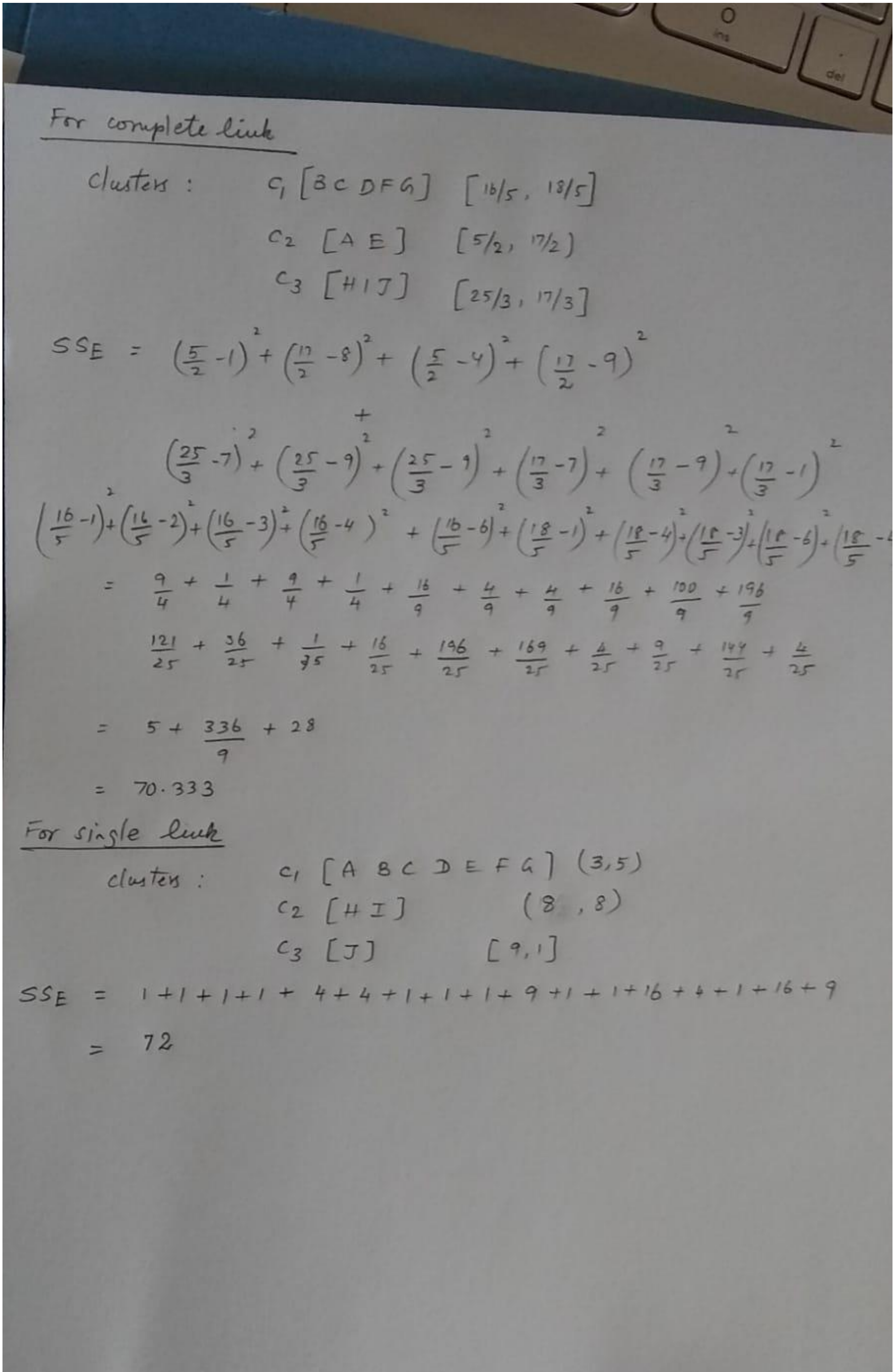
$$+$$

$$\left(\frac{31}{4}-6\right)^2 + \left(\frac{31}{4}-7\right)^2 + \left(\frac{31}{4}-9\right)^2 + \left(\frac{31}{4}-9\right)^2 + \left(\frac{21}{4}-4\right)^2 + \left(\frac{21}{4}-7\right)^2 + \left(\frac{21}{4}-9\right)^2 + \left(\frac{21}{4}-1\right)$$

$$= 1 + 1 + \frac{25}{9} + \frac{16}{9} + \frac{1}{9} + 4 + 1 + 1 + \frac{1}{9} + \frac{16}{9} + \frac{25}{9} + \frac{49}{16} + \frac{9}{16} + \frac{25}{16} + \frac{25}{16}$$

$$\frac{25}{16} + \frac{49}{16} + \frac{225}{16} + \frac{289}{16}$$

$$= 60.8333$$



For complete link

clusters : C_1 [B C D F G] [16/5, 18/5]

C_2 [A E] [5/2, 17/2)

C_3 [H I J] [25/3, 17/3]

$$SS_E = \left(\frac{5}{2}-1\right)^2 + \left(\frac{17}{2}-8\right)^2 + \left(\frac{5}{2}-4\right)^2 + \left(\frac{17}{2}-9\right)^2$$

$$+$$

$$\left(\frac{25}{3}-7\right)^2 + \left(\frac{25}{3}-9\right)^2 + \left(\frac{25}{3}-9\right)^2 + \left(\frac{17}{3}-7\right)^2 + \left(\frac{17}{3}-9\right)^2 + \left(\frac{17}{3}-1\right)^2$$

$$\left(\frac{16}{5}-1\right)^2 + \left(\frac{16}{5}-2\right)^2 + \left(\frac{16}{5}-3\right)^2 + \left(\frac{16}{5}-4\right)^2 + \left(\frac{16}{5}-6\right)^2 + \left(\frac{18}{5}-1\right)^2 + \left(\frac{18}{5}-4\right)^2 + \left(\frac{18}{5}-3\right)^2 + \left(\frac{18}{5}-6\right)^2 + \left(\frac{18}{5}-\right.$$

$$= \frac{9}{4} + \frac{1}{4} + \frac{9}{4} + \frac{1}{4} + \frac{16}{9} + \frac{4}{9} + \frac{4}{9} + \frac{16}{9} + \frac{100}{9} + \frac{196}{9}$$

$$\frac{121}{25} + \frac{36}{25} + \frac{1}{25} + \frac{16}{25} + \frac{196}{25} + \frac{169}{25} + \frac{4}{25} + \frac{9}{25} + \frac{144}{25} + \frac{4}{25}$$

$$= 5 + \frac{336}{9} + 28$$

$$= 70.333$$

For single link

clusters : C_1 [A B C D E F G] (3,5)

C_2 [H I] (8 , 8)

C_3 [J] [9, 1]

$$SS_E = 1+1+1+1+4+4+1+1+1+9+1+1+16+4+1+16+9$$

$$= 72$$

By looking at the SSE results, it's evident that K-Means is more optimal than Single link and Complete link hierarchical clustering.

From the above answer (2b), Single link clustering is proved to be good.

If we compare SSE of Single link and K-Means, K-means SSE is low.

On a general note, K-Means takes into consideration every cluster centre and then tries to get data points assigned to one of the closest clusters.

Hence, it's a better way of clustering than Hierarchial clustering.

3. (12 points) [Song Ju] For the transaction Table 1 given below, please answer the following questions:

TID Items Bought

T1 {B,D,F,H}

T2 {C,D,F,G}

T3 {A,D,F,G}

T4 {A,B,C,D,H}

T5 {A,C,F,G}

T6 {D,H}

T7 {A,B,E,F}

T8 {A,D,F,G,H}

T9 {A,C,D,F,G}

T10 {D,F,G,H}

T11 {A,C,D,E}

T12 {B,E,F,H}

T13 {D,F,G}

T14 {C,F,G,H}

T15 {A,C,D,F,H}

(a) (3 points) Explain what is frequent itemset and give an example of 2-itemset that is frequent itemset with support count = 8.

– An itemset from the given data whose support is greater than or equal to a minimum support threshold

Ex of 2-itemset that is frequent.

Here, support count =8.

{D,F} and {F,G}

(b) (3 points) Explain what is closed frequent itemset and list all of them with support count = 8.

Closed Frequent set :

The frequent itemsets for those none of their immediate supersets are having same support count as the itsemset.

Here, the sets that are closed frequent are :

A,D,F,H, DF, GF

(c) (3 points) Explain what is maximal frequent and list all of maximal itemset with support count = 8.

Maximal Frequent set:

The frequent itemsets whose none of the immediate supersets are frequent.

Here, the sets that are maximal frequent are:

A, H,DF, GF

(d) (3 points) Compute the support and confidence for association rule{D,F}→{G}.

Support : frequency of itemset {D,F,G}/ 15

Here, we have 15 transactions

= 6/15

=0.4

Confidence : frequency of itemset {D,F,G}/ frequency of itemset {D,F}

= 6/8

= 0.75

4. (13 points) [Association Analysis] [Ruth Okoilu] Consider the following market basket transactions shown in the Table 2 below.

Transaction ID Items ordered

1 {Flour, Eggs, Bread}

2 {Soda, Coffee}

3 {Flour, Butter, Milk, Eggs}

4 {Bread, Eggs, Juice, Detergent}

5 {Bread, Milk, Eggs}

6 {Eggs, Bread}

7 {Detergent, Milk}

8 {Coffee, Soda, Juice}

9 {Butter, Juice, Bread}

10 {Milk, Bread, Detergent}

For each of the following question, briefly explain your answers in 2-3 sentences.

(a) (2 points) How many items are in this data set? What is the maximum size of itemsets that can be extracted from this data set (only including itemsets that have ≥ 1 support)?

There are 9 items in this dataset.

The maximum size of itemset that can be extracted is 4.

EX :

{Bread, Eggs, Juice, Detergent}

{Flour, Butter, Milk, Eggs}

(b) (2 points) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

The maximum number of association rules are :

3^d – 2^(d+1)+1 where d is number of distinct items which is 9 in this case.

Therefore, 3^(9)-2^(10)+1

=19683-1024+1

18660

(c) (2 points) What is the maximum number of 2-itemsets that can be derived from this data set (including those have zero support)?

The maximum number of 2 itemsets that can be derived are 36.

Since there are 9 distinct itemsets and if we consider all possible combinations of length 2, 9C2 = 36.

(d) (3 points) Find an itemset (of size 2 or larger) that has the largest support.

The itemset {Eggs, Bread} has the largest support.

Support = #{Eggs, Bread}/total # of transactions

=4/10

=0.4

(e) (4 points) Given minconf = 0.5, find two pairs of items, a and b, such that the rules {a}→{b} and {b}→{a} have the same confidence, and their confidence is greater than or equal to the minconf threshold.

The 2 examples for the above requirement is :

1) Flour and Butter

If we consider Flour → Butter

Confidence = #{Flour, Butter}/#{Flour}

=1/2 = 0.5

If we consider Butter→ Flour

Confidence = # {Flour, Butter}/#{Butter}

= 1/2 =0.5

Here, for both rules the confidence is 0.5 which is equal to minconf given.

2) {Flour, Milk} and {Butter, Egg}

{Flour, milk} → {Butter, Egg}

Confidence = # {Flour, Milk, Butter, Egg} /# {Flour, Milk}

= 1/1

=1

If we consider

{Butter, Egg} → {Flour, milk}

Confidence = # {Flour, Milk, Butter, Egg} /#{Butter, Egg}

=1/1

=1

Here, for both rules the confidence is 1 which is greater than minconf given.

5. (20 points) [Apriori algorithm][Xi Yang] Consider the data set shown in Table 3 and answer the following questions using apriori algorithm.

TID	**Items**
t1	**A,C,D,E**
t2	**A,B,D,E**
t3	**C,E**
t4	**C,D**
t5	**A,B,D**
t6	**B,D,E**
t7	**A,C,D**
t8	**B,C,D,E**

(a) (10 points) Show (compute) each step of frequent item set generation process using apriori algorithm, with support count of 3.

Step 1 : generate Candidate set with itemset size 1

A-4

B-4

C-5

D-7

E-5

Since all the candidates have minimum support count, L1 would have all these candidates.

C2:

AB-2

AC-2

AD-3

AE-2

BC-1

BD-4

BE-3

CD-4

CE-3

DE-4

Eliminating the itsemsets whose support is less than 3, we get L2 :

AD

BD

BE

CD

CE

DE

Now, generating itemsets of size 3, we get C3:

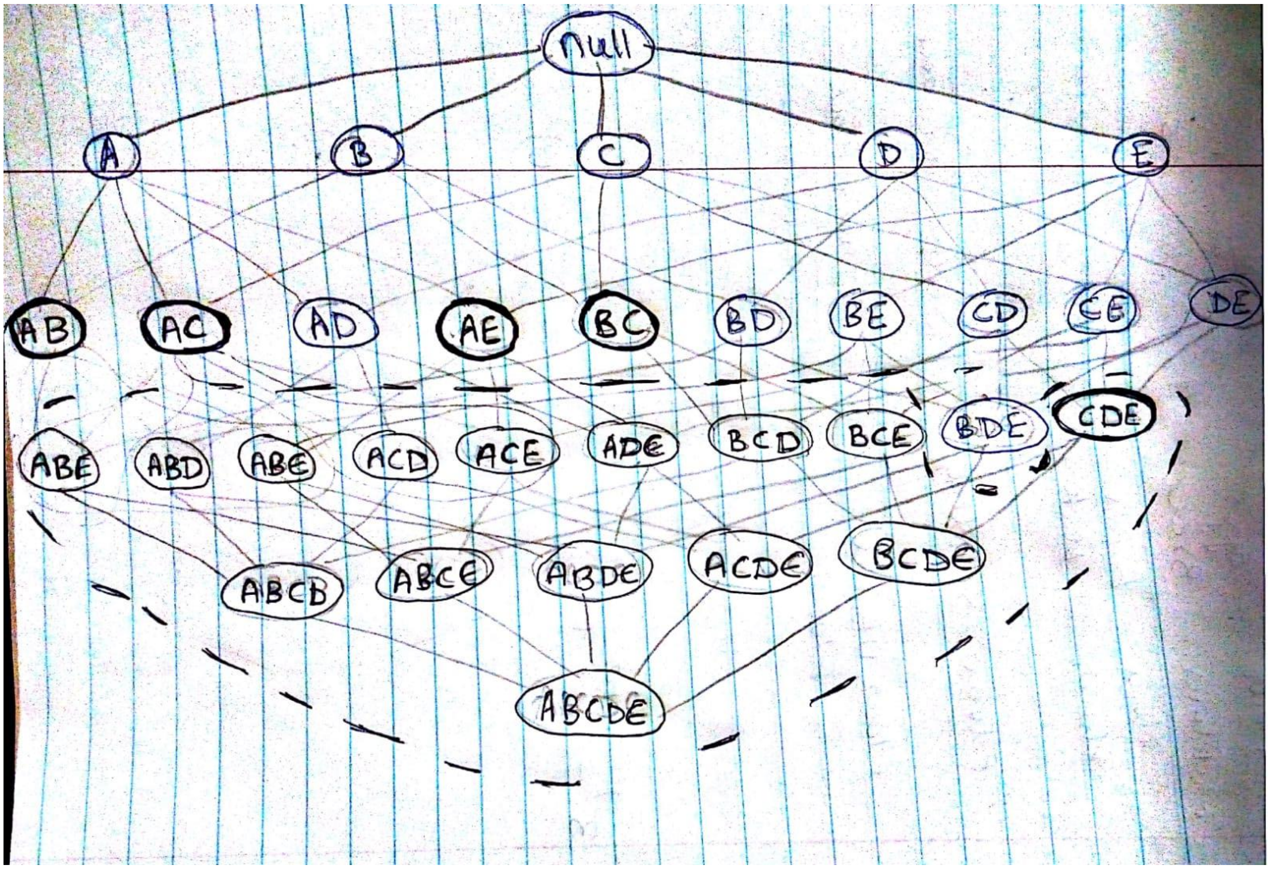
BDE-3

CDE-2.

Therefore the itemsets that are frequent are :

{A},{B},{C},{D},{E}, {A,D}, {B,D}, {B,E},{C,D},{C,E},{D,E},{B,D,E}

(b) (10 points) Show the lattice structure for the data given in table above, and mark the pruned branches if any. (Scanned hand-drawing is acceptable as long as it is clear.)



In the picture above, The dark blackened circles are pruned initially since their support count is less that minimum support given.

The section enclosed under dotted lines are the itemsets pruned since their subsets are not frequent.

For itemset of size 2, 4 nodes are discarded which are : AB, AC, AE, BC

Because of discarding them, except the itemsets : {B,D,E} and {C,D,E}, all others are discarded.

Among these 2, {C,D,E} support is less, so it's discarded.

Thus further itsemsets need not be checked, since all are pruned.

6. (22 points) [Frequent Pattern Tree][Song Ju] Consider the following data set shown in Table 4 and answer the following questions using FP-Tree.

TID	Items Bought
T1	{B,D,F,H}
T2	{C,D,F,G}
T3	{A,D,F,G}
T4	{A,B,C,D,H}
T5	{A,C,F,G}
T6	{D,H}
T7	{A,B,E,F}
T8	{A,D,F,G,H}
T9	{A,C,D,F,G}
T10	{(D,F,G,H}
T11	{A,C,D,E}
T12	{B,E,F,H}
T13	{D,F,G}
T14	{C,F,G,H}
T15	{A,C,D,F,H}

(a) (12 points) Construct an FP-tree for the set of transactions in the table below as the first step towards identifying the itemsets with minimum support count of 2 (at least 2 occurrences). Do not forget to include the header table that locates the starts of the corresponding linked item lists through the FP-tree. For consistency, please form your header table in the order of {F, D, G, H, A, C, B, E}

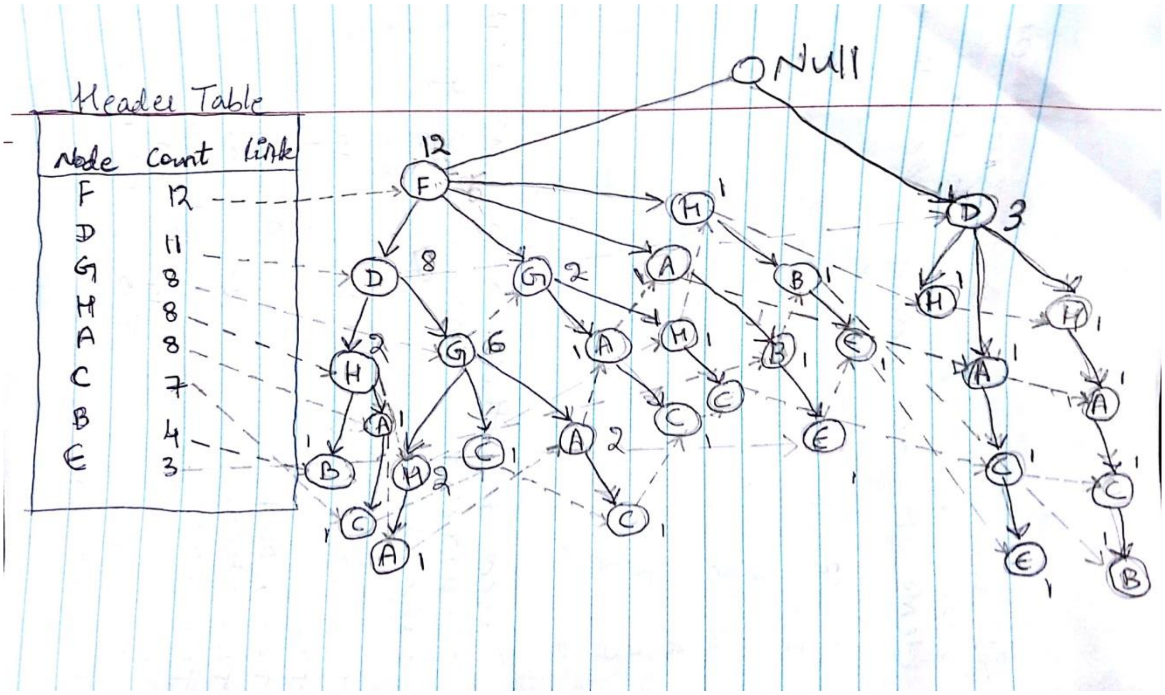
If we get frequency count of every item, the order of items in descending order is :

F	12
D	11
G	8
H	8
A	8
C	7
B	4
E	3

Based on this order, we will re-arrange items in every transaction.

T1	FDHB
T2	FDGC
T3	FDGA
T4	DHACB
T5	FGAC
T6	DH
T7	FABE
T8	FDGHA
T9	FDGAC
T10	FDGH
T11	DACE
T12	FHBE
T13	FDG
T14	FGHC
T15	FDHAC

Since every item has atleast minimum support count of 2, no item is discarded and below is the FP tree.



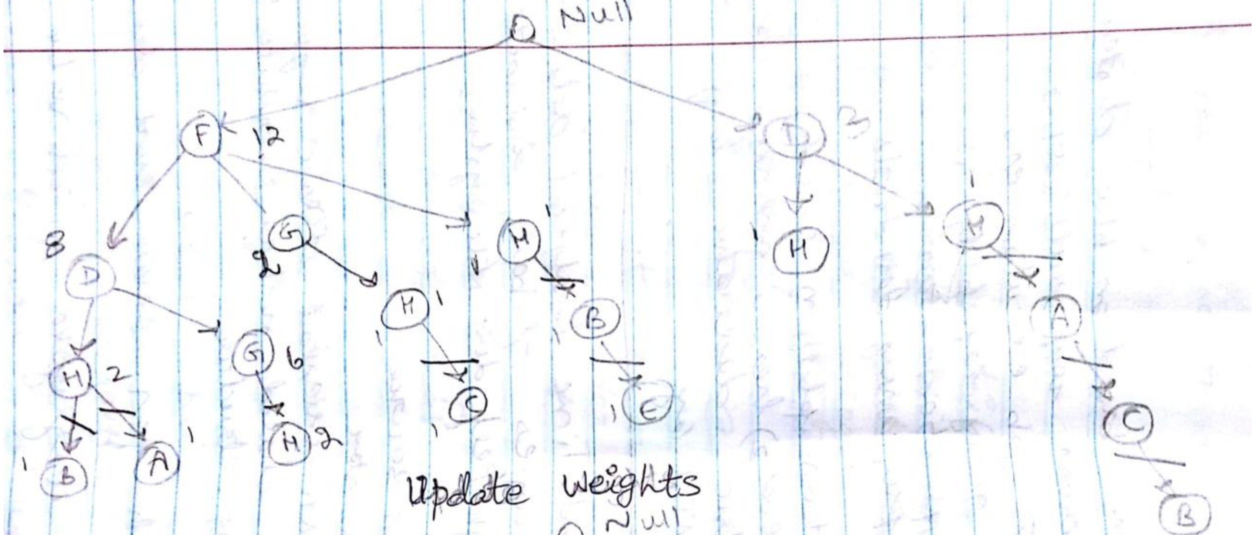
Header Table

Node	Count	Link
F	12	
D	11	
G	8	
H	8	
A	8	
C	7	
B	4	
E	3	

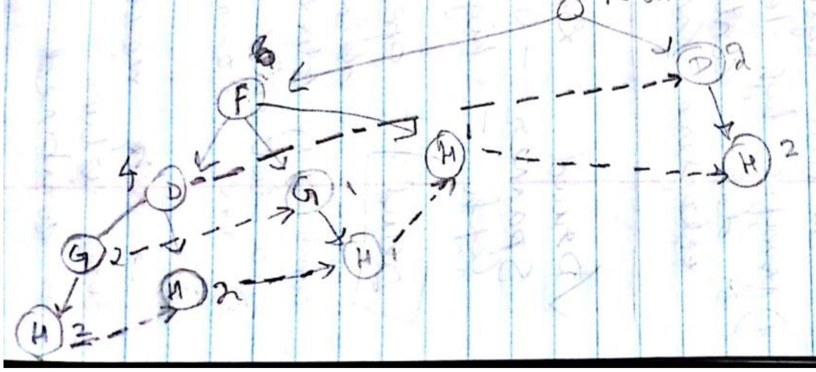
(b) (10 points) Using the FP-Tree constructed and support=3, generate all the frequent patterns with the base of item H step by step.

Taking the base as H, we have transactions in which end item ending with H or along the path, H can be there.

Conditional FP Tree on H



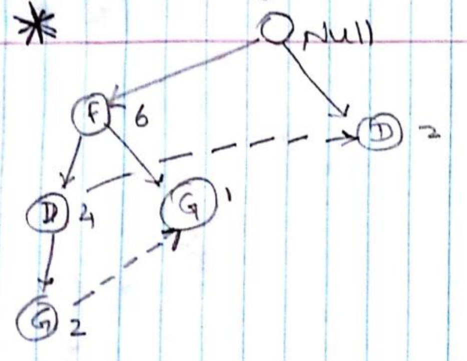
Update weights



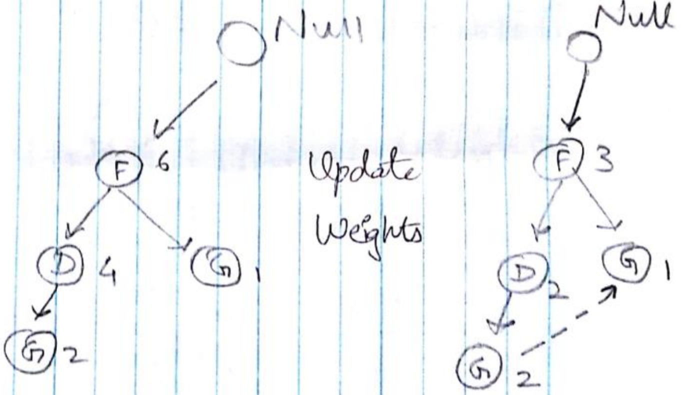
The counts for H is 8 which is ≥ 3.

So, we do not need any H nodes.

Pruning H nodes, we get :-

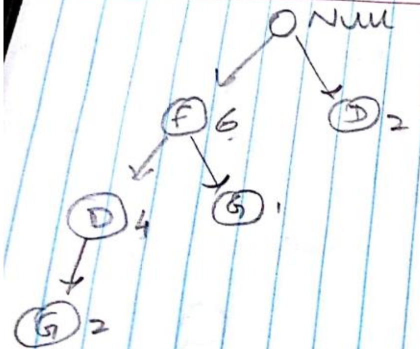


Now, we see if GH is frequent by extracting paths ending with G.

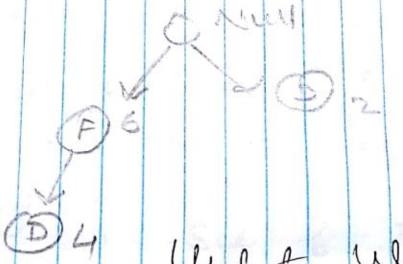


Here, the count of G = 3
⇒ Its frequent.

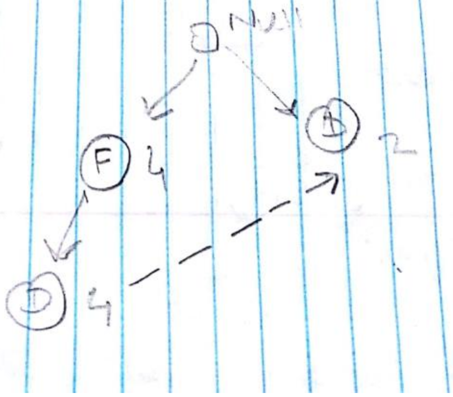
Now, ~~prune G~~, check for DH from graph marked with *



See for nodes ending with D.



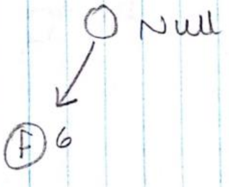
Update Weights



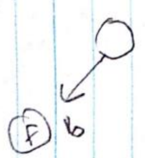
Here, the count for D is 6 which is ≥ 3.

∴ D is frequent.

Now check for FH from graph marked with *



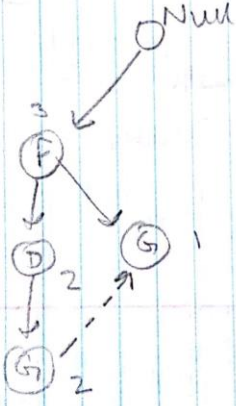
update weights



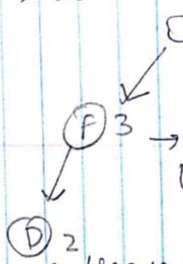
F count is 6 which is ≥ 3. ∴ F is frequent.

Now we have frequent conditional trees for H, DH, FH & GH.

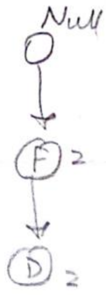
From GH conditional FP tree, we see if DGH is frequent



See if 'D' is frequent. Prune G & check

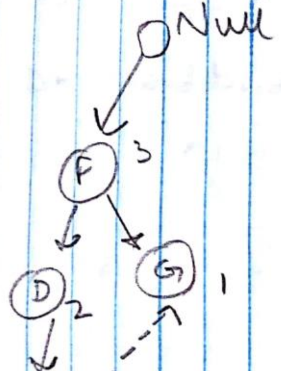


→ update weights

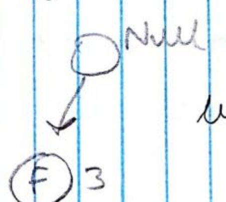


D = 2 ⇒ infrequent.

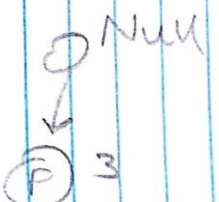
Now see if F is frequent on conditional GH tree.



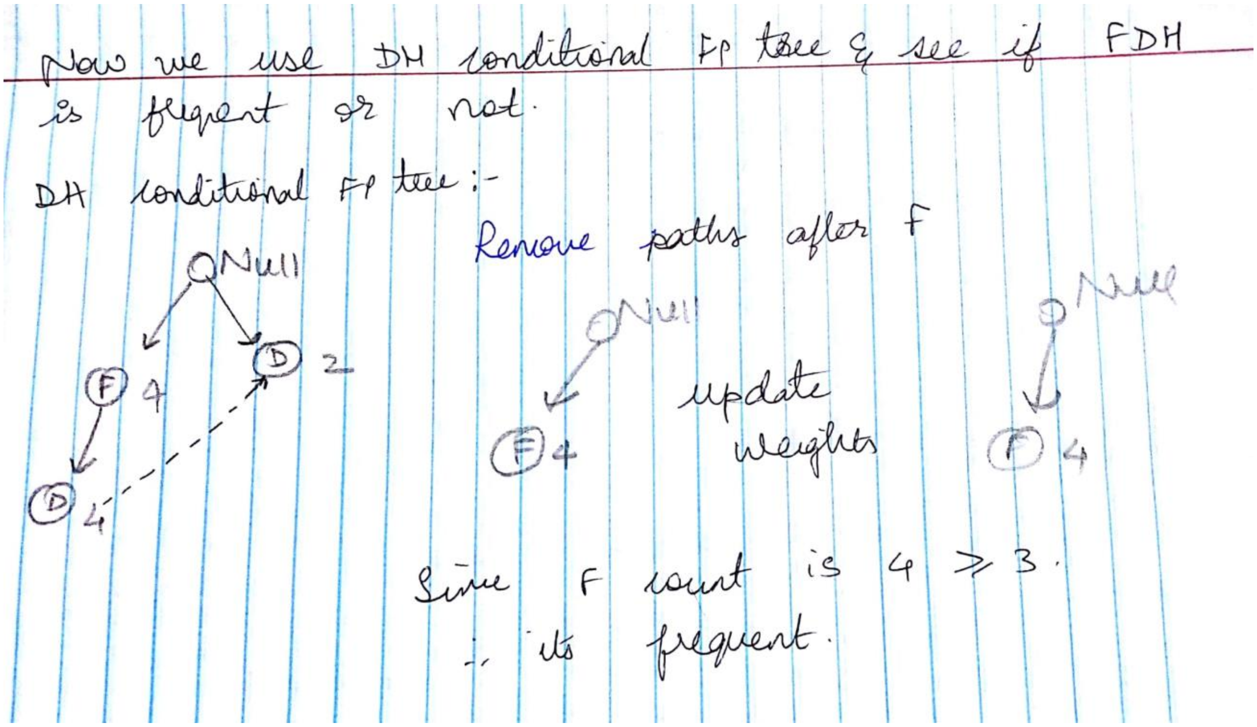
Removing all paths after F, we get



update weights



F is frequent as its count = 3 which is ≥ 3.



Now we use DH conditional FP tree & see if FDH is frequent or not.

DH conditional FP tree :-

Remove paths after F

update weights

Since F count is 4 ≥ 3.

∴ its frequent.

Null
F 4
D 2
D 4
Null
F 4
Null
F 4

Repeating this process and getting the Conditional tree of {FDH} and {FGH}, would anyway not give any nodes except null node, thus are not shown in the diagrams.

Therefore the frequent itemsets are : {H}, {G,H},{D,H},{F,H},{F,D,H},{F,G,H}.