ASSIGNMENT 3

Frequent item - sets

Exercise 1

Solution (1)

Given There are 1 to 100 ~~na~~ ~~st~~ numbers i.e. 100 itemsets and 100 baskets.

Item i is in basket b if ~~b % i = 0~~

$\frac{b}{i}$ has remainder as 0.

We can do it till 100 or we can do it at a certain point such that we understand the pattern

baskets b	Item sets i
1	{1}
2	{1, 2}
3	{1, 3}
4	{1, 2, 4}
5	{1, 5}
6	{1, 2, 3, 6}
7	{1, 7}
8	{1, 2, 4, 8}
9	{1, 3, 9}
10	{1, 2, 5, 10}

Scanned with CamScanner

baskets b	itemsets i
11	{1, 11}
12	{1, 2, 3, 4, 6, 12}
13	{1, 13}
14	{1, 2, 7, 14}
15	{1, 3, 5, 15}
16	{1, 2, 4, 8, 16}
17	{1, 17}
18	{1, 2, 3, 6, 18}
19	{1, 19}
20	{1, 2, 5, 10, 4, 20}
21	{1, 3, 7, 21}
22	{1, 2, 11, 22}

Support threshold = 5

~~Support threshold is how many times an item~~

Support ~~threshold~~ is in how many basket an item came. i.e. how many ~~times~~ ~~or~~ baskets an item i was a subset.

For example

Item 1 appears in every basket from 1 to 100.

Scanned with CamScanner

Hence
support of item 1 is 100

Frequent items are those where the items passes support threshold of 5. ~~so~~
According to above 1 ~~it~~ has support of 100 which is greater than or equal to support threshold of 5. Hence 1 is a frequent item.

Now find out ~~support for~~ items that has support more than 5 so that you find the ~~items~~ frequent itemsets.

Support of Item 2 = 50 → since it appears in 50 even baskets.

Support of item 3 = 33
Support of Item 4 = 100/4 = 25

As you can see whatever is the Quotient of $\frac{\text{basket}}{\text{item}}$, avoiding the remainder is the support.
So we generated a pattern.
Now since we want support greater than ~~5~~ or equal to 5, we cannot get

Scanned with CamScanner

5 or greater for items ~~more~~ greater than 20. because.

$$\frac{100}{20} = 5$$

$\therefore$ support of 20 = 5

~~at~~ which is valid

but then ~~at~~ for every other item ~~greater~~ greater than 20 we ~~get~~ will get support less than 5.

Hence, all the items from ~~1, 1 to 100~~ 1 to 20, are frequent items.

Hence, frequent items

{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}

Scanned with CamScanner

Solution (2)

Association rule $I \to j$: where I is a set of items and j is an item, implicates that if all the items in I appears in some basket then there is likeliness of j also appearing in the same basket.

The likeliness is given by confidence of rule $I \to j$

Confidence of $I \to j$ is defined as the ratio of support for $I \cup \{j\}$ ~~and~~ to the support for I.

In this question, this means that confidence is the ratio of number of baskets all items in I ~~appears to the number of baskets in~~ and item j appears to the number of baskets all items in I appears.

(a) $\{5, 7\} \to 2$

Here to find ~~confidence~~ confidence for $\{5, 7\}$

we find support for $\{5, 7\}$ and then support for $\{5, 7\} \cup \{2\}$.

Scanned with CamScanner

Support for $\{5, 7\}$ = $\{5\}$ and $\{7\}$ ~~will~~ will appear in baskets multiple of L.C.M. of 5 and 7 which is 35.

$\therefore$ 5 and 7 will appear in ~~3~~ baskets 35 and 70.

Hence

Support for $\{5, 7\} = 2$

Now, support for $\{5, 7\} \cup \{2\}$ = $\{5, 7\}$ and $\{2\}$ will appear together in ~~a~~ ~~basket~~ only basket 70.

Hence, Support ~~for~~ $\{5, 7\} \cup \{2\} = 1$.

~~the~~ $\therefore$ Confidence for $\{5, 7\} \rightarrow 2 = \dfrac{\text{Support for } \{5, 7\} \cup \{2\}}{\text{Support for } \{5, 7\}}$

$\therefore$ Confidence for $\{5, 7\} \rightarrow 2 = \dfrac{1}{2}$

$\therefore$ Confidence for $\{5, 7\} \rightarrow 2$ is $\dfrac{1}{2}$

Scanned with CamScanner

(b) $\{2,3,4\} \rightarrow 5$

Support for $\{2,3,4\}$ = $\{2,3,4\}$ will appear in baskets which are multiples of LCM 2, 3, 4 ~~which~~ i.e. 12.

Hence $\{2,3,4\}$ will appear in baskets 12, 24, 36, 48, 60, 72, 84, 96

$\therefore$ Support for $\{2,3,4\} = 8$

Support for $\{2,3,4\} \cup \{5\} = \{2,3,4,5\}$ will appear only in basket 60

Support for $\{2,3,4\} \cup \{5\} = 1$

$\therefore$ Confidence $\{2,3,4\} \rightarrow 5$ $= \dfrac{\text{Support for } \{2,3,4\} \cup \{5\}}{\text{Support for } \{2,3,4\}}$

$= \dfrac{1}{8}$

$\therefore$ Confidence for $\{2,3,4\} \rightarrow 5$ is $\dfrac{1}{8}$

Scanned with CamScanner

Exercise 3 Clustering

(1) One dimensional set of points

1, 4, 9, 16, 25, 36, 49, 64, 81

For performing hierarchical clustering,

1) we will merge two clusters having the closest distance

2) After 2 points are merged into cluster, the centroid of that cluster will be used to ~~find~~ compare the distance between cluster and any remaining datapoints.

Step 1

difference between two data points

1 4 9 16 25 36 49 64 81

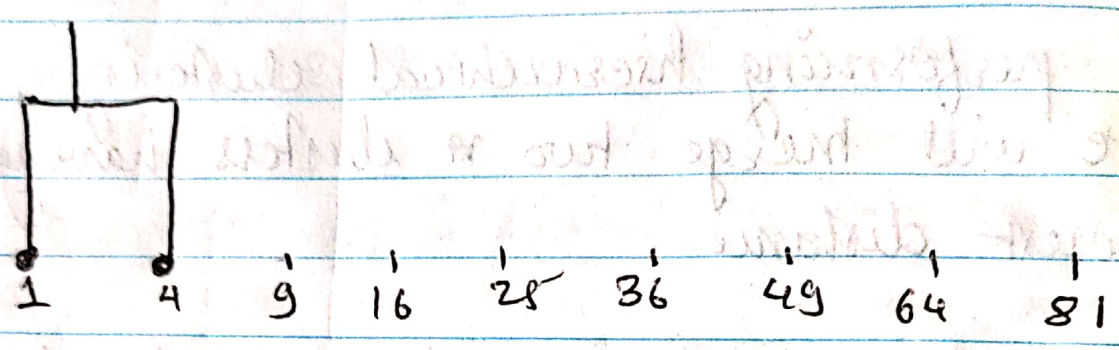
3 5 7 9 11 13 15 17

Now, as you can see that distance between 1 and 4 is closest so we merge them.

(1 4) 9 16 25 36 49 64 81

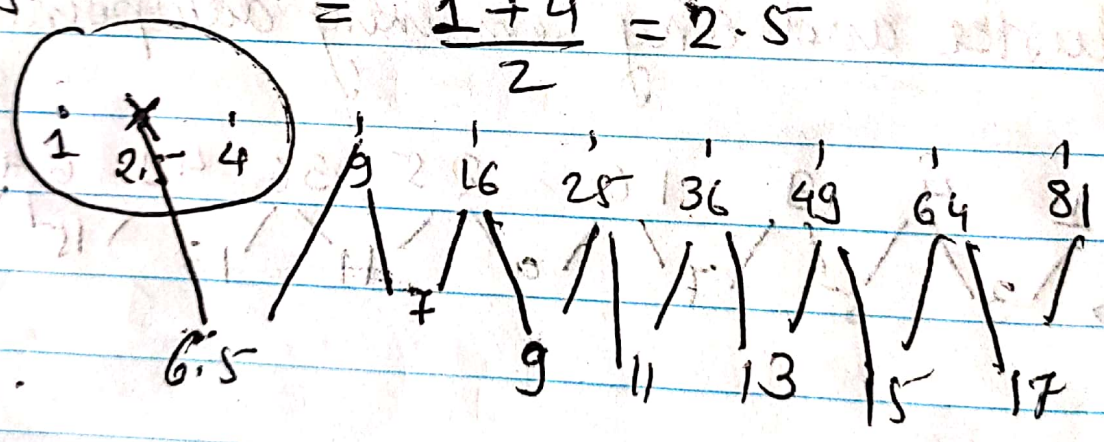
Scanned with CamScanner

Dendogram for step 1



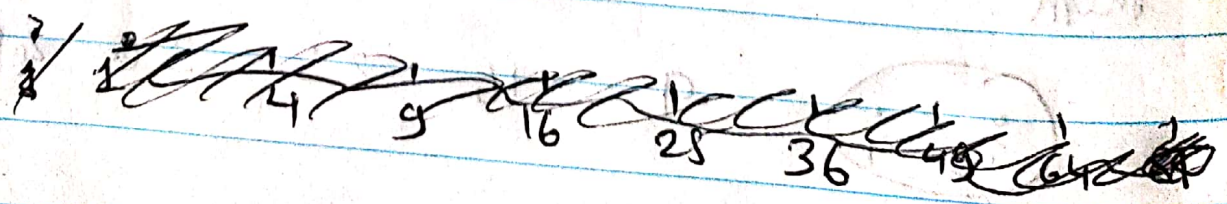
Step 2 Now 1 and 4 are cluster so we consider the average point to compare with other data points.

$$= \frac{1+4}{2} = 2.5$$

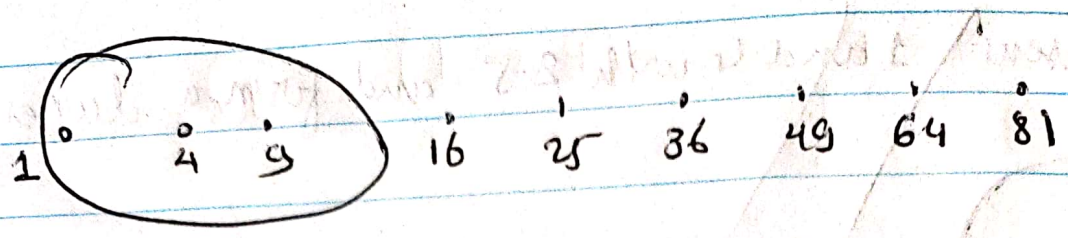


Difference between two points.

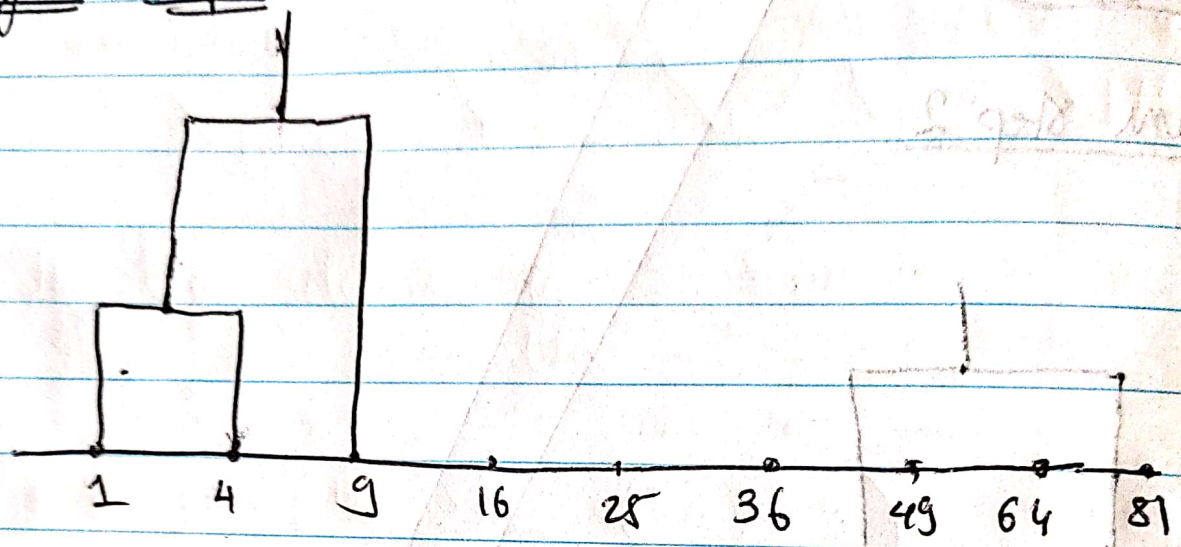
So now 2.5 and 9 are merged. as 6.5 is the closest distance.



Scanned with CamScanner



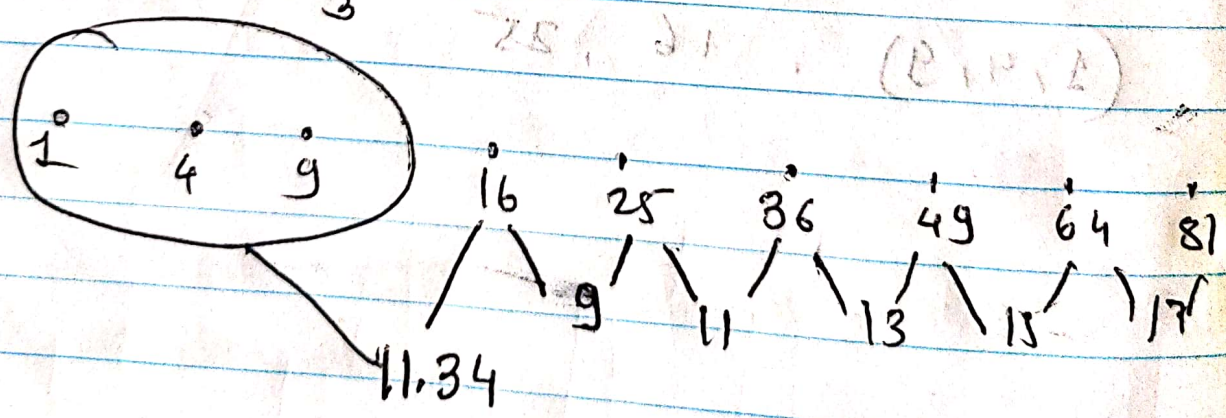
Dendrogram Step 2



~~Step 3~~ Iteration 3

Now ~~to~~ calculate the centroid of the cluster formed (1, 4, 9)

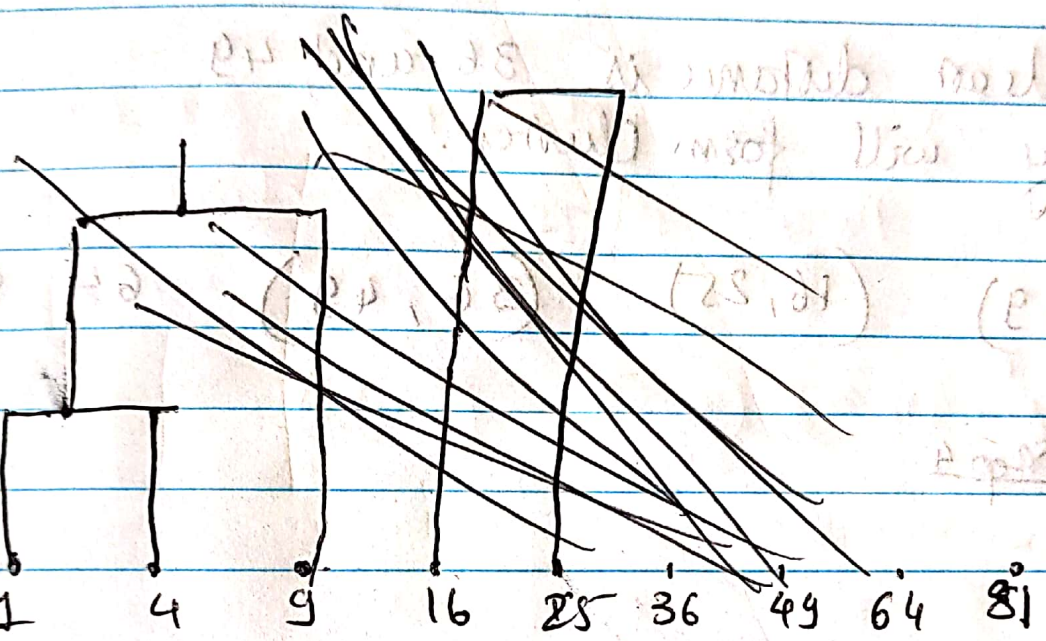
$= \frac{14}{3} = 4.66$

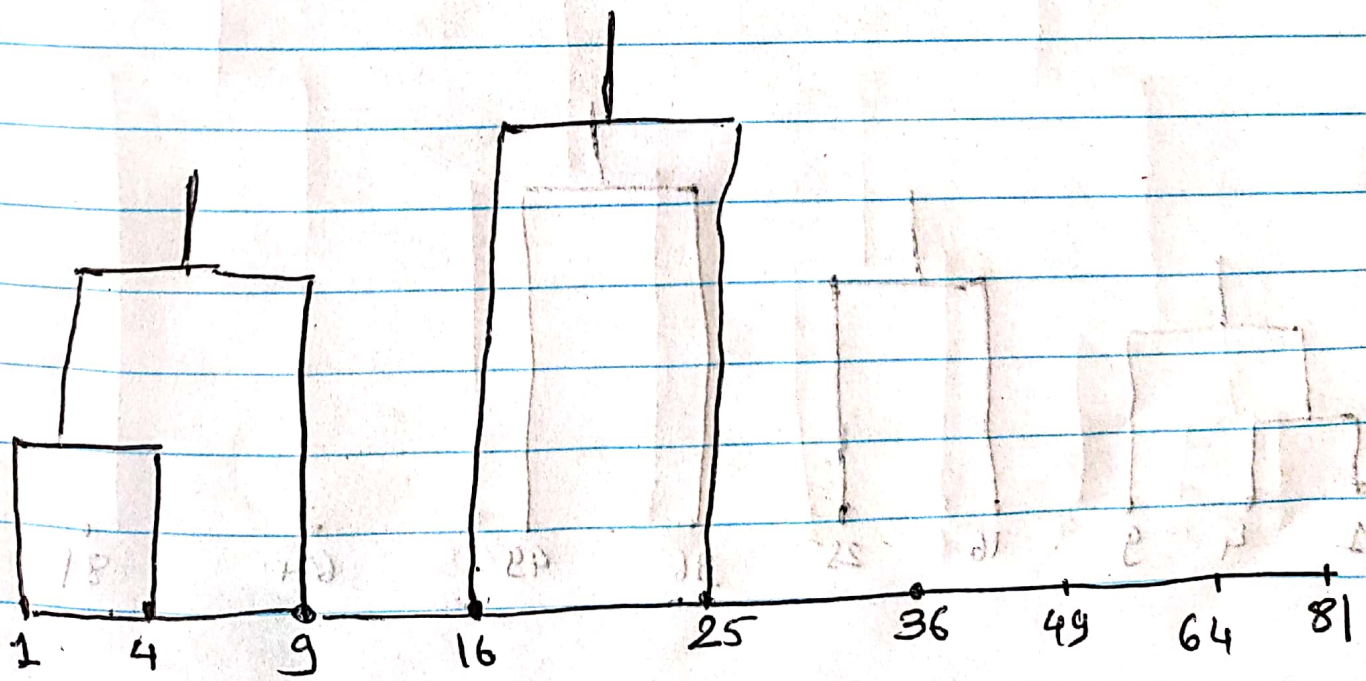


Here (16, 25) are the closest so they will form a cluster

(1, 4, 9) (16, 25) 36 49 64 81

Dendogram Step 3





Scanned with CamScanner

Iteration Centroid

Step 4

$C_1 = (1, 4, 9) \longrightarrow 4.66$

$C_2 = (16, 25) \longrightarrow 20.5$

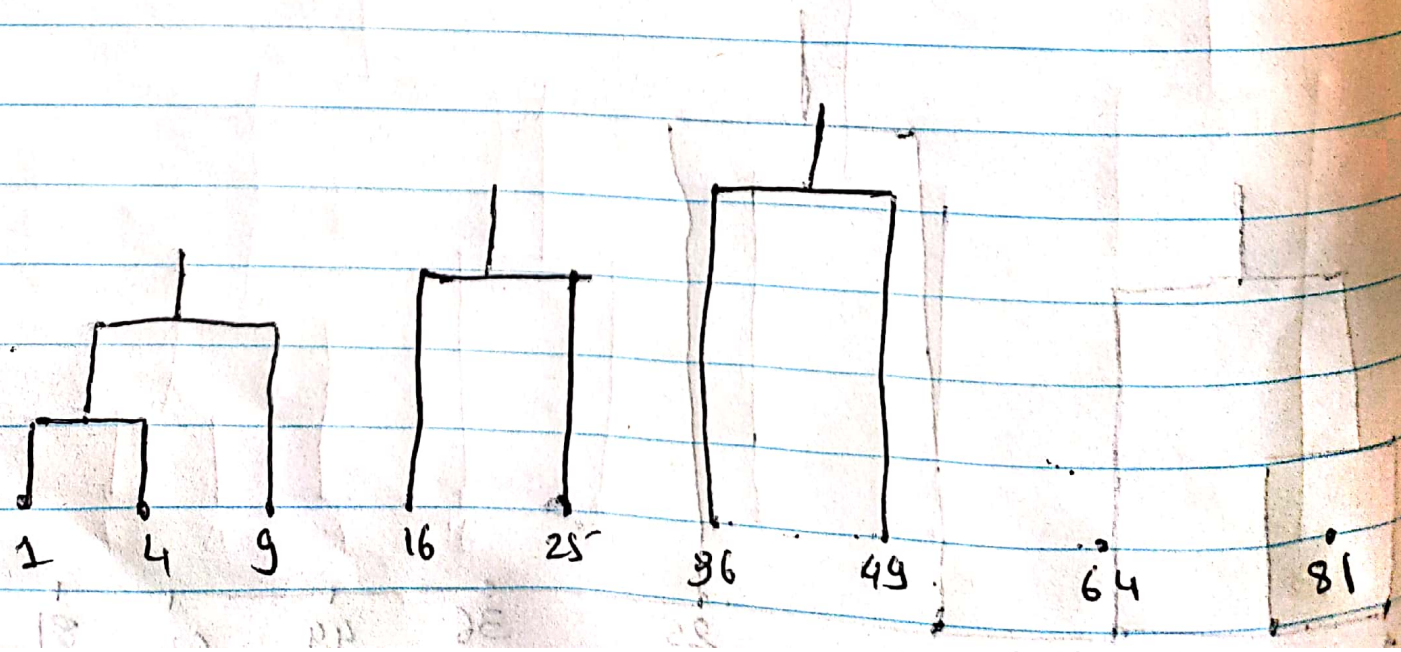
	(1, 4, 9)	(16, 25)	36	49	64	81
Centroids	4.66	20.5	36	49	64	81

15.84 15.5 13 15 17

The least distance is 36 and 49 so they will form cluster.

(1, 4, 9) (16, 25) (36, 49) 64 81

Dendogram Step 4



Scanned with CamScanner

Iteration ~~Step~~ 5

(1, 4, 9) (16, 25) (36, 49) 64 81

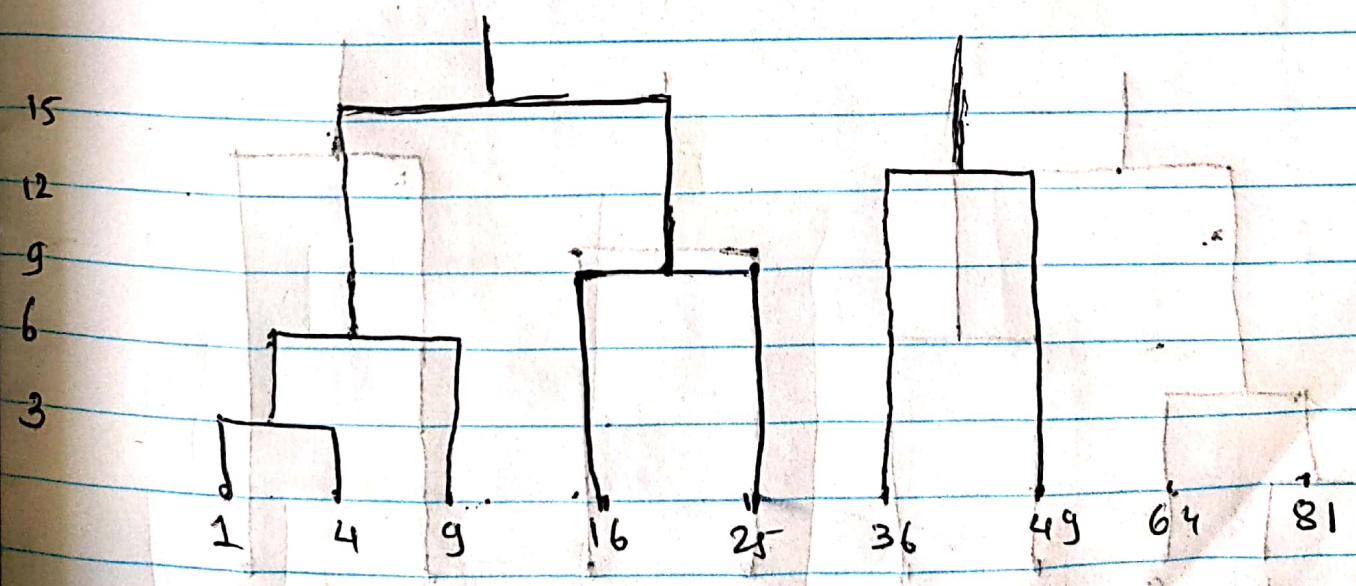
Centroids 4.66 20.5 42.5 64 81

Difference 15.84 22 21.5 17

Next cluster is (1, 4, 9) and (16, 25)

∴ (1, 4, 9, 16, 25) (36, 49) 64 81

Dendogram Iteration 5



Scanned with CamScanner

Iteration 6

(1, 4, 9, 16, 25) (36, 49) 64 81

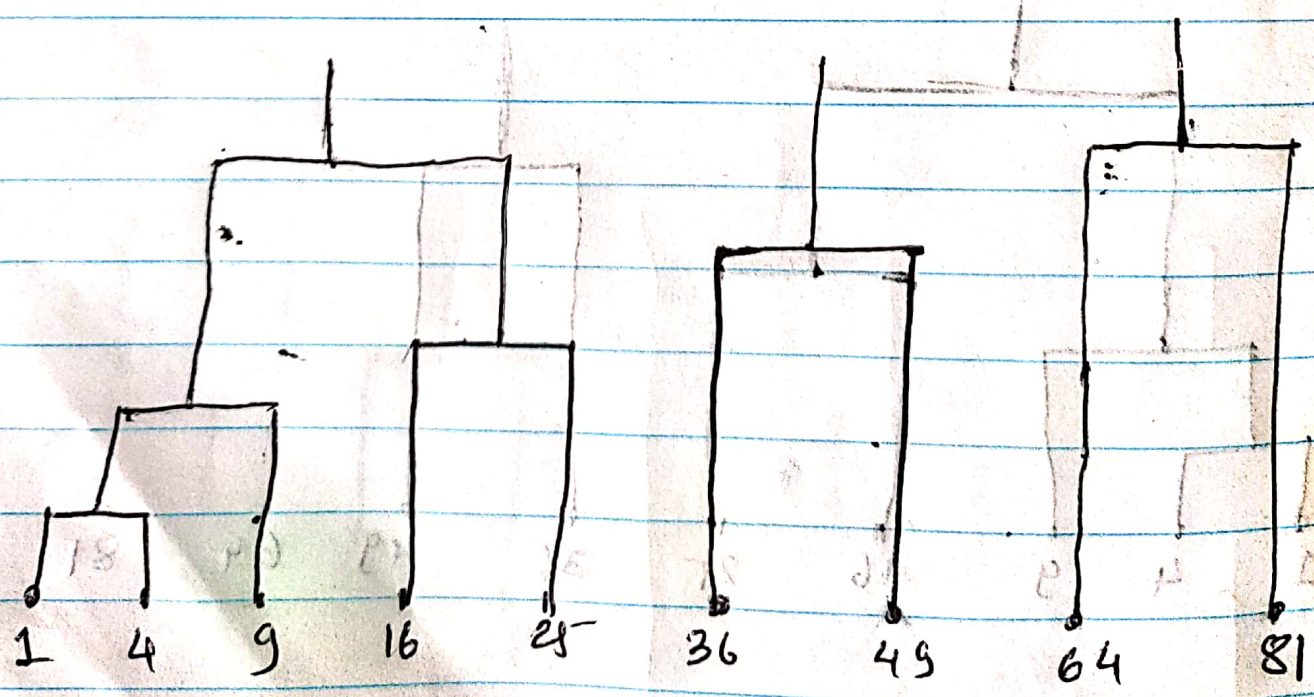
Centroids 11 42.5 64 81

Difference 31.5 21.5 17

The closest ~~cent~~ ~~clust~~ points are 64 and 81
Hence they will form a cluster.

(1, 4, 9, 16, 25) (36, 49) (64, 81)

Dendogram Iteration 6.



Scanned with CamScanner

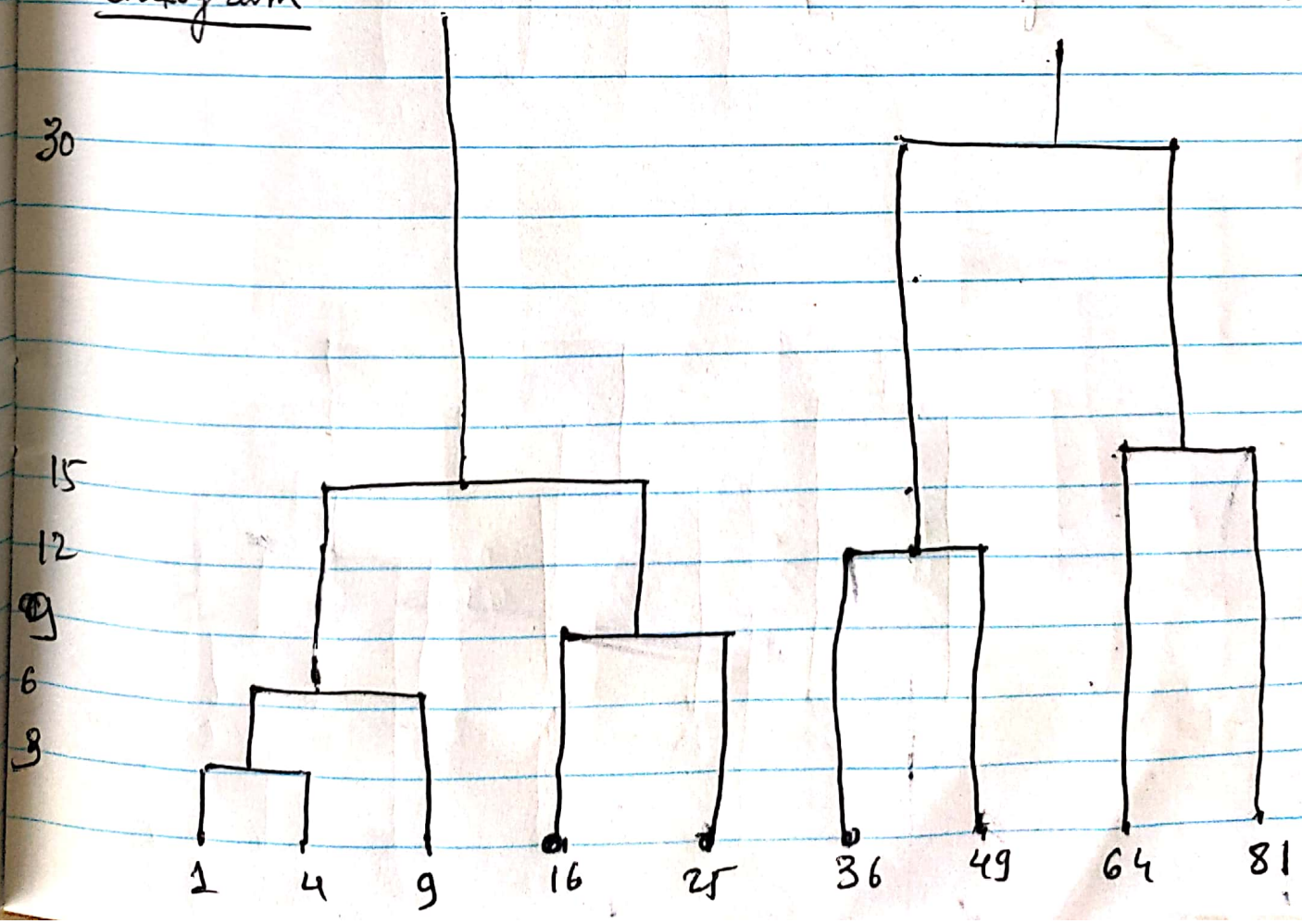
Iteration 7

	(1, 4, 9, 16, 25)	(36, 49)	(64, 81)
Centroids	11	42.5	72.5
Difference	31.5	30	

So Now the cluster is between (36, 49) and (64, 81). merge ~~the~~ them

(1, 4, 9, 16, 25) (36, 49, 64, 81)

Dendrogram



Scanned with CamScanner

Iteration 8

Since there are only two clusters remaining merge them.

(1, 4, 9, 16, 25) (36, 49, 64, 81)

Centroids 11 57.5

Difference 46.5

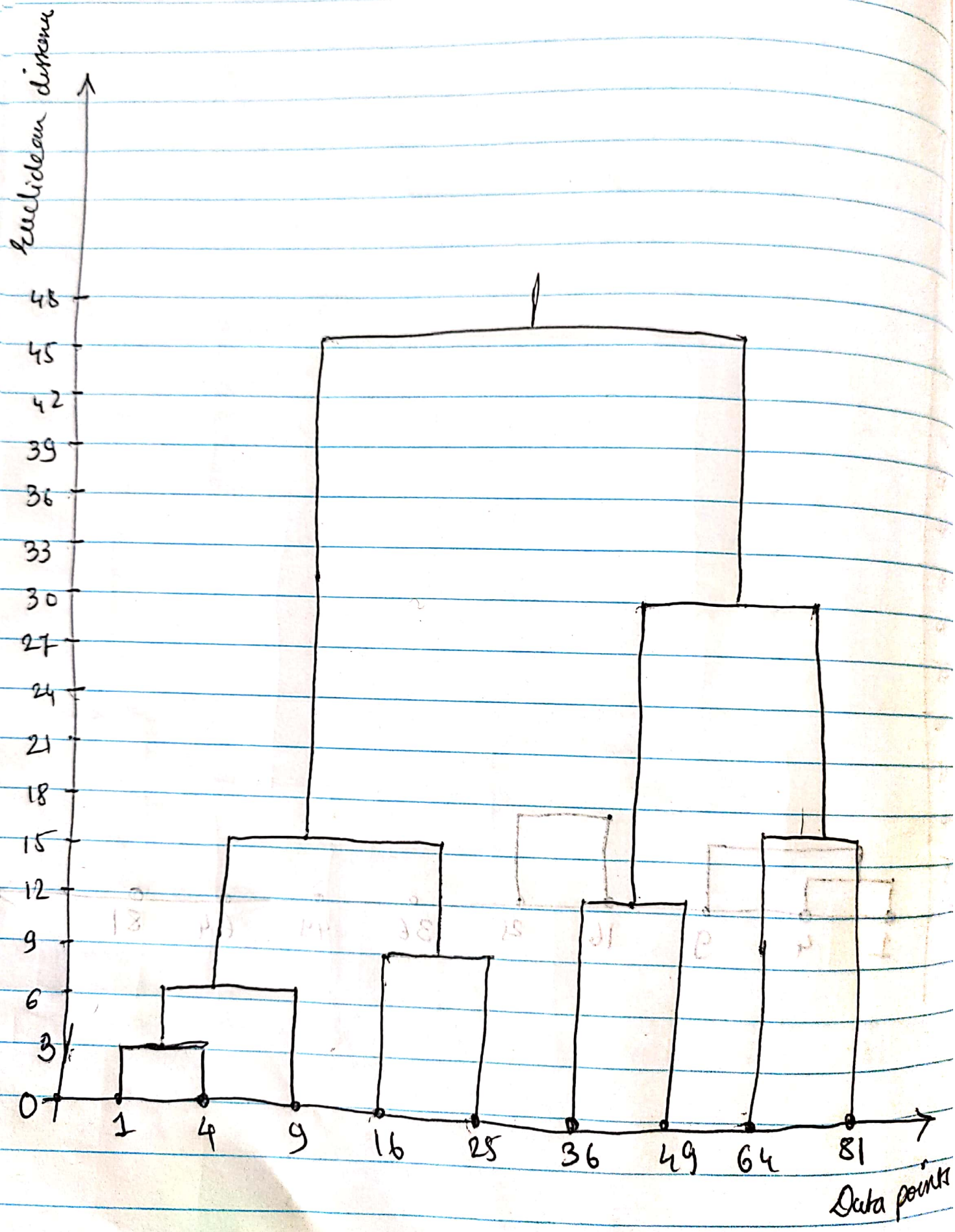
Dendrogram

(1, 4, 9, 16, 25, 36, 49, 64, 81)

This is the final step

Scanned with CamScanner

Dendogram final



Scanned with CamScanner

This is how the hierarchical clustering is done.

Scanned with CamScanner