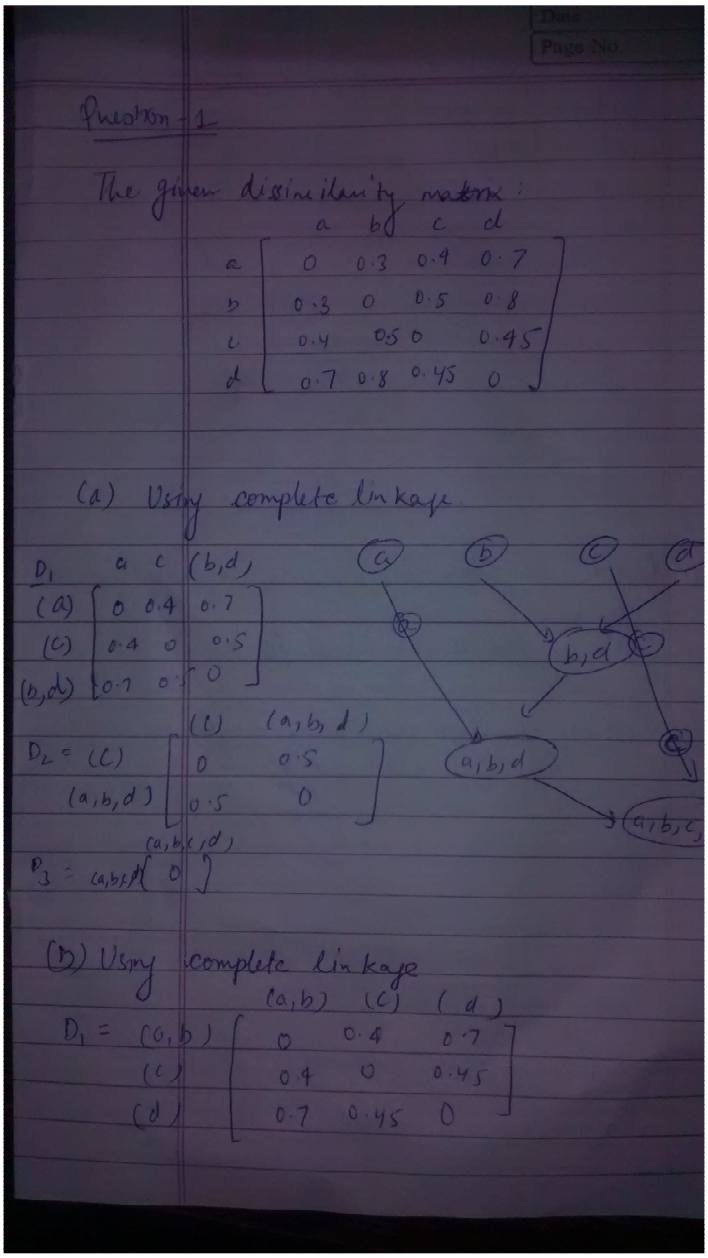
Report

Question 1



Question - 1

The given dissimilarity matrix:

	a	b	c	d
a	0	0.3	0.4	0.7
b	0.3	0	0.5	0.8
c	0.4	0.5	0	0.45
d	0.7	0.8	0.45	0

(a) Using complete linkage

D_1

	a	c	(b,d)
(a)	0	0.4	0.7
(c)	0.4	0	0.5
(b,d)	0.7	0.5	0

D_2

	(c)	(a,b,d)
(c)	0	0.5
(a,b,d)	0.5	0

D_3

	(a,b,c,d)
(a,b,c,d)	0

(b) Using complete linkage

D_1

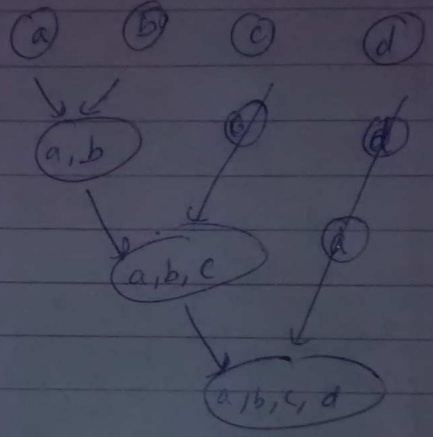
	(a,b)	(c)	(d)
(a,b)	0	0.4	0.7
(c)	0.4	0	0.45
(d)	0.7	0.45	0

Date :
Page No.

$D_2 =$

	(a,b,c)	d
(a,b,c)	0	0.45
(d)	0.45	0

$D_3 =$ [0]



(c) Results of two clusters in complete linkage,

{a, b, d} and {c}.

(d) Results of two cluster in single linkage

{a, b, c} and {d}

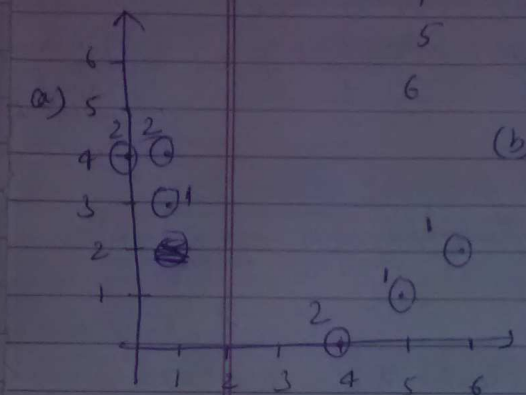
2.

Date :
Page No

(1) K = 2
M = 6
f = 2

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	6	2
5	5	1
6	4	0

(a)



(b) Random initialisation of clusters

~~(6, 2)~~ →
Cluster (1)
(6,2), (5,1), (1,3)
Cluster (2)
(0,4), (1,4), (4,0)

(c) Centroid of (1) $\left(\frac{6+5+1}{3}, \frac{2+1+3}{3}\right) = (4, 2)$

" (2) → $\left(\frac{0+1+4}{3}, \frac{4+4+0}{3}\right) = \left(\frac{5}{3}, \frac{8}{3}\right)$

(d) Closest distance for (6, 2) = (1) $\sqrt{(6-4)^2+(2-2)^2} = 4$
cluster 2 (2) $\sqrt{\left(6-\frac{5}{3}\right)^2+\left(\frac{8}{3}-2\right)^2}$

Distance for cluster (2) > Distance for cluster 1

So (6, 2) in cluster (1)
Similarly (5, 1) in cluster (1)
(1, 3) in cluster (2)
(0, 4) " (2)
(1, 4) " (2)
(4, 0) " (1)

Date :
Page No.

(d) New centroids for
Cluster(1) (6,2) (5,1) (4,0)

$$\Rightarrow \left(\frac{6+5+4}{3}, \frac{2+1+0}{3}\right) = (5,1)$$

Cluster(2) = (0,4), (1,9), (1,3)

$$= \left(\frac{0+1+1}{3}, \frac{4+9+3}{3}\right) = \left(\frac{2}{3}, \frac{11}{3}\right)$$

Now calculating the nearest centroids.

(6,2) in cluster(1) $\sqrt{\left(\frac{2}{3}-6\right)^2+\left(\frac{11}{3}-2\right)^2}$

$> \sqrt{(6-5)^2+(2-1)^2}$

Similarly.

(5,1) in cluster (1)

(4,0) " " (1)

and

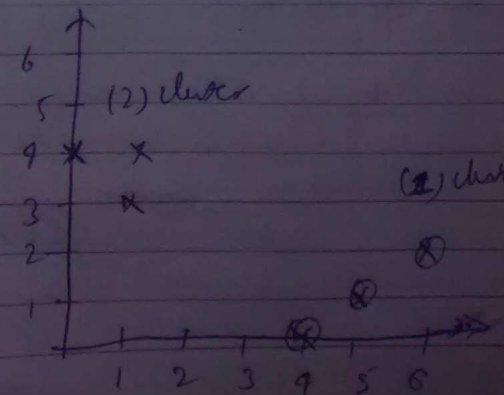
(0,4) in cluster 2

(1,9) in cluster 2

(1,3) in cluster 2.

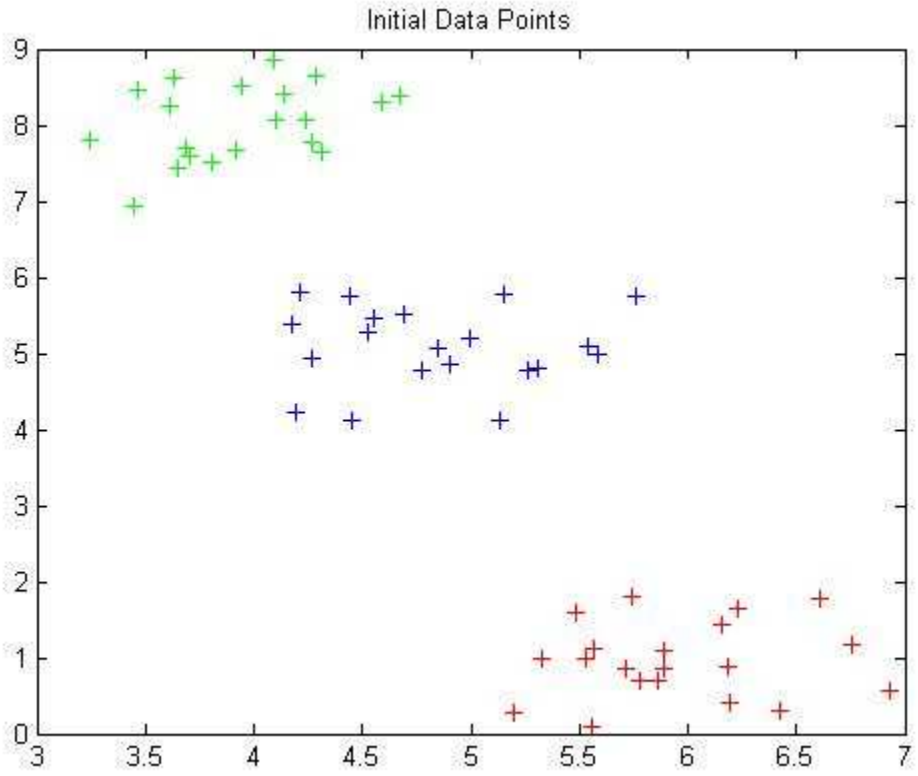
Since the clusters are not changing we can be stop here.

(f) Final clusters.

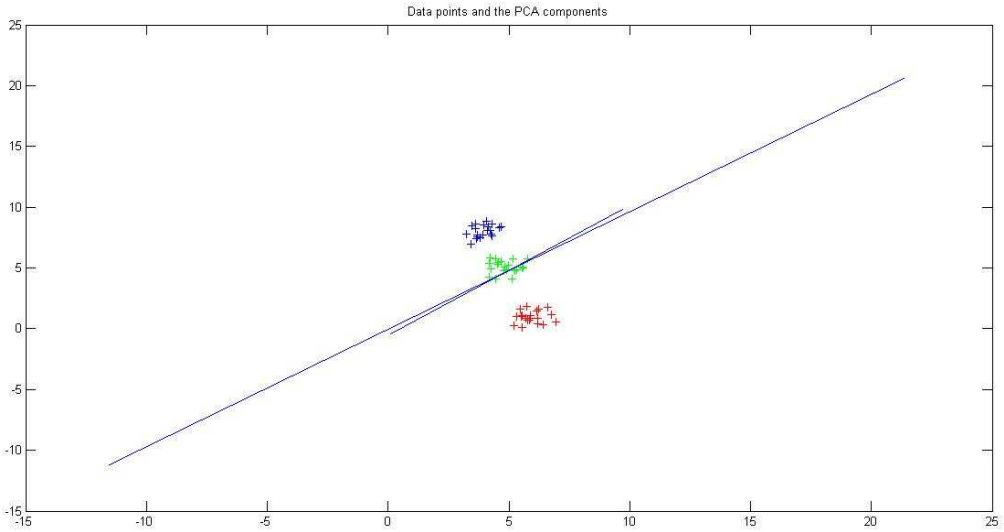


3.

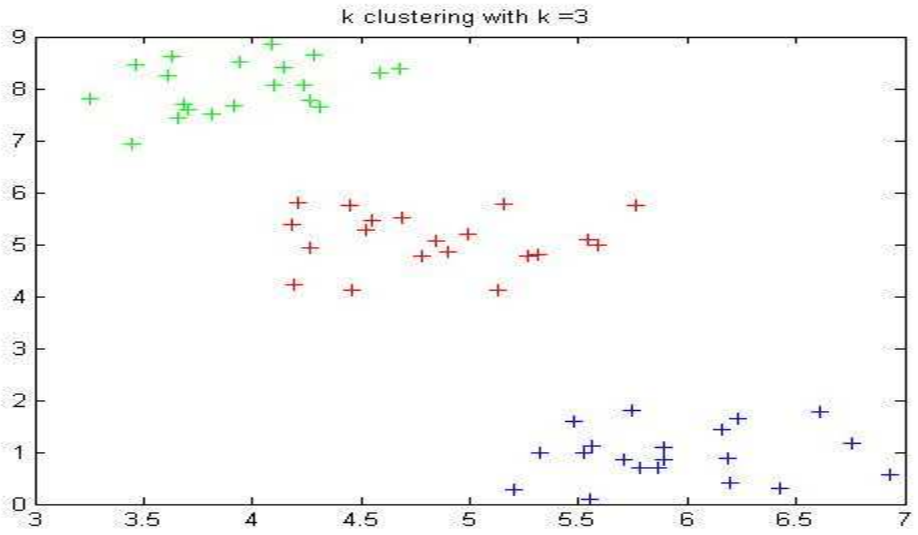
(a) The 60 observation initially with the data labels



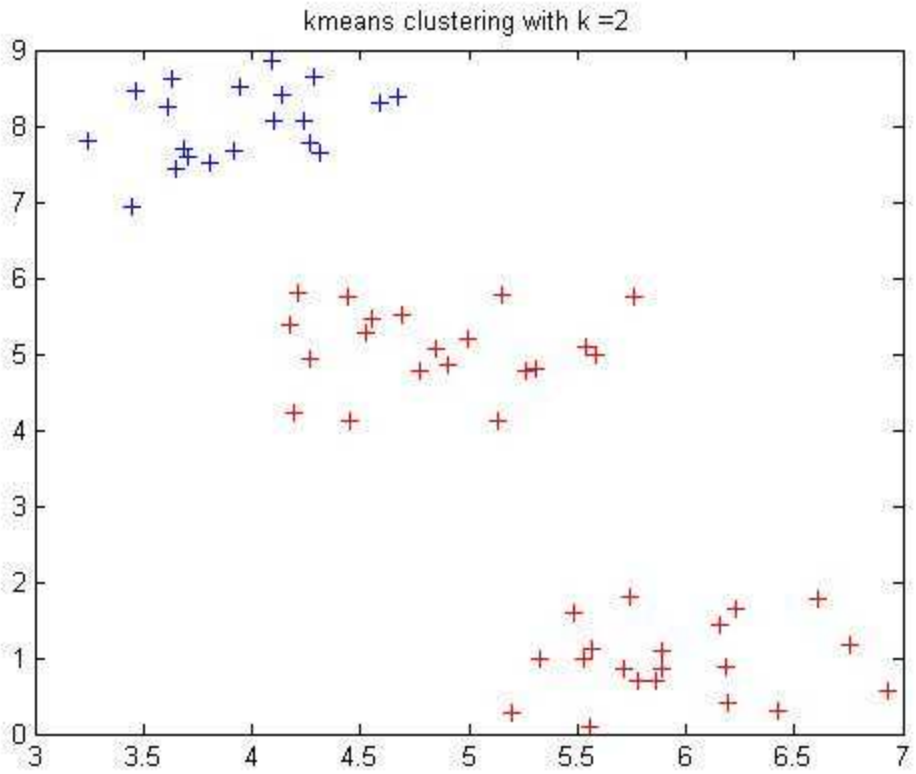
3(b) The data points with the PCA vector shown



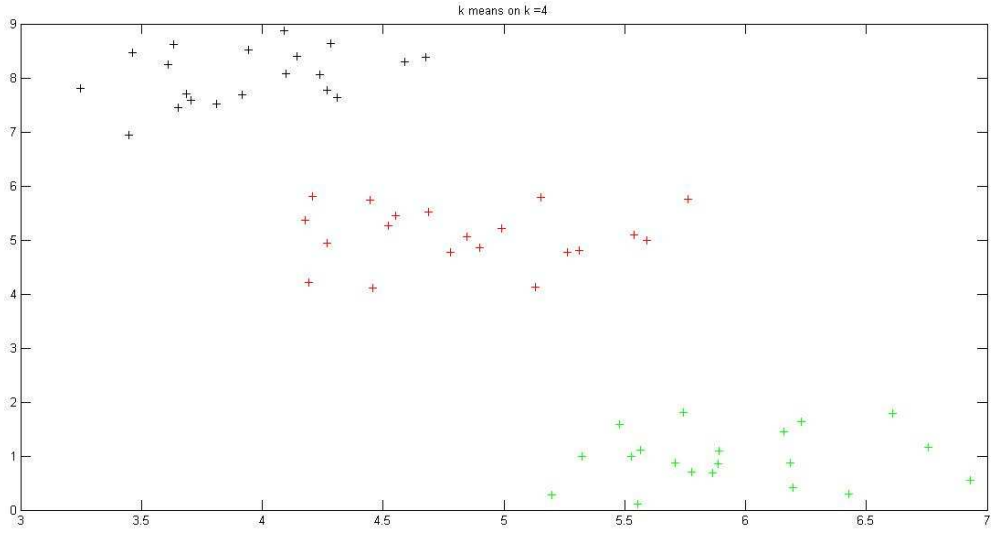
3(c) K means clustering on the data with k=3



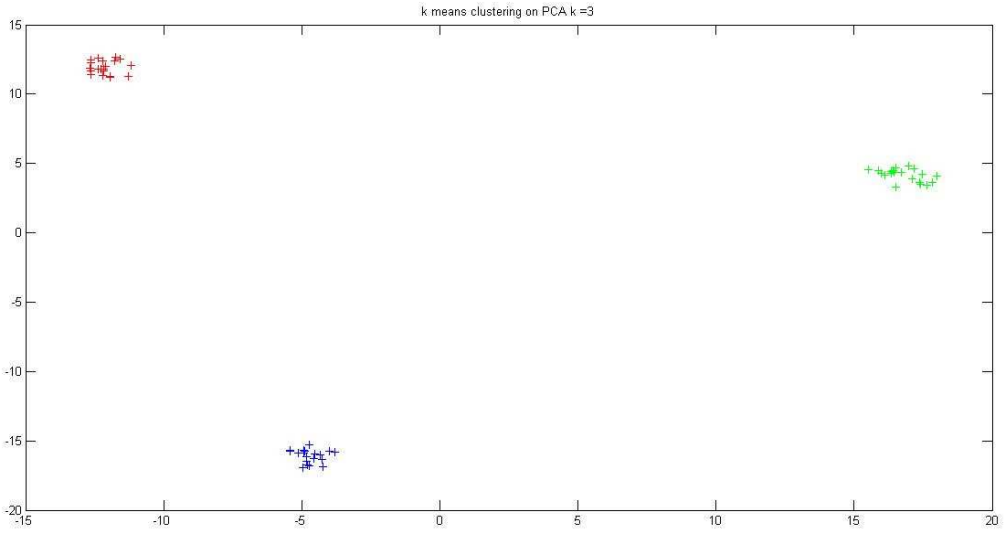
(d) k means clustering on K =2



(e) k means clustering on the data with k =4 on the original data

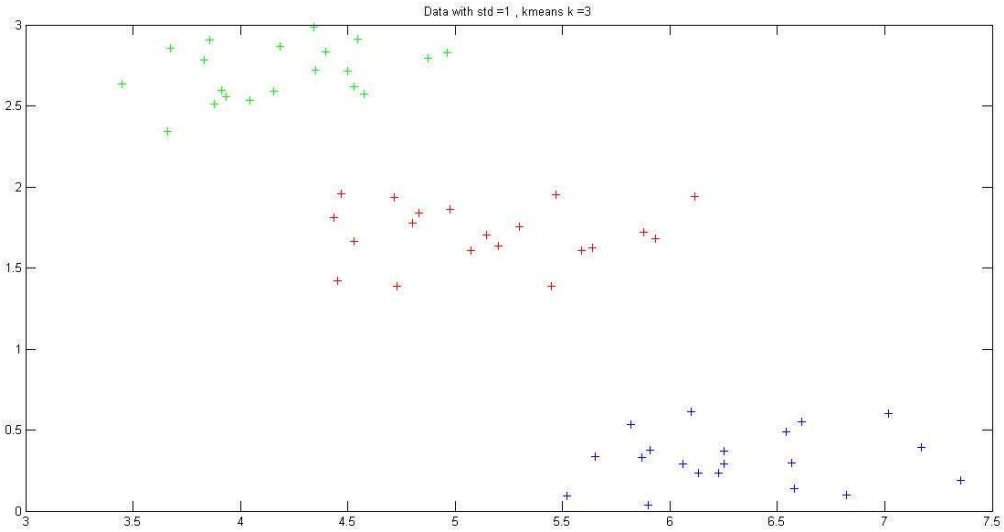


(f) k means clustering on the PCA points with k =3.

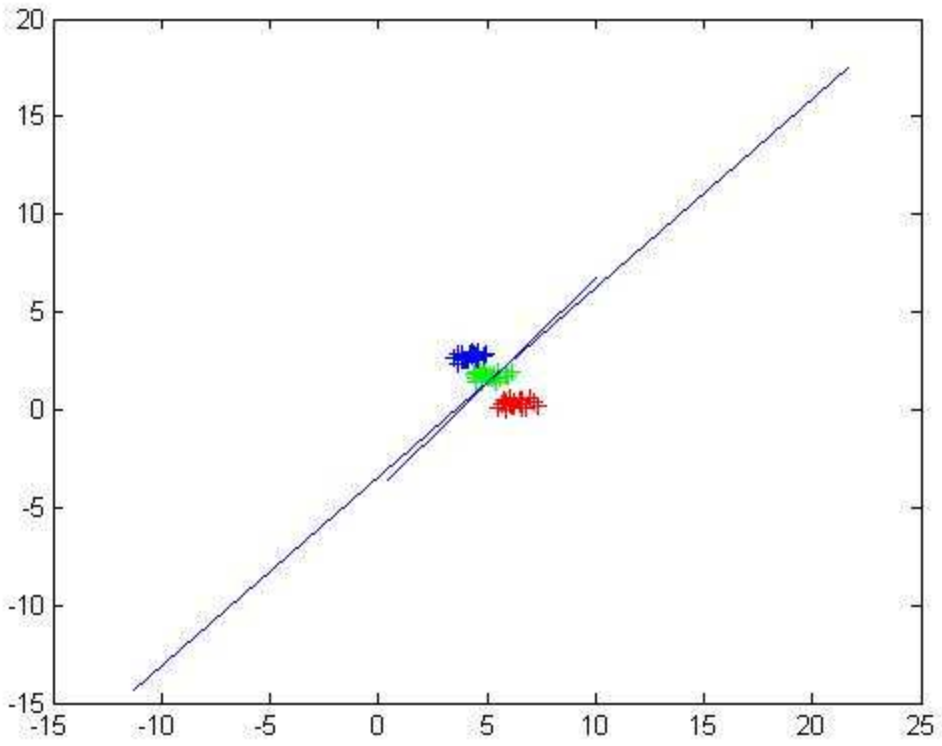


It can be seen in the PCA of the data points the points are more far separated than the original data .

(g) k means cluster on the points with std deviation = 1



With std deviation and PCA

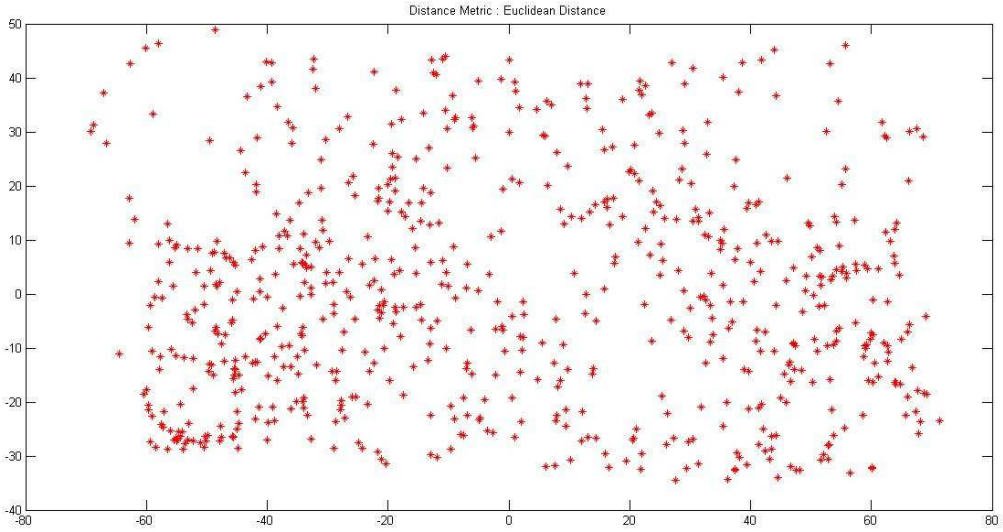


It can be seen by changing the standard deviation of the data to 1 the data points are more closer as the in the original data standard deviation was 0.5.

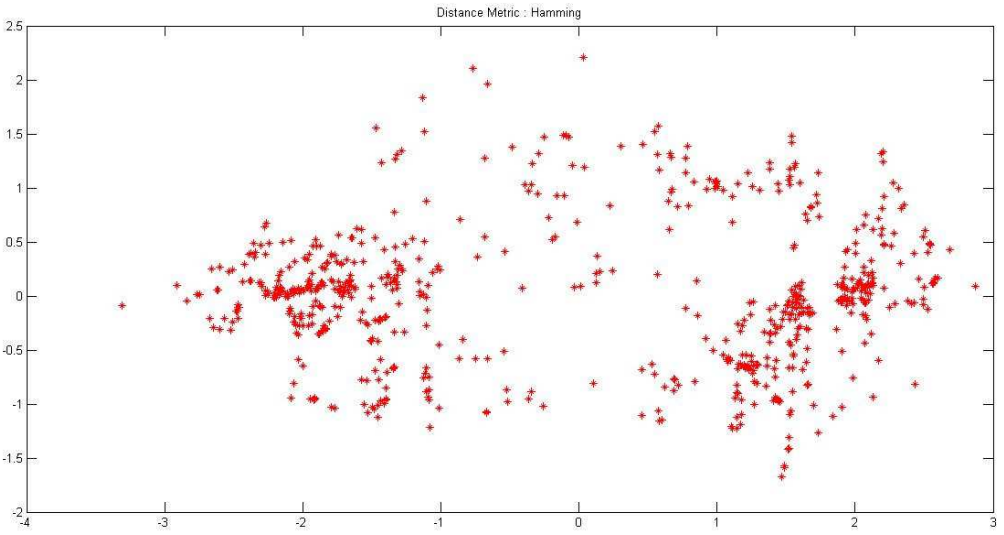
4.

(b)

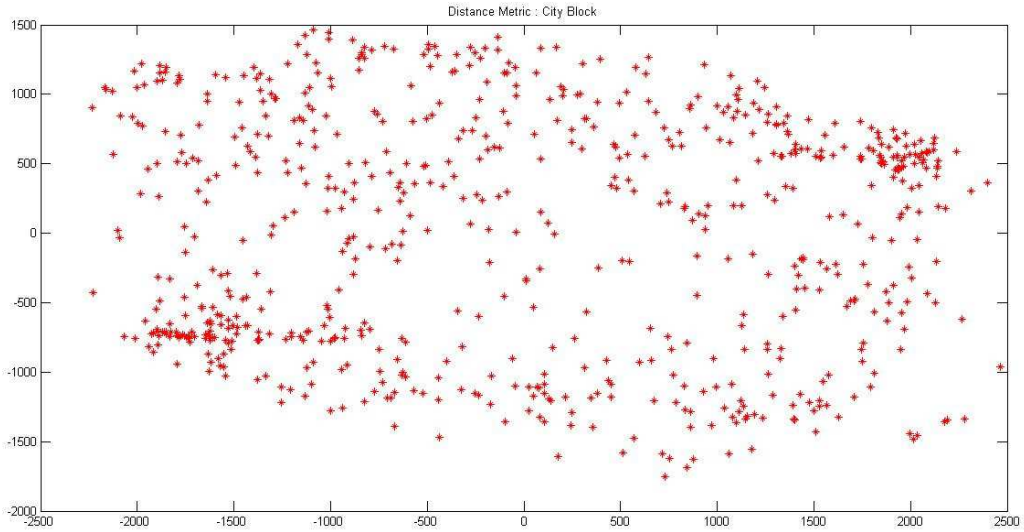
The 2D embedding with euclidean distance



(c) Embedding with distance metric as hamming



Embedding with distance metric as cityblock


Distance Metric : City Block
1500
1000
500
0
-500
-1000
-1500
-2000
-2500
-2000
-1500
-1000
-500
0
500
1000
1500
2000
2500