Name: Zhang Jingbin
Student NO:MB954741
CISC7019 Web Mining
Due: 10/04/2019

Assignment #2

Problem 1. (7.2.1) : Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged.
Solution:
Because the objects are one-dimensional set of points, we define the distance between two clusters as following equation $distance(a, b) = |a - b|$, the centroid of two clusters will be defined as following equation $centroid(C) = \frac{1}{|C|}\sum_{c \in C} c$.

Iteration 1:

centroid	1	4	9	16	25	36	49	64	81
distance	c1	c2	c3	c4	c5	c6	c7	c8	c9
c1	0	3	8	15	24	35	48	63	80
c2		0	5	12	21	32	45	60	77
c3			0	7	16	27	40	55	72
c4				0	9	20	33	48	65
c5					0	11	24	39	56
c6						0	13	28	45
c7							0	15	32
c8								0	17
c9									0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t. distance(c_i, c_j) \neq 0$
where c_1, c_2 have the smallest distance, so c_1, c_2 are merged to the same cluster. Specifically, $c_1 = \{1,\ 4\}$.
Therefore the result of iteration 1 will be like this:

$$\{\{1,\ 4\},\ \{9\},\ \{16\},\ \{25\},\ \{36\},\ \{49\},\ \{64\},\ \{81\}\}$$

Iteration 2:

centroid	2.5	9	16	25	36	49	64	81
distance	c1	c3	c4	c5	c6	c7	c8	c9
c1	0	6.5	13.5	22.5	33.5	46.5	61.5	78.5
c3		0	7	16	27	40	55	72
c4			0	9	20	33	48	65
c5				0	11	24	39	56
c6					0	13	28	45
c7						0	15	32
c8							0	17
c9								0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t.distance(c_i, c_j) \neq 0$
where c_1, c_3 have the smallest distance, so c_1, c_3 are merged to the same cluster. Specifically, $c_1 = \{1,\ 4,\ 9\}$.
Therefore the result of iteration 2 will be like this:

$$\{\{1,\ 4,\ 9\},\ \{16\},\ \{25\},\ \{36\},\ \{49\},\ \{64\},\ \{81\}\}$$

Iteration 3:

centroid	4.667	16	25	36	49	64	81
distance	c1	c4	c5	c6	c7	c8	c9
c1	0	11.33	20.33	31.33	44.33	59.33	76.33
c4		0	9	20	33	48	65
c5			0	11	24	39	56
c6				0	13	28	45
c7					0	15	32
c8						0	17
c9							0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t.distance(c_i, c_j) \neq 0$
where c_4, c_5 have the smallest distance 3, so c_4, c_5 are merged to the same cluster. Specifically, $c_4 = \{16,\ 25\}$.
Therefore the result of iteration 3 will be like this:

$$\{\{1,\ 4,\ 9\},\ \{16,\ 25\},\ \{36\},\ \{49\},\ \{64\},\ \{81\}\}$$

Iteration 4:

centroid	2.5	20.5	36	49	64	81
distance	c1	c4	c6	c7	c8	c9
c1	0	15.83	31.33	44.33	59.33	76.33
c4		0	15.5	28.5	43.5	60.5
c6			0	13	28	45
c7				0	15	32
c8					0	17
c9						0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t.distance(c_i, c_j) \neq 0$
where c_6, c_7 have the smallest distance, so c_6, c_7 are merged to the same cluster. Specifically, $c_6 = \{36,\ 49\}$.
Therefore the result of iteration 4 will be like this:

$$\{\{1,\ 4,\ 9\},\ \{16,\ 25\},\ \{36,\ 49\},\ \{64\},\ \{81\}\}$$

Iteration 6:

centroid	2.5	20.5	42.5	64	81
distance	c1	c4	c6	c8	c9
c1	0	15.83	31.33	59.33	76.33
c4		0	22	43.5	60.5
c6			0	21.5	38.5
c8				0	17
c9					0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t.distance(c_i, c_j) \neq 0$
where c_1, c_4 have the smallest distance, so c_1, c_4 are merged to the same cluster. Specifically, $c_1 = \{1,\ 4,\ 9,\ 16,\ 25\}$.
Therefore the result of iteration 6 will be like this:

$$\{\{1,\ 4,\ 9,\ 16,\ 25\},\ \{36,\ 49\},\ \{64\},\ \{81\}\}$$

Iteration 7:

centroid	11	42.5	64	81
distance	c1	c6	c8	c9
c1	0	31.5	53	70
c6		0	21.5	38.5
c8			0	17
c9				0

Found the minimal value in distance table.
$(c_i, c_j) = min(distance(c_i, c_j)), s.t.distance(c_i, c_j) \neq 0$

where c_8, c_9 have the smallest distance, so c_8, c_9 are merged to the same cluster. Specifically, $c_8 = \{64,\ 81\}$.

Therefore the result of iteration 7 will be like this:

$$\{\{1,\ 4,\ 9,\ 16,\ 25\},\ \{36,\ 49\},\ \{64,\ 81\}\}$$

Iteration 8:

centroid	11	42.5	72.5
distance	c1	c6	c8
c1	0	31.5	61.5
c6		0	30
c8			0

Found the minimal value in distance table.

$(c_i, c_j) = min(distance(c_i, c_j)), s.t. distance(c_i, c_j) \neq 0$

where c_6, c_8 have the smallest distance, so c_6, c_8 are merged to the same cluster. Specifically, $c_6 = \{36,\ 49,\ 64,\ 81\}$.

Therefore the result of iteration 8 will be like this:

$$\{\{1,\ 4,\ 9,\ 16,\ 25\},\ \{36,\ 49,\ 64,\ 81\}\}$$

Iteration 9:

centroid	11	57.5
distance	c1	c6
c1	0	46.5
c6	46.5	0

Found the minimal value in distance table.

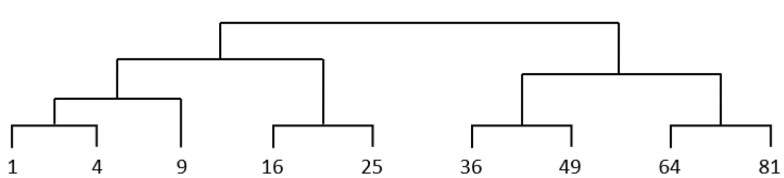
$(c_i, c_j) = min(distance(c_i, c_j)), s.t. distance(c_i, c_j) \neq 0$

where c_1, c_6 have the smallest distance, so c_1, c_6 are merged to the same cluster. Specifically, $c_1 = \{1,\ 4,\ 9,\ 16,\ 25,\ 36,\ 49,\ 64,\ 81\}$.

Therefore the result of iteration 9 will be like this:

$$\{\{1,\ 4,\ 9,\ 16,\ 25,\ 36,\ 49,\ 64,\ 81\}\}$$

After 9 iterations, we can get the result of hierarchical clustering of those items:



Problem 2. (7.2.2) : How would the clustering of Example 7.2 change if we used for the distance between two clusters:

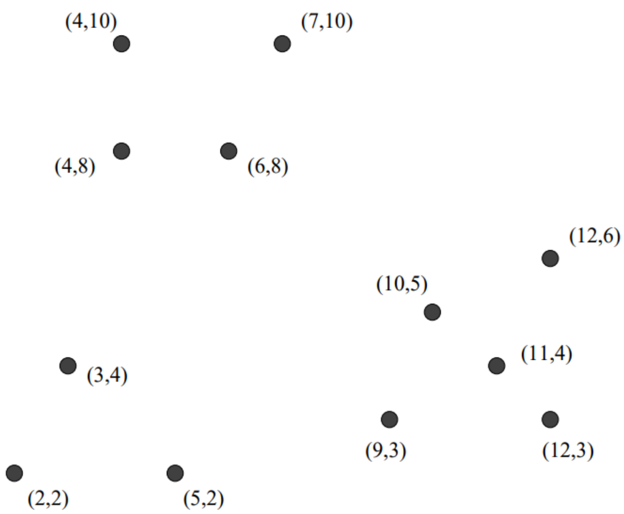


Figure 7.2: Twelve points to be clustered hierarchically

(a) The minimum of the distances between any two points, one from each cluster.

(b) The average of the distances between pairs of points, one from each of the two clusters.

Solution:

We use P0, P1,..., P11 to represent the data points

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
X	2	5	9	12	3	11	10	12	4	6	4	7
Y	2	2	3	3	4	4	5	6	8	8	10	10

Calculate the distance matrix:

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
P0	0	3	7.07	10.04	2.23	9.21	8.54	10.77	6.32	7.21	8.24	9.43
P1		0	4.12	7.07	2.82	6.32	5.83	8.06	6.08	6.08	8.06	8.24
P2			0	3	6.08	2.23	2.23	4.24	7.07	5.83	8.60	7.28
P3				0	9.05	1.41	2.82	3	9.43	7.81	10.63	8.60
P4					0	8	7.07	9.21	4.12	5	6.08	7.21
P5						0	1.41	2.23	8.06	6.40	9.21	7.21
P6							0	2.23	6.70	5	7.81	5.83
P7								0	8.24	6.32	8.94	6.40
P8									0	2	2	3.60
P9										0	2.82	2.23
P10											0	3
P11												0

(a) The minimum of the distances between any two points, one from each cluster.

Iteration 1:

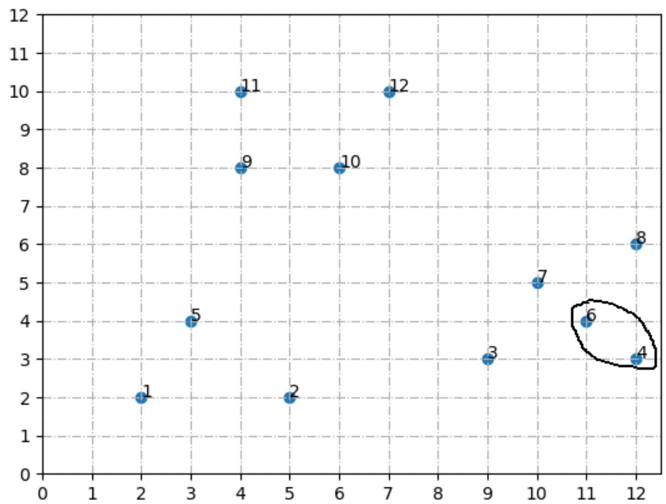
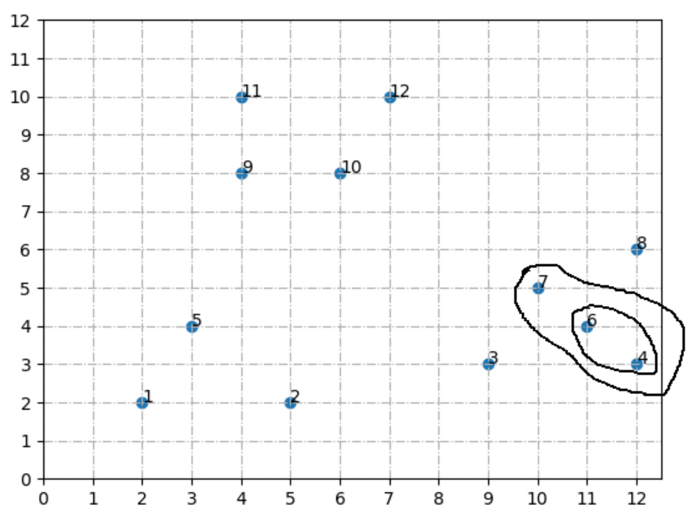
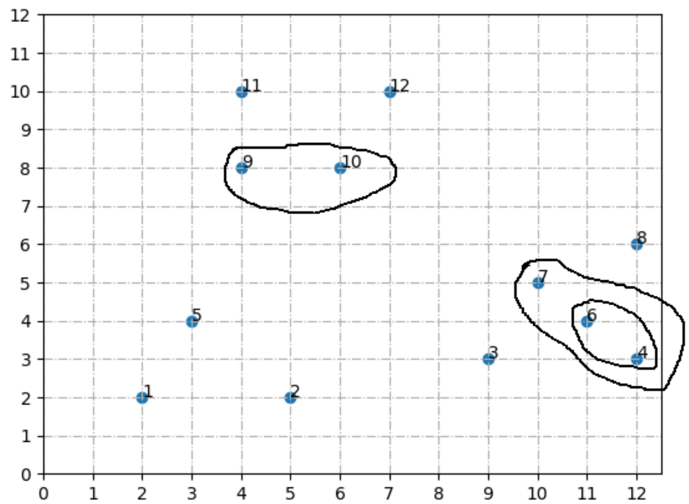


Figure 1: Latency-throughput for read workload

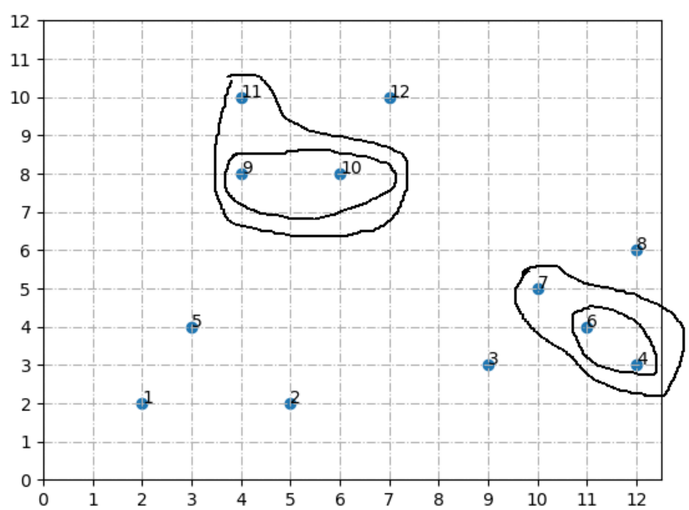
Iteration 2:



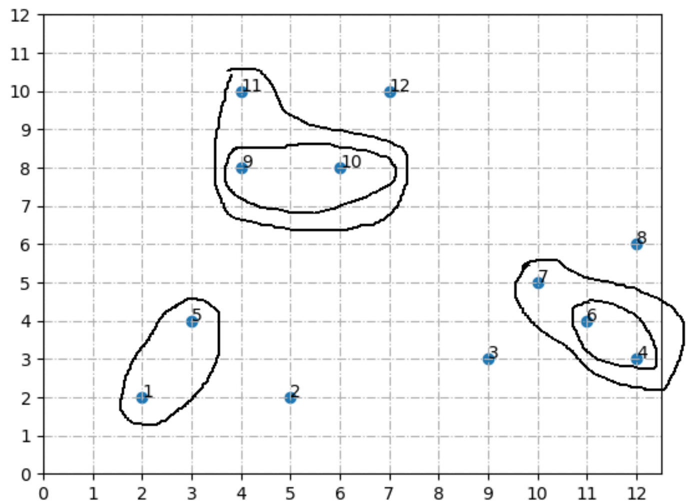
Iteration 3:



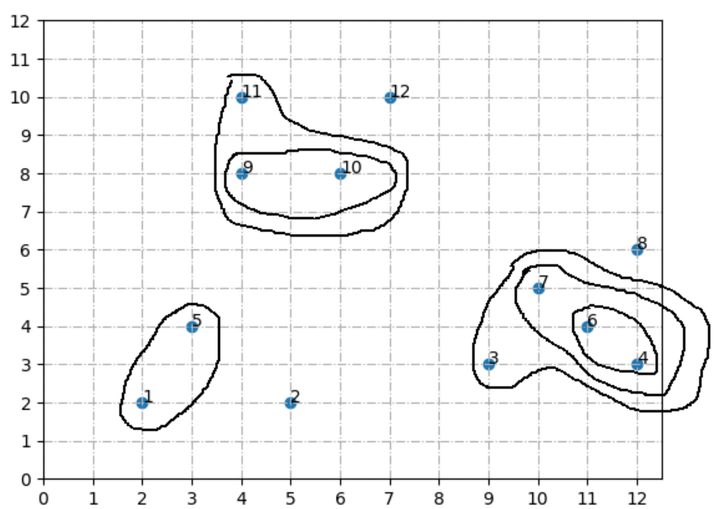
Iteration 4:



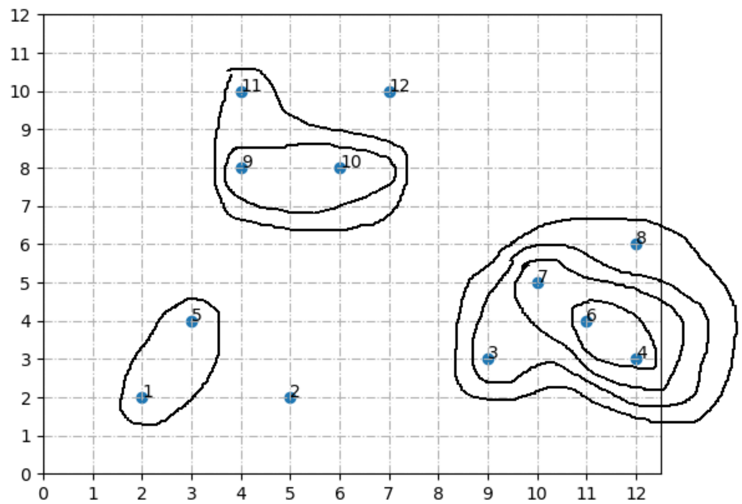
Iteration 5:



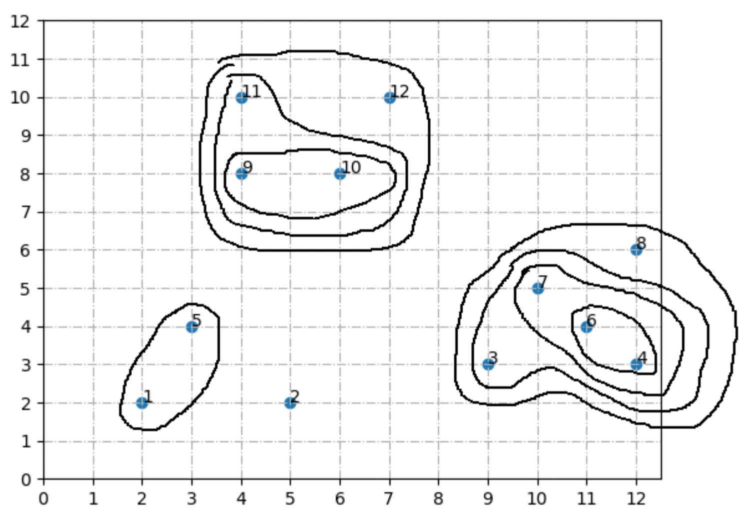
Iteration 6:



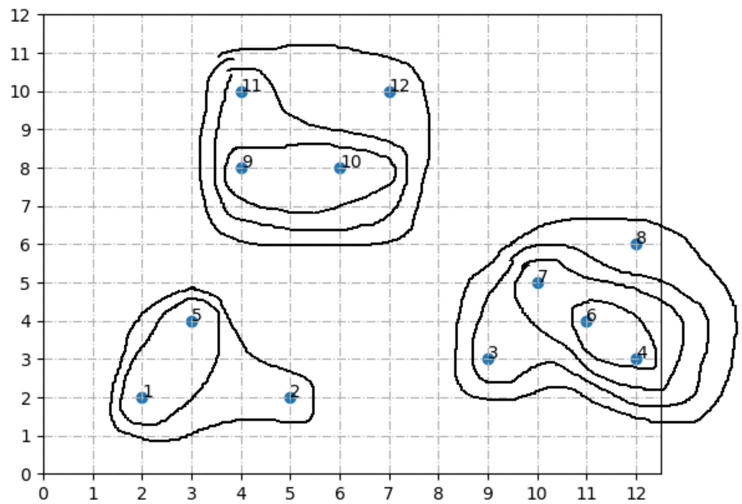
Iteration 7:



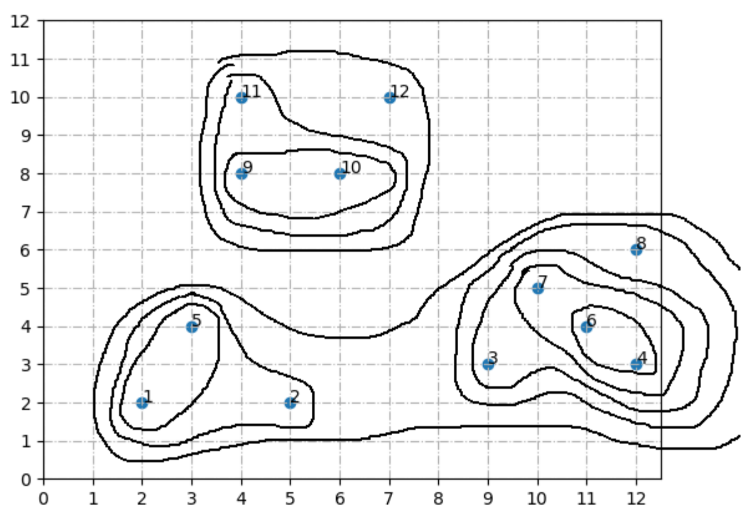
Iteration 8:



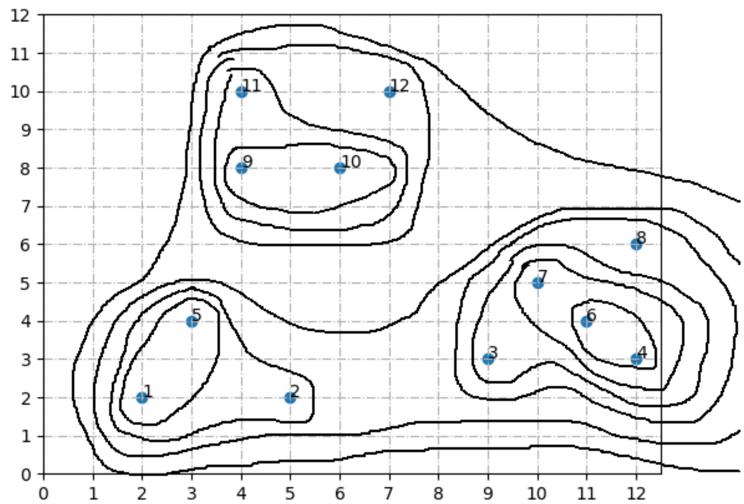
Iteration 9:



Iteration 10:

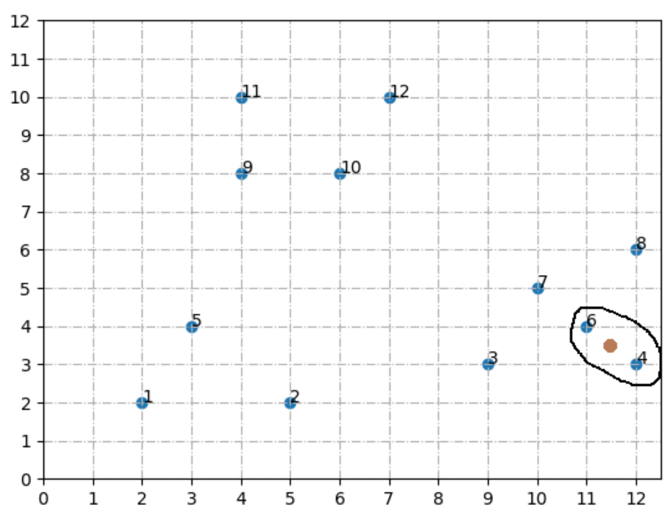


Iteration 11:

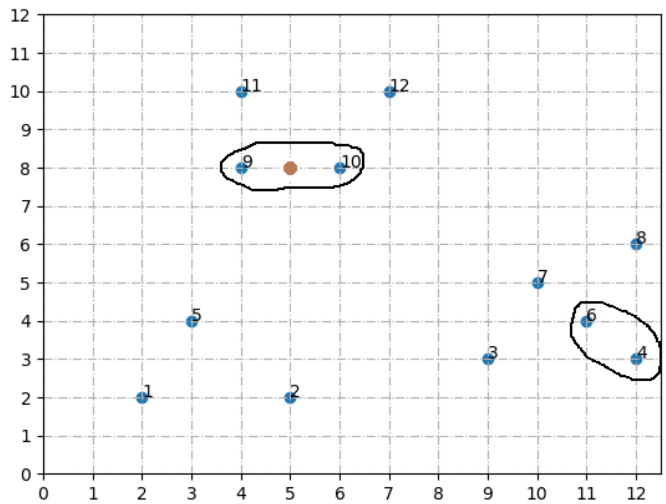


(b) The average of the distances between pairs of points, one from each of the two clusters.

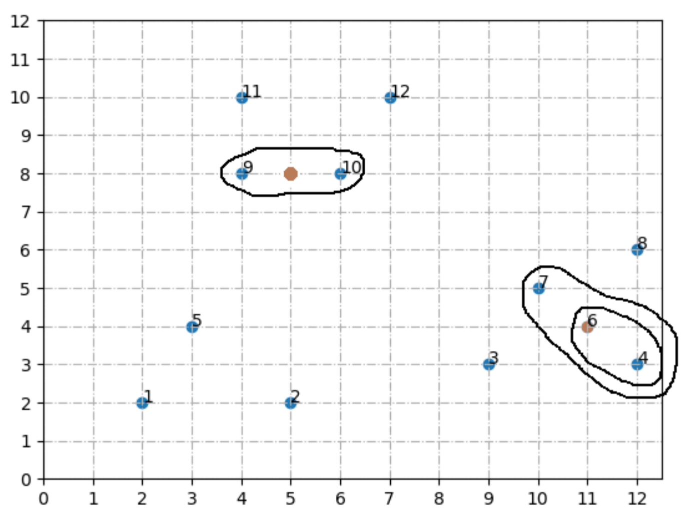
Iteration 1:



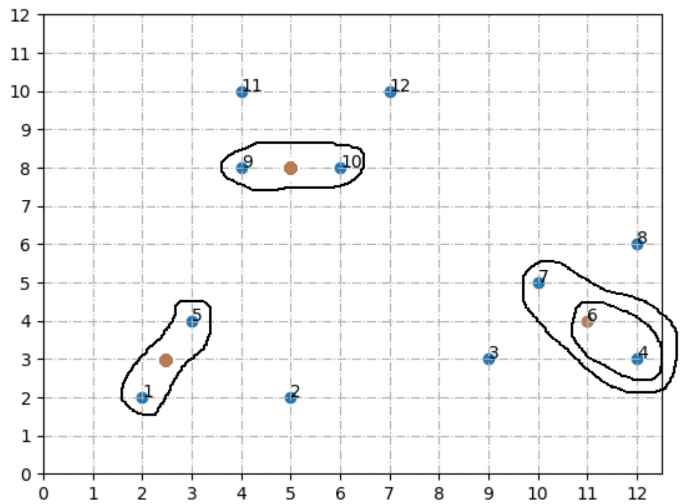
Iteration 2:



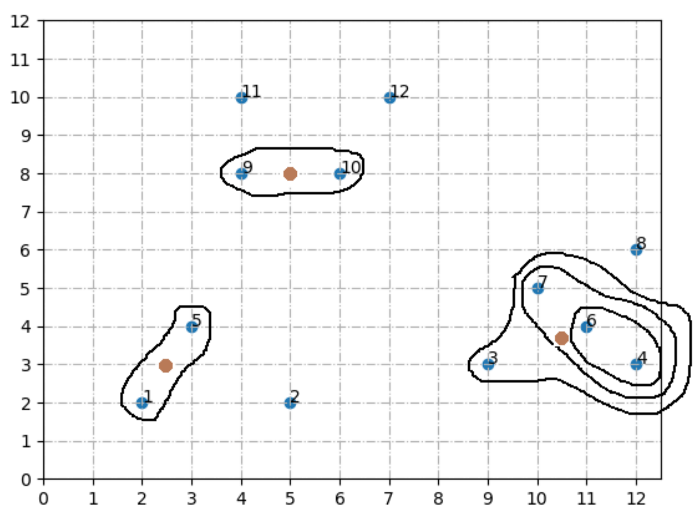
Iteration 3:



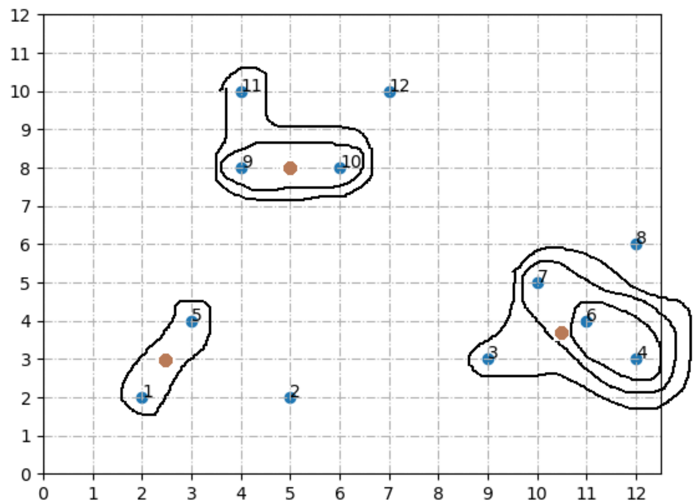
Iteration 4:



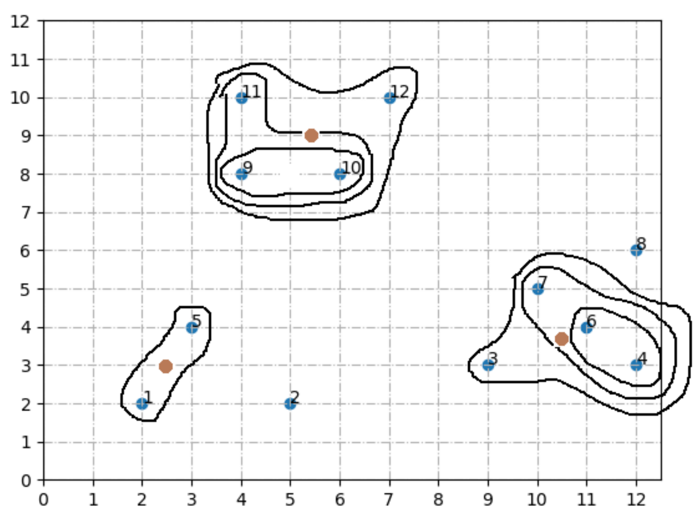
Iteration 5:



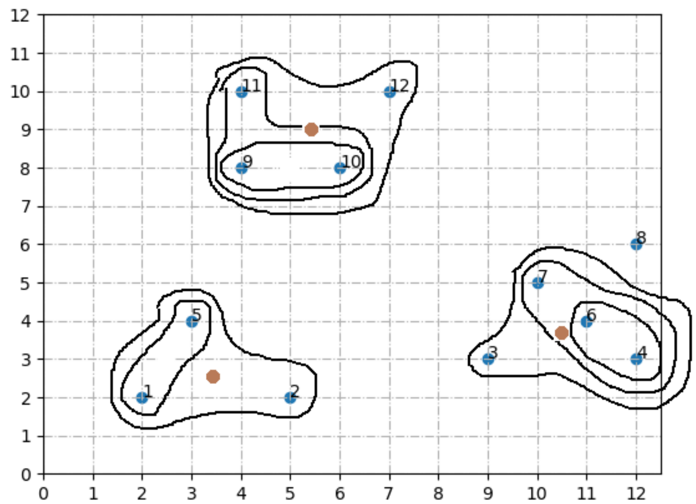
Iteration 6:



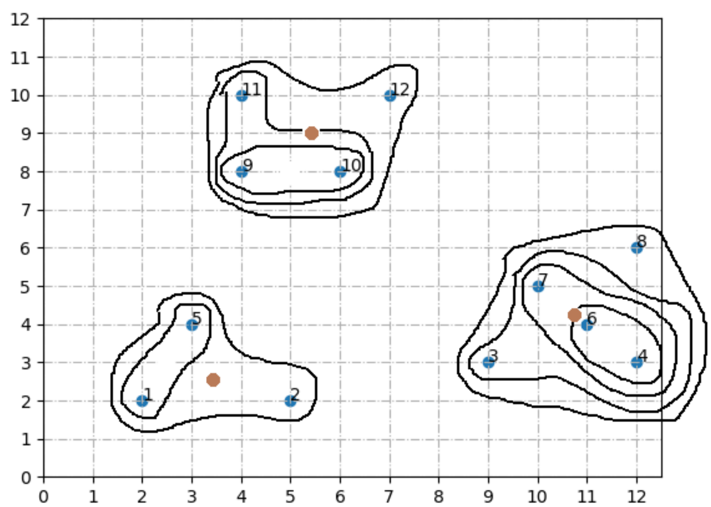
Iteration 7:



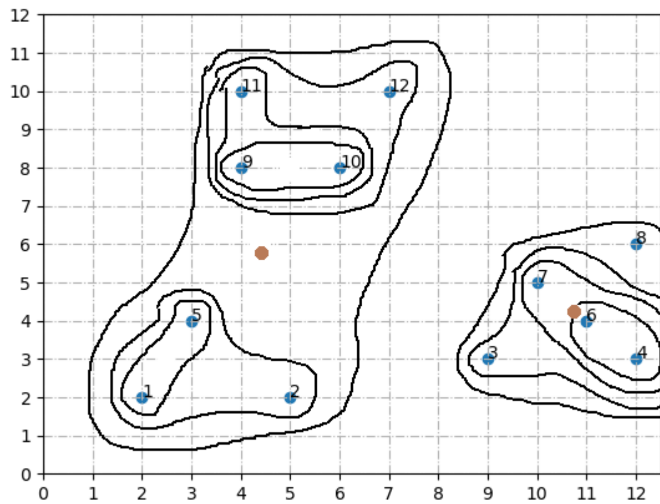
Iteration 8:



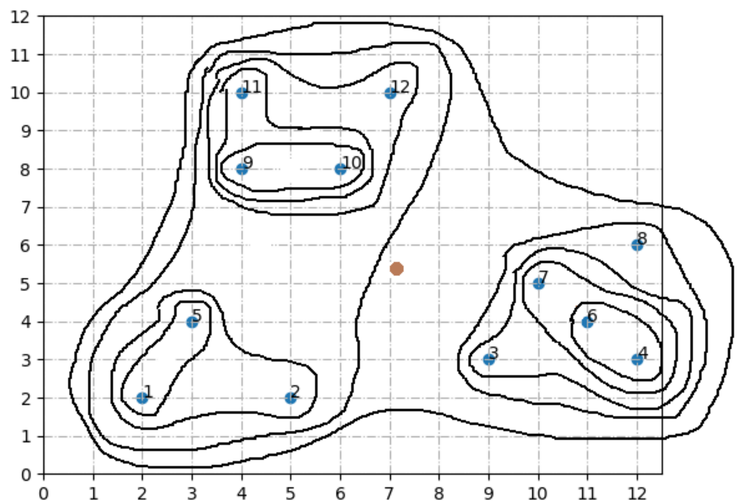
Iteration 9:



Iteration 10:



Iteration 11:



Problem 3. (7.3.1) : For the points of Fig. 7.8, if we select three starting points using the method of Section 7.3.2, and the first point we choose is (3,4), which other points are selected.

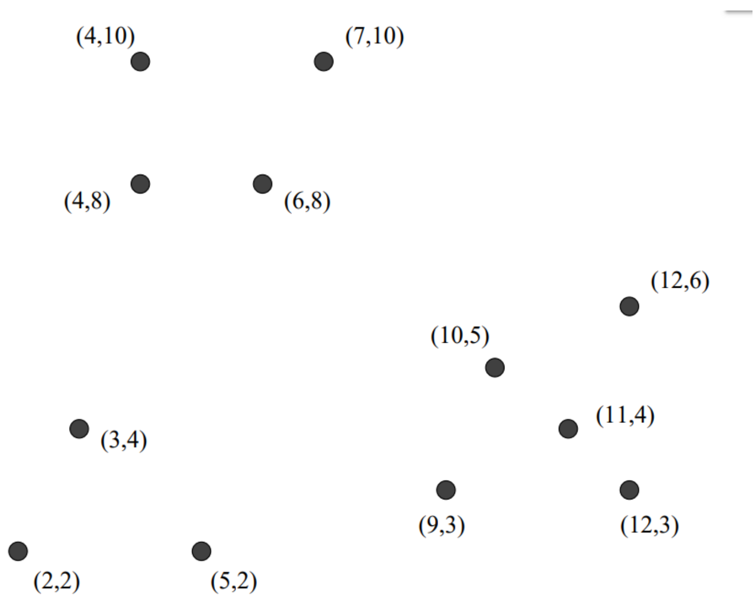


Figure 7.8: Repeat of Fig. 7.2

Solution:

First point is $(3,\ 4)$

The second point should be the farthest point from the first point. After calculating the distance from the point $(3,\ 4)$, the point $(12,\ 6)$ farthest from the point $(3,\ 4)$ can be obtained with the distance is 9.220.

The last point should be the farthest point from the first two points. After calculating the distance from the point $(3,\ 4)$ and point $(12,\ 6)$, the point $(7,\ 10)$ farthest from both the point $(3,\ 4)$ and point $(12,\ 6)$ can be obtained with the distance is 7.211 and 6.403.

Therefore the final starting points are $(3,\ 4)$, $(12,\ 6)$ and $(7,\ 10)$.