

# DATA MINING

## Exercises Clustering

---

Riccardo Guidotti



# K-Means

---

Ex 1

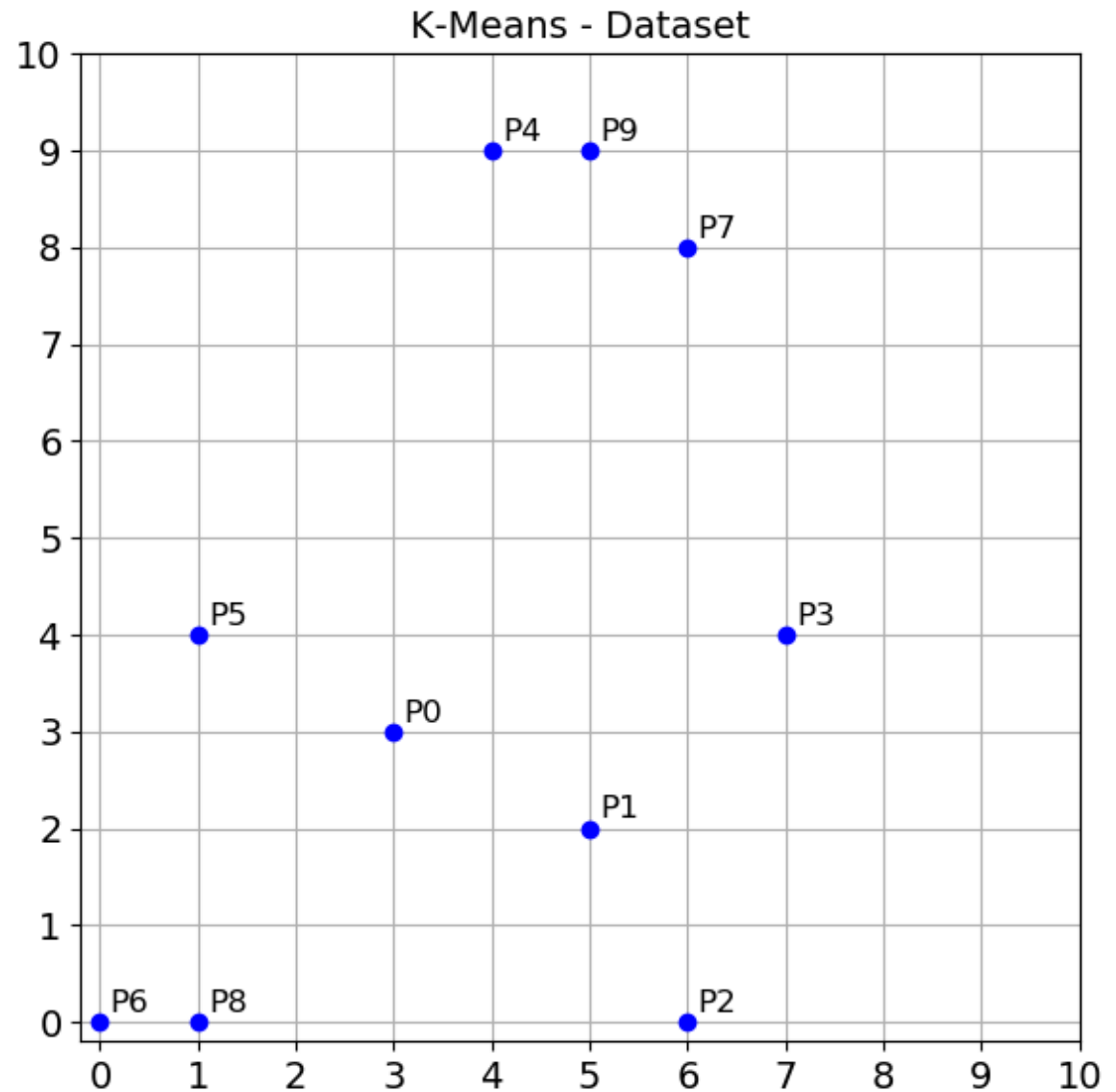
---

# K-means simulation

Initial centroids:

$$C1 = P1 = (5, 2)$$

$$C2 = P5 = (1, 4)$$



# Solution: Identify the Bisecting lines dividing the plane between pairs of centroids

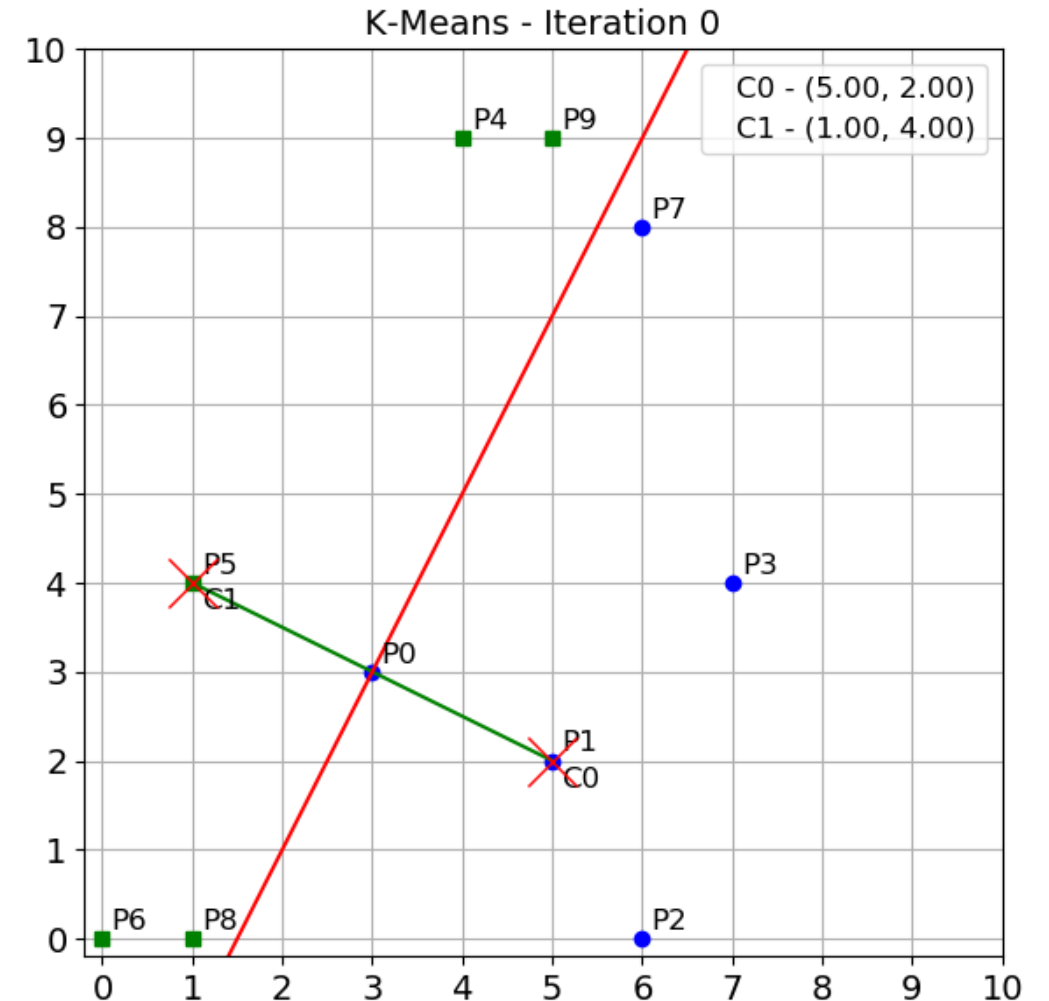
**Cluster 1:** P0,P1,P2,P3,P7

**Cluster 2:** P5,P4,P6,P8,P9

**Centrod1:**

**C1=** (5.40, 3.40)

**C2=** (2.20, 4.40)



# Solution: Identify the Bisecting lines dividing the plane between pairs of centroids

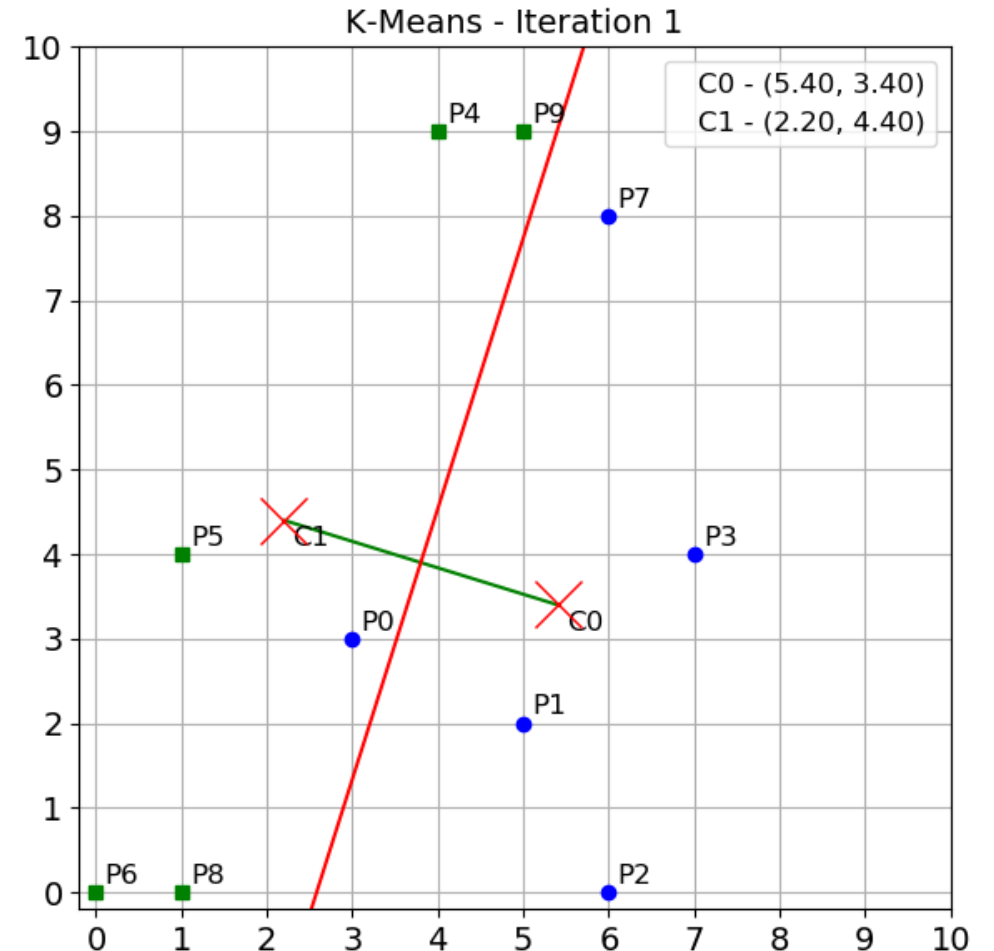
**Cluster 1:** P1,P2,P3,P7

**Cluster 2:** P5,P4,P6,P8,P9,P0

**Centrod1:**

**C1=** (6.00, 3.50)

**C2=** (2.33, 4.17)

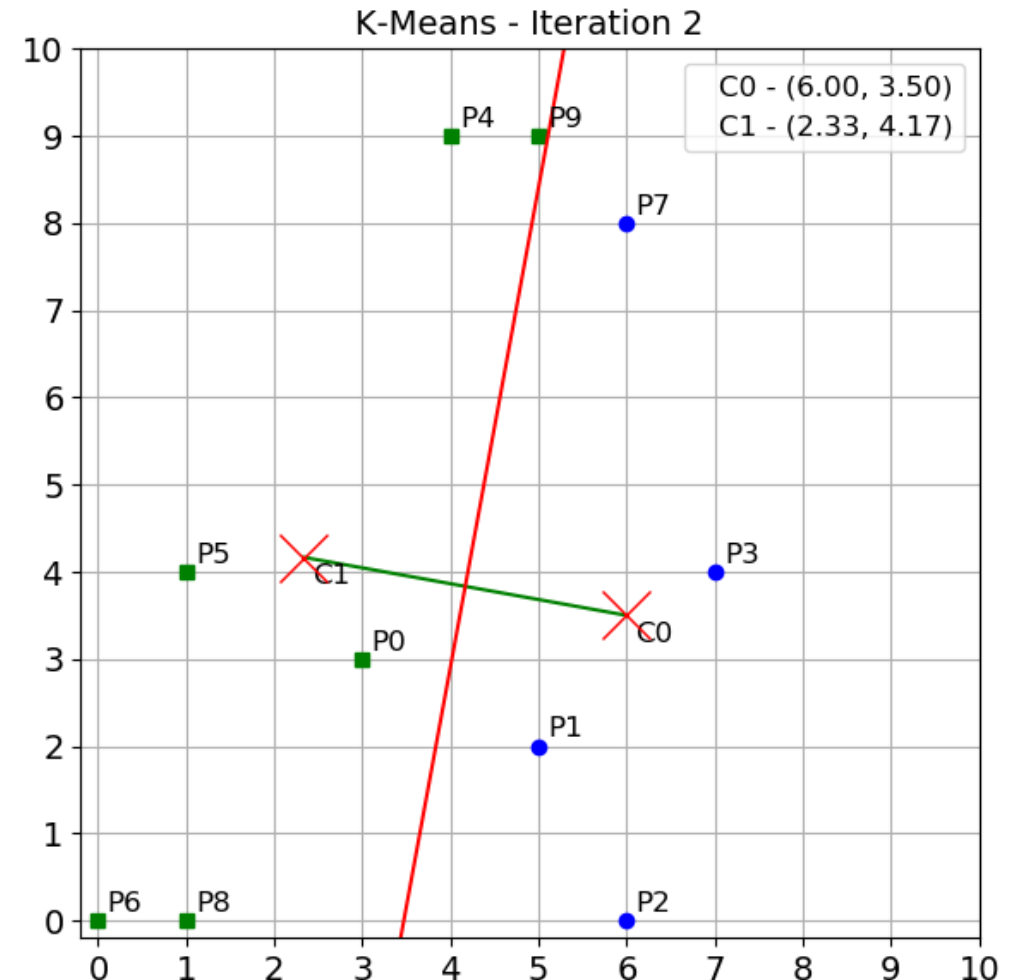


# Solution: Identify the Bisecting lines dividing the plane between pairs of centroids

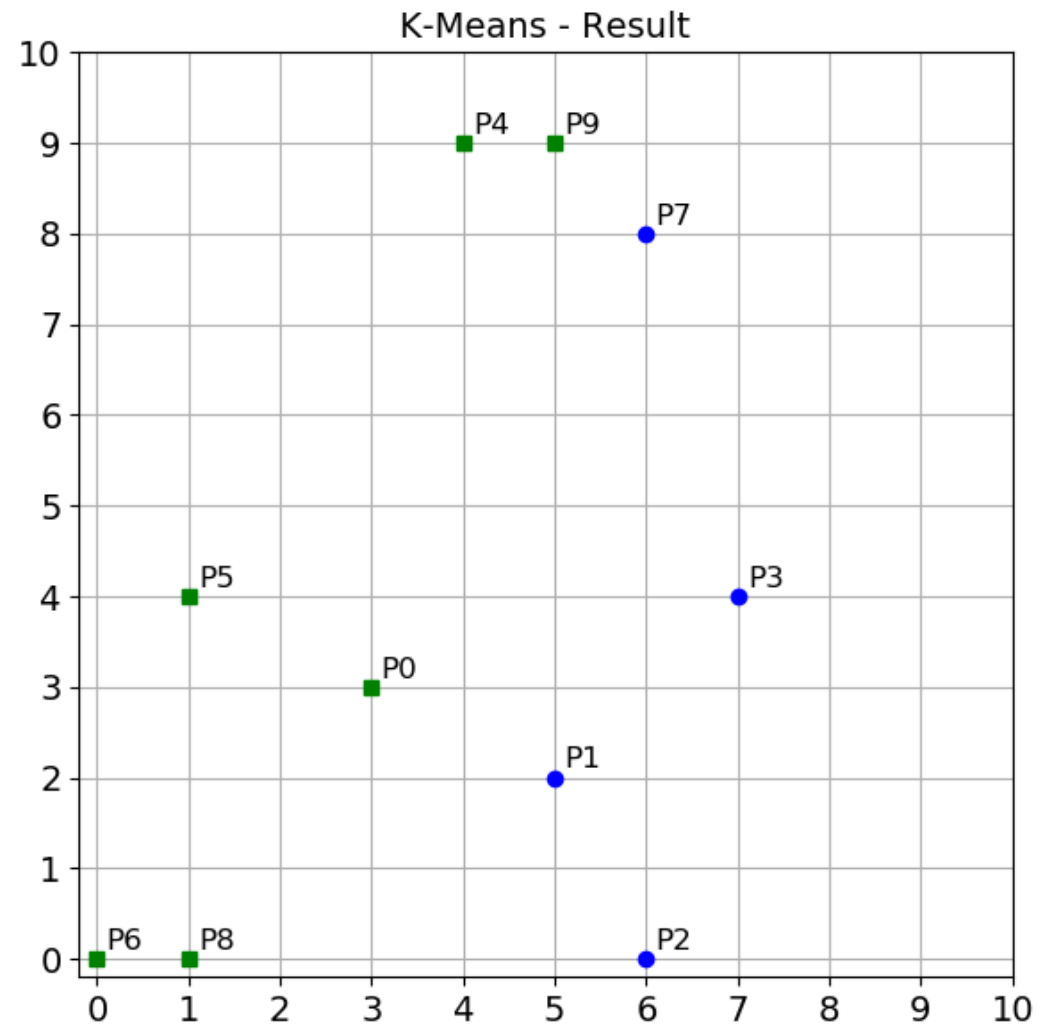
**Cluster 1:** P1,P2,P3,P7

**Cluster 2:** P5,P4,P6,P8,P9,P0

**The cluster composition does not change,  
so K-means stops**



# K-means result





Ex 2

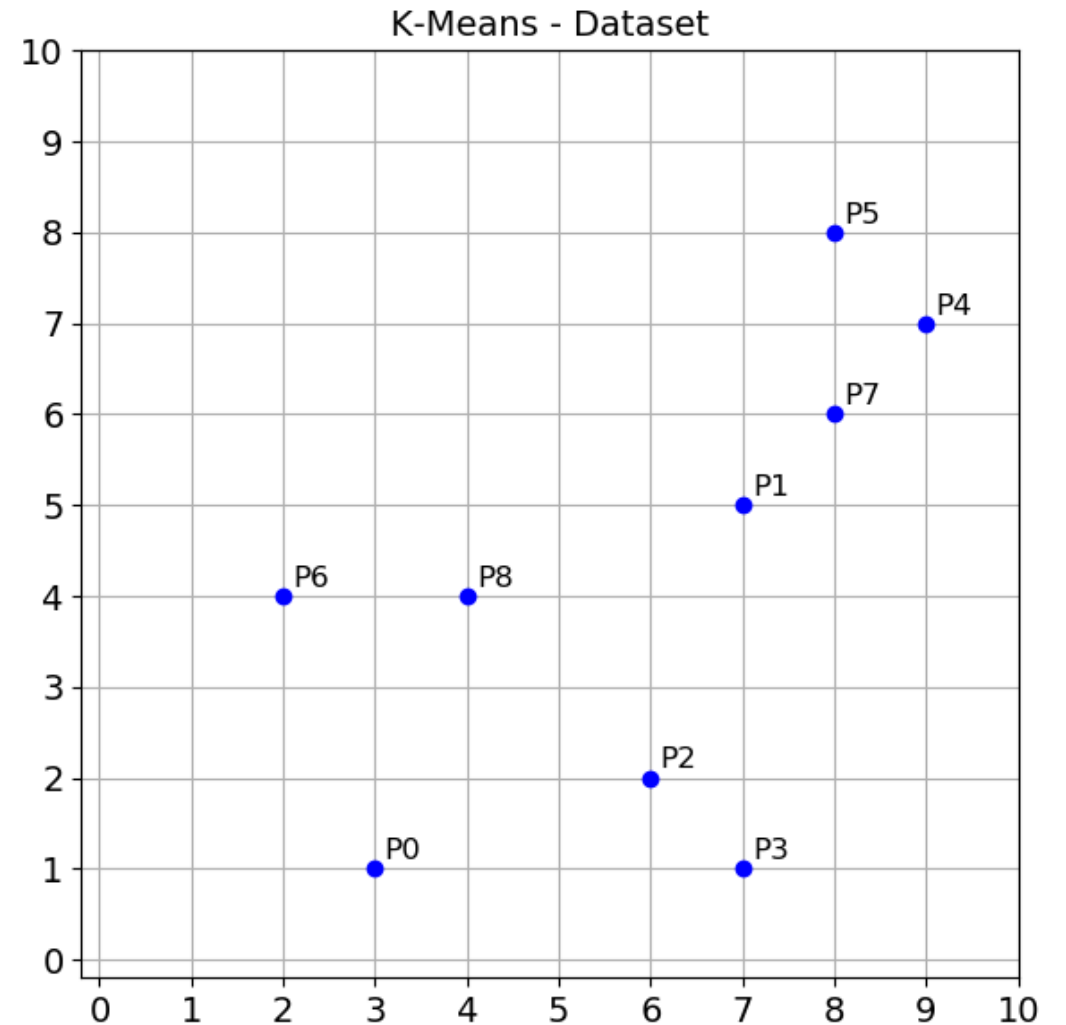
---

# K-means simulation

Initial centroids:

$$C1 = P2 = (6, 2)$$

$$C2 = P1 = (7, 5)$$



# Solution: Identify the Bisecting lines dividing the plane between pairs of centroids

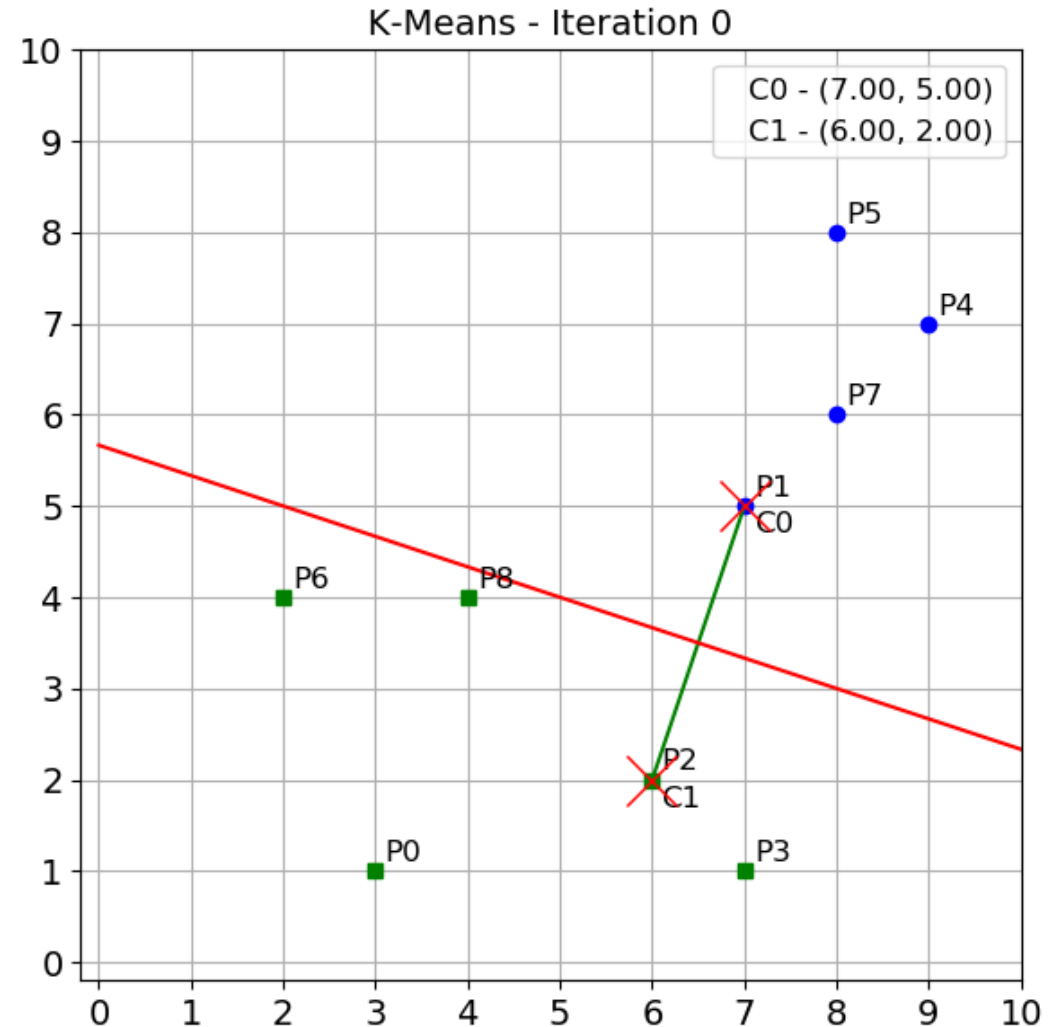
**Cluster 1:** P0,P2,P3,P6,P8

**Cluster 2:** P1,P4,P5,P7

**Centrod1:**

$$X1=(2+3+4+6+7)/5=4.4 \quad Y1=(4+1+4+2+1)/5=2.4$$

$$X2=(6+8+8+9)/4=8 \quad Y2=(5+6+8+7)/4=6.5$$

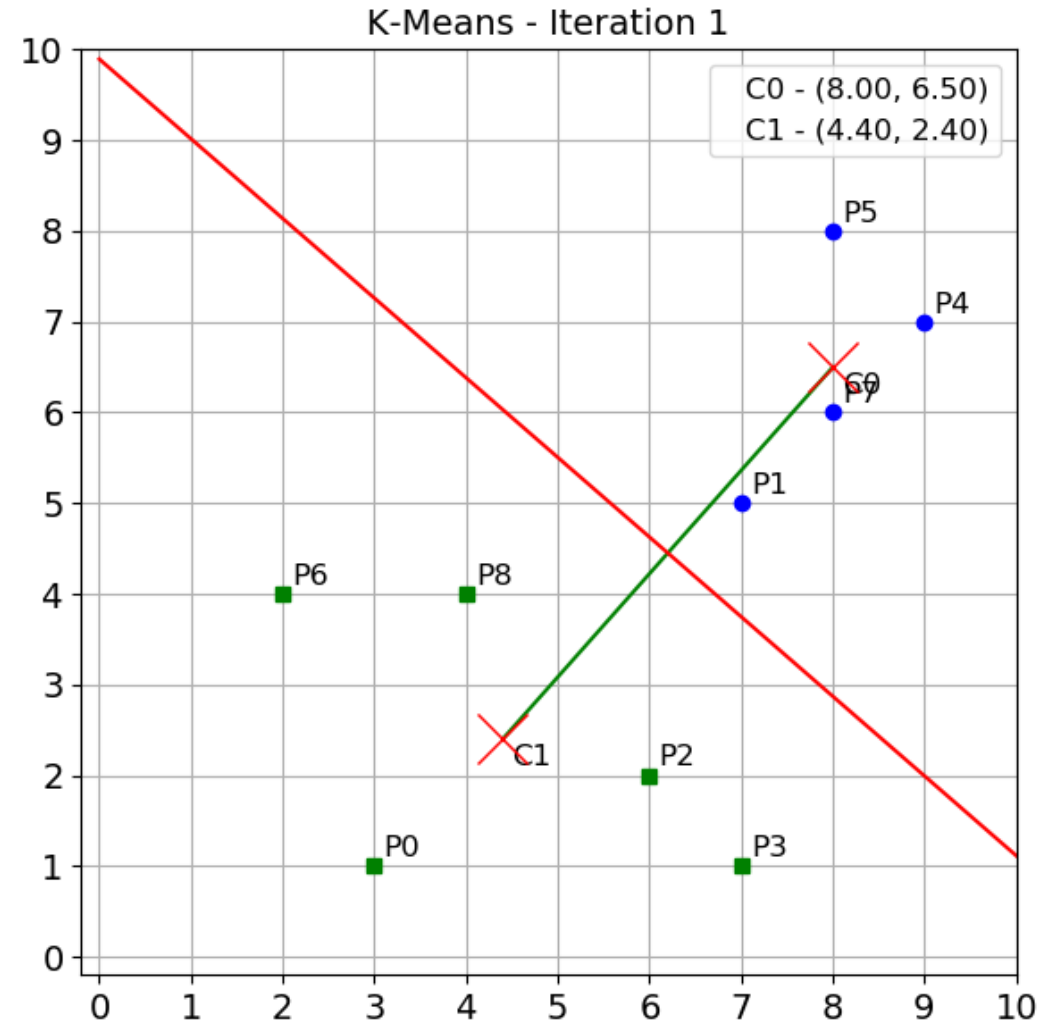


# Solution: Identify the Bisecting lines dividing the plane between pairs of centroids

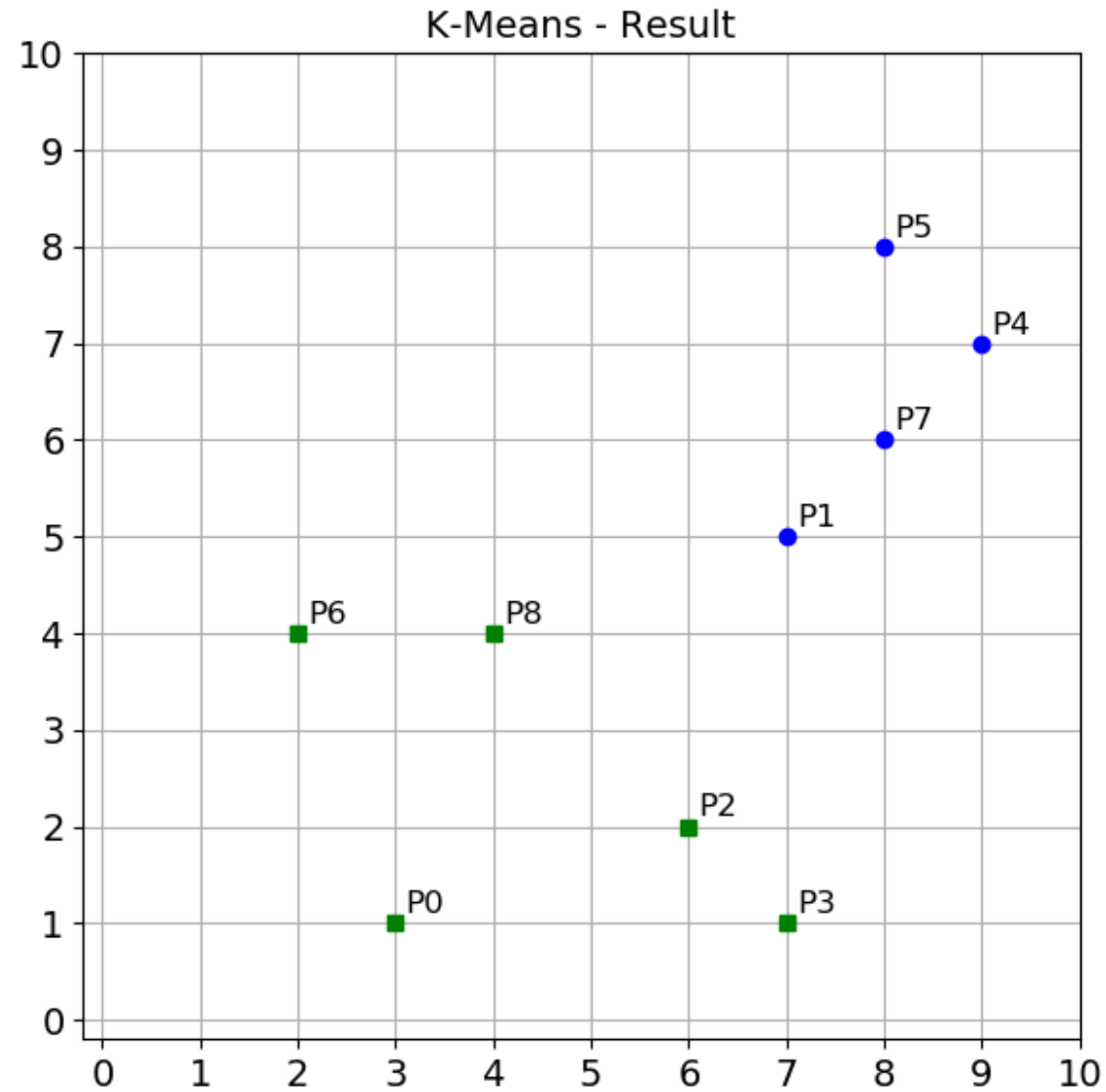
**Cluster 1:** P0,P2,P3,P6,P8

**Cluster 2:** P1,P4,P5,P7

**The cluster composition does not change,  
so K-means stops**



# K-means result

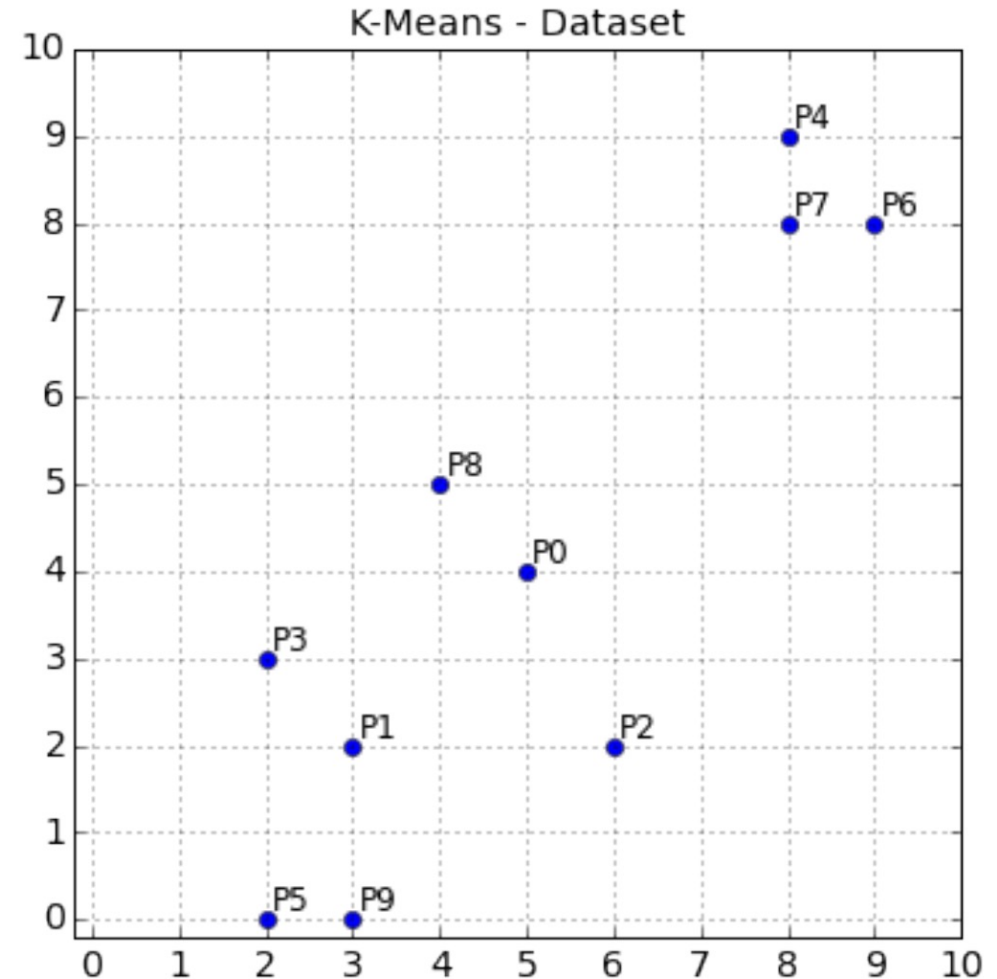


Ex 3

---

- Apply **K-means** to the dataset in the below table and figure using  $K=2$ , and the centroids  $c1=P2$  and  $c2=P9$ . Explain what happens in any iteration (**10 points**).
- Discuss the reason of the k-means termination (**4 points**).
- Identify another couple of initial centroids leading to the same clustering obtained in a) (**2 points**).

Points	X	Y
P0	5	4
P1	3	2
P2	6	2
P3	2	3
P4	8	9
P5	2	0
P6	9	8
P7	8	8
P8	4	5
P9	3	0



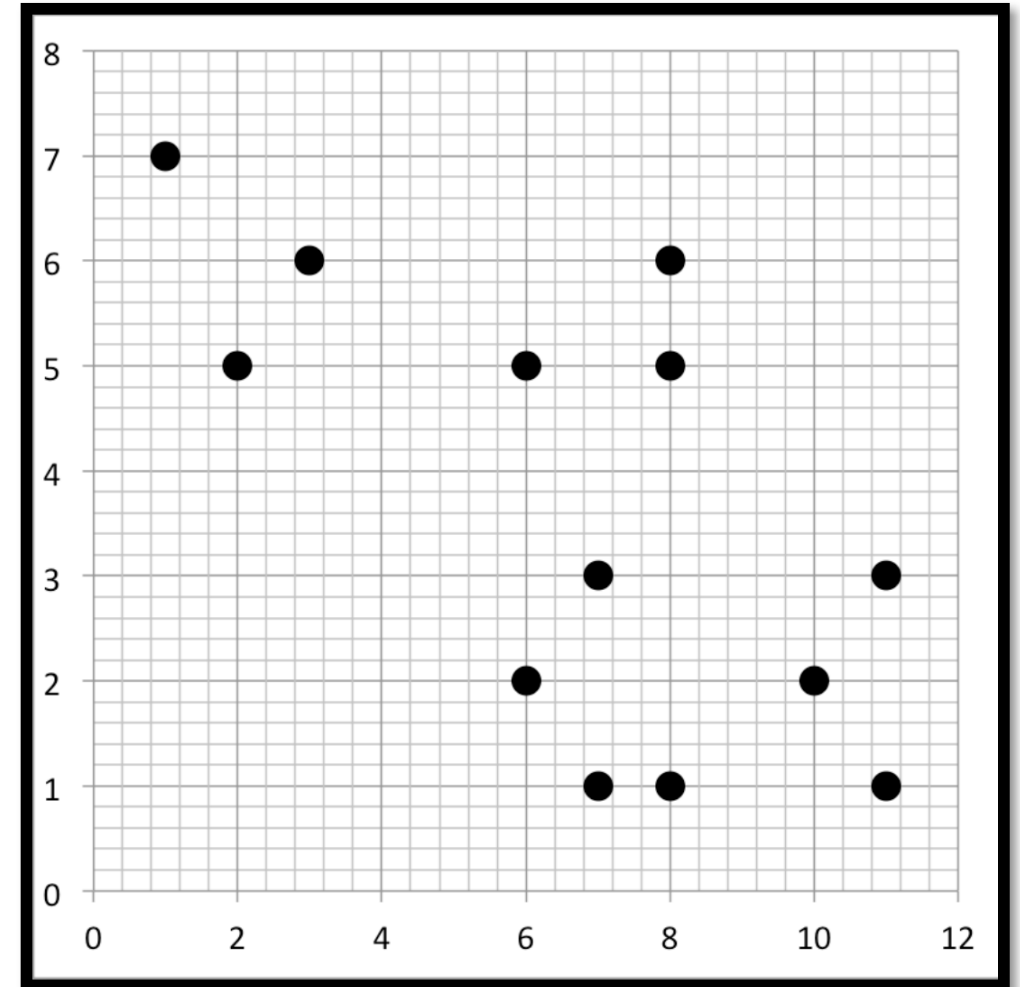
Ex 4

---



- a) Apply **K-means** to the dataset in the below table and figure using  $K=2$ , and the centroids  $c1=P5$  and  $c2=P8$ . Explain what happens in any iteration.
- b) Discuss the reason of the k-means termination

Points	X	Y
P1	1	7
P2	2	5
P3	3	6
P4	10	2
P5	11	1
P6	11	3
P7	6	2
P8	7	1
P9	7	3
P10	8	1
P11	6	5
P12	8	6
P13	8	5



# DBSCAN

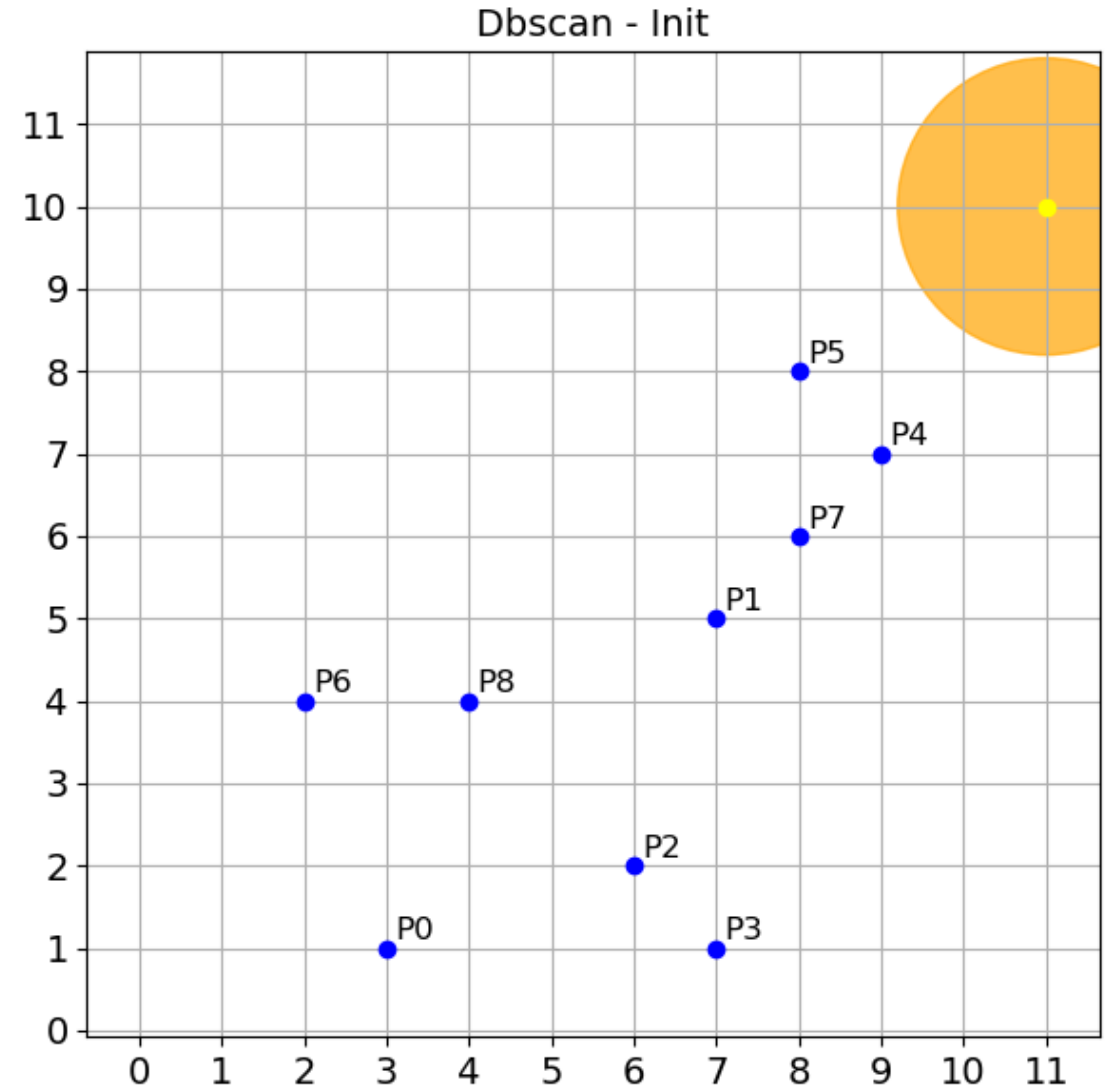
---

Ex 5

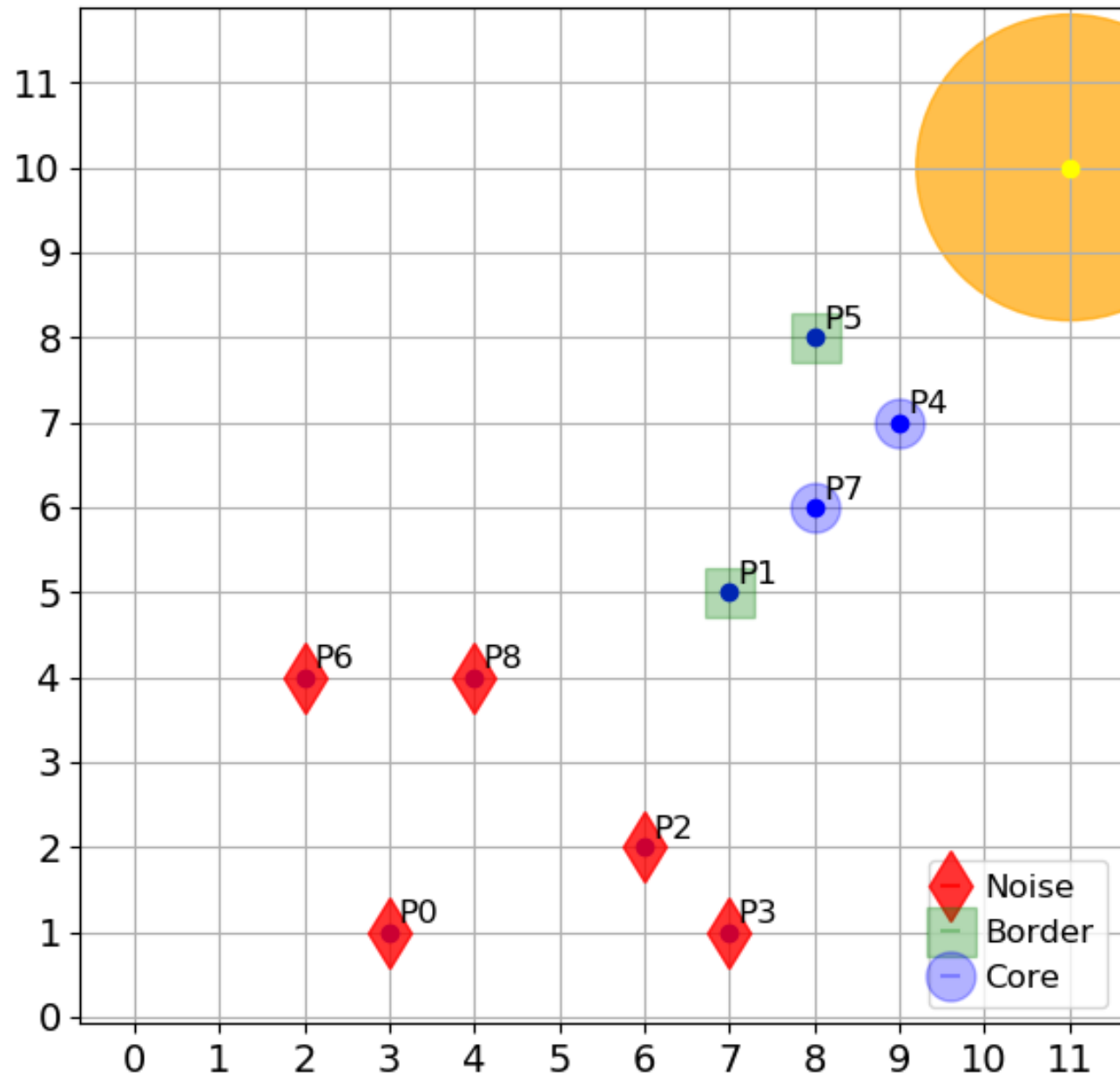
---

# DBSCAN simulation

- Eps=1.8
- MinPoints=3 (included the point)



Dbscan - Noise Points

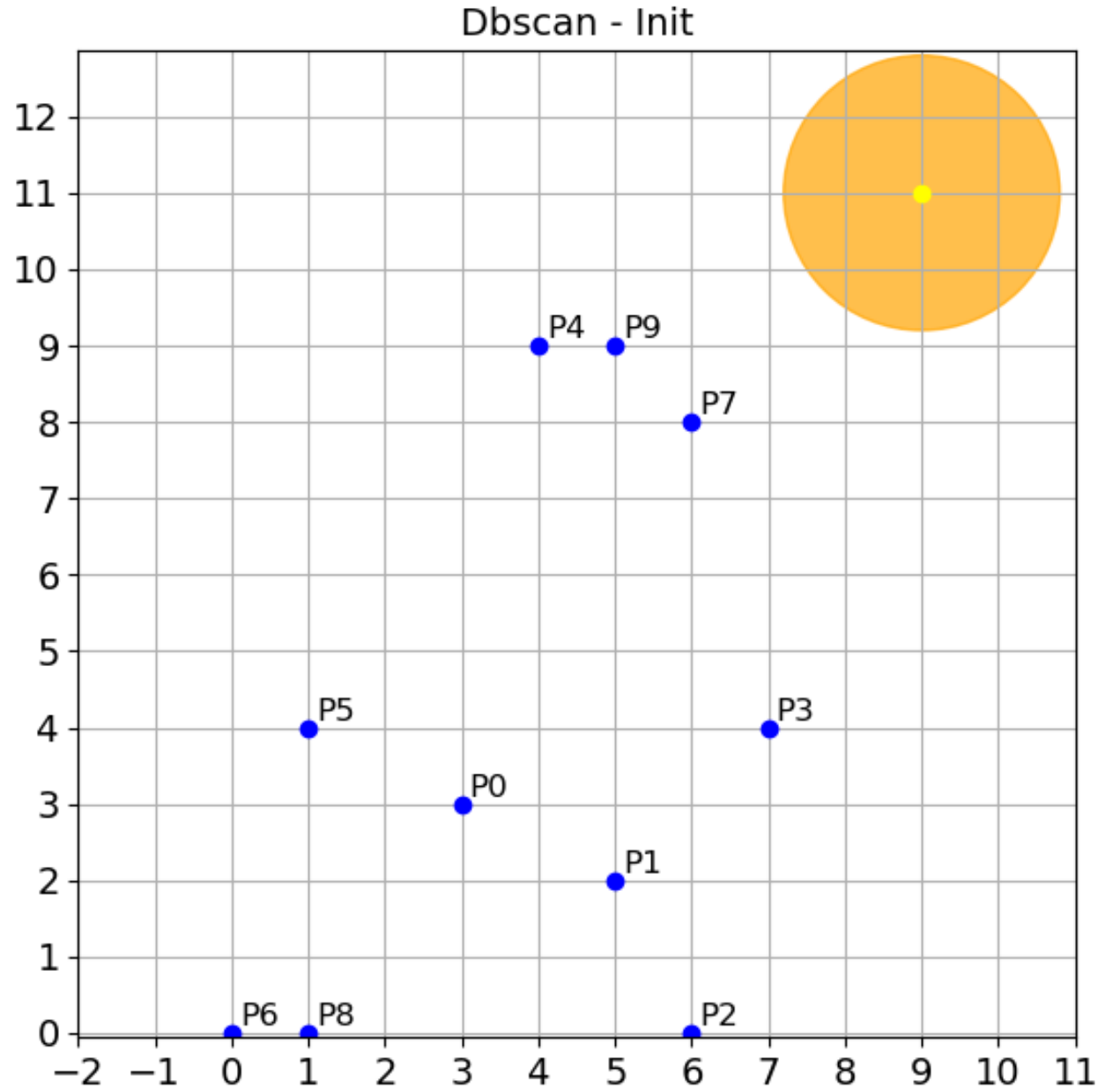


Ex 6

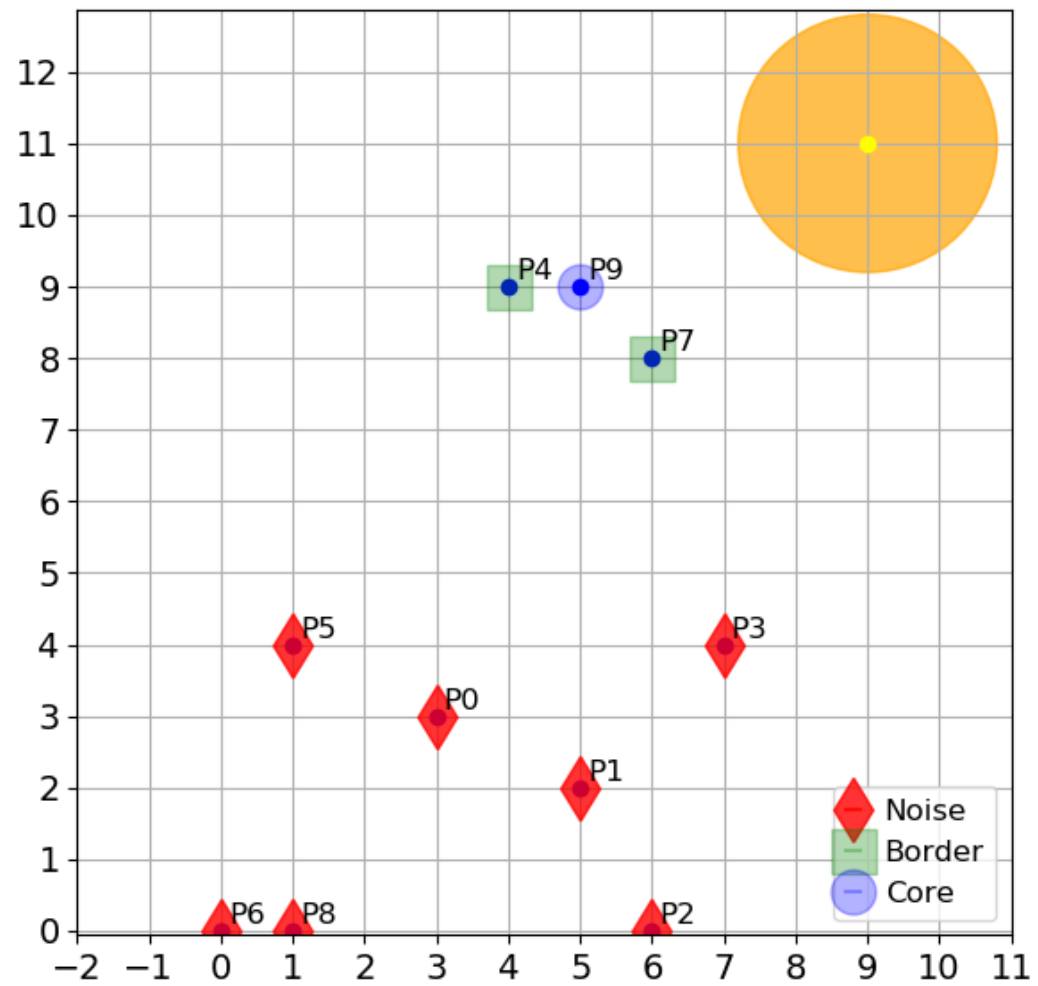
---

# DBSCAN simulation

- Eps=1.8
- MinPoints=3 (included the point)



Dbscan - Noise Points



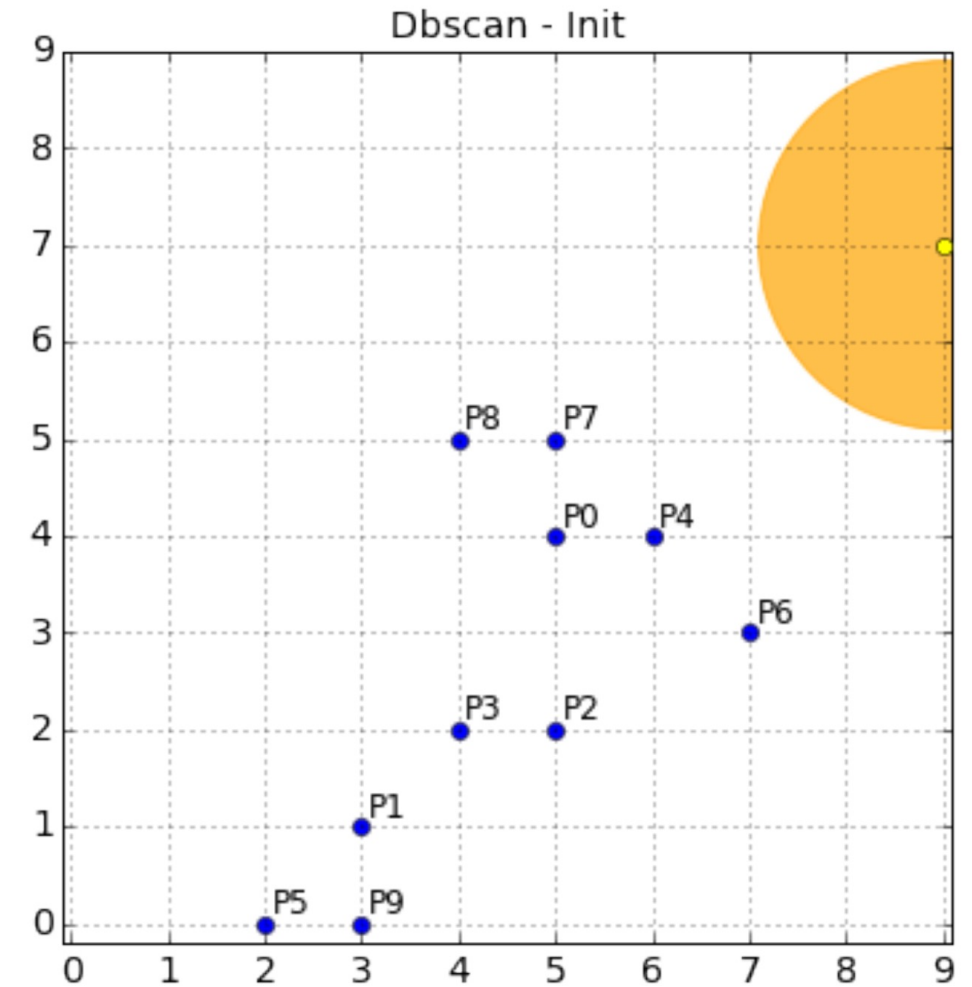


Ex 7

---

- Apply **Dbscan** on the data of previous exercise with radius  $\text{eps}=1.9$  and  $\text{minPts}=3$  (2 neighbor plus the point itself) and for each point specify if it is a core point, border point or noise **(10 points)**.
- Indicate the composition of the clusters obtained **(2 points)**.
- Which is the minimum  $\text{eps}$  to obtain a unique cluster? How many core and border points are presents in this new clustering? **(3 points)**.

Points	X	Y
P0	5	4
P1	3	1
P2	5	2
P3	4	2
P4	6	4
P5	2	0
P6	7	3
P7	5	5
P8	4	5
P9	3	0



# Hierarchical Clustering

---

Ex 8

---

# Hierarchical: Single-LINK

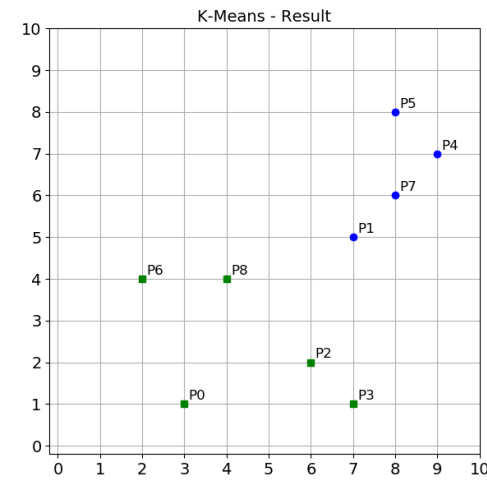
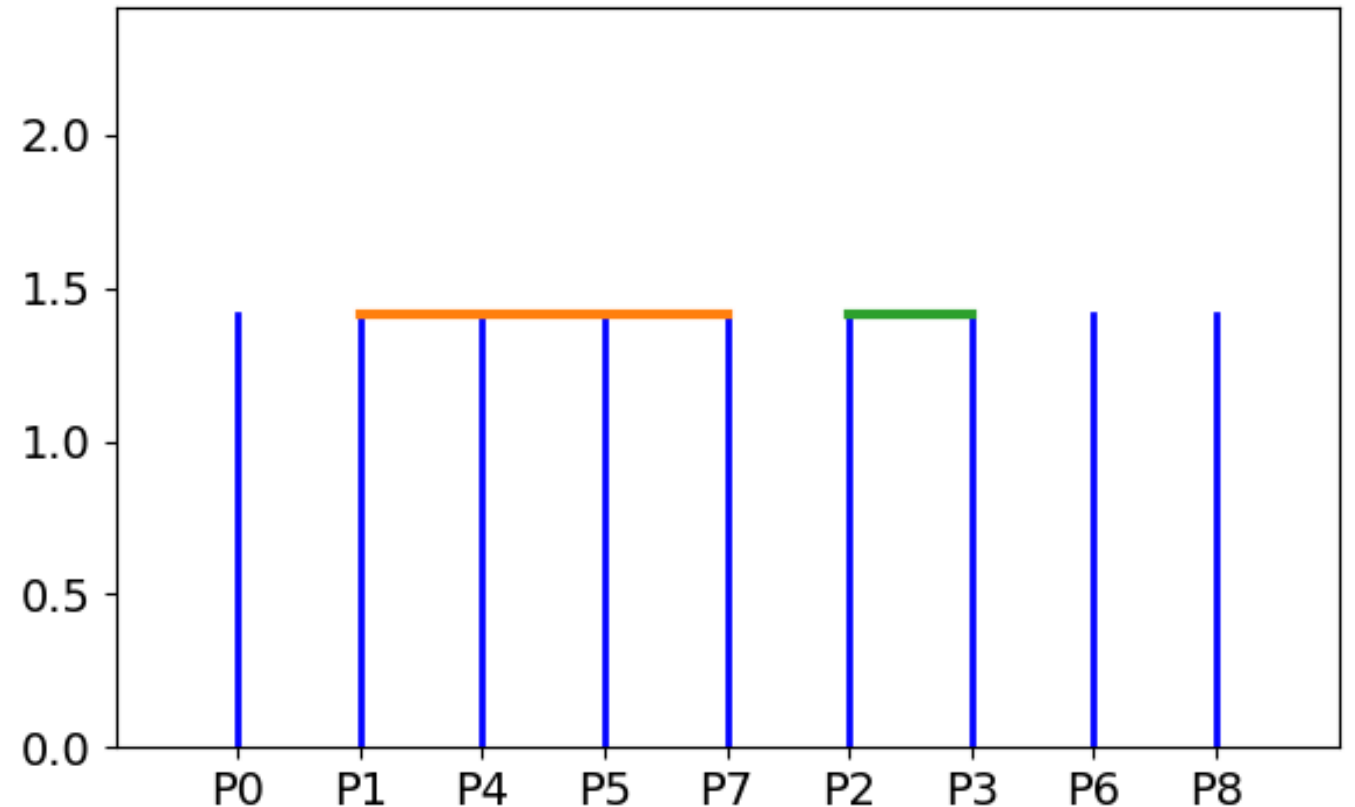
Euclidean Distance

0	5.66	3.16	4	8.49	8.6	3.16	7.07	3.16
5.66	0	3.16	4	2.83	3.16	5.1	<b>1.41</b>	3.16
3.16	3.16	0	<b>1.41</b>	5.83	6.32	4.47	4.47	2.83
4	4	<b>1.41</b>	0	6.32	7.07	5.83	5.1	4.24
8.49	2.83	5.83	6.32	0	<b>1.41</b>	7.62	1.41	5.83
8.6	3.16	6.32	7.07	<b>1.41</b>	0	7.21	2	5.66
3.16	5.1	4.47	5.83	7.62	7.21	0	6.32	2
7.07	<b>1.41</b>	4.47	5.1	1.41	2	6.32	0	4.47
3.16	3.16	2.83	4.24	5.83	5.66	2	4.47	0

**distance merge 1.41**

Min  
Distance

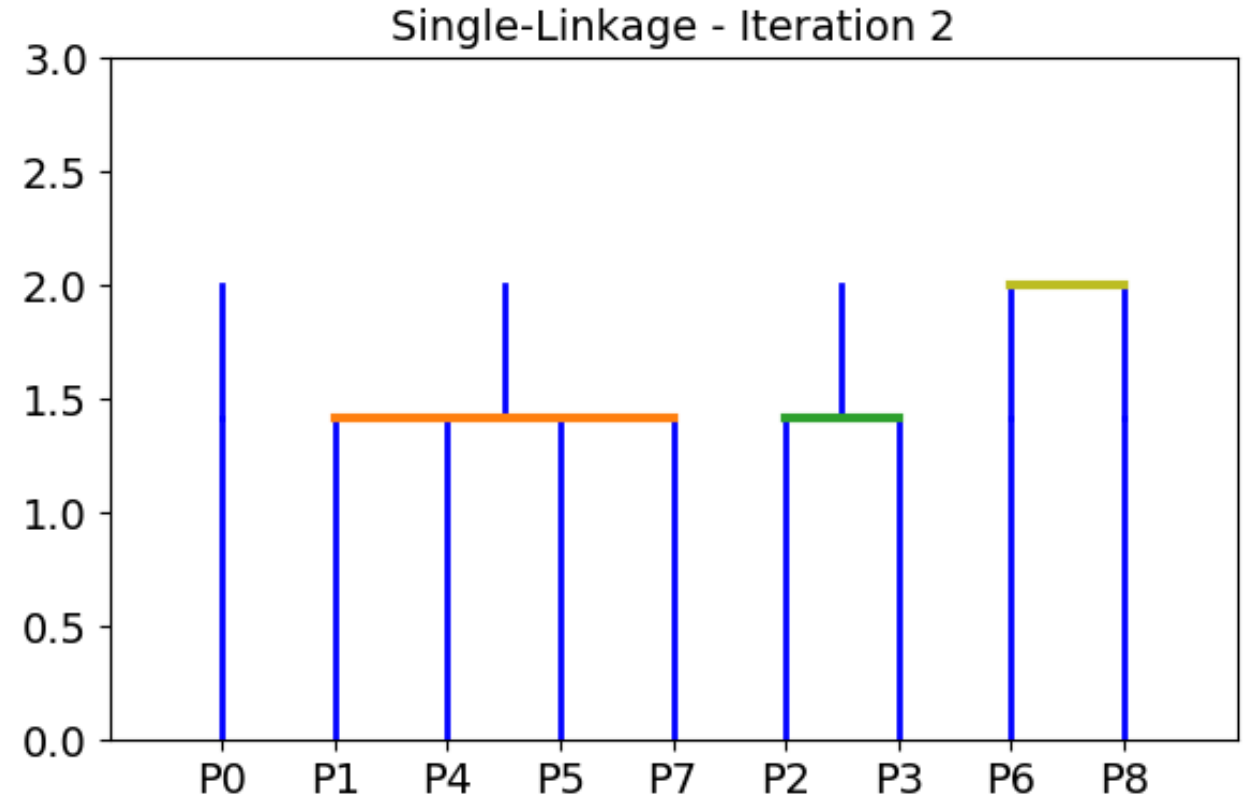
Single-Linkage - Iteration 1



# Hierarchical: **Single-LINK**- Euclidean Distance

[(0,)]	(1,4,5,7)	(2, 3),	(6,),	(8,)]
[0.	5.66	3.16	3.16	3.16]
[5.66	0.	3.16	5.1	3.16]
[3.16	3.16	0.	4.47	2.83]
[3.16	5.1	4.47	0.	2. ]
[3.16	3.16	2.83	2.	0. ]

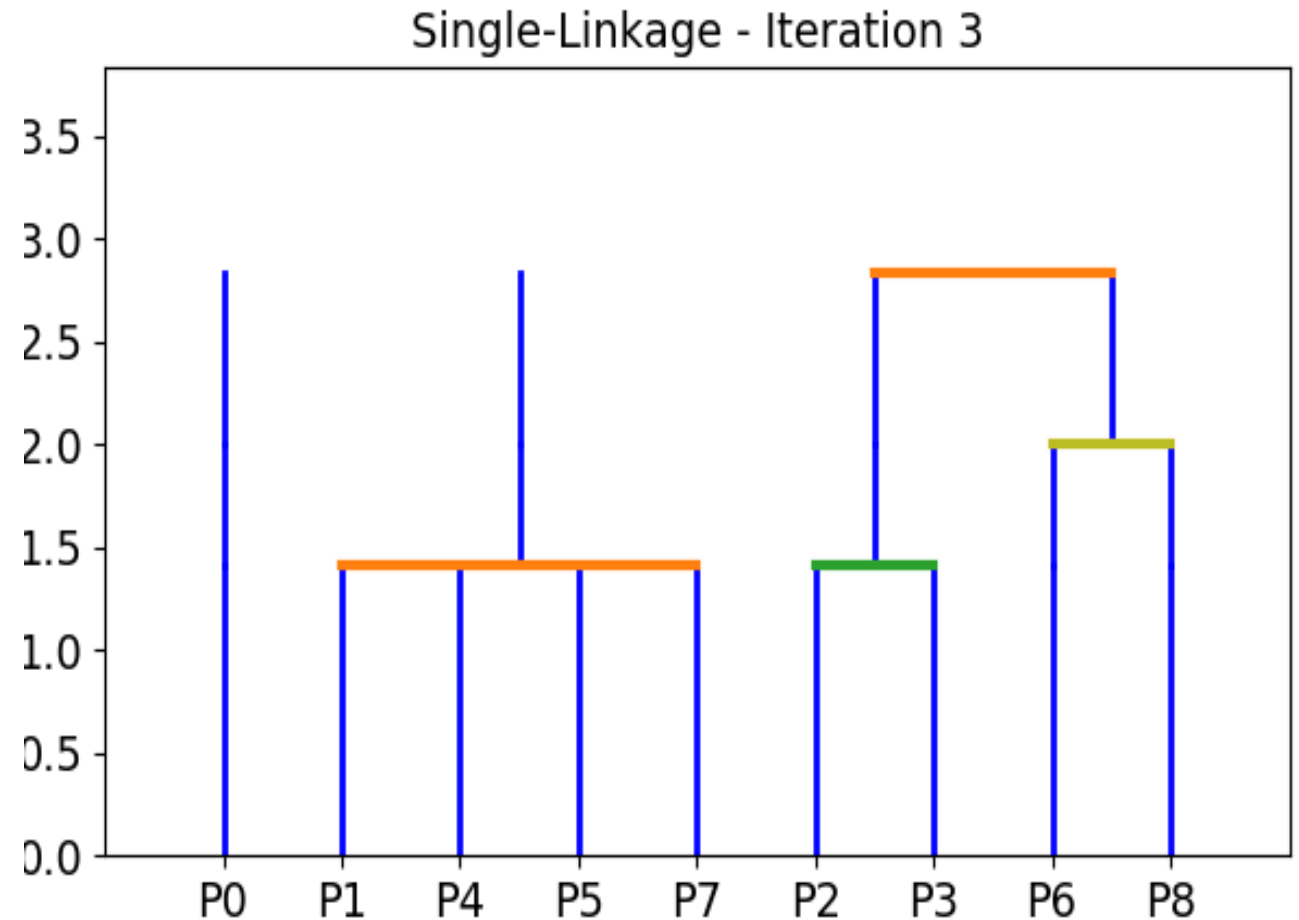
**distance merge 2.00**



# Hierarchical: Single-LINK- Euclidean Distance

[(0,),	(1,4,5,7)	(2, 3),	(6, 8)]
[0.	5.66	3.16	3.16]
[5.66	0.	3.16	3.16]
[3.16	3.16	0.	2.83]
[3.16	3.16	2.83	0. ]

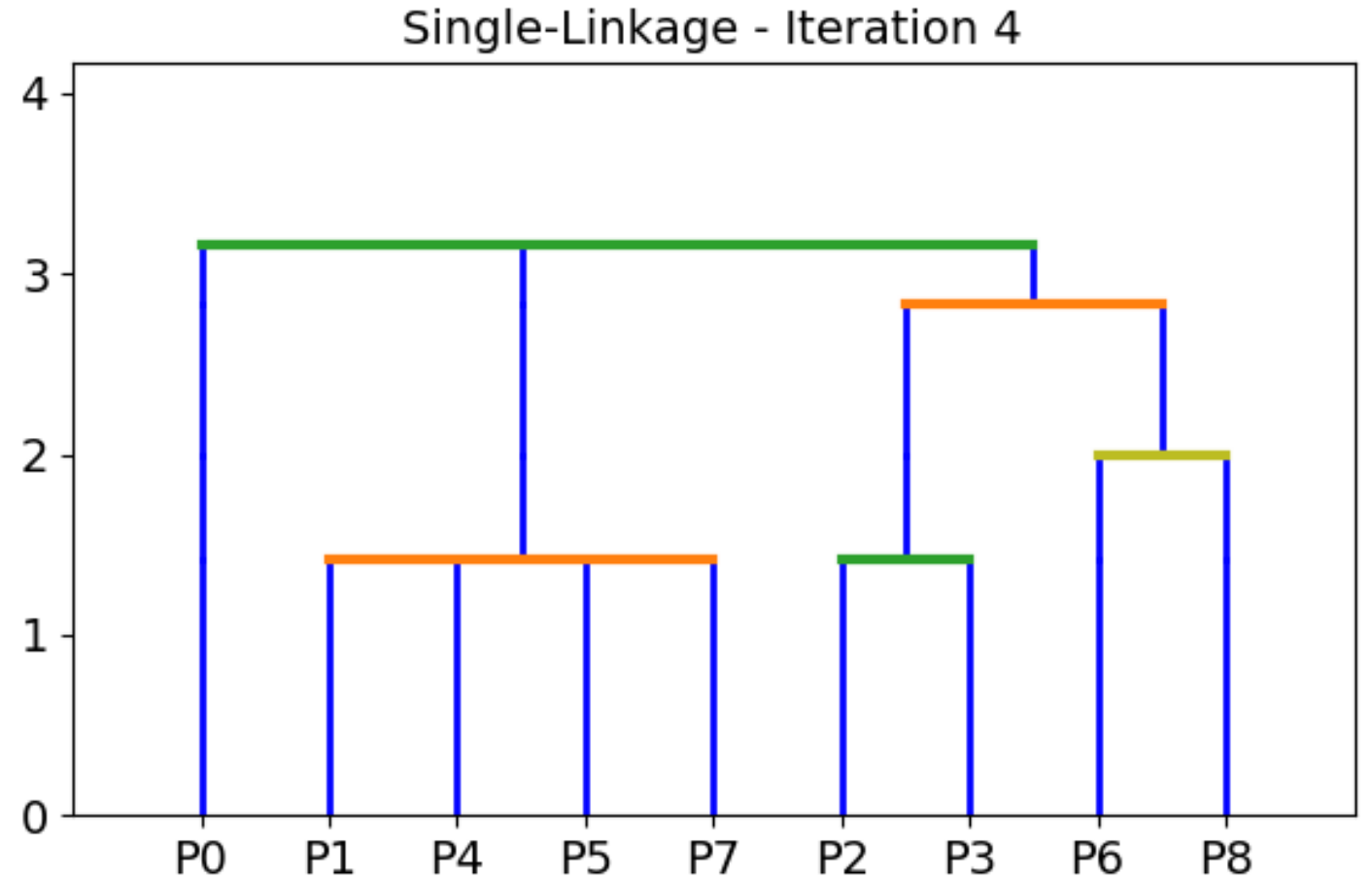
**distance merge 2.83**



# Hierarchical: Single-LINK- Euclidean Distance

[(0,)	(1,4,5,7)	(2,3,6,8)]
[0.	5.66	3.16]
[5.66	0.	3.16]
[3.16	3.16	0. ]

**distance merge 3.16**





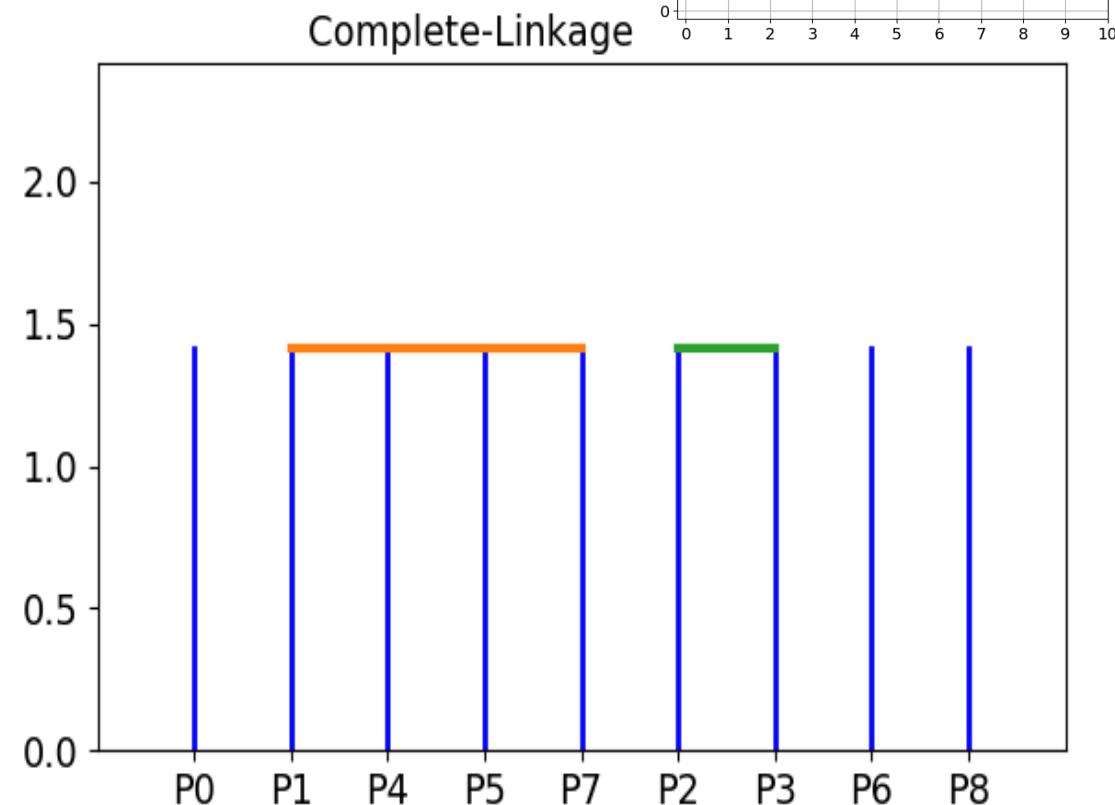
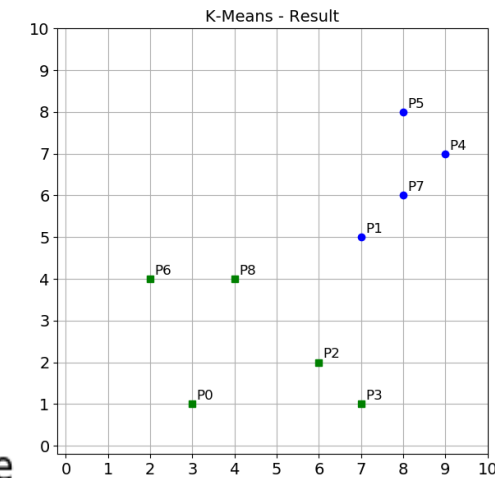
Ex 9

---

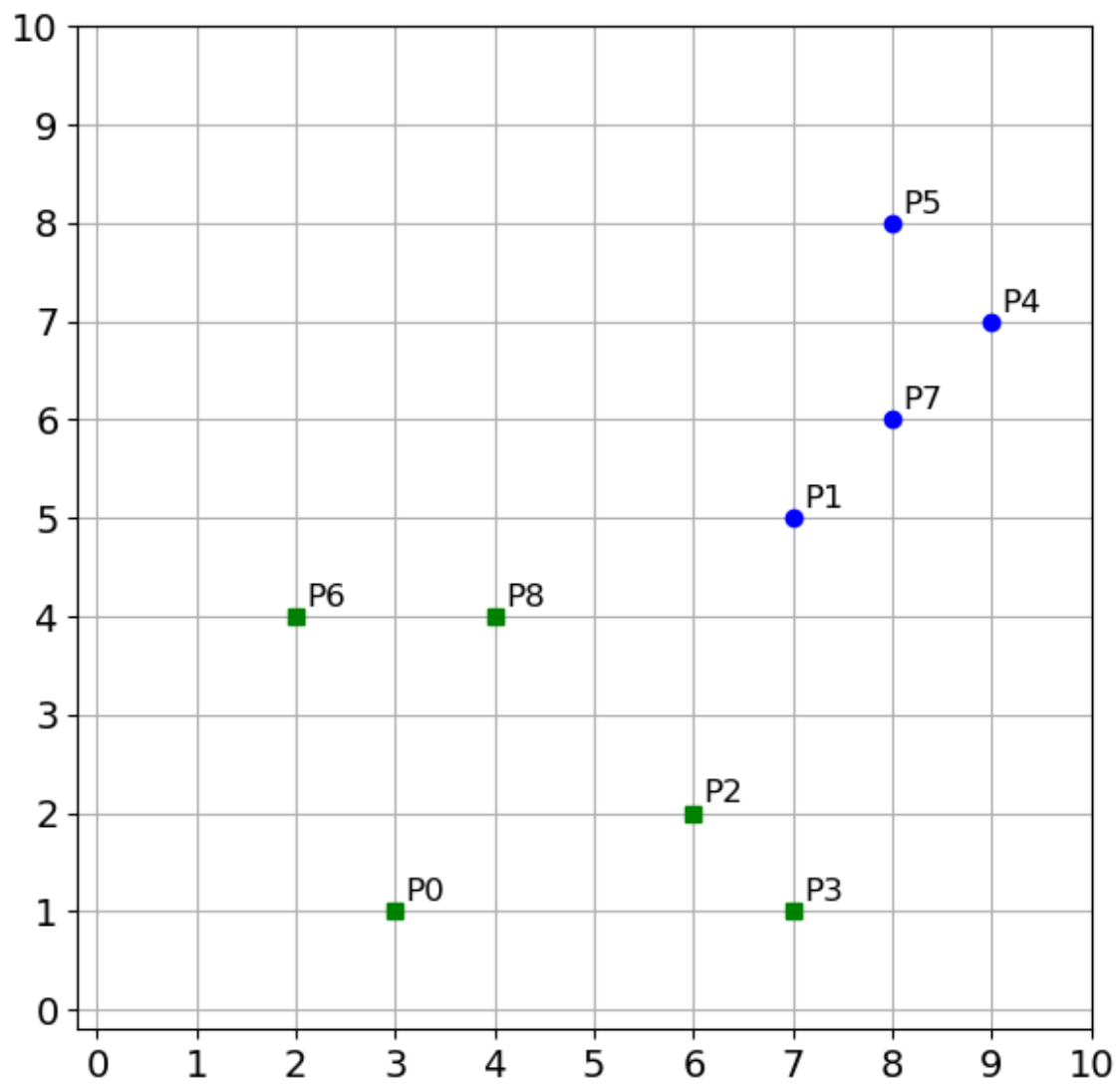
# Hierarchical: Complete-LINK

Euclidean Distance

0	5.66	3.16	4	8.49	8.6	3.16	7.07	3.16
5.66	0	3.16	4	2.83	3.16	5.1	<b>1.41</b>	3.16
3.16	3.16	0	<b>1.41</b>	5.83	6.32	4.47	4.47	2.83
4	4	<b>1.41</b>	0	6.32	7.07	5.83	5.1	4.24
8.49	2.83	5.83	6.32	0	<b>1.41</b>	7.62	1.41	5.83
8.6	3.16	6.32	7.07	<b>1.41</b>	0	7.21	2	5.66
3.16	5.1	4.47	5.83	7.62	7.21	0	6.32	2
7.07	<b>1.41</b>	4.47	5.1	1.41	2	6.32	0	4.47
3.16	3.16	2.83	4.24	5.83	5.66	2	4.47	0



Where is the “mistake” ????



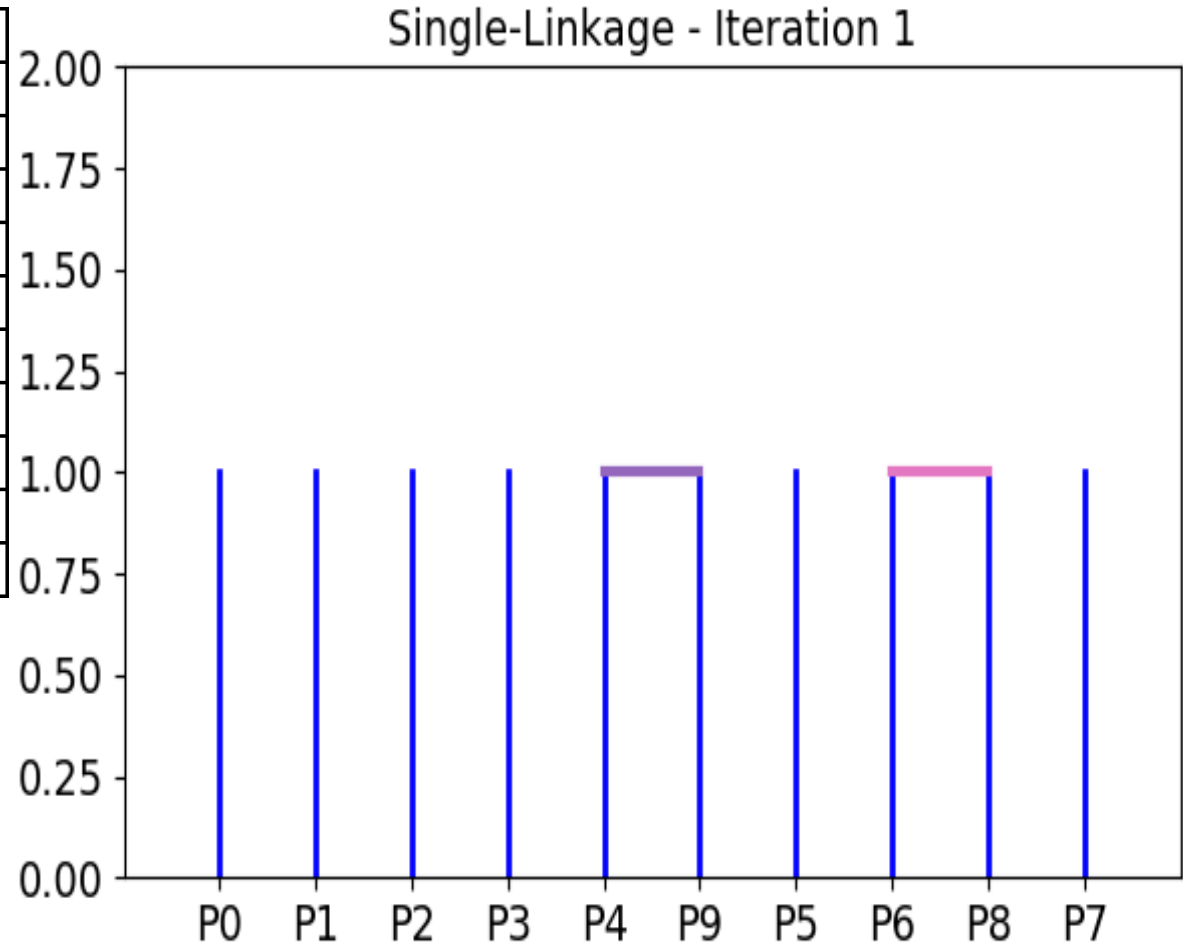
Ex 10

---

# Hierarchical: Single-LINK- Euclidean Distance

(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0.	2.24	4.24	4.12	6.08	2.24	4.24	5.83	3.61	6.32
2.24	0.	2.24	2.83	7.07	4.47	5.39	6.08	4.47	7.
4.24	2.24	0.	4.12	9.22	6.4	6.	8.	5.	9.06
4.12	2.83	4.12	0.	5.83	6.	8.06	4.12	7.21	5.39
6.08	7.07	9.22	5.83	0.	5.83	9.85	2.24	9.49	1.
2.24	4.47	6.4	6.	5.83	0.	4.12	6.4	4.	6.4
4.24	5.39	6.	8.06	9.85	4.12	0.	10.	1.	10.3
5.83	6.08	8.	4.12	2.24	6.4	10.	0.	9.43	1.41
3.61	4.47	5.	7.21	9.49	4.	1.	9.43	0.	9.85
6.32	7.	9.06	5.39	1.	6.4	10.3	1.41	9.85	0.

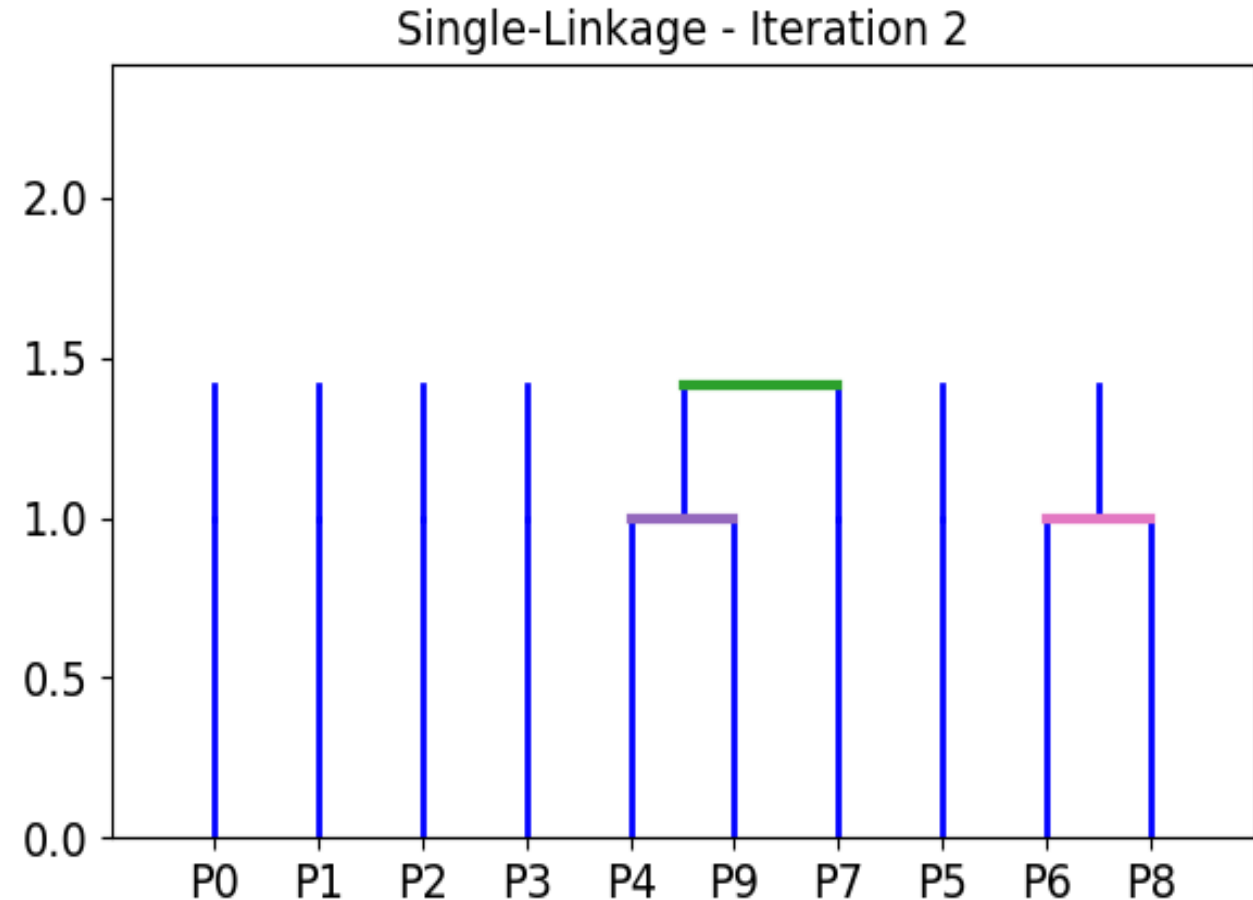
distance merge 1.00



# Hierarchical: Single-LINK- Euclidean Distance

( 0 )	( 1 )	( 2 )	( 3 )	(4,9)	( 5 )	(6,8)	( 7 )
[0.	2.24	4.24	4.12	6.08	2.24	3.61	5.83]
[2.24	0.	2.24	2.83	7.	4.47	4.47	6.08]
[4.24	2.24	0.	4.12	9.06	6.4	5.	8. ]
[4.12	2.83	4.12	0.	5.39	6.	7.21	4.12]
[6.08	7.	9.06	5.39	0.	5.83	9.49	1.41]
[2.24	4.47	6.4	6.	5.83	0.	4.	6.4 ]
[3.61	4.47	5.	7.21	9.49	4.	0.	9.43]
[5.83	6.08	8.	4.12	1.41	6.4	9.43	0. ]

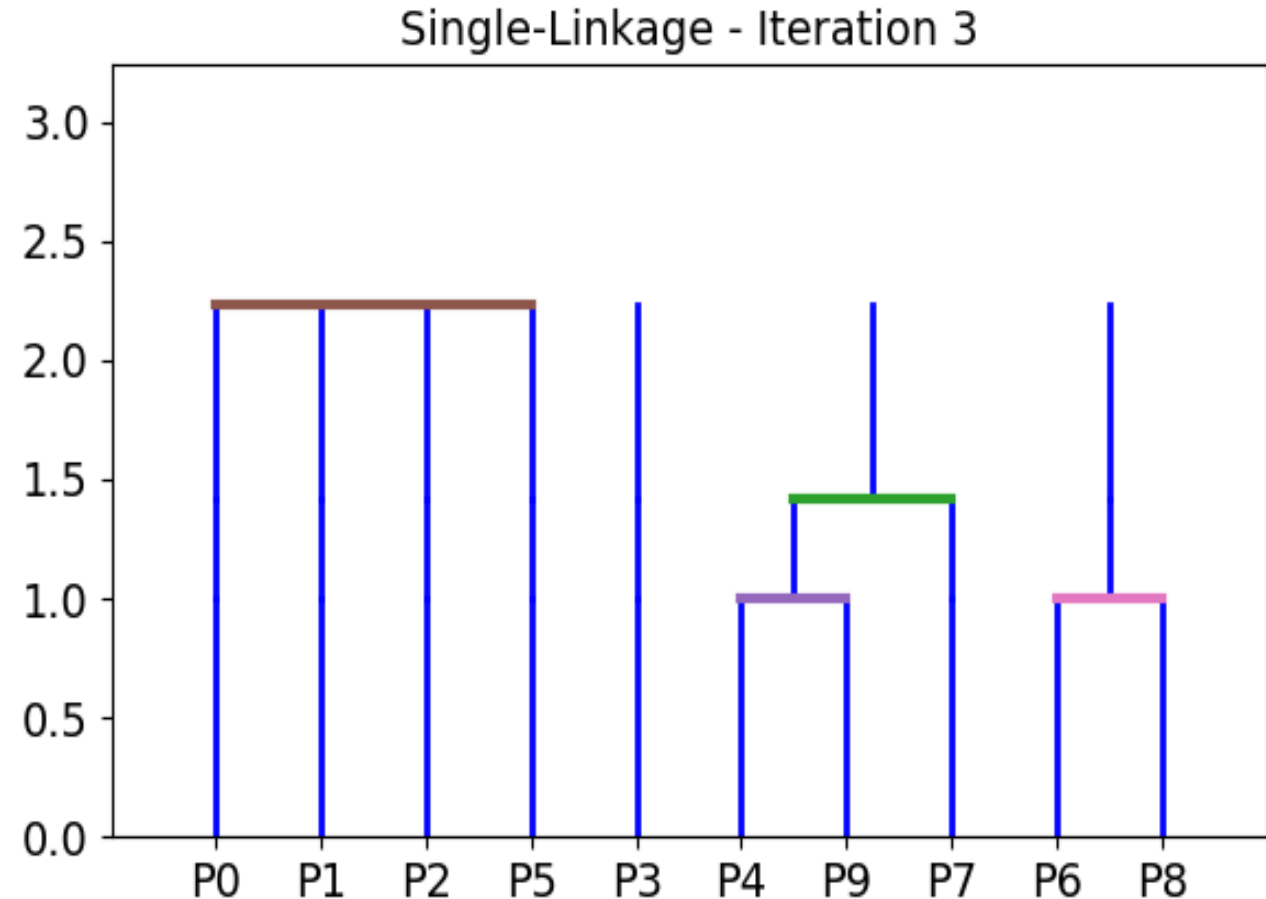
**distance merge 1.41**



# Hierarchical: Single-LINK- Euclidean Distance

( 0 )	( 1 )	( 2 )	( 3 )	(4,7,9)	( 5 )	(6,8)
[0.	2.24	4.24	4.12	5.83	2.24	3.61]
[2.24	0.	2.24	2.83	6.08	4.47	4.47]
[4.24	2.24	0.	4.12	8.	6.4	5. ]
[4.12	2.83	4.12	0.	4.12	6.	7.21]
[5.83	6.08	8.	4.12	0.	5.83	9.43]
[2.24	4.47	6.4	6.	5.83	0.	4. ]
[3.61	4.47	5.	7.21	9.43	4.	0. ]

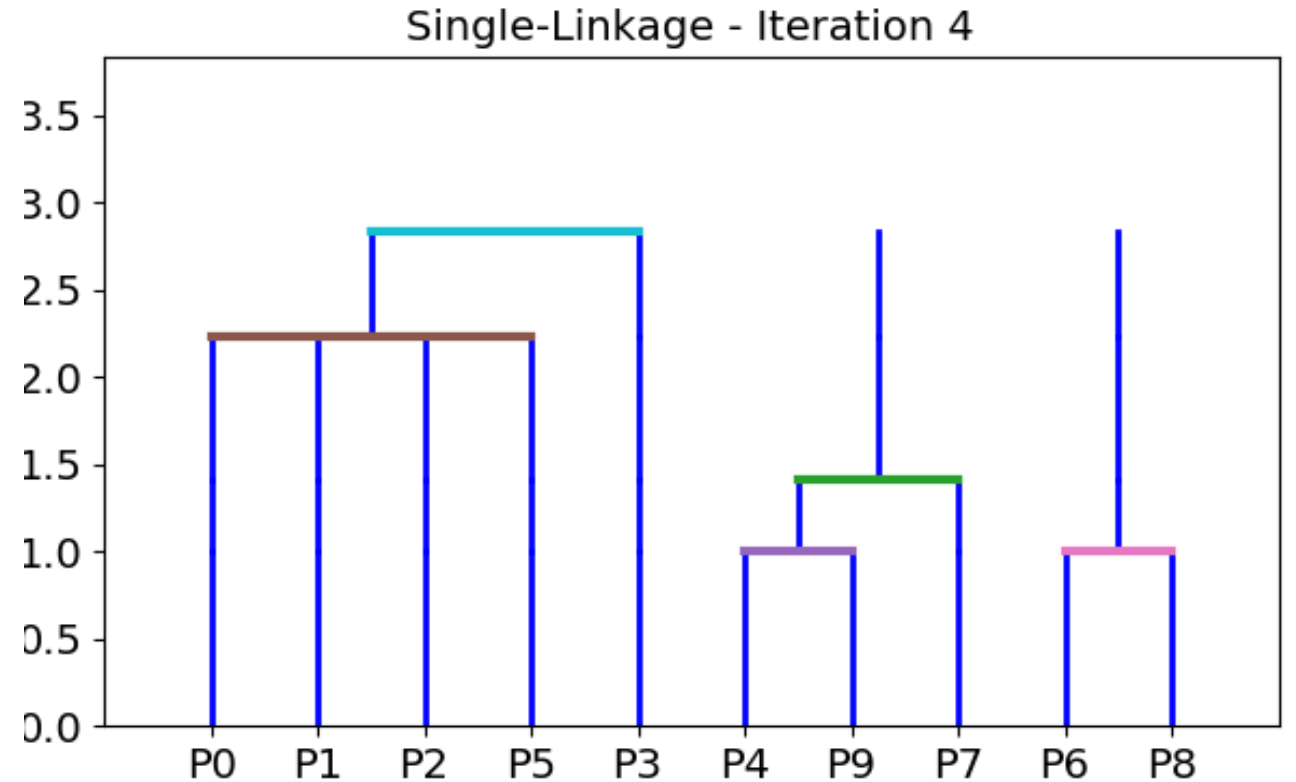
**distance merge 2.24**



# Hierarchical: Single-LINK- Euclidean Distance

(0,1,2,5 )	( 3 )	(4,7,9)	(6,8)
[0.	2.83	5.83	3.61]
[2.83	0.	4.12	7.21]
[5.83	4.12	0.	9.43]
[3.61	7.21	9.43	0. ]

**distance merge 2.83**

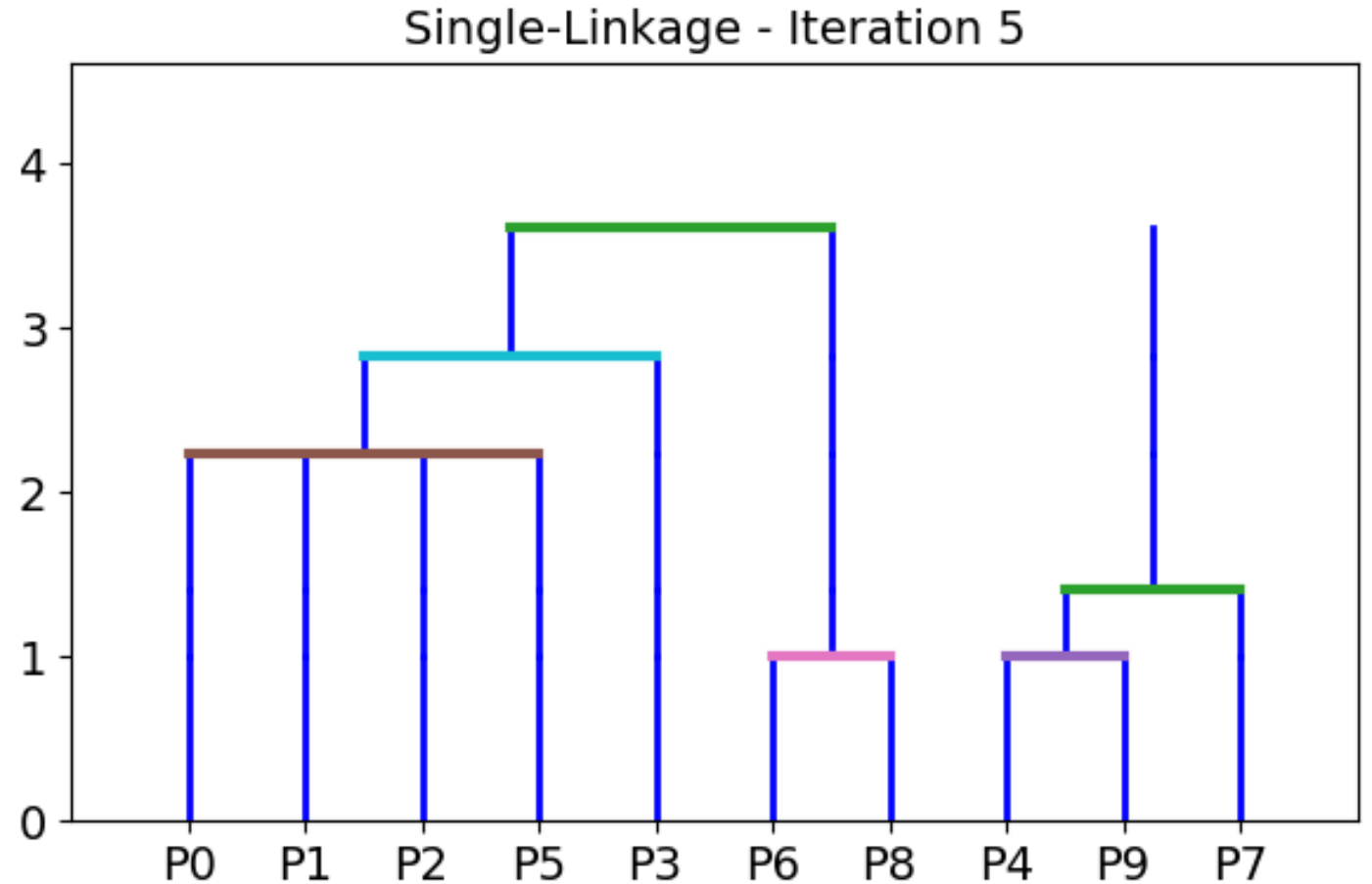




# Hierarchical: Single-LINK- Euclidean Distance

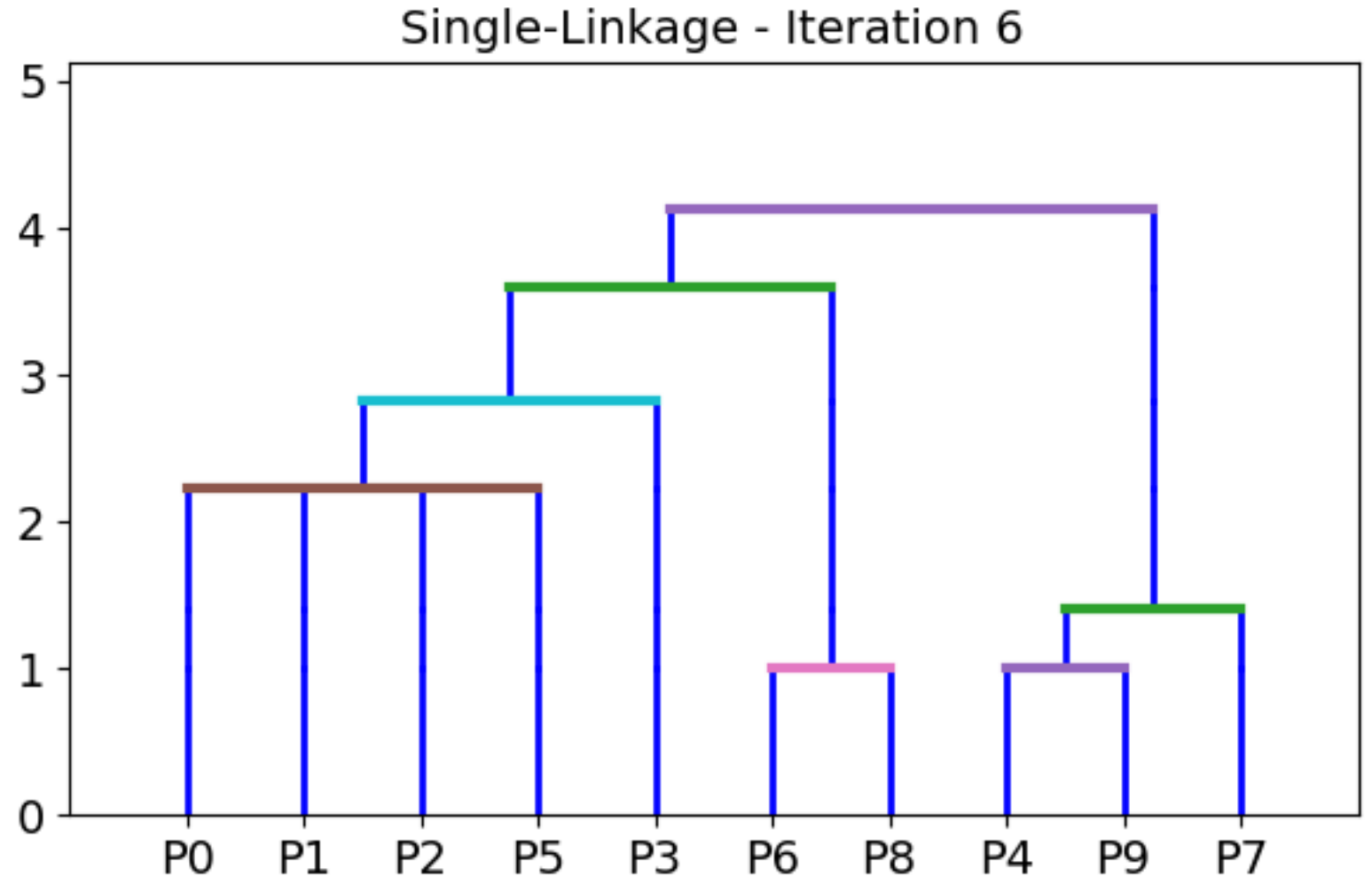
(0,1,2,3,5 )	(4,7,9)	(6,8)
[0.	4.12	3.61]
[4.12	0.	9.43]
[3.61	9.43	0. ]

**distance merge 3.61**



# Hierarchical: Single-LINK- Euclidean Distance

(0,1,2,3,5,6,8	(4,7,9)
[0.	4.12 ]
[4.12	0. ]



Ex 11/12

---

Consider the following points and use the Manhattan distance to solve the following exercises:

1. Apply the single-linkage HAC on the dataset and draw the corresponding dendrogram.
2. Apply the complete-linkage HAC on the dataset and draw the corresponding dendrogram.

P1	1	7
P2	6	2
P3	7	1
P4	7	3
P5	8	1
P6	6	5
P7	8	6
P8	8	5