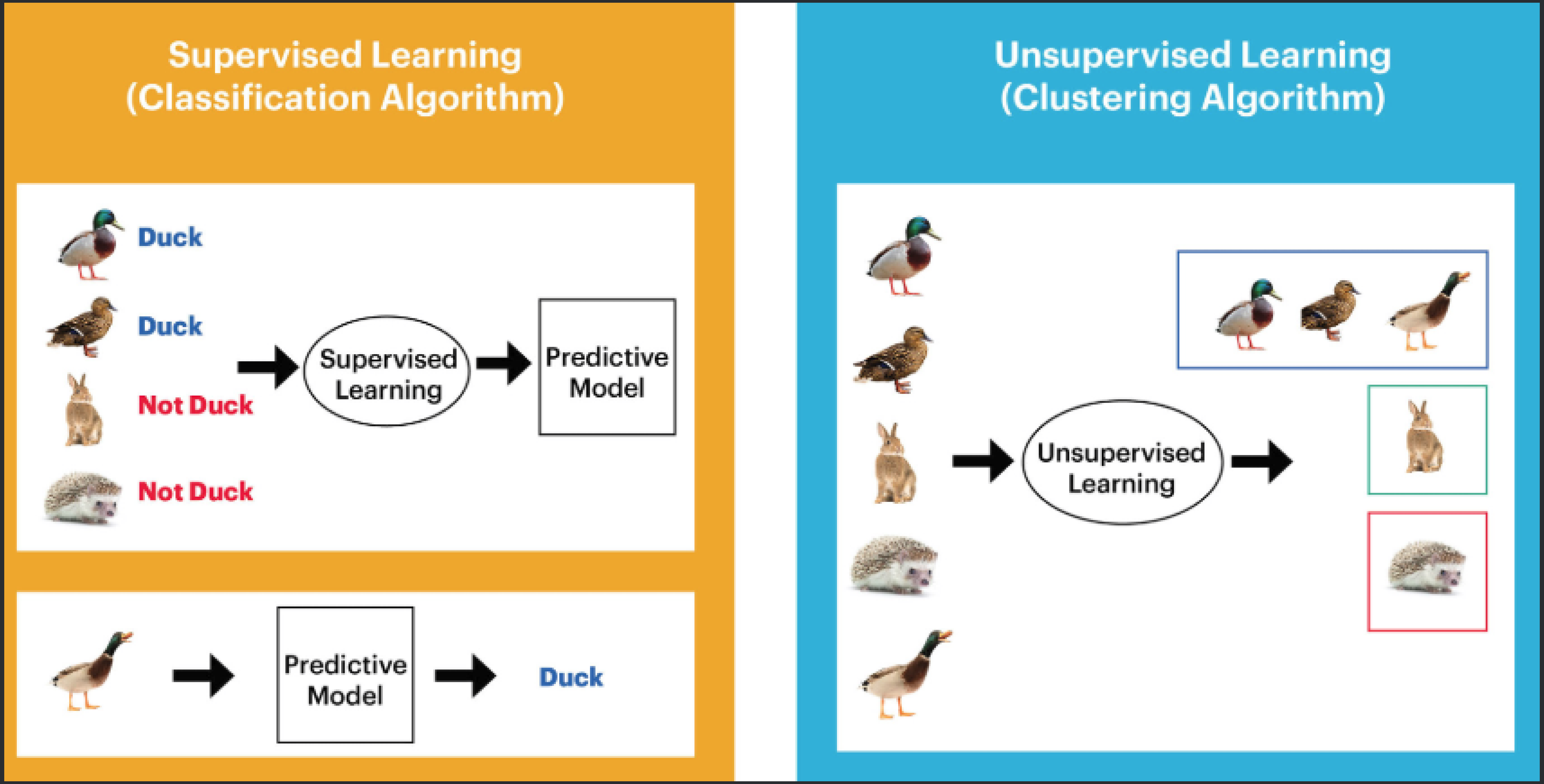
Unsupervised Learning


Supervised Learning
(Classification Algorithm)
Duck
Duck
Not Duck
Not Duck
Supervised Learning
Predictive Model
Predictive Model
Duck
Unsupervised Learning
(Clustering Algorithm)
Unsupervised Learning

Machine Learning

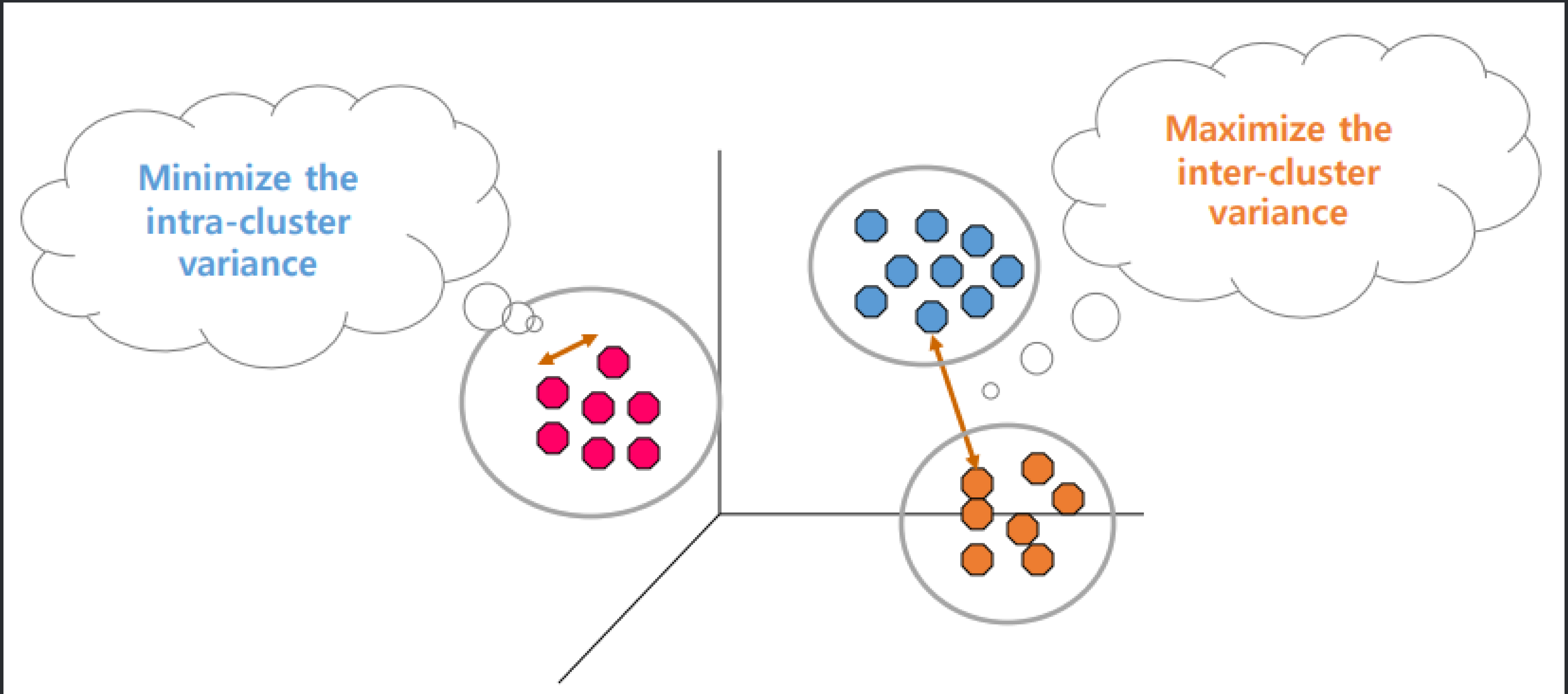
UN-SUPERVISED LEARNING

Clustering

클러스터링은 데이터에서 비슷한 객체들을 하나의 그룹으로 묶는 것

그럼 데이터가 비슷한 기준은?

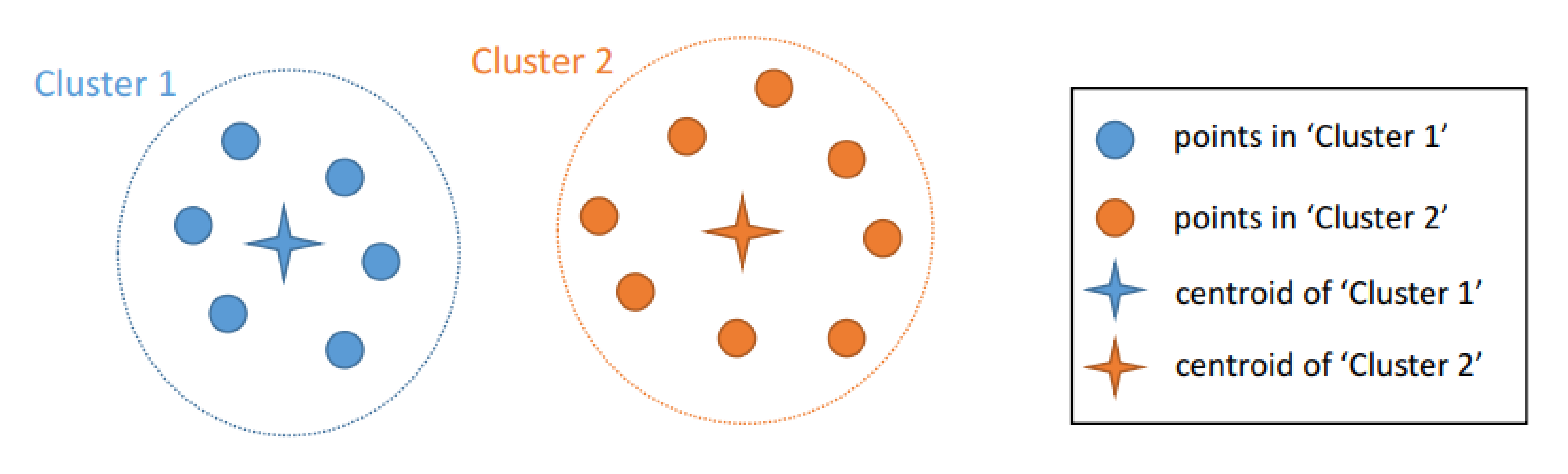
유사도 (거리) 정보 기반


Minimize the intra-cluster variance
Maximize the inter-cluster variance

클러스터링 알고리즘

- k-means clustering
- Hierarchical clustering
- Others
- Density-based spatial clustering of applications with noise (DBSCAN)
- Gaussian mixture model
- Self-organizing map (SOM)

k-means clustering


Cluster 1
Cluster 2
points in ‘Cluster 1’
points in ‘Cluster 2’
centroid of ‘Cluster 1’
centroid of ‘Cluster 2’

유사도

$d(X_i, X_j)$

유클리디안 거리 (L2 distance)

두 점 P와 Q가 각각 $P=(p_1, p_2, p_3, \ldots, p_n)$와 $Q=(q_1, q_2, q_3, \ldots, q_n)$의 좌표를 갖을 때 두 점 사이의 거리를 계산하는 유클리디안 거리 $(Euclidean distance)$공식은 다음과 같습니다.

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

유클리디안 거리 (L2 distance)

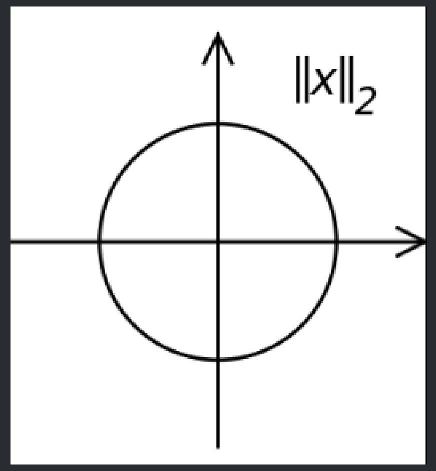
$$\frac{1}{1+Ed}$$

#	A	B	C	D
1	3	2	0	2
2	1	2	3	0
3	2	2	2	2
4	1	5	0	0

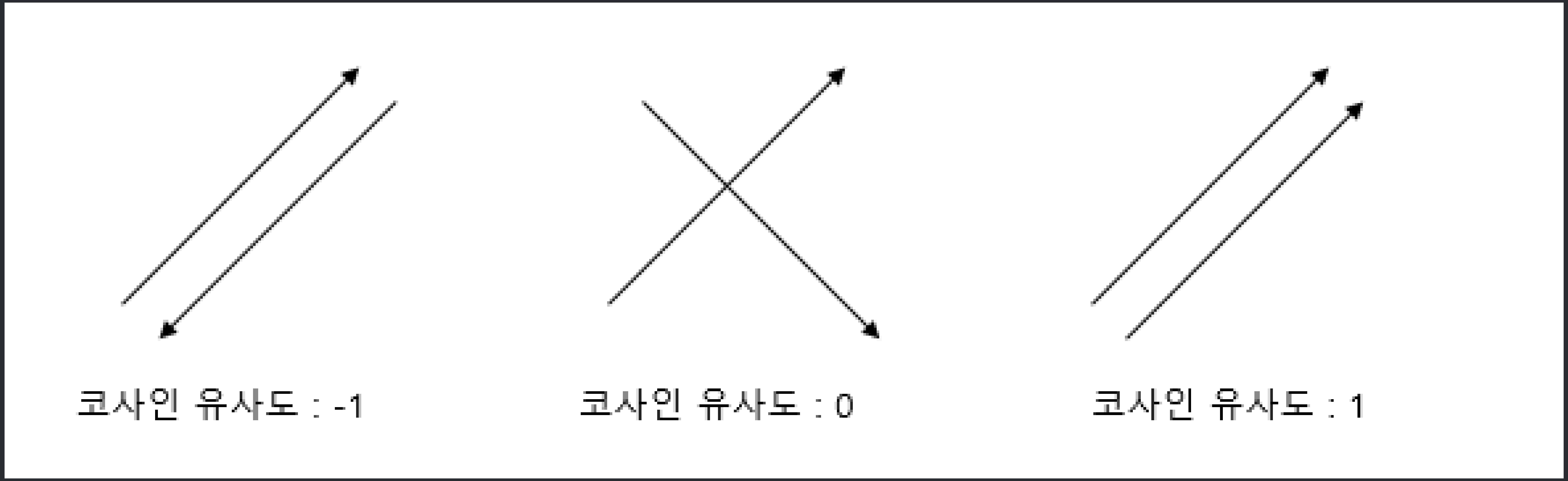
$$dist(D1, Q) = \sqrt{(3-1)^2 + (2-5)^2 + (0-0)^2 + (2-0)^2} = \sqrt{17}$$

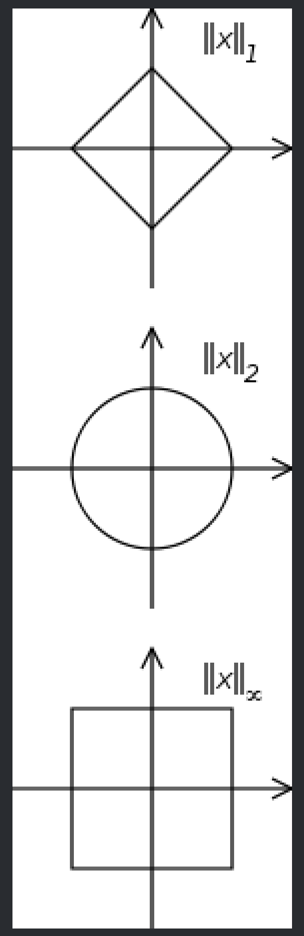
$$dist(D2, Q) = \sqrt{(1-1)^2 + (2-5)^2 + (3-0)^2 + (0-0)^2} = \sqrt{18}$$

$$dist(D3, Q) = \sqrt{(2-1)^2 + (2-5)^2 + (2-0)^2 + (2-0)^2} = \sqrt{18}$$



코사인 유사도?


코사인 유사도 : -1
코사인 유사도 : 0
코사인 유사도 : 1



$$\cos.similarity = \frac{A \cdot B}{\| A \| \; \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} \left(A_i \right)^2} \times \sqrt{\sum_{i=1}^{n} \left(B_i \right)^2}}$$

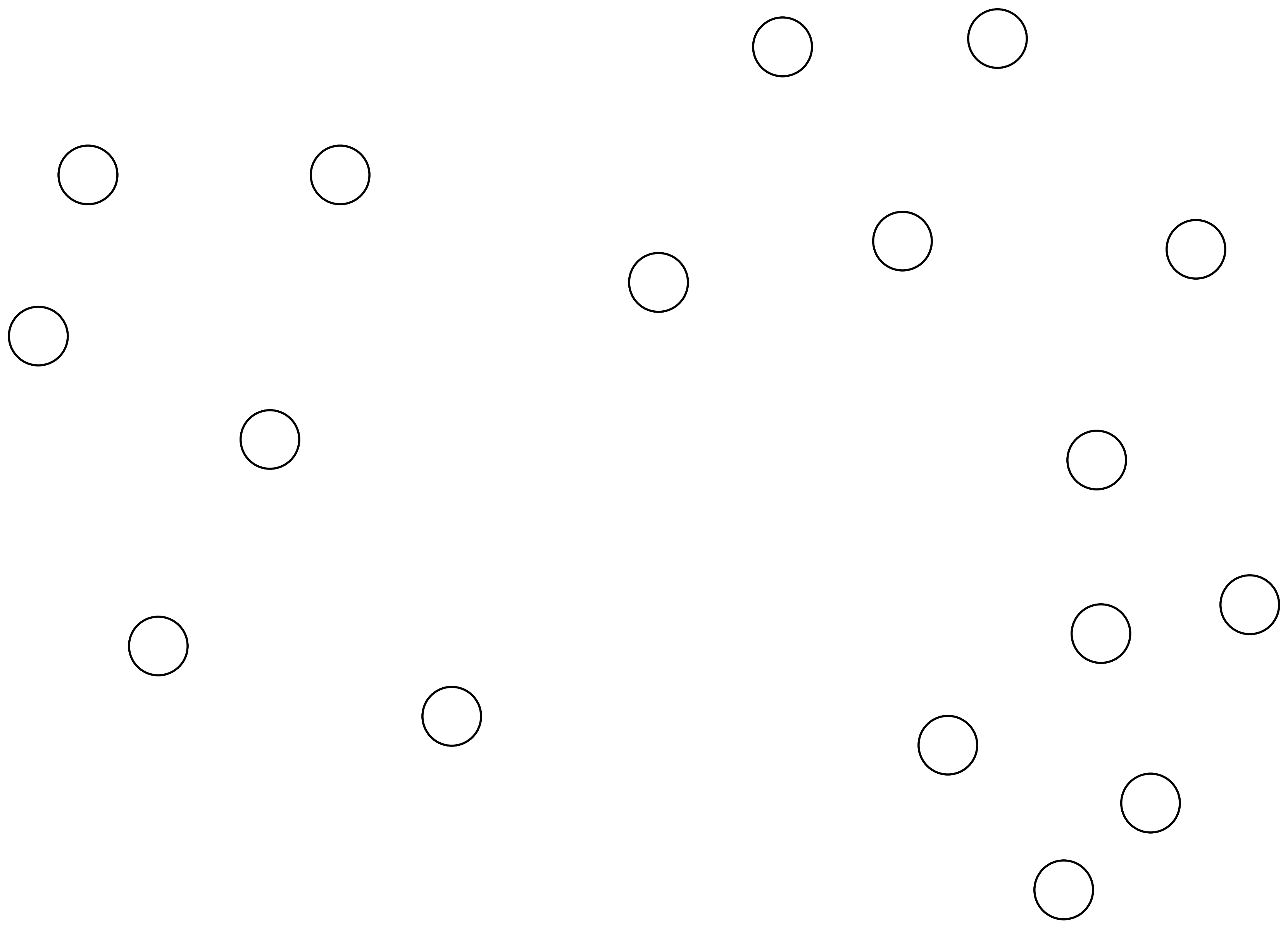
#	A	B	C	D	E	F	D	H
L1	8	8	2	2	0	0	0	0
L2	10	8	3	0	0	0	0	0
L3	0	0	3	2	8	6	6	8
L4	0	0	3	0	8	6	2	8

#	내적	NORM A	NORM B	NORMA * NORM B	Cos.Sim
L1 X L2	8*10 + 8*8 + 2*3 + 2*0 + 0*0 + 0*0 + 0*0 + 0*0	(8*8 + 8*8 + 2*2 + 2*2 + 0*0 + 0*0 + 0*0 + 0*0)^0.5	(10*10 + 8*8 + 3*3 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0)^0.5	11.6619*13.1529	150/153.3878
	150	11.6619	13.1529	153.3878	**0.9779**
L3 X L4	0*0 + 0*0 + 3*3 + 2*0 + 8*8 + 6*6 + 6*2 + 8*8	(0*0 + 0*0 + 3*3 + 2*2 + 8*8 + 6*6 + 6*6 + 8*8)^0.5	(0*0 + 0*0 + 3*3 + 0*0 + 8*8 + 6*6 + 2*2 + 8*8)^0.5	14.5945*13.3041	185/194.1667
	185	14.5945	13.3041	194.1667	**0.9528**
L1 X L3	8*0 + 8*0 + 2*3 + 2*2 + 0*8 + 0*6 + 0*6 + 0*8	(8*8 + 8*8 + 2*2 + 2*2 + 0*0 + 0*0 + 0*0 + 0*0)^0.5	(0*0 + 0*0 + 3*3 + 2*2 + 8*8 + 6*6 + 6*6 + 8*8)^0.5	11.6619*14.5945	10/170.1996
	10	11.6619	14.5945	170.1996	**0.0588**

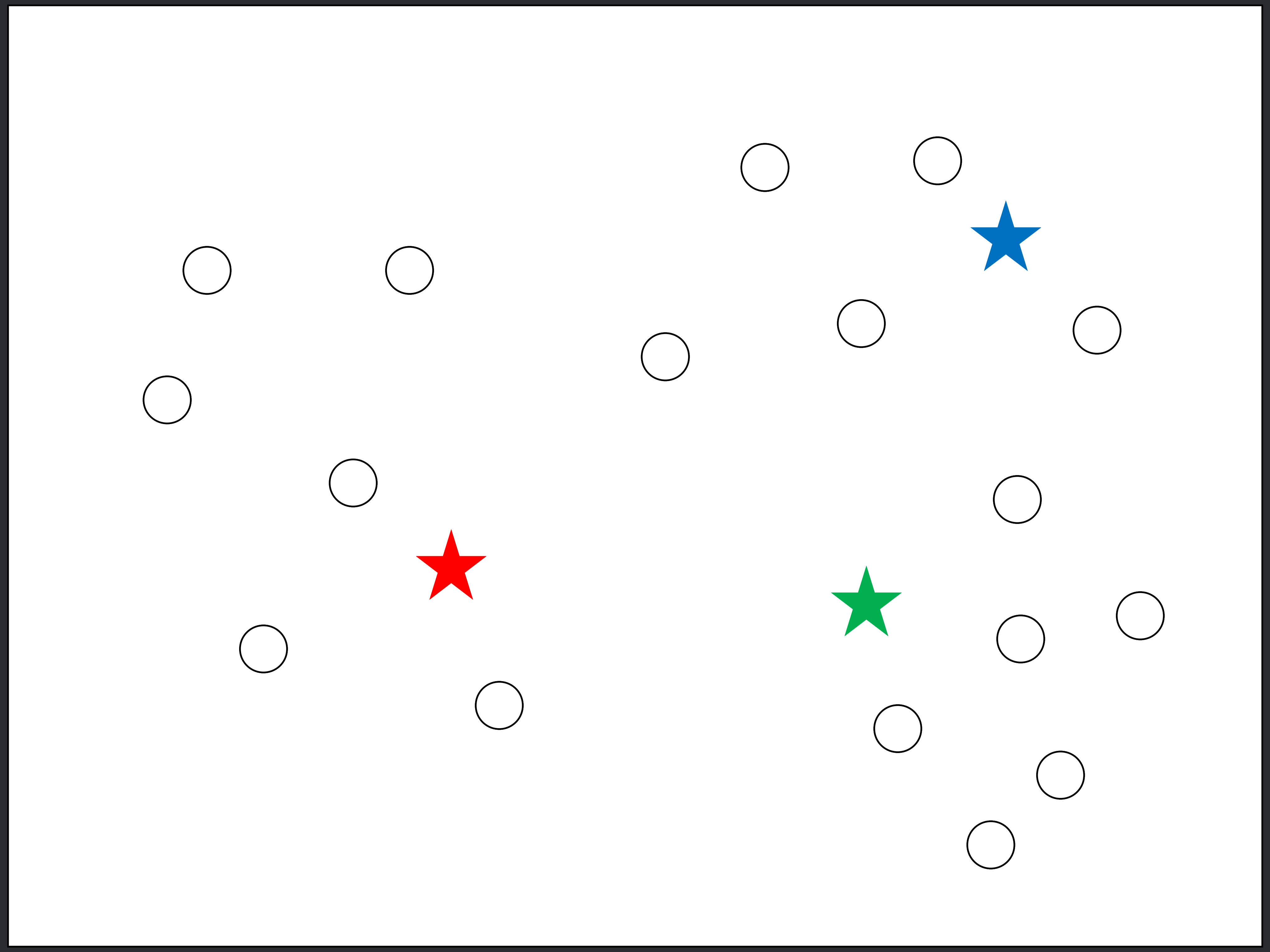
$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \mathbf{c}_i||^2$$

Algorithm 1 Basic K-means Algorithm.

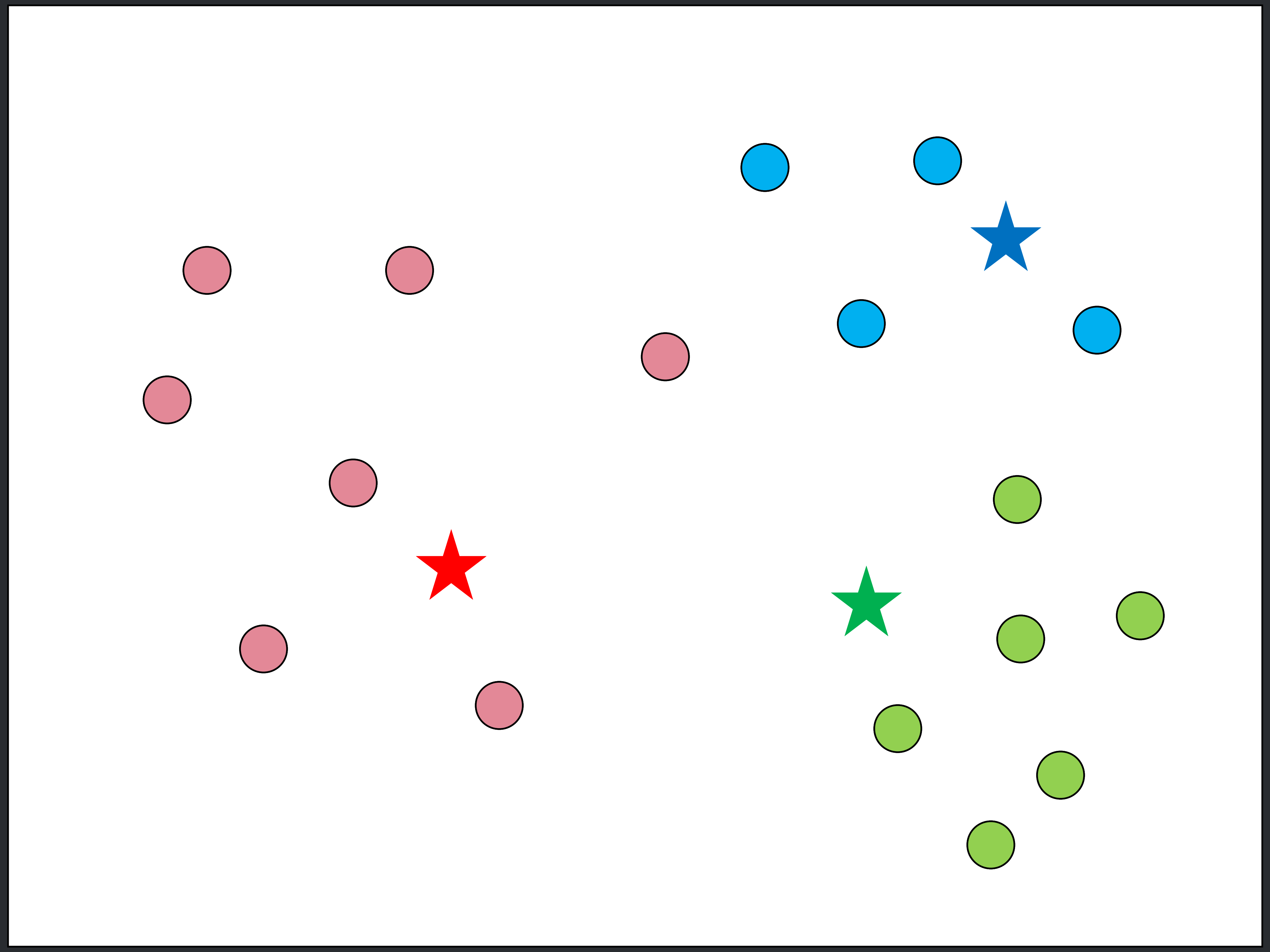
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change



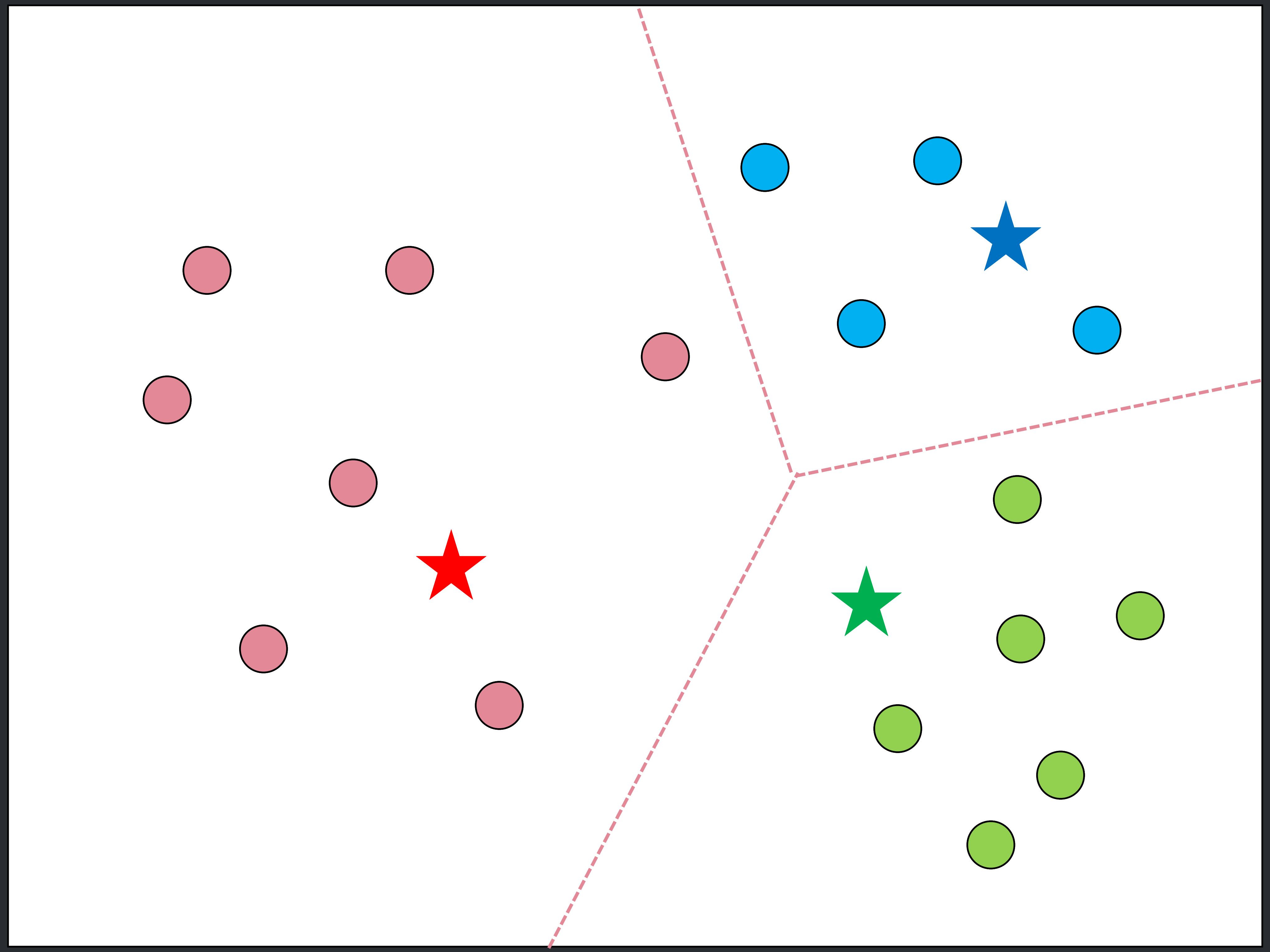
k = 3



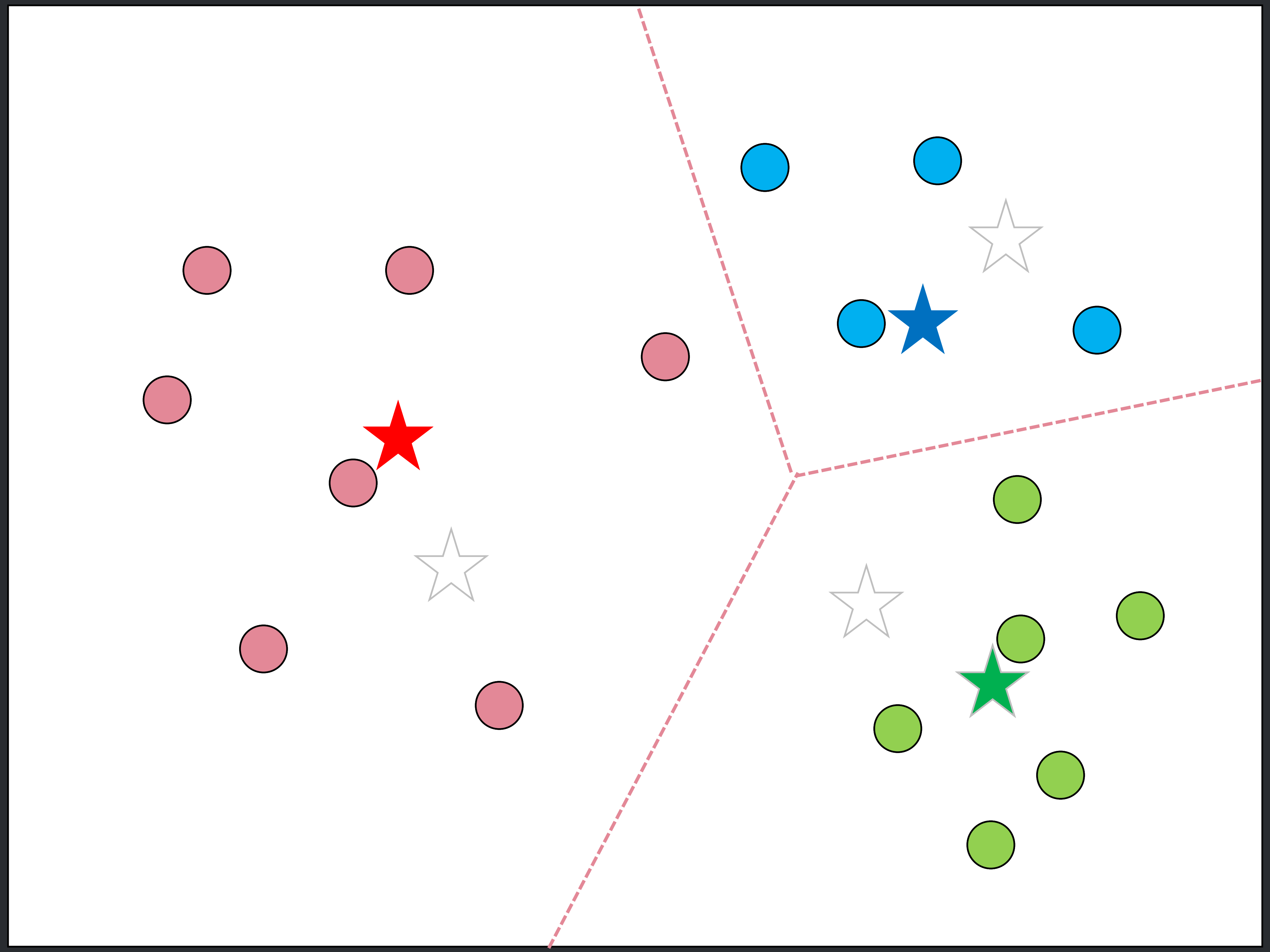
k = 3



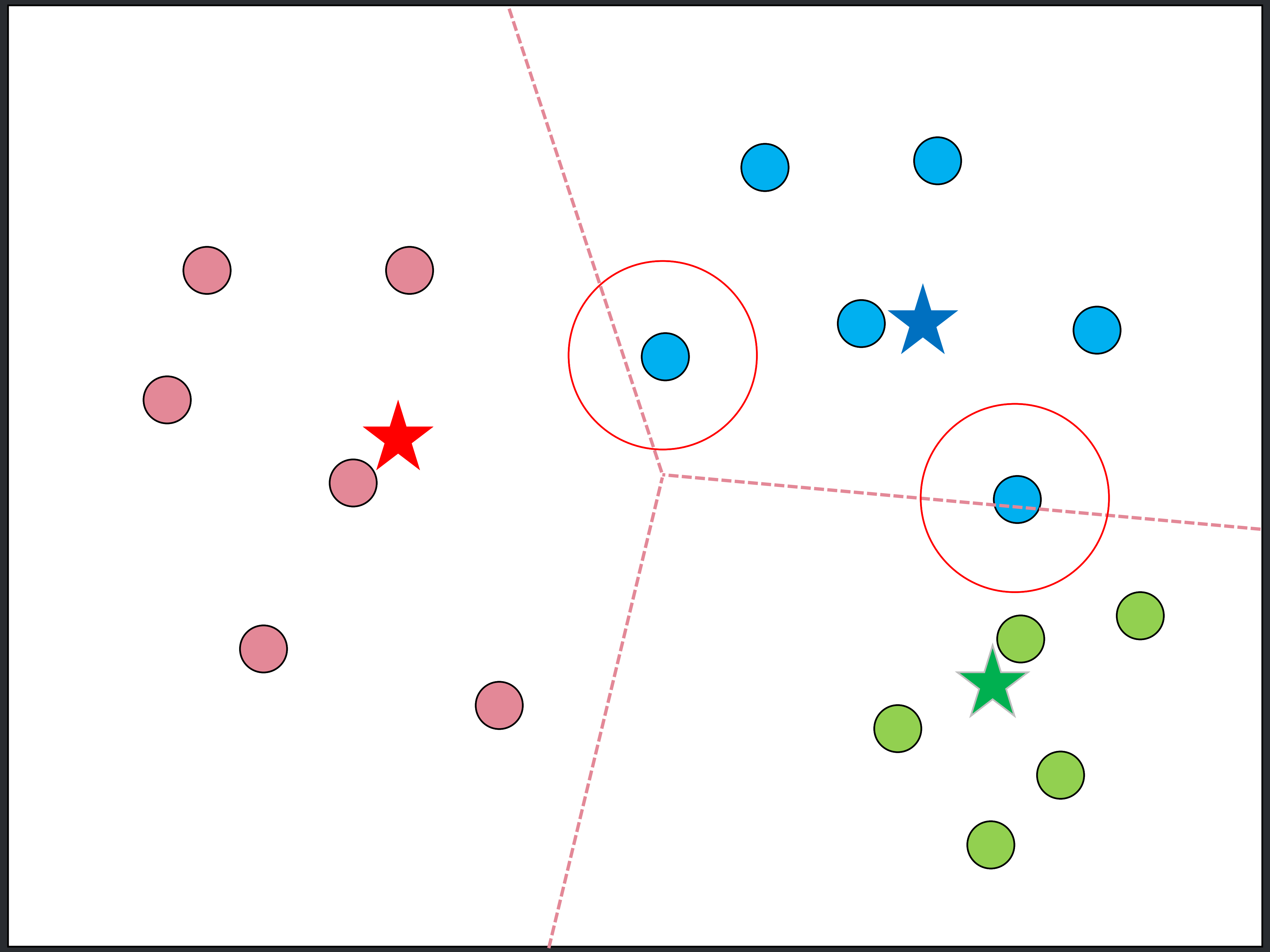
k = 3



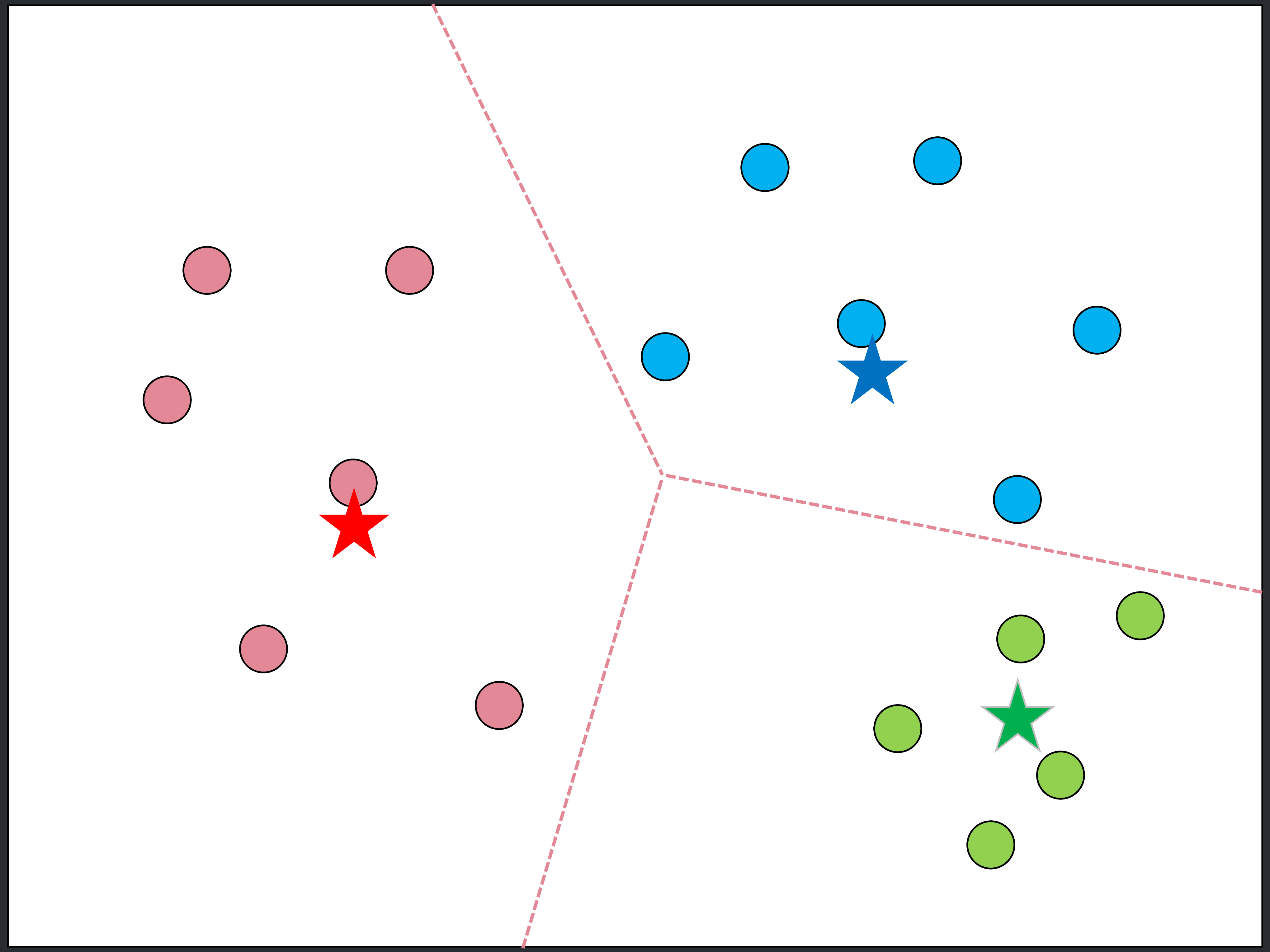
k = 3



k = 3



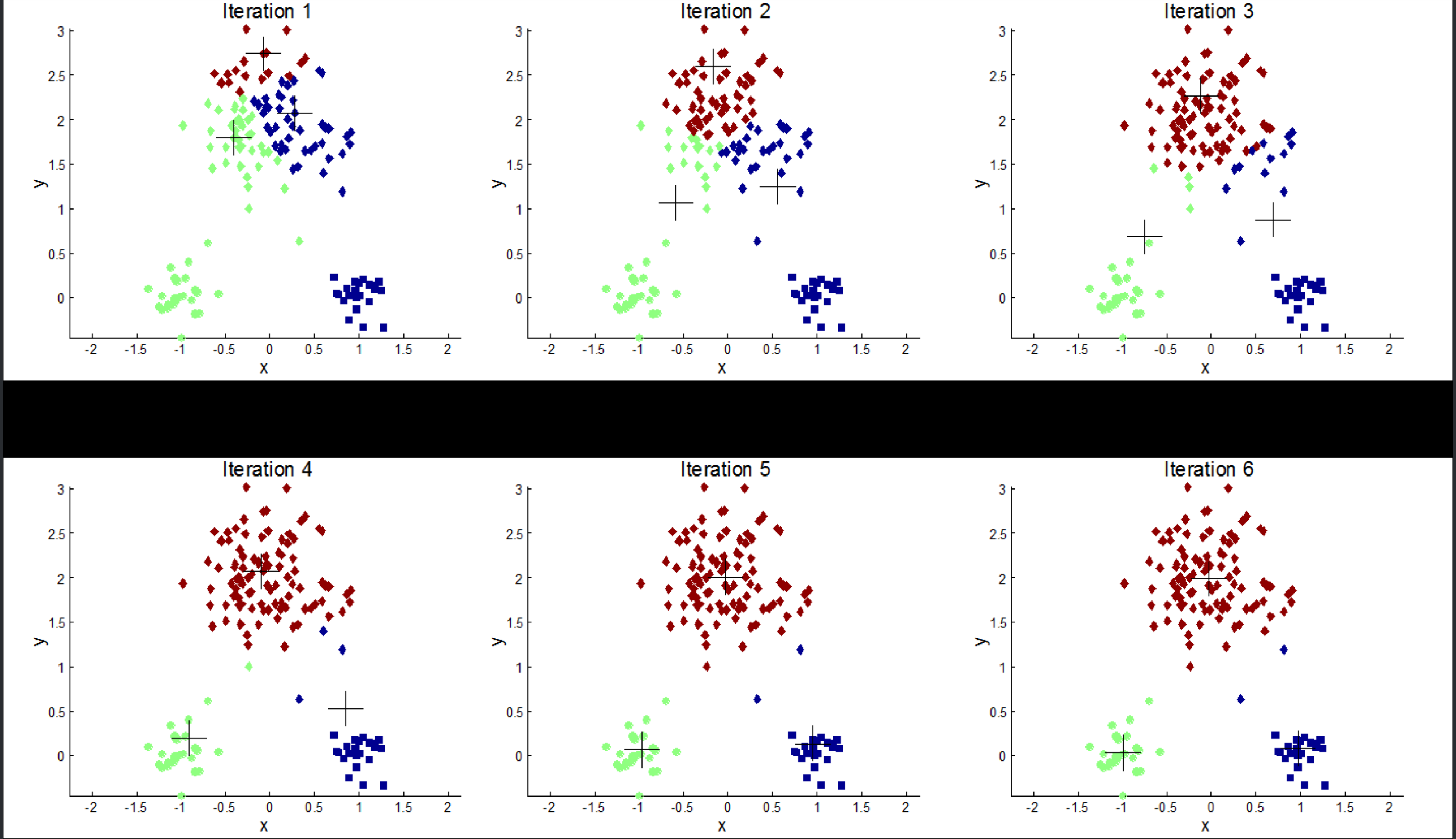
k = 3



epoch =2

Weak points	
1	Sensitive results from Initial points
2	Ball-shaped clusters
3	Sensitive to noise points

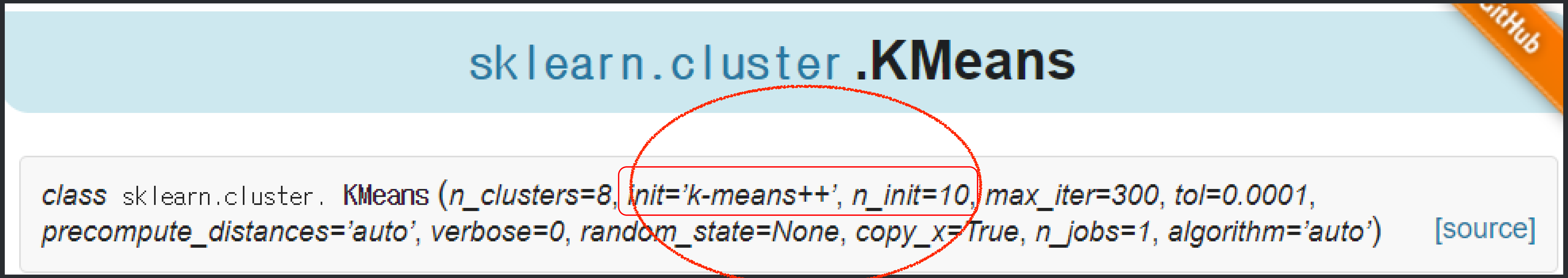
1. Sensitive results from Initial points


Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y


Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
x
y
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
0 0.5 1 1.5 2 2.5 3

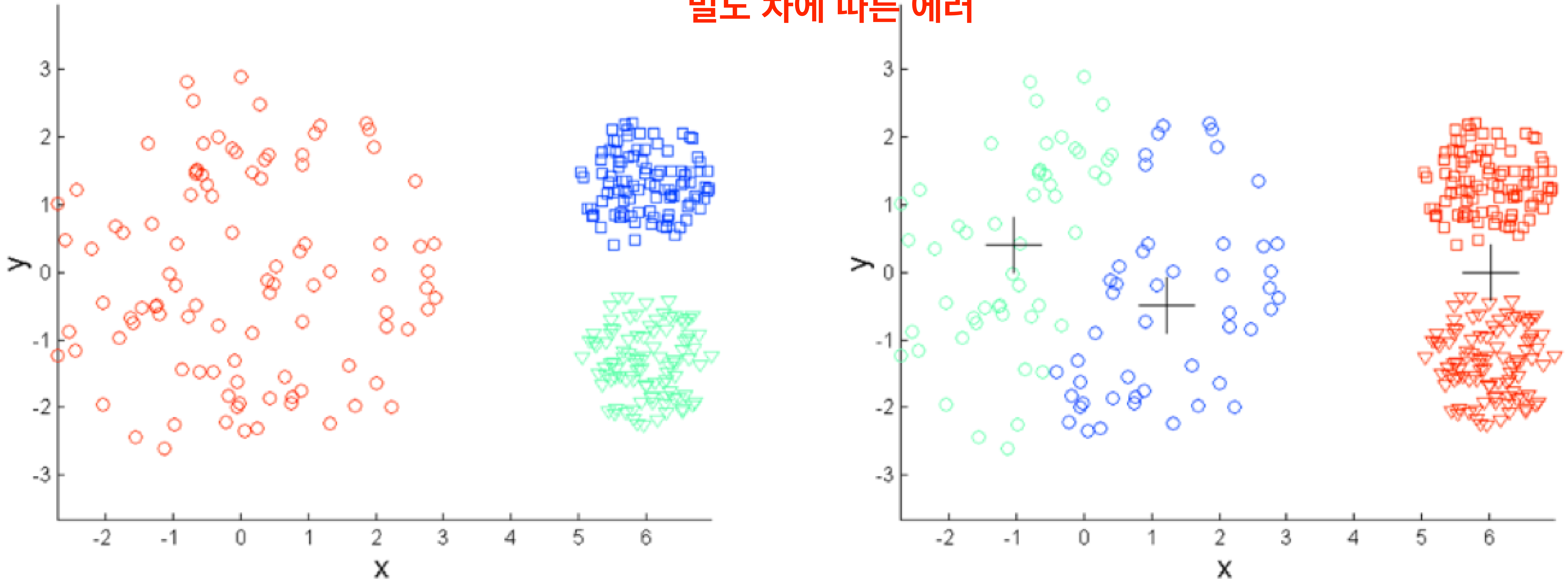
1. 해결법

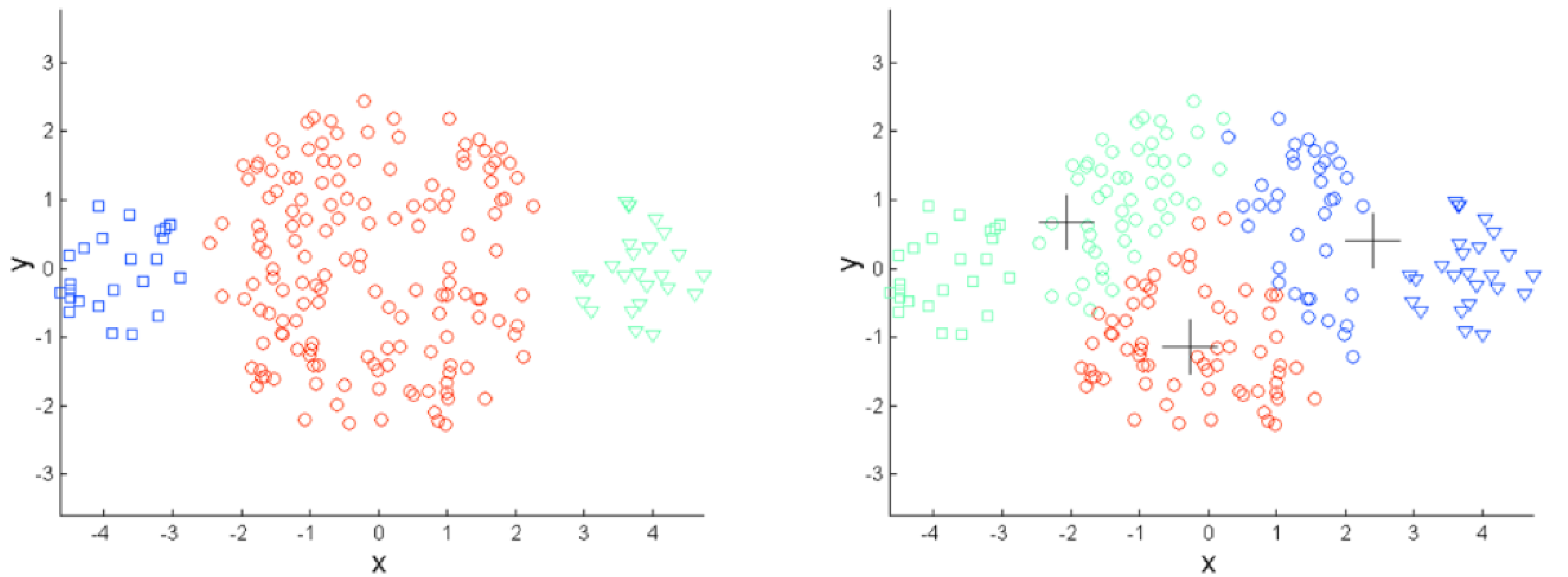
1	n_init
2	init='k-means++'



2. Ball-shaped clusters

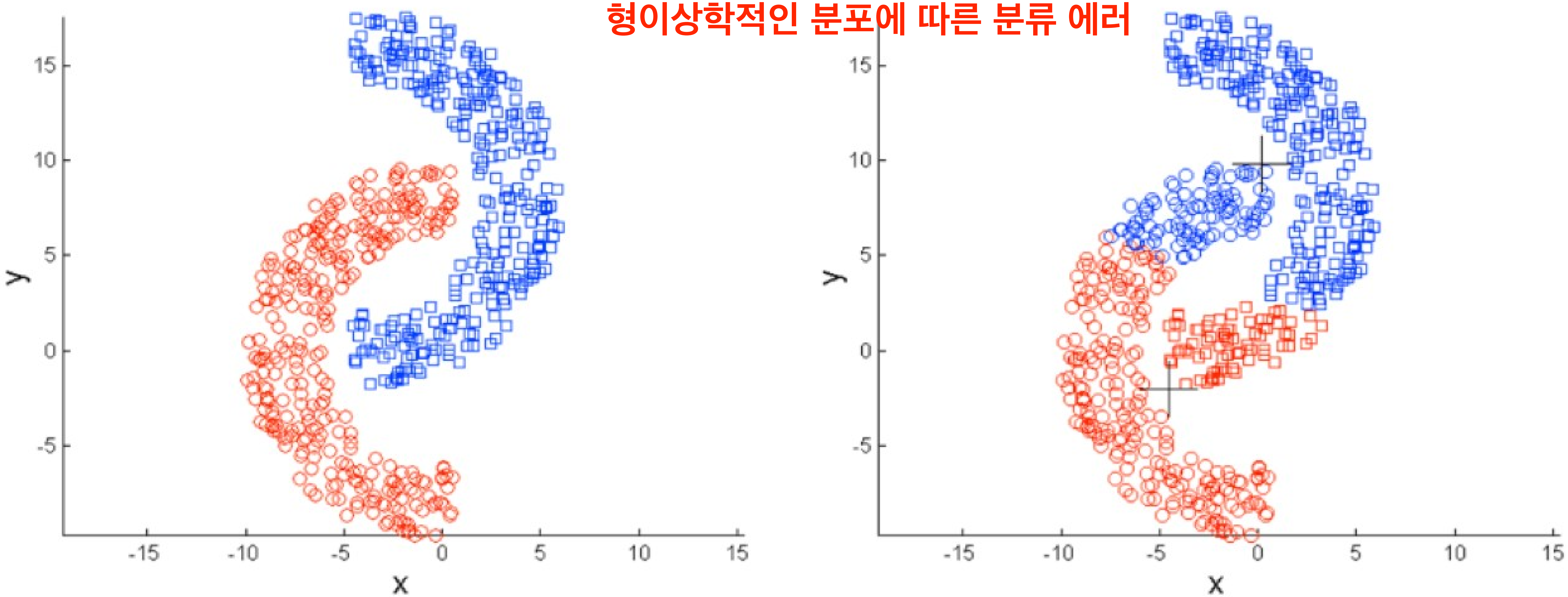
밀도 차에 따른 에러



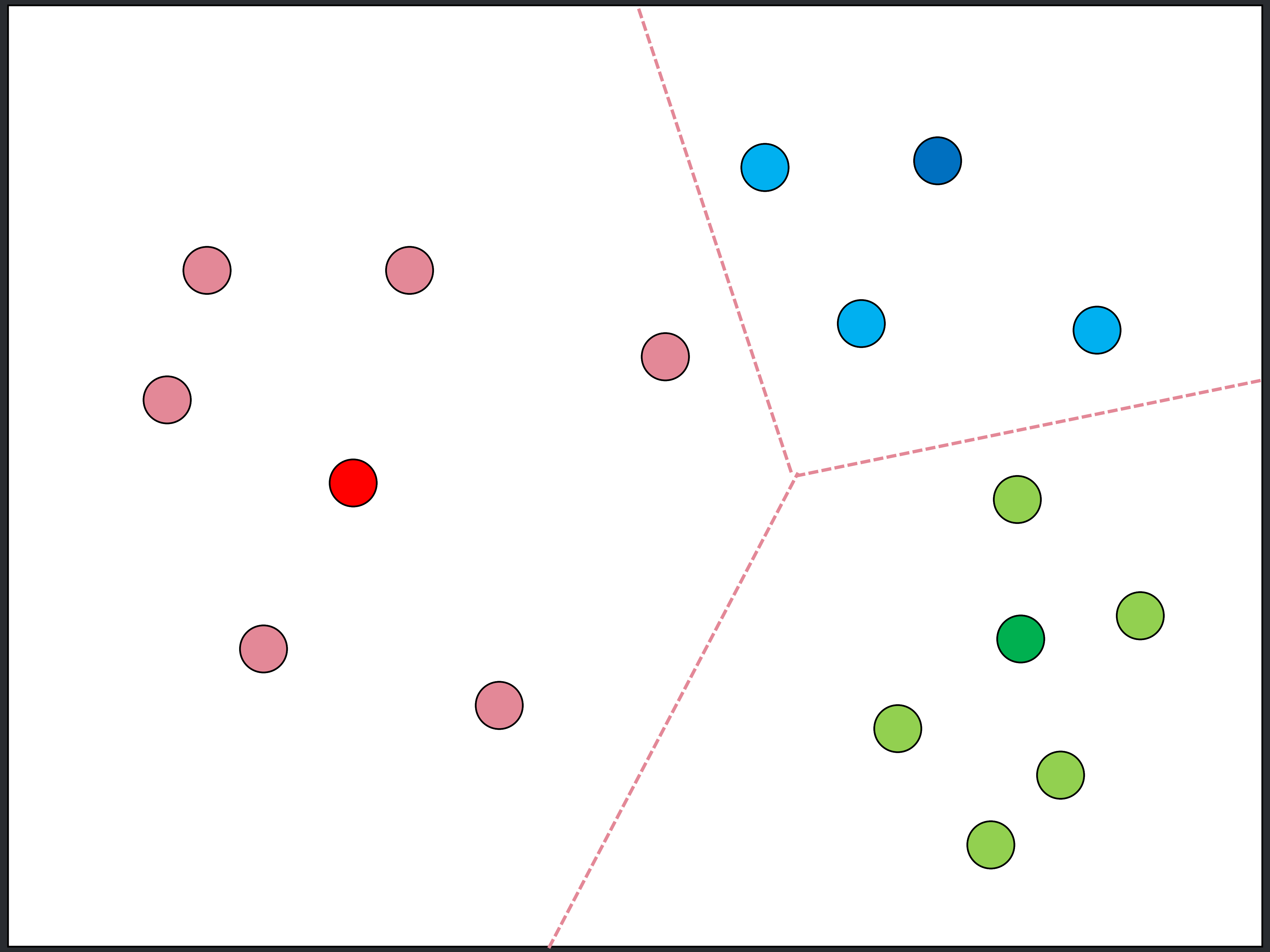


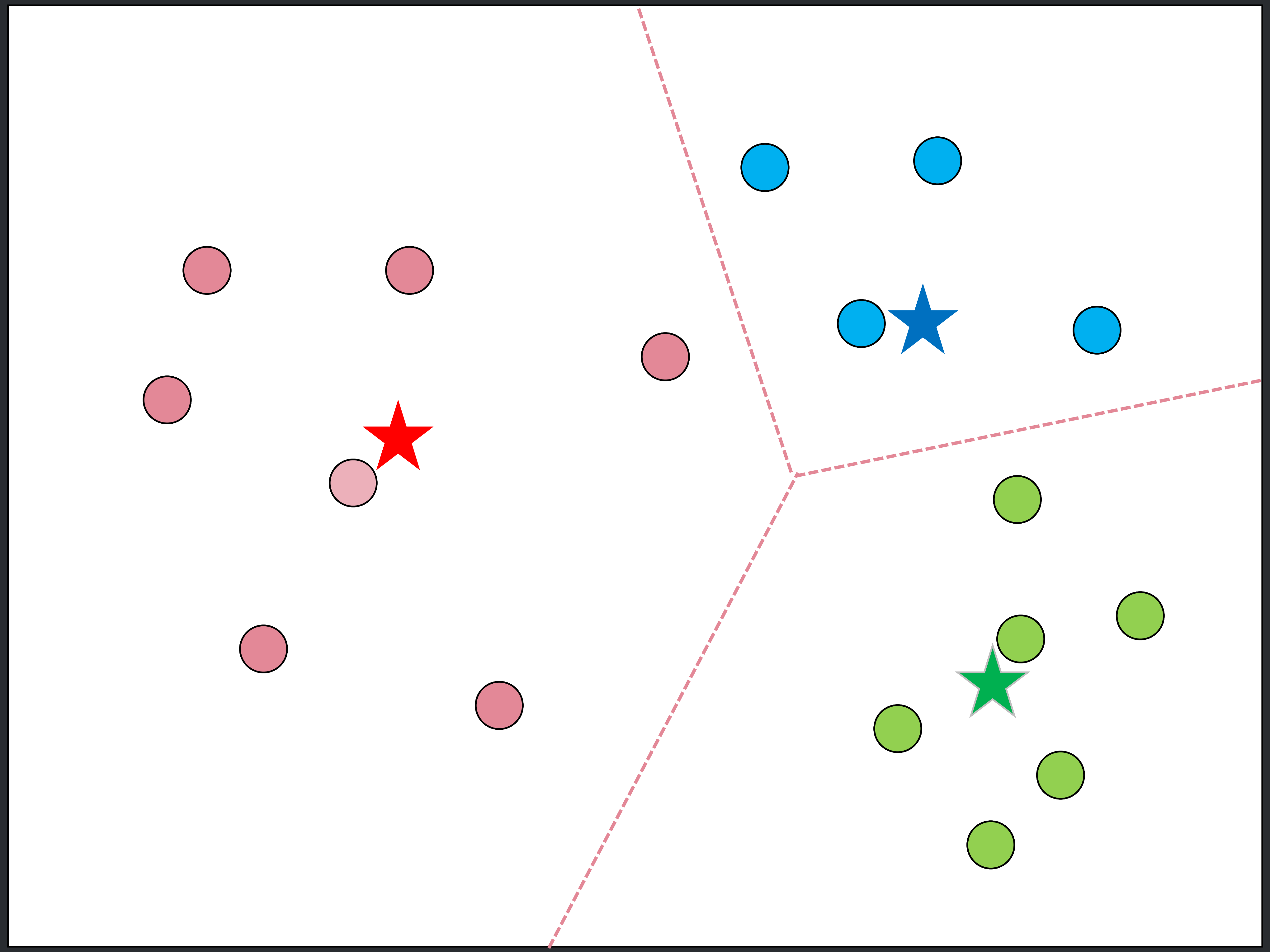
y
x
y
x

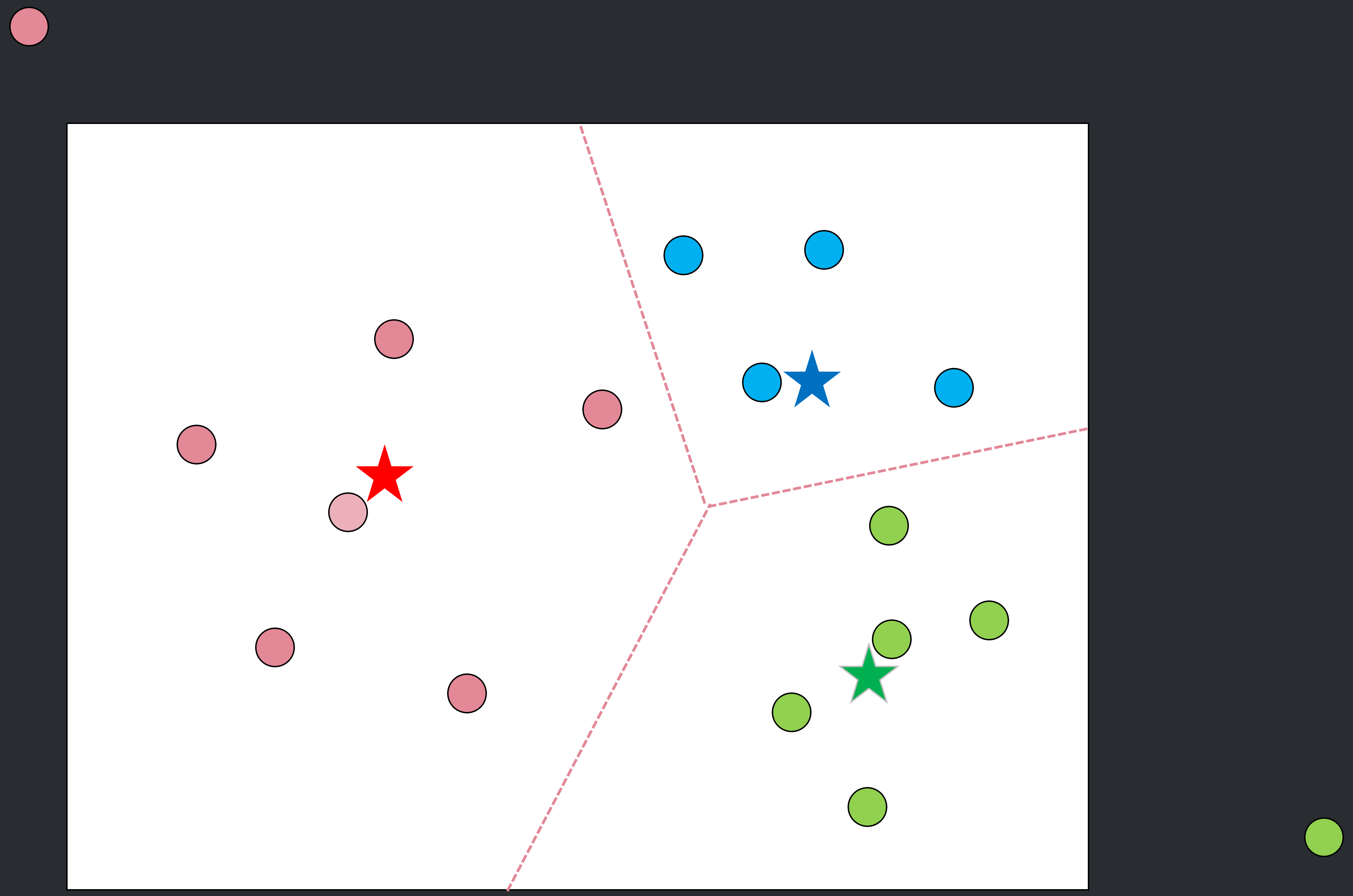
형이상학적인 분포에 따른 분류 에러

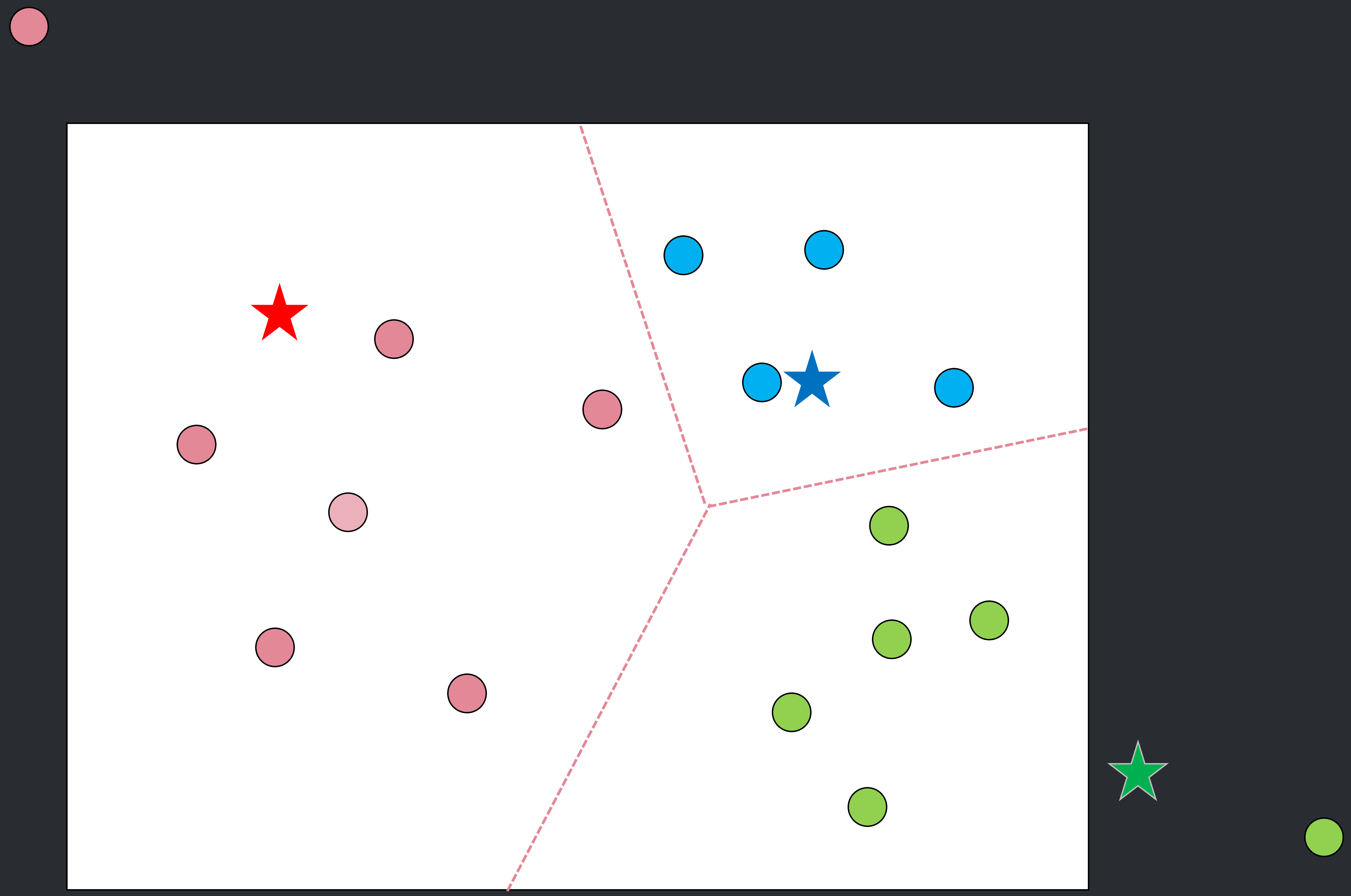


3. sensitive to noise points

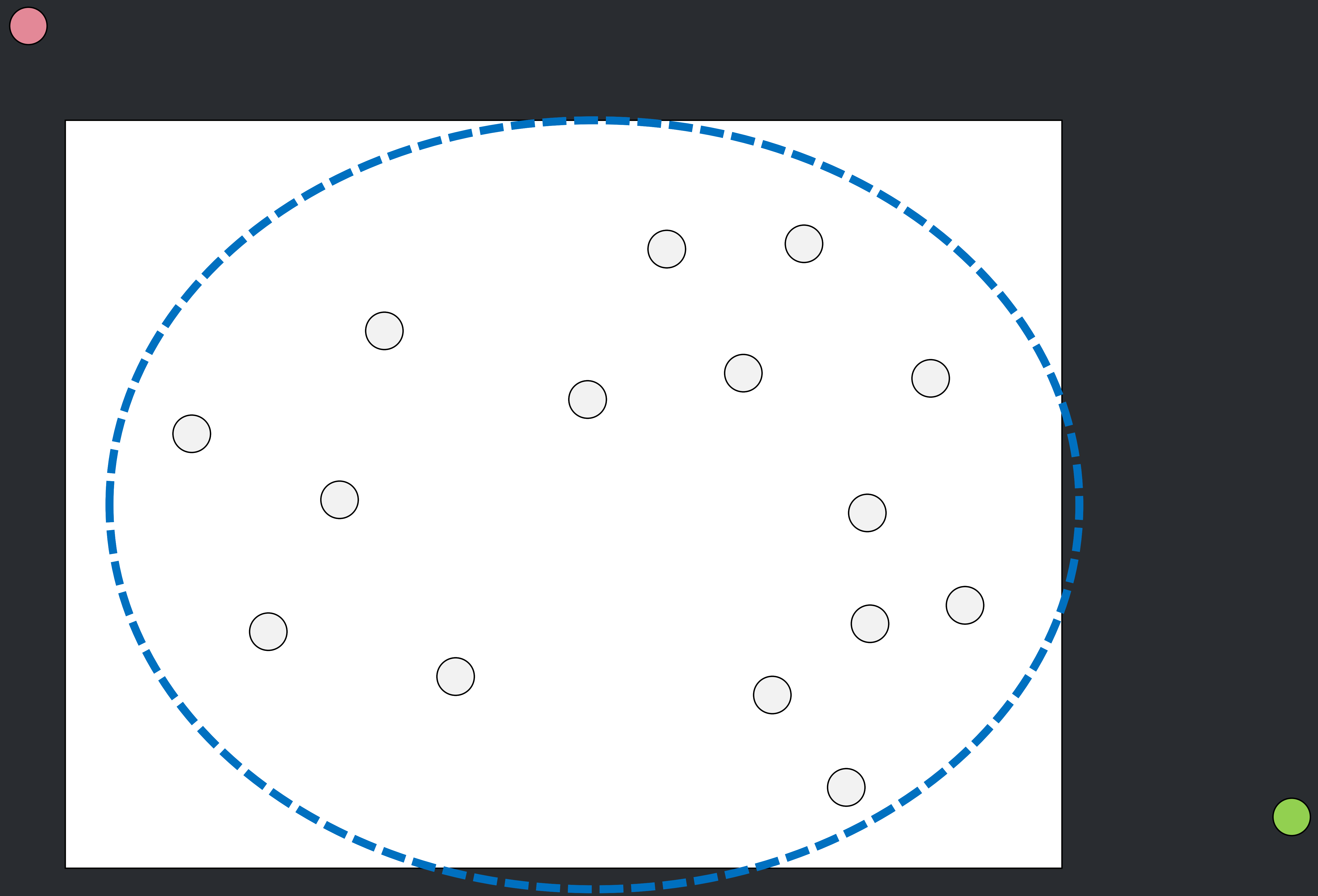


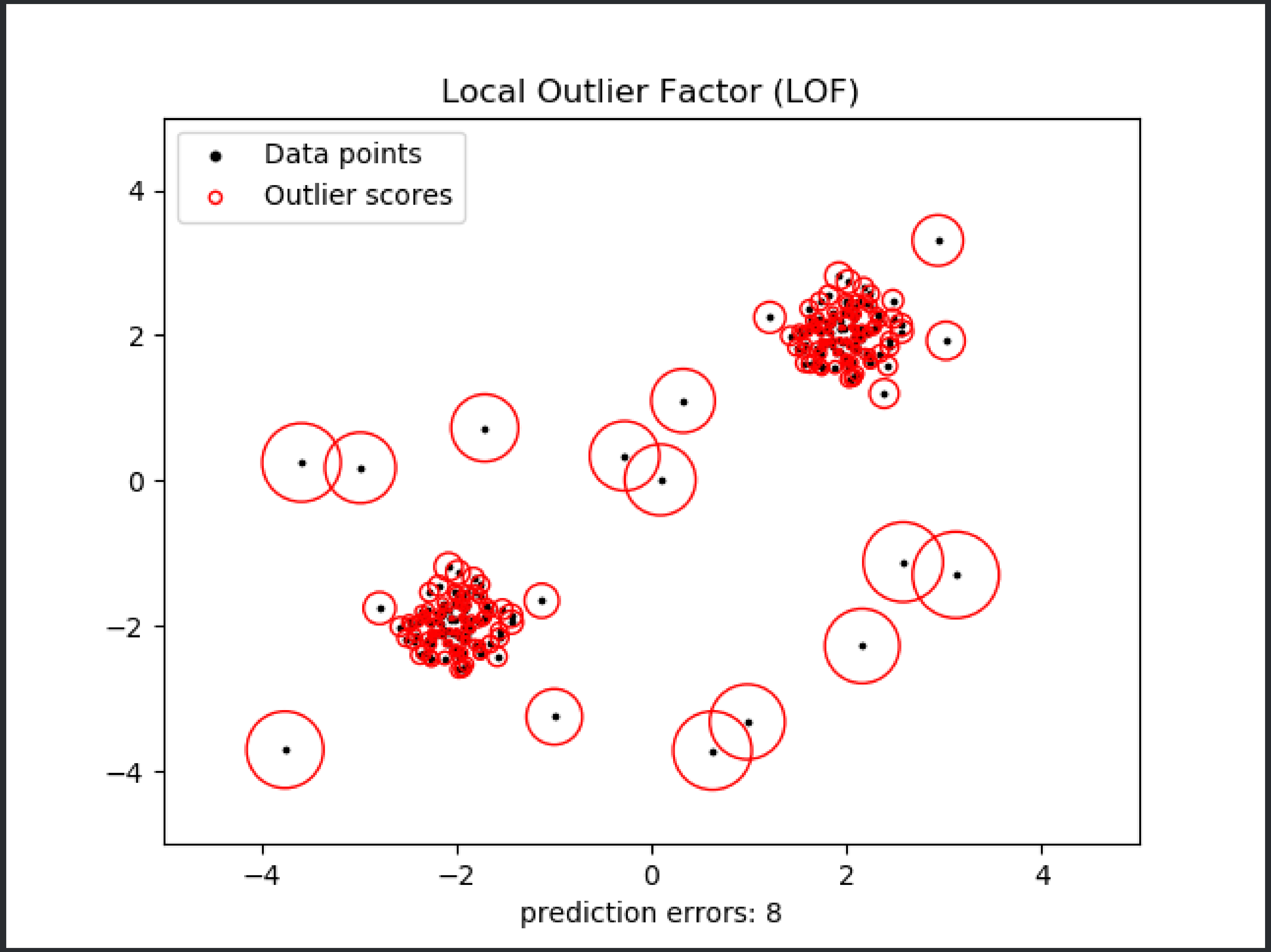






3. 해결법




Local Outlier Factor (LOF)
Data points
Outlier scores
4
2
0
−2
−4
−4
−2
0
2
4
prediction errors: 8

예측 전에 LOF 처리

k-means clustering 장단점

장점	계산이 쉽다. 다른 군집화 알고리즘에 비해 복잡도가 낮다
	구현이 쉽고 다양한 언어와 플랫폼에서 제공되는 알고리즘
단점	노이즈에 매우 민감
	군집 개수를 사전에 지정
	앞의 몇 가지 상황에서는 최적의 군집 구조를 찾기 어려움

Evaluation metrics for clustering

Sadly, there is no good way

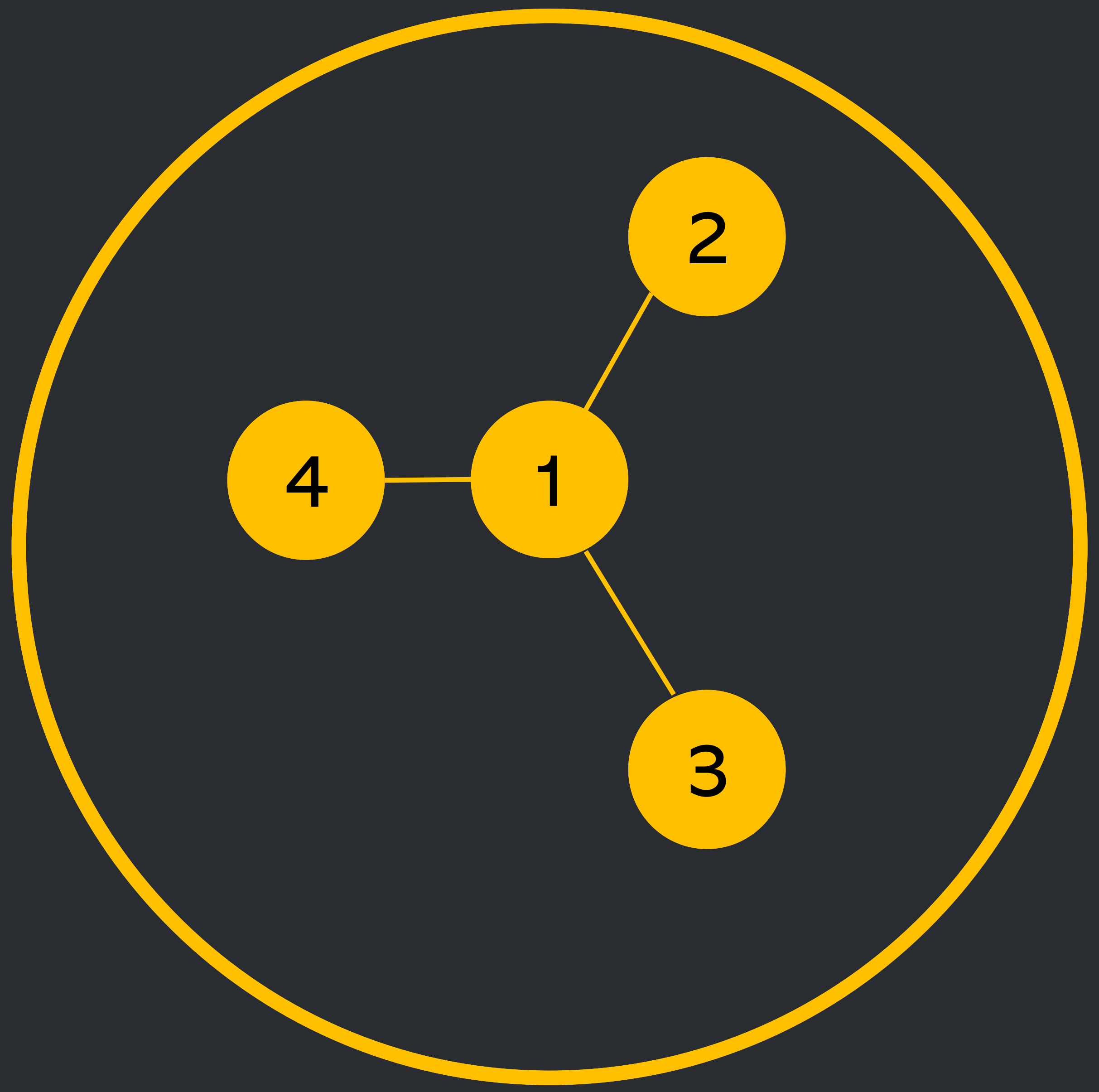
So,

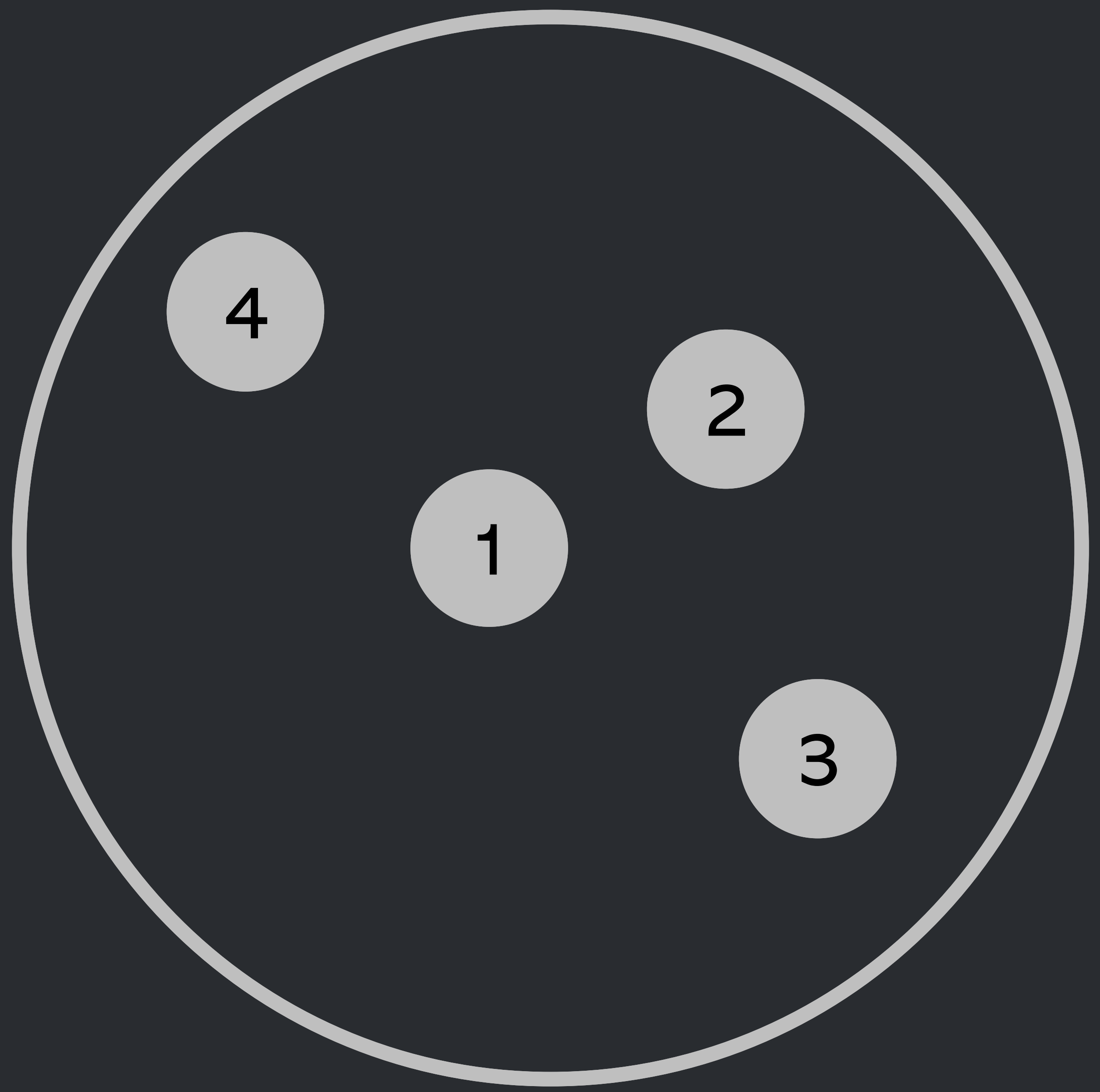
Sum of squared distance for each point to it’s assigned centroid

Silhouette score

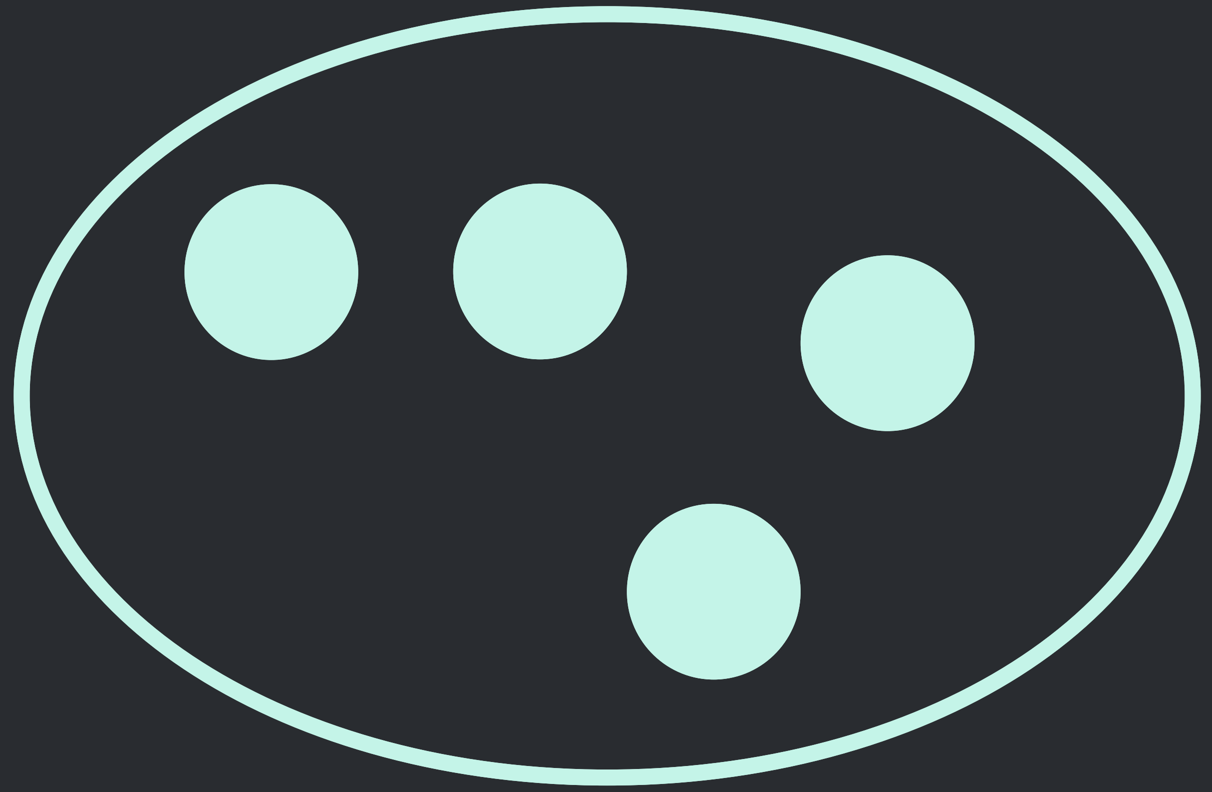
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(1)

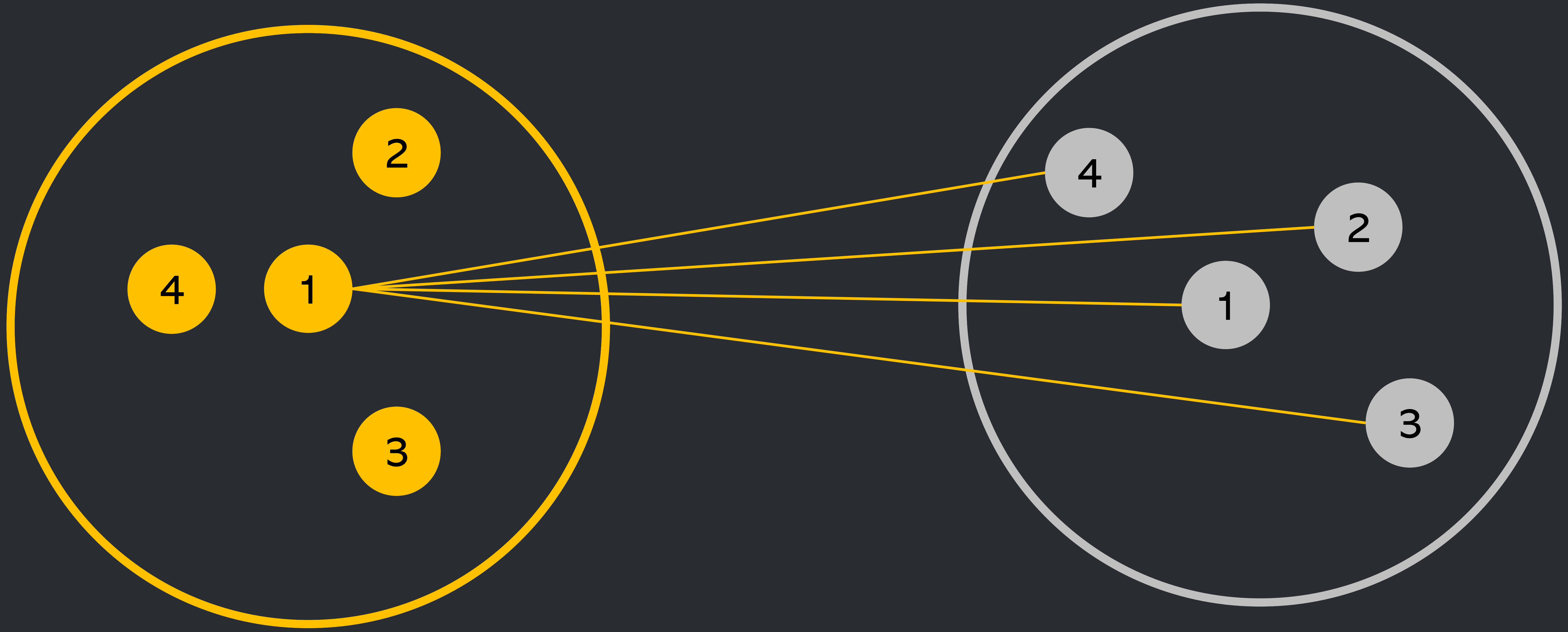




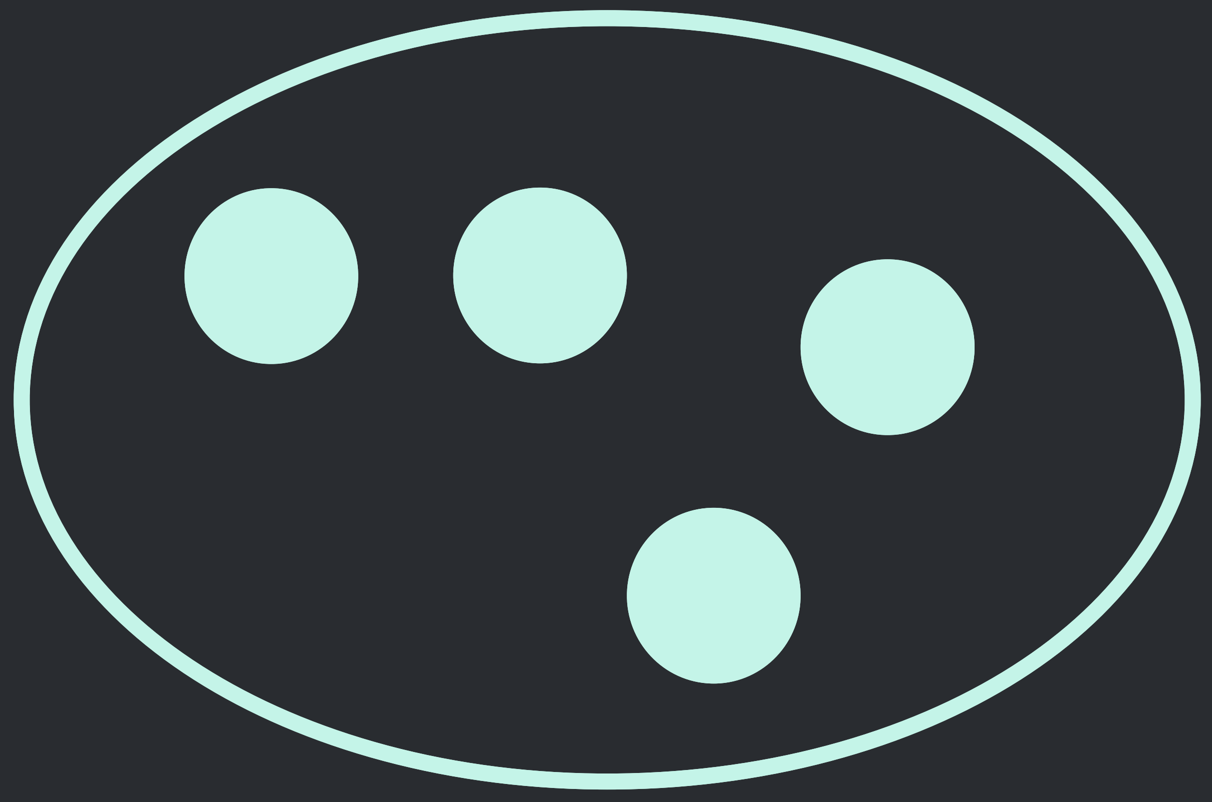
a(1) = a(1)에서 각 점들 간의 거리를 평균

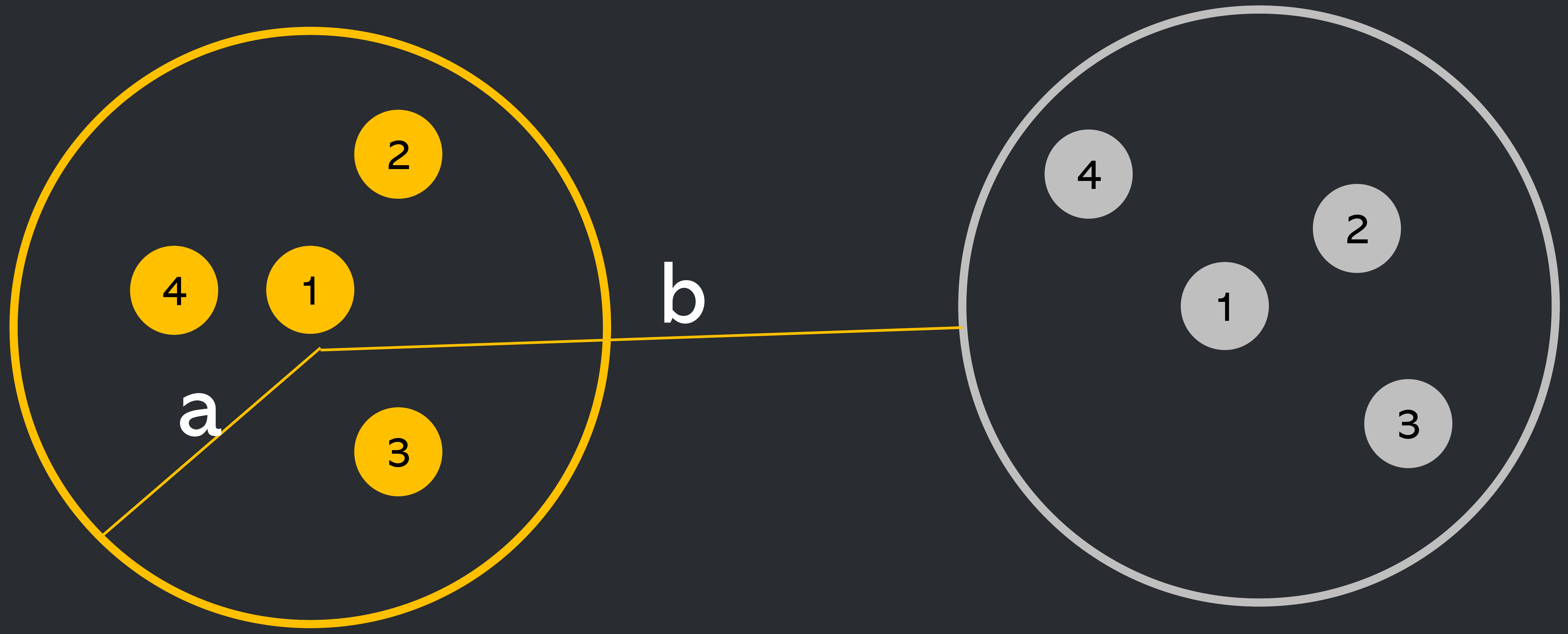


b(1)



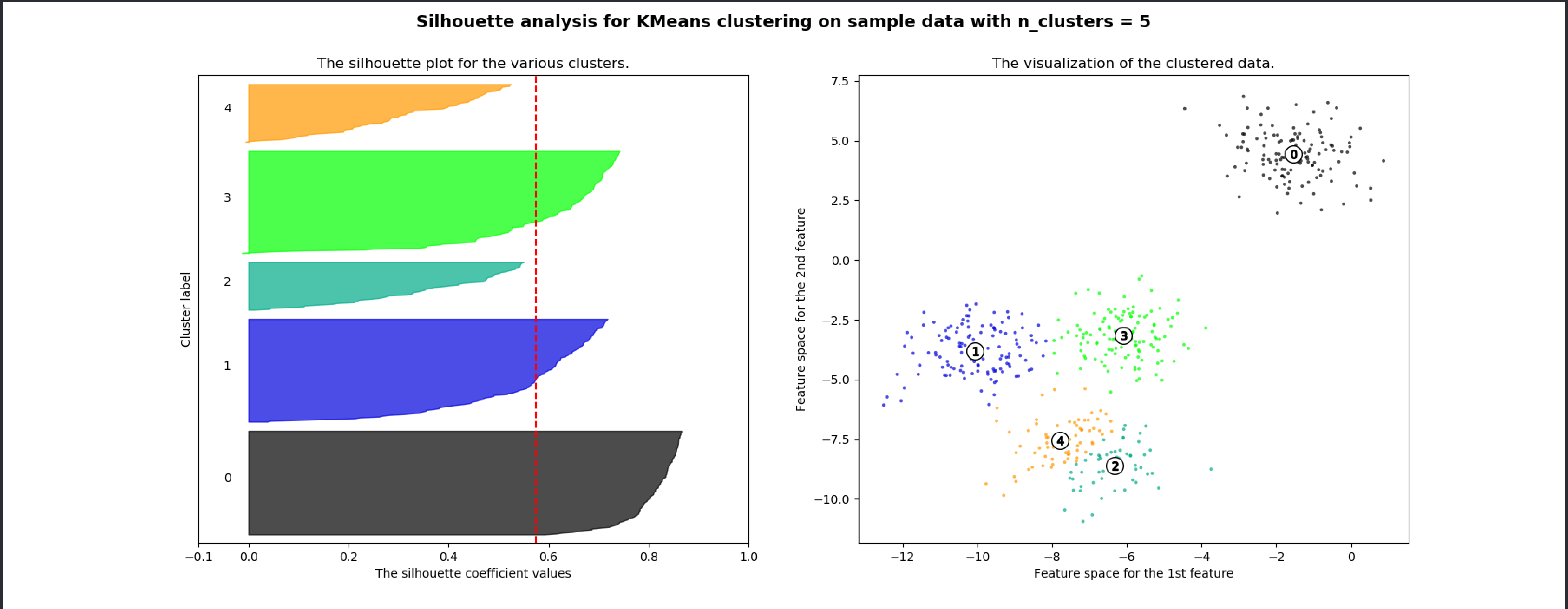
b(1) = b(1)에서 각 점들 간의 거리를 평균




2
4
1
a
3
b
4
2
1
3

$$s = \frac{b - a}{\max(a, b)}$$

$$-1 \leq s \leq 1$$


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12
−10
−8
−6
−4
−2
0
Feature space for the 1st feature