Machine Learning

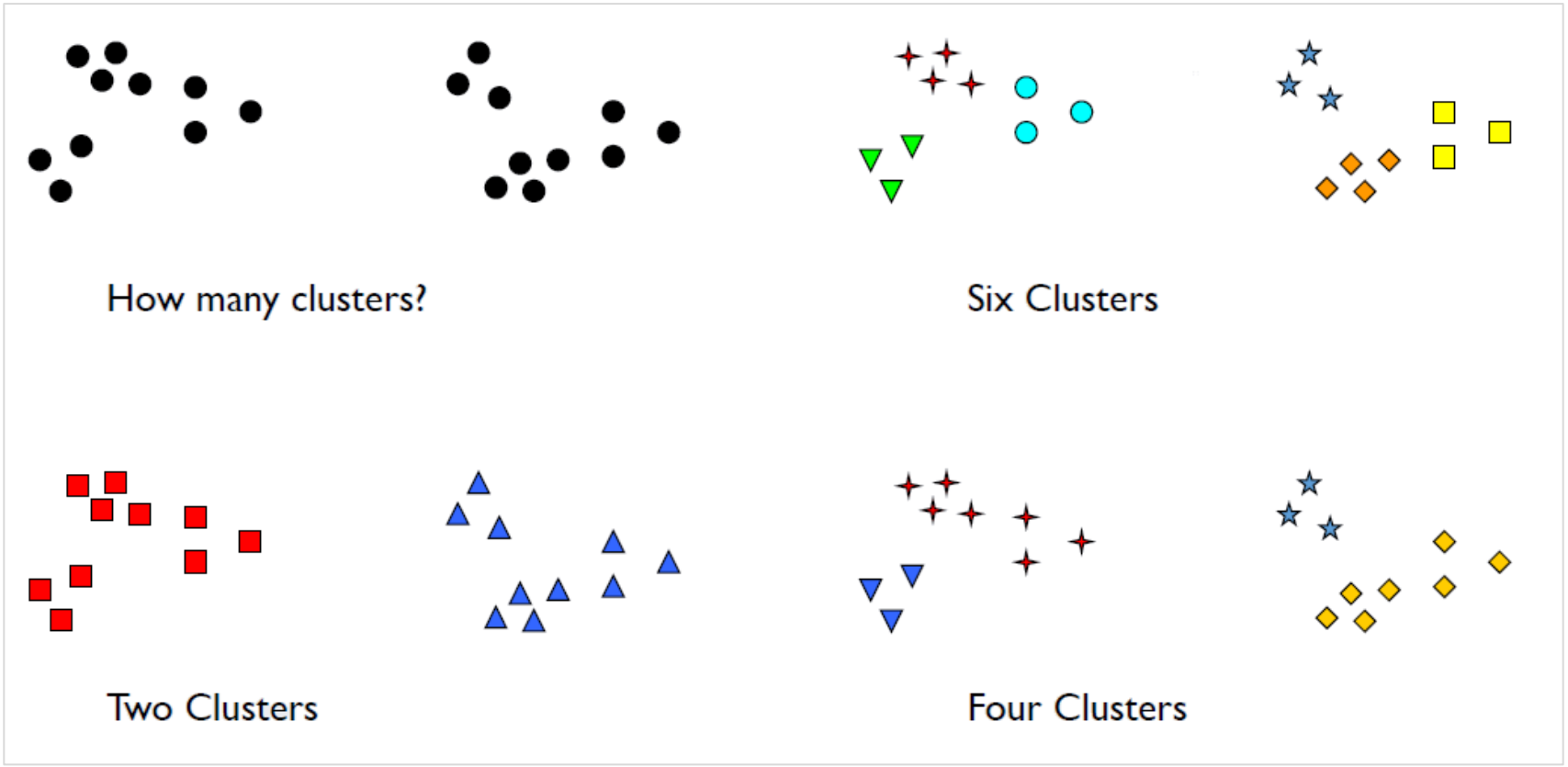
K-Means Algorithm

김선녕(ksycafe@gmail.com)

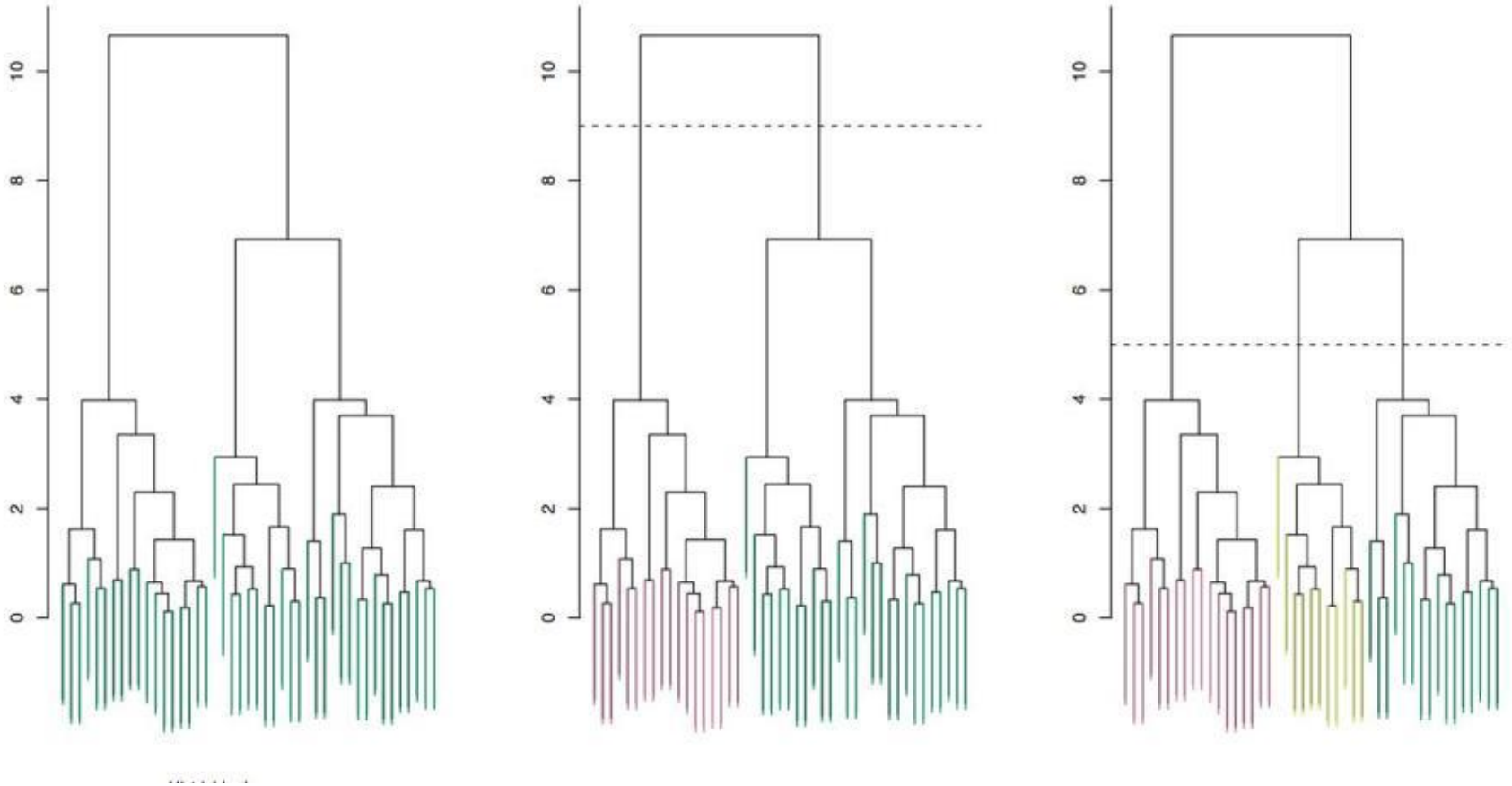
군집화(Clustering)

- 지식(정보)가 없는 상태에서 유사(similar)한 데이터를 모아 같은 그룹으로 묶는 일
  - 군집간 분산 최대화, 군집 내 분산 최소화
  - 유사도(similarity) 계산
- 레이블이 없는 대표적인 **비지도학습(unsupervised learning)**
  - 분류(classification) : 지도학습(supervised learning)
- Hierarchical Clustering 과 Partitional Clustering 가 있다.

분할 군집화(Partitional Clustering)

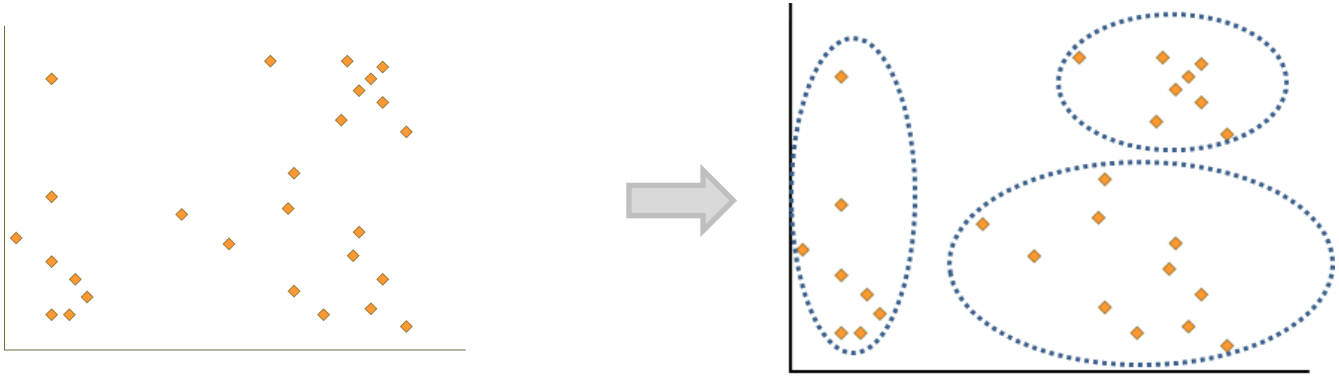


계층형 군집화(Hierarchical Clustering)



군집화의 응용

군집화의 응용



정지영상을 비슷한 것끼리 군집화

영상 분할(segmentation)

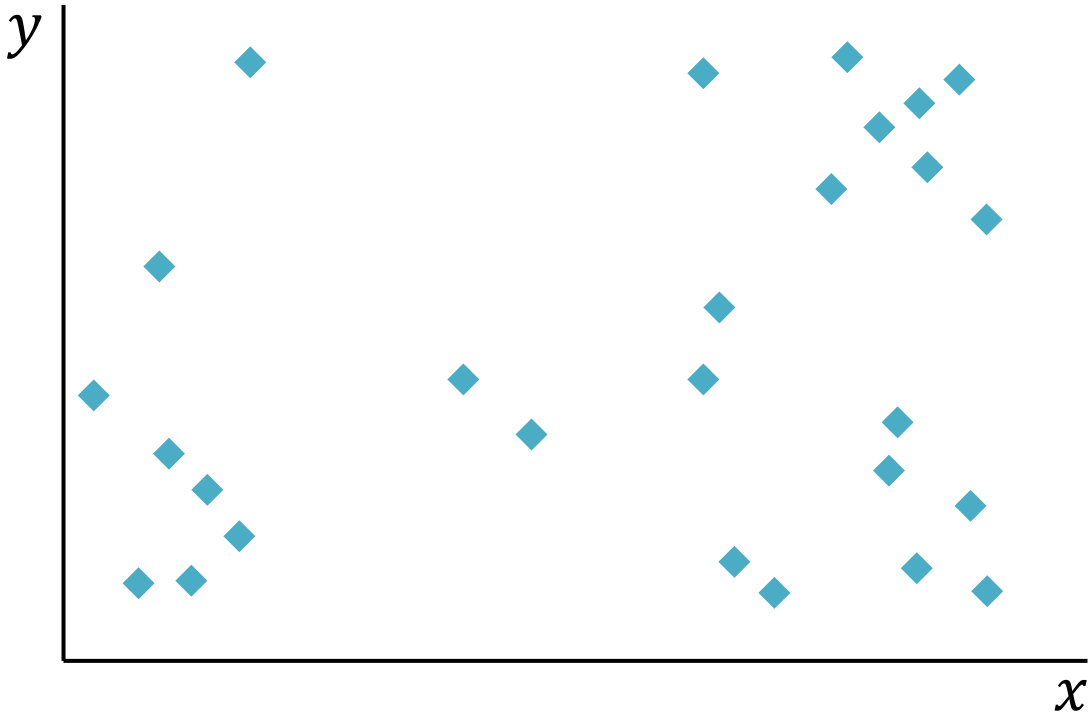
k-평균 알고리즘(k-means algorithm)

입력 : 훈련집합 $X = \{x_1, x_2, \cdots, x_n\}$, 군집의 개수 $\boldsymbol{k}$
출력 : 군집집합 $C = \{c_1, c_2, \cdots, c_{\boldsymbol{k}}\}$

[pseudo code]

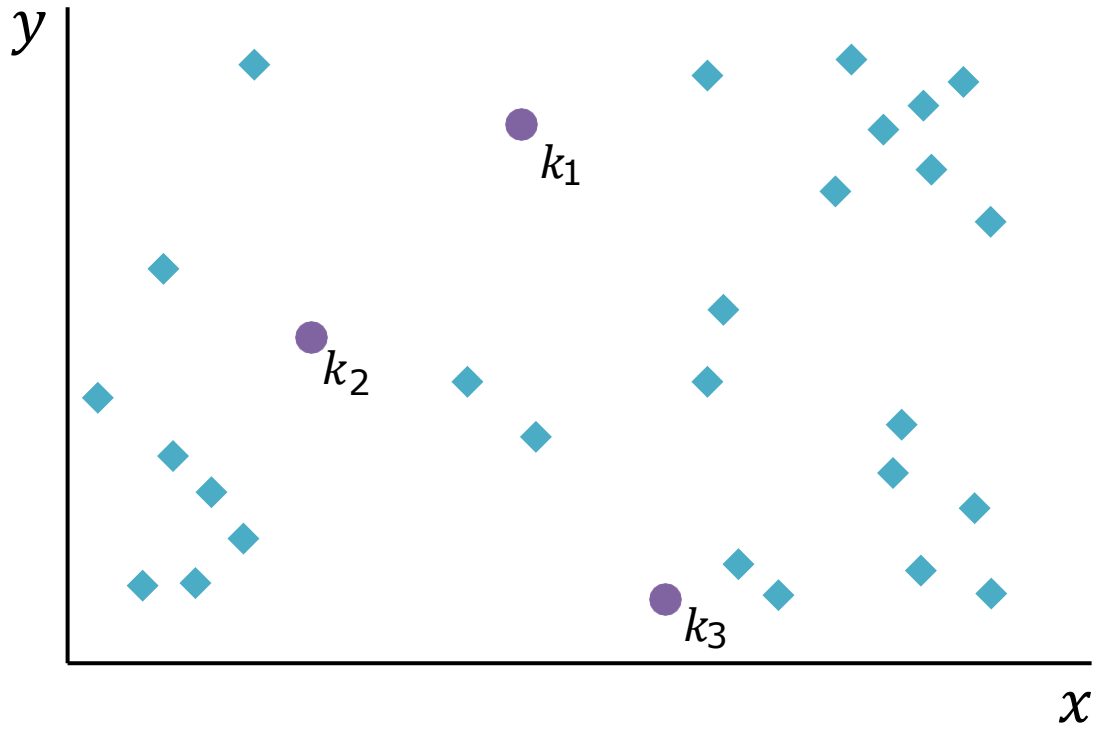
1: k개의 점들을 초기 중심점($centroid$)으로 선택한다.
2: ***repeat***
3: 각 점을 가장 가까운 중심점($centroid$)에 할당하여 k개의 군집을 형성한다.
4: 각 군집의 중심점을 다시 계산한다.
5: ***until*** 중심점이 바뀌지 않을 때까지

Suppose we wish to cluster these items

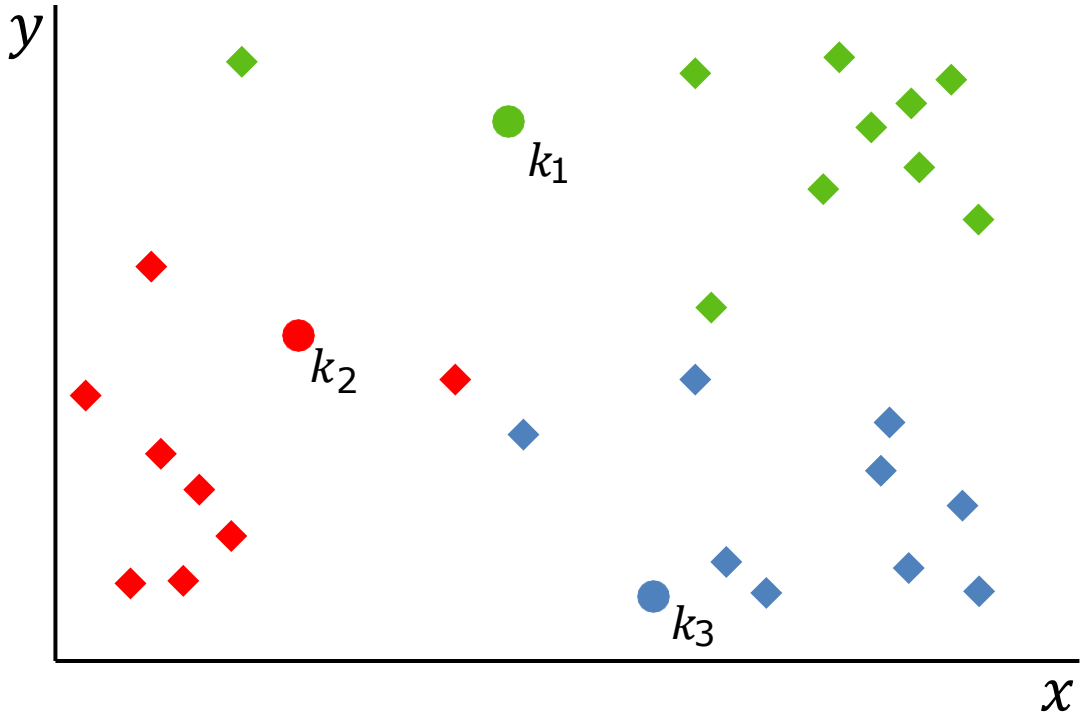


https://www3.nd.edu/~rjohns15/cse40647.sp14/www/home.php

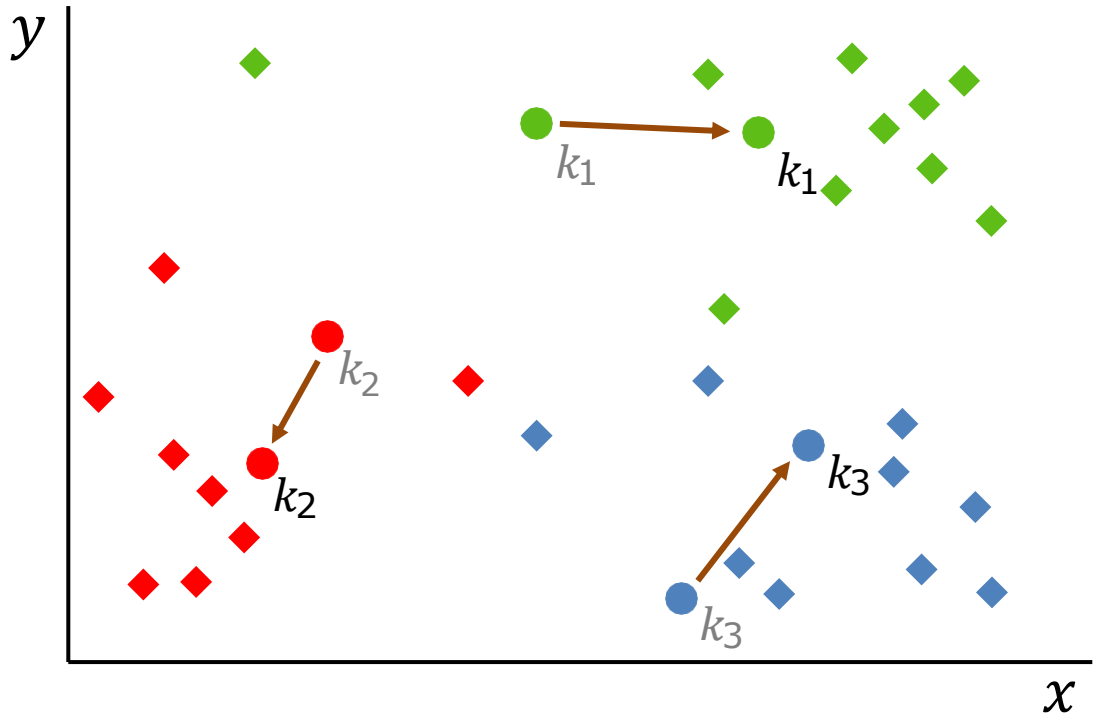
Pick 3 initial cluster centers at random



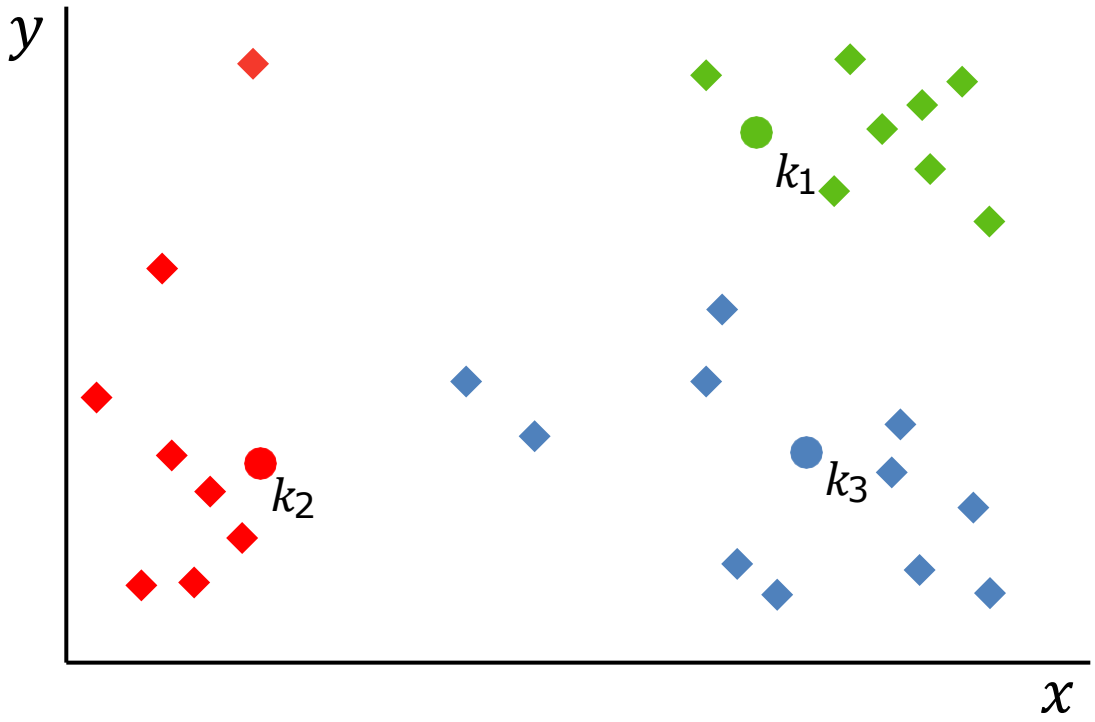
Assign each instance to the closest cluster center



Move each cluster center to the mean of each cluster.
Reassign points closest to a different new cluster center.

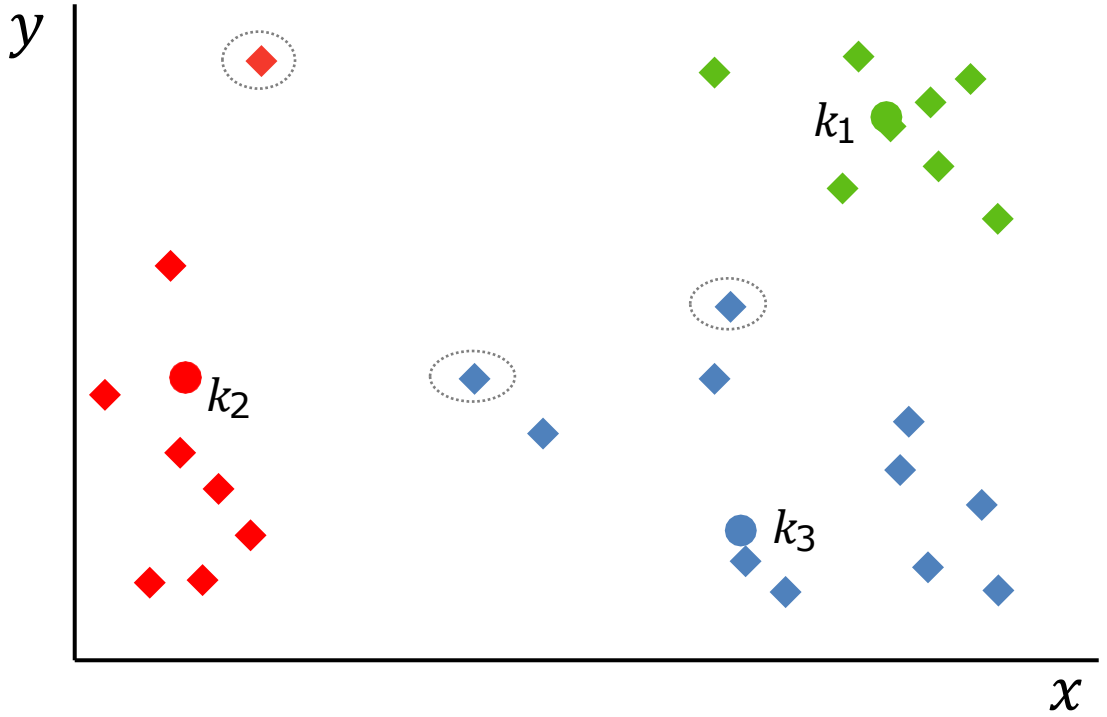


Recompute cluster means

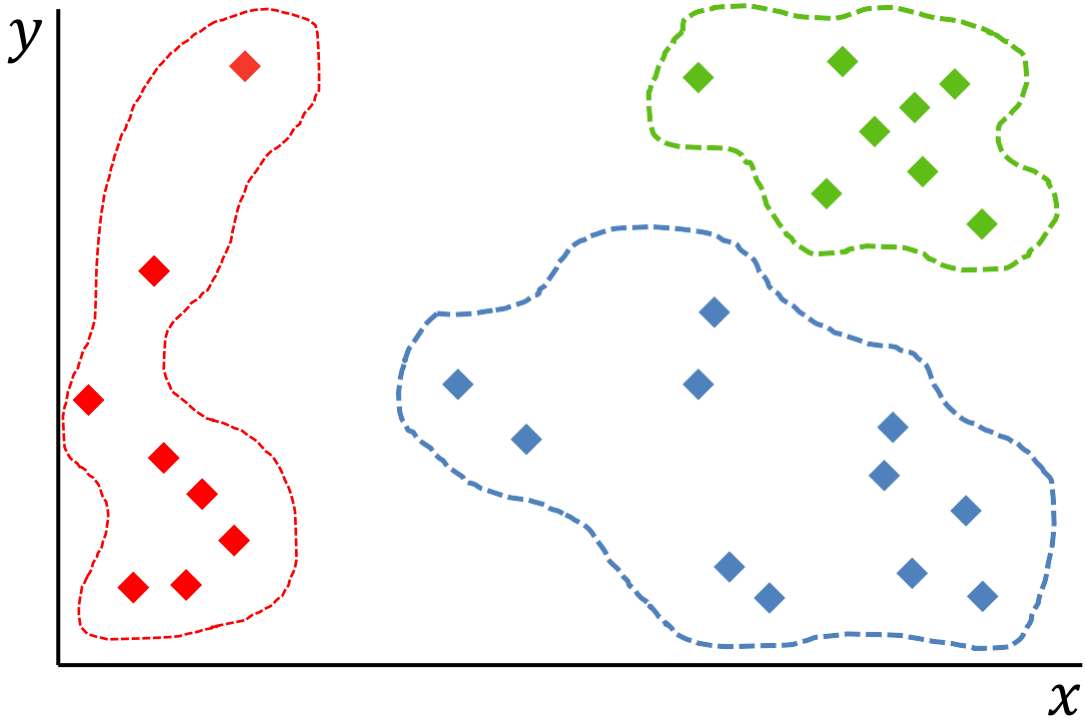


Reassign labels

No change. Done!



Final clusters



Evaluating Clusters

- 목적함수($objective\ function$)
  - 점들 사이 또는 점과 군집 중심점 간의 인접성에 의존
  - 클러스터의 품질은 오차 제곱의 합(ESS : $Error\ Sum\ of\ Squares$)으로 평가

$$ESS = \sum_{i=1}^{k} \sum_{x \in c_i} dist(c_i, x)^2$$

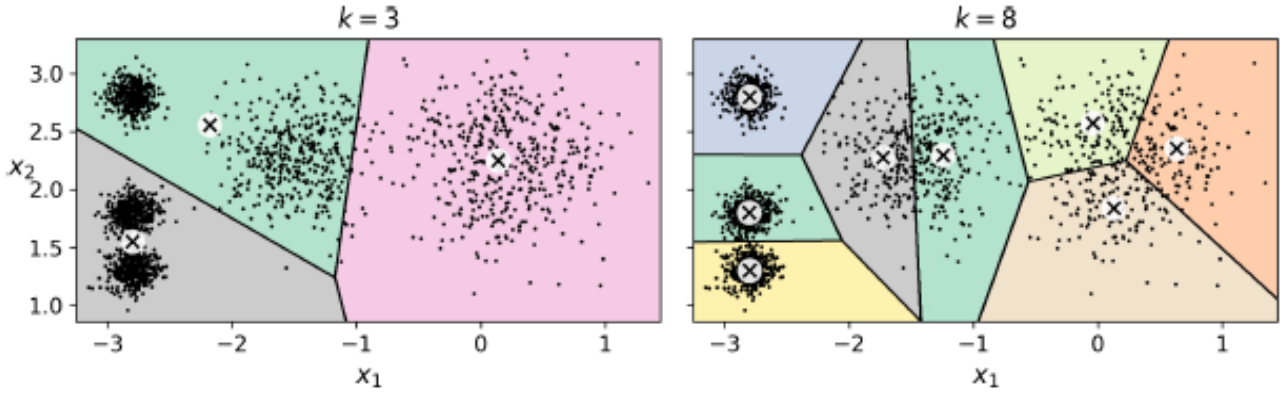
i번째 군집의 중심점(평균)은

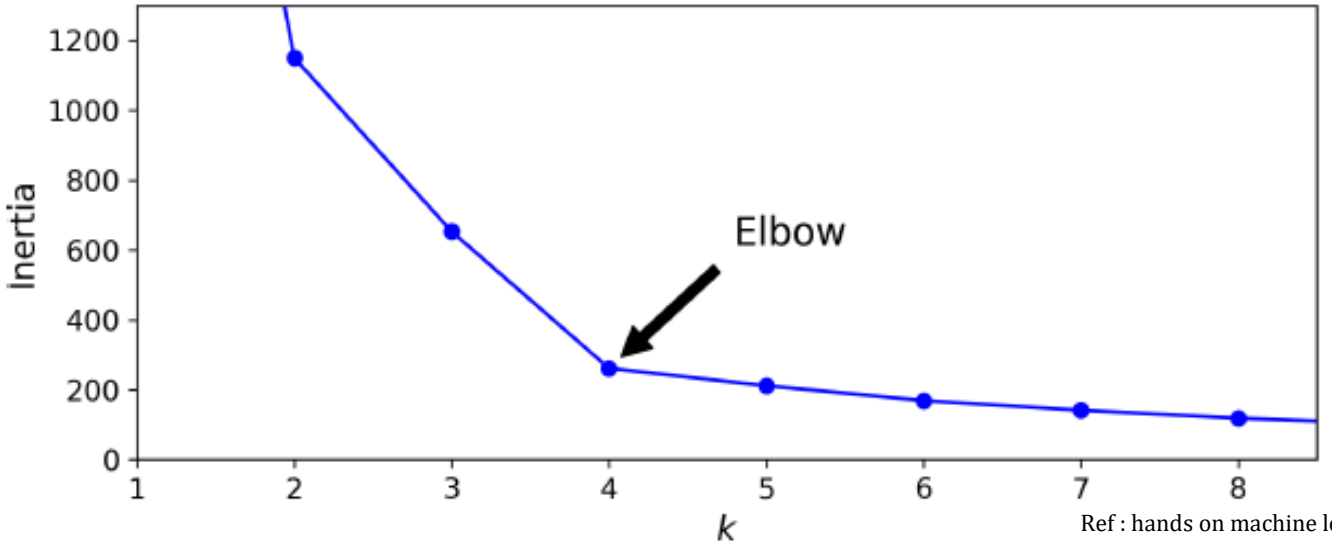
$$c_i = \frac{1}{m_i} \sum_{x \in c_i} x$$

- SSE의 최소화하는 k값을 선택하는 것이 중요

잘못된 클러스터 개수 선택

- k가 너무 작으면 별개의 클러스터를 합치고 k가 너무 크면 하나의 클러스터가 여러 개로 나뉜다
- 이너셔(inertia) : 각 샘플과 가장 가까운 센트로이드사이의 평균제곱거리

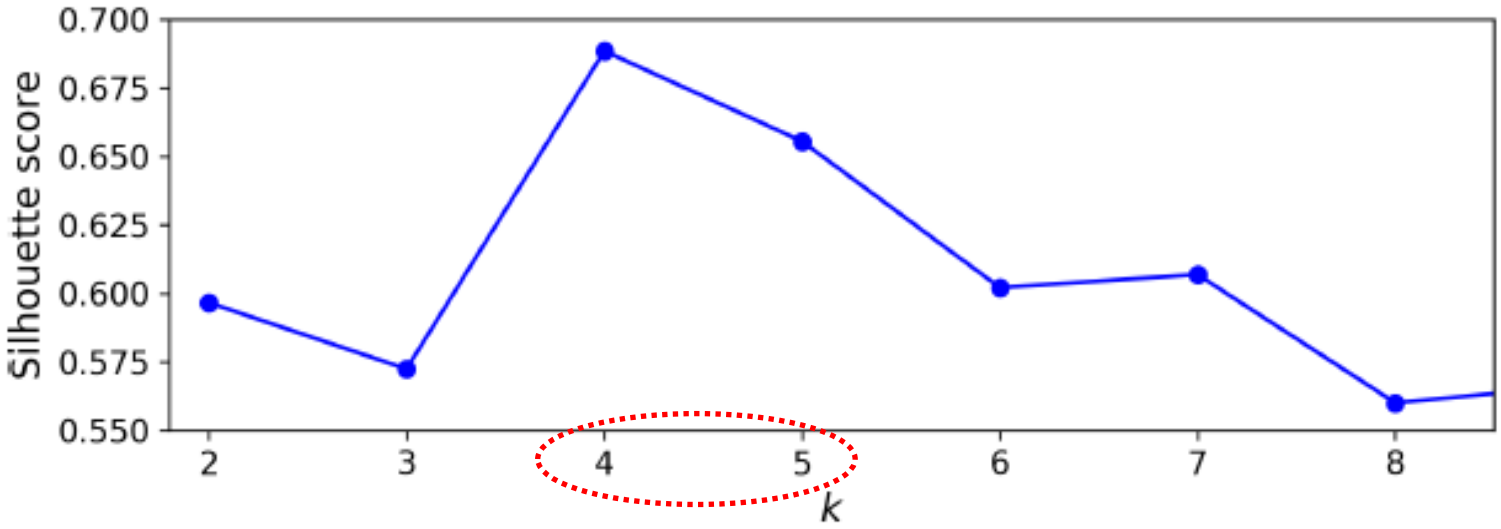




Ref : hands on machine learning, oreilly

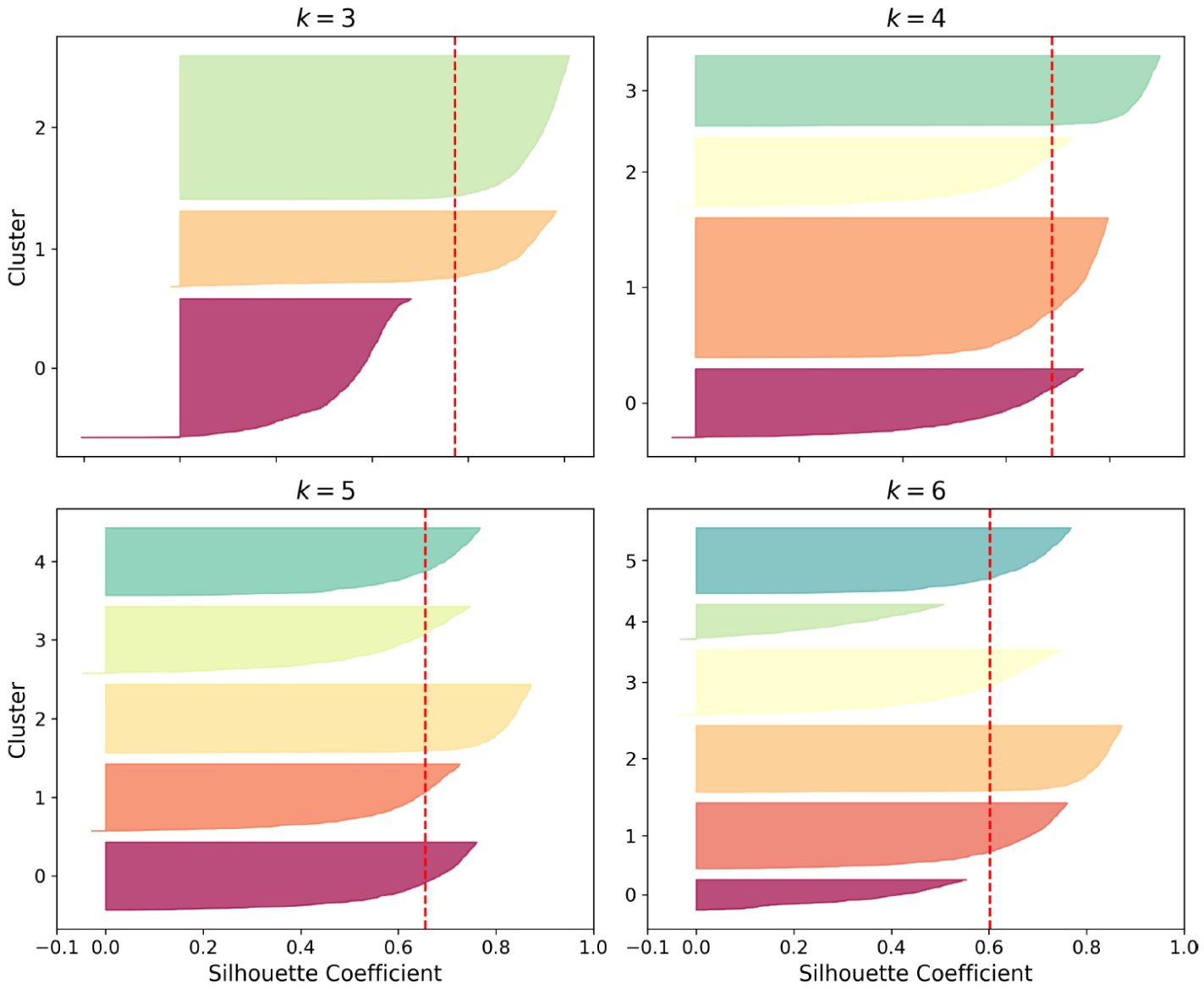
실루엣 점수

- 실루엣 계수 = $(b-a)/\max(a,b)$
  - a : 동일한 클러스터에 있는 다른 샘플까지 평균거리(클러스터 내부의 평균 거리)
  - b : 가장 가까운 클러스터까지 평균거리(가장 가까운 클러스터의 샘플까지 평균 거리)
  - $+1$에 가까우면 자신의 클러스터안에 잘 속해 있고 다른 클러스터와는 멀리 떨어져 있다는 뜻.
  - 0에 가까우면 클러스터 경계에 위치한다는 의미
  - -1에 가까우면 잘못된 클러스터라는 의미
- 실루엣 점수 : 실루엣 계수의 평균



Ref : hands on machine learning 2nd, oreilly

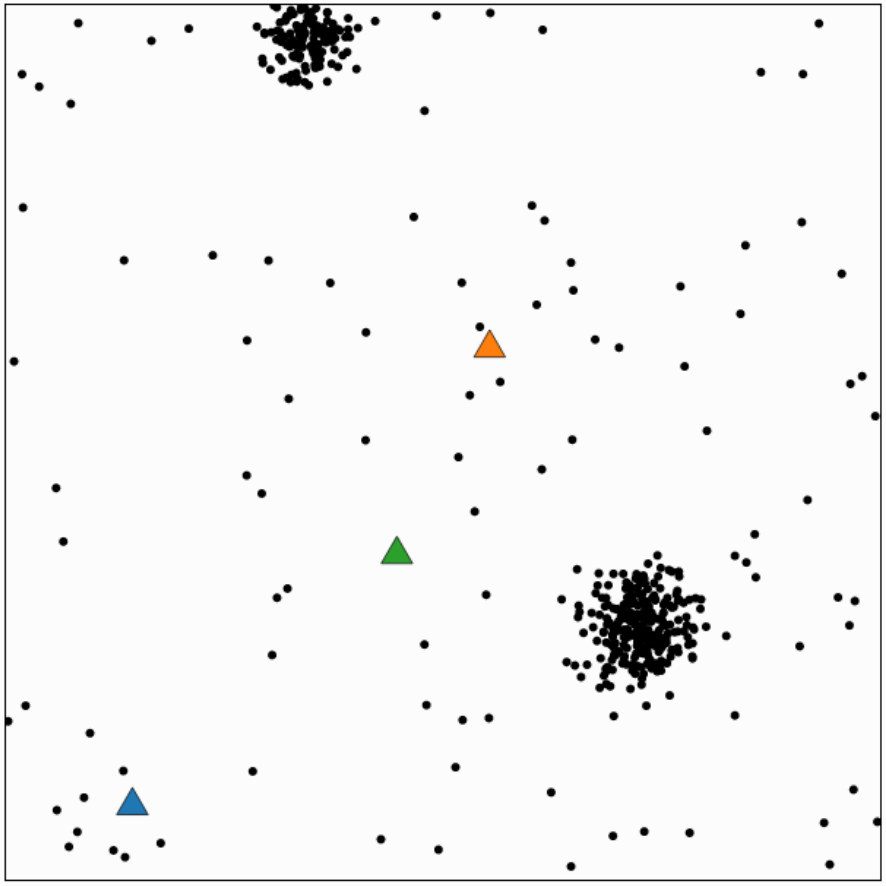
실루엣 다이어그램



Ref : hands on machine learning 2nd, oreilly

데모

https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html



Mean square point-centroid distance: not yet calculated

The k-means algorithm is an iterative method for clustering a set of N points (vectors) into k groups or clusters of points.

Algorithm

Repeat until convergence:

Find closest centroid

Find the closest centroid to each point, and group points that share the same closest centroid.

Update centroid

Update each centroid to be the mean of the points in its group.

Find closest centroid

Data

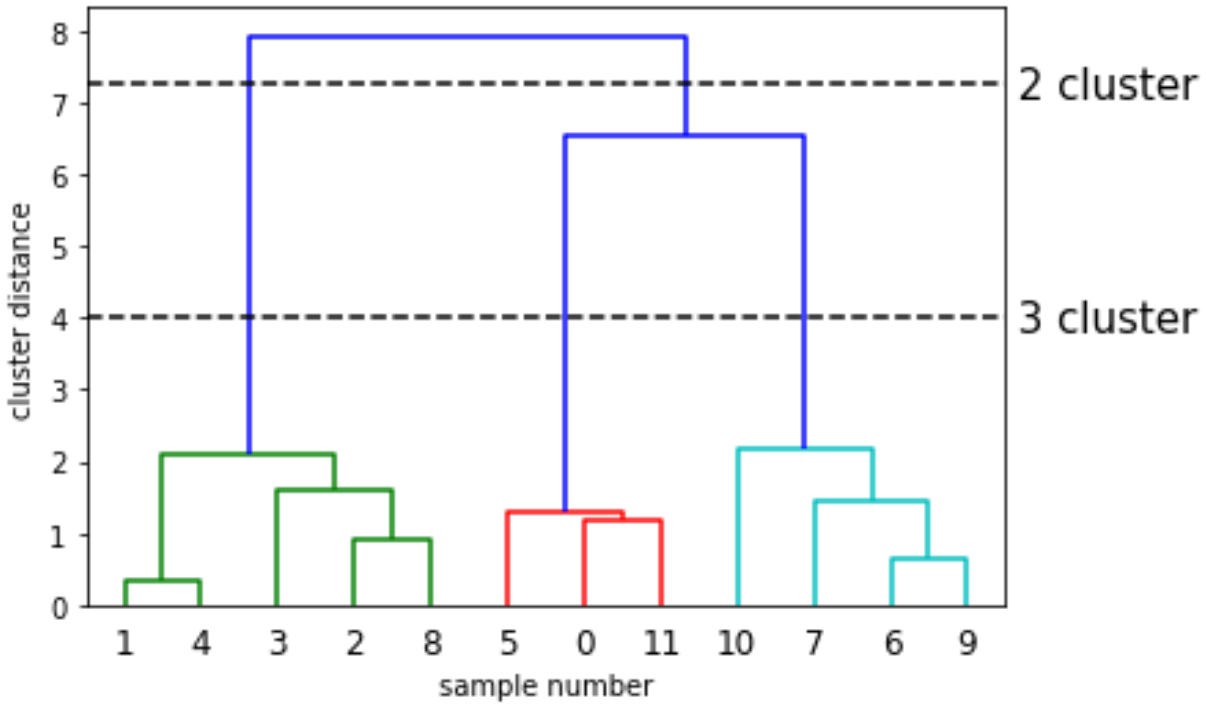
Clustered points — Random

Number of clusters : 3

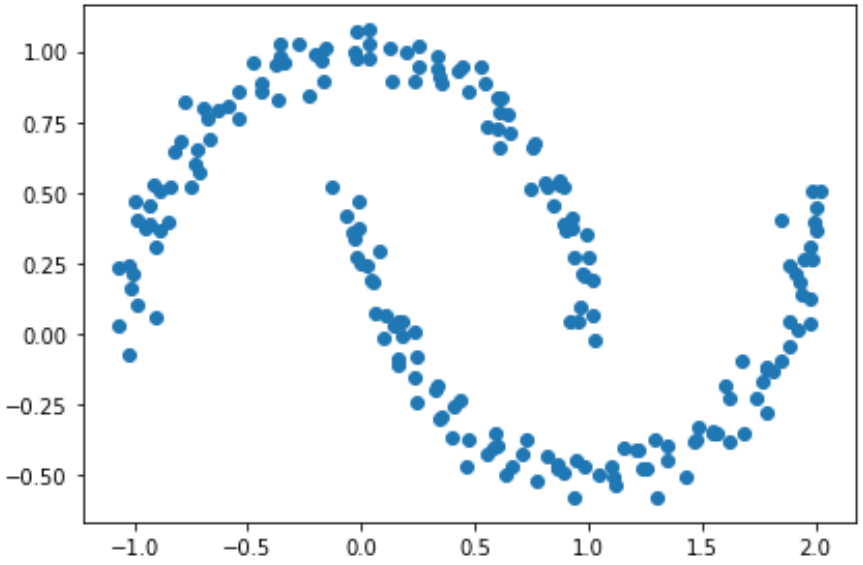
Number of centroids : 3

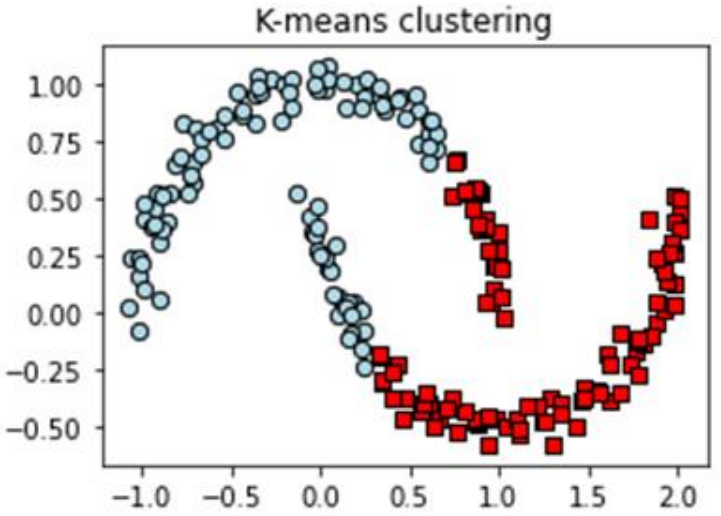
New points | New centroids

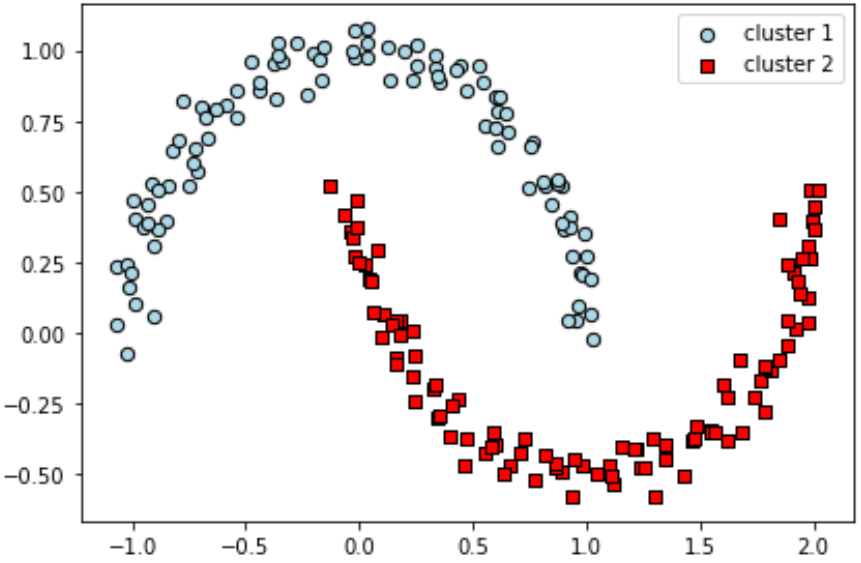
계층형 군집화(Hierarchical Clustering)

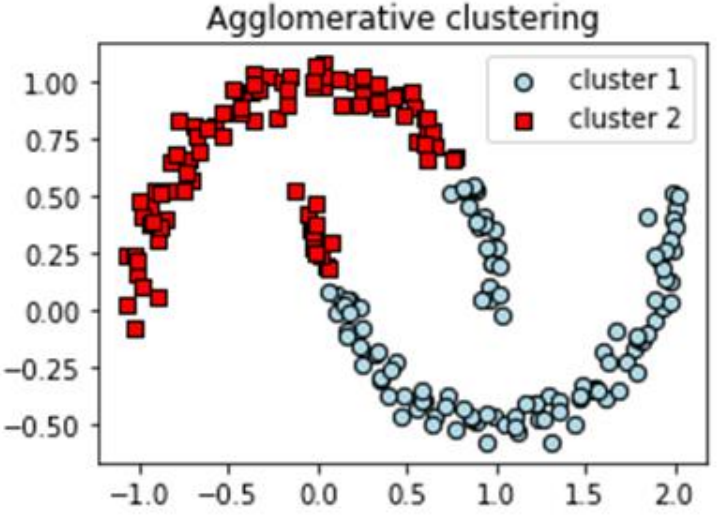


DBSCAN(*Density-Based Spatial Clustering of Application with Noise*)









Ref : Python Machine Learning - Second Edition - Packt