



Machine Learning Algorithm

- 지도 학습(Supervised Learning)

데이터에 대한 label(정답)이 주어진 상태에서 컴퓨터를 학습시키는 것

- 비지도 학습(Unsupervised Learning)

데이터에 대한 label(정답)이 주어지 않은 상태에서 컴퓨터를 학습시키는 것

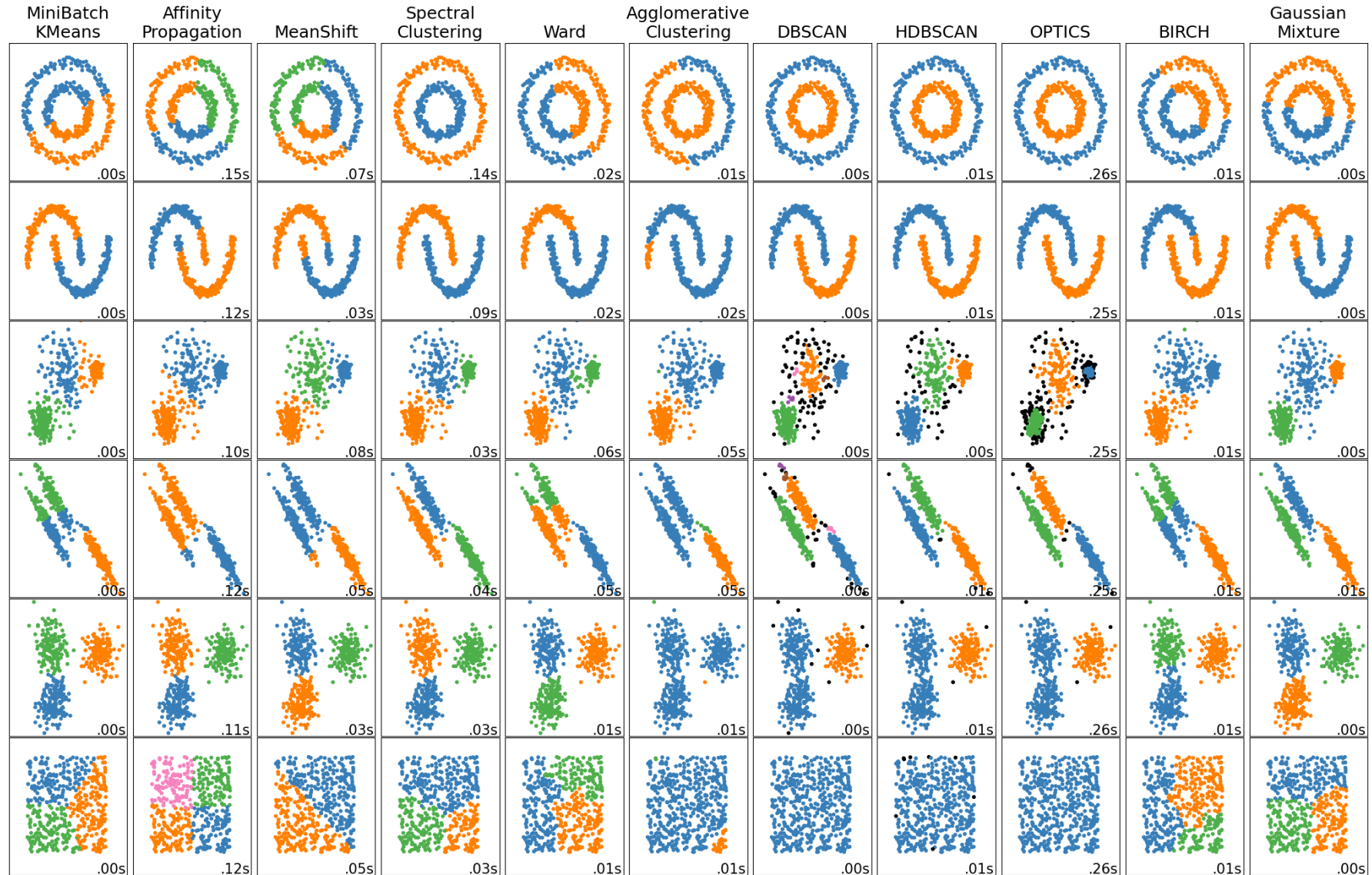
- 강화 학습(Reinforcement Learning)

현재 상태에서 어떤 행동을 취하는 것이 최적인지를 컴퓨터에 학습시키는 것

비지도 학습

- Clustering (군집화)
K-means, DBSCAN...
- Dimensionality reduction (차원 축소)
PCA, T-SNE...
- Visualization (시각화)
- Association Rule Learning (연관 규칙 학습)
Apriori, Eclat

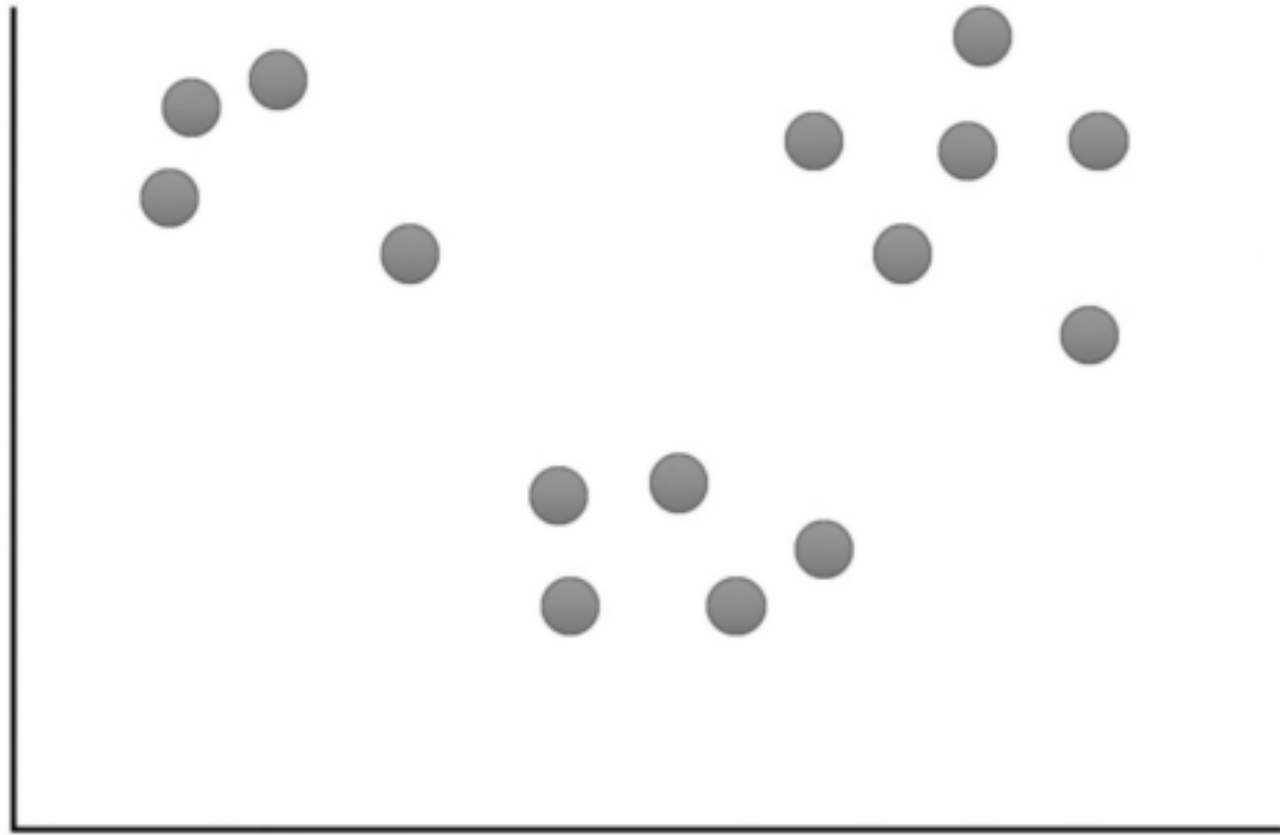
Clustering



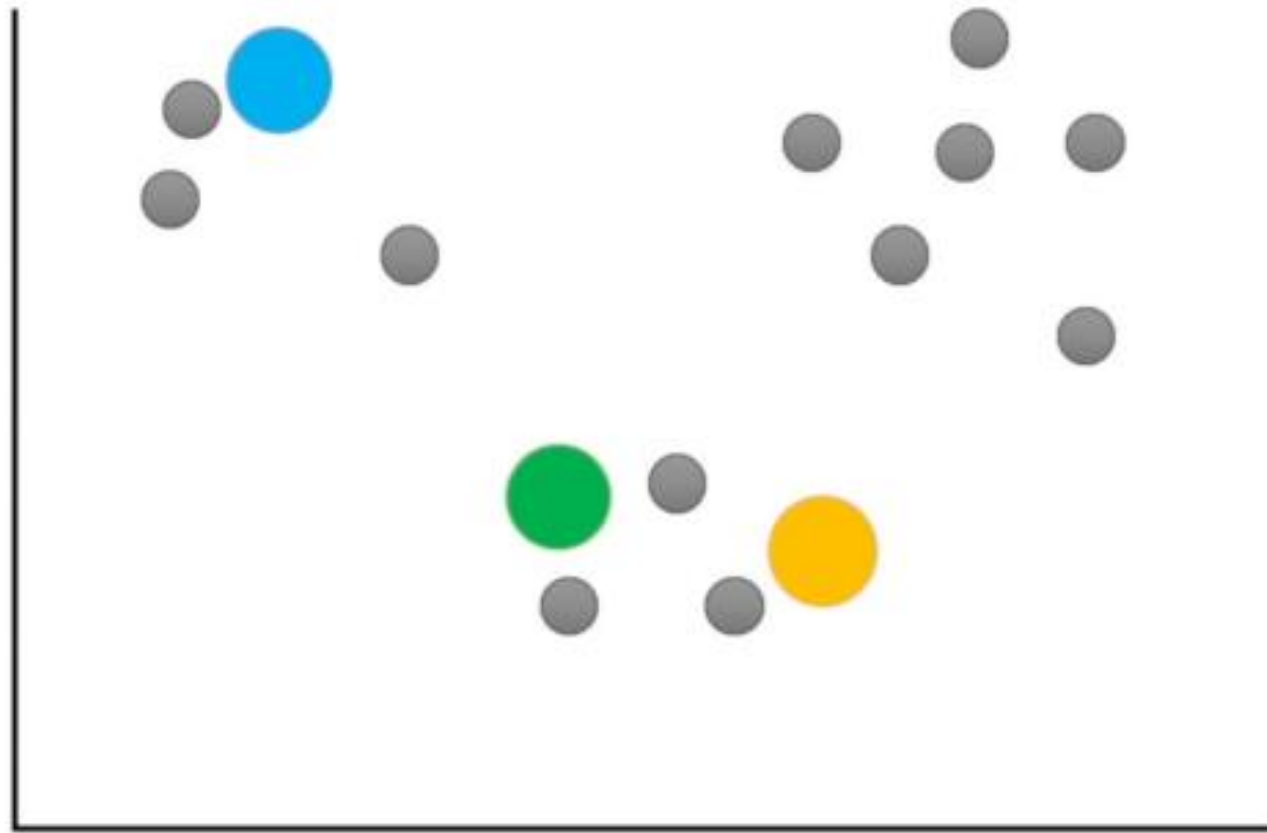
K-means

- K : Cluster의 개수
 - Means : Cluster의 중심과 데이터들의 평균 거리 *유클리드 거리
1. 원하는 클러스터의 수(K) 결정
 2. 무작위로 K개의 중심점(centroid)을 선정
 3. 각 데이터들을 가장 가까운 중심점의 cluster에 할당
 4. 각 cluster에 대하여 새로운 중심점을 계산
 5. 중심점이 변하지 않을 때까지 3-4의 과정 반복

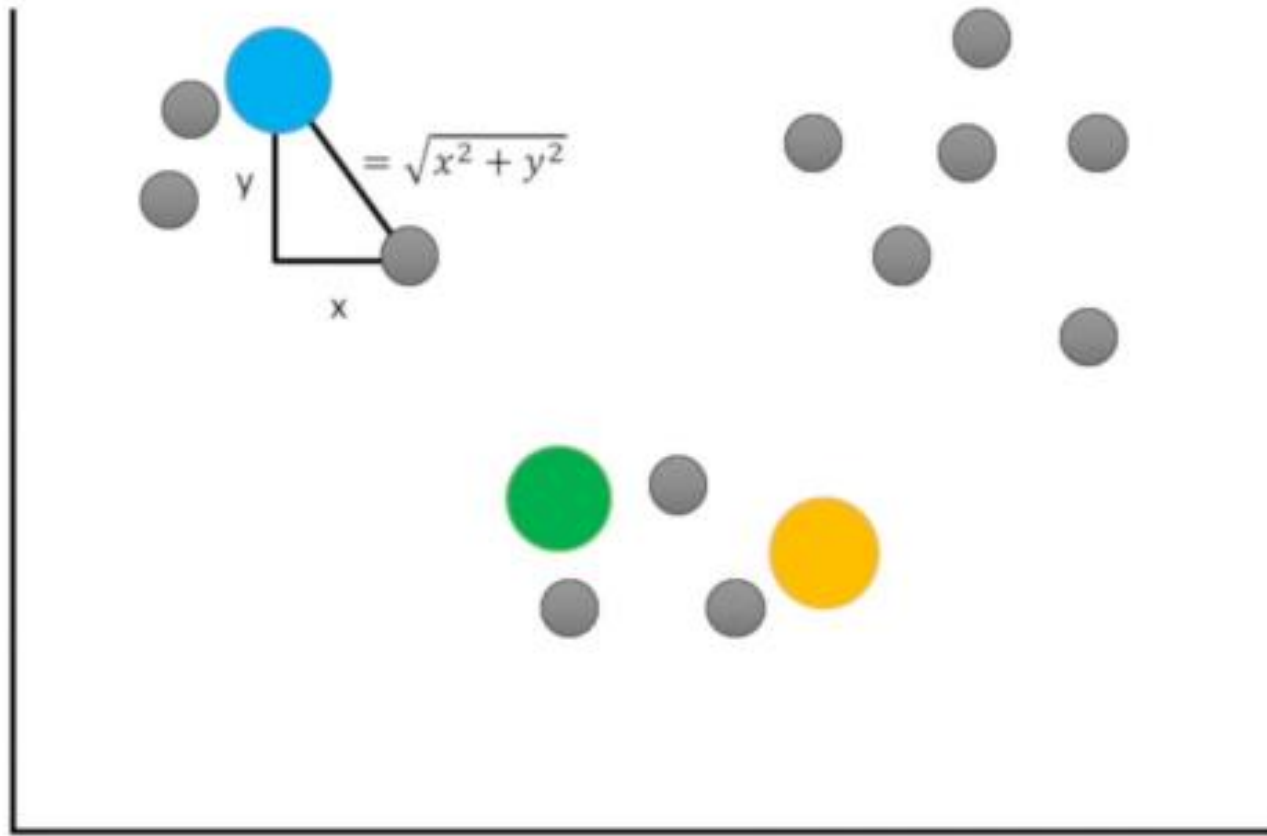
K-means



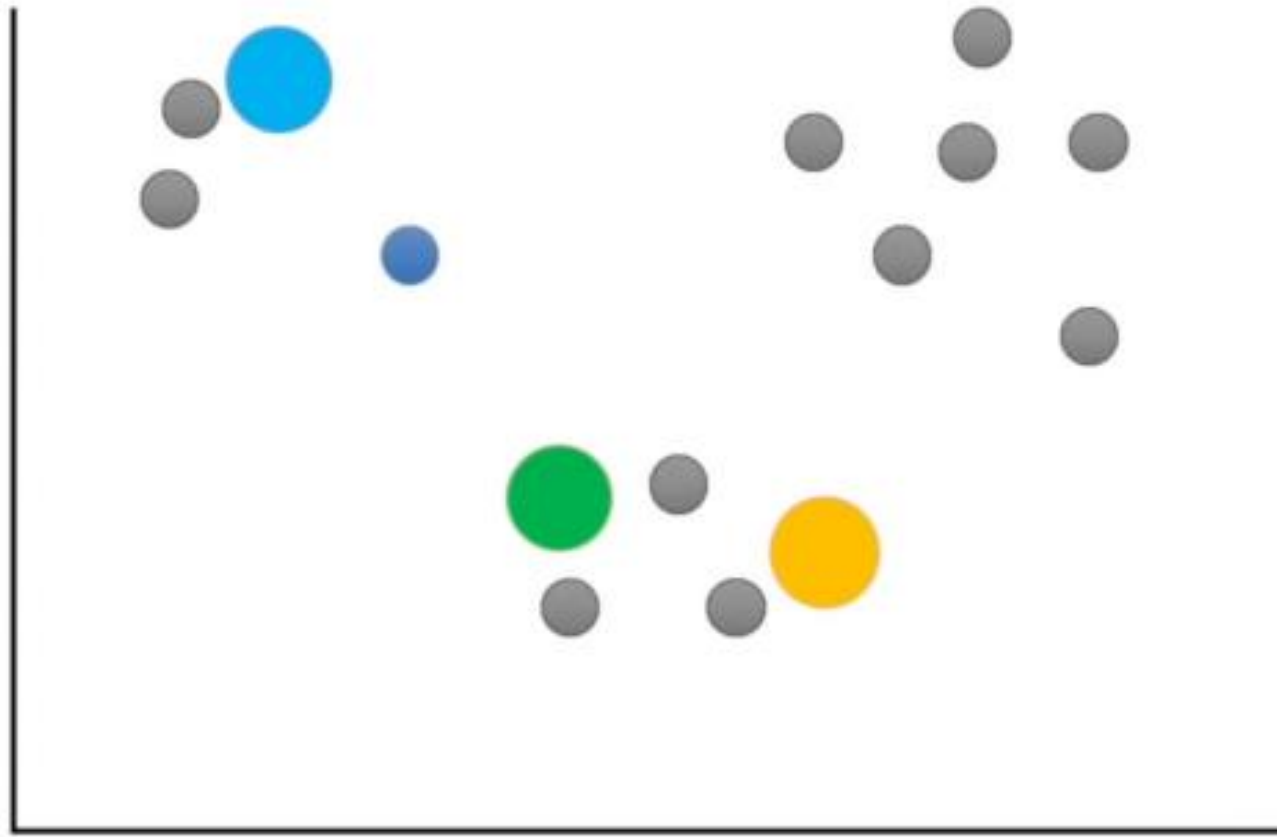
K-means



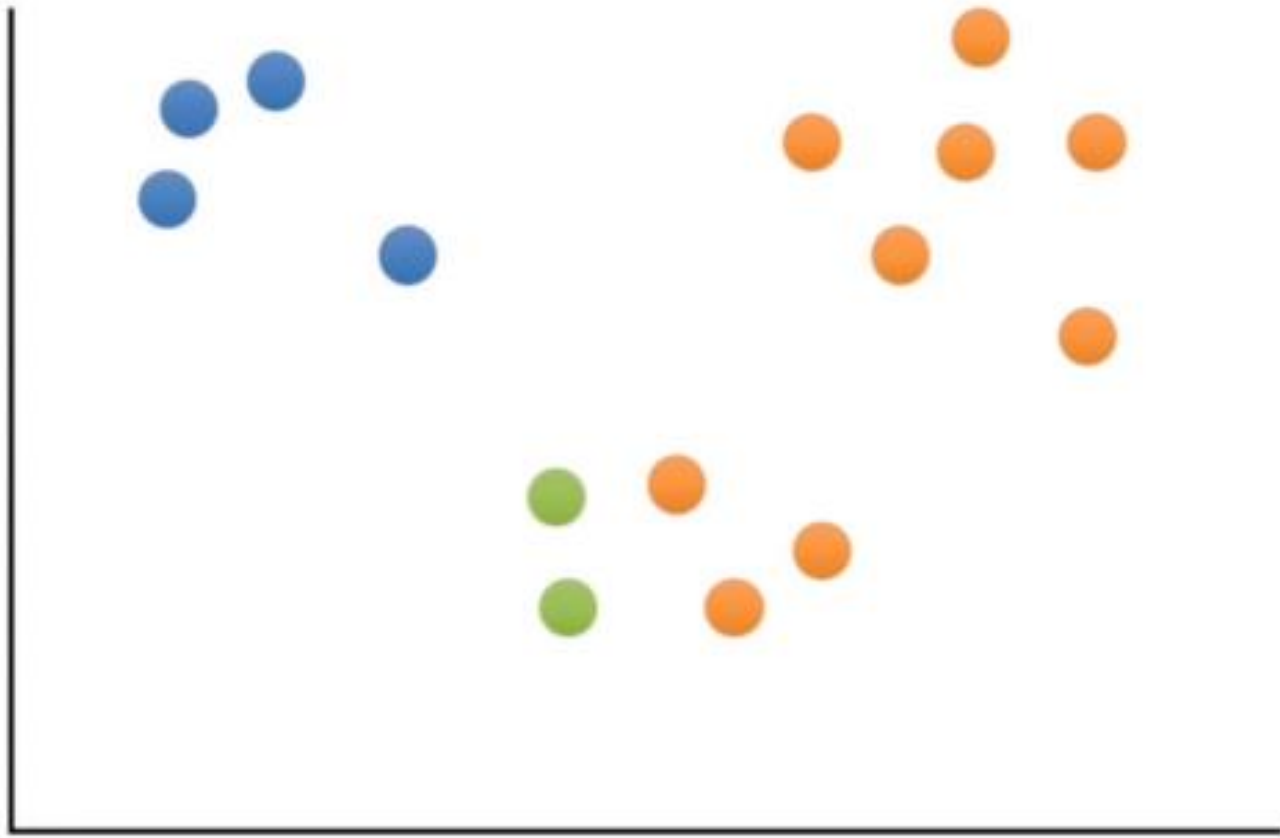
K-means



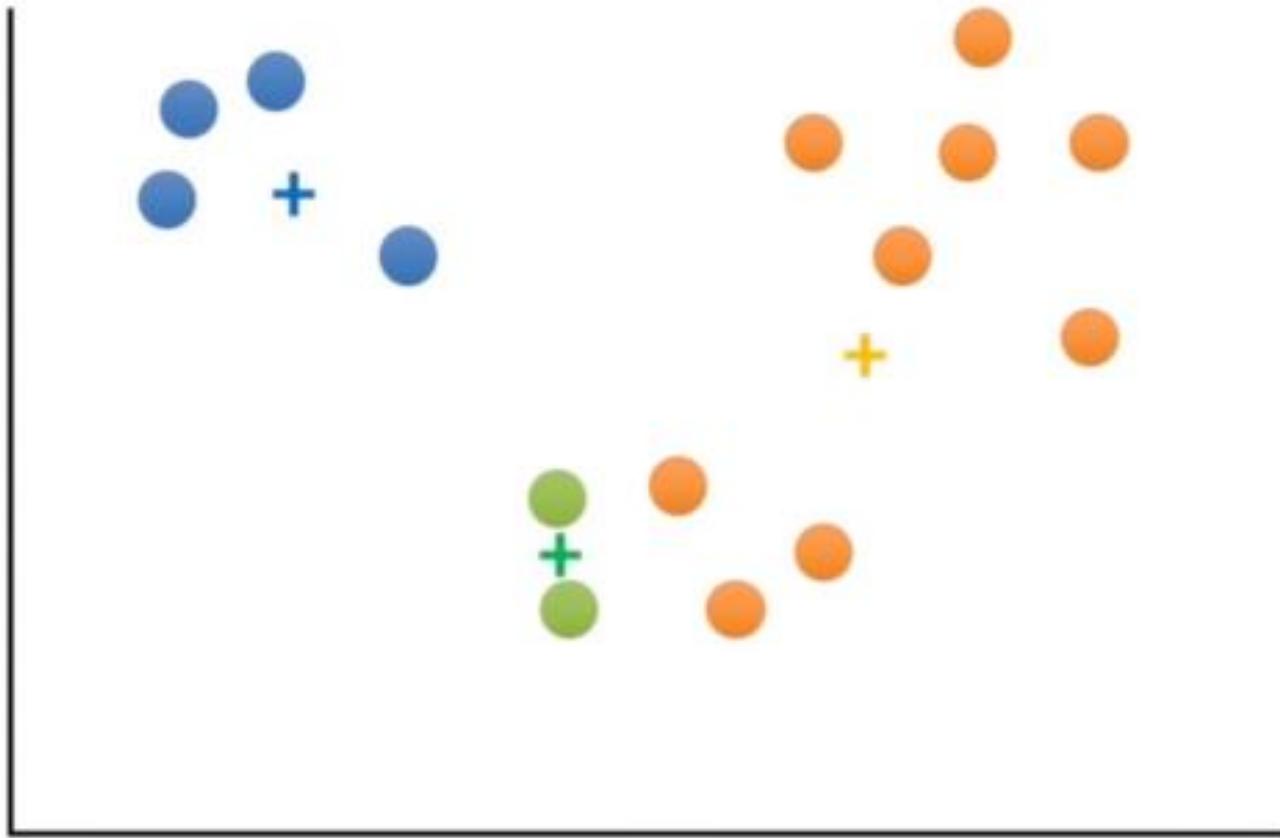
K-means



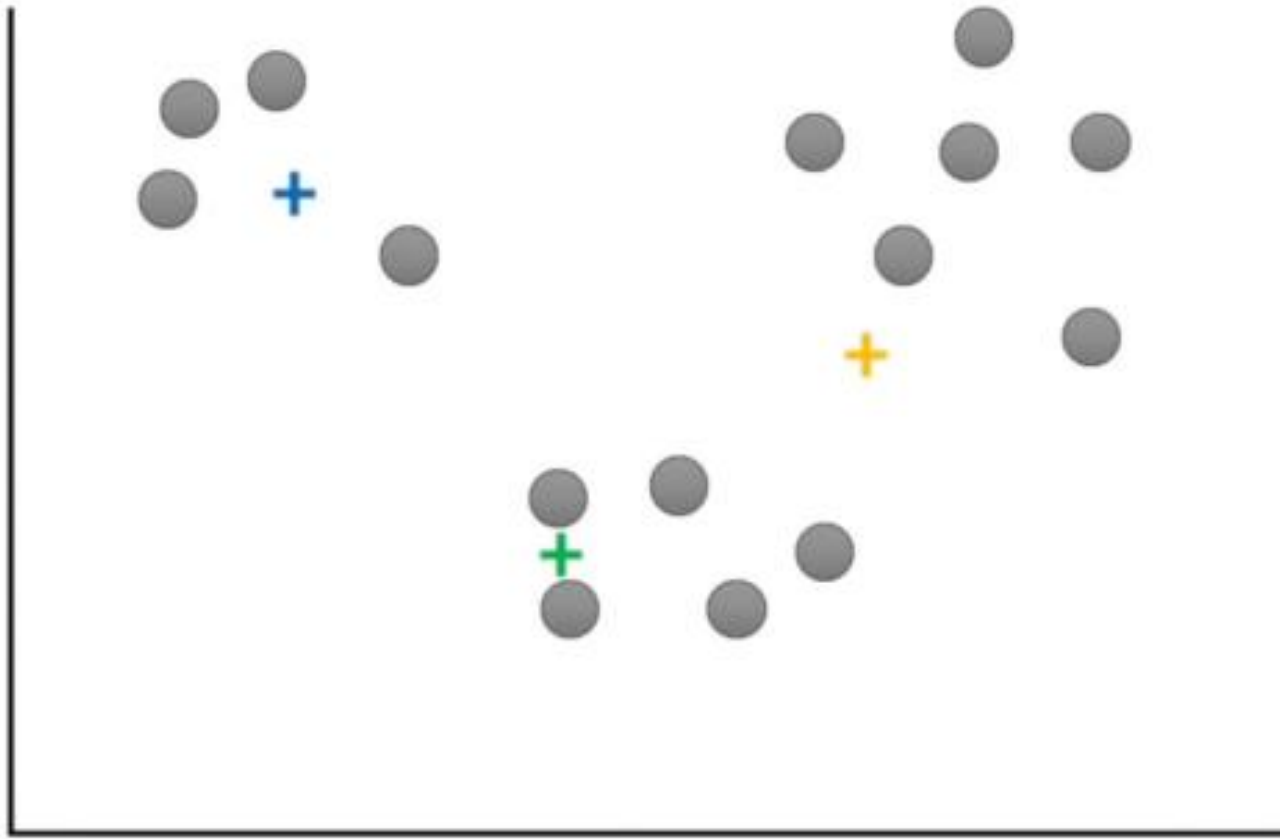
K-means



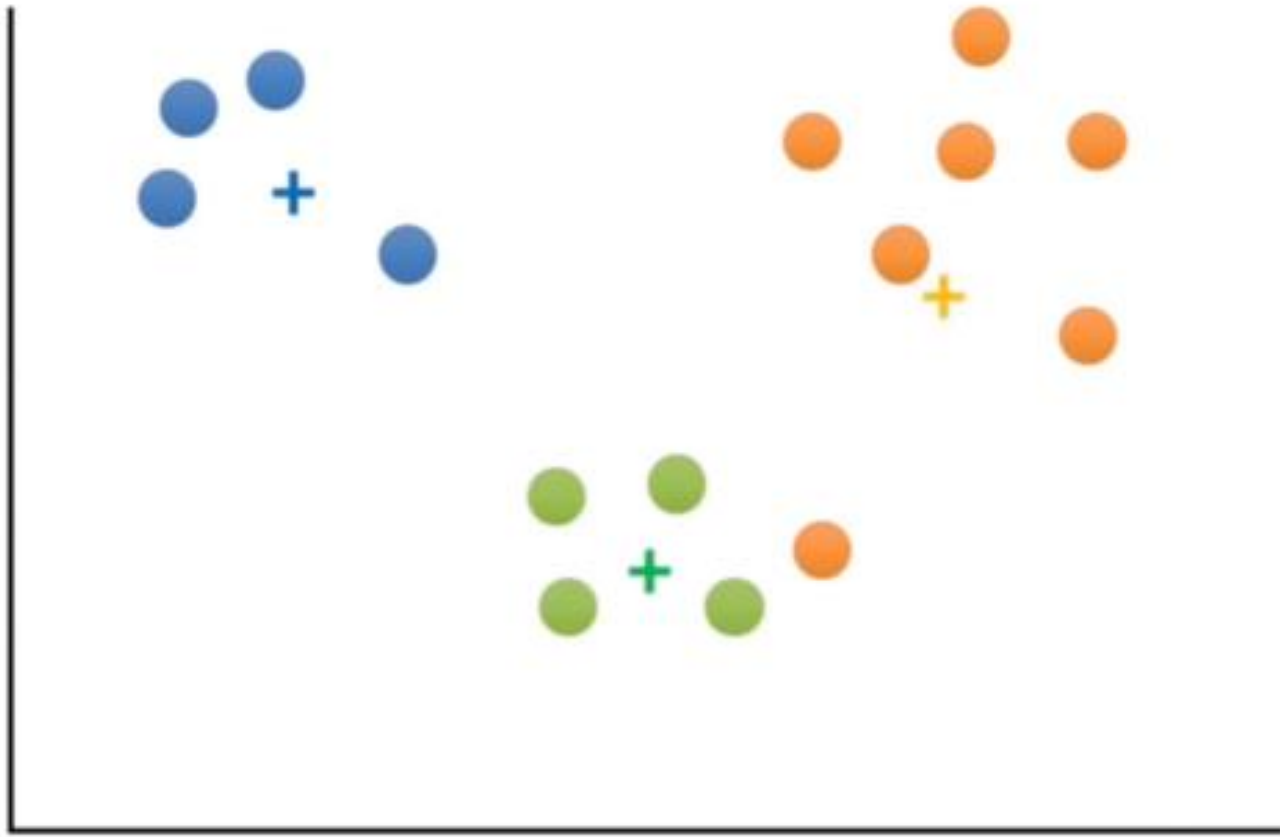
K-means



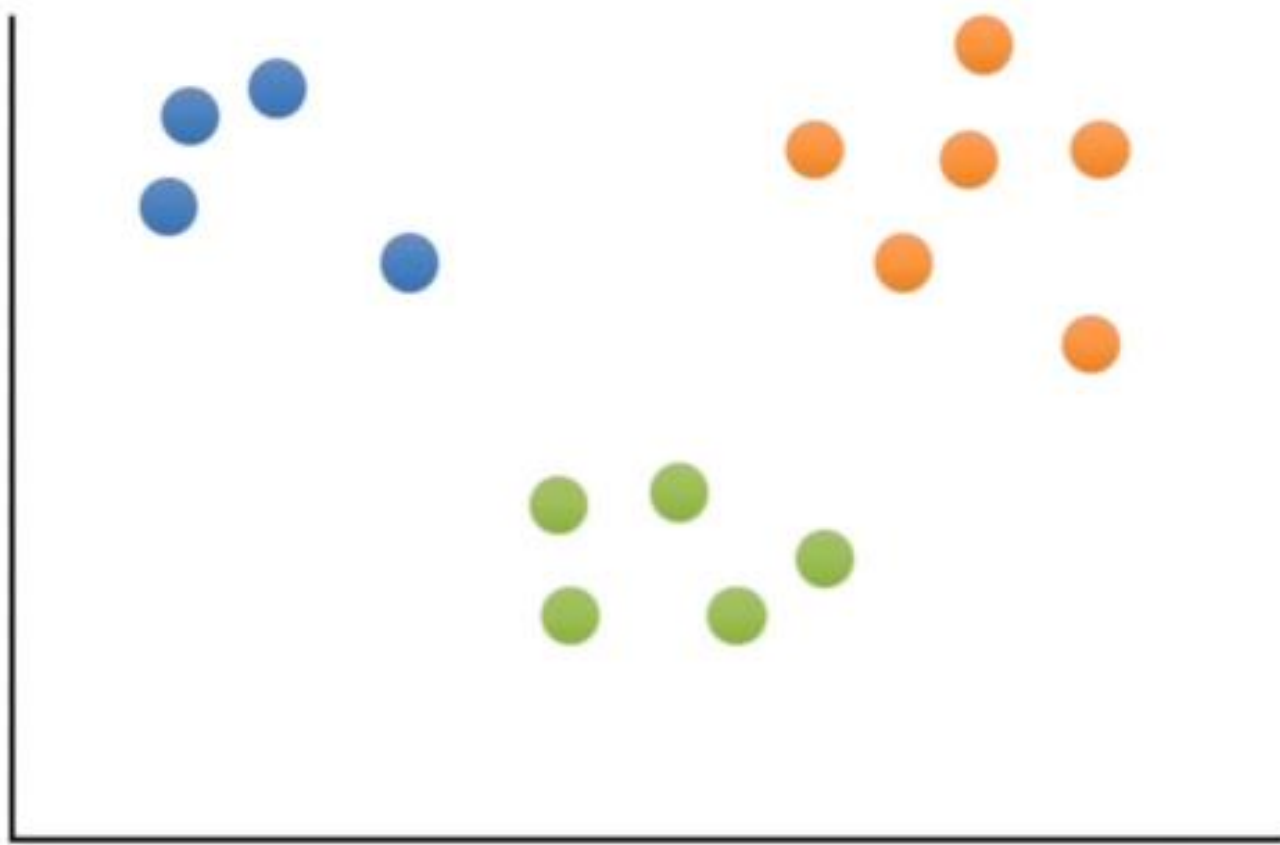
K-means



K-means



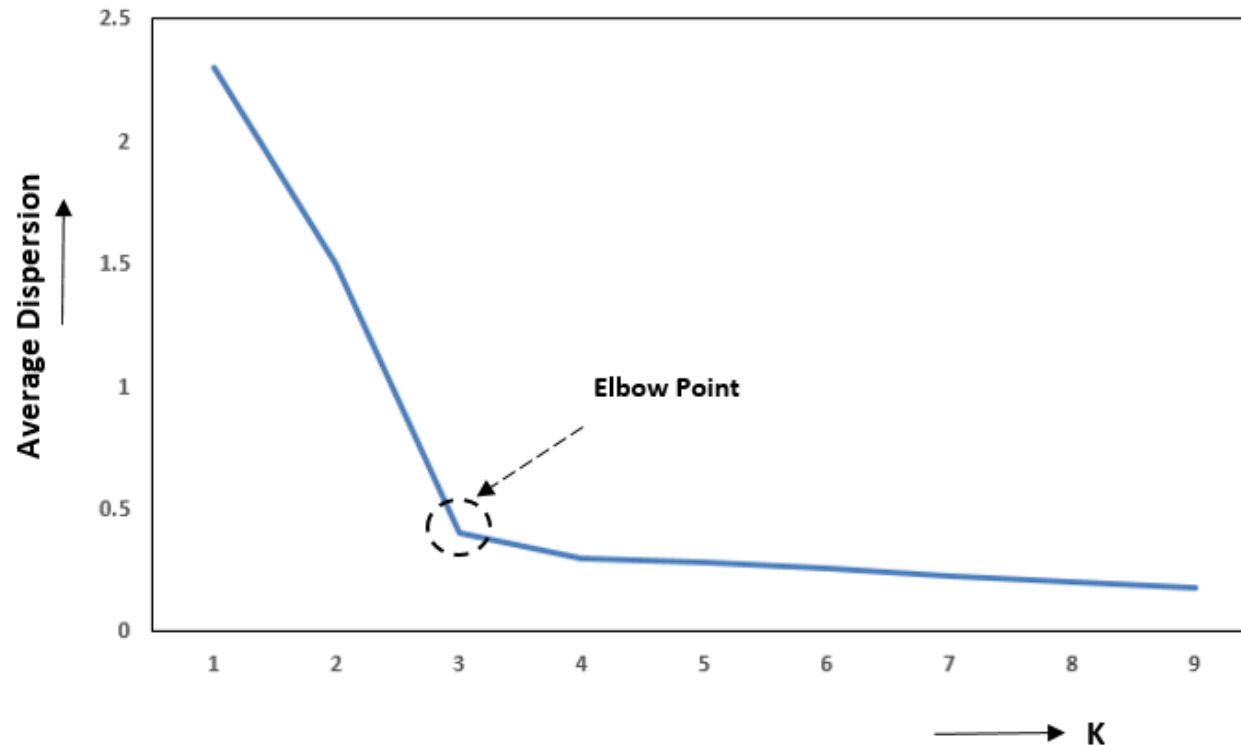
K-means



K-means

- 사전 지식이 없을 때의 k값 추정 : Elbow Method
- 변동량의 변화가 작아지는 지점의 군집의 수를 K로 지정

Elbow Method for selection of optimal “K” clusters



K-means

- 장점
 - 간단한 기준으로 clustering을 진행
 - 유연하고 효율적
- 단점
 - 계산이 단순하기 때문에 underfitting 발생 가능
 - 운이 없다면 최적의 cluster가 되지 않을 수 있음
 - K값을 미리 알고 있어야 함

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
밀도 기반 데이터 클러스터링 알고리즘
- 밀도 기반이란?
cluster를 찾을 때, 데이터들이 밀도 있게 모여 있는 곳을 찾는 알고리즘
= 서로 인접한 데이터들은 같은 cluster 안에 있음
- K-Means와의 차이점
DBSCAN : 서로 인접한 데이터들이 같은 cluster로 할당됨
K-means : cluster의 중심지와 가까운 데이터들이 같은 cluster로 할당됨

K-means vs DBSCAN

DBSCAN



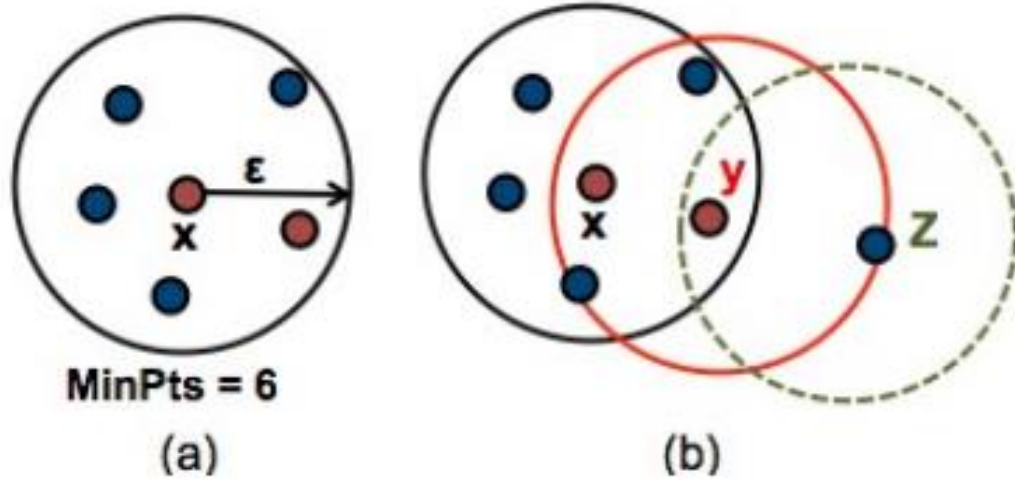
k-means



DBSCAN

- 파라미터
 - ϵ (epsilon) : 데이터로부터의 반경
 - minPts : cluster을 구성하는 데에 필요한 최소 데이터의 개수
- 데이터 포인트
 - core point : 한 점의 ϵ 반경 내에 minPts 이상의 개체가 포함된 점
 - border point : 한 점의 ϵ 반경 내에 minPts 보다 적은 개체를 포함하고 있지만,
적어도 하나의 core point의 반경에 속하는 점
 - noise point : core point 또는 border point가 아닌 점
 ϵ 반경 내에 minPts보다 적은 수의 개체를 포함 하는 점

DBSCAN



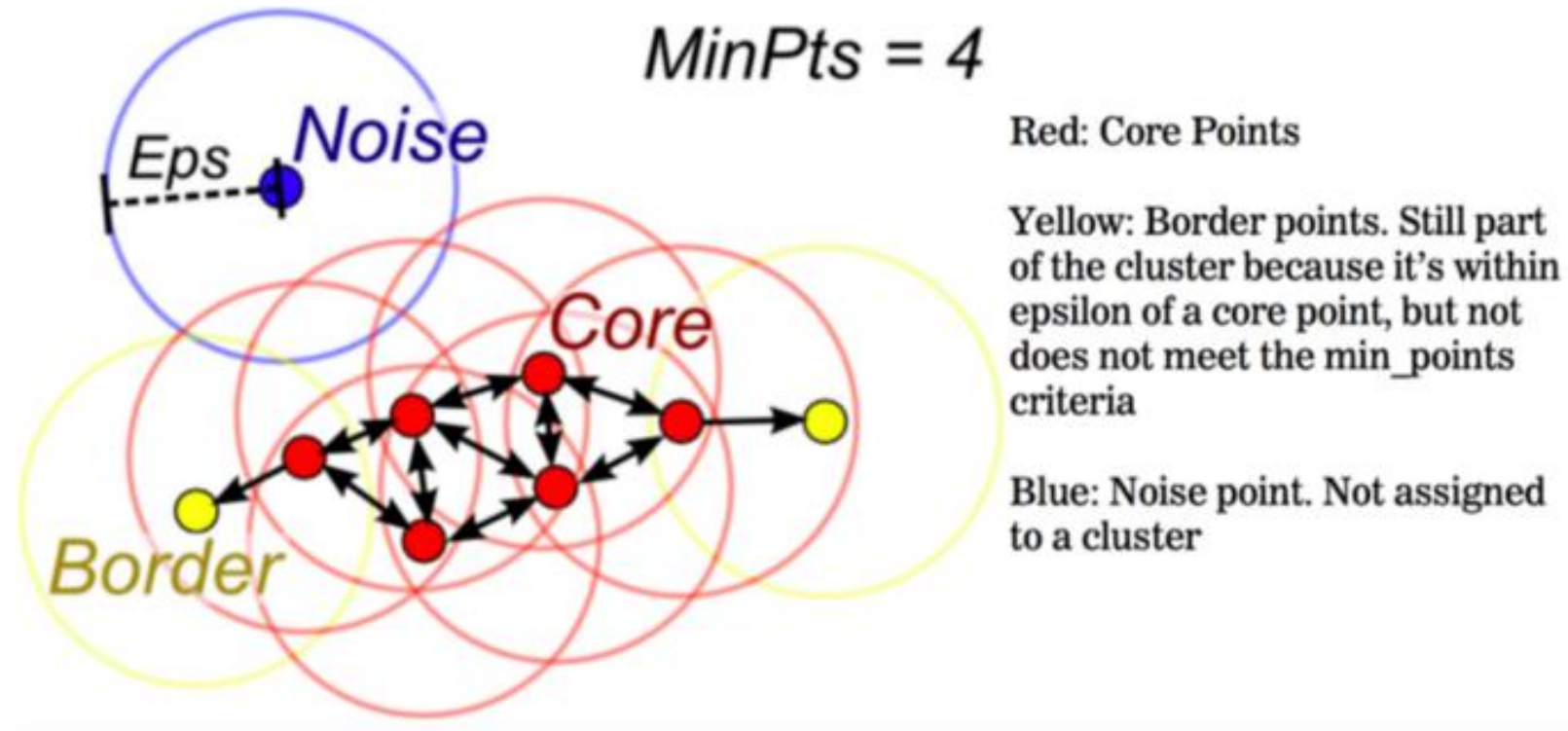
- 파라미터

- ϵ
- minPts

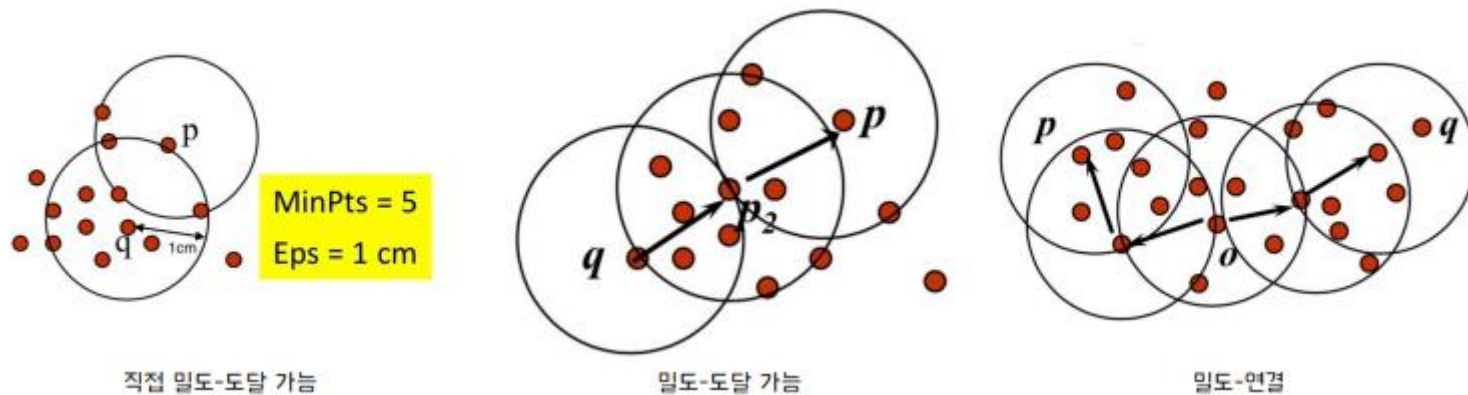
- 분류 포인트

- core point x
- border point y
- noise point z

DBSCAN



DBSCAN



- 직접 밀도-도달 가능

점 p 가 점 q 의 반경 안에 들어오고 점 q 가 코어 점일 때,
"점 p 가 점 q 로부터 직접 밀도-도달 가능한 관계에 있다"

- 밀도-도달 가능

점 p 와 점 q 사이에 p_1, p_2, \dots, p_n ($p_1 = q, p_n = p$)들이 있고,
점 p_{i+1} 이 점 p_i 로부터 직접 밀도-도달 가능하다면,
"점 p 는 q 로부터 밀도-도달 가능 관계에 있다"

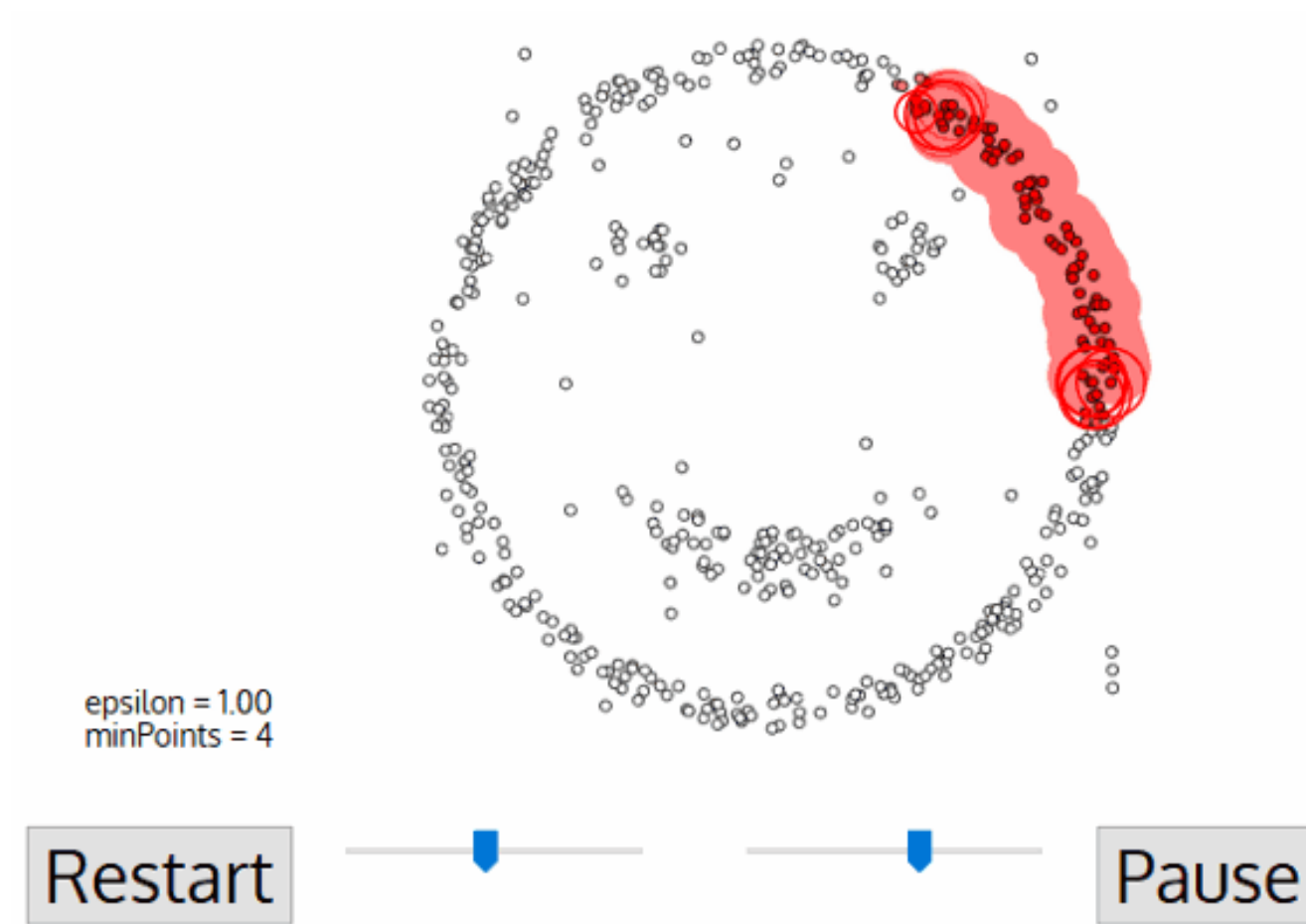
- 밀도-연결

어떤 점 O 로부터 다른 두 점 p, q 가 반경 내 minPts 조건 하에 밀도-도달 가능하다면,
"점 p 와 q 는 O 를 통해 서로 연결되어 있다"

DBSCAN

1. ϵ 과 minPts를 설정
2. 데이터로부터 core point를 만족하는 임의의 점 선택
3. 밀도-도달 가능한 점들을 뽑아 core point와 border point를 구분, 이에 속하지 않는 점들은 noise point로 분류
4. ϵ 반경 안에 있는 core point들을 서로 연결
5. 연결된 core point를 통해 하나의 cluster를 형성
6. 모든 border point들은 어느 하나의 cluster로 할당
(여러 cluster에 걸쳐 있는 경우, 반복 과정에서 먼저 할당된 cluster로 할당)

DBSCAN



DBSCAN

- 장점
 - non-linear한 clustering도 가능하다
 - cluster의 개수를 지정하지 않아도 된다
 - outlier에 크게 민감하지 않다
 - 시작 점에 의존성이 크지 않다
 - 속도가 빠르다
- 단점
 - 결과 값이 일정하지 않고 다양하게 나타난다
 - 여러 개의 cluster에 속하는 점들은 애매하게 분류된다
 - 유클리드 거리에 의존하여, 차원이 높아질수록 적당한 ϵ 을 찾기가 어렵다
 - 등성등성한 데이터에 대해서는 cluster 결과가 좋지 못하다
 - Normalizing의 의존성이 높다

Clustering 성능 평가

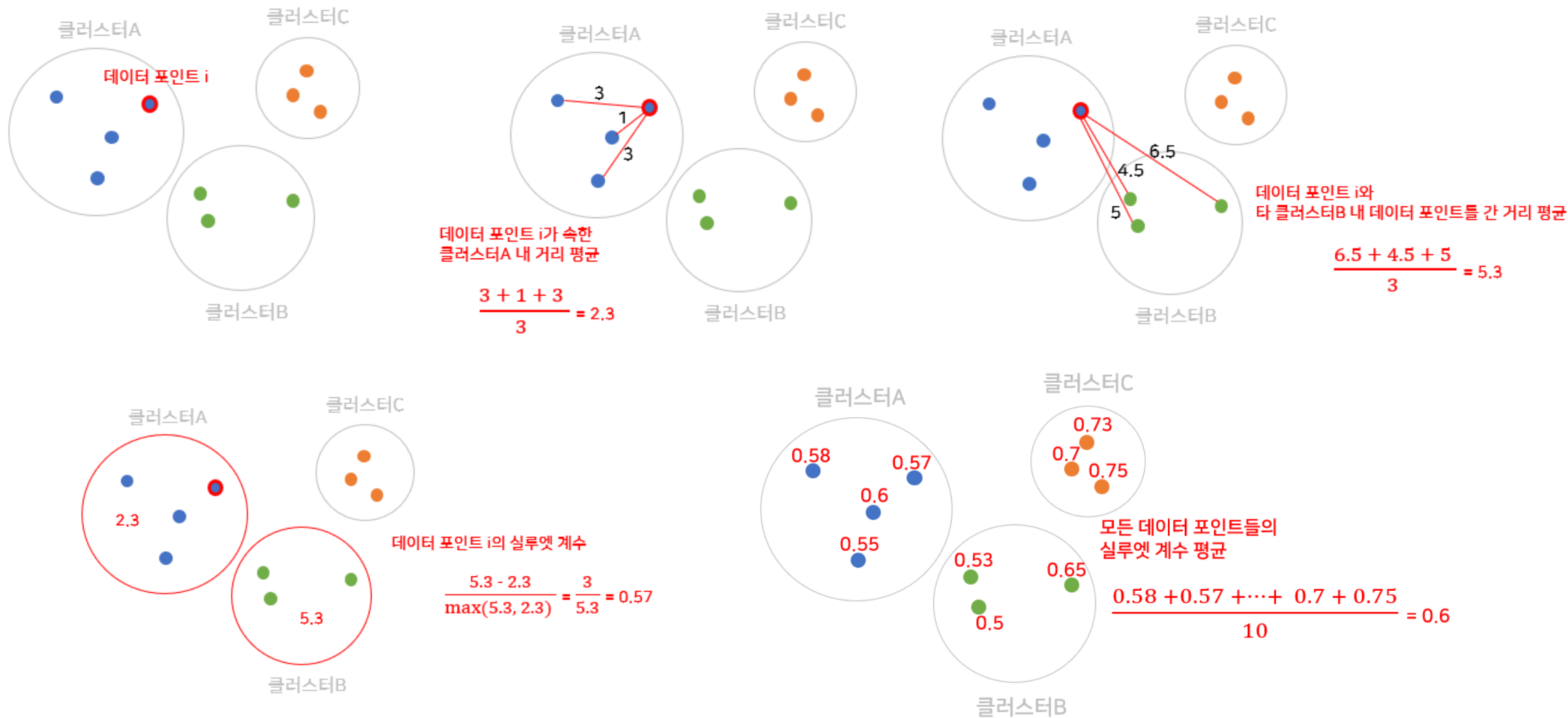
- Rand Index
- MI(mutual information)
- Homogeneity
- Completeness
- V-measure
- Silhouette Coefficient

A: 데이터와 그 데이터와 같은 그룹의 나머지 데이터 간의 평균 거리

B: 데이터와 그 데이터가 속하지 않은 가장 가까운 그룹의 데이터 간의 평균 거리

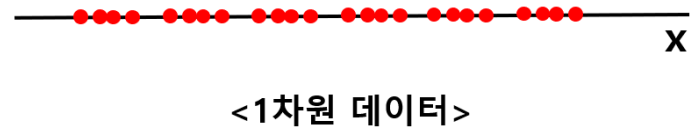
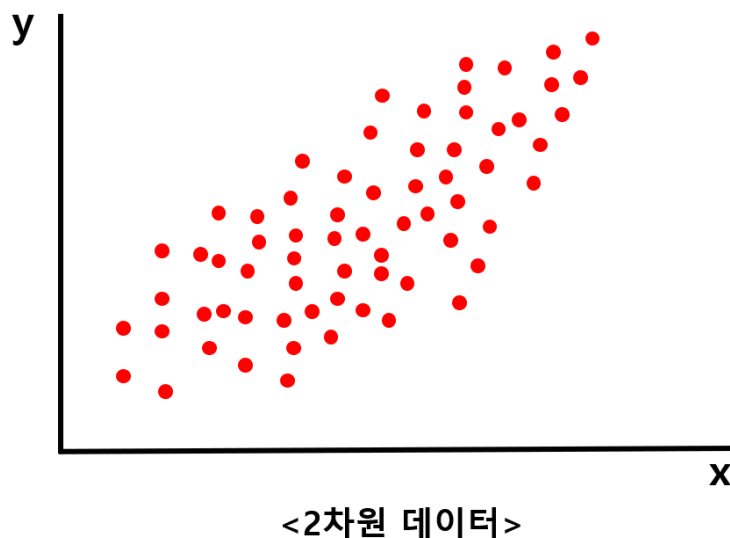
$$s = \frac{b - a}{\max(a, b)}$$

Clustering 성능 평가



PCA

- Principle Component Analysis (주성분분석)
- 각 데이터들의 feature들의 기여도를 분석하여 높은 feature만 추출하는 방법
차원을 축소하지만, 손실되는 정보는 최대한 적게 하는 것이 이 알고리즘의 목적이다



PCA

- 장점
 - 고차원의 데이터를 저차원으로 축소하기 때문에 효율적이다
- 단점
 - 때에 따라서 중요한 정보가 손실될 때가 있다
 - 비선형 데이터에서는 적절하지 못하다