비지도학습

머신러닝의 유형

유형	방법
지도학습 (Supervised Learning)	선형회귀, 로지스틱 회귀
	트리, 랜덤포레스트, 에이다부스트
	Naïve Bayes
	Support Vector Machine (SVM)
	인공신경망
	k-NN
비지도학습 (Unsupervised Learning)	군집분석: k-means, hierarchical, DBSCAN
	주성분 분석 (PCA), 비음수 행렬분해 (NMF)
	t-SNE
	연관성 분석

비지도학습의 활용

비지도학습의 활용:

- 학습 데이터의 구조적 이해를 위해서 활용할 수 있다.
- 새로운 데이터의 분류의 목적으로 활용할 수 있다.

Red? Blue?

k-means 클러스터 알고리즘

k-means 클러스터 알고리즘:

- 비지도학습.
- 목적은 관측값을 k 개의 군집 (클러스터)으로 분류하는 것.
- 군집에는 centroid 라고도 불리우는 중심점이 있음.
- 반복적 수렴 알고리즘 (Lloyd의 표준 알고리즘).
- 연속적 변수를 사용하며 거리의 개념이 필요하다.

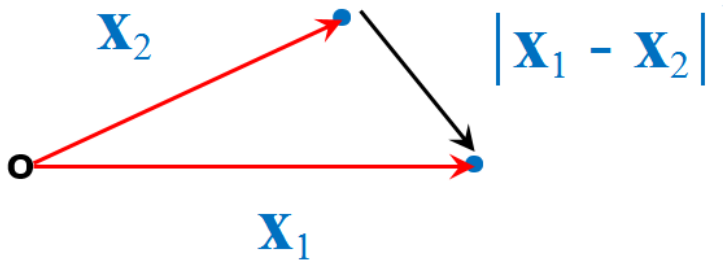
k-means 클러스터 알고리즘

k-means 클러스터 알고리즘: 장단점

장점	단점
✓ 직관적인 이해가 가능함. ✓ 모수에 대한 검정이 필요 없다. ✓ 쉽게 적용할 수 있다.	✓ 노이즈의 영향을 많이 받는다. ✓ 외상치 (outlier)에 비교적 민감하다.

거리의 척도 : 유클리드 거리

유클리드 거리 (Euclidian distance):



$$Euclidean\ distance = \sqrt{(x_{11} - x_{21})^2 + \cdots + (x_{1m} - x_{2m})^2}$$

거리의 척도 : 정의

다른 거리의 정의:

- 수치 변수인 경우:
  → 유클리드 거리.
  → 표준화 거리.
  → 마할라노비스 거리.
  → 체비셰프 거리.
  → 캔버라 거리.
  → 맨하탄 거리.
  → 민코우스키 거리.

- 유형 변수인 경우:
  → 자카드 거리, 등.

k-means 클러스터 알고리즘의 원리

다음과 같이 N개의 좌표로 나타내는 데이터 세트를 가정해 봅니다.

$$x_1, x_2, \cdots, x_N$$

두개 ($k = 2$) 의 클러스터를 목표로 설정해 봅니다.

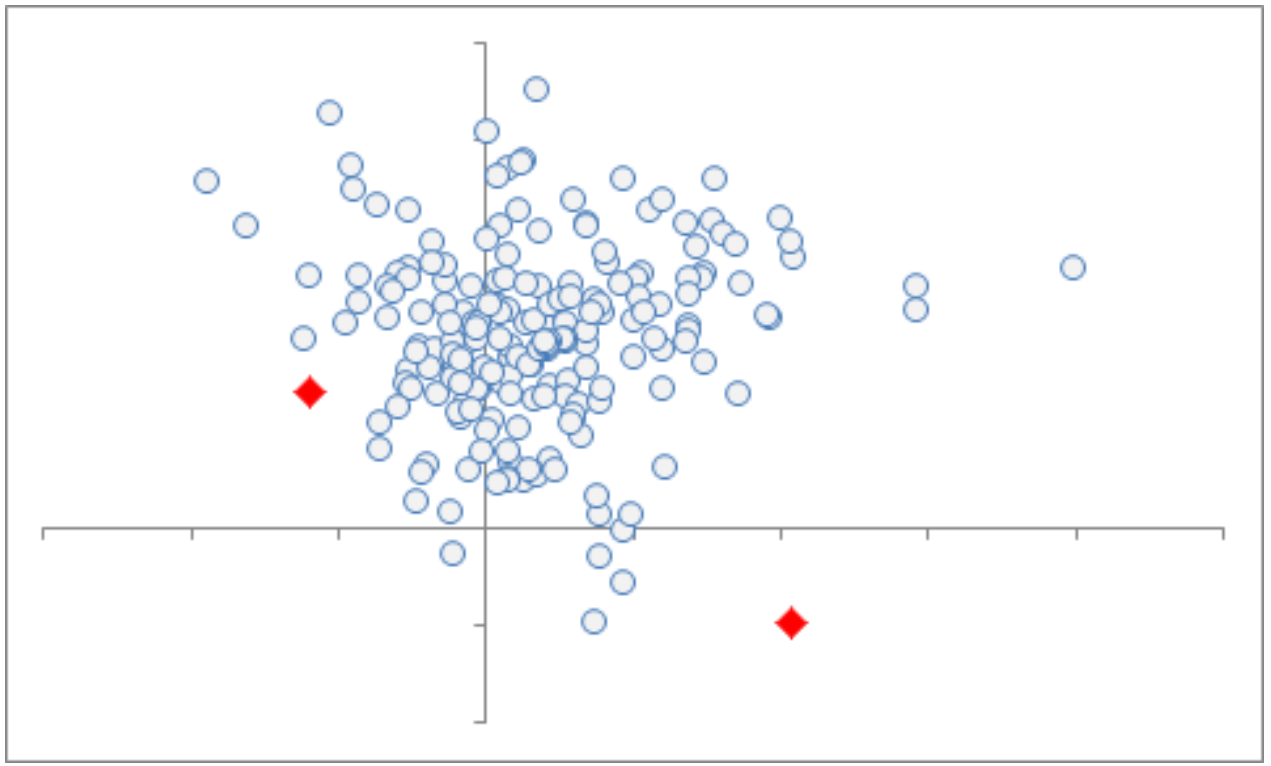
C_1 and C_2

k-means 클러스터 알고리즘의 원리

두개의 centroid의 위치를 랜덤으로 초기화 합니다.

μ_1 and μ_2

k-means 클러스터 알고리즘의 원리



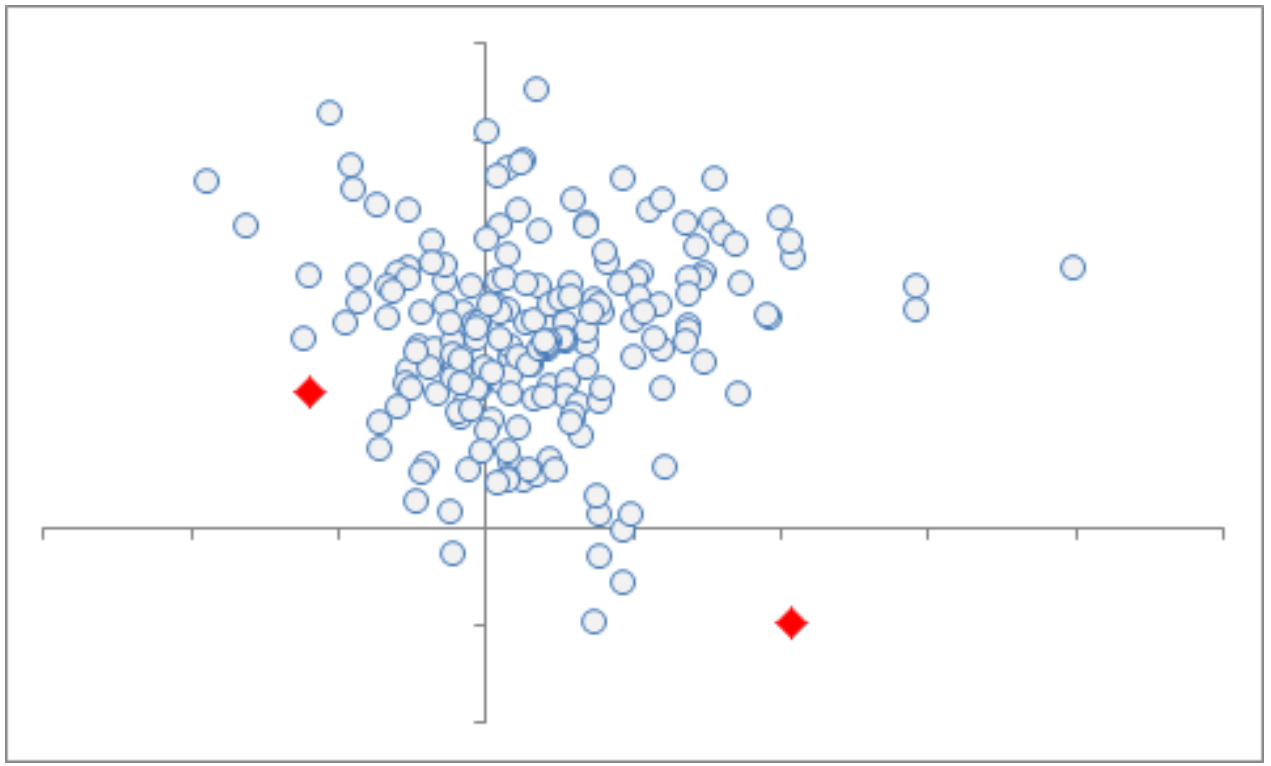
k-means 클러스터 알고리즘의 원리

그리고서는, centroid와 데이터 좌표 사이의 제곱 거리(*)가 최소화 되는 방향으로 centroid의 위치를 갱신 시킵니다.

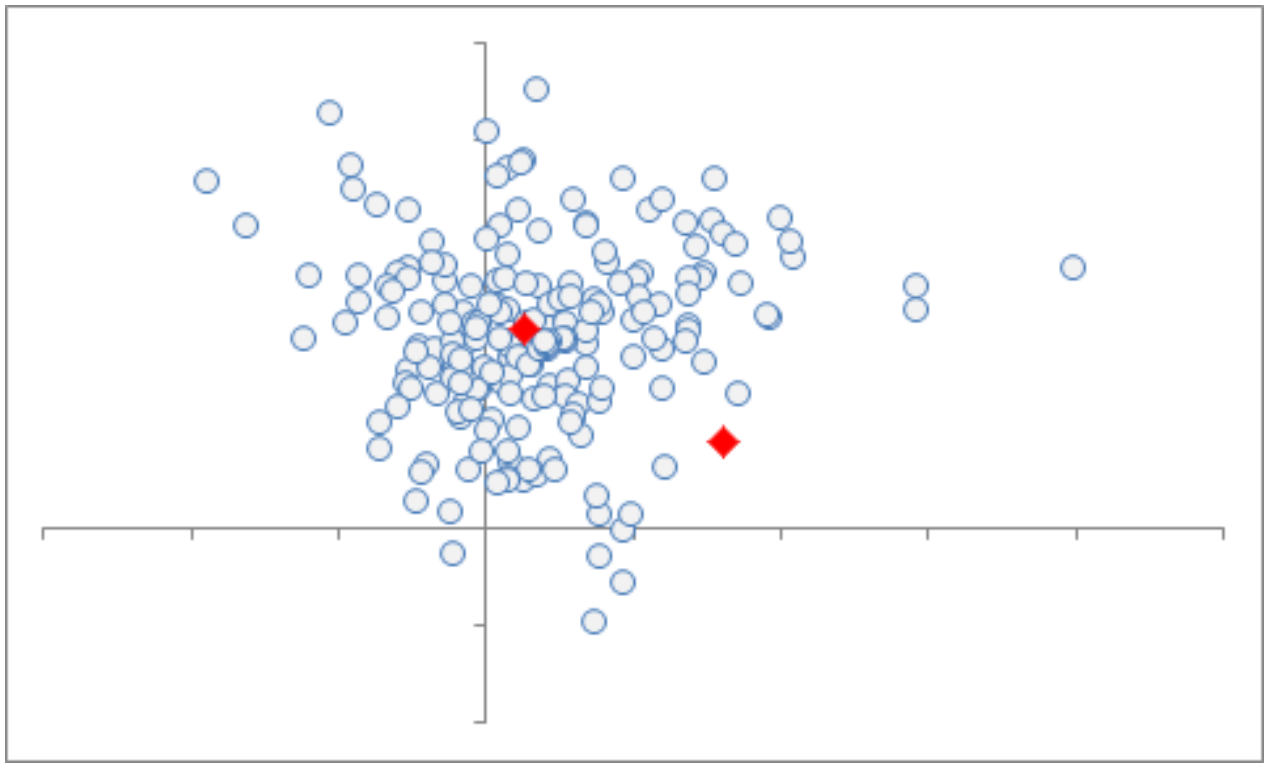
$$minimize \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

(*) “Sum of Square Distance Within”이라고 불리웁니다.

k-means 클러스터 알고리즘의 원리

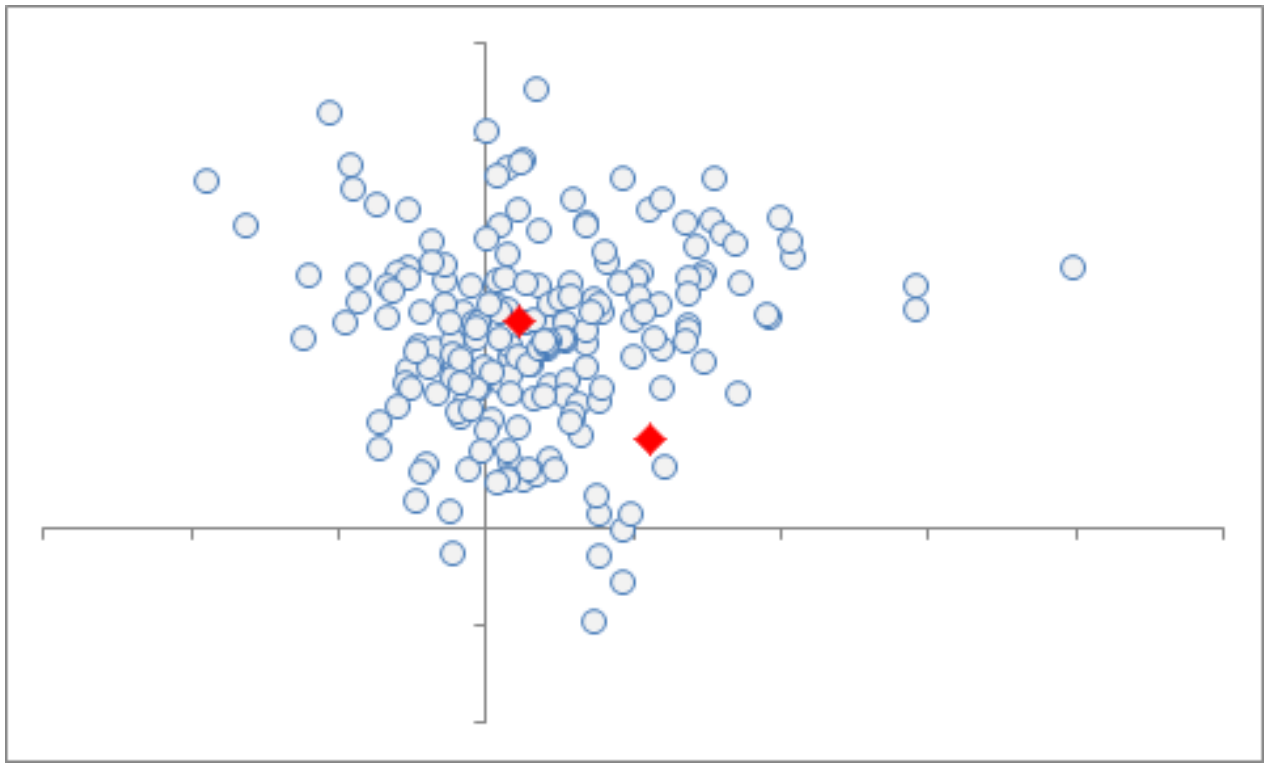


k-means 클러스터 알고리즘의 원리



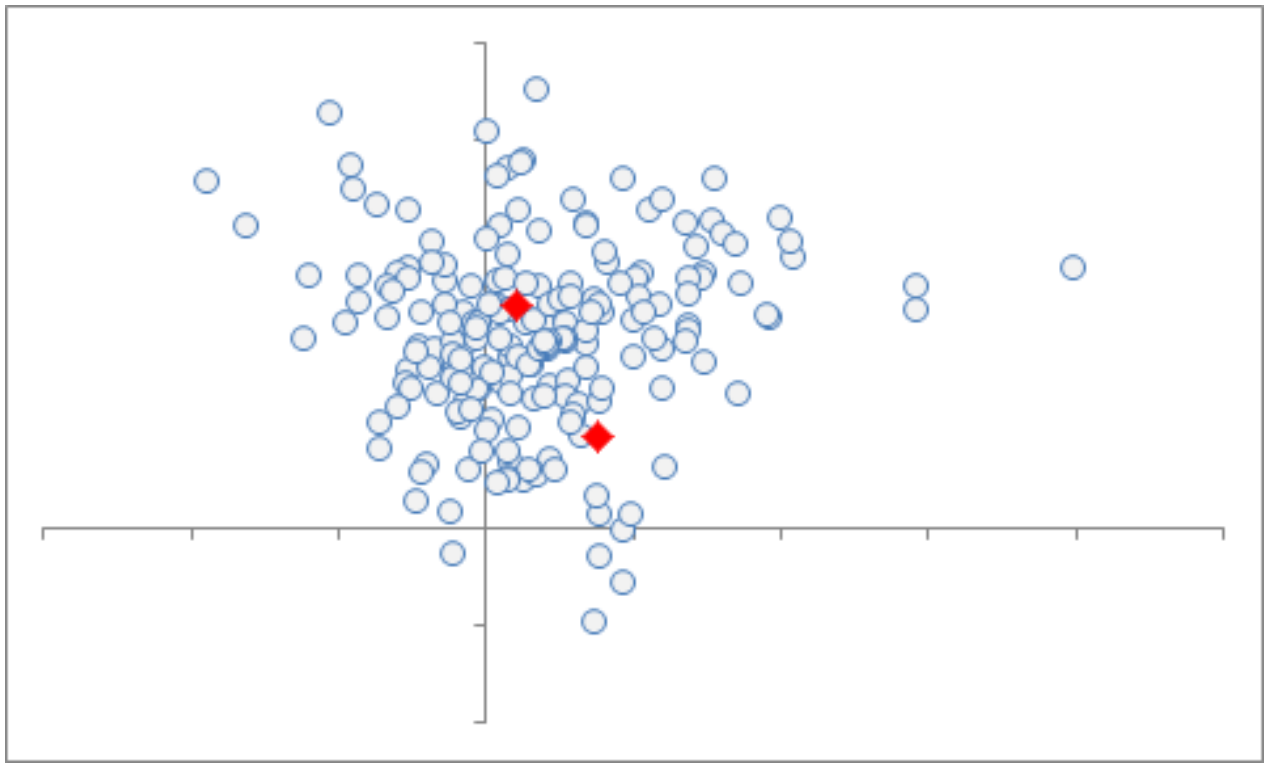
반복적 갱신.

k-means 클러스터 알고리즘의 원리



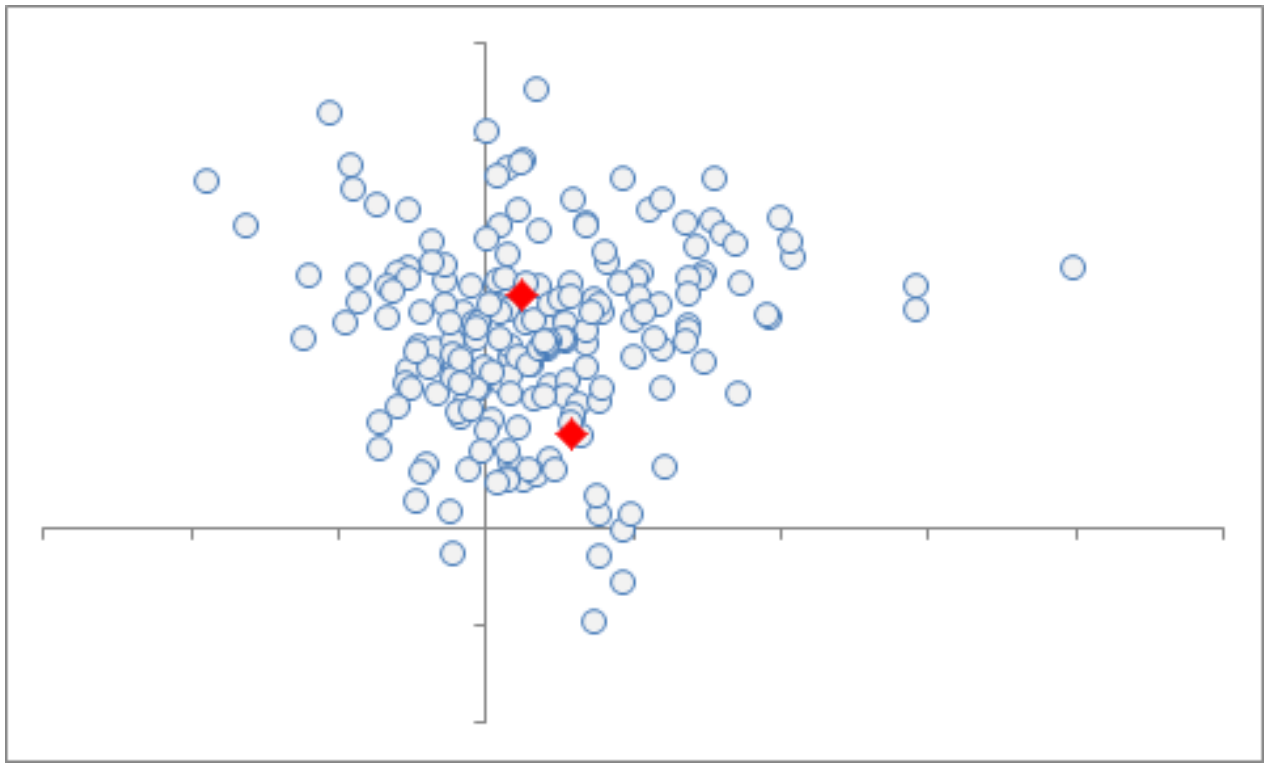
반복적 갱신.

k-means 클러스터 알고리즘의 원리



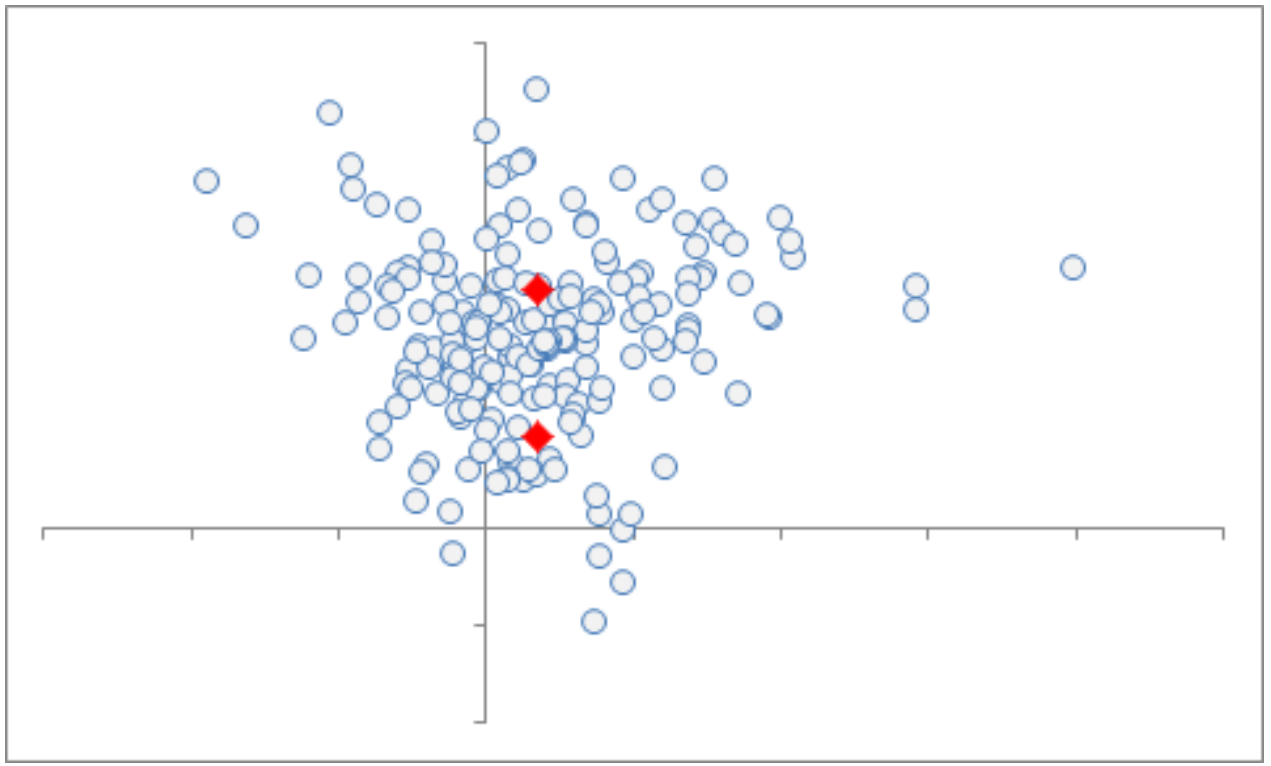
반복적 갱신.

k-means 클러스터 알고리즘의 원리



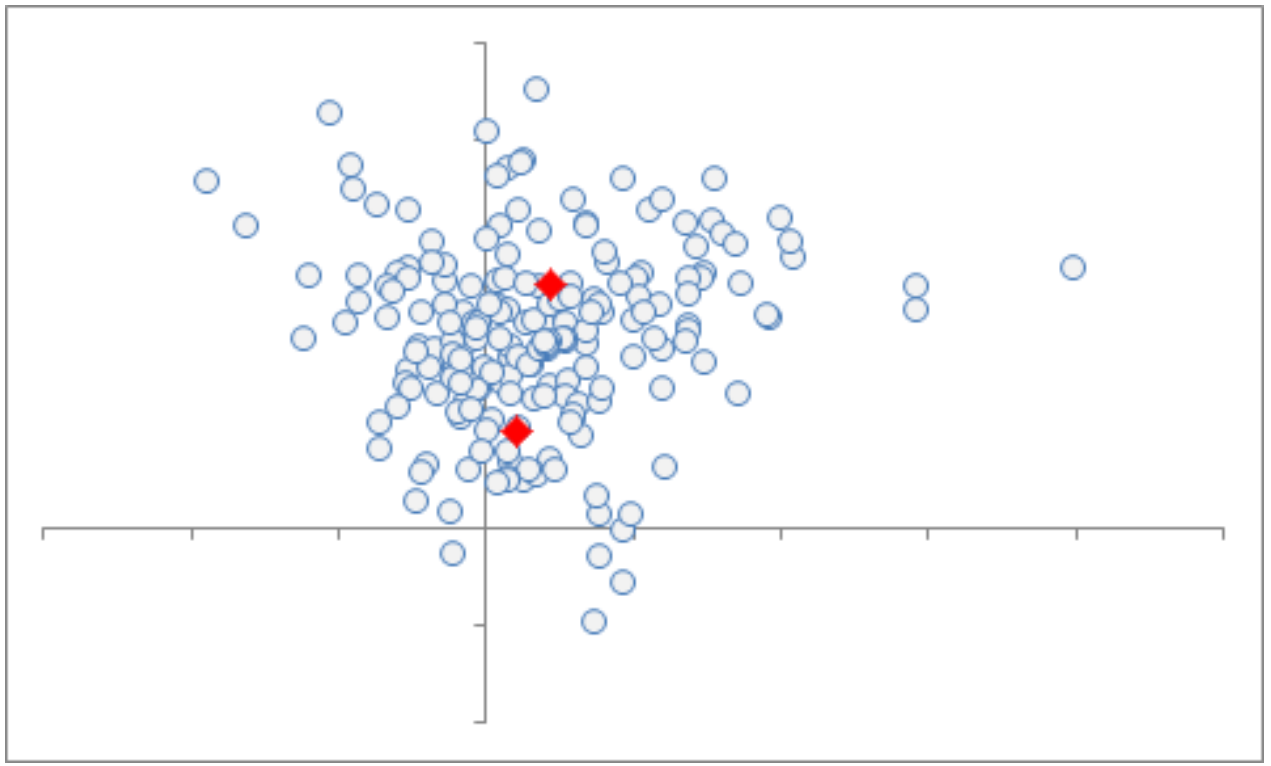
반복적 갱신.

k-means 클러스터 알고리즘의 원리



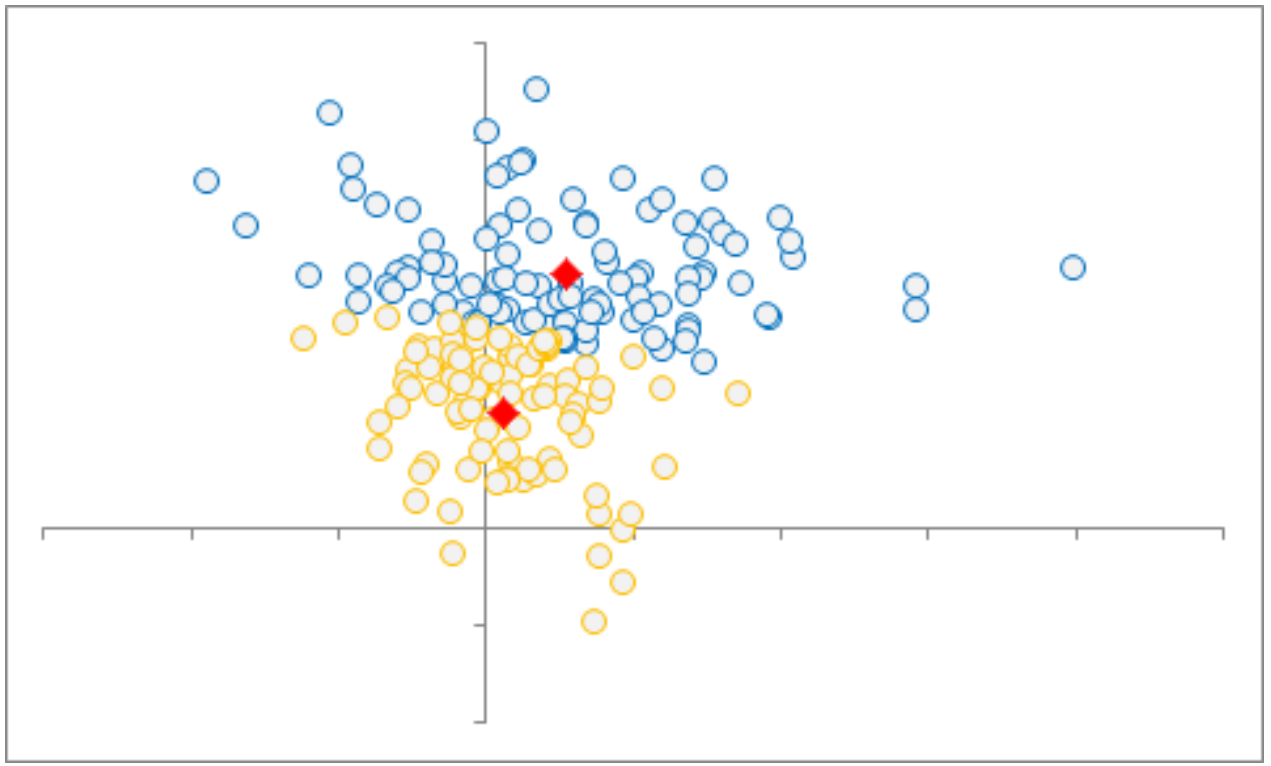
반복적 갱신.

k-means 클러스터 알고리즘의 원리



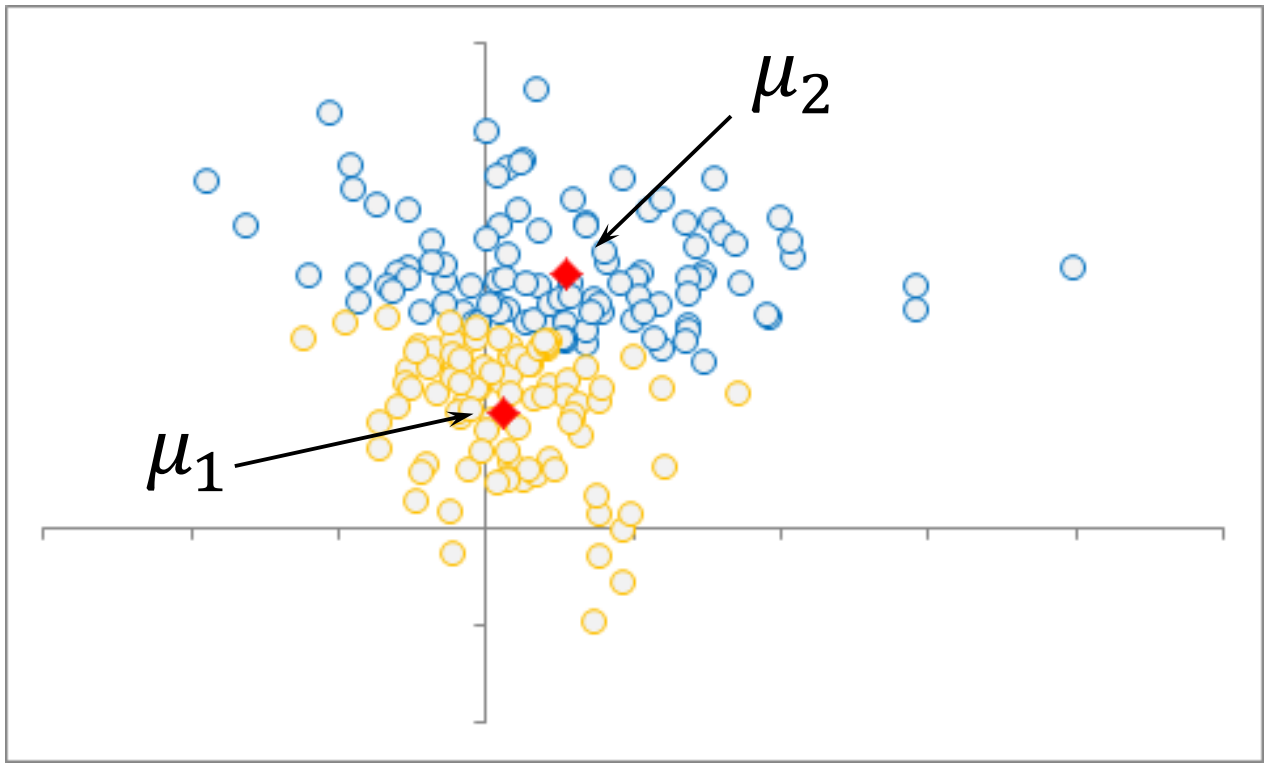
반복적 갱신.

k-means 클러스터 알고리즘의 원리



드디어 수렴!

k-means 클러스터 알고리즘의 원리



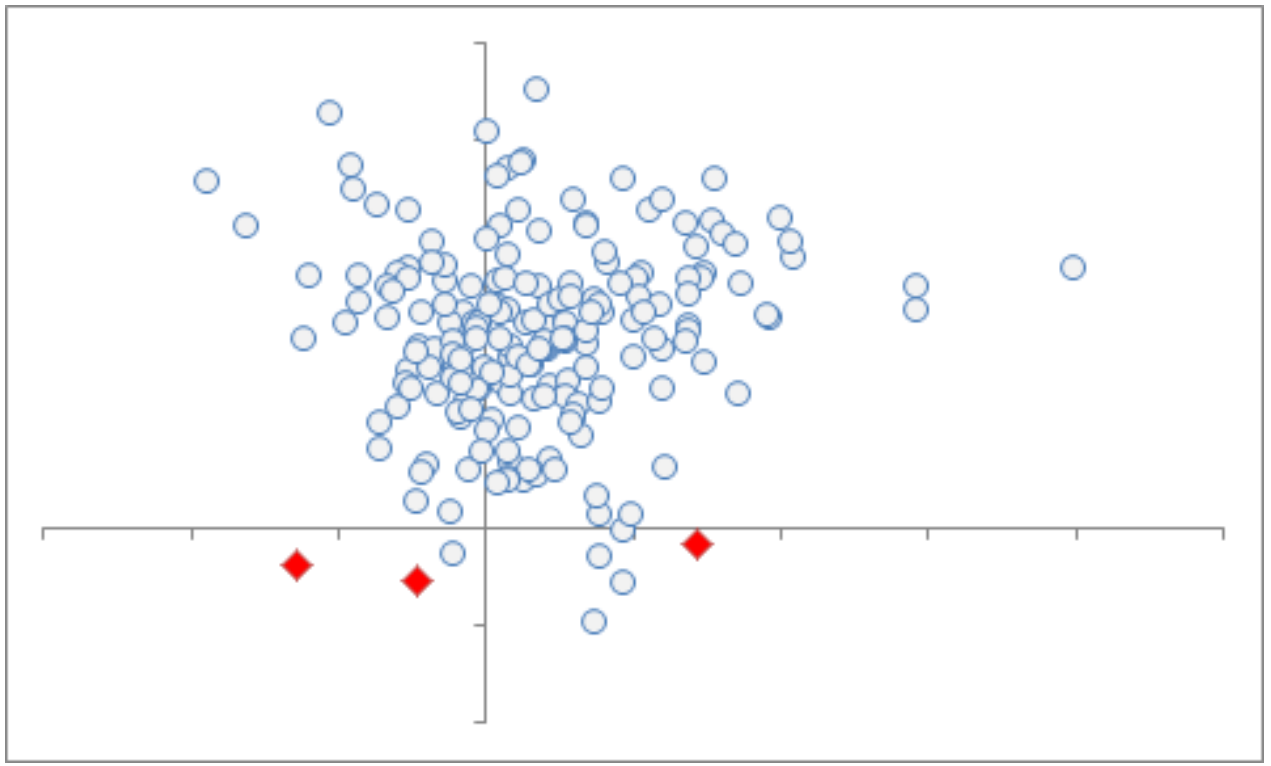
Centroid의 최종 위치.

k-means 클러스터 알고리즘의 원리

이제는 세개 ($k = 3$) 의 클러스터를 목표로 설정해 봅니다.

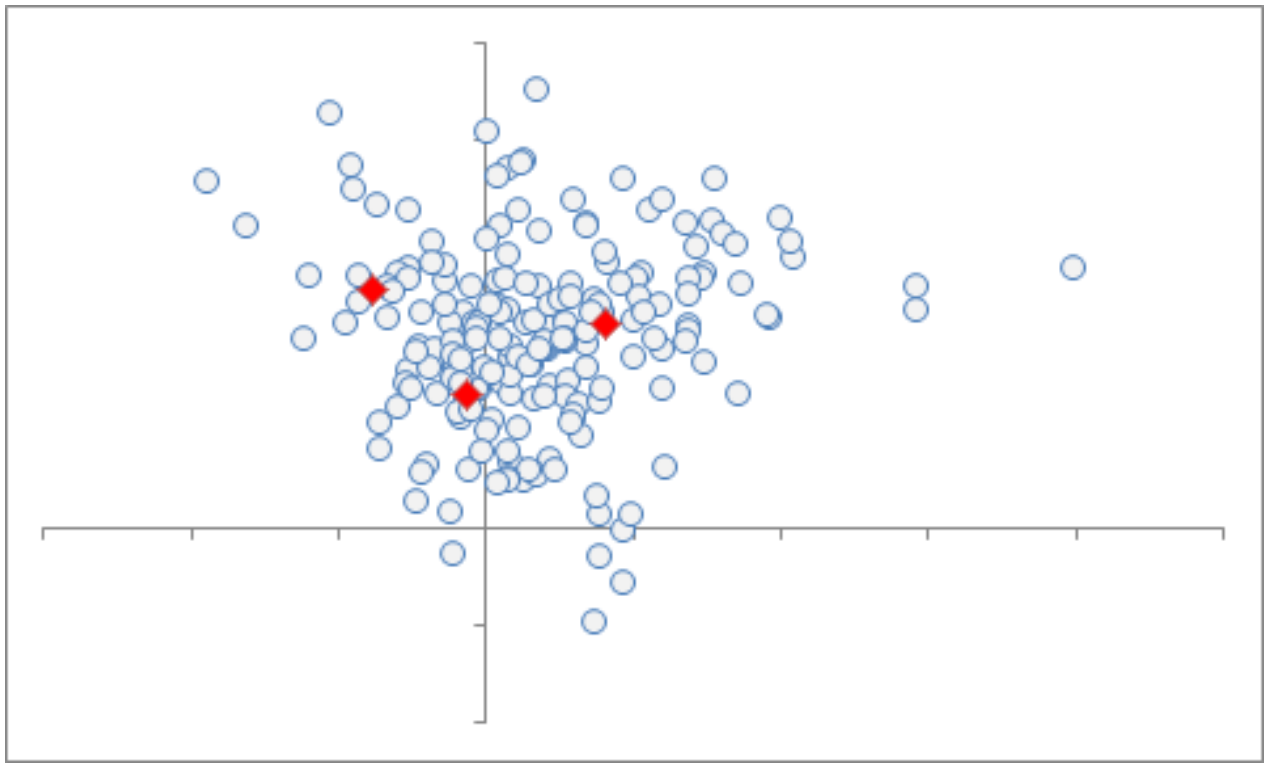
C_1 , C_2 and C_3

k-means 클러스터 알고리즘의 원리



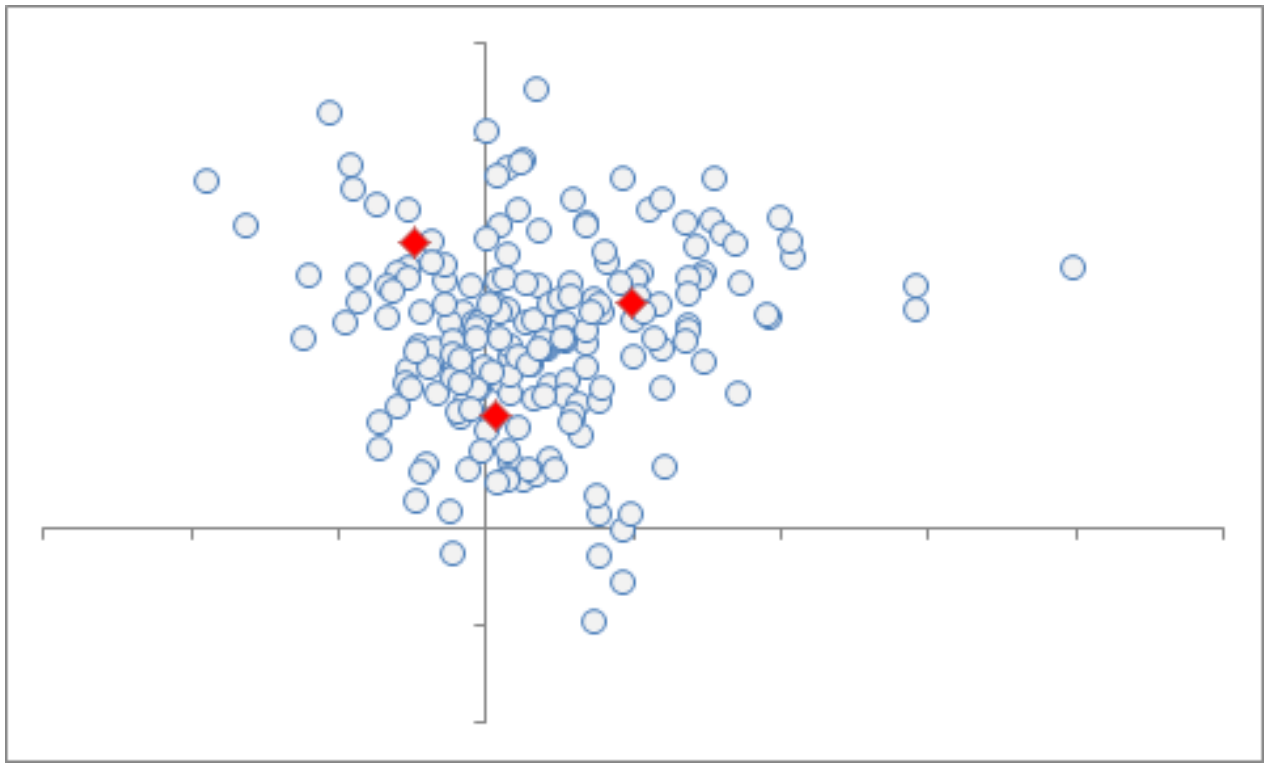
랜덤 초기화.

k-means 클러스터 알고리즘의 원리



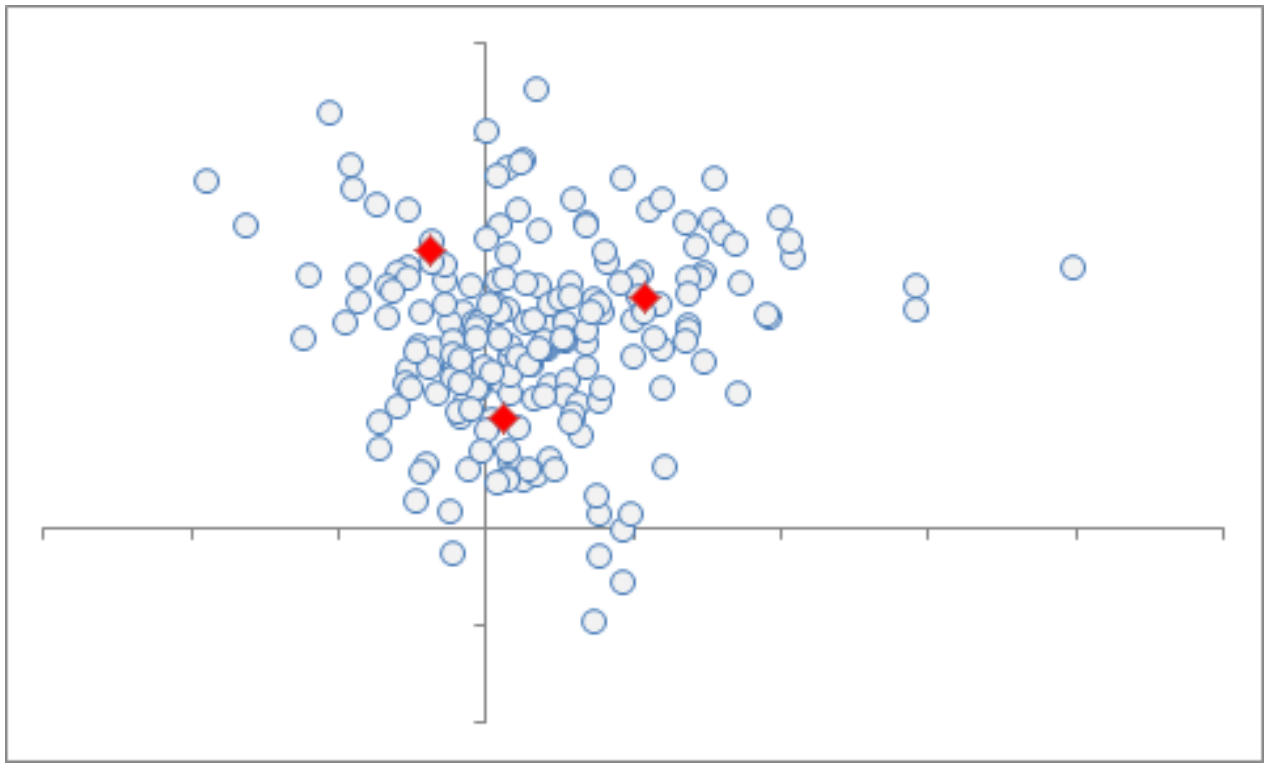
반복적 갱신.

k-means 클러스터 알고리즘의 원리



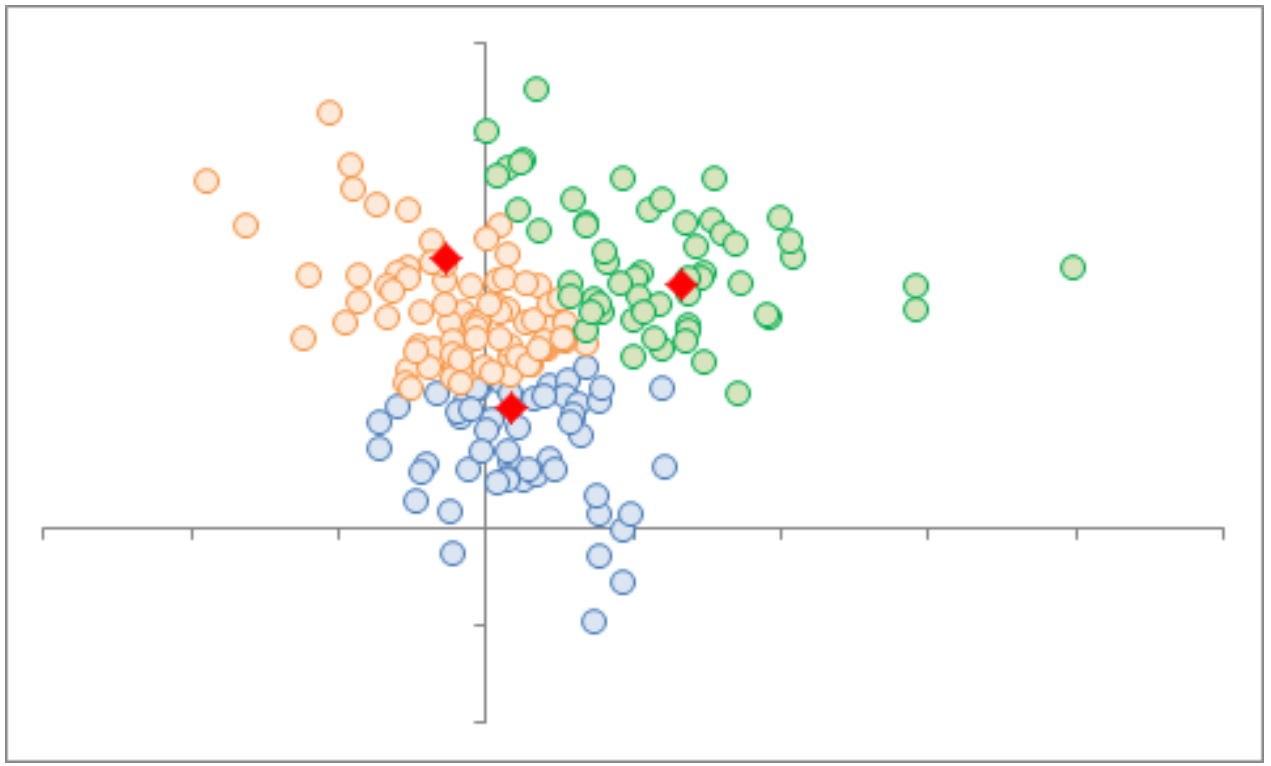
반복적 갱신.

k-means 클러스터 알고리즘의 원리



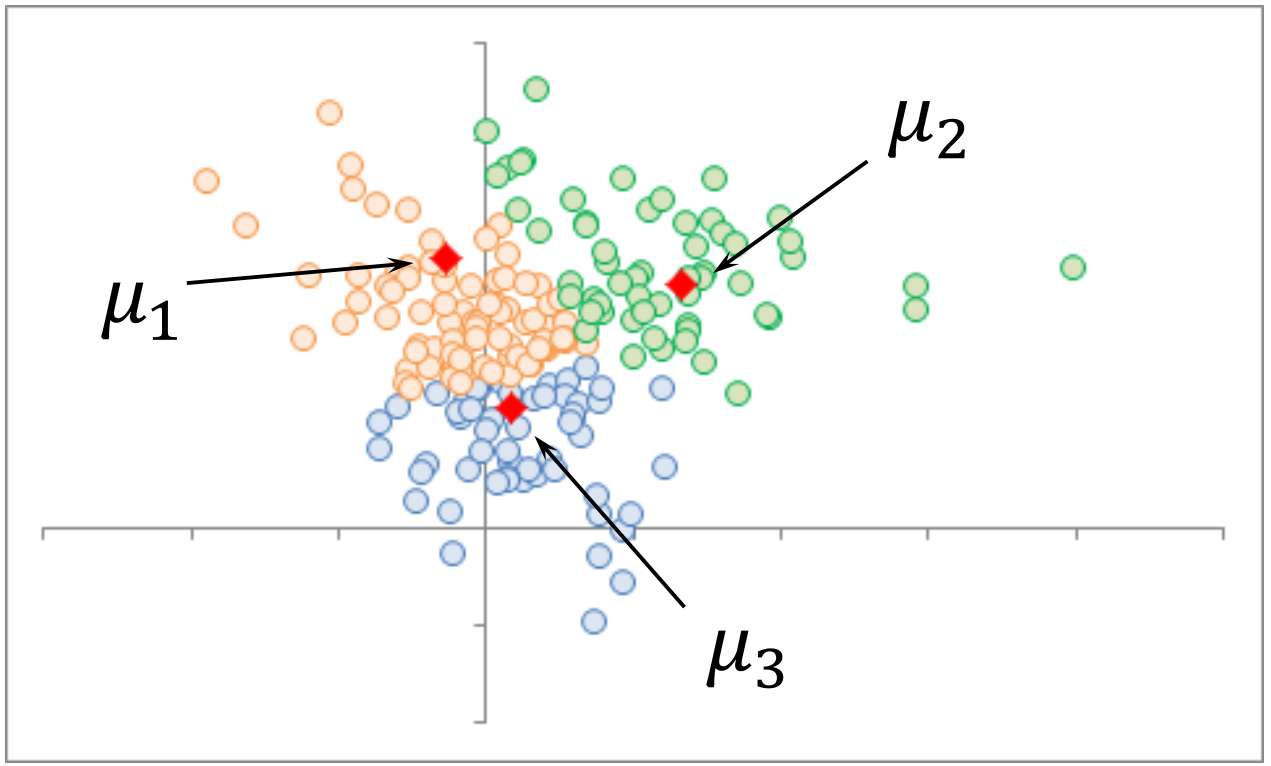
반복적 갱신.

k-means 클러스터 알고리즘의 원리



드디어 수렴!

k-means 클러스터 알고리즘의 원리



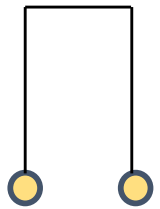
Centroid의 최종 위치.

계층적 군집화

계층적 군집화 (hierarchical clustering):

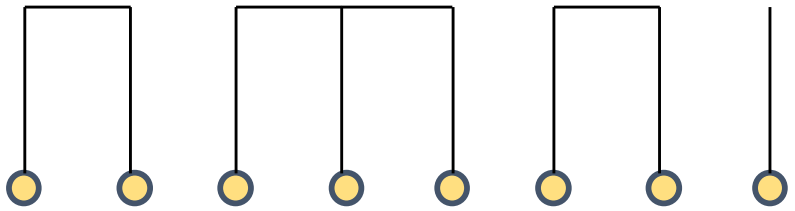
- 자율 학습.
- 병합군집 (agglomerative clustering) 알고리즘의 대표.
- 가까운 아이템끼리 순서대로 뭉쳐가는 형식.
- 위아래가 역전된 나무(tree)의 형상을 보임.

계층적 군집화



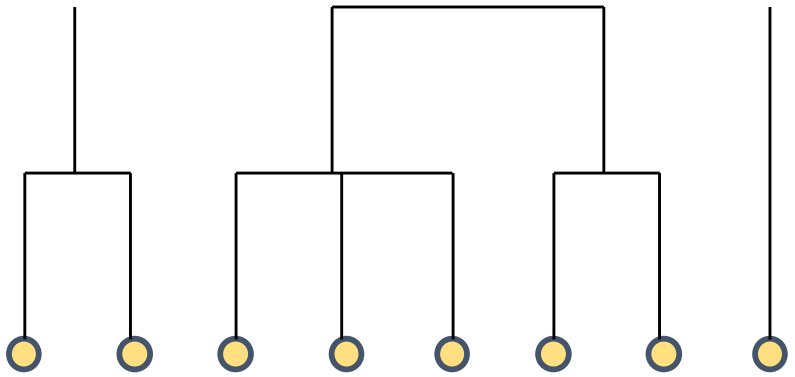
가장 가까운 아이템을 연결.

계층적 군집화



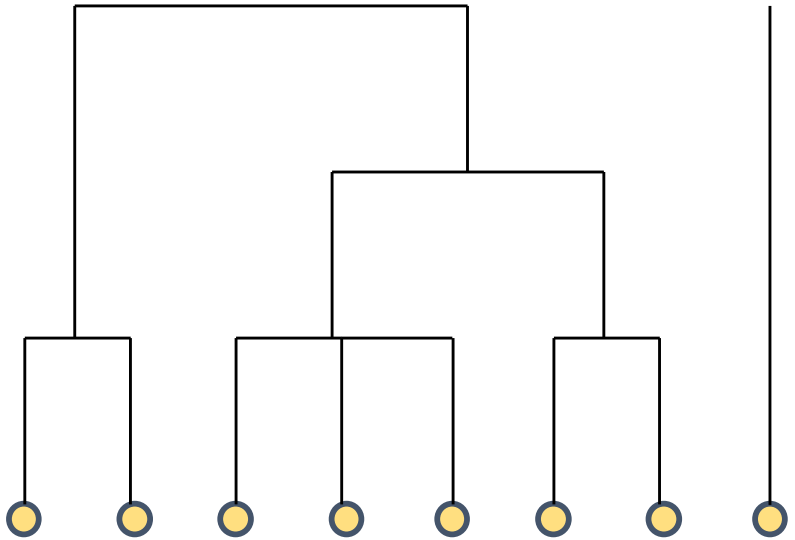
다양한 크기의 군집이 형성됨.

계층적 군집화



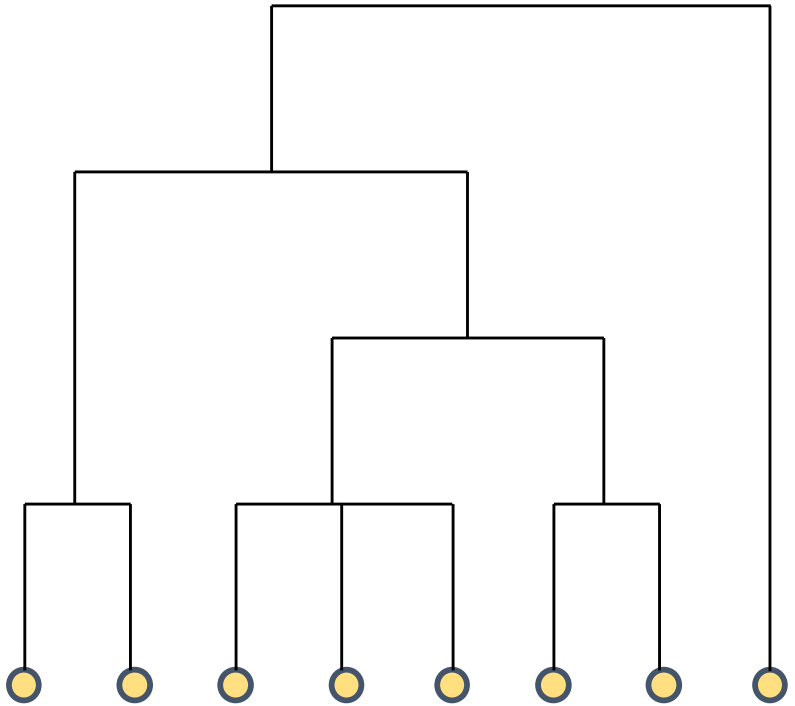
이제는 가까운 군집끼리 연결함.

계층적 군집화



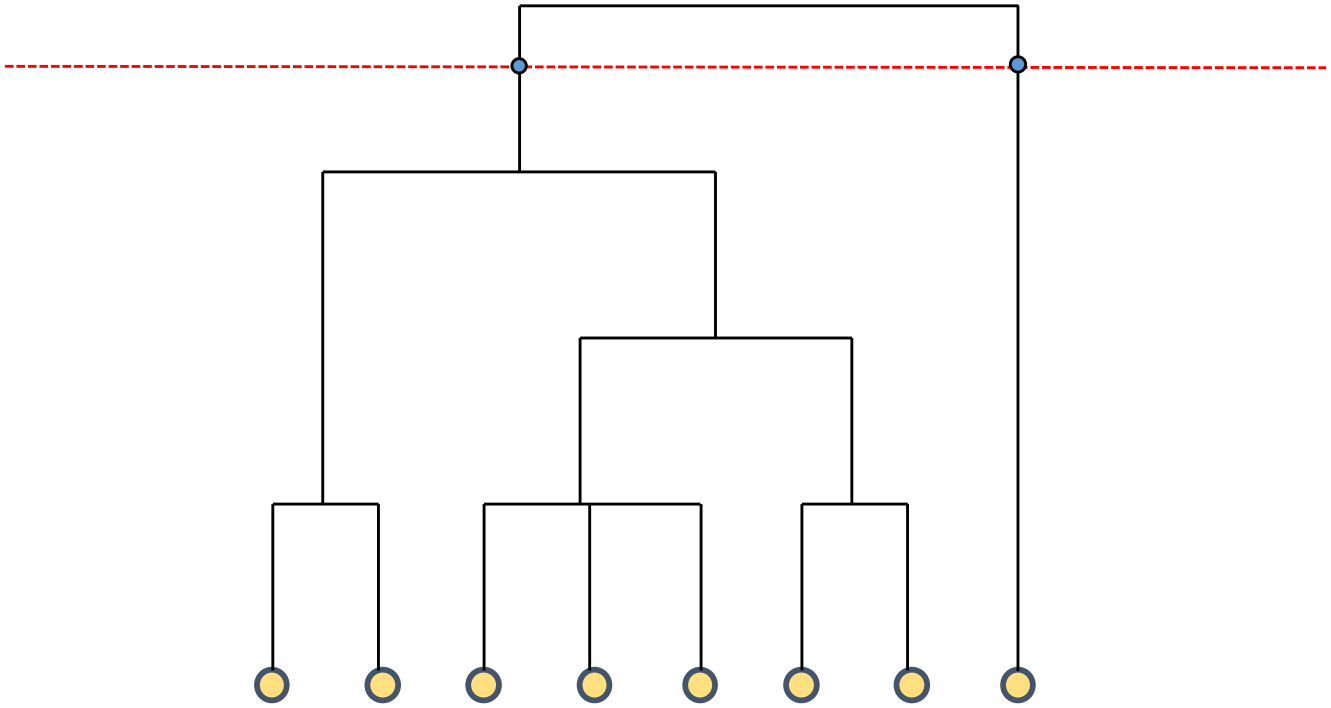
가까운 군집끼리 연결함.

계층적 군집화



궁극적으로는 하나의 덩어리(군집)으로 뭉침.

계층적 군집화



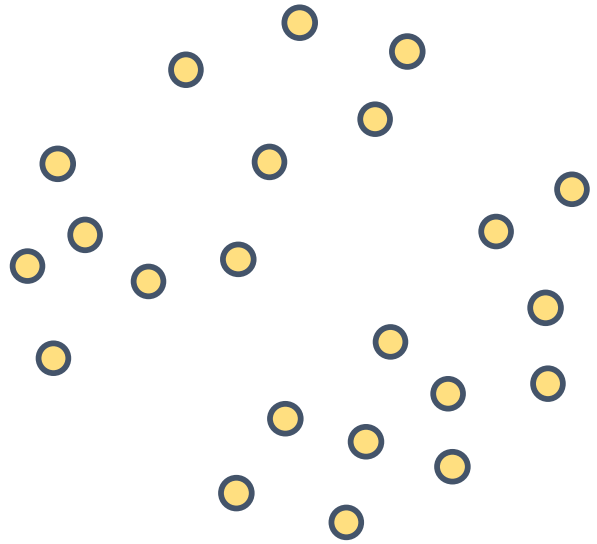
임의의 “높이”를 정하고 살펴봄.

DBSCAN 군집화

DBSCAN 군집화:

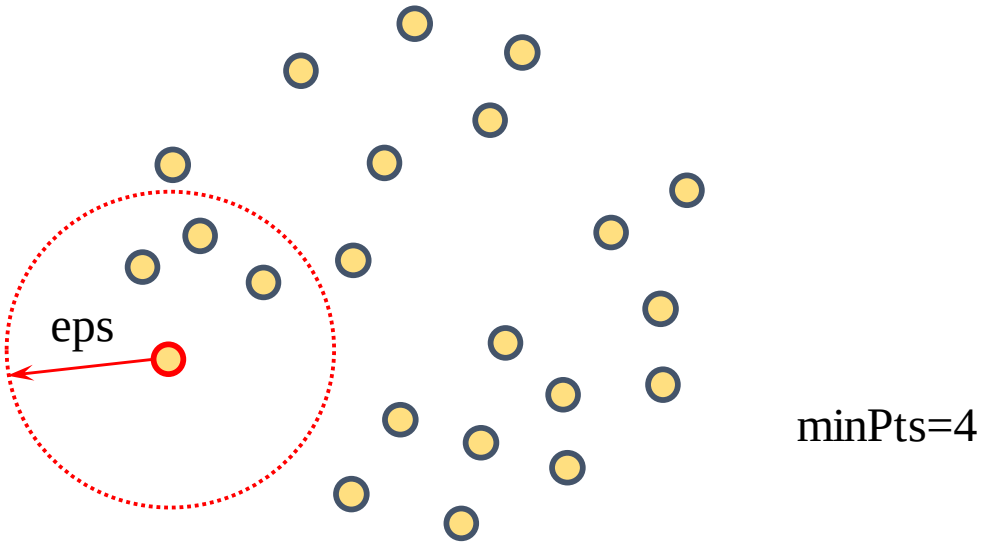
- 자율 학습.
- 1996년에 개발되어서 가장 효과적이고 많이 사용되는 군집분석방법.
- 밀도 (density)에 따라서 군집을 만들어 감.
- 엡실론(eps), 최소밀도(minPts), 등과 같은 파라미터를 정해 주어야 한다.
- 고밀도 지역이 연결되어 있으면 만족스러운 군집화가 어렵다는 단점이 있다.

DBSCAN 군집화



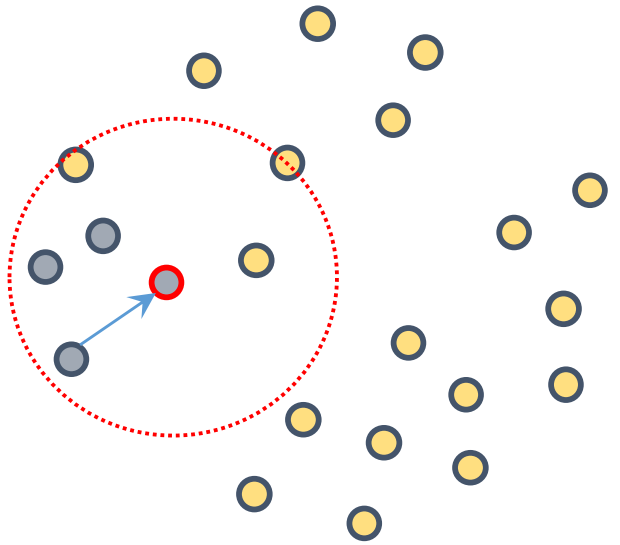
위와 같이 좌표가 분포되어 있다고 가정해 봅니다.

DBSCAN 군집화



반지름 eps 까지의 거리 안에 minPts이상의 좌표가 있는지 확인.

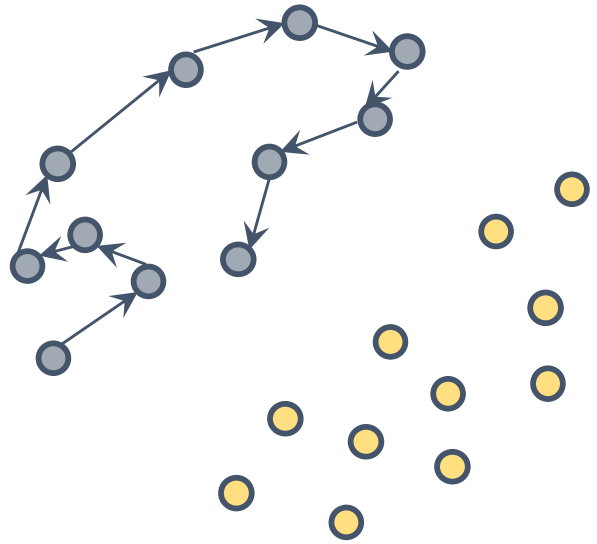
DBSCAN 군집화



minPts=4

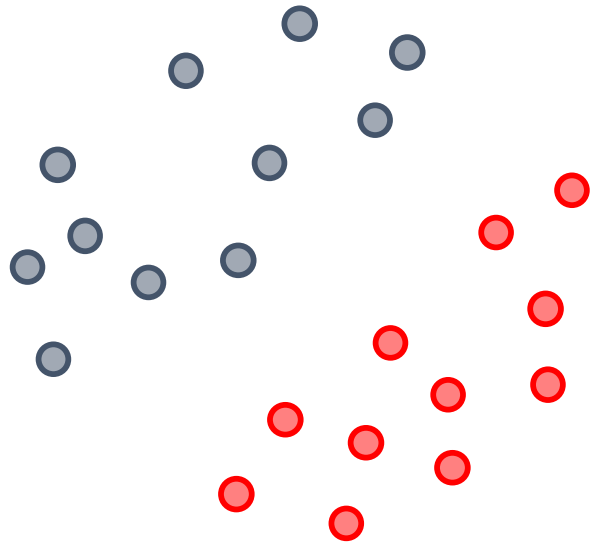
조건이 맞으면 다른점으로 옮겨가서 똑같은 조건 확인.

DBSCAN 군집화



반복적으로 실행해서 끊기지 않고 연결되는 좌표들이 군집을 형성함.

DBSCAN 군집화



완성!

t-SNE

t-SNE manifold learning:

- 주로 비지도 학습의 목적으로 사용된다.
- 군집화를 통해서 시각화를 향상시키는 알고리즘이다.
- 고차원 데이터를 2D 평면에 매핑하게 되는데 단순 투영은 아니다.
- 매핑 과정에서는 가까운 좌표끼리는 뭉치고 먼 좌표끼리는 더 떨어지게 만든다.

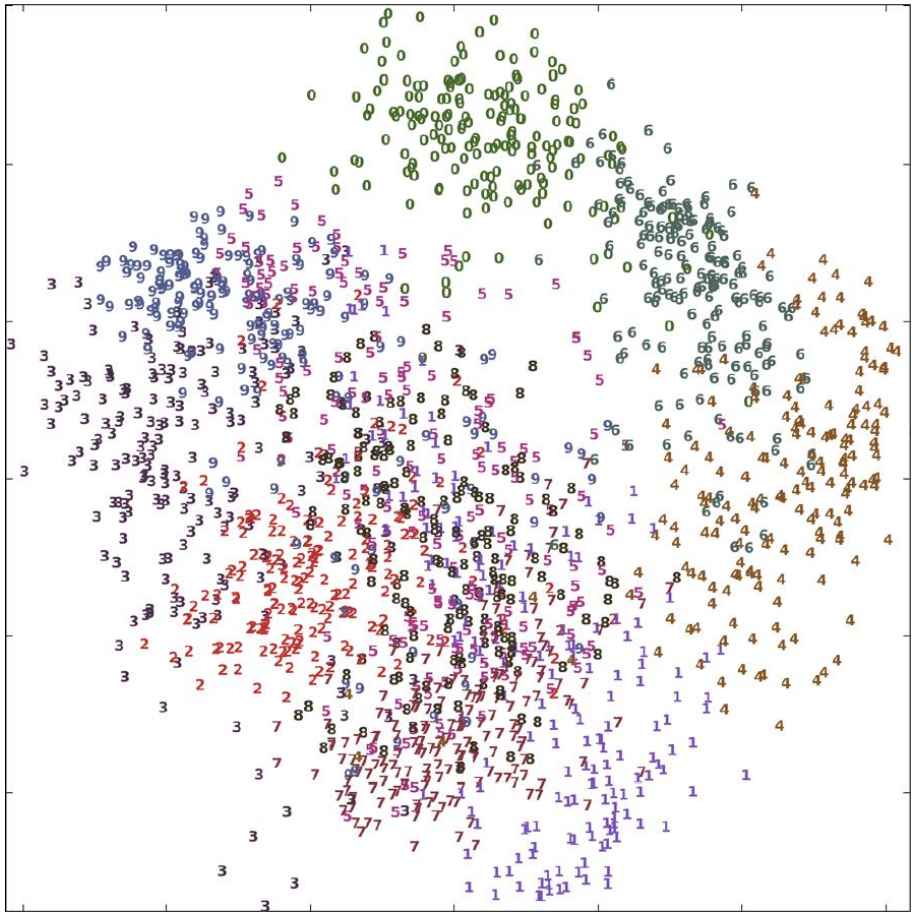
t-SNE

t-SNE manifold learning: 장단점

장점	단점
✓ PCA 사용 방법보다 더 만족스러운 시각화 결과를 보인다.	✓ 원 좌표가 변형된다. ✓ 좌표 자체의 의미가 퇴색된다.

t-SNE : PCA와 비교

t-SNE 과 PCA 비교: PCA의 주성분에 투영한 경우 군집들이 서로 뒤엉켜있다.



t-SNE : PCA와 비교

t-SNE 과 PCA 비교: t-SNE를 적용한 경우 군집들을 더욱 또렷히 분간할 수 있다.