

Part.05

Clustering

# | DBSCAN clustering

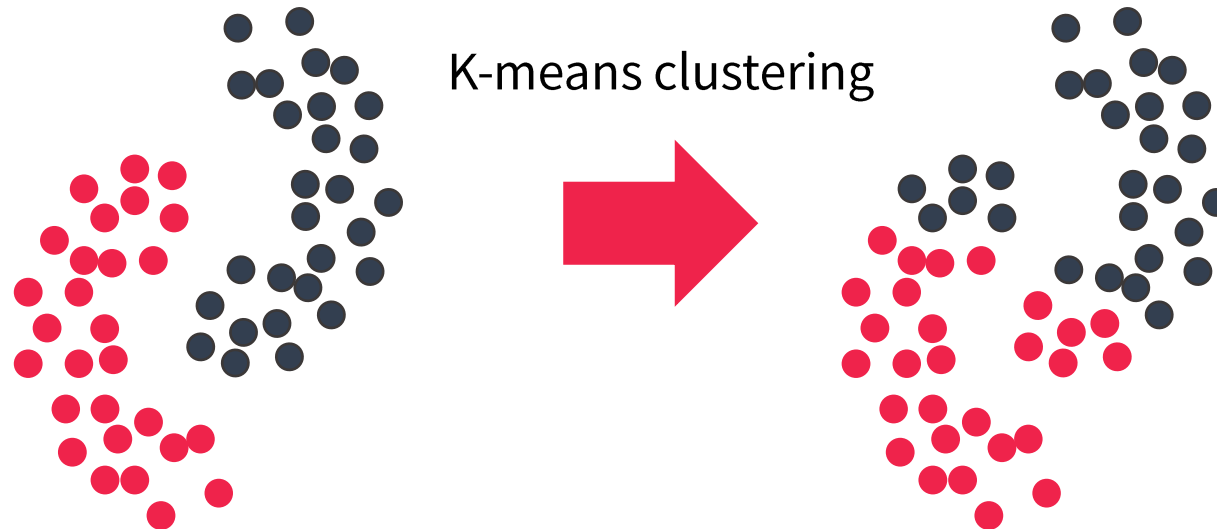
FASTCAMPUS  
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택

# I DBSCAN clustering 소개

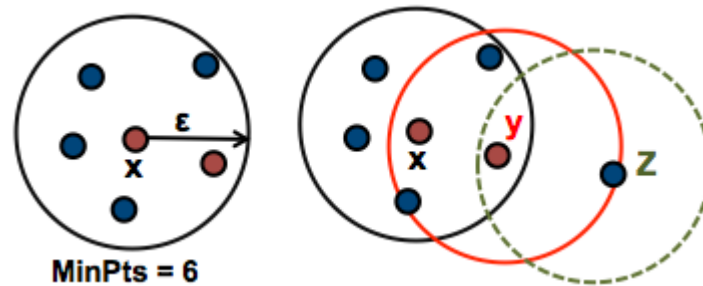
- K-means clustering의 단점
  - 초기 중심 값에 민감한 반응을 보임
  - 노이즈와 아웃라이어에 민감함
  - 군집의 개수 K를 설정하는 것에 어려움



# I DBSCAN clustering 소개

## ■ DBSCAN

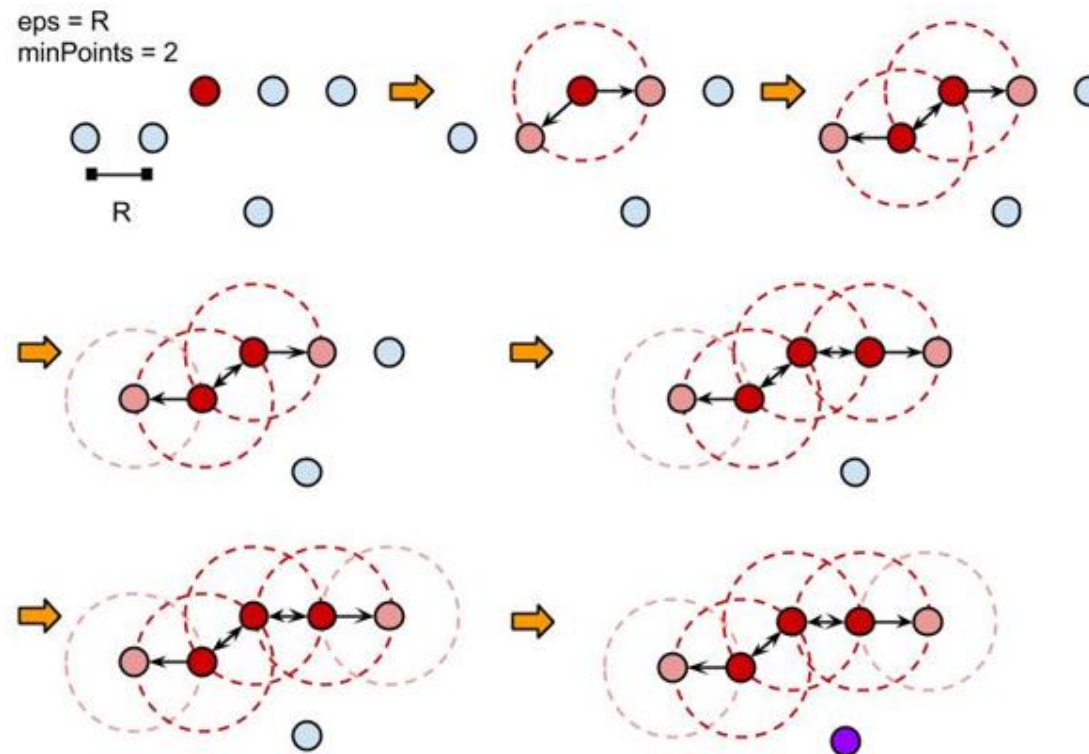
- Density-based Spatial Clustering of Applications with Noise
- 2014년 KDD 학회에서 상을 받은 알고리즘으로, density-based clustering 중 가장 유명하고 성능이 우수하다고 알려져 있음
- DBSCAN의 특징은 eps-neighbors와 MinPts를 사용하여 군집을 구성
  - Eps-neighbors: 한 데이터를 중심으로 epsilon( $\epsilon$ ) 거리 이내의 데이터들을 한 군집으로 구성
  - MinPts: 한 군집은 MinPts 보다 많거나 같은 수의 데이터로 구성됨  
만약 MinPts 보다 적은 수의 데이터가 eps-neighbors를 형성하면 노이즈(noise)로 취급함



# I DBSCAN clustering 소개

## ■ DBSCAN

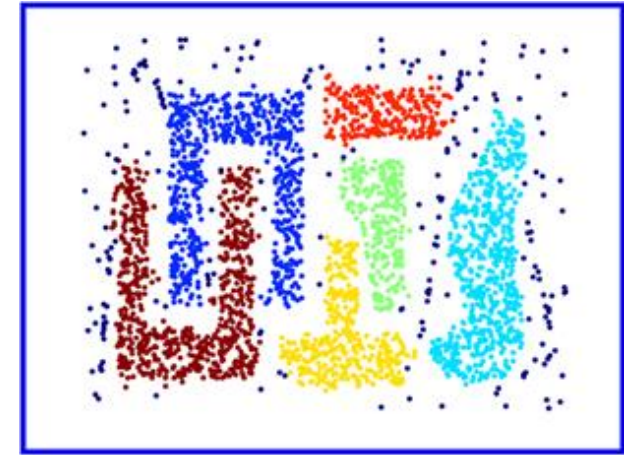
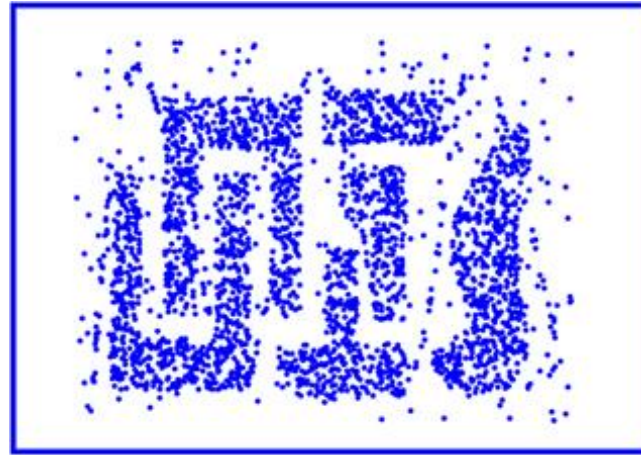
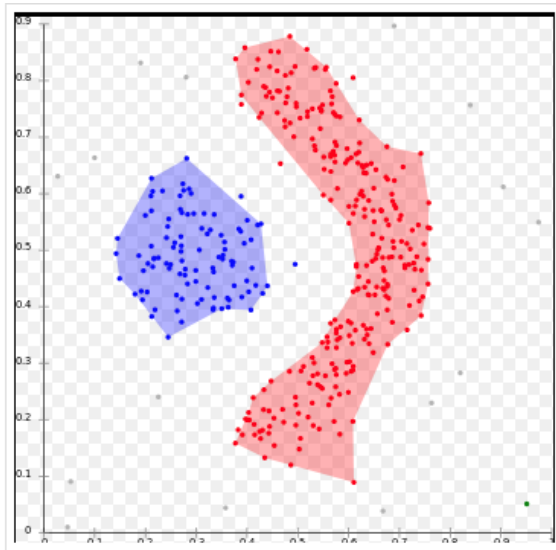
### DBSCAN 알고리즘 순서



# I DBSCAN clustering 소개

## ■ DBSCAN 예시

- 최소 거리  $\epsilon$  이내의 데이터들이 점진적으로 한 군집으로 합쳐지며 다양한 모양의 군집을 형성 (군집이 안 된 데이터는 노이즈로 취급) → outlier를 detection하는 분야에도 사용 가능



# I DBSCAN clustering 소개

- DBSCAN의 파라미터: MinPts, eps
  - DBSCAN은 군집 분석을 적용하고자 하는 데이터에 대한 이해도가 충분할 때 파라미터 설정이 쉬움
  - MinPts의 설정
    - 간단히 설정하는 경우에는 다음과 같음:  $\text{minPts} = \text{변수의 수} + 1$
    - MinPts는 3 이상으로 설정  
(1인 경우 데이터가 하나하나가 개별 군집 형성)
  - Eps의 설정
    - 너무 작은 경우, 상당 수의 데이터가 노이즈로 구분 될 수 있음
    - 너무 큰 경우, 군집의 수가 하나가 될 가능성이 있음
    - 일반적으로 K-nearest neighbor graph의 distances를 그래프로 나타낸 후 거리가 급격하게 **증가하는** 지점을 eps로 설정

# I DBSCAN clustering 소개

## ■ DBSCAN 장단점

### 장점

- ✓ K-means와 다르게 군집의 수를 설정할 필요가 없음
- ✓ 다양한 모양의 군집이 형성될 수 있으며, 군집끼리 겹치는 경우가 없음
- ✓ 노이즈 개념 덕분에 이상치에 대응이 가능
- ✓ 설정할 파라미터가 두 개(eps, minPts)로 적으며, 적용 분야에 대한 사전 지식이 있는 경우 비교적 쉽게 설정이 가능

### 단점

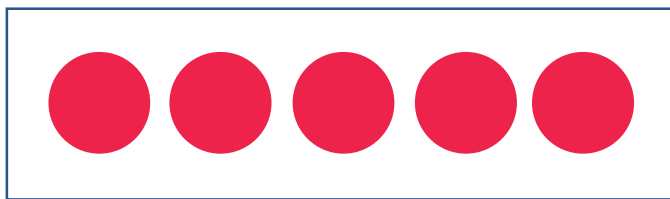
- ✓ 한 데이터는 하나의 군집에 속하게 되므로, 시작점에 따라 다른 모양의 군집이 형성됨
- ✓ Eps의 크기에 의해 DBSCAN의 성능이 크게 좌우됨
- ✓ 군집 별로 밀도가 다른 경우 DBSCAN을 이용하면 군집화가 제대로 이루어지지 않음

# I Clustering정리

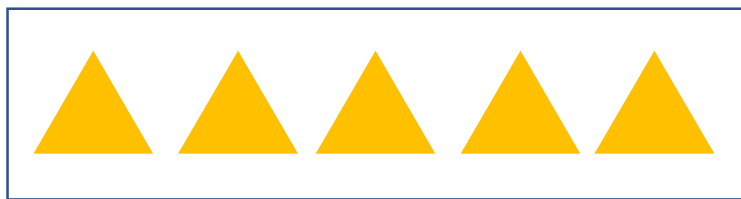
## ■ Clustering(군집분석)이란

- 각 데이터의 유사성을 측정하여 높은 대상 집단을 분류하고, 군집 간에 상이성을 규명하는 방법

전체데이터 (전체 뉴스기사)



군집1 (정치 관련 이슈)



군집2(스포츠 관련 이슈)



군집3(연예 관련 이슈)

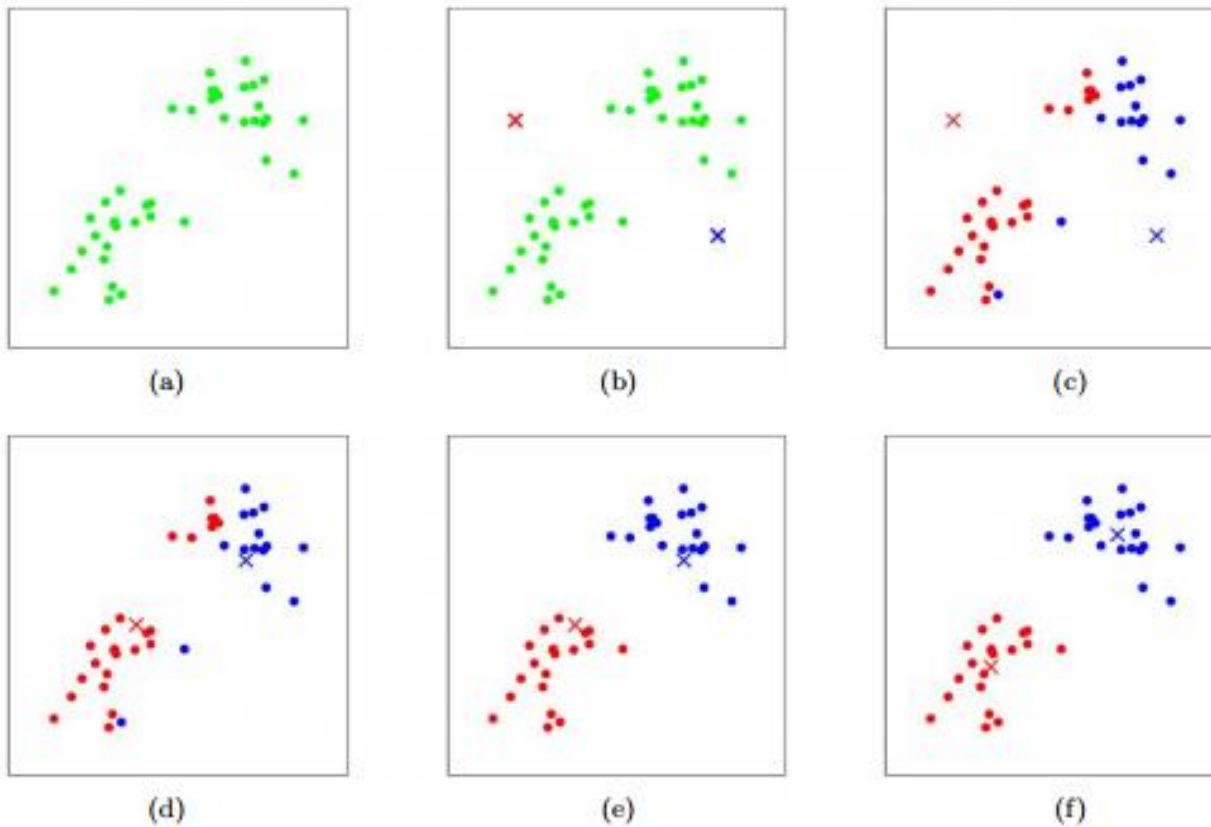
- 고객 segmentation을 통한 마케팅 활용 방안 / 군집 별 추가 분석수행



# I Clustering정리

## ■ K-means clustering

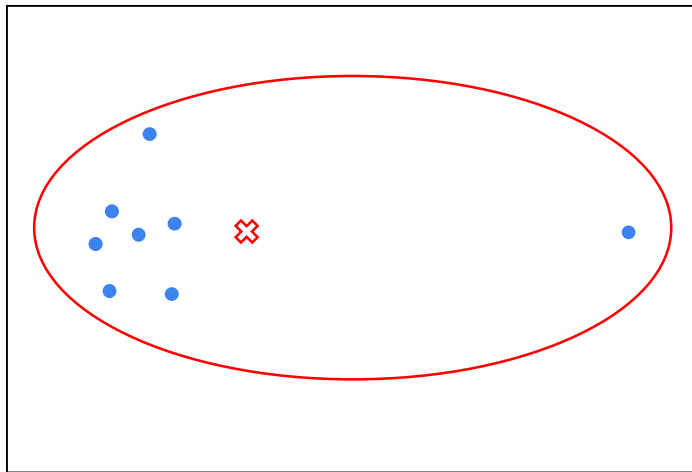
- Step1 – 각 데이터 포인트 i에 대해 가장 가까운 중심점을 찾고, 그 중심점에 해당하는 군집 할당
- Step2 – 할당된 군집을 기반으로 새로운 중심 계산, 중심점은 군집 내부 점들 좌표의 평균(mean)으로 함
- Step3 – 각 클러스터의 할당이 바뀌지 않을 때 까지 반복



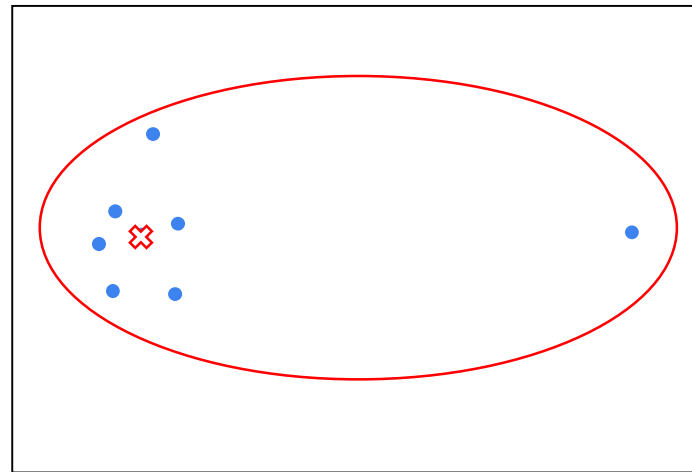
# I Clustering정리

## ■ K-medoids clustering

- K-means clustering의 변형으로, 군집의 무게 중심을 구하기 위해 데이터의 평균 대신 중간점(medoids)을 사용 (K-means보다 이상치에 강건한 성능을 보임)
- 아래 그림의 결과를 보면 K-medoids의 중앙점이 더 명확함 (이는 더 좋은 군집을 형성하게 될 가능성을 높임)



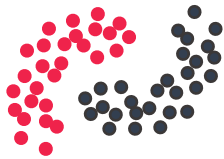
(a) Mean



(b) Medoid

# I Clustering정리

- K-means vs K-medoids

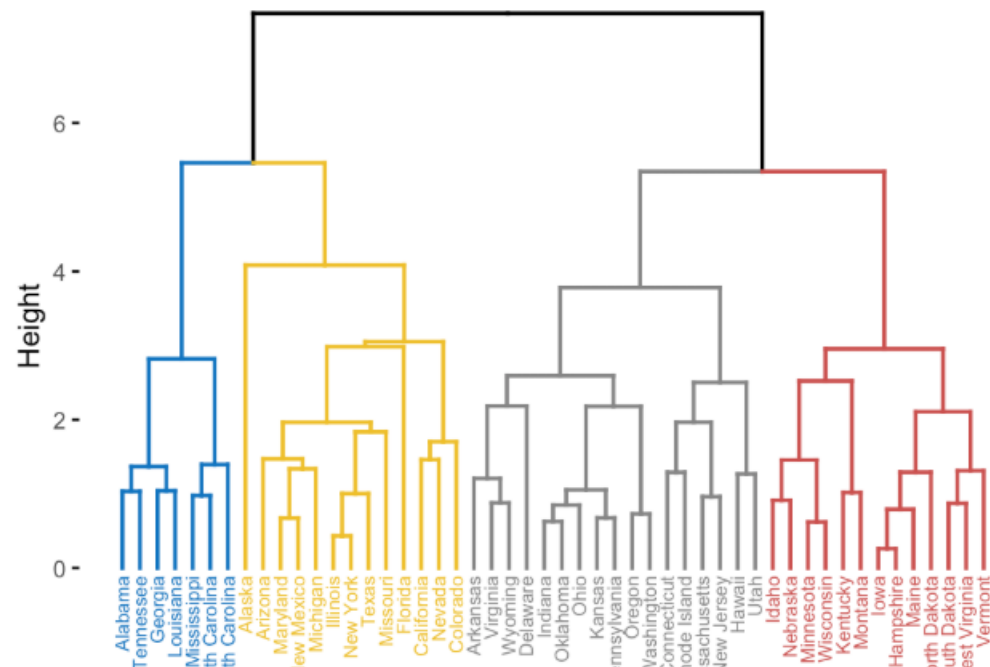
	K-means	K-medoids
중심	군집의 평균 값	군집 내 중앙 데이터
이상치	이상치가 전체 거리 평균 값에 영향을 주어 이상치에 민감함	K-means보단 덜 민감함
계산 시간	상대적으로 적은 시간이 소요	데이터 간 모든 거리 비용을 반복하여 계산해야 하므로 상대적으로 많은 시간이 소요
파라미터	군집의 개수 k, 초기 중심점	
군집 모양	원형의 군집이 아닌 경우 군집화를 이루기 어려움(아래 그림 참조) 	

# I Clustering정리

## ■ Hierarchical clustering

- 개체들을 가까운 집단부터 순차적/계층적으로 차근차근 묶어 나가는 방식
- 유사한 개체들이 결합되는 dendrogram 을 통해 시각화 가능
- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 cluster 형성
- 유사도 행렬 update

Cluster Dendrogram



Part.06

Class Imbalanced Problem

# | Class Imbalanced Problem이란

FASTCAMPUS  
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택