데이터마이닝팀

4팀

이진모
이은서
임주은
박지민
장이준

CONTENTS

Clustering

클러스터링

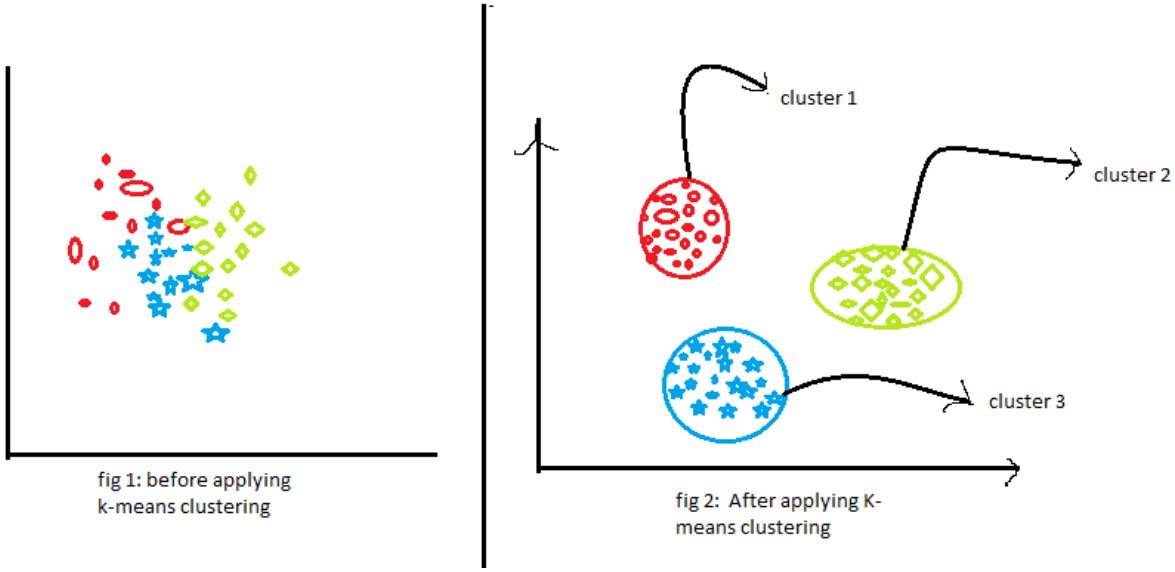


fig 1: before applying k-means clustering

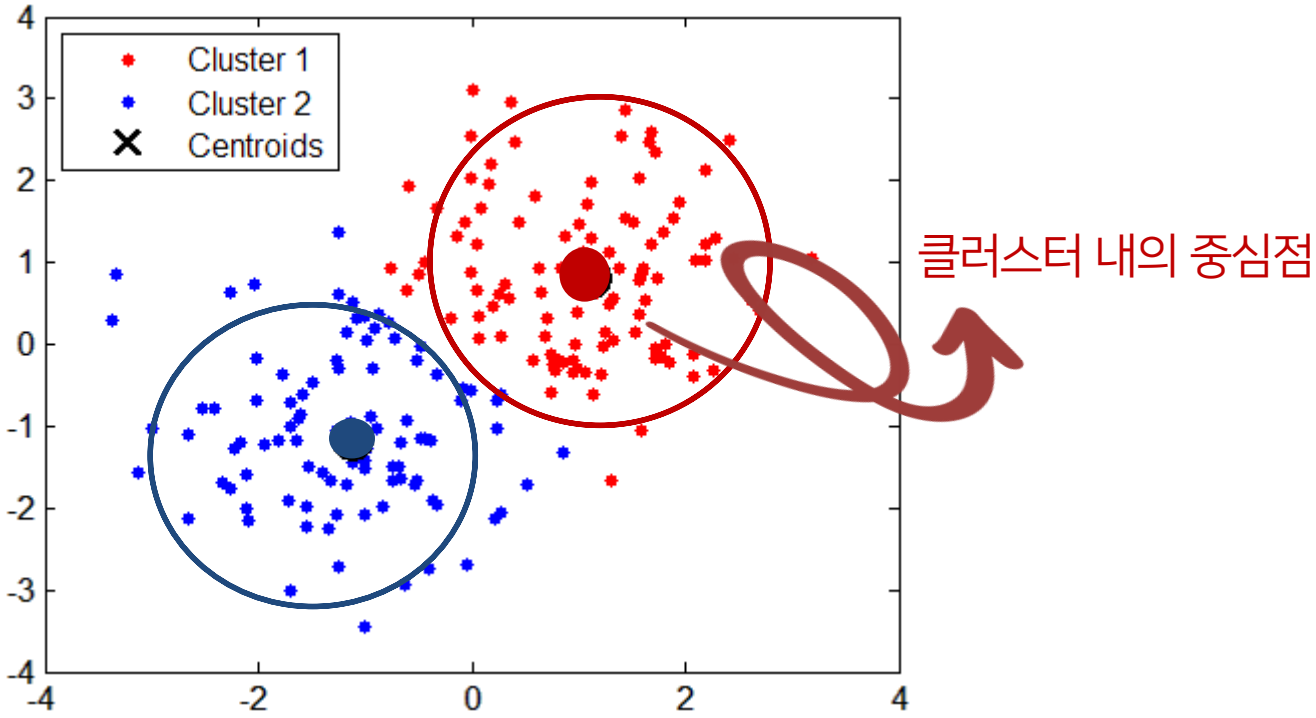
fig 2: After applying K-means clustering

주어진 데이터들의 특성을 고려해 개체들을

몇 개의 클러스터(부분 그룹)로 나누는 과정을 의미

→ 예측, 분류와 같은 명확한 목적이 아닌 탐색적 상황에 적합

K-Means Clustering



클러스터 내의 중심점과 클러스터 내 값들 사이 거리 분산을 최소화

K-Means Clustering

$$WCSS = \sum_{k=1}^{k} n_k \sum_{C(i)=k} \|x_i - \overline{x_k}\|^2$$

k번째 obs 개수

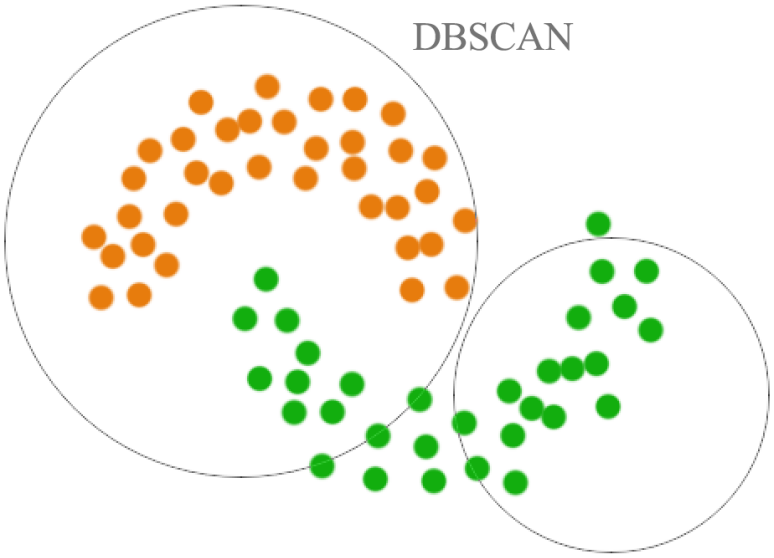
k번째 클러스터의 평균값과 obs간의 거리

클러스터의 평균값으로 중심점이 업데이트 되면서

그 때의 군집 내 분산이 작아지는지 확인함으로써 군집 형성

DBSCAN

Non-Hierarchical

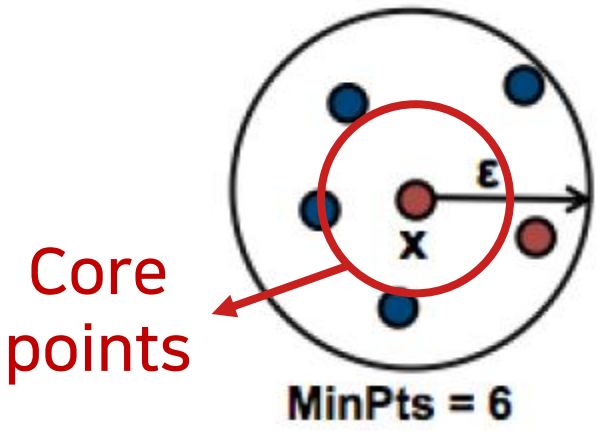


밀도 기반 알고리즘

같은 클러스터 데이터는 **서로 근접하게 분포**

DBSCAN

용어



MinPts

클러스터 구성 시 필요한 최소 데이터 수

엡실론 (ε)

데이터로부터의 반경

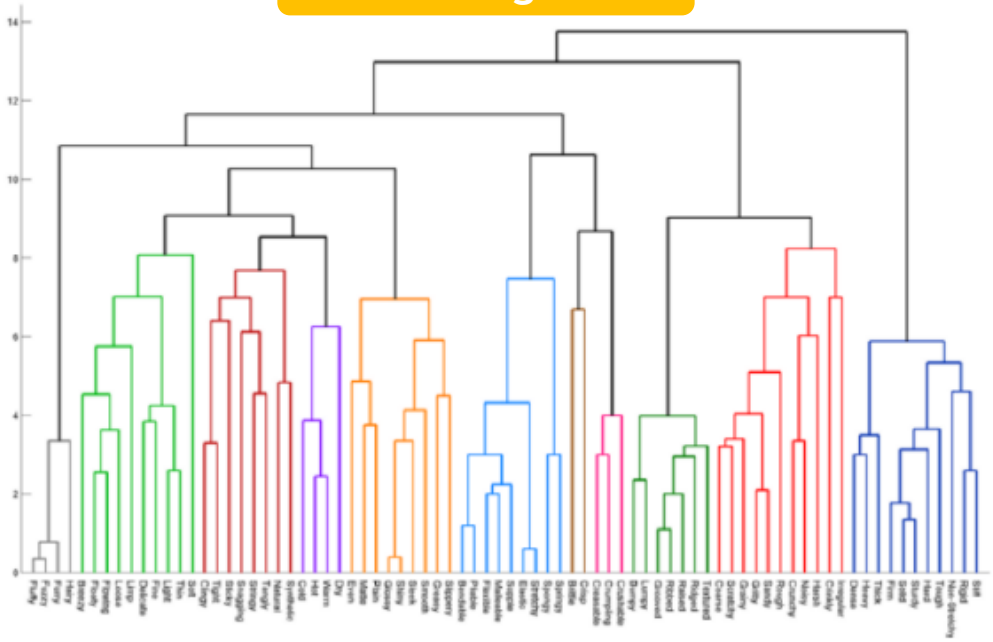
Core point (x)

엡실론 반경 내 MinPts 이상의 개체가 포함된 점

Hierarchical Clustering

계층적으로 유사한 개체와 클러스터링

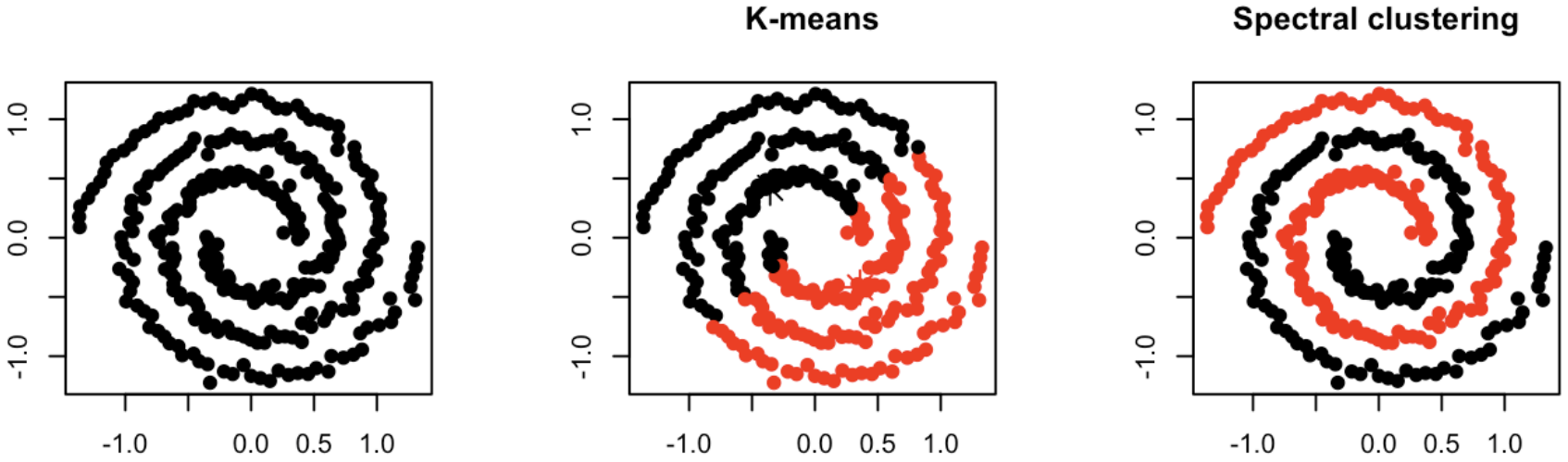
Dendrogram



클러스터링 개수 설정 없이 학습 가능

스펙트럴 클러스터링

그래프 기반의 계층적 클러스터링



더 낮은 차원의 데이터로 사영한 후에 클러스터링 진행

고유벡터 추출 후

양수 = 군집 A에 포함 vs 음수 = 포함 x

What is Recommendation System?

정보 필터링 기술의 일종

특정 고객이 관심을 가질만한 정보를 추천하는 시스템

콘텐츠 기반 추천

TF-IDF

$$TF - IDF = TF * log(\frac{n_D}{1 + n_t})$$

n_D : 전체 문서 수

n_t : 단어 t가 나온 문서 수

TF-IDF → 자연어 데이터로부터 feature를 뽑아내는 방법

(Term Frequency - Inverse Doc Frequency)

협업 필터링

사용자 기반 협업 필터링

$$Similarity(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \bar{r_a})(r_{b,p} - \bar{r_b})}{\sqrt{\sum_{p \in P}(r_{a,p} - \bar{r_a})^2}\sqrt{\sum_{p \in P}(r_{b,p} - \bar{r_b})^2}}$$

	주은	은서	이준	지민
진모	0.717	-0.271	0.427	0.561

피어슨 상관계수를 이용해서 진모와 다른 유저간의 유사도 계산

(주은이랑 잘 맞고 은서랑은 상극)

협업 필터링

잠재 요소 협업 필터링

	F1	F2	F3
주은	0.96	0.47	-0.76
은서	-0.03	0.84	-2.47
이준	2.38	0.11	-1.20
지민	0.59	1.10	-1.06

	어벤져스	포레스트 검프	매트릭스	엑시트	분노의 질주
F1	1.62	-0.79	1.04	1.07	1.43
F2	1.51	0.45	-0.06	0.12	-0.21
F3	-2.22	-1.85	0.43	1.18	-0.50

SVD와 SGD를 이용해서 sparse matrix를 효율적으로 대체해 행렬 분해한 예시