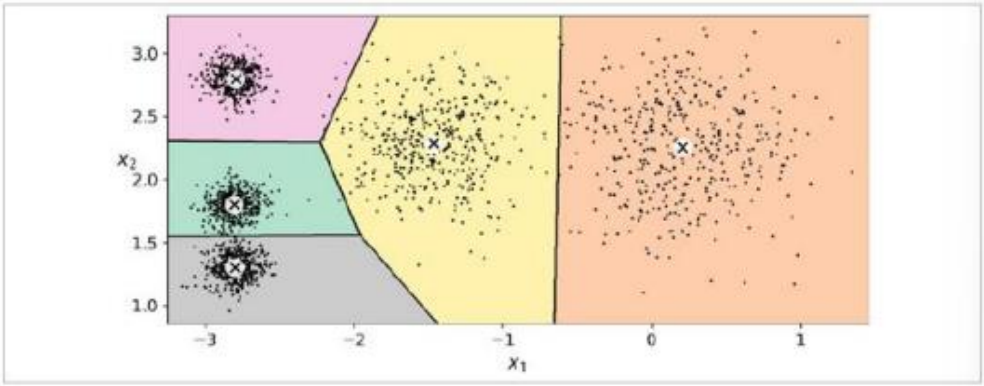
11주차 정리노트

김성욱, 서동현

<비지도 학습>

1. 군집

- 각 샘플은 하나의 그룹에 할당
- 비슷한 샘플을 구별해 하나의 클러스터(cluster) 또는 비슷한 샘플의 그룹으로 할당하는 작업
  - K-평균(로이드-포지 알고리즘)
    - 반복 몇 번으로 레이블이 없는 데이터셋을 빠르고 효율적으로 클러스터로 묶는 간단한 알고리즘
    - 결정경계



(보로노이 다이어그램)

    - 하드 군집, 소프트 군집
      - 하드 군집: 각 샘플에 대해 가장 가까운 클러스터를 선택.
      - 소프트 군집: 클러스터마다 샘플에 점수를 부여함. 샘플별로 각 군집 센트로이드와의 거리를 측정
    - 알고리즘
      - 처음에는 센트로이드를 랜덤하게 선정
        - 센트로이드는 각각 군집에서 중앙에 있는 값

- 수렴할 때까지 다음 과정 반복
  - 각 샘플을 가장 가까운 센트로이드에 할당
  - 군집별로 샘플의 평균을 계산하여 새로운 센트로이드 지정

◆ 센트로이드 초기화 방법

- 관성
  - K-mean 모델 평가 방법
  - 정의: 샘플과 가장 가까운 센트로이드와의 거리의 제곱의 합
  - 각 군집이 센트로이드에 얼마나 가까이 모여 있는가를 측정
- 좋은 모델 선택법
  - n번 학습 후 가장 낮은 관성을 갖는 모델 선택.
- K-평균++
  - 센트로이드를 무작위로 초기화하는 대신 특정 확률분포를 이용하여 선택

◆ elkan 알고리즘(속도 개선)

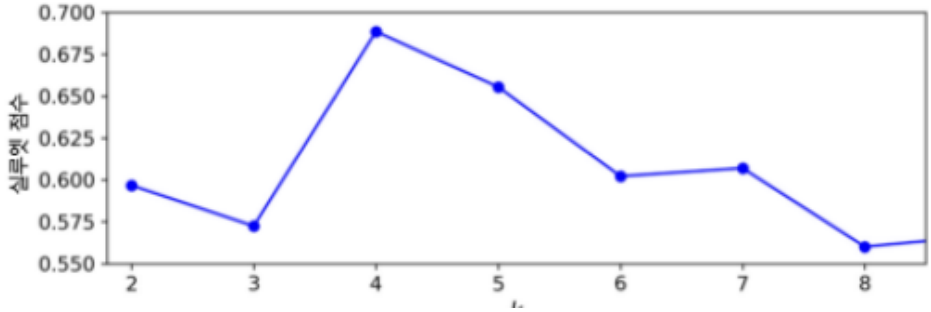
- 불필요한 거리 계산을 많이 피함으로 학습 속도 향상됨

◆ 미니배치 K-평균

- 전체 데이터셋 대신 각 반복마다 미니배치를 사용해 센트로이드를 조금씩 이동함

◆ 최적의 클러스트 개수 찾기

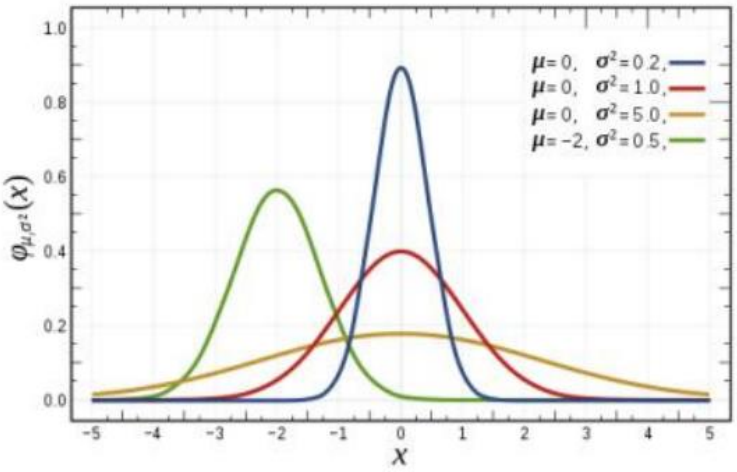
- 실루엣 점수와 클래스터 개수
  - 실루엣 점수: 모든 샘플에 대한 실루엣 계수와 평균
  - 실루엣 계수
    - • +1에 가까운 값: 자신의 클러스터 안에 포함되고, 다른 클러스터와는 멀리 떨어짐
    - 0에 가까운 값: 클러스터 경계에 위치
    - -1에 가까운 값: 샘플이 잘못된 클러스터에 할당됨



- DBSCAN
  - 밀집된 연속적 지역을 클러스터로 정의
  - 두개의 하이퍼파라미터 사용(eps, min_samples)
  - 핵심샘플과 군집
  - 이상치

가우시안 혼합

- 샘플이 파라미터가 알려지지 않은 여러 개의 혼합된 가우시안 분포에서 생성되었다고 가정하는 확률 모델
- 가우시안 분포 = 정규분포



- 클러스터
  - 하나의 가우시안 분포에서 생성된 모든 샘플들의 그룹
  - 일반적으로 타원형 모양
- 클러스터 개수 선택하기

- 이론적 정보 기준을 최소화하는 모델 선택 가능
  - 이론적 정보 기준
    - BIC
    - AIC

4-1. 비지도 학습의 활용 분야:

1. 데이터 분석
2. 고객 분류
3. 추천 시스템
4. 이미지 분할
5. 준지도 학습
6. 이상치 탐지
7. 차원 축소

5. 주의 사항:

1. 초기화에 민감한 K-평균
2. 밀도에 민감한 DBSCAN
3. 초기 클러스터링에 민감한 가우시안 혼합