6장 비지도학습

감사의 글

자료를 공개한 저자 오렐리앙 제롱과 강의자료를 지원한 한빛아카데미에게 진심어린 감사를 전합니다.

- 레이블이 없는 데이터 학습
    - 예제: 사진에 포함된 사람들 분류하기
- 용도
    - 군집화(clustering)
    - 이상치 탐지
    - 밀도 추정

- 군집화(clustering): 비슷한 샘플끼리 군집 형성하기
    - 데이터 분석
    - 고객분류
    - 추천 시스템
    - 검색 엔진
    - 이미지 분할
    - 준지도 학습
    - 차원 축소

- 이상치 탐지: 정상테이터 학습 후 이상치 탐지.
    - 제조라인에서 결함제품 탐지
    - 시계열데이터에서 새로운 트렌드 찾기

- 밀도 추정: 데이터셋 생성확률과정의 확률밀도함수 추정 가능
    - 이상치 분류: 밀도가 낮은 지역에 위치한 샘플
    - 데이터분석
    - 시각화

군집/군집화

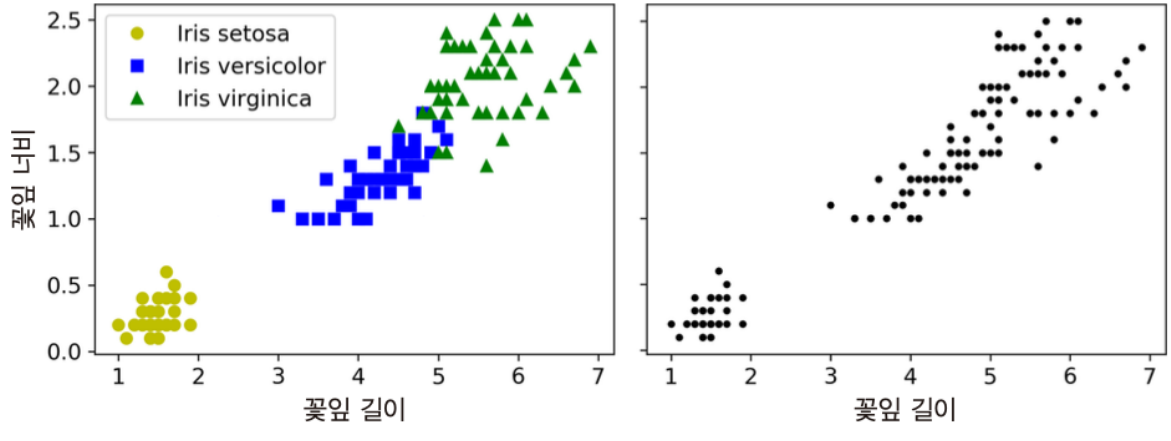
- 군집(클러스터, cluster): 유사한 샘플들의 모음(집합, 그룹)
- 군집화(clustering): 유사한 부류의 대상들로 이루어진 군집 만들기

분류 대 군집화

- 유사점: 각 샘플에 하나의 그룹 할당
- 차이점: 군집화는 군집이 미리 레이블(타깃)로 지정되지 않고 예측기 스스로 적절한 군집을 찾아내야 함.

예제

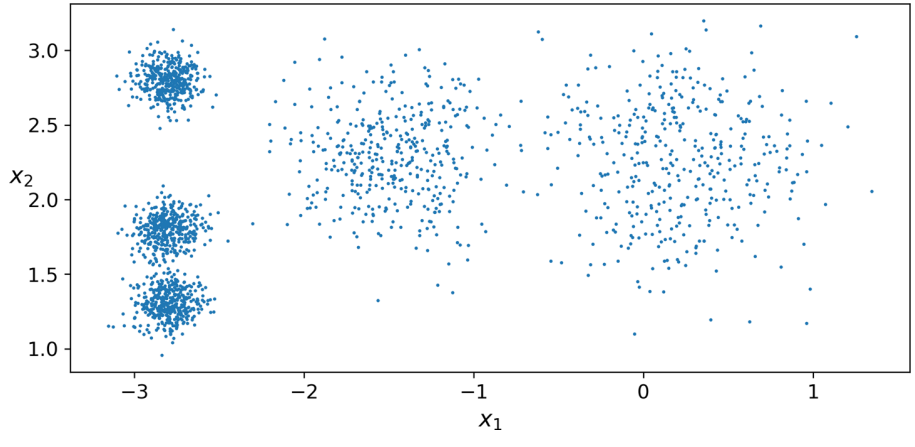
- 왼편: 분류
- 오른편: 군집화



K-평균

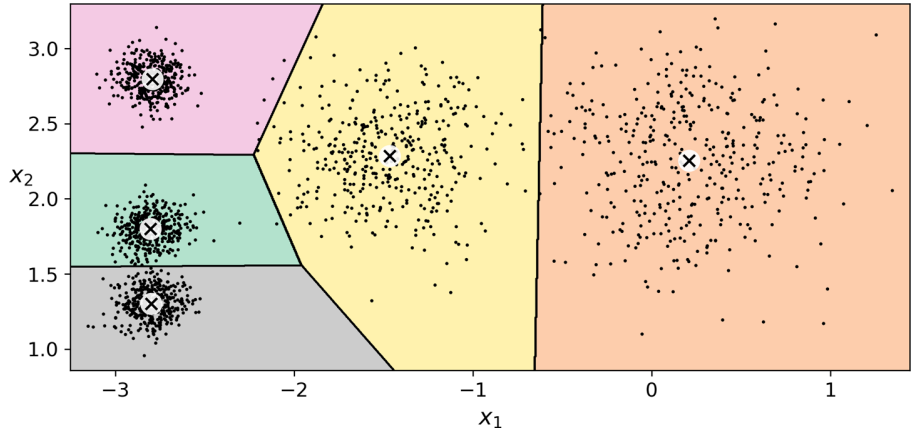
예제

- 샘플 덩어리 다섯 개로 이루어진 데이터셋



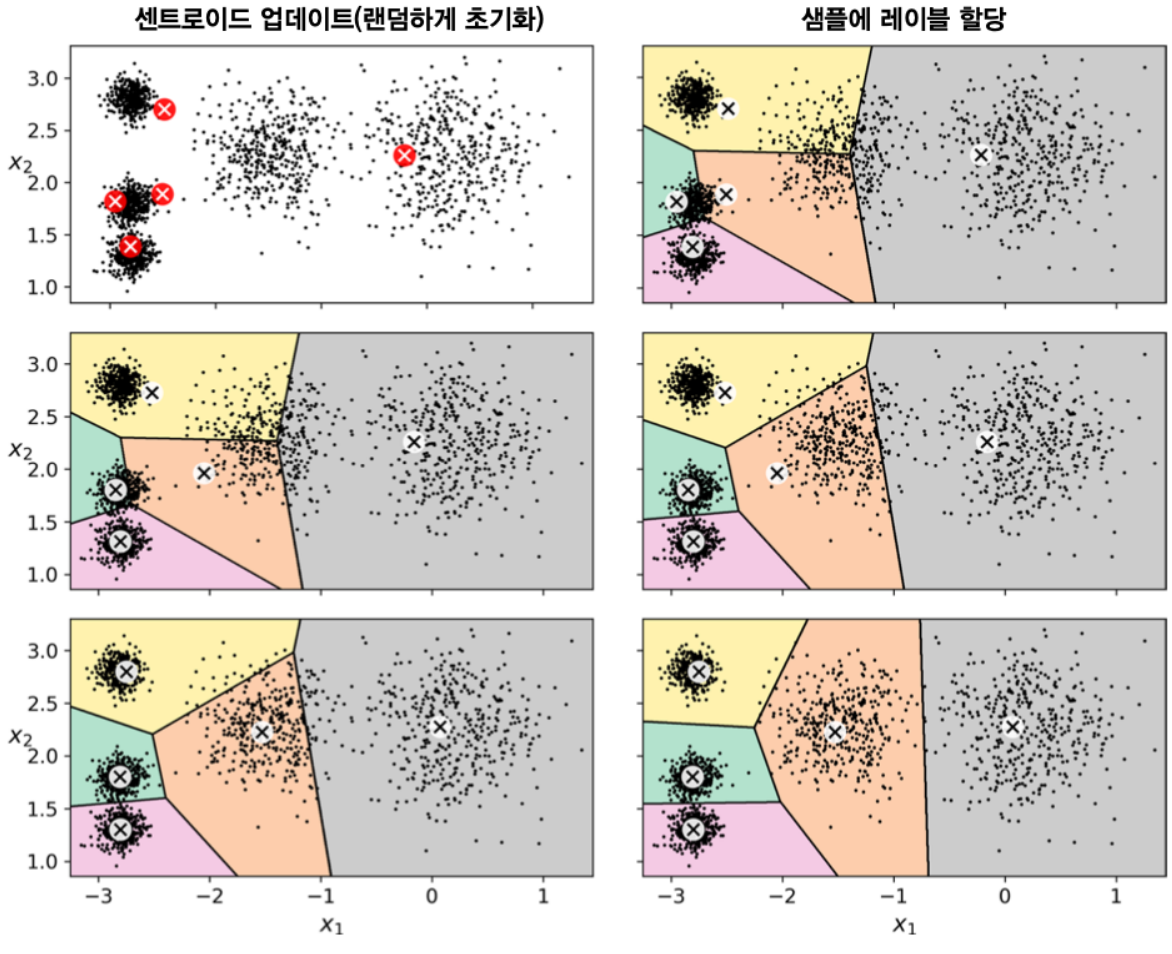
결정 경계

- 결과: 보로노이 다이어그램
    - 평면을 특정 점까지의 거리가 가장 가까운 점의 집합으로 분할한 그림
- 경계 부분의 일부 샘플을 제외하고 기본적으로 군집이 잘 구성됨.



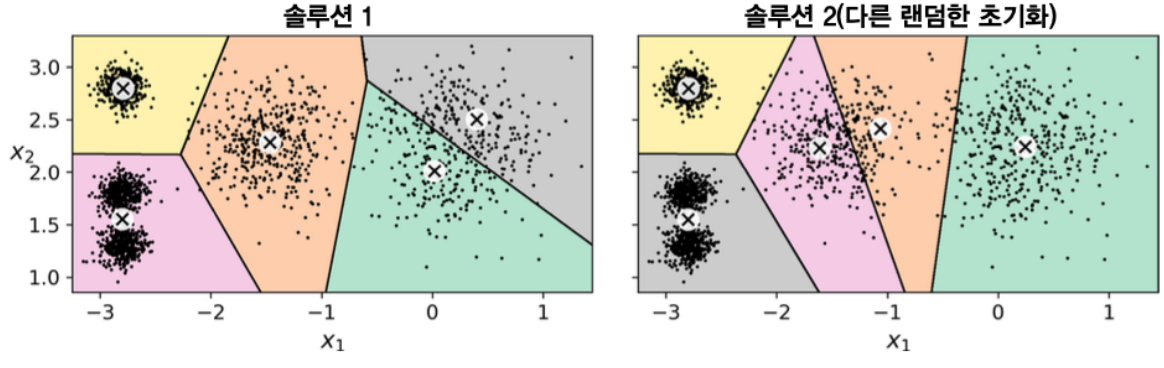
K-평균 알고리즘

- 먼저 k 개의 센트로이드 랜덤 선택
- 수렴할 때까지 다음 과정 반복
    - 각 샘플을 가장 가까운 센트로이드에 할당
    - 군집별로 샘플의 평균을 계산하여 새로운 센트로이드 지정


센트로이드 업데이트(랜덤하게 초기화)
샘플에 레이블 할당
x_2
x_1
3.0
2.5
2.0
1.5
1.0
−3
−2
−1
0
1

K-평균 알고리즘의 단점

- 초기 센트로로이드에 따라 매우 다른 군집화 발생 가능

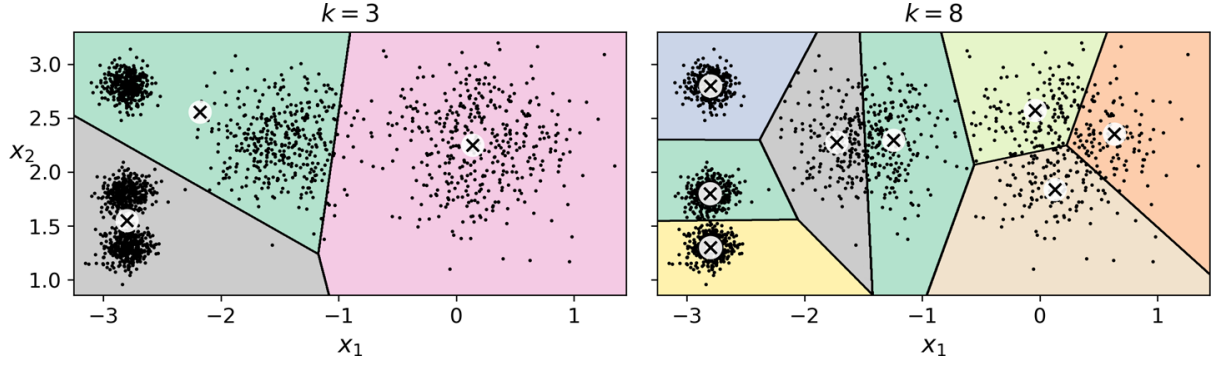


관성(inertia, 이너셔)

- k-mean 모델 평가 방법
- 정의: 샘플과 가장 가까운 센트로이드와의 거리의 제곱의 합
- 각 군집이 센트로이드에 얼마나 가까이 모여있는가를 측정
- `score()` 메서드가 측정. (음수 기준)

최적의 군집수 찾기

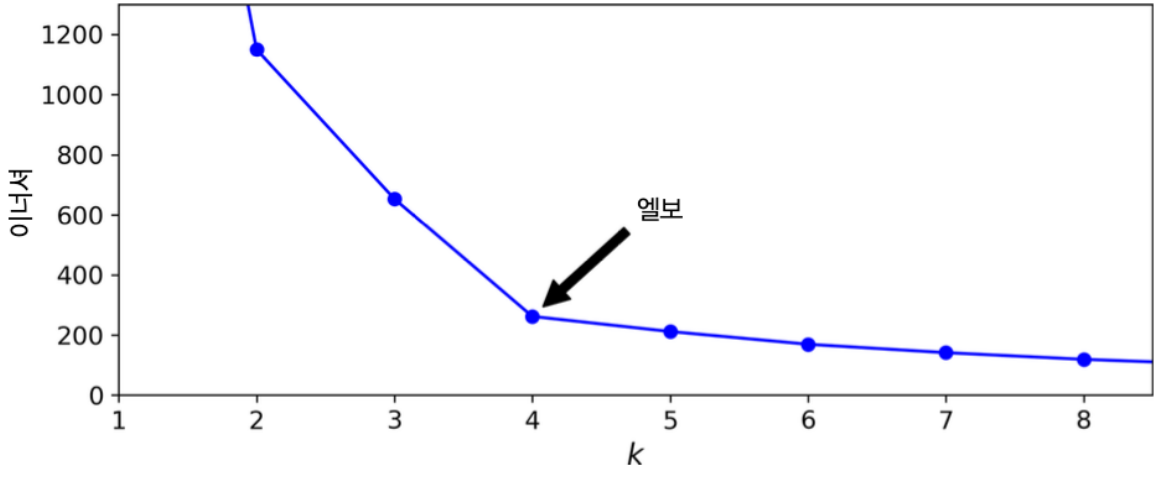
- 최적의 군집수를 사용하지 않으면 적절하지 못한 모델을 학습할 수 있음.



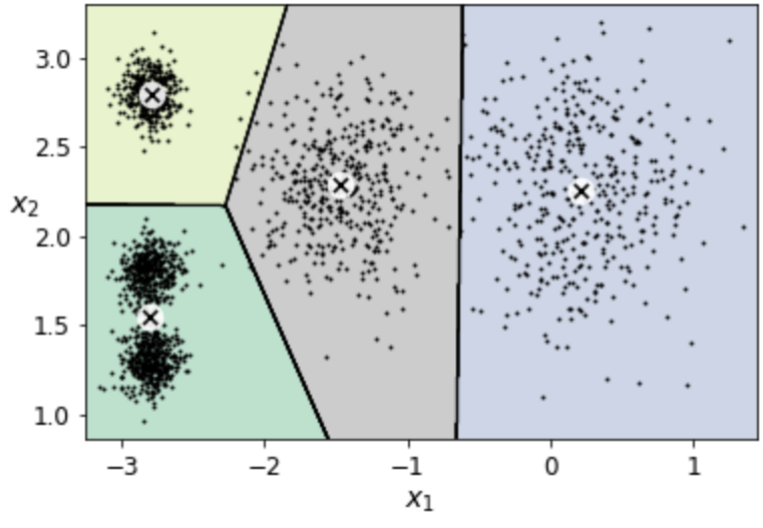
관성과 군집수

- 군집수 k가 증가할 수록 관성(inertia) 줄어듬.
- 따라서 관성만으로 모델을 평가할 없음.

- 관성이 더 이상 획기적으로 줄어들지 않는 지점의 군집수 선택 가능



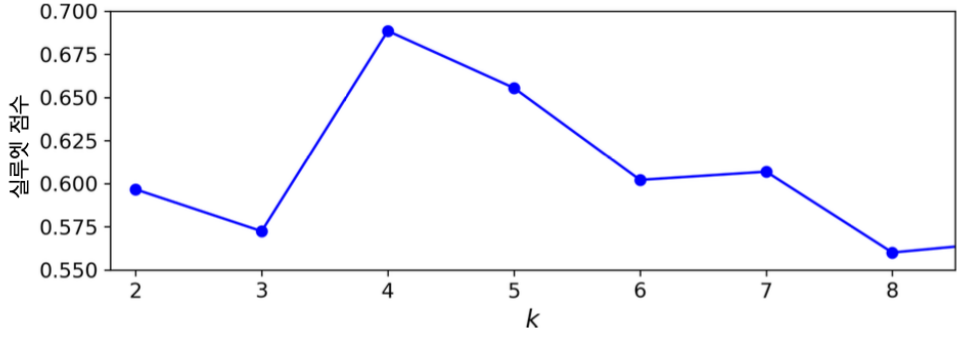
- 위 그래프에 의해 `k=4` 선택 가능.
- 하지만 아래 그림에서 보듯이 좋은 성능이라 말하기 어려움.



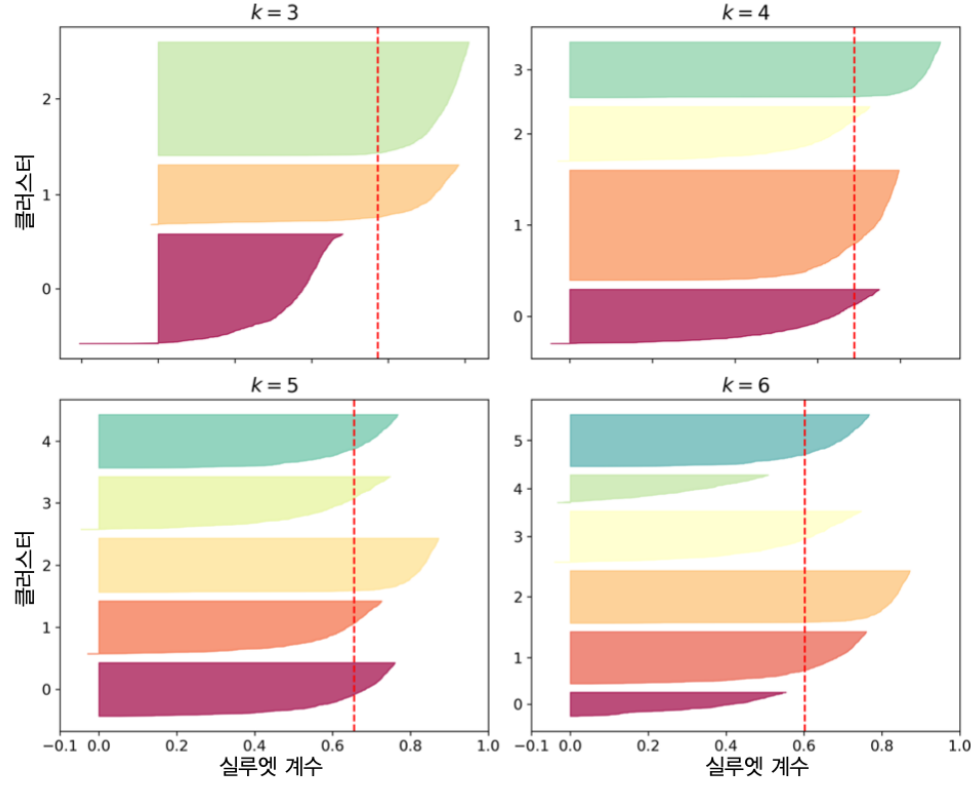
실루엣 점수와 군집수

- 샘플별 실루엣 계수의 평균값
- 실루엣 계수: -1과 1사이의 값
    - 1에 가까운 값: 적절한 군집에 포함됨.
    - 0에 가까운 값: 군집 경계에 위치
    - -1에 가까운 값: 잘못된 군집에 포함됨

- 아래 그림에 의하면 `k=5` 도 좋은 선택이 될 수 있음.



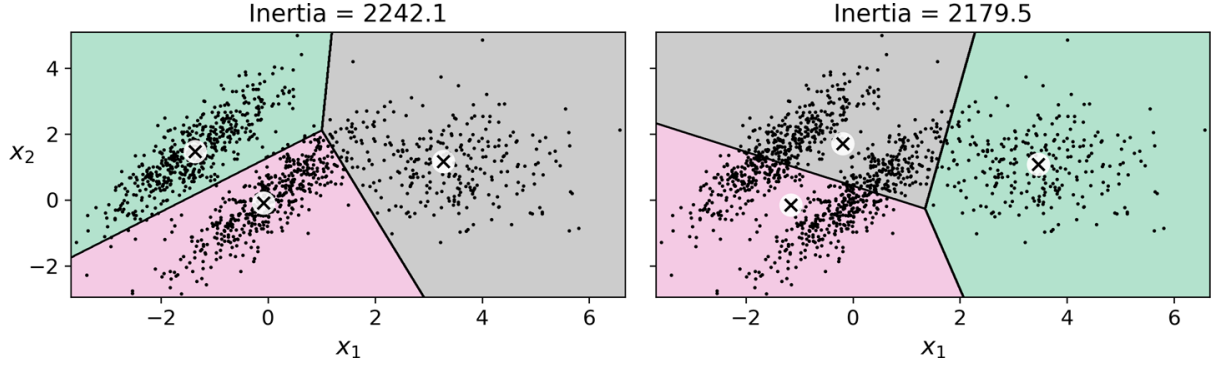
실루엣 다이어그램과 군집수



- 실루엣 다이어그램
    - 군집별 실루엣 계수 모음. 칼 모양.
    - 칼 두께: 군집에 포함된 샘플 수
    - 칼 길이: 군집에 포함된 각 샘플의 실루엣 계수
- 빨간 파선: 군집별 실루엣 점수
    - 대부분의 칼이 빨간 파선보다 길어야 함.
    - 칼의 두께가 서로 비슷해야, 즉, 군집별 크기가 비슷해야 좋은 모델임.
- 따라서 `k=5` 가 보다 좋은 모델임.

K-평균의 한계

- 최적의 모델을 구하기 위해 여러 번 학습해야 함.
- 군집수를 미리 지정해야 함.
- 군집의 크기나 밀집도가 다르거나, 원형이 아닐 경우 잘 작동하지 않음.



군집화 활용: 이미지 분할

이미지 분할

- 이미지를 여러 영역(segment)으로 분할하기
- 동일한 종류의 물체는 동일한 영역에 할당됨.
    - 자율주행: 보행자들을 모두 하나의 영역, 또는 각각의 영역으로 할당 가능
- 합성곱 신경망이 가장 좋은 성능 발휘

- 여기서는 K-평균을 이용하여 색상분할 실행
    - 인공위성 사진 분석: 전체 산림 면적 측정
    - 군집수가 중요함.

군집 수에 따른 변화