Chatpor9. 비지도 학습

(9.1 군집)

22.06.15

화학안전연구센터 최지원

9. 비지도 학습

- 비지도학습(unsupervised learning): **정답이 없는 데이터(라벨이 없음)**를 비슷한 특징끼리 군집화 하여 새로운 데이터에 대한 결과를 예측하는 방법

군집화(clustering)

- 유사성에 따라 데이터를 분할하는 것(e.g., 고객분류, 데이터 분석)

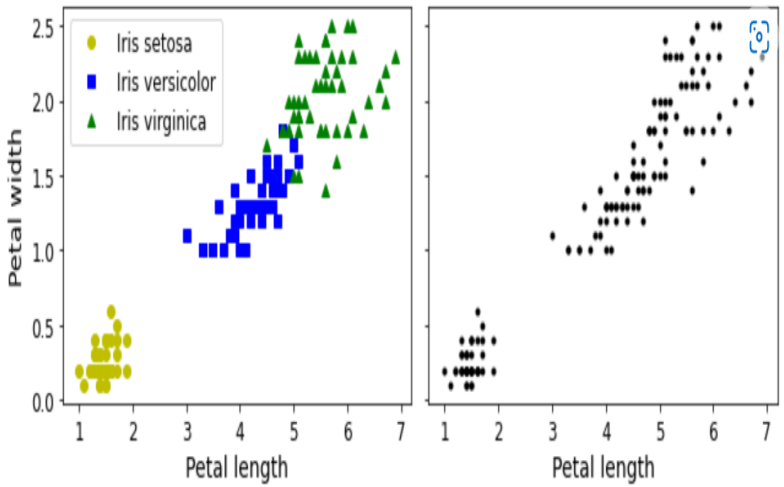
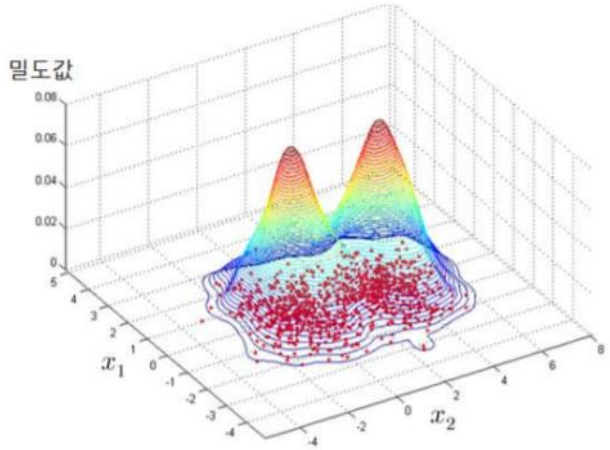


그림9-1. 분류(왼쪽), 군집(오른쪽)

밀도추정 (density estimation)

- 부류(class)별 데이터를 만들어 냈을 것으로 추정되는 확률분포를 찾는 것



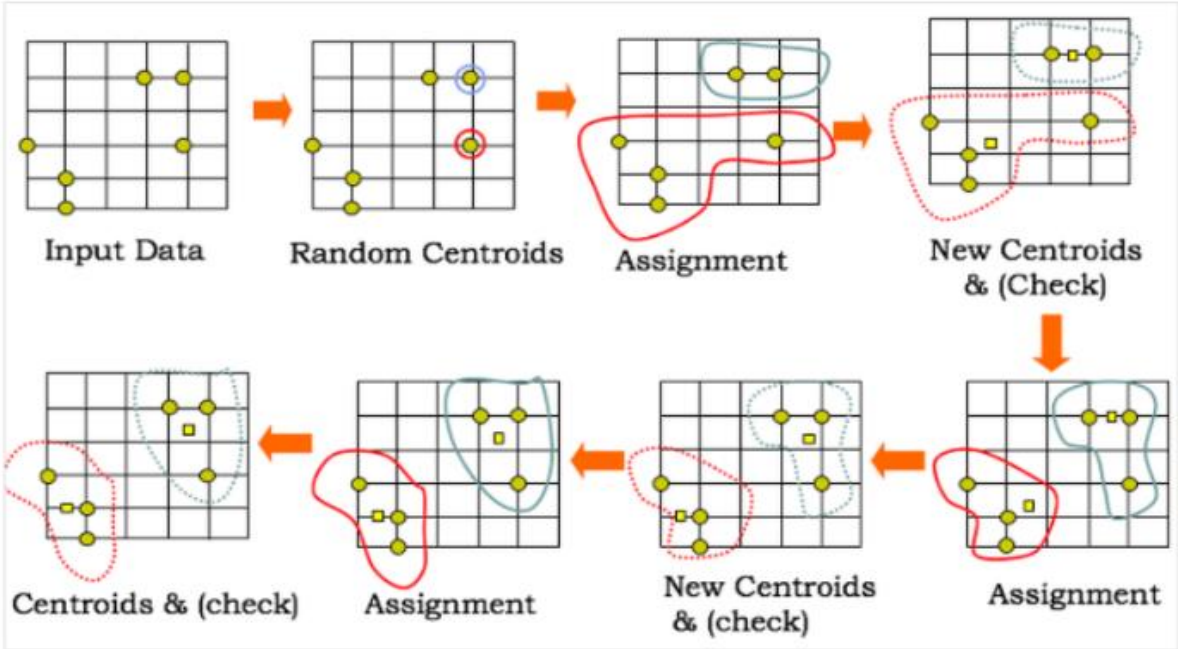
이상치(outlier) 탐지

- 다른 데이터와 크게 달라서 다른 메커니즘에 의해 생성된 것이 아닌지 의심스러운 데이터를 찾는 것

Ref.
https://gaebom.tistory.com/6
https://velog.io/@euisuk-chung

9.1.1 k-평균(k-means clustering)

- 각 클러스터의 중심(centroid, 센트로이드)을 찾고 가장 가까운 클러스터에 샘플을 할당하는 작업
  1) 최초에 임의의 점 k 개를 중심으로 지정
  2) 각 데이터를 k개의 점과 비교하여 가장 가까운 점이 있는 쪽으로 분류
  3) 모든 데이터를 k개 그룹으로 분류하고 나면, 각 그룹의 중심점을 계산
  4) 이전 단계에서 계산한 k개의 중심점을 이용해서 2), 3) 단계를 반복하며 3) 단계에서 갱신한 k개의 중심점이 이전 과정에서 사용한 중심들과 차이가 없거나 미리 정한 수준 이하로만 변하면 종료

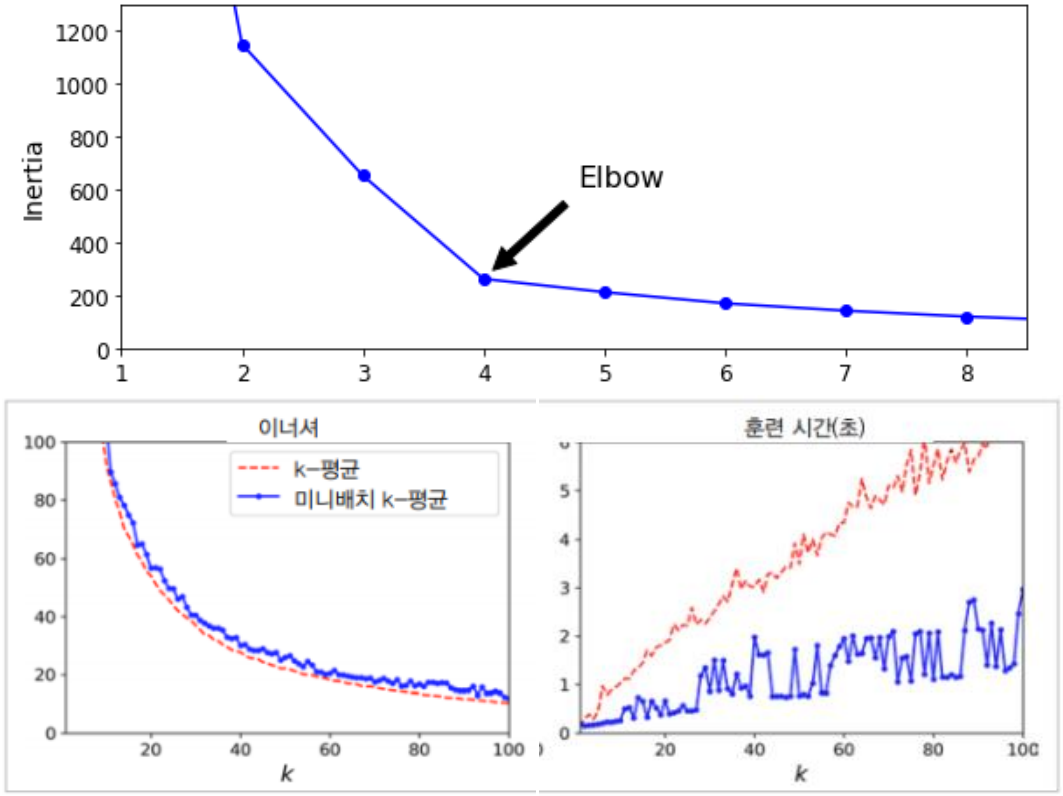


Ref. naver blog(Seungwoo's Daily)

9.1.1 k-평균(k-means clustering)

- 최적의 클러스터 개수는 k-평균 성능을 결정짓는 매우 중요한 요소임
  - 성능지표: 이너셔, 실루엣 점수

이너셔(inertia)



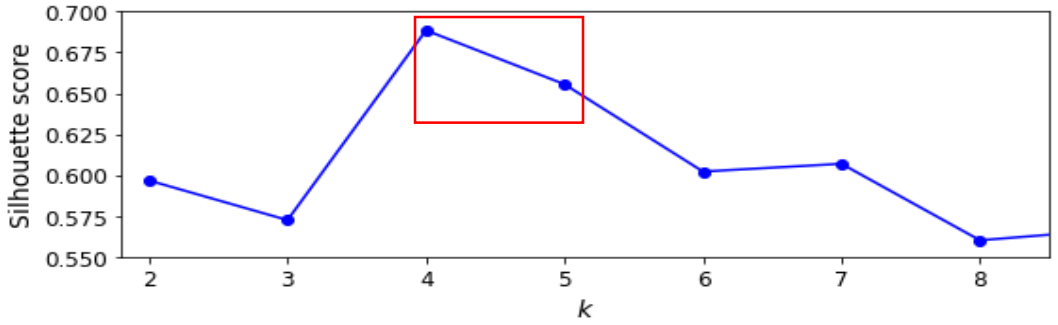
*미니배치 k-평균: 전체 데이터셋을 사용해 반복하지 않고 미니배치를 사용해 센트로이드를 조금씩 이동하여 클러스터 진행

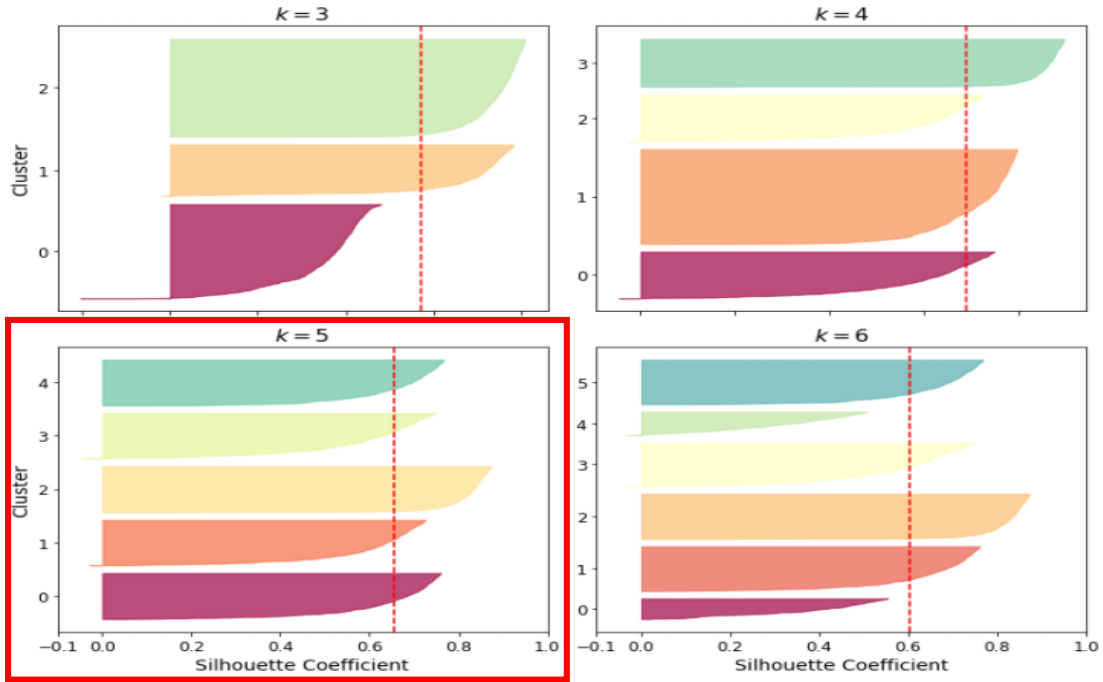
- 클러스터 중심과 클러스터에 속한 샘플 사이의 거리의 제곱의 합
  - 클러스터에 속한 샘플이 얼마나 가깝게 모여있는지 나타내는 값 (낮을수록 좋음)
- K=4(Elbow)를 최적의 클러스터로 선정
  - k<4: 가파른 이너셔 감소
  - k>4: 완만한 이너셔 감소
- 최적의 클러스터를 선정하기 위한 정보 제한적

9.1.1 k-평균(k-means clustering)

- 최적의 클러스터 개수는 k-평균 성능을 결정짓는 매우 중요한 요소임
  - 성능지표: 이너셔, 실루엣 점수

실루엣 점수(silghouette score)





- 모든 데이터에 대한 실루엣 계수의 평균

$$\text{실루엣계수} = (\boldsymbol{b} - \boldsymbol{a})/\max(\boldsymbol{a}, \boldsymbol{b})$$

  - a: 클러스터 내부의 평균거리; b: 가장 가까운 클러스터의 샘플까지 평균거리
  - 실루엣계수는 -1에서 +1까지 바뀔 수 있음
    1) +1에 가까우면 자신의 클러스터 안에 잘 속해 있고 다른 클러스터와 멀리 떨어져 있음
    2) 0에 가까우면 클러스터 경계에 위치
    3) -1에 가까우면 잘못된 클러스터에 할당

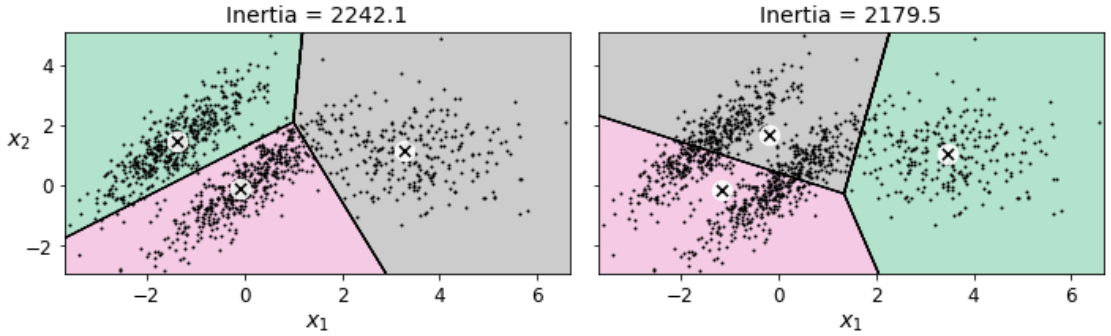
- 실루엣 점수만 보면 k=4가 클러스터개수로 적합

- 세부적인 분석을 위한 다이어그램까지 고려 시 최종적으로 k=5로 선택하는 것이 적합
  - 모든 클러스터 실루엣점수(빨간파선)보다 높음(k=4, k=5 만족)
  - 모든 클러스터 크기(그래프 높이) 비슷(k=5 만족)

9.1.2 k-평균의 한계

- 장점:
  1) 속도가 빠르고 확장성이 용이함
  2) 데이터에 대한 사전 정보가 필요하지 않음
  3) 사전에 특정 변수에 대한 역할 정의가 필요하지 않음
- 단점:
  1) 최적이 아닌 솔루션을 피하려면 알고리즘을 여러 번 실행해야 함
  2) 초기 클러스터링 수 결정하는데 어려움이 있음
  3) 클러스터 수가 적합하지 않으면 결과해석의 어려움
  4) 군집된 데이터 크기나 밀집도가 서로 다르거나 원형이 아닐 경우 잘 동작하지 않음

k-평균을 실행하기 전에 입력 특성의 스케일을 맞추는 것이 중요

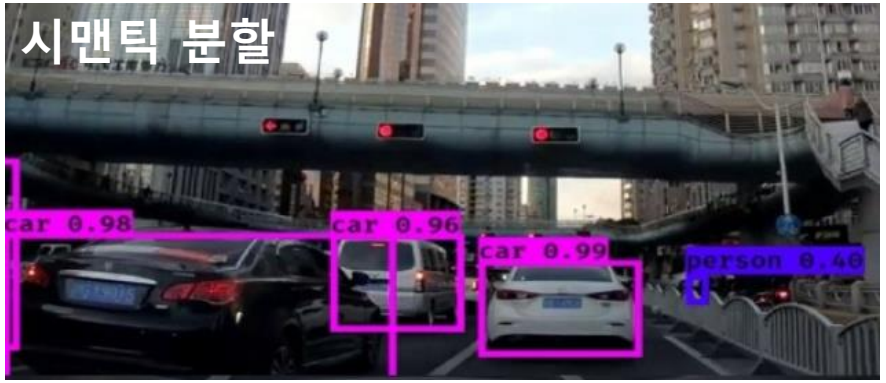


k-평균이 세 개의 타원형 클러스터를 적절히 구분하지 못함

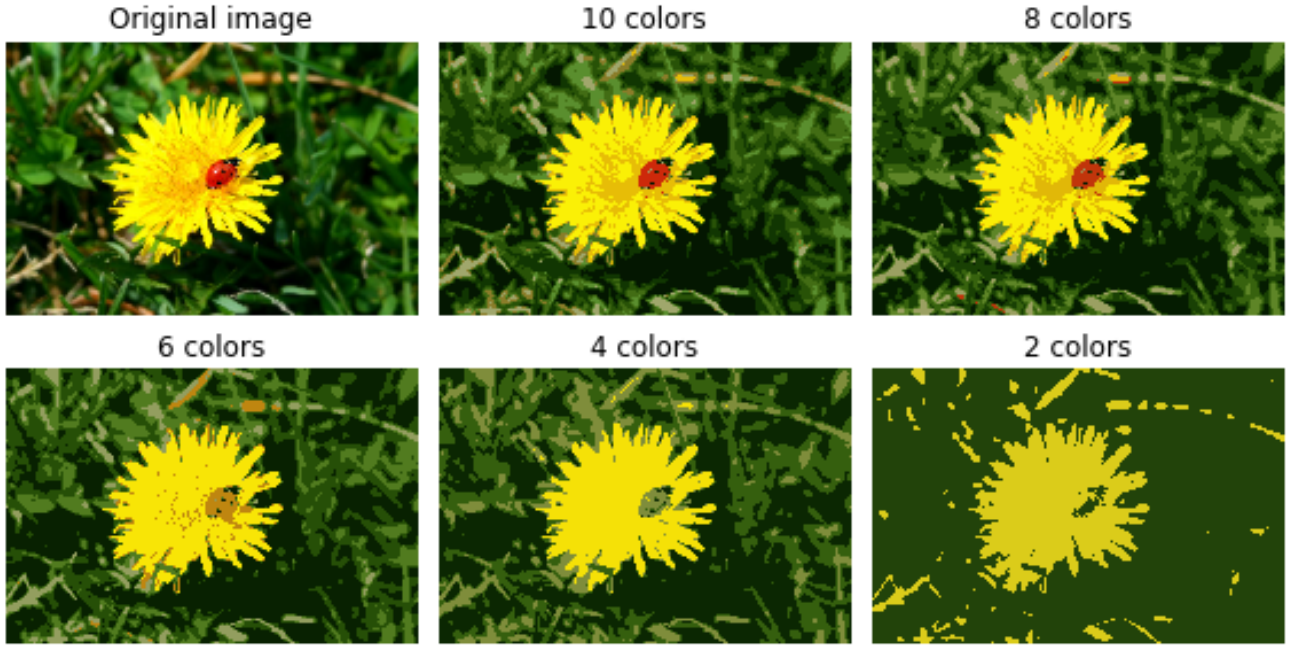
9.1.3 군집을 사용한 이미지 분할

- 이미지 분할: 이미지를 세그먼트 여러 개로 분할하는 작업
  1) 시맨틱 분할: 동일한 종류의 물체에 속한 모든 픽셀은 같은 세그먼트에 할당
  2) 인스턴스 분할: 이미지로부터 객체의 영역을 파악하는 세그먼트 할당
  3) 색상 분할: 동일한 색상을 가진 픽셀을 같은 세그먼트에 할당

     - 8개보다 클러스터 개수가 작으면 무당벌레(빨간색) 클러스터를 만들지 못함(∵k-평균이 비슷한 크기의 클러스터를 만드는 경향성)



Ref. naver blog(WeGo)



다양한 클러스터 개수로 k-평균을 사용해 만든 이미지 분할

9.1.4 군집을 사용한 전처리

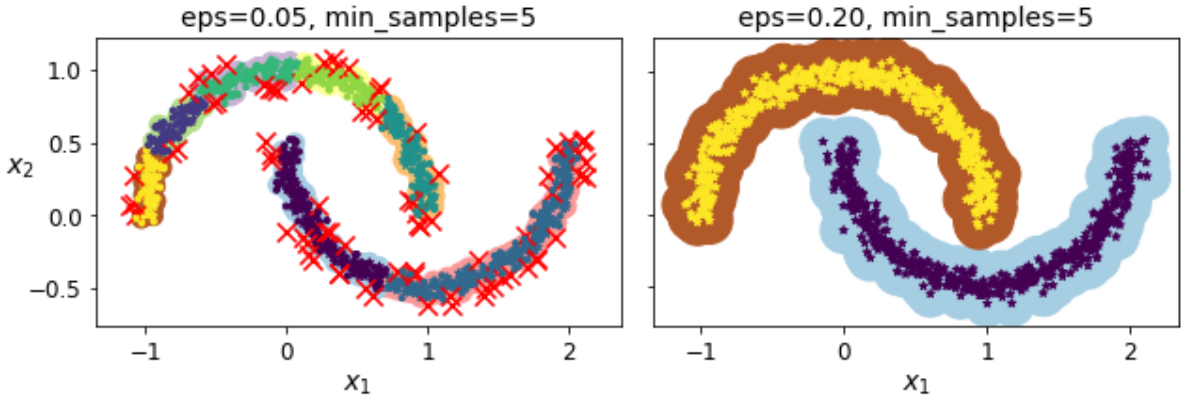
- 지도 학습 알고리즘(e.g., 로지스틱 회귀 모델)을 적용하기 전에 전처리 단계로 사용할 수 있음

9.1.5 군집을 사용한 준지도 학습

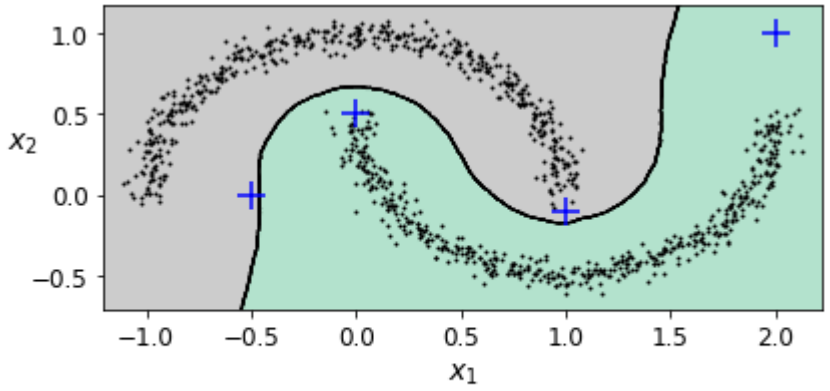
- 레이블이 없는 데이터가 많고 레이블이 있는 데이터가 적을 때 사용

9.1.6 DBSCAN

- 동일한 클래스에 속하는 데이터는 서로 밀접하게 분포
  1) 알고리즘이 각 샘플에서 작은 거리인 ε (입실론) 내에 샘플이 몇 개 놓여 있는지 센다(이 지역을 샘플의 ε-이웃 라고 부름)
  2) (자기 자신 포함) ε-이웃 내에 적어도 min_samples개 샘플이 있다면 이를 핵심 샘플로 간주 (i.e., 핵심 샘플은 밀집된 지역에 있는 샘플)
  3) 핵심 샘플의 이웃에 있는 모든 샘플은 동일한 클러스터에 속함. 이웃에는 다른 핵심 샘플이 포함될 수 있음. 따라서 핵심 샘플의 이웃의 이웃은 계속해서 하나의 클러스터 를 형성
  4) 핵심 샘플이 아니고 이웃도 아닌 샘플은 이상치로 판단



두 가지 다른 이웃 반경을 사용한 DBSCAN 군집

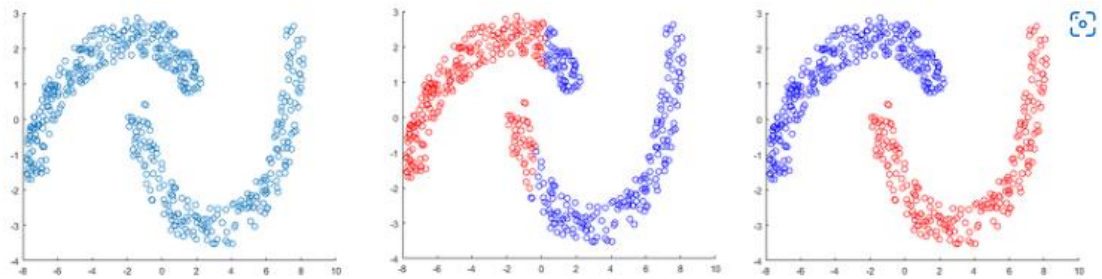


두 클러스터 사이의 결정 경계

Ref. https://untitledtblog.tistory.com/146.

9.1.6 DBSCAN

- 장점
  1) k-mean와 달리 클러스터 수를 정하지 않아도 됨
  2) 클러스터 밀도에 따라 클러스터를 서로 연결하기 때문에 기하학적 모양을 갖는 군집도 찾을 수 있음
  3) Noise point를 통하여 이상치(outlier) 검출이 가능
- 단점
  1) 해당 알고리즘은 새로운 샘플에 대한 클러스터를 예측할 수 없음
  2) 사용자가 필요한 예측기를 선택해야 함



[그림 1] Two moons 데이터셋에 대한 k-means clustering과 DBSCAN의 군집화 결과 (같은 색의 데이터는 같은 클래스에 군집되었다는 것을 의미함).

9.1.7 다른 군집 알고리즘

- 비지도학습(병합군집: 각 데이터 포인트를 하나의 클러스터로 지정하고 지정된 개수의 클러스터가 남을 때까지 가장 비슷한 두 클러스터를 합쳐 나가는 알고리즘
- 계층적 군집화(BIRCH): 비슷한 군집끼리 묶어 가면서 최종적으로는 하나의 케이스가 될 때까지 군집
- 평균-이동: 확률 밀도함수가 피크인 점을 군집의 중심으로 지속적으로 움직이면서 군집화 수행
- 유사도전파: 자신을 대표할 수 있는 비슷한 샘플에 투표하여 알고리즘이 수렴하면서 각 대표와 투표한 샘플이 클러스터를 형성
- 스펙트럼 군집: 샘플 사이의 유사도 행렬을 받아 차원을 축소하여 저차원 공간에서 또 다른 군집 알고리즘 사용