기계학습 비지도 학습 정리노트

4조- (이희구, 유제우)

※ 4조 의견

질문) 비지도 학습에는 어떤 것들이 있어?

대부분의 머신러닝 애플리케이션이 지도 학습 기반이지만, 사용할 수 있는 데이터는 대부분 레이블이 없다. 이럴 때 비지도 학습이 유용하다. 비지도 학습에는 군집, 이상치 탐지, 밀도 추정 등이 있다.

질문) 군집에 대해 설명해줘

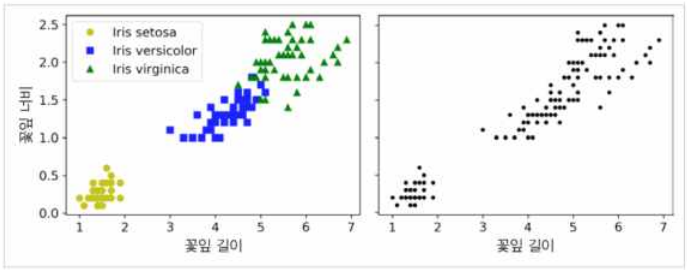


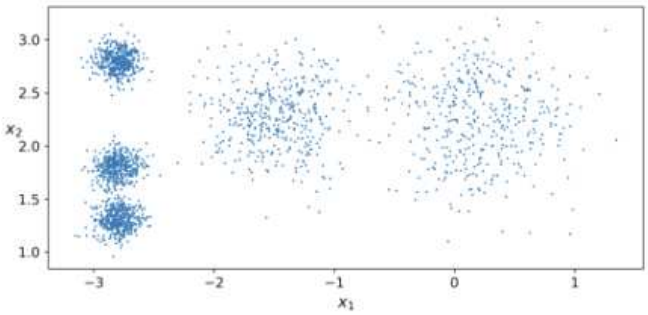
그림 9-1 분류(왼쪽) 대 군집(오른쪽)

각 샘플은 하나의 그룹에 할당되고, 비슷한 샘플을 구별해 하나의 클러스터 또는 비슷한 샘플의 그룹으로 할당하는 작업이다.

질문) 군집을 사용하는 다양한 어플리케이션에는 어떤 것들이 있을까?

고객 분류로는 추천시스템이 있고, 그 외에도 데이터 분석, 샘플의 친화성을 측정하는 차원 축소 기법, 제조분야에서 결함 감지를 하거나 부정 거래 감지에 활용하는 이상치 탐지, 준지도 학습, 검색 엔진, 이미지 분할 등이 있다.

질문) K-평균이 뭐야?



K-평균은 반복 몇 번으로 레이블이 없는 데이터셋을 빠르고 효율적으로 클러스터로 묶는 간단한 알고리즘이다. 예제로 샘플 덩어리 다섯 개로 이루어진 데이터 셋을 확인할 수 있다.

질문) 결정 경계란?

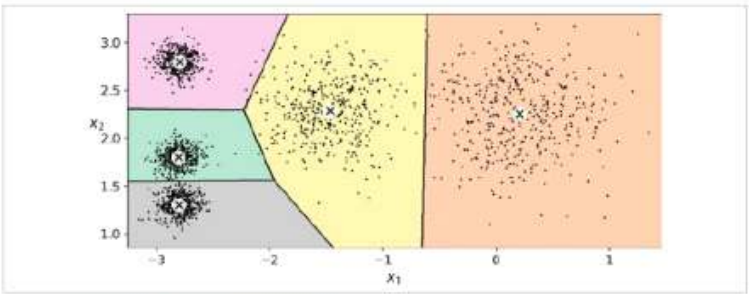


그림 9-3 k-평균의 결정 경계 (보로노이 다이어그램)

평면을 특정 점까지의 거리가 가장 가까운 점의 집합으로 분할한 그림이다.

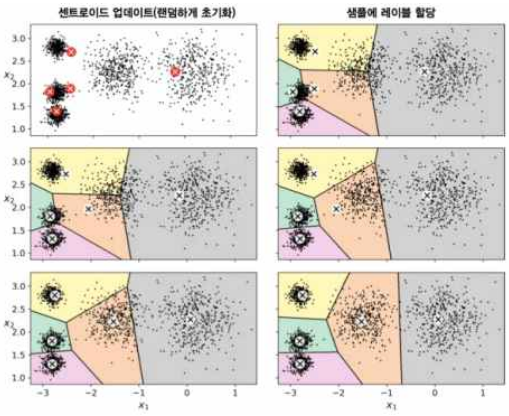
질문) K-평균 알고리즘의 단점은 뭐야?

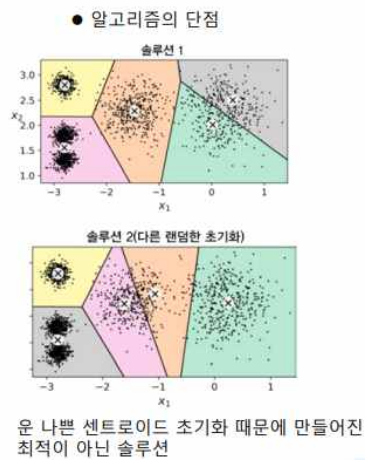
샘플과 센트로이드까지의 거리만 고려되기 때문에 군집의 크기가 서로 많이 다르면 잘 작동하지 않는다.

질문) 하드 군집과 소프트 군집에 대해 설명해줘

하드 군집은 각 샘플에 대해 가장 가까운 클러스터를 선택하고, 소프트 군집은 클러스터마다 샘플에 점수를 부여하여 샘플별로 각 군집 센트로이드와의 거리를 측정한다.

질문) K-평균 알고리즘 구동은 어떻게 돼?





처음에는 센트로이드를 랜덤하게 선정한다. 그 후 수렴할 때까지 다음 과정을 반복한다.
각 샘플을 가장 가까운 센트로이드에 할당하고, 군집별로 샘플의 평균을 계산하여 새로운 센트로이드를 지정한다.

질문) 관성에 대해 설명해줘

k-mean 모델 평가 방법이다. 정의로는 샘플과 가장 가까운 센트로이드와의 거리의 제곱의 합이다. 각 군집이 센트로이드에 얼마나 가까이 모여 있는가를 측정한다. score() 메서드가 측정을 한다.

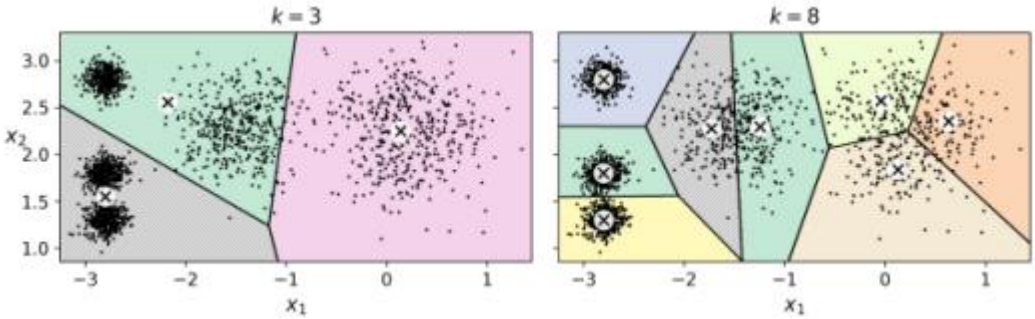
질문) 좋은 모델을 선택하려면 어떻게 해야 돼?

다양한 초기화 과정을 실험한 후에 가장 좋은 것을 선택한다. 여기서는 n_init = 10이 기본값으로 사용된다. 10번 학습 후 가장 낮은 관성을 갖는 모델을 선택한다.

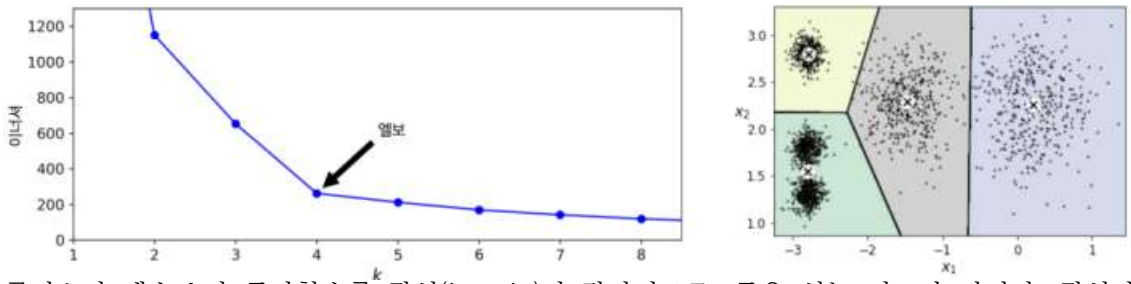
질문) K-평균++ 는 또 뭐야?

센트로이드를 무작위로 초기화하는 대신 특정 확률분포를 이용하여 선택한다. 센트로이드들 사이의 거리를 크게 할 가능성이 높아진다. 또한 KMeans 모델의 기본 값으로 사용된다.

질문) 최적의 클러스터 개수 설정 방법은?

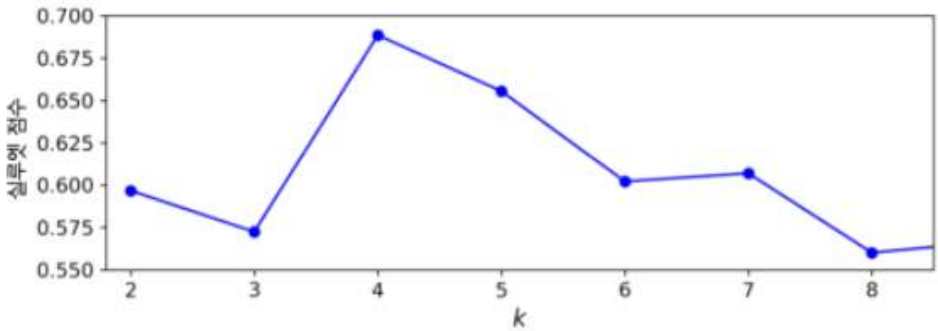


최적의 군집수를 사용하지 않으면 적절하지 못한 모델을 학습할 수 있다.



클러스터 개수 k가 증가할수록 관성(inertia)이 작아지므로, 좋은 성능 지표가 아니다. 관성만으로 모델을 평가할 수 없다. 관성이 더 이상 획기적으로 줄어들지 않는 지점의 클러스터 개수를 선택한다. (k=4 선택 가능)

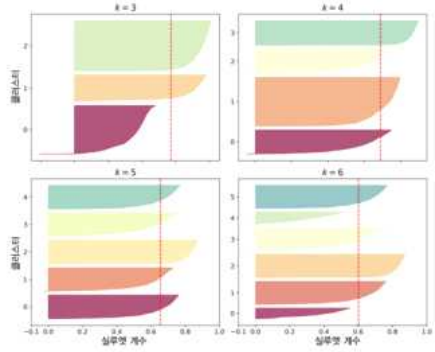
질문) 실루엣 점수와 클러스터 개수는?



[그림9-9] 실루엣 점수를 사용해 클러스터 개수 k를 선택하기

실루엣 점수는 모든 샘플에 대한 실루엣 계수의 평균이다. 실루엣 계수는 -1과 +1사이의 값이다. +1에 가까운 값은 자신의 클러스터 안에 포함되고, 다른 클러스터와는 멀리 떨어진다. 0에 가까운 값은 클러스터 경계에 위치하게 된다. -1에 가까운 값은 샘플이 잘못된 클러스터에 할당된다. [그림9-9] 에서 볼 때 k=4가 좋은 선택이지만, k=5도 좋은 선택이 될 수 있다.

질문) 실루엣 다이어그램과 클러스터 개수는?



실루엣 다이어그램은 클러스터별 실루엣 계수의 모음으로 칼 모양의 그래프이다. 칼 두께는 클러스터에 포함된 샘플의 개수를 나타내고, 칼 길이는 클러스터에 포함된 샘플의 실루엣 계수를 나타낸다. 빨간 파선은 클러스터 계수에 해당하는 실루엣 점수이다. 대부분의 칼이 빨간 파선보다 길어야 한다. 낮으면 다른 클러스터랑 너무 가깝기 때문이다. 칼의 두께가 서로 비슷해야, 즉, 클러스터별 크기가 비슷해야 좋은 모델이다. 실루엣 다이어그램 상에서 k=5가 보다 좋은 모델임을 알 수 있다.

질문) k-평균의 한계는 뭐야?

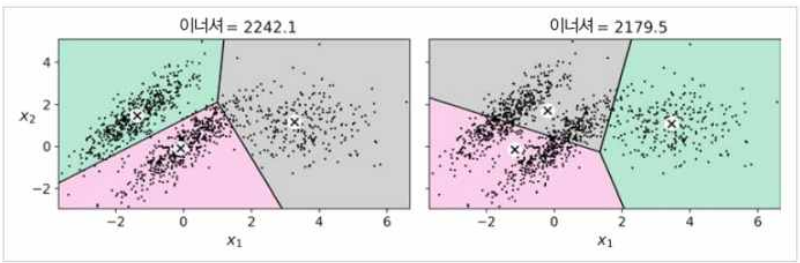


그림 9-11 k-평균이 세 개의 타원형 클러스터를 적절히 구분하지 못합니다.

k-평균은 속도가 빠르고 확장이 용이하다는 장점이 있지만, 완벽한 것은 아니다. K- 평균은 최적이 아닌 솔루션을 피하려면 알고리즘을 여러 번 실행해야 한다. 또한, 클러스터 개수를 미리 지정해야 하고 클러스터의 크기나 밀집도가 다르거나, 원형이 아닐 경우 잘 작동하지 않는다. 데이터에 따라서 잘 수행할 수 있는 클러스터 알고리즘이 다르다.

질문) 군집을 사용한 분할에는 어떤 것들이 있어?

이미지를 세그먼트 여러 개로 분할하는 작업인 이미지 분할, 동일한 종류의 물체에 속한 모든 픽셀은 같은 세그먼트에 할당하는 시맨틱 분할, k평균을 이용하여 분할하는 색상 분할 등이 있다.

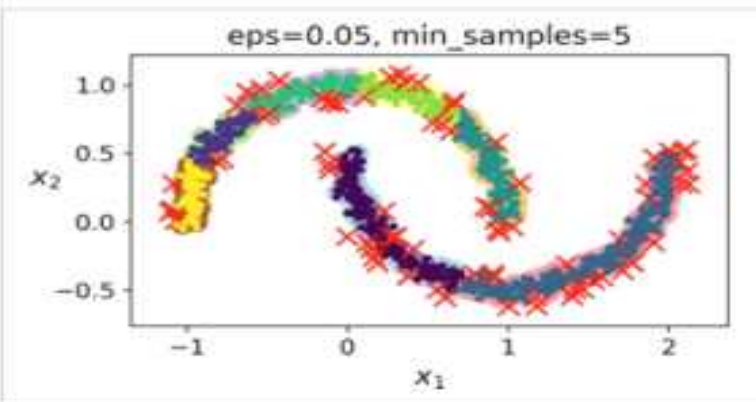
질문) 군집을 사용한 준지도 학습은 언제 사용해?

레이블이 없는 샘플이 많고 레이블이 있는 샘플이 적을 때 사용한다.

```
y_representative_digits = np.array([
    0, 1, 3, 2, 7, 6, 4, 6, 9, 5,
```

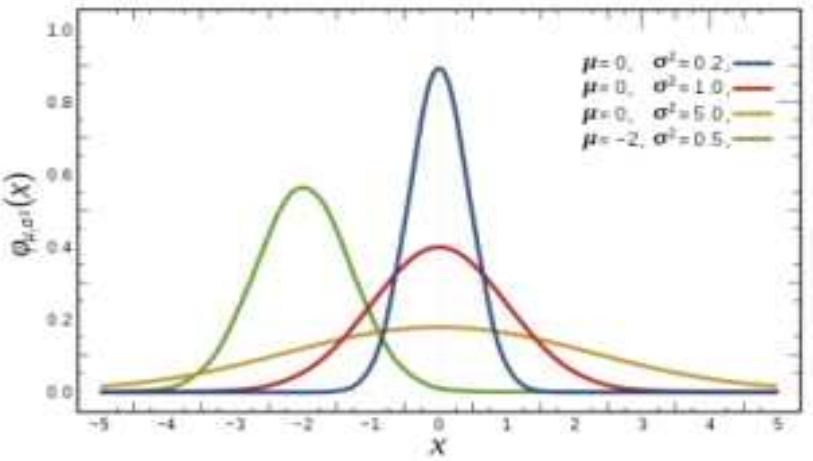
50개의 이미지를 보고 수동으로 레이블을 할당한다. 50개의 샘플을 레이블 할당하여 학습된 모델의 성능은 92.2% 정도가 나타났다.

질문) DBSCAN의 장점과 단점은 뭐야?



매우 간단하면서 매우 강력한 알고리즘이다. 군집의 모양과 개수에 상관이 없다. 이상치에 안정적이고 군집간의 밀집도가 크게 다르면 모든 군집 파악이 불가능하다.

질문) 가우시안 혼합 모델이 뭐야?



샘플이 파라미터가 알려지지 않은 여러 개의 혼합된 가우시안 분포에서 생성되었다고 가정하는 확률 모델이다. 가우시안 분포는 정규분포이다. 종 모양의 확률밀도함수를 갖는 확률분포이다.

질문) 가우시안 혼합에서 GMM 활용은 어떻게 하는거야?

```
from sklearn.mixture import GaussianMixture

gm = GaussianMixture(n_components=3, n_init=10, random_state=42)
gm.fit(X)
```

GaussianMixture 모델을 적용하고, n_components로 군집 수를 지정하고, n_init로 모델 학습 반복 횟수를 지정한다. 파라미터(평균값, 공분산 등)를 무작위로 추정한 후 수렴할 때까지 학습시킨다.

- EM 알고리즘이 추정한 파라미터를 확인

```
gm.weights_
```

```
array([0.39025715, 0.40007391, 0.20966893])
```

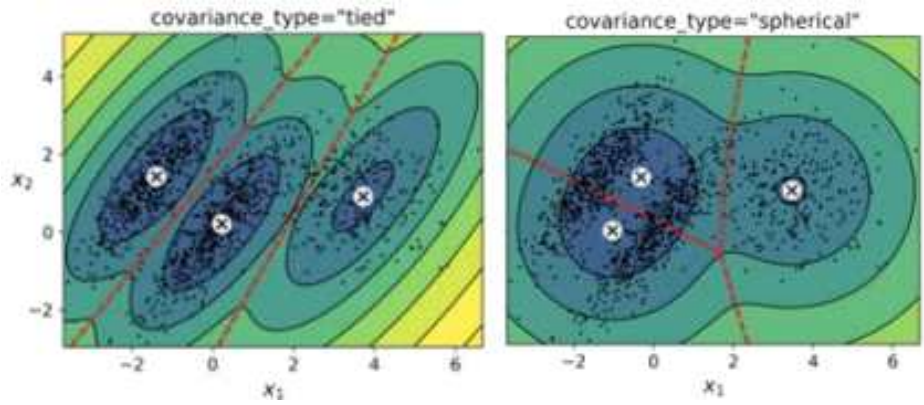
```
gm.means_
```

```
array([[ 0.05131611,  0.07521837],
       [-1.40763156,  1.42708225],
       [ 3.39893794,  1.05928897]])
```

```
gm.covariances_
```

```
array([[[ 0.68799922,  0.79606357],
        [ 0.79606357,  1.21236106]],
```

질문) GMM 모델 규제?

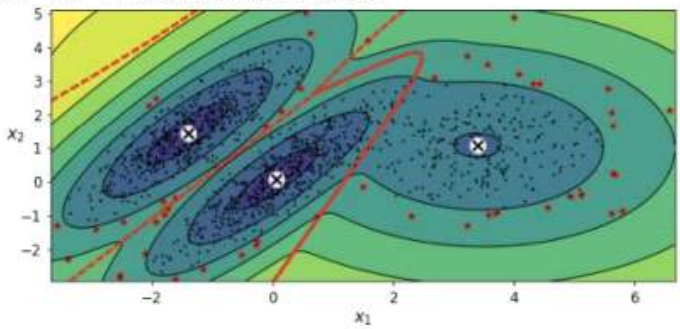


특성수가 크거나, 군집수가 많거나, 샘플이 적은 경우 최적 모델 학습이 어렵다. 공분산(covariance)에 규제를 가해서 학습을 도와줄 수 있다.

질문) 이상치 탐지는 어떻게 해?

```
densities = gm.score_samples(X)
density_threshold = np.percentile(densities, 4)
anomalies = X[densities < density_threshold]
```

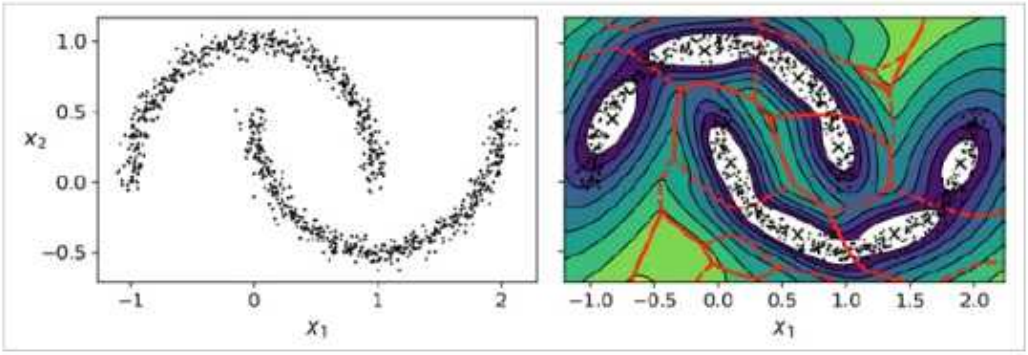
```
그림 저장: mixture_anomaly_detection_plot
```



이상치 탐지는 보통과 많이 다른 샘플을 감지하는 작업이다. 가우시안혼합 모델을 활용한 이상치 탐지는 밀도가 임계값보다 낮은 지역에 있는 샘플을 이상치로 간주 가능하다.

질문) 가우시안 혼합 모델의 장단점 알려줘

장점은 타원형 클러스터에 잘 작동한다는 것이다. 단점은 다른 모양을 가진 데이터 셋에서는 성능이 좋지 않다.



달모양 데이터에 적용하는 경우

억지로 타원을 찾으려 시도한다. 2개가 아니라 8개의 클러스터를 찾는다.

※ 저희 조는 K평균, 가우시안 혼합 모델, DBSCAN이 각각 어떤 상황에서 어떻게 쓰이는지 고민해봤고, 정확한 이해를 위해 조사를 진행하였습니다.

[K-평균]

K-평균 클러스터링은 데이터를 k개의 클러스터로 그룹화하는 비지도 학습 알고리즘입니다. 이 알고리즘은 각 클러스터의 중심을 찾아 해당 중심과의 거리를 최소화하는 방식으로 동작합니다. K-평균은 클러스터의 수(k)가 명확하게 알려져 있을 때 유용하고, 클러스터 크기가 비슷하고 밀집도가 균일한 경우에 잘 작동합니다. 클러스터링 결과는 각 데이터 포인트가 어떤 클러스터에 속하는지를 나타냅니다. 중심과의 거리가 최소인 클러스터에 할당된 데이터는 해당 클러스터에 속한다고 판단할 수 있습니다. 하지만, 초기 중심값에 따라 결과가 달라질 수 있습니다. 클러스터의 모양이 원형이 아닐 경우에는 성능이 떨어질 수 있습니다.

[가우시안 혼합 모델]

가우시안 혼합 모델은 데이터를 여러 가우시안 분포로 모델링하여 클러스터를 형성하는 알고리즘입니다. 각 데이터 포인트는 여러 클러스터에 속할 확률을 갖습니다. 가우시안 혼합 모델은 클러스터의 형태가 다양하거나 크기가 다를 때 유용합니다. 또한, 클러스터 간의 공분산이 다를 경우에도 적합합니다. 각 데이터 포인트는 각 클러스터에 속할 확률을 가지며, 가장 높은 확률을 가진 클러스터에 할당됩니다. 클러스터의 모양이 다양하게 모델링 될 수 있습니다. 하지만, 초기 매개변수 설정에 민감하며 수렴이 상대적으로 느릴 수 있습니다. 또한 이상치에 대한 민감성이 있을 수 있습니다.

[DBSCAN]

DBSCAN은 데이터 포인트의 밀도를 기반으로 클러스터를 형성하는 알고리즘으로, 클러스터의 수를 사용자가 사전에 지정할 필요가 없습니다. DBSCAN은 클러스터의 형태나 크기가 다양하게 분포하고 있을 때 사용합니다. 이상치를 감지하고자 할 때 유용하며, 클러스터의 밀도가 다를 때 효과적입니다. 클러스터는 데이터 포인트의 밀도가 높은 지역으로 형성되며, 이상치는 밀도가 낮은 지역으로 분류됩니다. 클러스터 간의 밀도가 다르게 설정될 수 있습니다.
하지만, 하이퍼파라미터 설정에 민감하며, 밀도가 일정하지 않은 데이터에 대해 잘 작동하지 않을 수 있습니다. 또한 거리 측정 방식에 따라 결과가 달라질 수 있습니다.