**DBXSCAN 알고리즘 하이퍼파라미터**

# 1. ε (Epsilon)

## 정의 및 역할

ε는 두 데이터 포인트가 같은 클러스터에 속하기 위해 가져야 하는 최대 거리입니다. 쉽게 말해, 이 거리를 기준으로 주변의 데이터 포인트들을 탐색하여 클러스터를 형성합니다. ε 값이 작을수록 더 작은 반경 내에서 데이터 포인트를 묶어 클러스터를 형성하게 되며, 클러스터의 크기가 작아집니다. 반대로 ε 값이 크면, 더 넓은 영역에서 포인트를 묶어 클러스터가 형성되어 클러스터 크기가 커집니다.

## 설정 방법

작은 ε: 클러스터가 너무 작아지고, 많은 노이즈 포인트가 발생할 수 있습니다. 이는 과소 적합(underfitting)의 위험을 증가시킵니다.  
큰 ε: 클러스터가 너무 커져 다른 클러스터와 병합되거나, 데이터 포인트 간의 명확한 구분이 어려워질 수 있습니다. 이는 과적합(overfitting)의 위험을 증가시킵니다.

## 설정 시 고려 사항

데이터의 밀도 분포를 이해하는 것이 중요합니다. ε 값을 설정하기 전에 데이터셋의 특성을 분석하는 것이 좋습니다. 일반적으로 k-거리 그래프(k-distance graph)를 사용하여 ε를 설정합니다. 이는 각 포인트에 대해 k번째 이웃과의 거리를 그래프로 나타내고, 가장 급격한 변화가 나타나는 지점이 최적의 ε 값으로 간주됩니다.

# 2. MinPts

## 정의 및 역할

MinPts는 유효한 클러스터로 간주되기 위해 특정 반경 내(ε)에 존재해야 하는 최소 데이터 포인트 수입니다. 이 값은 클러스터가 형성될 수 있는 최소 밀도를 결정하는 중요한 역할을 합니다.

## 설정 방법

작은 MinPts: 너무 작은 값을 설정하면, 매우 작은 클러스터들이 형성될 수 있습니다. 이 경우, 데이터가 지나치게 세분화되어 불필요한 클러스터가 다수 생성될 수 있습니다.  
큰 MinPts: 너무 큰 값을 설정하면, 클러스터가 잘 형성되지 않거나, 클러스터로 인식되지 않고 노이즈로 분류되는 데이터가 많아질 수 있습니다.

## 설정 시 고려 사항

일반적으로 MinPts는 데이터셋의 차원(Dimension)에 따라 설정합니다.  
- 2차원 데이터에서는 일반적으로 4 이상의 값을 권장합니다.  
- 고차원 데이터에서는 차원이 증가할수록 MinPts 값을 증가시켜야 합니다.  
데이터의 특성 및 데이터 포인트의 밀도 분포를 이해하고 이에 맞게 설정하는 것이 중요합니다.

# 3. 하이퍼파라미터 튜닝 전략

## 3.1 실험적 접근

처음에는 경험적 규칙을 사용하여 적절한 ε와 MinPts 값을 선택합니다. 이후, 다양한 ε와 MinPts 값을 사용하여 여러 번 실험해 보고, 결과를 비교하여 최적의 조합을 찾습니다.

## 3.2 시각화 활용

k-거리 그래프와 클러스터링 결과를 시각화하여 ε와 MinPts의 영향을 분석합니다. 클러스터의 수, 노이즈 포인트의 수, 클러스터의 크기 등을 시각적으로 분석하여 최적의 값을 찾아냅니다.

## 3.3 도메인 지식 적용

데이터를 잘 이해하고 있는 도메인 전문가의 지식을 활용하여 ε와 MinPts 값을 설정할 수 있습니다. 도메인 지식은 특정 클러스터의 크기, 밀도, 기대되는 노이즈 수준 등에 대한 이해를 기반으로 합니다.

# 4. 최적화의 중요성

DBXSCAN의 하이퍼파라미터 설정은 클러스터링의 성능과 결과에 결정적인 영향을 미칩니다. 따라서 데이터의 특성, 밀도, 그리고 기대되는 결과를 고려하여 신중하게 설정해야 합니다. ε와 MinPts의 조합은 데이터셋의 특성에 따라 다르므로, 다양한 시도를 통해 최적의 값을 찾아내는 것이 중요합니다.