

DBSCAN 심층 분석 보고서

Density-Based Spatial Clustering of Applications with Noise

1. 개요 및 핵심 철학

DBSCAN은 데이터가 위치한 공간의 '**밀도(Density)**'를 기준으로 군집을 형성하는 알고리즘입니다. K-Means처럼 거리에만 의존하지 않기 때문에, 데이터가 어떤 기하학적 모양(초승달 모양, 도넛 모양 등)을 하고 있더라도 군집을 잘 찾아냅니다. 또한 이름에 포함된 'Noise'에서 알 수 있듯이, **노이즈 데이터를 효과적으로 감지**하여 제외할 수 있습니다.



핵심 컨셉: "사람들이 붐비는 곳"

"친구가 많은 사람끼리 모인다."

특정 반경 내에 이웃이 충분히 많이 있으면 '핵심 포인트'로 인정하고 군집을 확장해 나갑니다. 반대로 주변에 아무도 없는 데이터는 과감히 '노이즈(Outlier)'로 분류합니다.

⚙️ 2. 상세 작동 메커니즘

DBSCAN은 두 가지 핵심 파라미터(ϵ : 반경, MinPts: 최소 이웃 수)를 사용하여 점들을 세 가지 유형으로 분류합니다.

1. 핵심 점 (Core Point)

반경 ϵ 내에 자신을 포함하여 MinPts개 이상의 점이 있는 경우입니다. 군집의 중심이 되는 점들입니다.

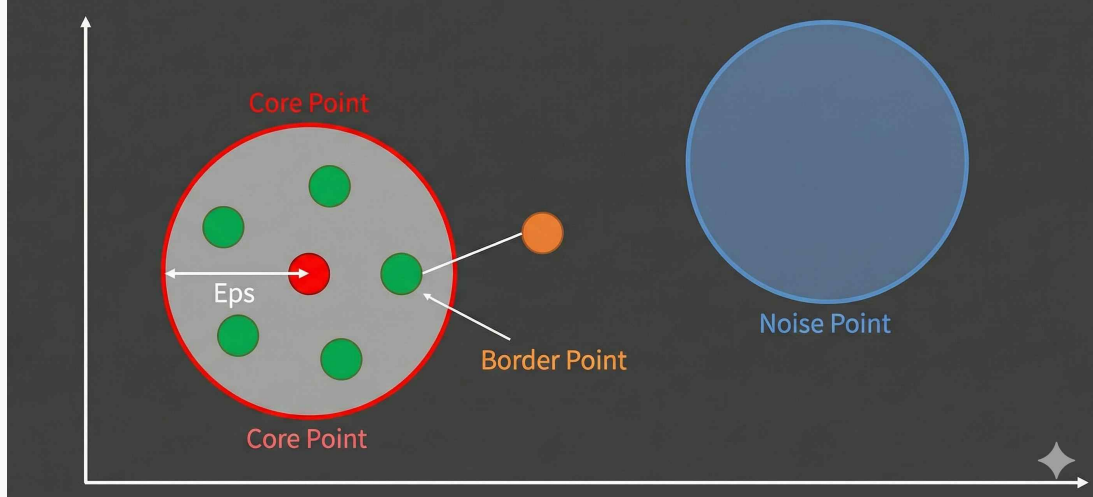
2. 경계 점 (Border Point)

핵심 점의 반경 내에는 있지만, 스스로는 MinPts 개수를 채우지 못한 점입니다. 군집의 외곽을 형성합니다.

3. 노이즈 점 (Noise Point)

어떤 핵심 점의 반경 내에도 없고, 스스로도 이웃이 부족한 점입니다. 이는 이상치(Outlier)로 간주되어 **군집에서 제외(-1 레이블)**됩니다.

DBSCAN 밀도에 따른 점의 분류 (Core, Border, Noise)



🖼️ 그림 1: 밀도에 따른 점의 분류 (Core, Border, Noise)

3. K-Means vs DBSCAN

K-Means는 가장 대중적이지만 모양과 이상치에 취약합니다. DBSCAN은 이를 완벽히 보완합니다.

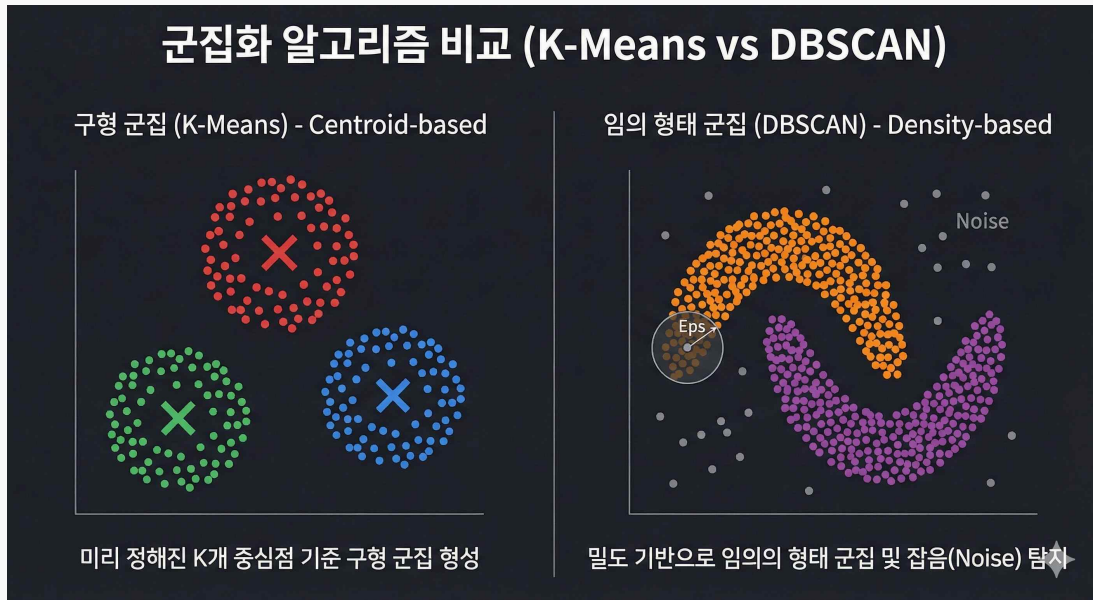


그림 2: 구형 군집(K-Means) vs 임의의 형태 군집(DBSCAN)

항목	K-Means	DBSCAN
군집 개수(K)	사용자가 미리 지정해야 함	자동으로 결정됨 (데이터 밀도에 따라)
군집 형태	원형(Spherical) 군집만 잘 찾음	불규칙한 모양 (반달, 꼬불꼬불 등)도 잘 찾음
노이즈 처리	모든 점을 강제로 군집에 할당 (이상치에 영향 많이 받음)	밀도가 낮은 점은 노이즈로 분류하여 버림
속도	매우 빠름 ($O(N)$)	상대적으로 느림 ($O(N \log N) \sim O(N^2)$)

 4. 장단점 심층 분석 장점 (Pros)

- ✓ **군집 개수 자동 탐지:** K를 몇 개로 할지 고민할 필요가 없습니다.
- ✓ **강력한 이상치 제어:** 노이즈 데이터가 군집 형성을 방해하지 않도록 걸러냅니다.
- ✓ **기하학적 유연성:** 데이터 분포가 복잡하고 기괴한 모양이어도 정확하게 군집을 분리합니다.

 단점 및 주의사항

- ✗ **파라미터 민감성:** ϵ 과 MinPts 값을 어떻게 설정하느냐에 따라 결과가 극적으로 달라집니다. (도메인 지식 필요)
- ✗ **밀도 차이:** 데이터 간의 밀도가 균일하지 않으면(어떤 군집은 뾰뾰하고 어떤 군집은 널널하면) 잘 동작하지 않습니다.

5. 결론 (Conclusion)

“

DBSCAN은
"데이터의 진짜 모양"을
찾아주는 탐험가입니다.

”

최종 요약

- ✓ 이상치(Outlier) 탐지가 필요한 프로젝트에 매우 유용합니다.
- ✓ 지리 공간 데이터(위도/경도) 분석에 특히 강력합니다.
- ✓ 데이터의 차원이 너무 높으면(Curse of Dimensionality) 성능이 떨어질 수 있습니다.