Posgrados CUCEA
El mejor lugar para el talento

DBSCAN: Density-Based Spatial Clustering

Maestría en ciencias de los datos

Centro Universitario de Ciencias Económico-Administrativas

Por: Ilse Arredondo Reyes

Fecha 26 de Abril del 2025

MCD Maestría en Ciencia de los Datos

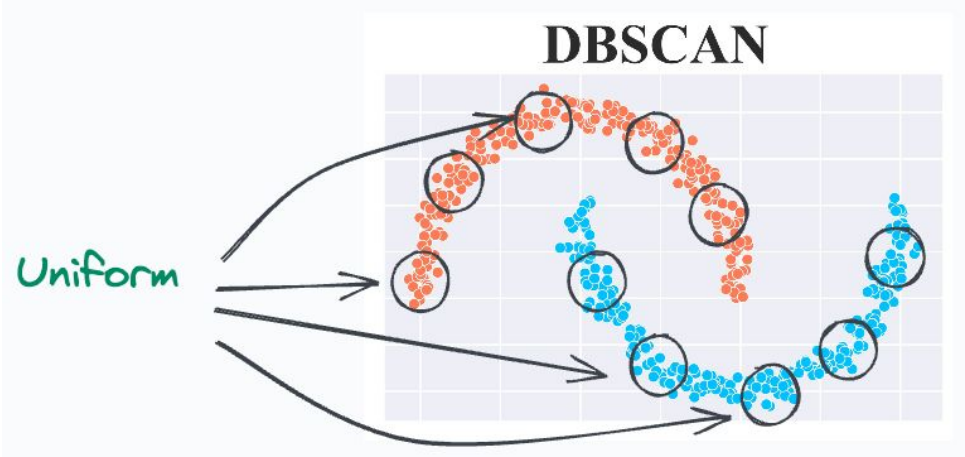
¿Qué es el Clustering?

Objetivo del Clustering: Agrupar puntos de datos similares sin conocimiento previo de los grupos (Aprendizaje No Supervisado).

¿Por qué Agrupar? Descubrir patrones ocultos, segmentar datos, detección de anomalías, reducción de datos.

Desafío Común: Muchos algoritmos (como K-Means) asumen que los clústeres son esféricos y tienen dificultades con formas complejas o ruido.

Entra DBSCAN: Un enfoque basado en densidad diseñado para superar estas limitaciones.



DBSCAN

Agrupamiento Espacial Basado en Densidad de Aplicaciones con Ruido (Density-Based Spatial Clustering of Applications with Noise).

Idea Central: Define los clusters como regiones continuas de alta densidad de puntos, separadas por regiones de baja densidad.

Características Clave:

- No requiere especificar el número de clústeres de antemano.
- Puede encontrar clústeres de formas arbitrarias.
- Identifica explícitamente puntos de ruido (outliers).
- Basado en dos parámetros clave: ϵ (eps) y MinPts.

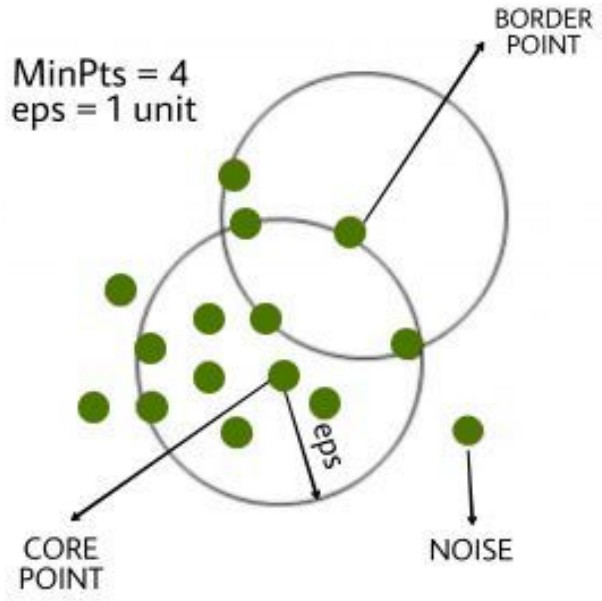
Conceptos Clave (1/2) - Parámetros

Parámetro 1: ϵ (Épsilon / eps)

- Representa un umbral de distancia.
- Define el radio de una "vecindad" alrededor de cada punto de datos..

Parámetro 2: MinPts (Puntos Mínimos)

- El número mínimo de puntos de datos requeridos dentro de la vecindad ϵ de un punto para que éste sea considerado "denso".
- Este conteo incluye al propio punto.



Conceptos Clave (2/2) - Clasificación de Puntos

Basado en ϵ y MinPts, los puntos se clasifican en tres tipos:

- **Punto Central (Core Point):** Un punto que tiene al menos MinPts puntos (incluyéndose a sí mismo) dentro de su vecindad ϵ. Estos puntos están en el interior de un clúster.

- **Punto Frontera (Border Point):** Un punto que no es un punto central (menos de MinPts vecinos) pero se encuentra dentro de la vecindad ϵ de un punto central. Estos puntos están en el borde de un clúster.

- **Punto Ruido (Noise Point / Outlier):** Un punto que no es ni central ni frontera. No es lo suficientemente dense ni está lo suficientemente cerca de una región densa.

Conceptos Centrales - Alcanzabilidad (Reachability)

Cómo se forman los clústeres: conectando puntos densos.

Directamente Alcanzable por Densidad (Directly Density-Reachable):

El punto q es directamente alcanzable por densidad desde el punto p si: q está dentro de la vecindad ϵ de p.

p es un punto central.

Alcanzable por Densidad (Density-Reachable): El punto q es alcanzable por densidad desde el punto p si existe una cadena de puntos p1,p2,...,pn donde p1=p, pn=q, y cada pi+1 es directamente alcanzable por densidad desde pi.

Conectado por Densidad (Density-Connected): Los puntos p y q están conectados por densidad si existe un punto central o tal que tanto p como q son alcanzables por densidad desde o.

Definición de Clúster: Un clúster DBSCAN es un conjunto maximal de puntos conectados por densidad. Cualquier punto no alcanzable desde un punto central se considera ruido.

Las Matemáticas Detrás de DBSCAN

Métrica de Distancia: Cómo medimos la "cercanía".

Más común: Distancia Euclidiana en espacio -dimensional:

$$d(p,q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$$

Se pueden usar otras métricas (Manhattan, Coseno, etc.) dependiendo del tipo de datos y el dominio. La elección de la métrica es crucial.

Consulta de Vecindad- (Nϵ(p)): El conjunto de puntos dentro de la distancia ϵ del punto p.

Consulta de Vecindad- (Nϵ(p)): El conjunto de puntos dentro de la distancia ϵ del punto p.

$$N_\epsilon(p) = q \in D \mid d(p,q) \leq \epsilon$$

Donde D es el conjunto de datos.

Condición de Punto Central: Un punto p es un punto central si el tamaño (cardinalidad) de su vecindad ϵ alcanza o supera MinPts.

Nota: Encontrar eficientemente para todos los puntos es clave para el rendimiento (a menudo utiliza índices espaciales como k-d trees o R-trees).

$$|N_\epsilon(p)| \geq MinPts$$

ALGORITMO - PASO A PASO

1. Inicializar todos los puntos como no visitados.
2. Iterar sobre cada punto p no visitado en el conjunto de datos D:

    a. Marcar p como visitado.
    b. Encontrar su vecindad ϵ: Nϵ(p).
    c. Si $|N_\epsilon(p)| < MinPts$:
        i. Marcar p (temporalmente) como RUIDO.
            1. Sino (p es un punto central)
        ii. Crear un nuevo clúster C. Añadir p a C.
        iii. Inicializar un conjunto semilla S con todos los puntos en Nϵ(p) (excluyendo p mismo).
        iv. Mientras S no esté vacío:
            1. Seleccionar y eliminar un punto q de S.
            2. Si q fue marcado como RUIDO, cambiar su estado y añadirlo al clúster C.
            3. Si q no ha sido visitado:
                a. Marcar q como visitado.
                b. Encontrar su vecindad ϵ: Nϵ(q).
                c. Si $|N_\epsilon(q)| \geq MinPts$ (q también es un punto central)

            * Añadir todos los puntos de Nϵ(q) que no estén visitados o marcados como RUIDO al conjunto semilla S.

                d. Si q aún no está asignado a ningún clúster, añadir q al clúster C.

3. Fin del Bucle. Todos los puntos están ahora asignados a un clúster o marcados como RUIDO.

HIPERPARÁMETROS

- **eps (ϵ):**
  - **Impacto:** Controla cuán cerca deben estar los puntos para ser considerados vecinos. Afecta el tamaño y número de clústeres.
  - *Demasiado pequeño:* La mayoría de los puntos se vuelven ruido; muchos clústeres pequeños.
  - *Demasiado grande:* Los clústeres se fusionan; menos clústeres, más grandes.
  - **Ajuste:** A menudo se elige usando un *gráfico de distancia-k*. Calcular la distancia al k-ésimo vecino más cercano (donde k=MinPts−1) para todos los puntos. Ordenar estas distancias y buscar el punto de "codo" en la gráfica.
- **MinPts:**
  - **Impacto:** Controla la densidad mínima requerida para formar un núcleo de clúster. Afecta la sensibilidad al ruido.
  - *Demasiado pequeño:* Podrían formarse clústeres dispersos; se identifica menos ruido.
  - *Demasiado grande:* Requiere mayor densidad; clústeres más pequeños podrían omitirse y etiquetarse como ruido.
  - **Ajuste:** A menudo se establece basado en el conocimiento del dominio o la dimensionalidad (D). Heurísticas comunes: MinPts≥D+1 o MinPts=2×D. Valores más grandes hacen los resultados más robustos pero podrían omitir clústeres más pequeños.
- **metric:**
  - La función de distancia utilizada (ej., 'euclidean', 'manhattan'). La elección depende mucho de la naturaleza de las características de los datos.

Métricas de Rendimiento (Evaluación)

Evaluar el clustering es inherentemente subjetivo sin etiquetas de verdad fundamental (ground truth).

Métricas Internas (Evalúan basado en la estructura del clúster):

Silhouette Score: Mide cuán similar es un objeto a su propio clúster en comparación con otros clústeres. Rango de -1 a 1. Más alto es mejor (clústeres bien separados). Considera compacidad y separación.

Davies-Bouldin Index: Ratio de la dispersión intra-clúster a la separación inter-clúster. Más bajo es mejor (clústeres compactos y lejos entre sí).

Calinski-Harabasz Index (Criterio de Ratio de Varianza): Ratio de la suma de la dispersión inter-clúster a la dispersión intra-clúster. Más alto es mejor.

Métricas Externas (Requieren etiquetas verdaderas - para benchmarking/pruebas):

Adjusted Rand Index (ARI): Mide la similitud entre las agrupaciones verdadera y predicha, corregida por azar. Rango de -1 a 1. Más alto es mejor.

Normalized Mutual Information (NMI): Mide la dependencia mutua entre las agrupaciones verdadera y predicha, normalizada. Rango de 0 a 1. Más alto es mejor.

Homogeneidad, Completitud, V-measure: Evalúan si los clústeres contienen solo miembros de una única clase (homogeneidad), si todos los miembros de una clase están en el mismo clúster (completitud), y su media armónica (V-measure). Rango [0, 1]. Más alto es mejor.

Nota sobre el Ruido: Algunas métricas pueden necesitar ajuste o interpretación ya que DBSCAN etiqueta explícitamente puntos de ruido (a menudo asignados a la etiqueta de clúster -1), lo cual podría no ser directamente comparable en cálculos de métricas estándar.

VENTAJAS

No necesita especificar el número de clústeres: Descubre clústeres orgánicamente basado en la densidad.

Encuentra formas arbitrarias: No está limitado a clústeres convexos/esféricos como K-Means.

Robusto al ruido: Identifica y maneja explícitamente outliers.

Conceptualmente intuitivo: Basado en ideas comprensibles de densidad y alcanzabilidad.

Independiente del orden (en su mayoría): Los puntos centrales y de ruido siempre se determinan de la misma manera. Los puntos frontera podrían teóricamente asignarse a diferentes clústeres dependiendo del orden de procesamiento, pero esto es raro en la práctica.

Desventajas

Sensibilidad a los Parámetros: El rendimiento depende mucho de elegir buenos valores de y MinPts, lo cual puede ser no trivial.

Dificultades con clústeres de densidad variable: No puede agrupar eficazmente conjuntos de datos donde diferentes regiones tienen densidades muy diferentes usando una única configuración global de ϵ y MinPts. (Extensiones como OPTICS o HDBSCAN abordan esto).

"Maldición de la Dimensionalidad": Las medidas de distancia se vuelven menos significativas en espacios de muy alta dimensión, impactando la estimación de densidad.

Complejidad Computacional: Puede ser $O(n^2)$ en el peor caso sin indexación espacial. Con indexación (como k-d trees), la complejidad promedio es a menudo O(n log (n)), pero la construcción del índice puede ser costosa.

Aplicaciones en el Mundo Real

Análisis de Datos Geográficos:

Identificación de puntos calientes (hotspots) de crimen.

Encontrar clústeres de puntos de interés (ej., restaurantes, tiendas).

Análisis de la distribución espacial de enfermedades o eventos.

Detección de Anomalías:

Identificación de transacciones fraudulentas (puntos ruido).

Detección de intrusiones en tráfico de red.

Encontrar productos defectuosos a partir de datos de sensores.

Biología y Medicina:

Agrupamiento de datos de expresión génica.

Análisis de poblaciones celulares en citometría de flujo.

Procesamiento de Imágenes:

Segmentación de imágenes basada en densidad/color de píxeles.

Sistemas de Recomendación:

Agrupación de usuarios con comportamientos similares (aunque a menudo de alta dimensión).

Conclusión y Resumen

DBSCAN es un potente algoritmo de clustering no supervisado basado en densidad.

Fortalezas clave: Maneja formas arbitrarias, identifica ruido, no requiere pre-especificar el número de clústeres.

Se basa en conceptos intuitivos: vecindad-, MinPts, puntos centrales/frontera/ruido, alcanzabilidad por densidad.

Desafíos principales: Ajuste de parámetros (ϵ, MinPts) y clústeres de densidad variable.

Ampliamente utilizado para datos espaciales, detección de anomalías y escenarios donde las formas de los clústeres son complejas.