DBSCAN

¿Qué es DBSCAN?

DBSCAN significa **Density-Based Spatial Clustering of Applications with Noise**.

Es un algoritmo de agrupamiento basado en densidad: encuentra "grupos" de puntos que están muy juntos y separa el "ruido" o puntos aislados.

¿Cómo funciona DBSCAN? (Explicado paso a paso)

1. Parámetros clave

- **ε (epsilon):** Radio de vecindad (la distancia máxima para considerar que dos puntos son vecinos).
- **minPts:** Número mínimo de puntos que debe tener una región para ser considerada un grupo denso (un "cluster").

2. Tipos de puntos

- **Punto central (core point):** Tiene al menos `minPts` puntos (incluyéndose a sí mismo) en un radio de `ε`.
- **Punto frontera (border point):** No tiene suficientes puntos en su vecindad, pero está cerca de un core point. Se unen al *cluster*, pero no servirán para incluir otros puntos en el *cluster*.
- **Ruido (noise point):** No es ni core ni border; está aislado.

3. Algoritmo

- Se visita cada punto del dataset.
- Si un punto tiene al menos `minPts` vecinos en su radio `ε`, se inicia un nuevo cluster con ese punto.
- Todos los puntos alcanzables dentro de ese radio, y los que a su vez sean "core points", se van agregando al cluster (es como expandir una mancha de pintura).
- Si un punto no es alcanzable y tampoco es core, se etiqueta como **ruido**.

Visualización sencilla

Imagina que tienes muchas monedas tiradas en una mesa:

- Si pones una moneda sobre otra y hay muchas juntas, forman una **mancha** o “cluster”.
- Las monedas que están solas o muy separadas son consideradas **ruido**.

Ventajas de DBSCAN

- Detecta clusters de **forma arbitraria** (no necesariamente redondos o de igual tamaño).
- No necesitas especificar el número de clusters desde el inicio (a diferencia de k-means).
- Identifica **ruido** o valores atípicos automáticamente.

Limitaciones

- Elegir los valores correctos de `ε` y `minPts` puede ser complicado.
- No funciona bien si los clusters tienen densidades muy diferentes.

Ejemplo práctico

Supón que tienes datos de localización de personas en una ciudad:

- Usas DBSCAN para detectar **zonas densamente pobladas** (clusters), por ejemplo, dónde se forman aglomeraciones (como conciertos o fiestas).
- Las personas solas, fuera de esas zonas, son clasificadas como **ruido**.

Resumido en pseudocódigo:

1. Para cada punto, mira cuántos puntos tiene cerca (`ε`).
2. Si cumple con `minPts`, forma un cluster.
3. Expande ese cluster con todos los puntos vecinos.
4. Marca como ruido lo que no pertenezca a ningún cluster.