Aprendizaje No Supervisado

Maestría en Ciencia de Datos

Lucas Fernández Piana

Primavera 2022

Universidad de San Andrés

DBSCAN

IDEAS

Pensar en cluster por densidades nos da un nuevo punto de vista sobre el problema que no está relacionado directamente con la distancia o disimilaridad.

IDEAS

Pensar en cluster por densidades nos da un nuevo punto de vista sobre el problema que no está relacionado directamente con la distancia o disimilaridad.

Filosofía

”Los clusters están definidos por regiones de alta concentración o densidad de puntos en el espacio que a su vez se encuentran separadas por regiones de baja concentración.”

IDEAS

Estos métodos fueron creados para captar clusters de diferentes formas.

DBSCAN

Las siglas **DBSCAN** provienen de Density-Based Spatial Clustering of Applications with Noise.

DBSCAN

Las siglas **DBSCAN** provienen de Density-Based Spatial Clustering of Applications with Noise.

Intuición

- Para un punto dentro del cluster, la densidad alrededor de ese punto debería exceder un umbral.
- El conjunto de puntos dentro de un cluster debería estar espacialmente relacionados.

A partir de estas intuiciones formalicemos los conceptos ...

DEFINICIONES

Sea (E, d) un espacio métrico donde tengo mi conjunto de datos D.

- ϵ-Vecindad (ϵ-Neighborhood) de un punto p de radio ϵ se define como

$$N_\epsilon(p) = \{q \in D : \ d(p, q) < \epsilon\}.$$

- *MinPts* : es un número natural que funciona como umbral. Este es el nombre que usualmente se encuentra en la literatura.

DEFINICIONES

Dados ϵ y *MinPts* podemos definir tres clases de puntos en nuestros datos:

- Punto Núcleo (core-point): decimos que p es un punto núcleo si su entorno $N_\epsilon(p)$ contiene más de *MinPts* datos. Es decir, si $\#N_\epsilon(p) \geq MinPts$.
- Punto Borde (border-point): decimos que q es un punto borde si su entorno $N_\epsilon(q)$ contiene menos de *MinPts* datos y existe un punto núcleo p tal que $q \in N_\epsilon(p)$. O sea que q no es núcleo, pero hay algún punto núcleo que lo contiene en su vecindad.
- Ruido (noise-point): decimos que o es ruido si no es punto núcleo o borde.

DEFINICIONES

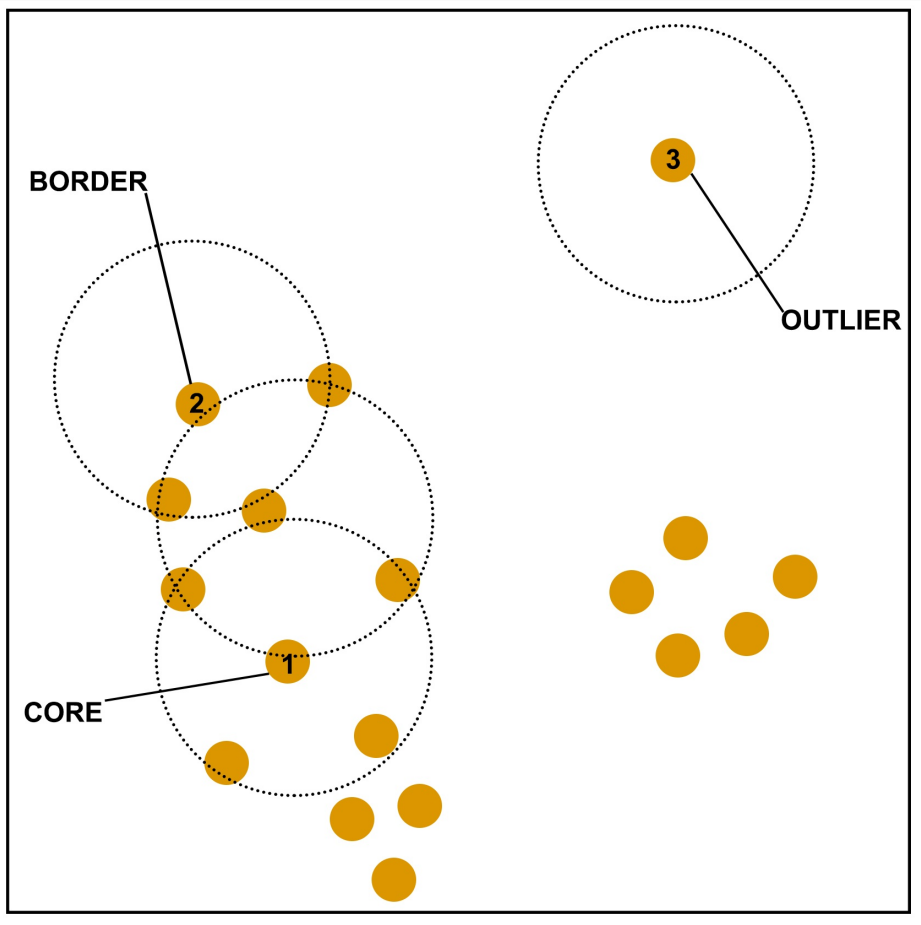

Figura 1: MinPts=5

DEFINICIONES

- Un punto q es directamente alcanzable mediante densidad desde un punto p (q is directly density-recheable from p), si p es un punto núcleo y $q \in N_\epsilon(p)$.
- Un punto q es alcanzable mediante densidad desde un punto p (q is density-recheable from p) si existe una sucesión $p_1, \ldots, p_n$ tales que $p_1 = q$, $p_n = p$ y p_{i+1} directamente alcanzable desde p_i.

DEFINICIONES

- Un punto q es directamente alcanzable mediante densidad desde un punto p (q is directly density-recheable from p), si p es un punto núcleo y $q \in N_\epsilon(p)$.
- Un punto q es alcanzable mediante densidad desde un punto p (q is density-recheable from p) si existe una sucesión $p_1, \ldots, p_n$ tales que $p_1 = q$, $p_n = p$ y p_{i+1} directamente alcanzable desde p_i.

Observación: la relación de ser alcanzable es asimétrica. Es decir, es posible que q sea alcanzable desde p y al mismo tiempo p no es alcanzable desde q.

DEFINICIONES

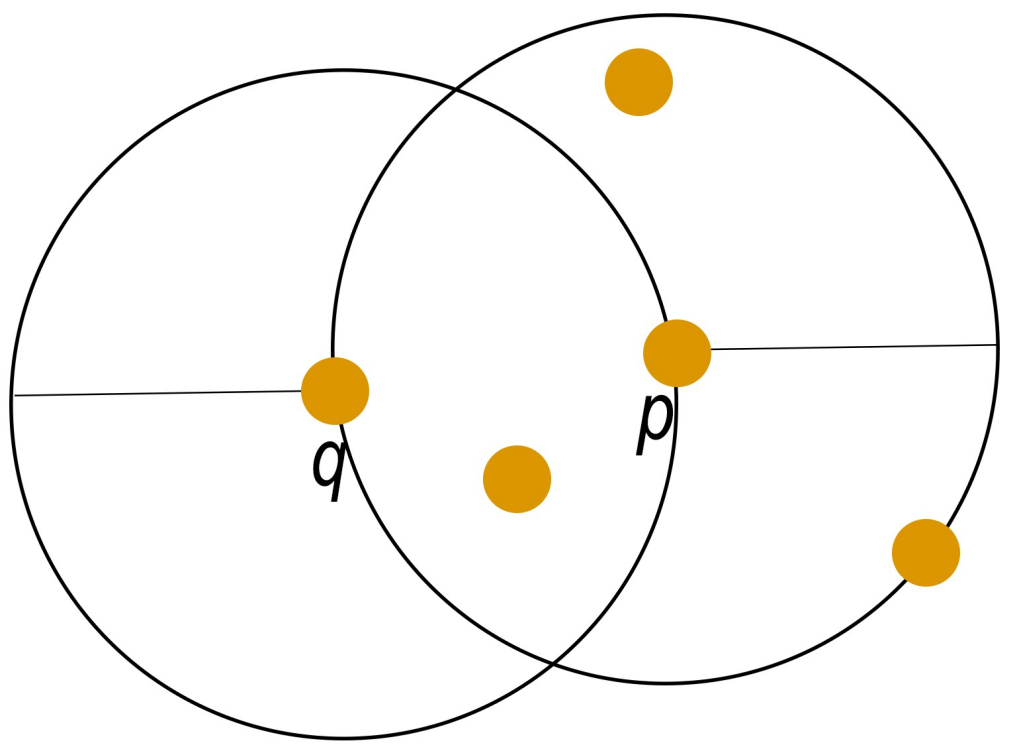

Figura 2: MinPts=4

DEFINICIONES

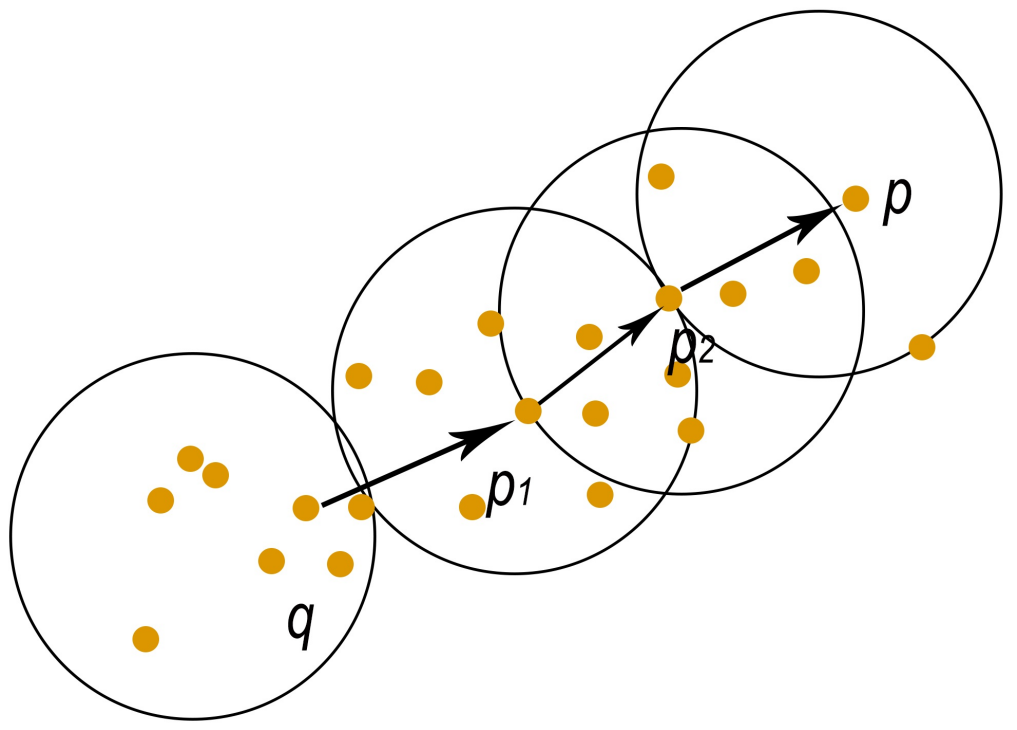

Figura 3: MinPts=7

Conectividad

Decimos que p está conectado mediante densidad a un punto q (density-conected) si existe un punto o tal que p y q son alcanzables desde o.

Conectividad

Decimos que p está conectado mediante densidad a un punto q (density-conected) si existe un punto o tal que p y q son alcanzables desde o.

Observación: esta relación sí es simétrica. Es decir, si p está conectado a q, entonces q está conectado a p.

DEFINICIONES

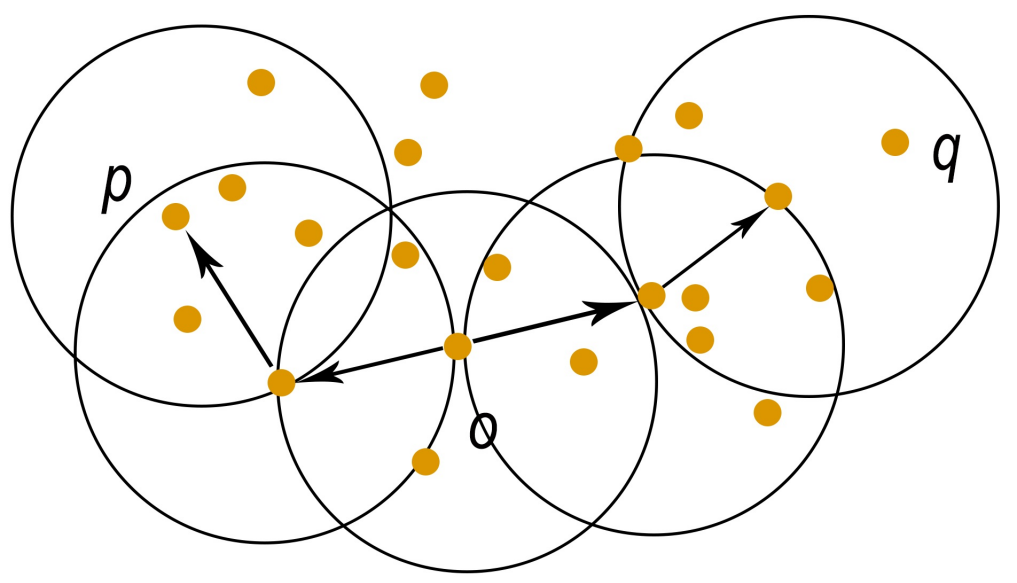

Figura 4: MinPts=7

DEFINICIONES

Cluster

Un cluster C es un subconjunto de D que satisface dos condiciones:

- **Maximalidad:** $\forall\ p, q$ se cumple que si $p \in C$ y q es alcanzable desde p, entonces $q \in C$.
- **Conectividad:** $\forall\ p, q \in C$ se cumple que p está conectado a q.

Ruido

Sean $C_1, \ldots, C_K$ los clusters en D definimos como **ruido** al conjunto de puntos que no pertenecen al ningun cluster, es decir, $Ruido = D - \cup_{i=1}^{K} C_i$.

ALGORITMO

Fijados ϵ y *MinPts* :

1. Asignar a cada punto su correspondiente categoría: núcleo, borde o ruido.
2. Eliminar los puntos que son catalogados como ruido.
3. Juntar los puntos núcleo que son alcanzables en un cluster.
4. Juntar los puntos bordes y asignarlos a su correspondiente cluster.

HEURISTICA

¿Cómo seleccionamos ϵ?

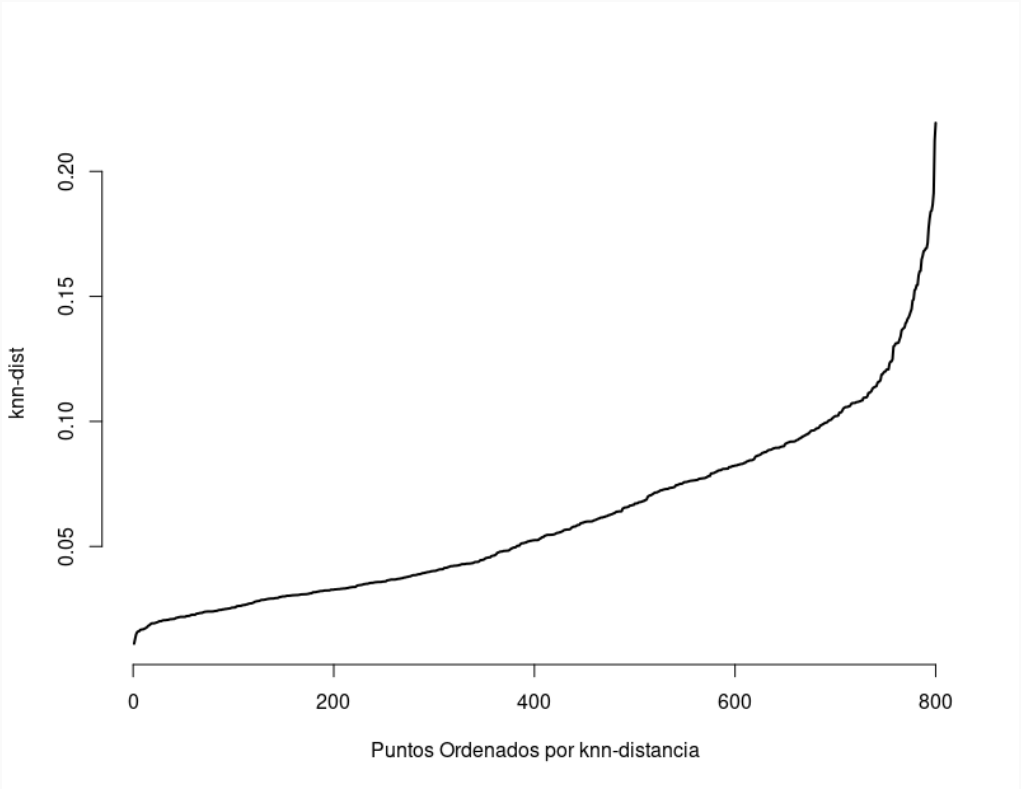

Figura 5: Distancia de cada punto a su k-ésimo vecino más cercano

HEURISTICA

¿Cómo seleccionamos ϵ?

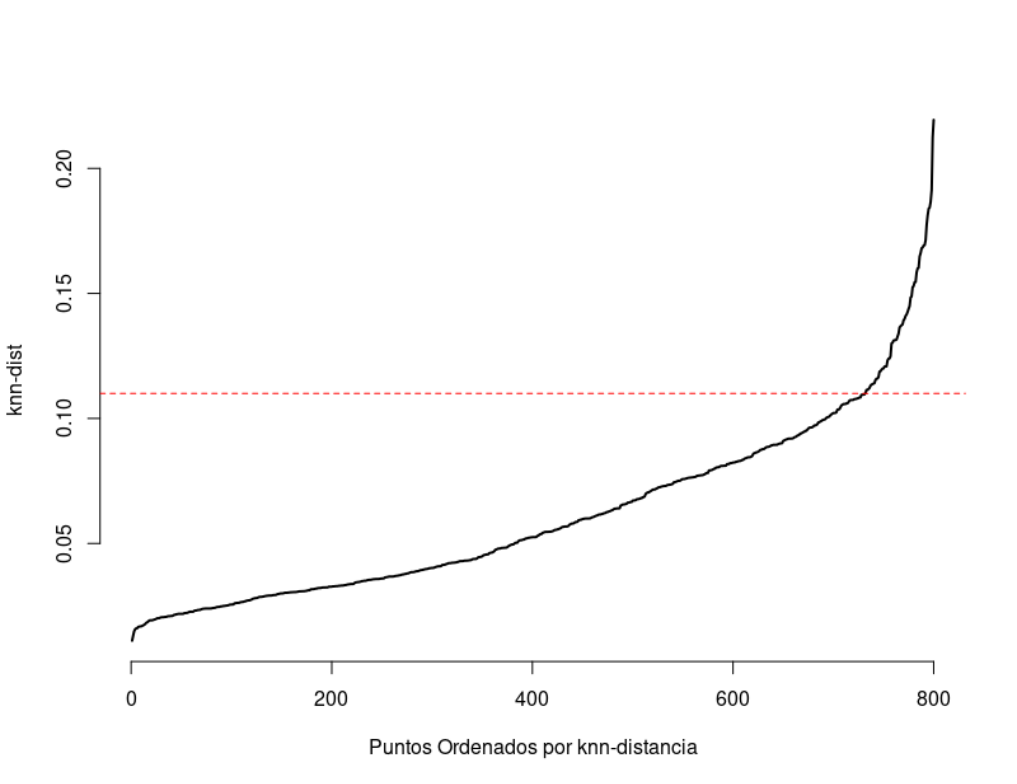

Figura 6: Corte en donde cambia la pendiente

MÁS MÉTODOS

Más métodos disponibles:

- OPTICS
- DENCLUE
- VDBSCAN
- DVBSCAN
- DBCLASD
- ST-DBSCAN

Rupanka, B. & Samarjeet, B. 2013. *"A Survey of Some Density Based Clustering Techniques"*. National Conference on Advancements in Information, Computer and Communication. DOI: 10.13140/2.1.4554.6887.

COFFEE BREAK!