

# EST-297

## Métodos de Clustering basados en Densidad

Juan Zamora O.

Junio, 2024.



PONTIFICIA  
UNIVERSIDAD  
CATÓLICA DE  
VALPARAÍSO

# Estructura de la Presentación

- 1 Revisitando técnicas basadas en representantes
- 2 DBSCAN Clustering
- 3 OPTICS Clustering
- 4 HDBSCAN Clustering

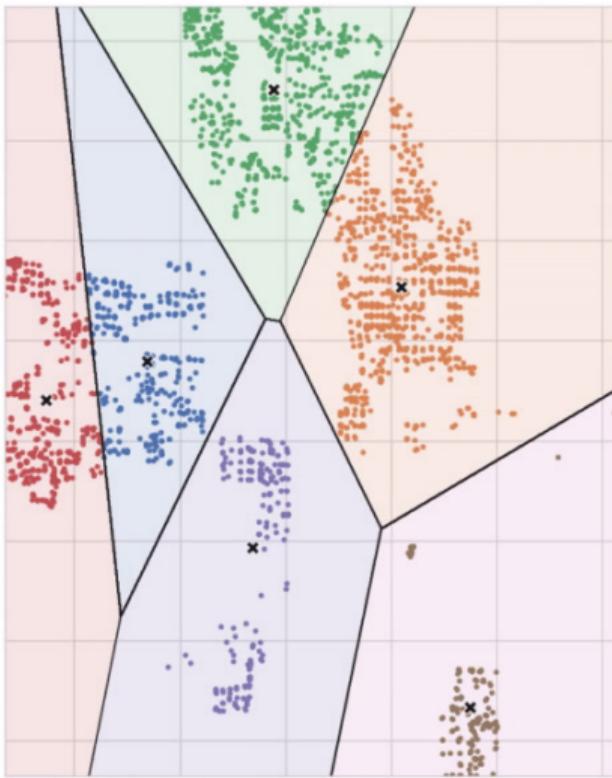
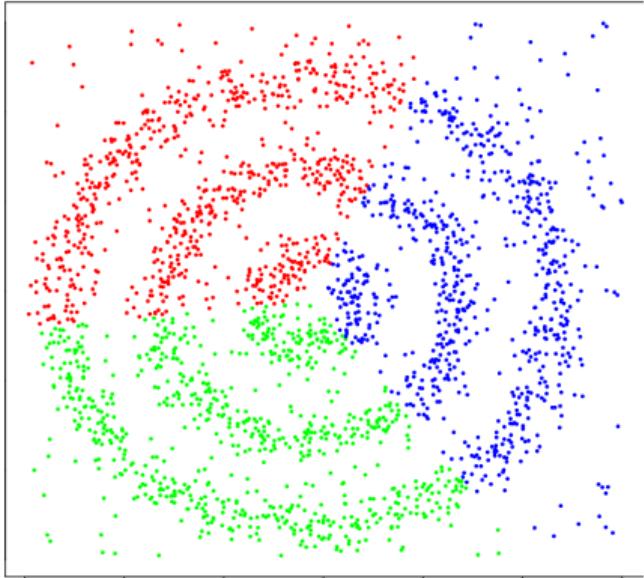
# Revisitando técnicas basadas en representantes

## **Supuesto** bastante usado

- Grupos/Clusters generados a partir de una distribución o mezcla de distribuciones simétricas (e.g. Normal)
- Expectation Maximization, K-Means y extensiones por mencionar algunos

Enfoque basado en densidad no supone forma específica.

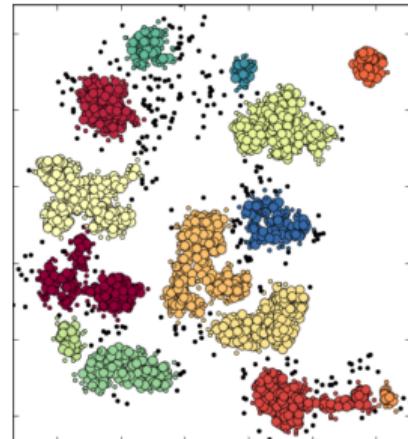
- Útil por ejemplo en datos geo-espaciales.



- Minimización de la distancia al cuadrado entre los puntos y sus respectivos centroides produce a un particionamiento tipo polígono de Voronoi. Esto no solo ocurren en 2D
- **Urge** contar con mecanismos de descubrimiento de grupos
  - con forma arbitraria.
  - identificando ruido
  - que no requieran de  $k$

# Clustering basado en densidad

- *Grupos basados en densidad* se definen como áreas densas y conectadas, separadas entre sí por áreas de menor densidad.
- El ruido se define como áreas con densidad menor que la de los clusters
- Noción de *Localidad* asociada a la definición de cluster
  - Posibilita encontrar regiones densas con forma arbitraria



- Puede ser considerado como un método no paramétrico
  - No hace supuestos respecto del número de grupos o su distribución

**Un algoritmo de clustering basado en densidad debe responder algunas preguntas:**

- ¿Como se estima la densidad?
- ¿Como se define la conectividad?

## DBSCAN Clustering

- Diseñado para descubrir clusters y ruído en los datos.
- Agrupa las observaciones mediante basándose en un umbral para el radio de búsqueda de vecinos y el número minimo de vecinos requeridos para identificar puntos clave.
  - Puntos clave  $\sim \text{core-points}$
- Cada cluster debe tener al menos un *core-point*
- *core-points* son aquellos con vecindarios ( $\epsilon$ -vec.) densos
- Un *core-point* es aquel cuyo vecindario contiene *al menos MinPts* puntos
  - Es decir, punto cuya densidad excede un determinado umbral.
- Ruído: Aquellos puntos que no pertenecen a ningún cluster

- Cuenta la cantidad de puntos en vecindarios de radio fijo ( $\epsilon$ )

$$N_\epsilon(p) = \{q \in \mathbf{D} | dist(p, q) \leq \epsilon\}$$

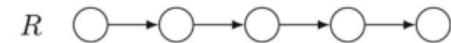
- Considera dos puntos conectados cuando son vecinos recíprocos
- Un punto  $q$  es alcanzable de manera directa (*directly-density-reachable*) por un *core-point*  $p$  si se encuentra en su vecindario,

$$|N_\epsilon(p)| \geq MinPts \text{ y } q \in N_\epsilon(p)$$

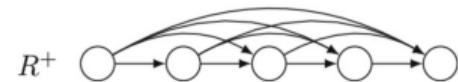
- Alcance ( $\text{density-reachable}(R)$ ) queda dado por la clausura transitiva de la relación  $\text{directly-density-reachable}(\text{DR})$

$$qRp \exists p_1, \dots, p_m \text{ con } p_1 = p \text{ y } p_m = q \text{ t.q. } p_{i+1} \text{DR} p_i$$

- Dos puntos  $p$  y  $q$  están conectados por densidad (*density-connected*) si existe otro  $r$  a partir del cual ambos son *density-reachable*
- Un **cluster** es entonces un conjunto de puntos conectados por densidad (*density-connected(C)*)
  - maximal respecto de *density-reachability*



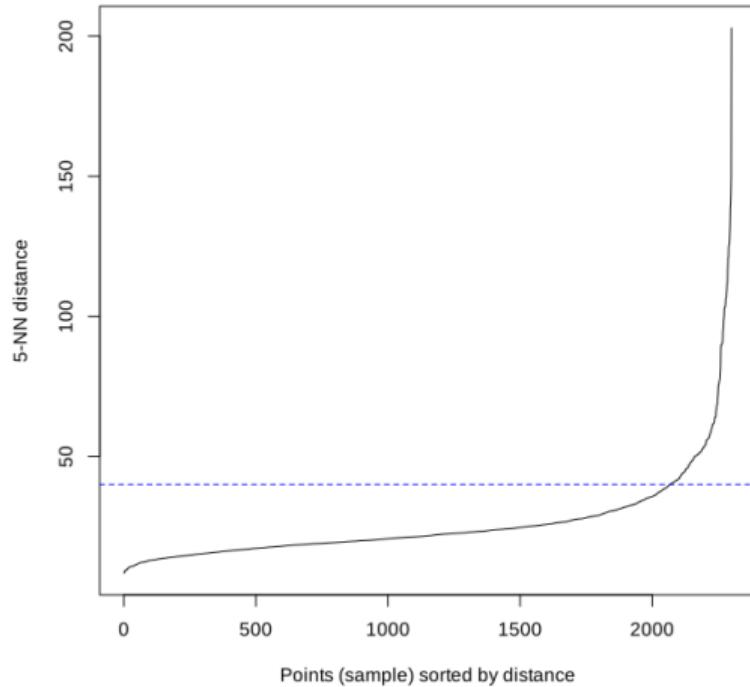
$$qCp, \exists r \text{ t.q. } rRp \wedge rRq$$



- Ruido se define como aquel conjunto de puntos que no pertenecen a ningún cluster

# Parámetros del método

- Número mínimo de puntos ( $MinPts$ ): Menor número requerido para formar un cluster
  - Fijado a un valor mayor que el número de dimensiones de los datos
- $\epsilon$  ( $eps$ ): Distancia máxima a la que pueden estar dos puntos para seguir formando parte del mismo cluster.
  - Estimado mediante un gráfico de distancia a los  $k$ -vecinos
    - Se calcula la distancia de cada punto a su  $k$ -ésimo vecino más cercano
    - Se ordenan estos valores (menor a mayor) y grafican
  - Buscar la “rodilla” en la curva (valor sobre el que las distancias empiezan a desviarse hacia los valores atípicos)



# Conclusiones

- Complejidad  $O(n \log n)$
- Incorpora identificación de ruído
- Densidad puede variar entre clusters ... problema!
- Problemas en alta dimensionalidad
- Algunas extensiones son OPTICS y HDBSCAN

# OPTICS Clustering

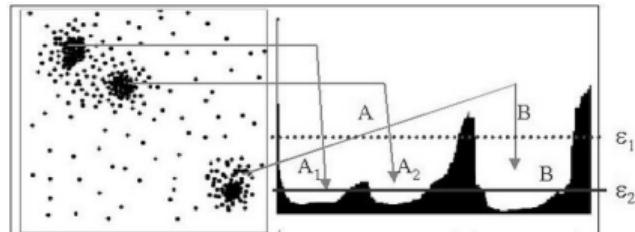
- Difícil caracterizar estructura intrínseca de grupos mediante parámetros globales de densidad
- Pueden existir grupos de diversa densidad en distintas regiones del espacio
- *Idea base:* Grupos de mayor densidad están contenidos en grupos con menor densidad
- Se construyen *simultaneamente* grupos con diferentes densidades para un valor fijo de  $MinPts$

# Conceptos relevantes

OPTICS introduce dos conceptos:

- 1 *Core-distance*: Valor  $\epsilon$  más pequeño para un punto  $p$  que lo convierte en *core-point*.
- 2 *reachability-distance*: Distancia más pequeña entre un par de puntos  $p$  y  $q$  que los hace directamente alcanzables

Además usa un gráfico (*reachability plot*) que muestra la densidad y conectividad de los puntos. **Útil** para distinguir clusters

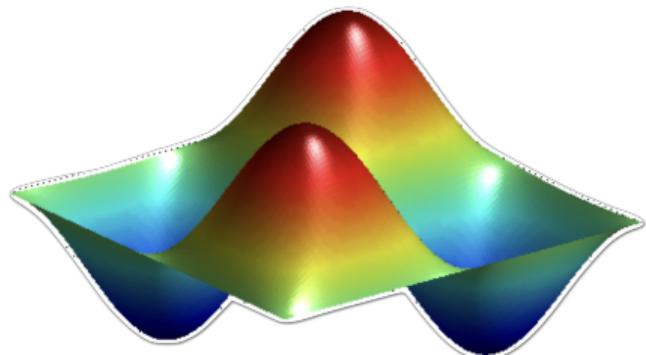


# HDBSCAN Clustering

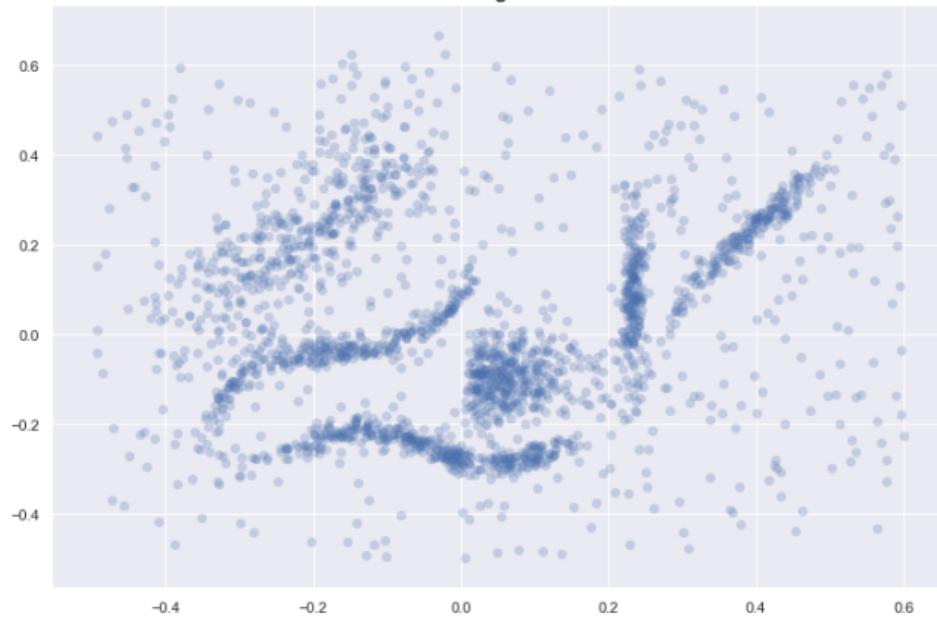
- Objetivo: Convertir DBSCAN en un método jerárquico
- Aproxima las densidades locales
- Puede ser visto como DBSCAN clustering a lo largo de todos los valores de  $\epsilon$

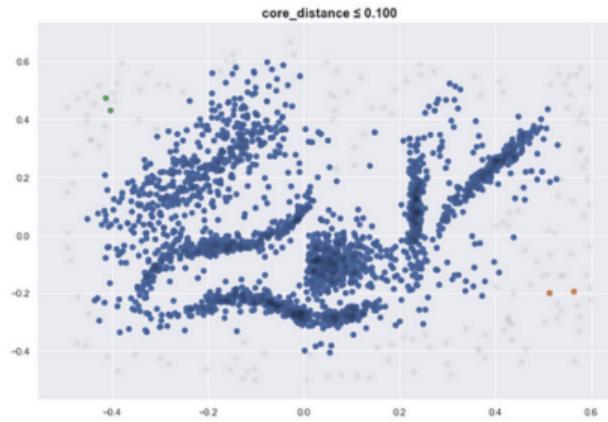
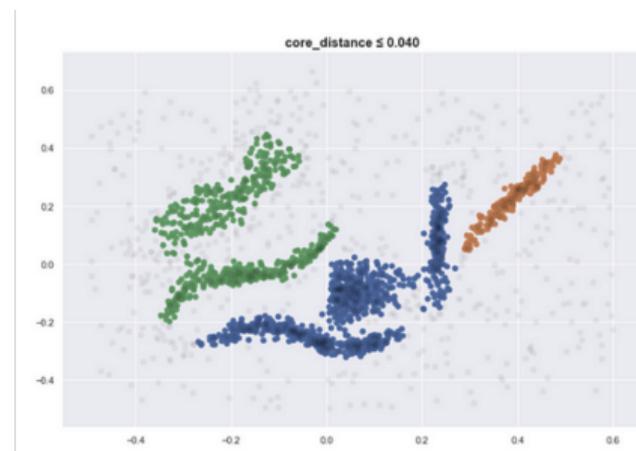
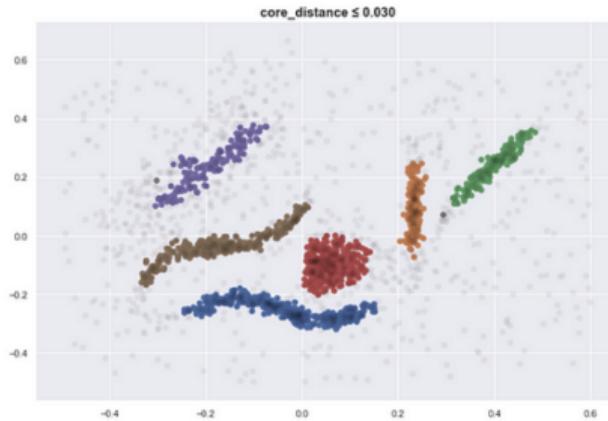
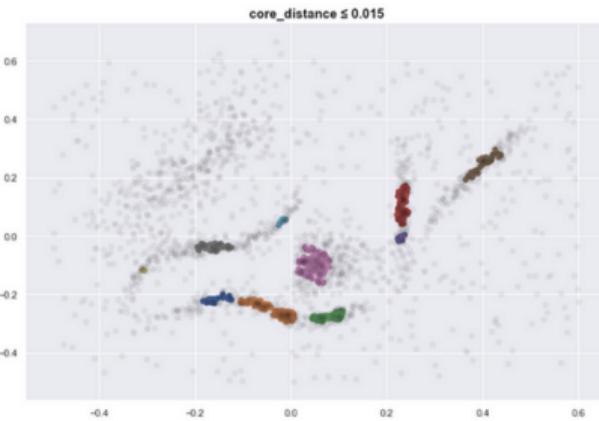
## Procedimiento

- En lugar de fijar un valor de  $\epsilon$ , se determina el número  $k$  de vecinos deseados y se encuentra el mínimo  $\epsilon$  que permitiría contener estos  $k$  vecinos
- Estas distintas distancias se denominan *core distances*
  - Puntos con *core distances* más pequeñas, se encuentran en regiones más densas
  - Puntos con *core distances* más grandes, se encuentran en regiones más dispersas porque se tuvo que ampliar el umbral de distancia para encontrar suficientes vecinos
- Encontrar regiones resulta similar a encontrar curvas de nive



Clustering Data Set





- A medida que se disminuye el umbral de *core-distances* van apareciendo grupos menos densos
  - Emergen nuevos clusters y eventualmente otros se fusionan
- Dos puntos se conectarán dependiendo de la distancia mutua de alcance

$$mrd_k(a, b) = \max\{coredist_k(a), coredist_k(b), dist(a, b)\}$$

- Esta métrica “aleja” puntos cercanos en regiones poco densas
  - Esto hace el agrupamiento más robusto al ruido
- Puntos en regiones densas no se ven afectados
- Usando esta nueva métrica se construye el MST

- MST : Subgrafo minimamente conectado
- Luego, se extrae una jerarquía de componentes conexos a partir del MST
  - Se ordenan los arcos de manera creciente
  - Se fusionan dos grupos por cada arco

