

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: Ordering Points To Identify the Clustering Structure
- 4 HDBSCAN Clustering
- 5 Identificación de tendencias
- 6 Validación de resultados de Clustering

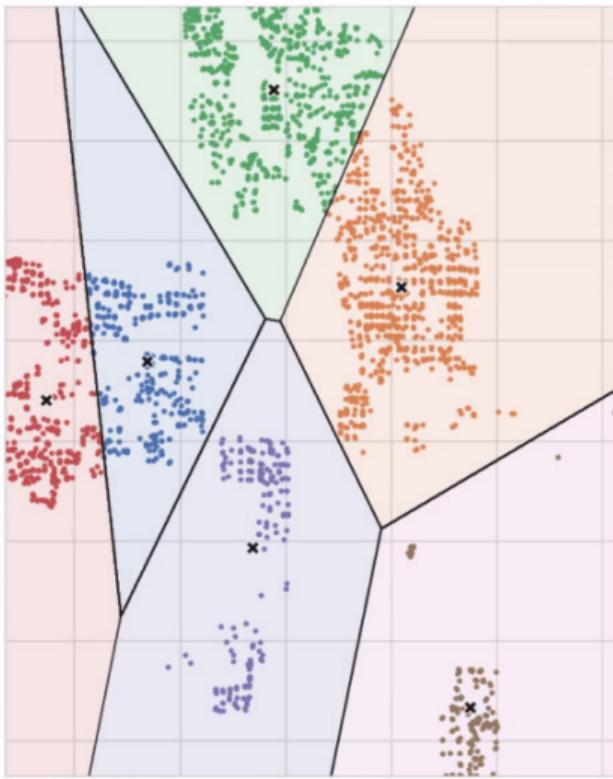
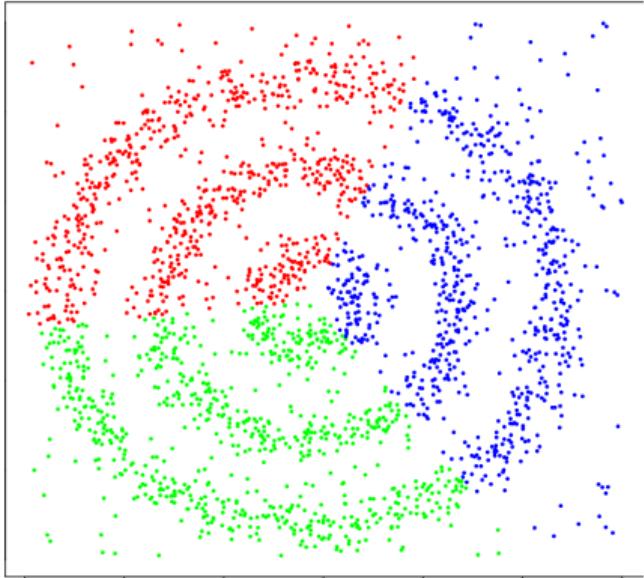
Revisitando técnicas basadas en representantes

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.

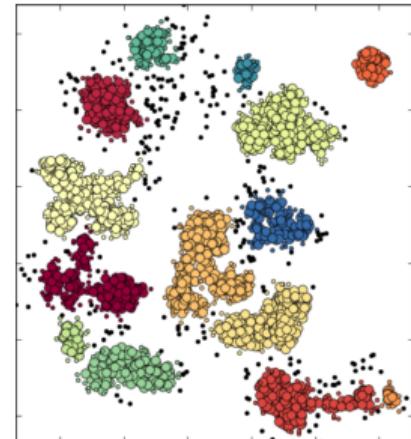
- Útil por ejemplo en datos geo-espaciales.



- 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 ruído 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 (ϵ -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 (ϵ)

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

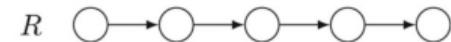
- Considera dos puntos conectados cuando son vecinos recíprocos
- Un punto q es alcanzable de manera directa (*directly-density-reachable*) por un *core-point* p si se encuentra en su vecindario,

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

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

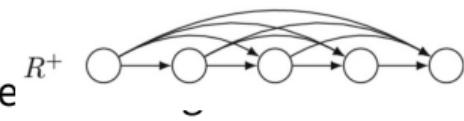
$$qRp \exists p_1, \dots, p_m \text{ con } p_1 = p \text{ y } p_m = q \text{ t.q. } p_{i+1} \text{DR} p_i$$

- 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*



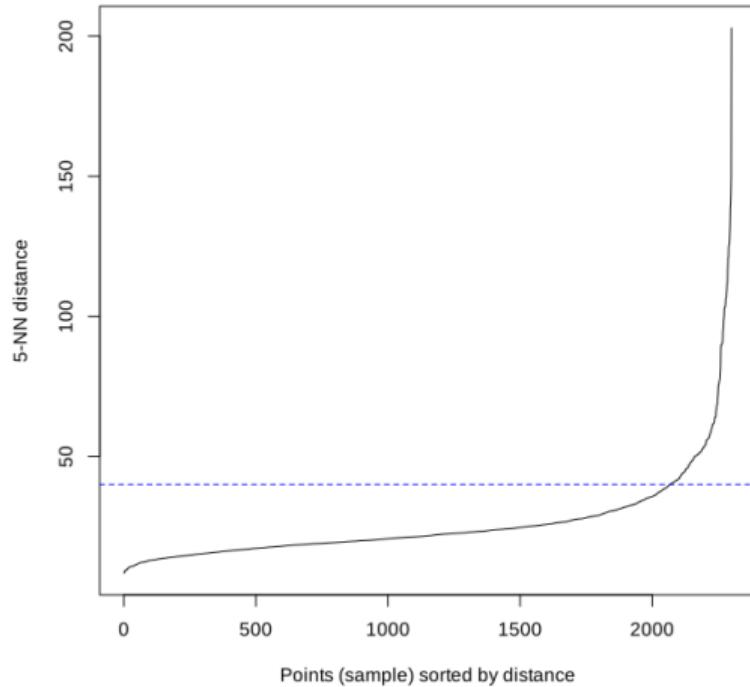
$$qCp, \exists r \text{ t.q. } rRp \wedge rRq$$

- Ruido se define como aquel conjunto de puntos que no pertenece a 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
- ϵ (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 ruido
- Densidad puede variar entre clusters ... problema!
- Problemas en alta dimensionalidad
- Algunas extensiones son OPTICS y HDBSCAN

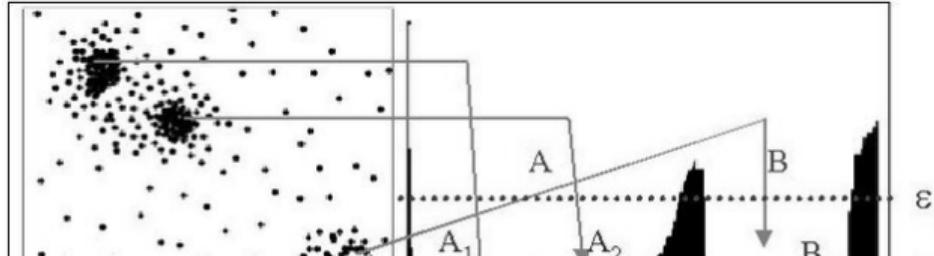
OPTICS: Ordering Points To Identify the Clustering Structure

- 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 y ventajas

- **Concepto:** OPTICS es una generalización de **DBSCAN** que aborda la dificultad de elegir el parámetro ϵ (radio). Genera un orden de los puntos basado en su **densidad** y su **distancia de alcanzabilidad** (reachability distance).
- **Salida:** Produce un “**diagrama de alcanzabilidad**” (reachability plot) que visualiza la estructura jerárquica de clústeres. Los valles en el diagrama corresponden a clústeres.
- **Ventajas:**
 - No requiere un valor global de ϵ .
 - Puede descubrir clústeres de **diferentes densidades**.
 - Identifica **ruido** de manera efectiva.

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



Parámetros y Ajuste

- **min_samples (Mínimo de muestras):**

- **Definición:** El número mínimo de puntos requeridos para formar un núcleo denso.
- **Ajuste:** Un valor más alto requiere clústeres más densos. Típicamente se elige entre 2 y el doble de la dimensionalidad de los datos. Se puede empezar con un valor bajo (ej. 5) y aumentarlo.

- **max_eps (Máximo épsilon):**

- **Definición:** El radio máximo para considerar vecinos. Limita la distancia máxima para buscar vecinos. No es el ϵ global de DBSCAN.
- **Ajuste:** Un valor demasiado pequeño puede resultar en clústeres no detectados. Un valor demasiado grande puede hacer que el cálculo sea lento y unir clústeres distintos. A menudo se establece en `inf` (infinito) para permitir a OPTICS explorar todas las posibles densidades, o un valor grande basado en la escala de tus datos.

- **metric (Métrica de distancia):**

- **Definición:** La función de distancia utilizada (ej., euclidiana, manhattan, etc.).
- **Ajuste:** Depende de la naturaleza de tus datos. La euclidiana es común para datos numéricos.

HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise

Concepto y Ventajas

- **Concepto:** Basado en DBSCAN, pero construye una **jerarquía de clústeres** a partir de un árbol de conectividad mínima (MST) de los datos, usando la “distancia de conectividad mutua inversa”.
- **Aislamiento de clústeres:** Identifica la **estabilidad de los clústeres** a través de diferentes umbrales de densidad, seleccionando los clústeres más “estables”.
- **Ventajas:**
 - **No requiere el parámetro ϵ .**
 - Puede encontrar clústeres de diferentes densidades.
 - Maneja el **ruido** de manera robusta.
 - Generalmente más eficiente que OPTICS para extraer clústeres.

Parámetros y Ajuste

- **min_cluster_size (Tamaño mínimo del clúster):**
 - **Definición:** El número mínimo de puntos para que se considere un clúster válido.
 - **Ajuste:** Un valor más pequeño detectará clústeres más pequeños y posiblemente más ruidosos. Un valor más grande solo identificará clústeres substanciales. Depende del dominio del problema; empezar con 2-10 es razonable.
- **min_samples (Mínimo de muestras / min_points):**
 - **Definición:** El número mínimo de puntos para considerar un punto como un “núcleo” de un clúster. También influye en el suavizado del árbol de conectividad.
 - **Ajuste:** Similar a min_samples en OPTICS. Un valor más alto hace que los clústeres sean más “densos” y ruidosos. A menudo se elige igual o ligeramente menor que min_cluster_size.

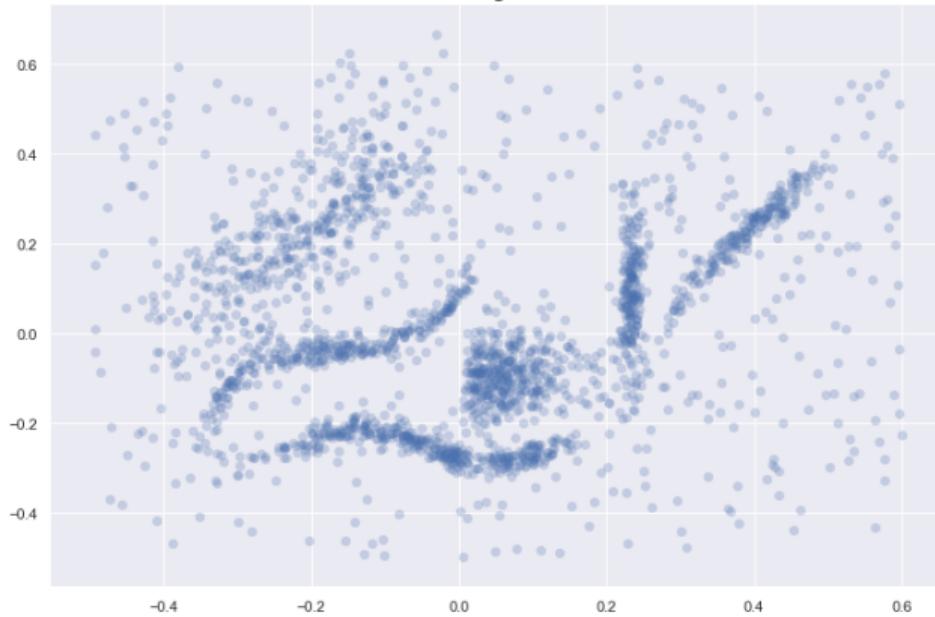
- **cluster_selection_epsilon:**

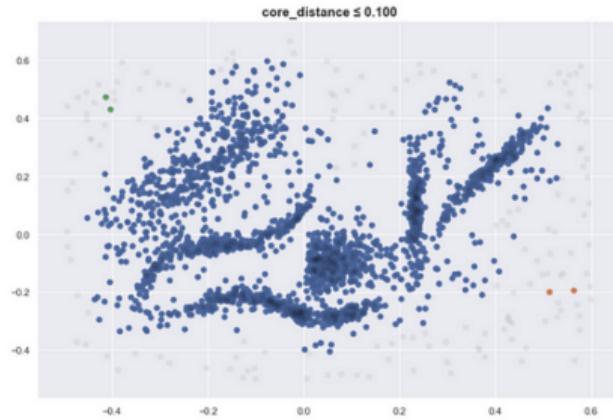
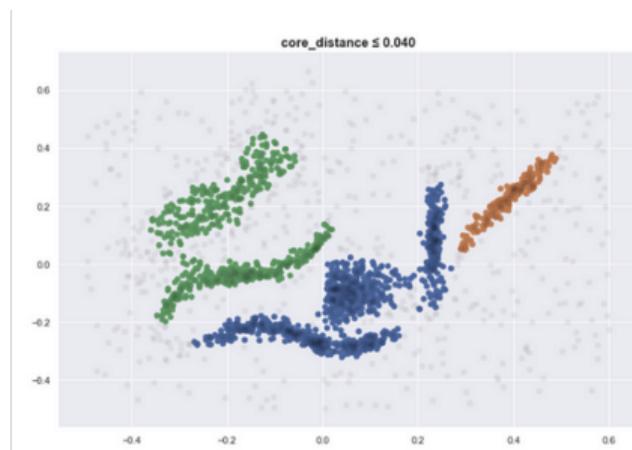
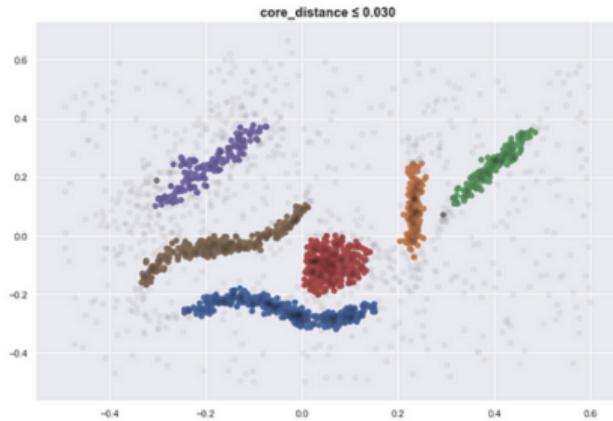
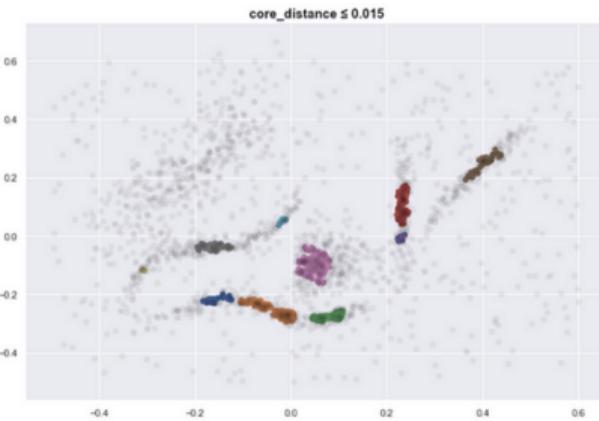
- **Definición:** Un umbral de distancia que permite que puntos adyacentes se unan a clústeres existentes, incluso si no son parte de un núcleo denso. Rara vez se usa y no se recomienda modificarlo al principio.
- **Ajuste:** Generalmente se deja en su valor por defecto (0.0). Solo se ajusta en casos muy específicos donde se necesita una relajación en la definición de densidad.

- **metric (Métrica de distancia):**

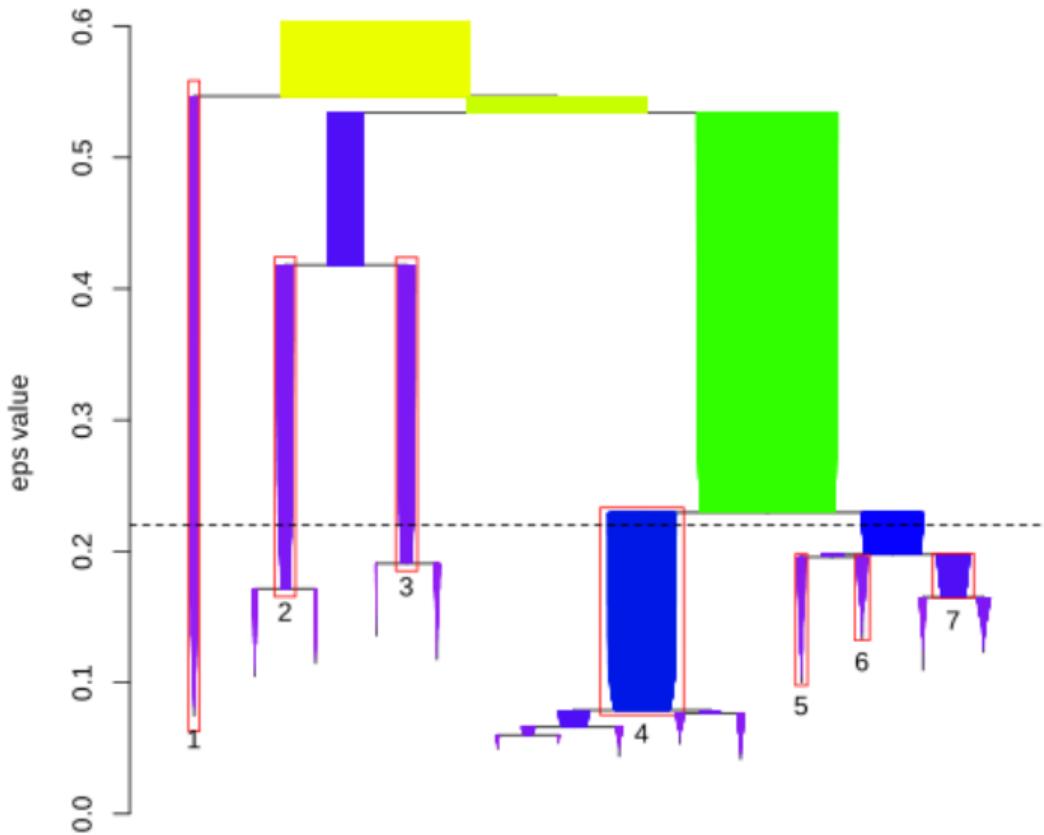
- **Definición:** La función de distancia utilizada.
- **Ajuste:** Similar a OPTICS. La euclidiana es la más común.

Clustering Data Set





HDBSCAN*



Comparación: OPTICS vs. HDBSCAN

Característica	OPTICS	HDBSCAN
Concepto Principal	Ordenamiento de puntos basado en densidad para construir un diagrama de alcanzabilidad.	Jerarquía de clústeres basada en la estabilidad de la densidad.
Salida	Diagrama de alcanzabilidad (jerarquía explícita).	Asignación de clústeres para cada punto (jerarquía implícita).
Necesidad de ϵ	Requiere <code>max_eps</code> (límite superior de búsqueda).	No requiere ϵ .
Robustez a densidades variables	Excelente.	Excelente.
Extracción de clústeres	Requiere un post-procesamiento del diagrama (ej., usando <code>cluster_optics_dbscan</code> o manualmente).	Realiza la extracción de clústeres automáticamente.
Parámetros Clave	<code>min_samples</code> , <code>max_eps</code>	<code>min_cluster_size</code> , <code>min_samples</code>
Ventajas	Flexibilidad, visualización detallada de la estructura.	Más automatizado, robusto, más rápido en la extracción de clústeres.
Desventajas	Extracción de clústeres puede ser manual/complicada.	<code>min_cluster_size</code> y <code>min_samples</code> pueden ser difíciles de ajustar al principio.
Uso Típico	Cuando se necesita una comprensión profunda de la estructura de densidad y la relación jerárquica.	Cuando se necesita una asignación directa de clústeres con ruido, con énfasis en la velocidad y robustez.

Consideraciones Finales

- **Ajuste de Parámetros Global:** Ambos métodos se benefician de un buen conocimiento del dominio de los datos. La validación interna (ej. coeficientes de silueta adaptados a ruido) o el análisis visual pueden guiar el ajuste.
- **Recomendación:** Si bien OPTICS proporciona una visión más granular, HDBSCAN es a menudo el preferido para la mayoría de las aplicaciones debido a su capacidad para extraer clústeres de forma automática y robusta sin la necesidad de definir un ϵ global.

Identificación de tendencias

- Un método de clustering intentará encontrar grupos **aún cuando no existan**

¿Existe o no una tendencia de agrupamiento en los datos?

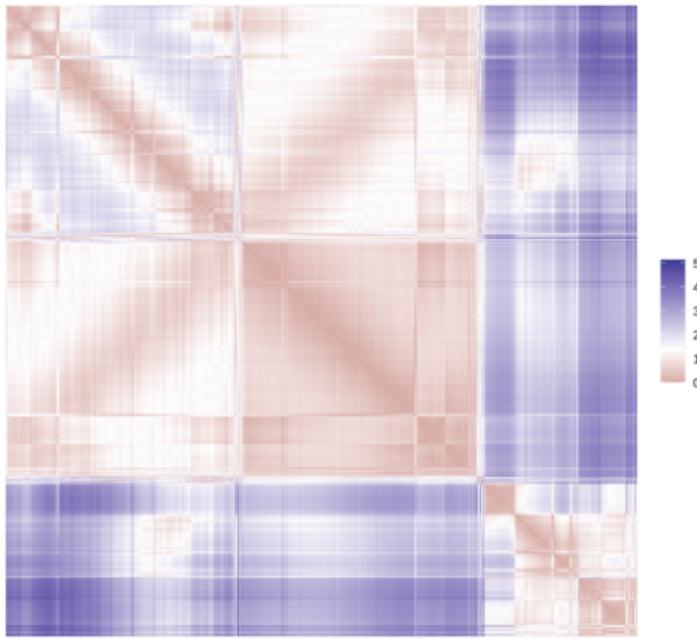
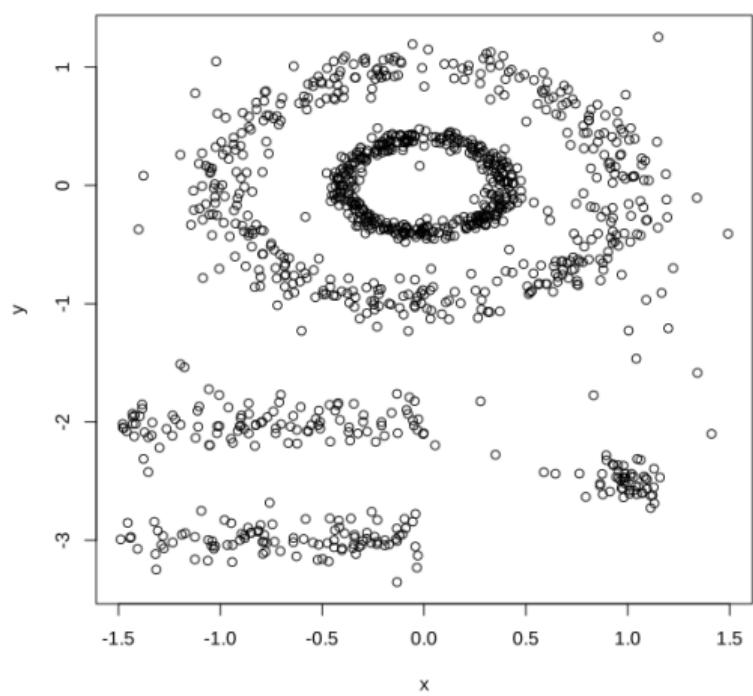
- Una manera que no realiza muchos supuestos es la evaluación visual de tendencia (**VAT e improved VAT**)

Método VAT

Este método consiste en:

1. Calcular la matriz de distancia entre los objetos del conjunto de datos
2. Re-ordenar las filas/columnas de la matriz de manera que los objetos similares queden próximos. Este proceso genera una matriz de distancias ordenada (MDO) .
3. Mostrar la MDO.

VAT permite detectar tendencia de manera visual contando los bloques cuadrados a lo largo de la diagonal



- 1 - Bezdek, J. C., & Hathaway, R. J. (2002, May). VAT: A tool for visual assessment of (cluster) tendency. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) (Vol. 3, pp. 2225–2230). IEEE.
- 2 - Wang, L., Nguyen, U. T., Bezdek, J. C., Leckie, C. A., & Ramamohanarao, K. (2010, June). iVAT and aVAT: enhanced visual analysis for cluster tendency assessment. In Pacific–Asia Conference on Knowledge Discovery and Data Mining (pp. 16–27). Berlin, Heidelberg: Springer Berlin Heidelberg.

Validación de resultados de Clustering

- El objetivo es poder medir la calidad de solución entregada por un método
- Estas medidas se categorizan en 3 grupos:
 - ① Medidas internas: No usa referencia externa, solo propiedades intrínsecas de la solución
 - ② Medidas externas: Compara solución con un patrón externo, por ejemplo etiquetas de clase. Permite medir en qué medida el agrupamiento coincide con lo esperado.

Medidas internas

- A menudo reflejan cohesión/compactitud, conectividad y separación entre particiones
- **Cohesión:** Que tan cercanos están los objetos dentro de cada grupo.
 - Los distintos indices se basan en medidas de distancia
 - **Variación intra cluster** es un ejemplo de indicador de cohesión
- **Separación:** Que tan bien separado se encuentra un cluster de los otros.
 - Distancias entre centroides o representantes de cada grupo
 - Distancias mínimas entre pares de objetos en ambos clusters
- **Conectividad:** Grado de conectividad de los clusters basado en k-vecinos más cercanos.
 - Principio: Items en el vecindario debieran compartir el mismo cluster.



Coeficiente de Silhouette

- Compara las distancias internas entre los objetos en un cluster con las distancias al cluster más cercano
- Para el objeto i , la distancia promedio mas corta a objetos del cluster más cercano b_i

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- Valores de S_i cercanos a 1 indican una asignación correcta. Valores cercanos a 0 indica que la observación se encuentra entre 2 grupos. Valores negativos indican una asignación incorrecta.

Indice Dunn

- Contrast la separación entre los grupos con la dispersión interna de cada uno.
- Calcula
 - Distancia más pequeña entre objetos de clusters distintos (α)
 - Distancia mayor entre objetos del mismo cluster (β)

$$D = \frac{\alpha}{\beta}$$

- A mayor valor, mayor separación entre los grupos relativa a sus diámetros

Medidas externas

- En general operan sobre la matriz de contingencia entre clusters y clases indicadas en el conjunto de datos
- Purity (P): Cada cluster se etiqueta según la clase más frecuente. Se suma la cantidad de objetos correctamente etiquetados y se divide por el total de objetos.
 - Ojo: Aumenta a medida que aumenta la cantidad de clusters

$$P = \frac{1}{n} \sum_k \max_j |W_k \cap C_j|$$

- Rand Index (RI): Considera los pares de objetos que son asignados dentro del mismo o en distinto cluster.
 - V_p : Objetos similares que son asignados al mismo cluster
 - V_n : Objetos disimiles asignados en clusters distintos
 - F_p : Objetos disimiles asignados al mismo cluster
 - F_n : Objetos similares que son asignados en clusters distintos

$$RI = \frac{V_p + V_n}{V_p + V_n + F_p + F_n}$$