Computación Científica

Juan Sebastián Valencia Villa
juan.valencia72@eia.edu.co

¿Qué es un modelo?

Es una función matemática que mapea un conjunto de características o entradas en una salida. La salida puede ser un vector, una variable categórica o una variable numérica.

$$y \sim f(x_n)$$

$$y = \alpha x_1 + \beta x_2 + \ \ldots + \ \theta x_n$$

Parámetros

Valores variables que modifican las características para generar una salida.

PROBLEMA:

Identificar los parámetros

Problema de Agrupación

Objetivo: Encontrar patrones dentro de un conjunto de observaciones. Las observaciones dentro de un grupo tienen similitud y suficiente diferencia entre las de otros grupos.

Requiere implementar medidas de distancia para validar la similitud

Agrupamiento por particiones (Partitioning Clustering): Se requiere que el usuario identifique la cantidad de grupos previamente.	K – Means K – Medoids K – Modes
Agrupamiento Jerárquico (Hierarchichal Clustering): No requiere que el usuario de forma anticipada indique el número de grupos.	Divisive Agglomerative
Métodos por densidad: Modifican o combinan metodologías previas.	Fuzzy C – Means DBSCAN GMM

Medidas de Distancia

Distancia Euclidiana
Longitud del segmento entre dos puntos

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

Distancia de Manhattan
Diferencia absoluta entre cada dimensión

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

Correlación
Mantenimiento de los mismos patrones de comportamiento

$$d(p,q) = 1 - corr(p,q)$$

Distancia Coseno
Mide la similitud por orientaciones

$$cos(\alpha) = \frac{x.y}{\|x\|\|y\|}$$

Coeficiente de ajuste simple
Se usa cuando las observaciones son binarias

$$SMC = \frac{número\ de\ coincidencias}{número\ total\ de\ atributos}$$

Cantidad de grupos

- Método Elbow (codo): Se encuentra el número de parámetros que minimice la métrica de desempeño.
- Método average silhouette: Se encuentra el número de parámetros cuando se maximiza el coeficiente de silhouette

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

$$a_i = promedio\ de\ las\ distancias\ entre\ la\ observación\ i\ y\ las\ demás\ observaciones\ del\ clúster$$

$$b_i = la\ menor\ de\ las\ distnacias\ promedio\ entre\ la\ observación\ i\ y\ el\ resto\ de\ clústers$$

K - means

Agrupa las observaciones en un número predefinido de K grupos de forma que la suma de las varianzas internas de los grupos sea la menor posible.

Algoritmo:

1. Especifique el número K de grupos que se quieren crear.
2. Seleccionar de forma aleatoria k observaciones del conjunto de datos como centroides iniciales.
3. Asignar cada una de las observaciones al centroide más cercado.
4. Para cada un de los grupos generados en el paso 3, recalcular su centroide.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance la cantidad de iteraciones definidas.

K - medoids

Método más robusto que k – means, en este caso el centro del clúster lo define una observación del grupo (medoid).

Algoritmo:

1. Seleccionar k observaciones aleatorias como medoids iniciales.
2. Calcular la matriz de distancia entre todas las observaciones si esta no se ha calculado anteriormente.
3. Asignar cada observación a su medoid más cercano.
4. Para cada uno de los grupos creados, comprobar si la distancia media del grupo mejora con otro medoid.
5. Si al menos uno de los medoid ha cambiado en el paso 4, volver al paso 3, de lo contrario finalizar el proceso.

Agrupamiento jerárquico aglomerativo

El método inicia con todas las observaciones separadas, cada una crea un grupo individual. Los grupos se van combinando hasta que quede en uno solo.

Algoritmo:

1. Considerar cada una de las observaciones como un clúster individual.
2. Calcular la distancia entre cada posible par de clústers. Los más similares se fusionan (se elige la métrica de distancia).
3. Cortar la estructura del dendograma a una altura determinada.

El proceso es iterativo hasta que quede en un solo grupo.

Agrupamiento jerárquico divisivo

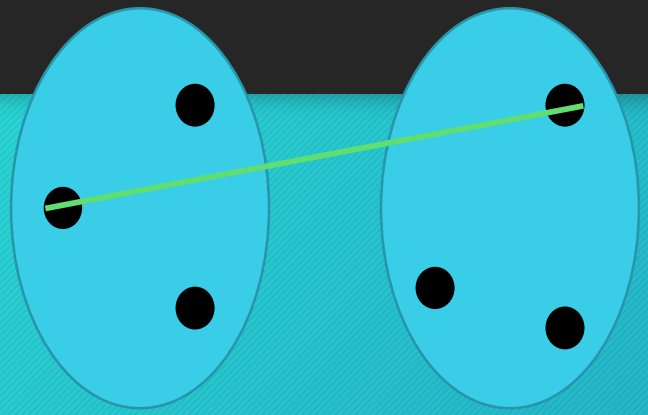
Este método inicia como un único clúster que contiene todas las observaciones que se va dividiendo en cada iteración.

Algoritmo

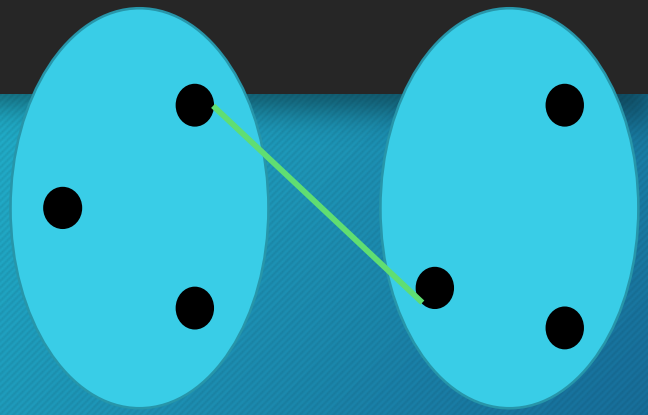
1. Todas las observaciones forman un único clúster.
2. Calcular para cada clúster la mayor de las distancias entre pares de observaciones.
3. Seleccionar el clúster con mayor diámetro.
4. Calcular la distancia media de cada observación con respecto a las demás.
5. La observación más distante forma un nuevo clúster.
6. Se reasignan las observaciones restantes al nuevo clúster.

Se repite hasta que se tengan n clústers.

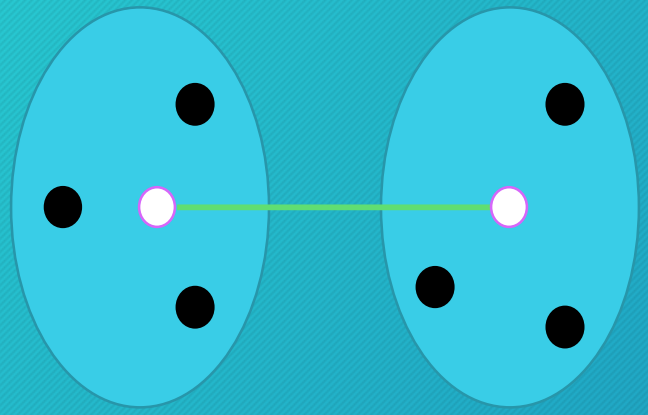
Mecanismos de asociación o enlace



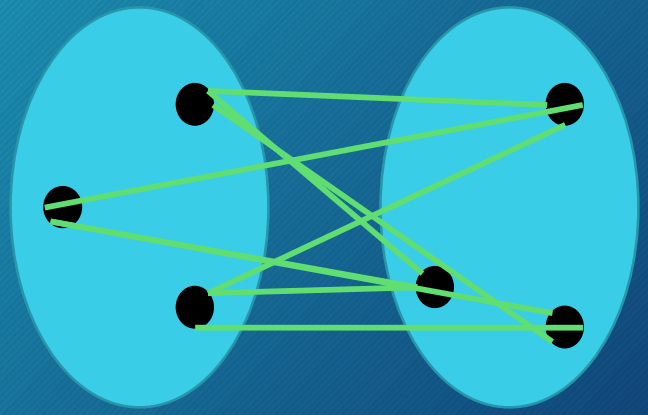
Completa



Simple



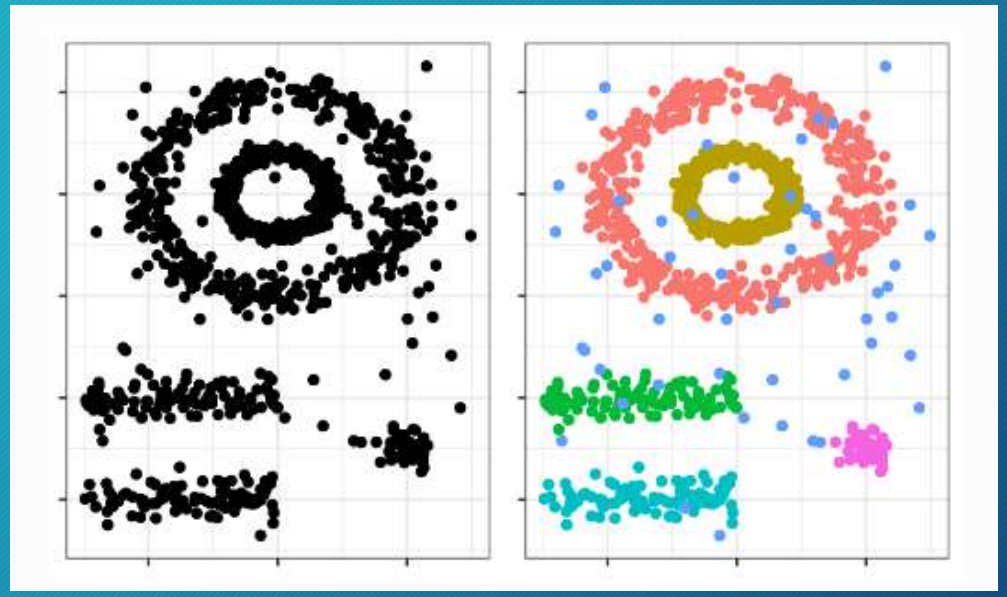
Centroide



Promedio

Método de búsqueda - DBSCAN

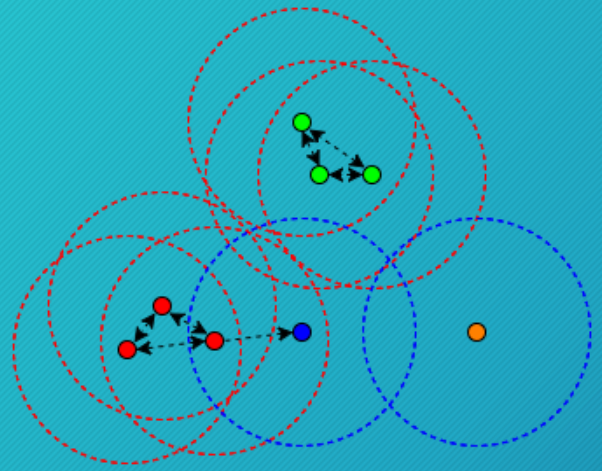
Density Based Spatial Clustering of Applications with Noise, fue presentado como un método intuitivo siguiendo el modo como los seres humanos identifica regiones de altas densidades y las diferencia de las regiones de baja densidad.



Método de búsqueda - DBSCAN

Elementos a tener en cuenta:

- Puntos de núcleo: Son individuos que satisfacen la cantidad mínima de vecinos en el radio determinado.
- Puntos de frontera o alcanzables: Son individuos que no son puntos de núcleo pero que se encuentran dentro del radio definido.
- Puntos de ruido: individuos que no son ni de núcleo ni de frontera.



Algoritmo:

1. Se establece el valor de épsilon (radio) y muestras mínimas.
2. El algoritmo da inicio en un punto aleatorio.
3. En ese punto aleatorio se utilizan el radio y las muestras mínimas y se establece si es un punto de núcleo, de frontera o ruido.
4. Una vez determinado, se pasa a un siguiente punto.
5. Se determinan entonces todos los puntos si se asocian a un punto de núcleo o si quedan como ruido.

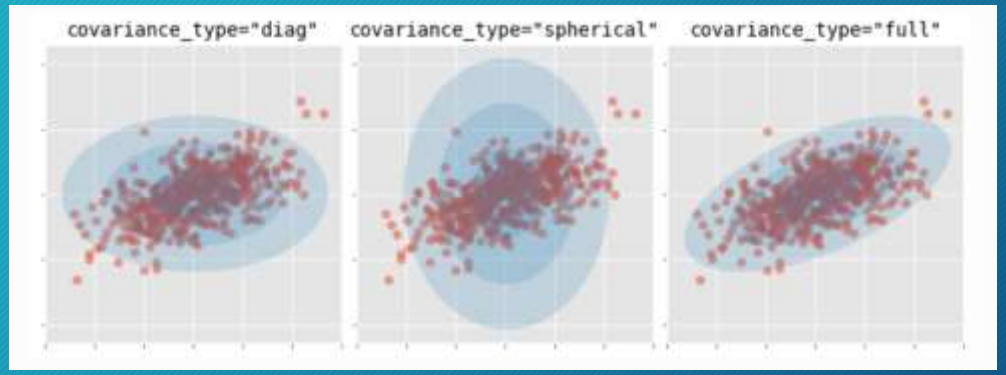
Parámetros:

- Épsilon : máxima distancia entre dos grupos.
- Mínimas muestras: número mínimo de individuos alrededor para determinar la creación de un grupo.
- Métrica: medida de la distancia.
- Algoritmo: mecanismo de cálculo de la distancia.

Gaussian Mixture Models

Modelo probabilístico el que se considera que las observaciones siguen una distribución de probabilidad formada por la combinación de múltiples distribuciones normales.

Se entiende como una generalización de Kmeans que devuelve la probabilidad de pertenencia de un individuo a un grupo.



Utiliza el algoritmo de **Expectation - Maximization** para determinar la función de probabilidad del grupo mediante un valor medio y una matriz de covarianzas.

Gaussian Mixture Models

Tipos de matrices de covarianza:

TIED	DIAGONAL	SPHERICAL	FULL
• Todos los grupos comparten la misma matriz de covarianza.	• Las covarianzas de cada grupo son distintas en todas las dimensiones. • Las elipses generadas quedan alineadas con los ejes.	• Las covarianzas de cada grupo son las mismas en todas las dimensiones. • Genera grupos de diferente tamaño, pero todos esféricos.	• Todos los grupos pueden modelarse mediante diferentes matrices de covarianza.