ALGORITMOS DE AGRUPAMIENTO

CLUSTERING

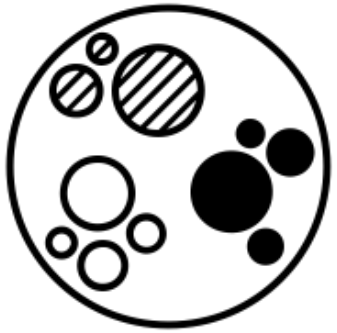
Diferencias entre agrupamiento y clasificación

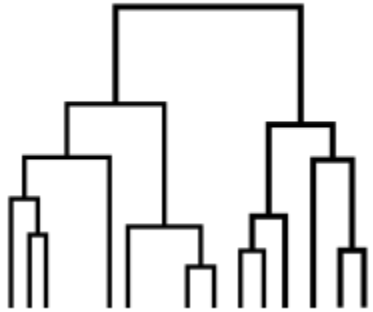
En términos generales, la **clasificación** es un modelo de aprendizaje supervisado donde cada instancia de datos de entrenamiento pertenece a una clase en particular.

Sin embargo, en el **clustering**, los datos no están etiquetados y el proceso no está supervisado. Podemos agrupar clientes similares, y asignarlos a un clúster, en función de si comparten atributos similares, como edad, educación, etc.

¿Qué es?

El **clustering** significa encontrar agrupaciones en un conjunto de datos, sin supervisión.

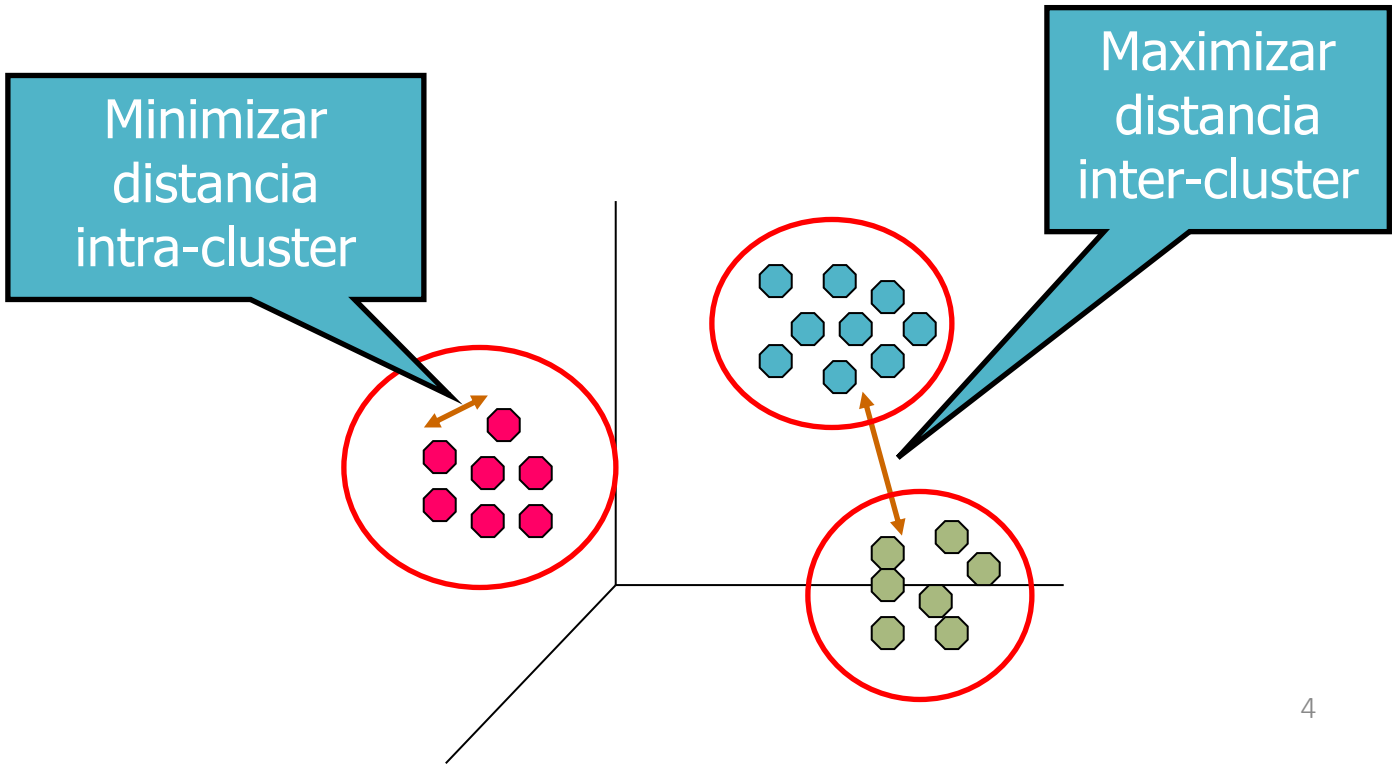




Un **clúster** es un grupo de puntos de datos u objetos en un conjunto de datos que son similares a otros objetos en el grupo y diferentes a los puntos de datos en otros grupos.

OBJETIVO

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos:



Ejemplos Prácticos

En la **industria minorista**, el clustering se utiliza para encontrar asociaciones entre clientes sobre sus características demográficas y usar esa información para identificar patrones de compra de varios grupos de clientes.

En **grupos bancarios**, los analistas encuentran grupos de transacciones normales para encontrar los patrones fraudulentos de uso de tarjeta de crédito. Además, usan clustering para identificar clusters de clientes, por ejemplo, para encontrar clientes leales, frente a clientes abandonados.

En la **industria de seguros**, el clustering se utiliza para la detección de fraudes en el análisis de reclamos, o evaluar el riesgo de seguro de ciertos clientes en función de sus segmentos.

En **Publication Media**, el clustering se utiliza para clasificar automáticamente las noticias en función de su contenido, o etiquetar noticias, luego agruparlas, para recomendar artículos de noticias similares a los lectores.

En **medicina**: se puede utilizar para caracterizar el comportamiento del paciente, en función de sus características similares, para identificar terapias médicas exitosas para diferentes enfermedades.

En **biología**: el clustering se usa para agrupar genes con patrones de expresión similares, o agrupar marcadores genéticos para identificar los lazos familiares.

Algoritmos

La **agrupación basada en particiones** es un grupo de algoritmos de agrupación que produce esferas de grupos, como k-Means, k-Median o Fuzzy c-Means. Estos algoritmos son relativamente eficientes y se utilizan para bases de datos de tamaño medio y grande.

Los algoritmos de **agrupación jerárquica** producen árboles de agrupaciones. Son intuitivos y generalmente son buenos para usar con conjuntos de datos de pequeño tamaño.

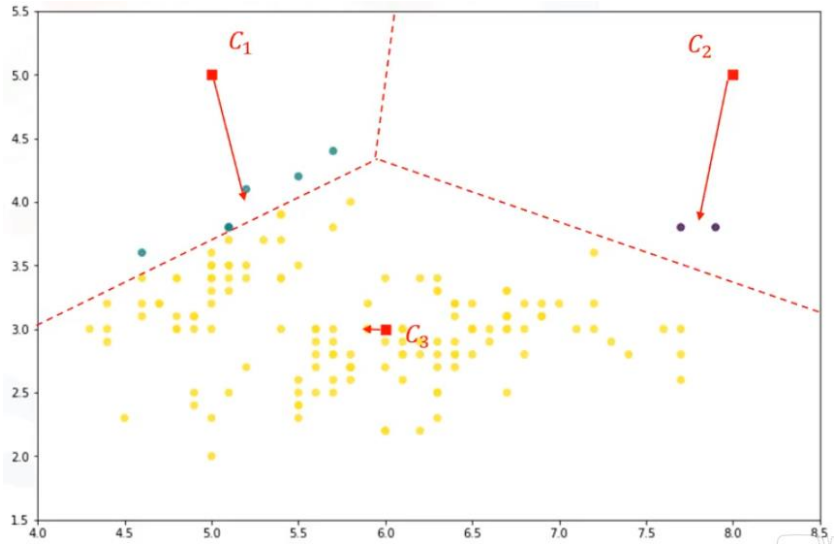
Los algoritmos de **agrupamiento basados en la densidad** producen grupos de formas arbitrarias. Son especialmente buenos cuando se trata de grupos espaciales o cuando hay ruido en su conjunto de datos, por ejemplo, el algoritmo DBSCAN.

Los resultados obtenidos dependerán del algoritmo de agrupamiento seleccionado, el conjunto de datos disponible y la medida de similitud utilizada para comparar objetos.

Cluster Particiones (K-means)

Descripción

Una agrupación basada en particiones, es decir, divide los datos en subconjuntos sin ninguna estructura interna.



Objetivo

Formar particiones de tal manera que muestras similares vayan a una partición y muestras disimilares estén en diferentes particiones

Procedimiento

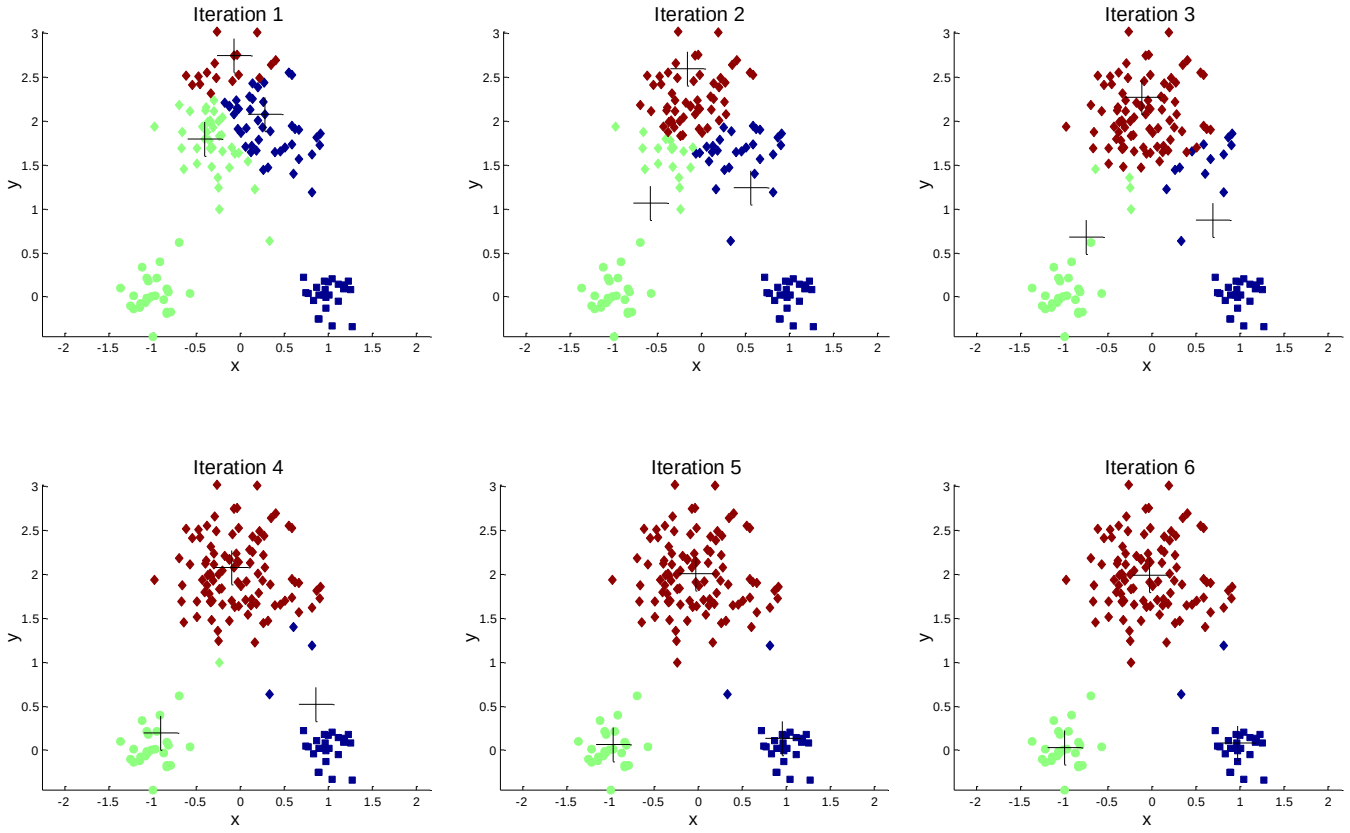
1. Número de clusters conocido (**k**)

2. Cada cluster tiene asociado un centroide (centro geométrico del cluster).

3. Los puntos se asignan al cluster cuyo centroide esté más cerca (utilizando cualquier métrica de distancia).

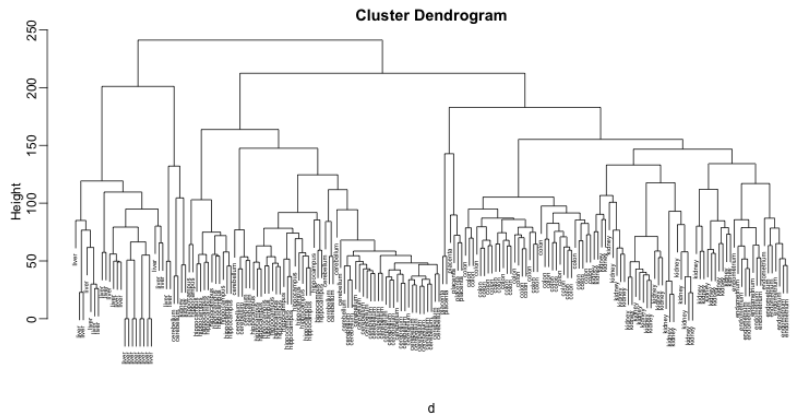
4. Iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a clusters, hasta que los centroides dejen de cambiar.

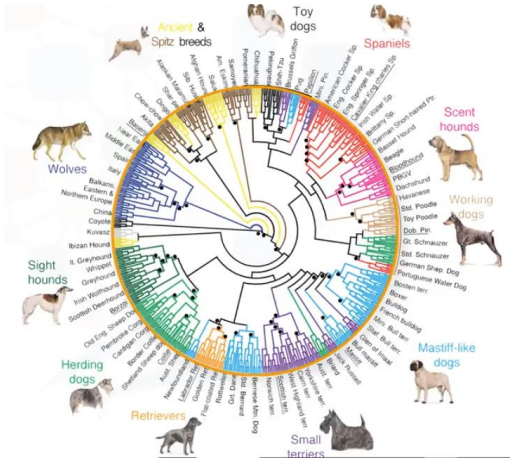
Iteraciones



Cluster Jerárquico (Dendograma)

Los algoritmos de clústeres jerárquicos construyen una jerarquía de clústeres en el que cada nodo es un clúster que consta de los clústeres de sus nodos hijas.





Las estrategias para la agrupación jerárquica en general se dividen en dos tipos: Divisivo y Aglomerativo.

Estrategias de agrupación

Descendente, por lo que se inicia con todas las observaciones en un clúster grande y se dividen en partes más pequeñas. Imagina divisivo como "dividir" el clúster.

Aglomerativo es lo contrario de divisivo, por lo que es de abajo hacia arriba, donde cada observación se inicia en su propio clúster y los pares de clústeres se fusionan a medida que avanzan en la jerarquía. La aglomeración significa manipular o recoger cosas, que es exactamente lo que esto hace con el clúster.

Procedimiento

1. Crear n clusteres, uno para cada punto de datos.

2. Cada punto se asigna como un cluster.

3. Cacular la matriz de distancia/proximidad, que será n por n en la primera tabla.

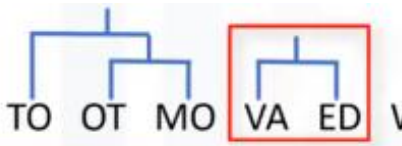
4. Iterativamente ejecutamos la “unión” de los dos grupos más cercanos y actualizamos la matriz de proximidad con los nuevos valores.

5. Se termina cuando se alcanza el número especificado de clusteres, o sólo hay un cluster que queda, con el resultado almacenado en un dendrograma.

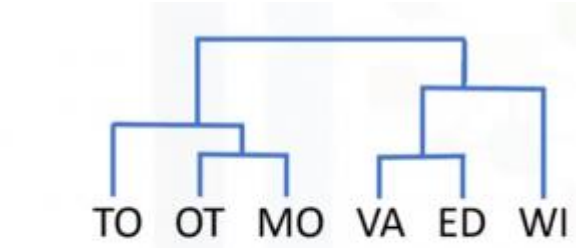
Iteraciones

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						

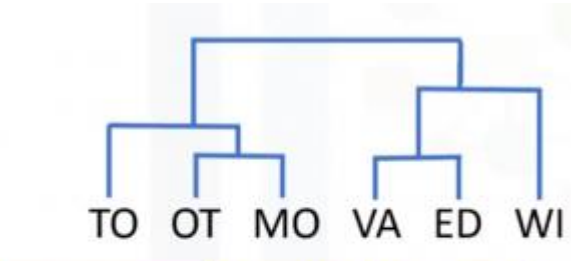
	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					



	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		

Comparación de cluster jerárquico con k-means

Ventajas

1. No es necesario especificar el número de clusteres necesarios para el algoritmo.

2. El clustering jerárquico es fácil de implementar.

3. El dendrograma producido es muy útil para comprender los datos.

Desventajas

1. El algoritmo nunca puede deshacer ninguna de los pasos anteriores. Si, por ejemplo, el algoritmo agrupa 2 puntos, y más tarde vemos que la conexión no era buena, el programa no puede deshacer ese paso.

2. La complejidad del tiempo para el clustering puede dar lugar a tiempos de cálculo muy largos, en comparación con algoritmos eficientes, como K-Means.

3. Si tenemos un conjunto de datos grande, puede ser difícil determinar el número correcto de clusteres por el dendrograma.

Cluster jerárquico vs con k-means

Los k-means son más eficientes para los conjuntos de datos grandes.

El clustering jerárquico no requiere que se especifique el número de clusteres.

El clustering jerárquico da más de un particionamiento en función de la resolución, mientras que k-Means sólo da un solo particionamiento de los datos.

El clustering jerárquico siempre genera los mismos clusteres, en contraste con la k-Means que devuelve distintos clusteres cada vez que se ejecuta debido a la inicialización aleatoria de los centroides.

Cluster basado en densidad (DBSCAN)

Localiza regiones de alta **densidad** que están separadas una de la otra por regiones de baja densidad.

Densidad, en este contexto, se define como el número de puntos dentro de un radio especificado.

Un tipo específico y muy popular de clustering basado en la densidad es DBSCAN. Por sus siglas en inglés Density-based Spatial Clustering of Applications with Noise" (Clustering espacial basado en densidad de aplicaciones con ruido).

Procedimiento

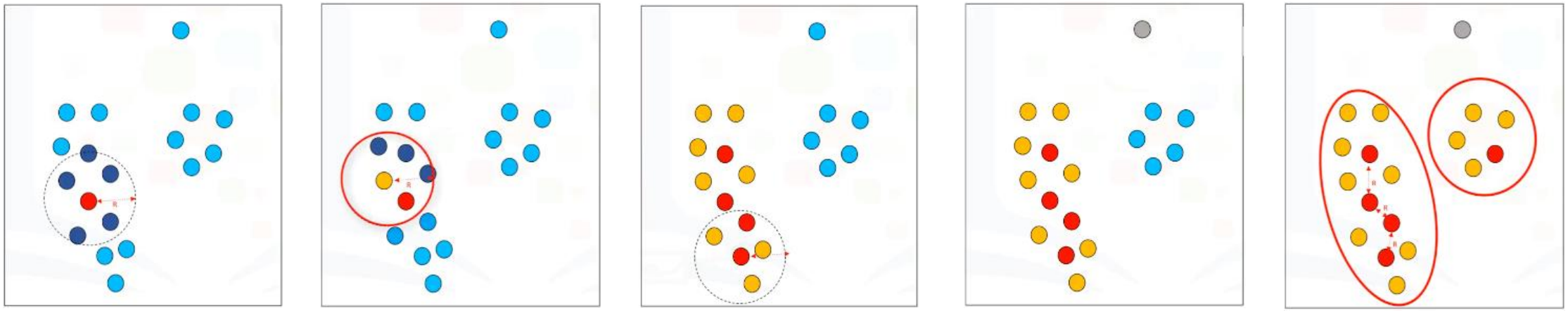
DBSCAN trabaja en la idea de que si un punto en particular pertenece a un cluster, debería estar cerca de un montón de otros puntos en ese cluster.

Funciona en base a 2 parámetros: Radio y Puntos Mínimos.

1. El radio, R, determina una longitud específica que, si incluye suficientes puntos dentro de él, lo llamamos de “Área densa”.
2. Los Puntos Mínimos, M, determina la cantidad mínima de datos que queremos alrededor de otra observación para definir un cluster.

Cada punto de nuestro conjunto de datos puede ser un punto central (core), fronterizo (border) o atípico (outlier). Toda la idea detrás del algoritmo DBSCAN es visitar cada punto, y encontrar su tipo.

Iteraciones



Método Elbow

El método Elbow, también conocido como método del codo, sigue una estrategia comúnmente empleada para encontrar el valor óptimo de un hiperparámetro.

La idea es probar un rango de valores del hiperparámetro en cuestión, representar gráficamente los resultados obtenidos con cada uno, e identificar aquel punto de la curva (codo) a partir del cual la mejora deja de ser notable.

En los casos de agrupamiento, como por ejemplo K-means, las observaciones se agrupan de una forma tal que se minimiza la varianza total intra-cluster. El método Elbow calcula la varianza total intra-cluster en función del número de clusters y escoge como óptimo aquel valor a partir del cual añadir más clusters apenas consigue mejoría.

ndice de Silueta

El método de índice de silueta se considera como número óptimo de clusters aquel que maximiza la media del coeficiente silueta de todas las observaciones.

El **coeficiente silueta** (s_i) cuantifica cómo de buena es la asignación que se ha hecho de una observación comparando su similitud con el resto de observaciones de su cluster frente a las de los otros clusters. Su valor puede estar entre -1 y 1, siendo valores próximos a 1 un indicativo de que la observación se ha asignado al cluster correcto.

Procedimiento

•Calcular el promedio de las distancias (llámese (a_i)) entre la observación i y el resto de observaciones que pertenecen al mismo *cluster*. Cuanto menor sea a_i, mejor ha sido la asignación de i a su *cluster*.

•Calcular la distancia promedio entre la observación i y el resto de *clusters*. Entendiendo por distancia promedio entre i y un determinado *cluster* C como la media de las distancias entre ii y las observaciones del *cluster* C.

•Identificar como b_i a la menor de las distancias promedio entre i y el resto de *clusters*, es decir, la distancia al *cluster* más próximo (*neighbouring cluster*).

•Calcular el valor de *silueta* como:

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