

Aprendizaje automático II

Máster Universitario en Informática Industrial y Robótica

Tema 2

Aprendizaje no supervisado: agrupamiento

Óscar Fontenla Romero

Escuela Politécnica de Ingeniería de Ferrol

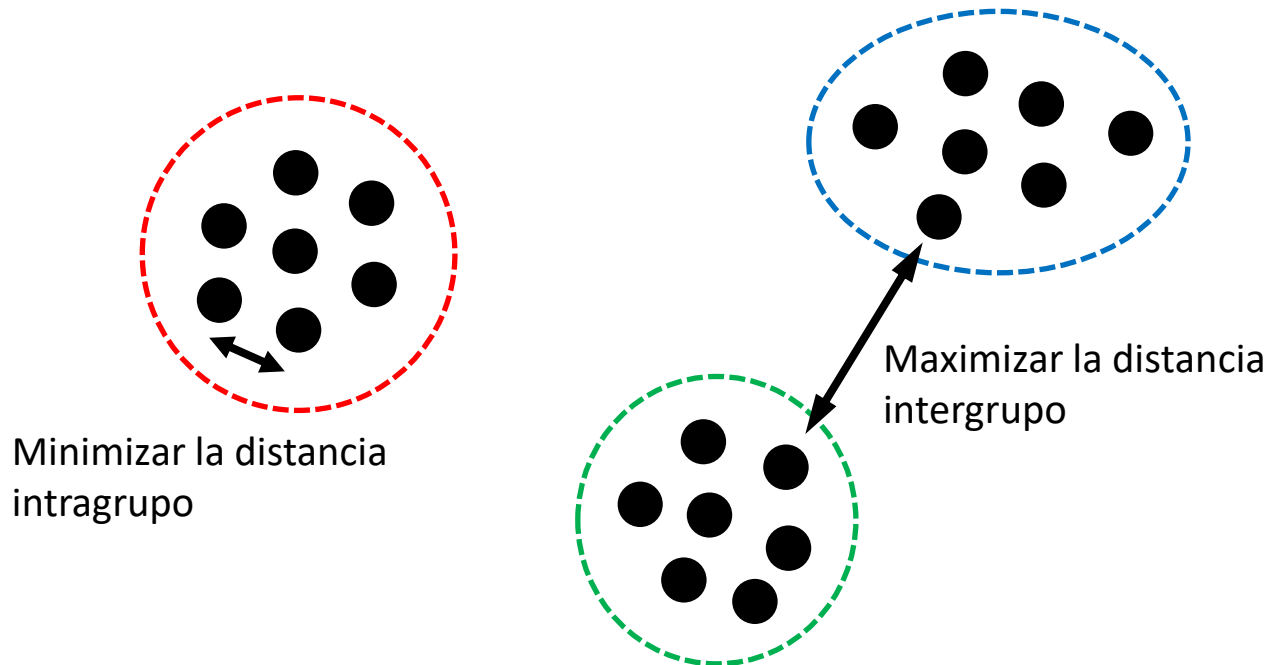
<http://www.udc.es/epef>

- Introducción al aprendizaje no supervisado: análisis clúster
- Medidas de similitud
- Tipos de métodos de agrupamiento (clustering):
 - Jerárquico
 - Por particiones: k -medias
 - Por densidad: DBSCAN

- Aprendizaje no supervisado:
 - El algoritmo lleva a cabo un proceso iterativo de aprendizaje **no guiado**
 - A diferencia del supervisado, los datos de entrenamiento no están clasificados: no existen etiquetas de clase
 - Existe una gran variedad de problemas del mundo real que no disponen de datos etiquetados por lo que habrá que usar técnicas no supervisadas
 - Dentro de este tipo de algoritmos, el agrupamiento (**clustering**), es el más utilizado, ya que particiona los datos en grupos que posean características similares entre sí

Análisis clúster

- Conjunto de técnicas empleadas para clasificar un conjunto de datos (objetos) en grupos homogéneos:
 - Datos similares en el mismo grupo (clúster)
 - Datos distintos entre grupos diferentes



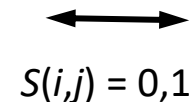
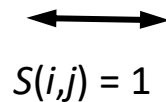
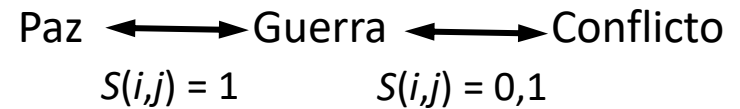
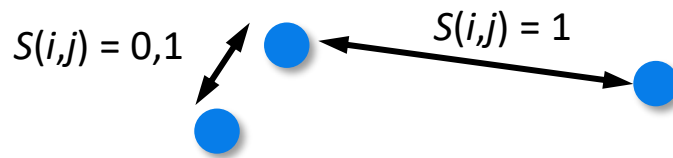
- Diferencia entre el análisis clúster y el discriminante:
 - Aprendizaje no supervisado
 - Los grupos son desconocidos a priori y es lo que se quiere determinar
- Usos principales:
 - Herramienta de análisis de datos
 - Etapa de preprocesado para otros algoritmos

■ Aplicaciones:

- Marketing y ventas: segmentación de clientes
- Filtros de Spam
- Categorización de documentos por temas similares
- Búsqueda de usuarios similares en redes sociales
- Detección de fraudes y gestión de riesgos
- Identificación de *fake news*
- Agrupamiento de genes con funciones similares
- Segmentación de imágenes, etc.

Análisis clúster: medida de similitud

- Para conocer si dos datos (objetos) son similares entre sí será necesario adoptar una **medida de similitud**



Análisis clúster: medida de similitud

- El resultado del agrupamiento (clústeres) depende de la medida de similitud empleada entre los datos considerados
- Todo depende de lo que se considere como **similar**
- Ejemplo:
 - Una baraja de cartas españolas se podría dividir de distintos modos: palos, figuras o números, etc.

Análisis clúster: medida de similitud

- Medida de similitud: se suelen emplear medidas de distancia porque verifican las siguientes propiedades:

1. No negatividad: $d(i,j) \geq 0$ para todo i y j

2. Propiedad reflexiva: $d(i,j) = 0$ si y sólo si $i = j$

3. Propiedad simétrica: $d(i,j) = d(j,i)$ 

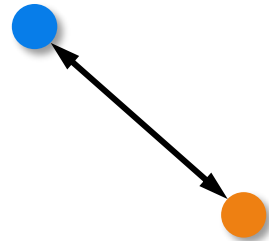
4. Desigualdad triangular: $d(i,j) \leq d(i,k) + d(k,j)$

Análisis clúster: métricas de distancia

■ Distancia Euclídea:

- Es la más distancia más empleada
- Definición:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$$



- Inconveniente:
 - No es invariante a la escala: las distancias calculadas pueden estar sesgadas según las unidades de las variables
 - Por ello, es recomendable normalizar los datos antes de utilizar esta medida de distancia

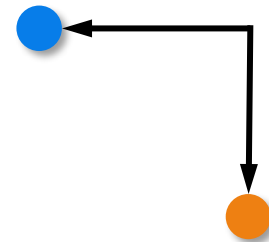
Análisis clúster: métricas de distancia

■ Distancia manhattan:

- Distancia entre dos datos considerando que sólo pueden moverse en ángulos rectos
- No hay ningún movimiento diagonal involucrado en el cálculo de la distancia

- Definición:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^m |x_i - y_i|$$



- Apropriada cuando el conjunto de datos tiene variables discretas y/o binarias, ya que tiene en cuenta las “rutas” realistas que podrían formarse con los valores de esas variables

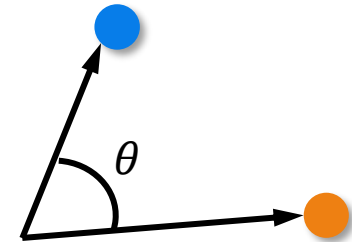
Análisis clúster: métricas de distancia

■ Distancia coseno:



- Es el coseno del ángulo formado por los dos datos
- Es una medida de la orientación: la magnitud no es relevante

$$d(\mathbf{x}, \mathbf{y}) = \cos(\theta) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

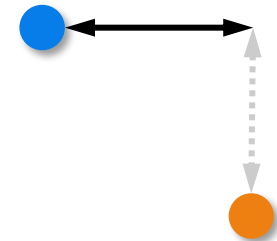


- Se utiliza con frecuencia para análisis de textos cuando los datos se representan mediante el recuentos de palabras:
 - Si en un documento una palabra ocurre con más frecuencia que en otro no significa necesariamente que esté más relacionado con esa palabra (pueden ser, por ejemplo, textos de diferentes longitudes)
- Inconveniente: no es apropiada si la magnitud de los vectores es relevante, no simplemente su dirección

■ Distancia Chebyshev:

- Es la mayor diferencia entre dos datos a lo largo de cualquiera de los eje de coordenadas
- Definición:

$$d(\mathbf{x}, \mathbf{y}) = \max_i (|x_i - y_i|)$$

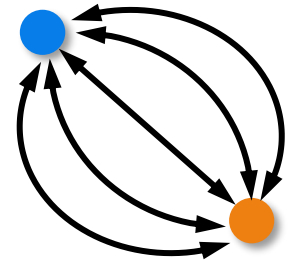


- Medida menos habitual en general, solo apropiada para casos muy concretos, por ejemplo, en juegos o problemas de logística (movimiento de grúas)

Análisis clúster: métricas de distancia

■ Distancia Minkowski:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^m |x_i - y_i|^p \right)^{\frac{1}{p}}$$



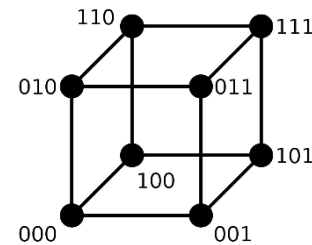
- Medida que generaliza a otras:
 - Si $p=1 \rightarrow$ Distancia manhattan
 - Si $p=2 \rightarrow$ Distancia Euclídea
- Normalmente el valor p se establece entre 1 y 2, ya que para valores de $p < 1$ no satisface la desigualdad triangular
- Ventaja: medida flexible al poder adaptarla al problema con el valor del parámetro p
- Inconveniente: encontrar el valor óptimo del parámetro p , ya que puede ser computacionalmente bastante costoso

Análisis clúster: métricas de distancia

■ Distancia Hamming:

- Es el número de valores que son diferentes entre dos vectores

Vector 1	1	0	1	0	0
Vector 2	1	1	1	1	0



- Normalmente se utiliza para comparar dos vectores binarios de igual longitud, aunque también se puede usar para comparar dos cadenas de caracteres
- Ventaja: apropiada para medir la distancia entre variables categóricas
- Desventaja: no se puede usar directamente si los dos vectores o cadenas no tienen la misma longitud

- Familias de algoritmos de agrupamiento (*clustering*):
 - Clustering **jerárquico**: descomposición jerárquica del conjunto de datos
 - Agrupamiento por **particiones**: crean particiones sucesivas del conjunto de datos
 - Métodos basados en **densidad**: emplean funciones de densidad y conectividad

Clustering jerárquico

- Los datos no se particionan en grupos (*clusters*) de una sola vez, sino que se van haciendo particiones sucesivas a distintos niveles de agregación o agrupamiento
- Tipos principales
 - Aglomerativos:
 - Comienzan con tantos grupos como datos
 - En cada paso, se combinan el par de grupos más cercanos hasta que sólo quede un grupo o k grupos (número dado de antemano)
 - Disociativos:
 - Comienzan con un único grupo que engloba a todos los datos
 - En cada paso, se parte un grupo hasta que cada grupo contenga un único dato o queden en total k grupos

- Algoritmo aglomerativo básico:

Inicialización:

1. Crear tantos grupos (clústeres) como datos
2. Calcular la matriz de similitudes/distancias entre los grupos

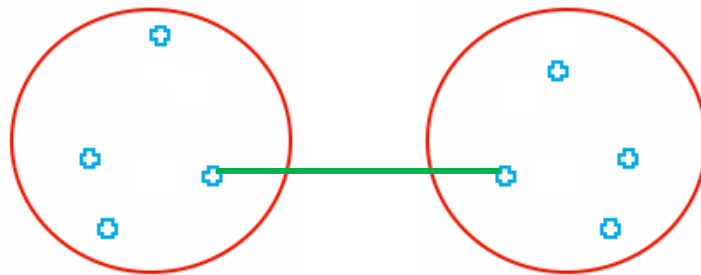
Repetir:

Combinar los dos grupos más cercanos

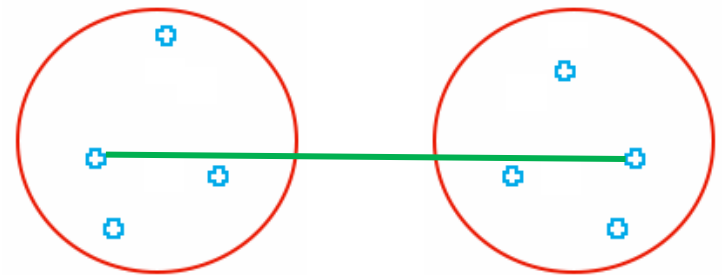
Actualizar la matriz de similitudes/distancias hasta que sólo queden k grupos

Clustering jerárquico

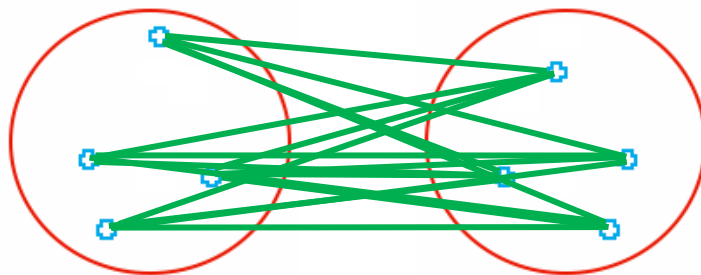
- Algunas alternativas para medir la distancia entre grupos:



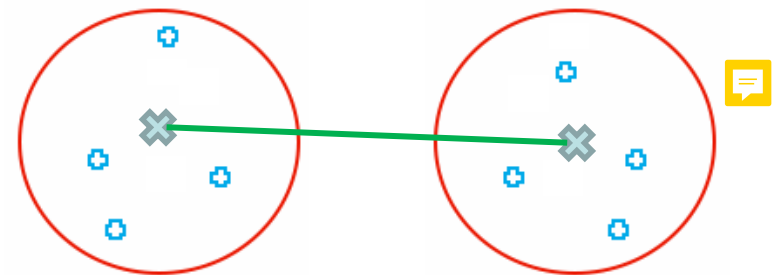
**Distancia mínima
(single linkage)**



**Distancia máxima
(complete linkage)**



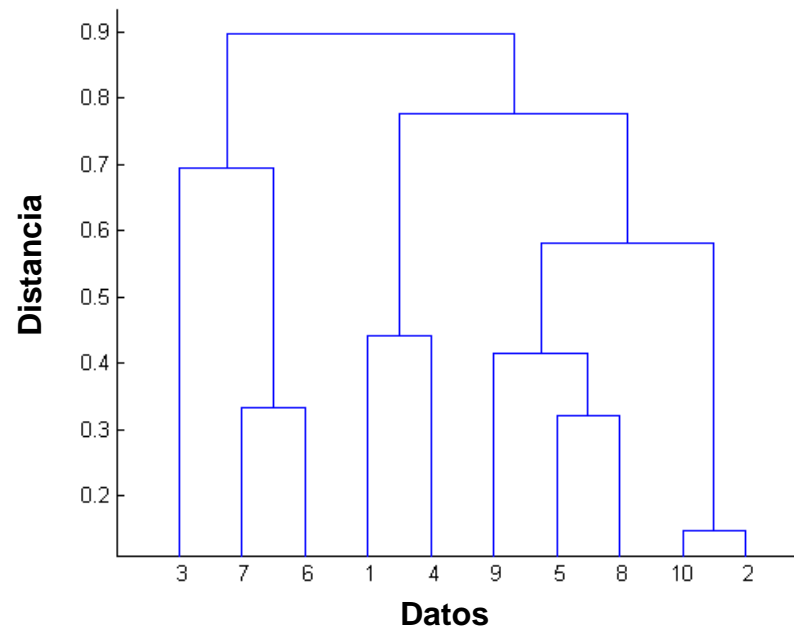
**Distancia promedio
(average linkage)**



Distancia entre centroides

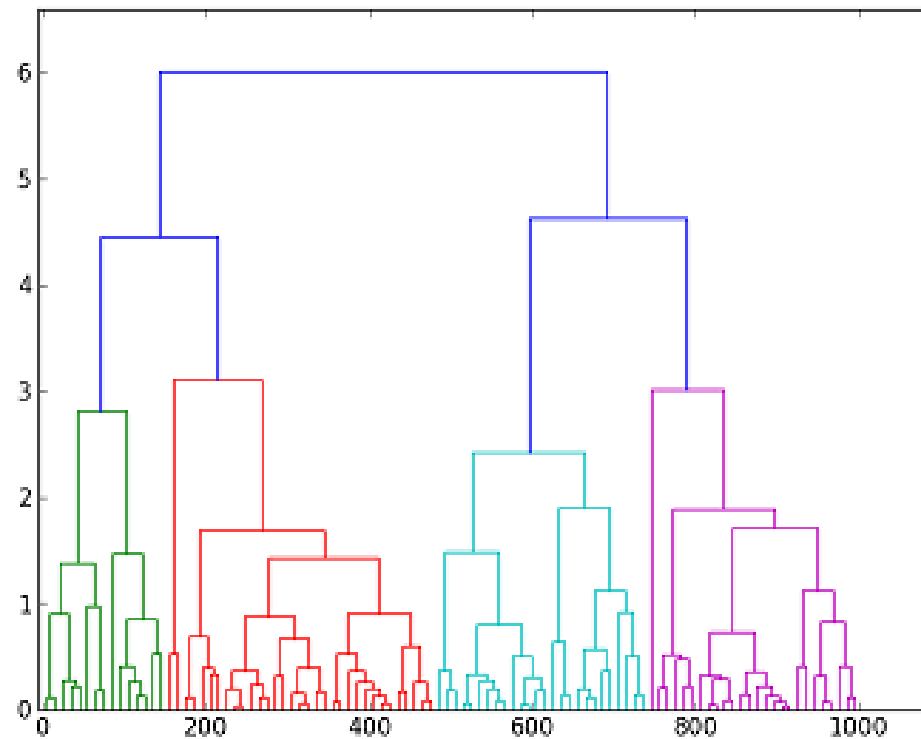
Clustering jerárquico

- Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de **dendrograma**:
 - Permite seguir de forma gráfica el procedimiento de unión
 - Muestra los grupos que se van uniendo, el nivel concreto en que lo hacen y el valor de la medida de asociación entre los grupos



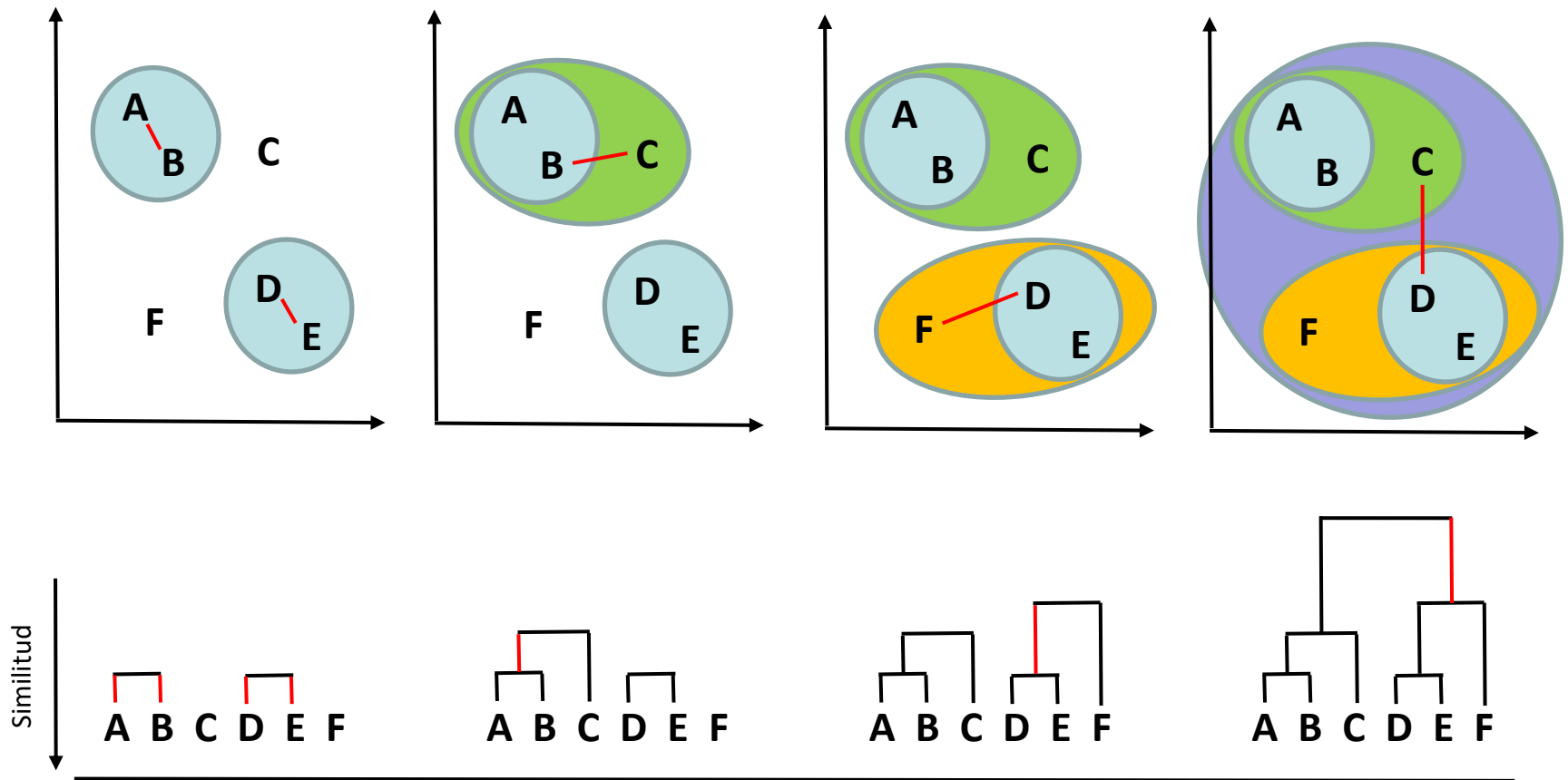
Clustering jerárquico

- El dendrograma puede ayudar a determinar el número adecuado de agrupamientos (aunque en muchos casos no será fácil)



Clustering jerárquico

- Ejemplo de creación del dendrograma:



Clustering jerárquico


- Ventajas:

- No es necesario indicar de antemano el número de grupos
- La naturaleza jerárquica se relaciona muy bien con la intuición humana en algunos dominios

- Inconvenientes:

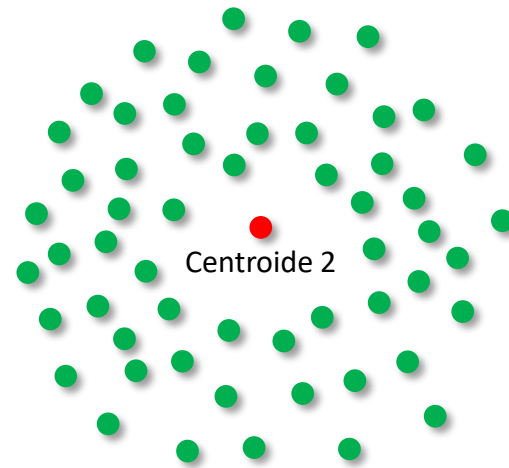
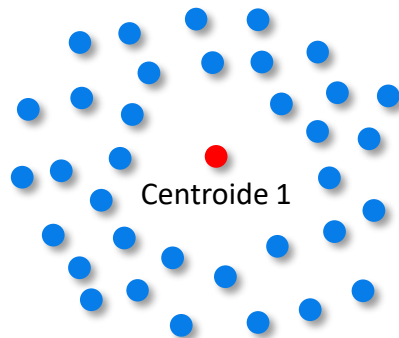
- No es escalable para grandes conjuntos de datos: $O(n^2)$, siendo n el número de datos

Agrupamiento por particiones

- Encontrar una partición en k grupos (predeterminados por el usuario) que optimice un criterio elegido 
- En general no es factible realizar un análisis exhaustivo de todas las posibles soluciones para obtener la partición globalmente óptima
- Por tanto, se buscan soluciones subóptimas obtenidas mediante algoritmos iterativos
- Estos métodos reasignan los datos a un grupo de forma iterativa partiendo de un agrupamiento inicial

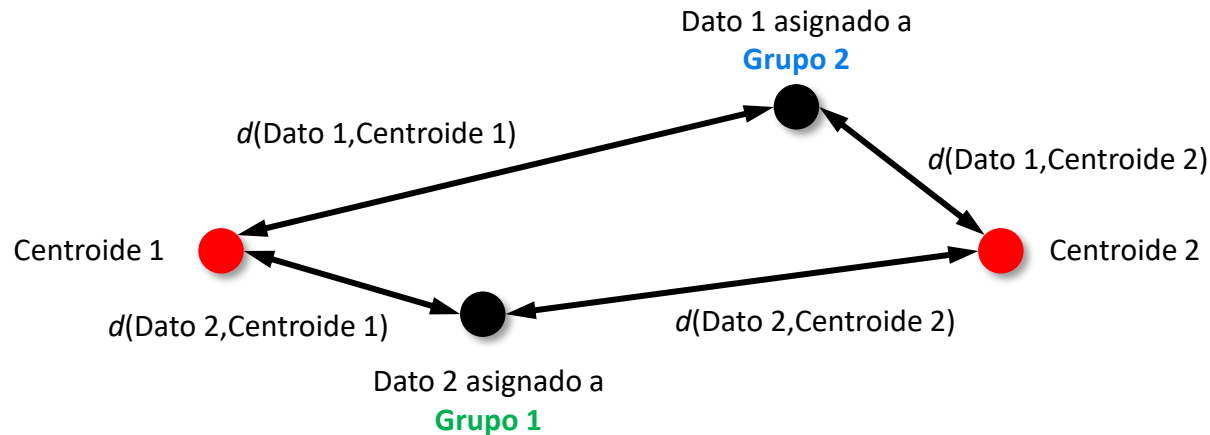
Agrupamiento por particiones: k -medias

- En k -medias (k -means) el número de grupos (k) debe ser preestablecido
- Cada grupo tiene asociado un centroide (centro geométrico del clúster)



Agrupamiento por particiones: k -medias

- Los datos se asignan al grupo cuyo centroide esté más cerca empleando para ello una métrica de distancia



- El método actualiza los centroides, de forma iterativa, en función de las asignaciones de datos a grupos hasta que los centroides dejen de cambiar

Agrupamiento por particiones: k -medias

- El objetivo del método es minimizar la varianza total intragrupo, o el error cuadrático total entre los datos de entrenamiento (\mathbf{x}_i) y los centroides (\mathbf{c}_j) asociados:

$$\min_{\mathbf{c}_j} \sum_{i=1}^n \sum_{j=1}^k p_{ij} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

donde k es el número de grupos, n es el número de datos y

$$\begin{cases} p_{ij} = 1, & \text{si } \mathbf{x}_i \text{ pertenece al grupo } j \\ p_{ij} = 0, & \text{si } \mathbf{x}_i \text{ no pertenece al grupo } j \end{cases}$$

Agrupamiento por particiones: k -medias

Algoritmo:

Inicialización: seleccionar k centroides aleatoriamente

hacer

Calcular las distancias de todos los datos a los k centroides

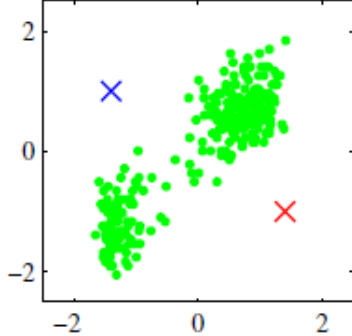
Formar k grupos, asignando cada dato al centroide más cercano

Recalcular los nuevos centroides con los datos de cada grupo

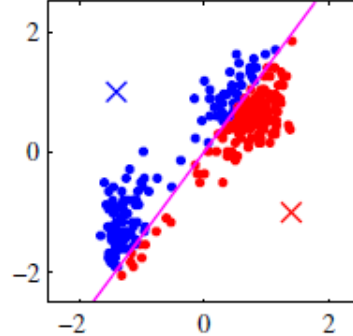
mientras cambien los centroides

k-medias: ejemplo

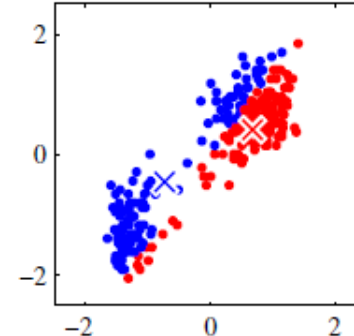
Inicialización de centroides



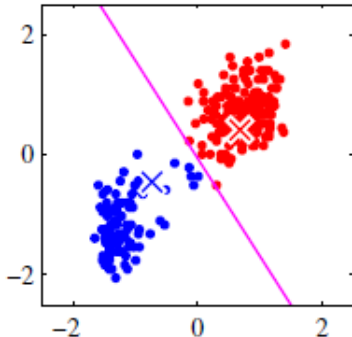
Paso 1: agrupamiento de datos



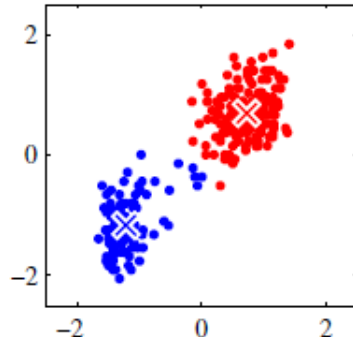
Paso 1: recalcular centroides



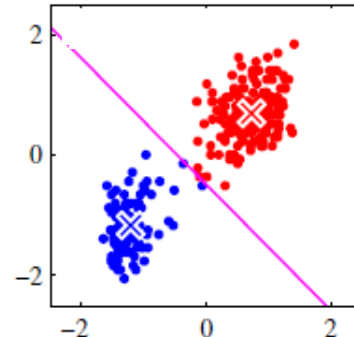
Paso 2: agrupamiento de datos



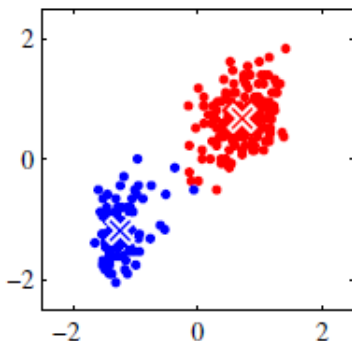
Paso 2: recalcular centroides



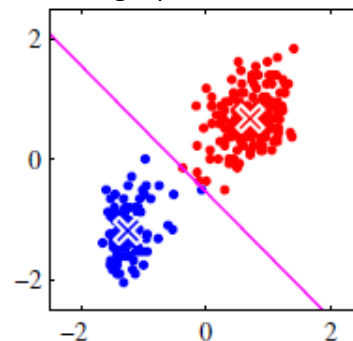
Paso 3: agrupamiento de datos



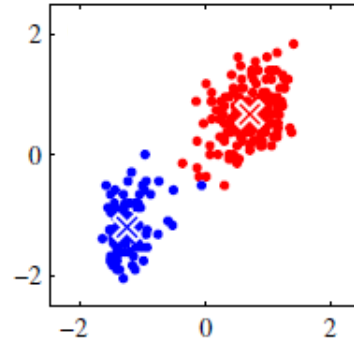
Paso 3: recalcular centroides



Paso 4: agrupamiento de datos



Paso 4: recalcular centroides



- Inconvenientes:
 - Es muy sensible a la elección inicial de los centroides, los grupos obtenidos pueden variar de forma significativa
- Posibles soluciones:
 - Realizar varias ejecuciones con varios conjuntos de centroides iniciales y comparar resultados
 - Estimar a priori unos buenos centroides:
 - Elegir una muestra y aplicar un método jerárquico
 - Métodos específicos: k-means++

k -medias: inicialización mediante k -means++

- Idea intuitiva del método:
 - Seleccionar centroides que estén alejados entre sí
 - Esto aumenta la posibilidad de elegir centroides iniciales que se encuentren en diferentes grupos



k -medias: inicialización mediante k -means++

- Pasos del método:

1. Elegir aleatoriamente un centroide inicial, \mathbf{c}_1 , del conjunto de datos X
2. Calcular para cada punto x la distancia menor, $d(\mathbf{x})$, al centroide más cercano a él en el conjunto actual de centroides
3. Elegir el siguiente centroide, \mathbf{c}_i , seleccionando $\mathbf{c}_i = \mathbf{x}' \in X$ con probabilidad:

$$\frac{d(\mathbf{x}')^2}{\sum_{x \in X} d(\mathbf{x})^2}$$

La probabilidad de elegirlo es directamente proporcional a la distancia desde el centroide más cercano a él

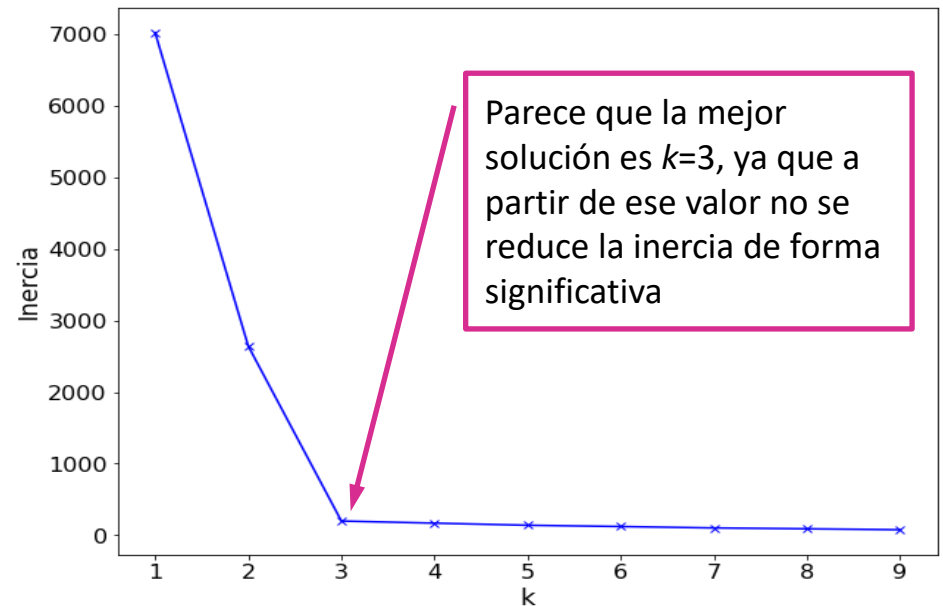
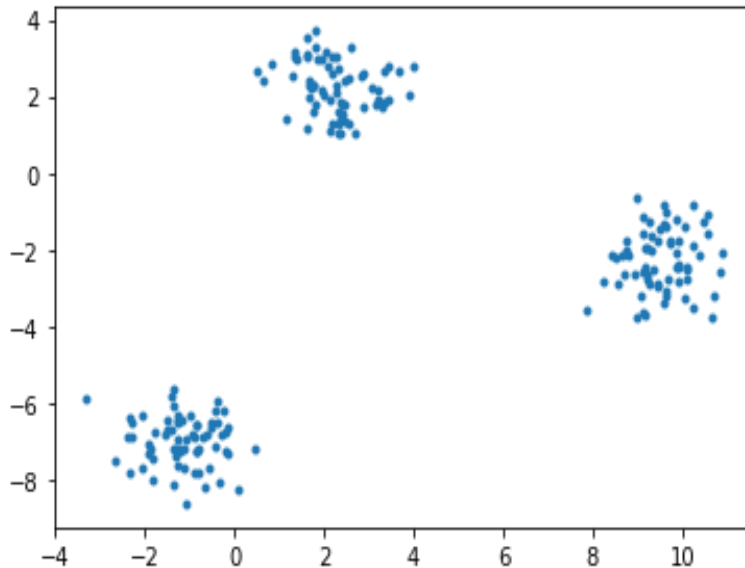
3. Repetir los pasos 2 y 3 hasta que tengamos un total de k centroides

- Determinación del valor de k :
 - No hay una regla general
 - Una posible estrategia sencilla es el método *elbow*:
 - Comenzar con $k = 1$ y aumentar k gradualmente
 - Para cada valor de k , mostrar gráficamente la evolución de la suma de las distancias al cuadrado de los datos a su centroide más cercano (inercia)
 - Elegir el valor de k a partir del cual no disminuye significativamente el valor de la inercia: existe un “codo” en la gráfica

- Ejemplo de determinación del valor de k con el método *elbow*:

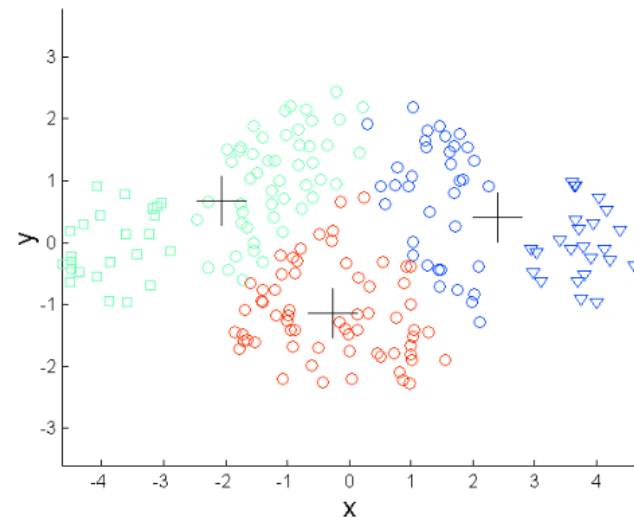
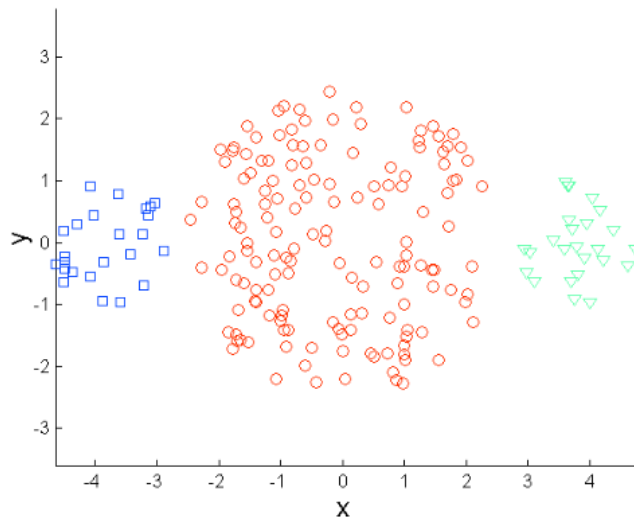


Datos a agrupar



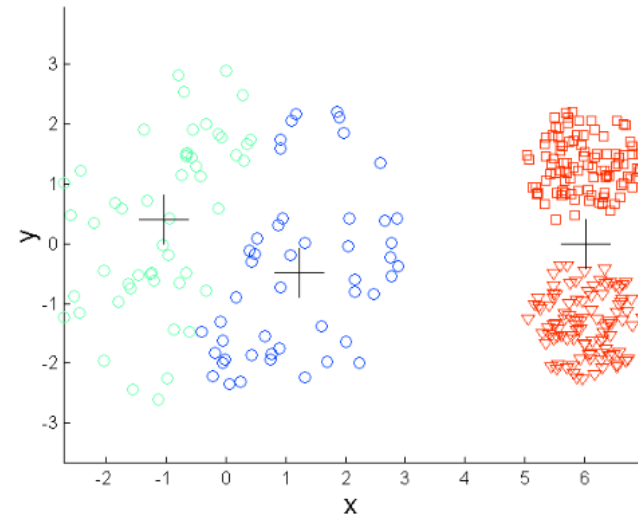
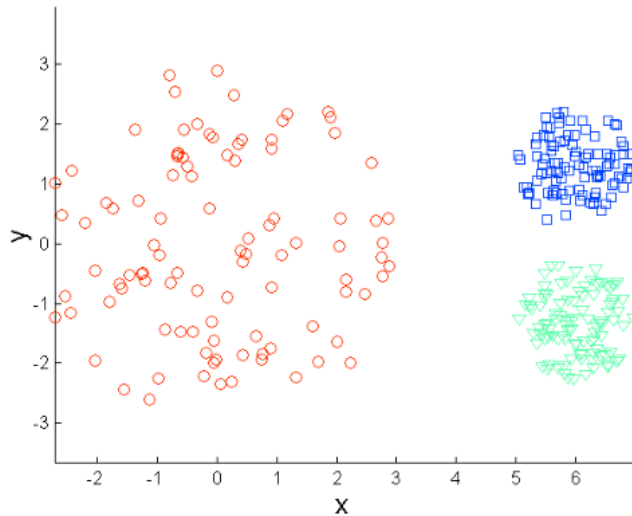
- El método no funciona bien cuando los clústeres son:

1. De distinto tamaño



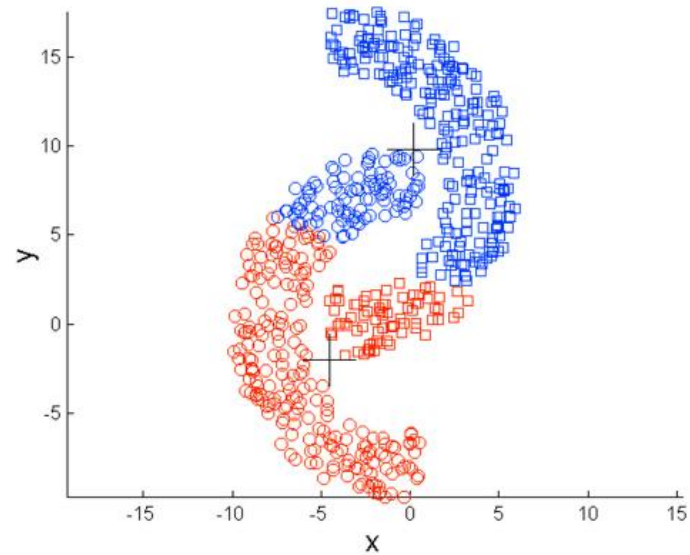
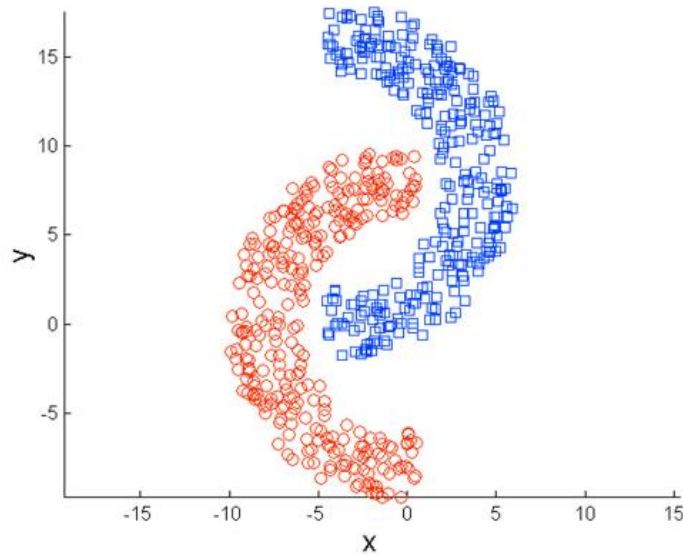
- El método no funciona bien cuando los clústeres son:

2. De diferente densidad



- El método no funciona bien cuando los clústeres son:

3. No convexos

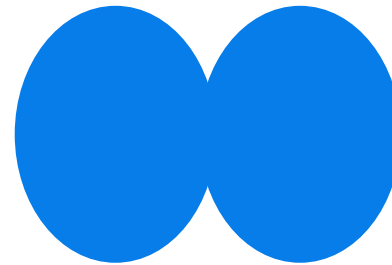
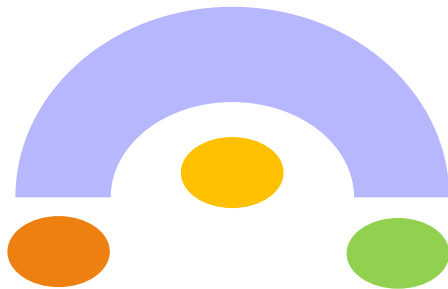


Agrupamiento basado en densidad

- Los métodos basados en densidad identifican grupos / regiones contiguas de alta densidad de puntos que están rodeadas por áreas poco densas
- Cada una de las regiones densas identificadas se asocia con un clúster / grupo
- Los puntos de datos en las regiones de separación de baja densidad de puntos generalmente se consideran ruido / valores atípicos

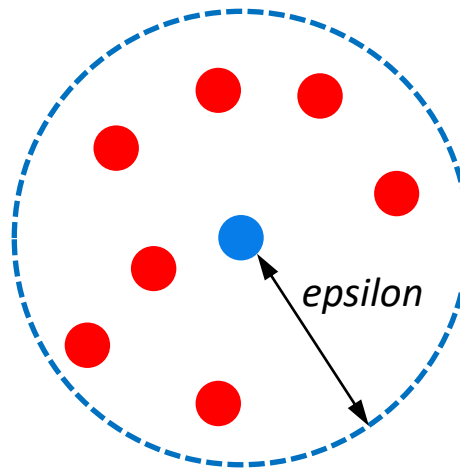
Agrupamiento basado en densidad

- El clustering basado en densidad es útil cuando los grupos tienen formas irregulares, están entrelazados o hay ruido / outliers en los datos



Agrupamiento basado en densidad: DBSCAN

- Es un algoritmo que implementa la noción de densidad mediante un procedimiento sencillo
- Densidad de cada punto: número de datos que están dentro de un radio específico, dado por un hiperparámetro (*epsilon*)

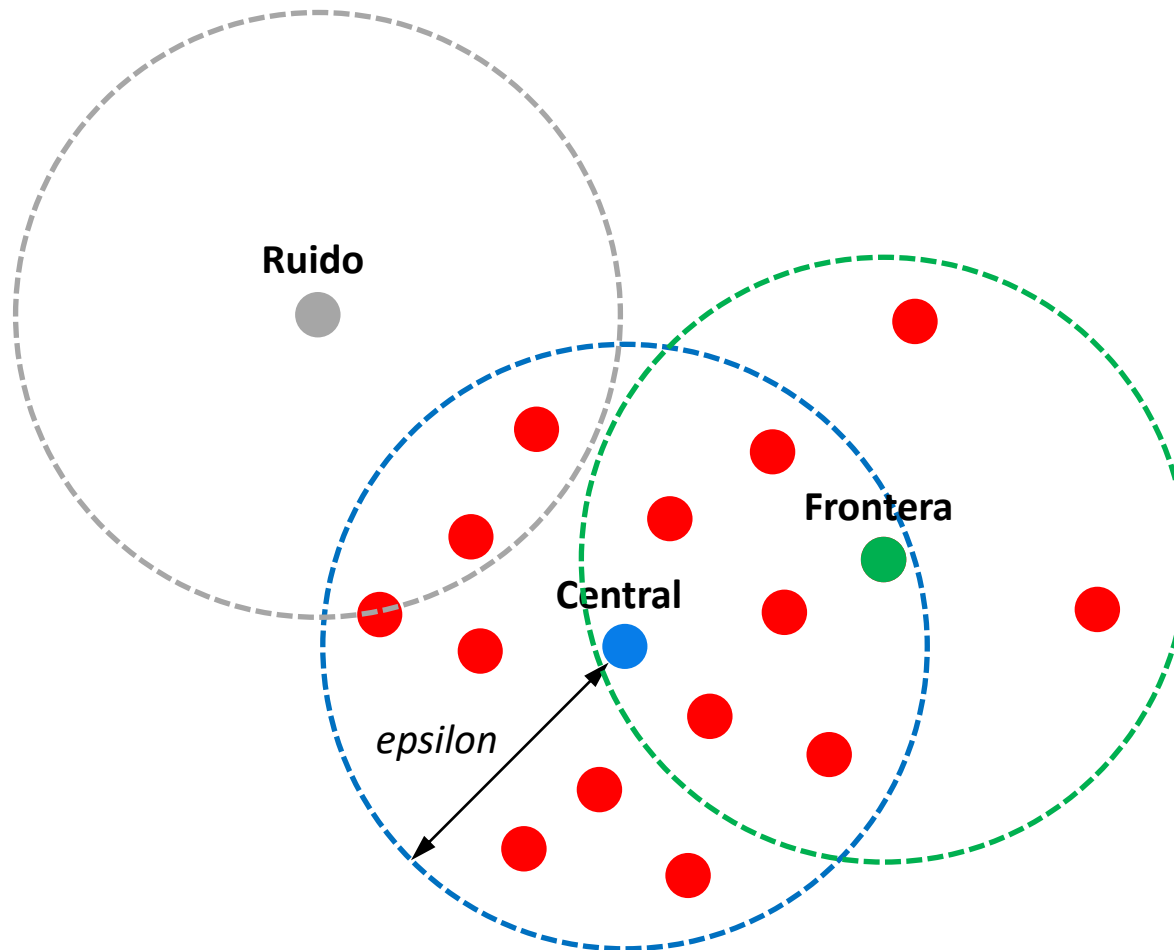


Agrupamiento basado en densidad: DBSCAN

- Dependiendo del valor de densidad, y de un hiperparámetro que indica un número mínimo de puntos (minPts), cada punto se clasifica en uno de 3 tipos posibles:
 - **Central** (*core*): tienen un número mínimo de puntos (minPts) en su vecindario de radio *epsilon*
 - **Frontera** (*border*): tienen menos de minPts puntos en su vecindario de radio *epsilon*, pero están en el vecindario de algún punto central
 - **Ruido** (*noise*): los que no son centrales ni frontera

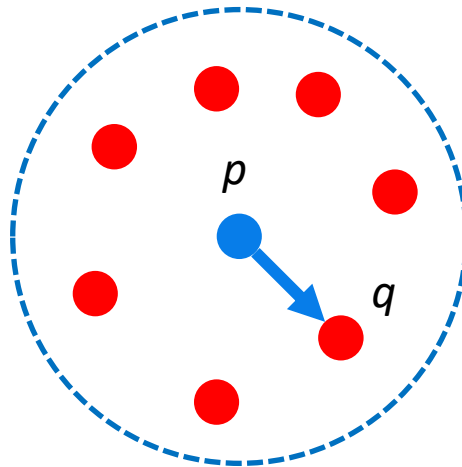
Agrupamiento basado en densidad: DBSCAN

- Ejemplo gráfico de los 3 tipos de puntos (minPts=10):



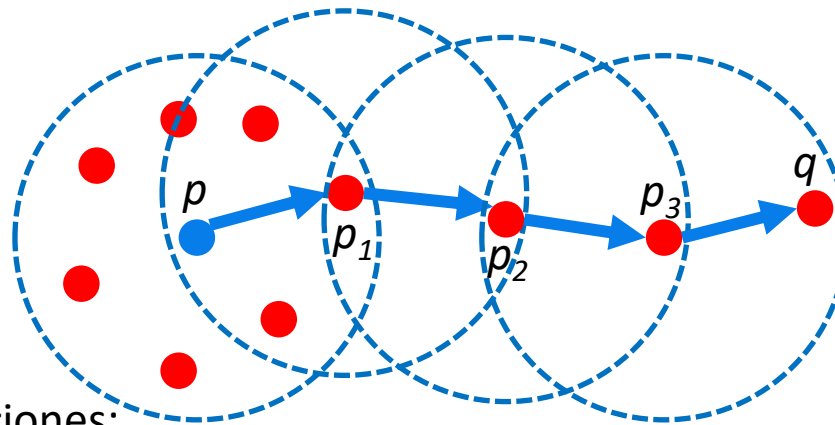
Agrupamiento basado en densidad: DBSCAN

- Conceptos asociados al método
 - **Accesibilidad directa:** un punto q es directamente accesible desde un punto p ($p \rightarrow q$) si se cumplen las dos siguientes condiciones:
 - p es un punto central
 - q se encuentra a una distancia menor de *epsilon* de él



Agrupamiento basado en densidad: DBSCAN

- Conceptos asociados al método
 - **Accesibilidad por densidad:** un punto q es accesible por densidad desde p , dado un ϵ y minPts , si existe una cadena de puntos p_1, p_2, \dots, p_n tal que $p \rightarrow p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n \rightarrow q$

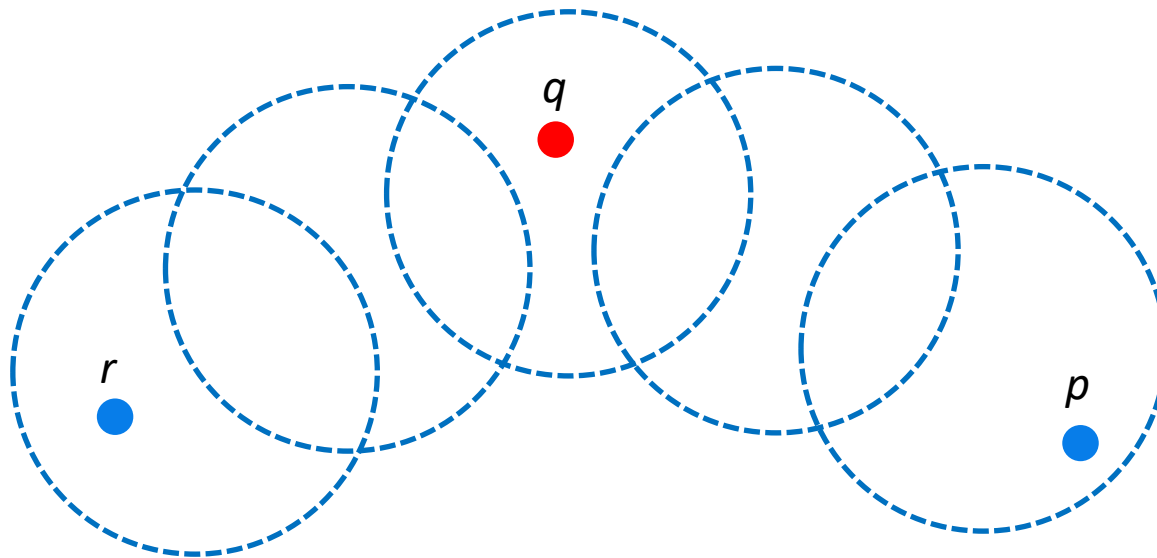


Observaciones:

- Por definición de la accesibilidad directa, todos los puntos p, p_1, p_2, \dots, p_n , deben ser centrales
- La accesibilidad por densidad no es necesariamente simétrica: puede ser que p no sea accesible por densidad desde q si q no es un punto central

Agrupamiento basado en densidad: DBSCAN

- Conceptos asociados al método
 - **Conectividad por densidad:** un punto p está conectado por densidad con un punto r si cumple las siguientes condiciones:
 - Existe un punto q tal que p es accesible por densidad desde q
 - El punto r es accesible por densidad desde q



Agrupamiento basado en densidad: DBSCAN

Algoritmo:

$i = 1$

hacer

 Seleccionar un punto p del conjunto de puntos M

 Encontrar el conjunto de puntos P que están conectados por densidad con p

si $P = \{\}$

$M = M \setminus \{p\}$

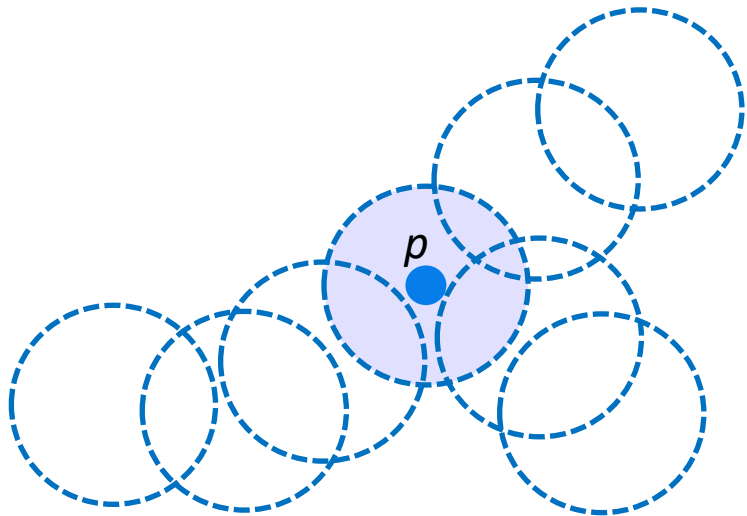
sino

$C_i = P$

$i = i + 1$

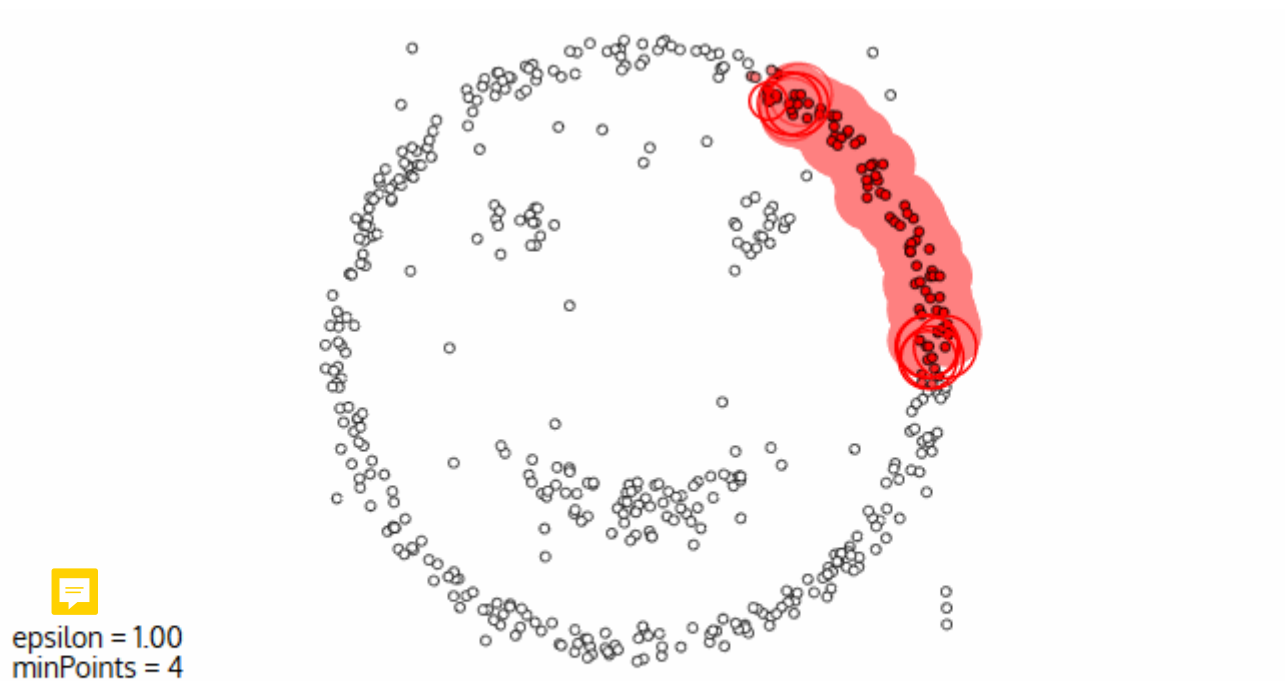
$M = M \setminus P$

mientras $M \neq \{\}$



Agrupamiento basado en densidad: DBSCAN

- Ejemplo de creación de clústeres:



Fuente:

<https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html>

Agrupamiento basado en densidad: DBSCAN

- Elección de parámetros:
 - El comportamiento del algoritmo depende de la elección de sus parámetros (*epsilon* y minPts)
 - Si estos parámetros no se fijan adecuadamente se obtendrán clústeres poco útiles
 - Alguna regla heurística propuesta para el parámetro minPts (Sander et al., 1998):

$$\text{MinPts} = 2 * \text{dimensión}$$

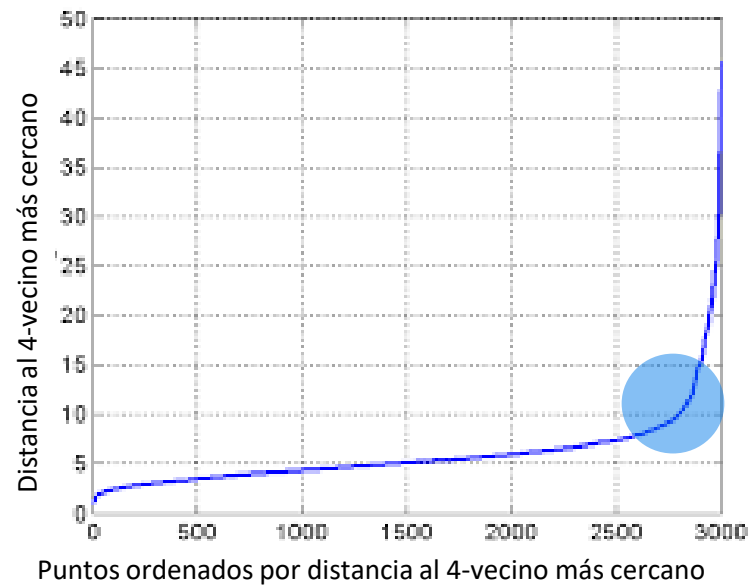
(Sander et al., 1998) Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1998. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. Data Mining and Knowledge Discovery 2, 2 (1998), 169–194. DOI: <http://dx.doi.org/10.1023/A:1009745219419>

Agrupamiento basado en densidad: DBSCAN

- Elección de parámetros:
 - Un procedimiento empleado habitualmente para fijar el valor de *epsilon* usa el concepto de k-distancia
 - La k-distancia de un punto se define como la distancia al k-ésimo punto más cercano
 - Pasos:
 - Calcular la k-distancia de cada punto, siendo $k = \text{minPts}$
 - Ordenar las k-distancias de menor a mayor y mostrarlas gráficamente

Agrupamiento basado en densidad: DBSCAN

- Elección de parámetros:
 - El valor en el que se produce un cambio brusco en la curva es un valor apropiado para *epsilon*



Agrupamiento basado en densidad: DBSCAN

■ Ventajas:

- No requiere que el usuario indique el número de clústeres
- Es determinista: siempre genera los mismos clústeres cuando se dan los mismos datos en el mismo orden, aunque los resultados pueden diferir cuando se proporcionan en un orden diferente
- Apropiado para clústeres de formas arbitrarias
- Robusto al ruido

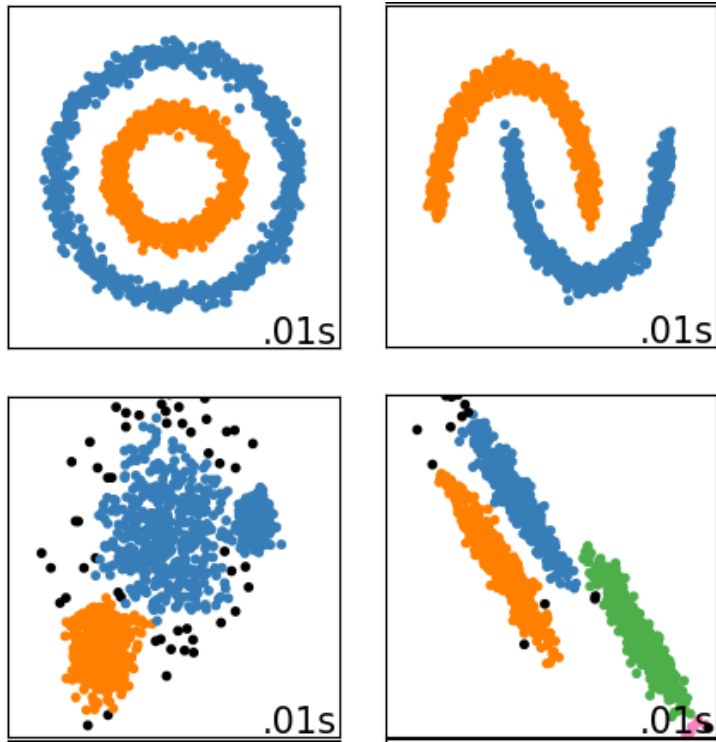
■ Inconvenientes:

- Conjuntos de datos con distintas densidades son problemáticos
- Es necesario ajustar de forma apropiada el valor de *epsilon* y minPts

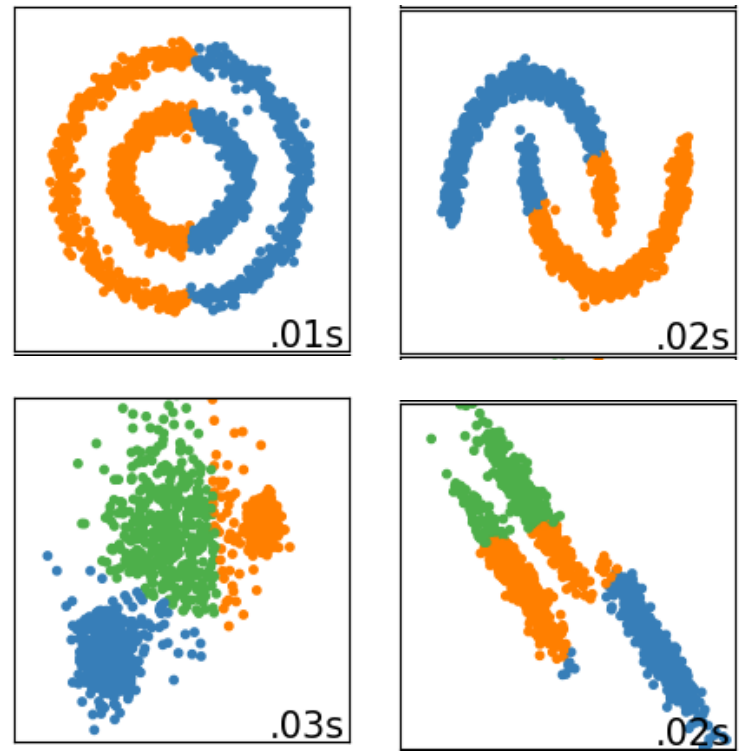
Agrupamiento basado en densidad: DBSCAN

- Comparativa de DBSCAN y *k*-medias:

DBSCAN



K-medias



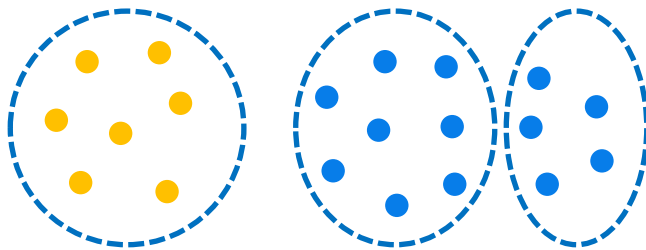
Evaluación de la calidad del agrupamiento

- ¿Cómo evaluar los resultados del agrupamiento cuando se disponen de las etiquetas reales de cada dato (evaluación **extrínseca**)?
- Más complicado que en la clasificación supervisada. Problemas:
 - **Correspondencia**: en el agrupamiento los identificadores de clúster se asignan de forma arbitraria ya que es un aprendizaje no supervisado
 - **Inconsistencia**: el número de grupos puede ser diferente del número de clases reales
- ¿Cómo encontrar la correspondencia entre la clase real del dato y un agrupamiento dado?

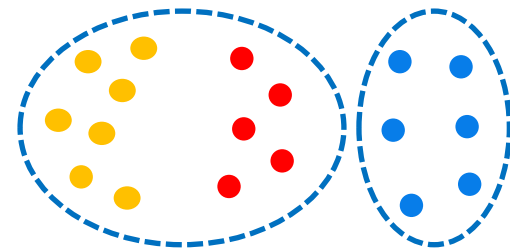
Evaluación extrínseca de la calidad del agrupamiento

■ Medida V (*V-measure*)

- Emplea los conceptos de homogeneidad y completitud
- **Homogeneidad**: cada grupo contiene sólo miembros de una sola clase
- **Integridad**: todos los miembros de una clase se asignan al mismo grupo



Este agrupamiento es homogéneo
pero no íntegro



Este agrupamiento es íntegro
pero no homogéneo

- Medida V (*V-measure*)

Es la media armónica entre homogeneidad e integridad:

$$v = \frac{(1 + \beta) * \text{homogeneidad} * \text{integridad}}{\beta * \text{homogeneidad} + \text{integridad}}$$

β : relación de peso atribuido a cada elemento

- Si $\beta = 1$, la homogeneidad y la integridad tienen el mismo peso
- Si $\beta > 1$, la integridad se pondera más fuertemente
- Si $\beta < 1$, la homogeneidad se pondera más fuertemente

- Medida V (*V-measure*)
 - Es independiente a los valores de las etiquetas: una permutación de los valores de las etiquetas de clase o clúster no cambiará el valor de la medida
 - Los valores de esta medida están en el rango $[0, 1]$
 - Un valor de 1 implica un etiquetado perfecto

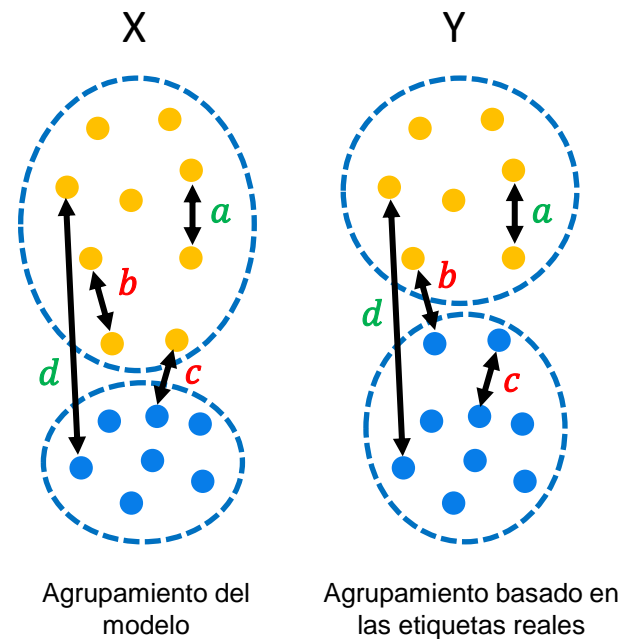
- Ventajas de la medida V (*V-measure*):
 - Interpretación intuitiva: se puede analizar cualitativamente en términos de homogeneidad e integridad
 - No se hace ninguna suposición sobre la estructura del clúster: no se asumen formas determinadas
- Inconvenientes de la medida V (*V-measure*):
 - No está normalizada respecto a un etiquetado aleatorio: dependiendo del número de datos, grupos obtenidos y clases reales, un etiquetado completamente aleatorio no siempre dará los mismos valores de homogeneidad, integridad y, por tanto, de medida V

Evaluación extrínseca de la calidad del agrupamiento

- Índice de Rand (*Rand Index*)

Proporciona una medida de similitud entre dos agrupamientos (X e Y) considerando todos los posibles pares de muestras y contando los pares que se asignan en el mismo o en diferentes grupos

X/Y	Y: pares en el mismo grupo	Y: pares en grupos diferentes
X: pares en el mismo grupo	<i>a</i>	<i>b</i>
X: pares en grupos diferentes	<i>c</i>	<i>d</i>



- Índice de Rand (*Rand Index*)

Definido como:

$$RI(X, Y) = \frac{a + d}{a + b + c + d}$$

Los valores de esta medida están en el rango [0, 1]

Las agrupaciones similares tienen un índice Rand alto y el valor 1 es la puntuación de coincidencia perfecta

- Ventajas del índice de Rand (*Rand Index*)
 - Fácil de interpretar: es proporcional al número de pares de muestras cuyas etiquetas son las mismas o son diferentes en ambos agrupamientos
 - Tiene un rango acotado
 - No se hace ninguna suposición sobre la estructura (forma) del agrupamiento: se puede usar para comparar todo tipo de algoritmos de agrupamiento

- Desventaja del índice de Rand (*Rand Index*)
 - No tiene en cuenta el posible acuerdo por casualidad lo que provoca que suela estar cerca de 1 incluso si los agrupamientos difieren significativamente
 - En la práctica, a menudo hay una mayoría de pares de elementos a los que se les asigna la etiqueta de par diferente tanto en el agrupamiento predicho como en el real, lo que resulta en una alta proporción de etiquetas de pares que están de acuerdo, lo que conduce posteriormente a una puntuación alta

- Índice de Rand ajustado (*Adjusted Rand Index*)

$$ARI(X, Y) = \frac{RI(X, Y) - E[RI]}{\max(RI) - E[RI]} \quad \text{---} \rightarrow \text{RI esperado}$$

- Los valores de esta medida están en el rango $[-1, 1]$:
 - Tiene un valor cercano a 0 en el caso de un etiquetado aleatorio, independientemente del número de agrupaciones y muestras
 - Tiene un valor exactamente 1 cuando las agrupaciones son idénticas (hasta una permutación)
- Es una medida simétrica

Evaluación extrínseca de la calidad del agrupamiento

- Ejemplos de valores de RI y RI ajustado:

```
clase_real      = [0, 0, 0, 1, 1, 1, 1]
agrupamiento    = [0, 0, 1, 1, 2, 2, 2]

rand_score(clase_real, agrupamiento)      —————> 0,71
adjusted_rand_score(clase_real, agrupamiento) —————> 0,38
```

```
clase_real      = [0, 0, 0, 1, 1, 1, 1]
agrupamiento    = [1, 1, 2, 2, 3, 3, 3]

rand_score(clase_real, agrupamiento)      —————> 0,71
adjusted_rand_score(clase_real, agrupamiento) —————> 0,38
```

Cambio en las etiquetas predichas

```
clase_real      = [0, 0, 0, 0, 1, 1]
agrupamiento    = [0, 1, 2, 4, 4, 4]

rand_score(clase_real, agrupamiento)      —————> 0,47
adjusted_rand_score(clase_real, agrupamiento) —————> -0,11
```

Etiquetas muy mal emparejadas

Evaluación de la calidad del agrupamiento

- ¿Cómo evaluar los resultados del agrupamiento cuando **no** se disponen de las etiquetas reales de cada dato?
- También conocida como evaluación **intrínseca** del agrupamiento

- Coeficiente de silueta (*Silhouette Coefficient*)

Empleada para estudiar la distancia de separación entre los grupos resultantes

Se calcula utilizando la distancia media dentro del grupo (a) y la distancia media del grupo más cercano (b) para cada muestra:

$$s = \frac{b - a}{\max(a, b)}$$

a : distancia media entre una muestra y todos los demás puntos del mismo grupo

b : distancia media entre una muestra y todos los demás puntos en el siguiente grupo más cercano

- Coeficiente de silueta (*Silhouette Coefficient*)
 - Los valores de esta medida están en el rango $[-1, 1]$
 - Una puntuación alta de este coeficiente indica un modelo con clústeres mejor definidos:
 - Valores cercanos 1 indican que la muestra está lejos de los grupos vecinos
 - Valores cercanos a 0 indican clústeres superpuestos
 - Valores negativos generalmente indican que una muestra se ha asignado al grupo equivocado, ya que otro grupo es más similar