Big Data y Minería de Datos

Unidad 3. Aprendizaje computacional

U3.3. Aprendizaje no supervisado

Autores: José Antonio Ruipérez Valiente (jruiperez@um.es),
Mariano Albaladejo González (mariano.albaladejog@um.es) y
Manuel Jesús Gómez Moratilla (manueljesus.gomezm@um.es)

UNIVERSIDAD DE
MURCIA

Unidad 3. Aprendizaje computacional
Estructura

UNIVERSIDAD DE MURCIA

- Fundamentos del aprendizaje no supervisado
- Algoritmos de agrupamiento
- Evaluación de resultados de agrupamiento

Big Data y Minería de Datos
Contenidos

UNIVERSIDAD DE MURCIA

Introducción a la minería de datos

Técnicas de inteligencia artificial

Reglas
Fundamentos lógicos
Sistemas basados en reglas
Reglas de asociación

Aprendizaje computacional
Fundamentos
Regresión y clasificación
No supervisado

Procesado del lenguaje natural
Fundamentos
Representación y tópicos
Sentimientos y NER

Introducción al big data

UNIVERSIDAD DE MURCIA

Unidad 3. Aprendizaje computacional

FUNDAMENTOS DEL APRENDIZAJE NO SUPERVISADO

Unidad 3. Aprendizaje computacional
El problema básico

UNIVERSIDAD DE MURCIA

- Vamos a tener un conjunto de observaciones sin etiquetas
  - ¡No es supervisado!
- El agrupamiento (clustering) será el proceso de asignarlos a grupos (clusters)
- Criterio: Poner las observaciones similares juntos y los diferentes en otros grupos
- Objetivo: Obtener un conjunto de grupos similares
- Va a depender de los datos de partida:
  - ¿Cómo calcular la similitud entre distintos tipos de datos?
  - Numéricas, categóricas, ordinales, documentos, imágenes…
- Tema estudiado desde los años 60-70 desde distintos campos
  - Empezando por taxonomías numéricas
  - Posteriormente incluido como parte de la IA

- Tenemos que agrupar observaciones similares... ¿cómo operacionalizamos la similitud?
- Aplicaremos las métricas de distancia, en la que dado un conjunto de m elementos $\Omega = \{1, 2, 3, \dots, m\}$, calcularemos las distancias entre dos elementos $d(i, j) = dij$
- Tiene algunas propiedades importantes
  - $d(i, j) \geq 0, \forall\, i, j \in \Omega$
  - $d(i, i) = 0, \forall\, i \in \Omega$
  - $d(i, j) = d(j, i) \forall\, i, j \in \Omega$
- Ejemplos de distancias vistas antes como la euclídea, Manhattan, Hamming…

- Es difícil proporcionar una categorización ajustada, porque algunas categorías tienen solapes
  - Métodos utilizan características de distintas categorías
- Métodos particionales: Particionamiento en un solo nivel con separación excluyente de grupos
- Métodos jerárquicos: Descomposición jerárquica, aglomerativa o divisiva
- Métodos basados en distribuciones: Basados en nociones de densidad de probabilidad (en lugar de distancias)
- Métodos basados en malla: Basados en la cuantización de un espacio en un número finito de celdas

UNIVERSIDAD DE MURCIA

Unidad 3. Aprendizaje computacional

ALGORITMOS DE AGRUPAMIENTO

- Dado un conjunto de *m* objetos, el particionamiento construye *k* particiones, donde cada partición representa un clúster
  - Cada grupo debe contener al menos un objeto, $k \leq n$
  - Los métodos básicos tienen separación exclusiva
  - La mayoría son basados en distancias, basado en un particionamiento inicial y una relocalización iterativa
- Elementos mismo grupo están cerca/relacionados y los de diferentes grupos alejados/diferentes (hay otros criterios)
- Alcanzar resultado óptimo global es computacionalmente prohibitivo (demasiadas iteraciones, se alcanzan locales)
- Para grupos con formas específicas no funcionan bien
- Ok para bases de datos pequeñas/medianas
- El número *k* de grupos puede ser conocido o no
  - Normalmente es un parámetro del algoritmo

Medir coherencia de cada grupo

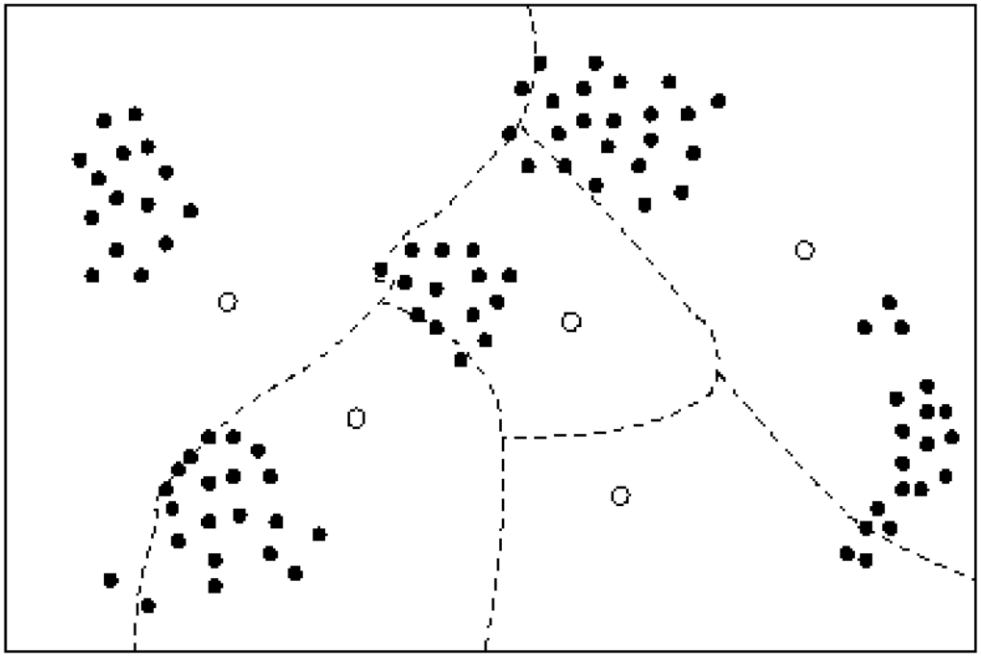
- Es necesario un criterio para medir la coherencia de los grupos
  - Criterio global: Representan grupo mediante un prototipo, asignando elementos a prototipos más cercanos
  - Criterio local: Forman grupos utilizando estructura local de los datos (por ej. zonas de alta densidad)
- Resultado: definir un encoder C, tal que: $C(i) = k \Leftrightarrow xi \in k$
- Objetivo: encontrar un encoder $C^{*}(i)$ que optimice criterio
- La distancia total entre los puntos de un conjunto será:
  - Intercluster: $B(C)\ \frac{1}{2}\sum_{k=1}^{K}\sum_{i:C(i)=k}\sum_{i'C(i')\neq k} d(xi, x_{i'})$
  - Intracluster: $\mathrm{W}(C)\ \frac{1}{2}\sum_{k=1}^{K}\sum_{i:C(i)=k}\sum_{i'C(i')=k} d(xi, x_{i'})$
- Para obtener un buen agrupamiento (encoder), debemos:
  - Maximizar la distancia intercluster $B(C)$, lo más separados
  - Minimizar la distancia intracluster $\mathrm{W}(C)$, lo más compactos

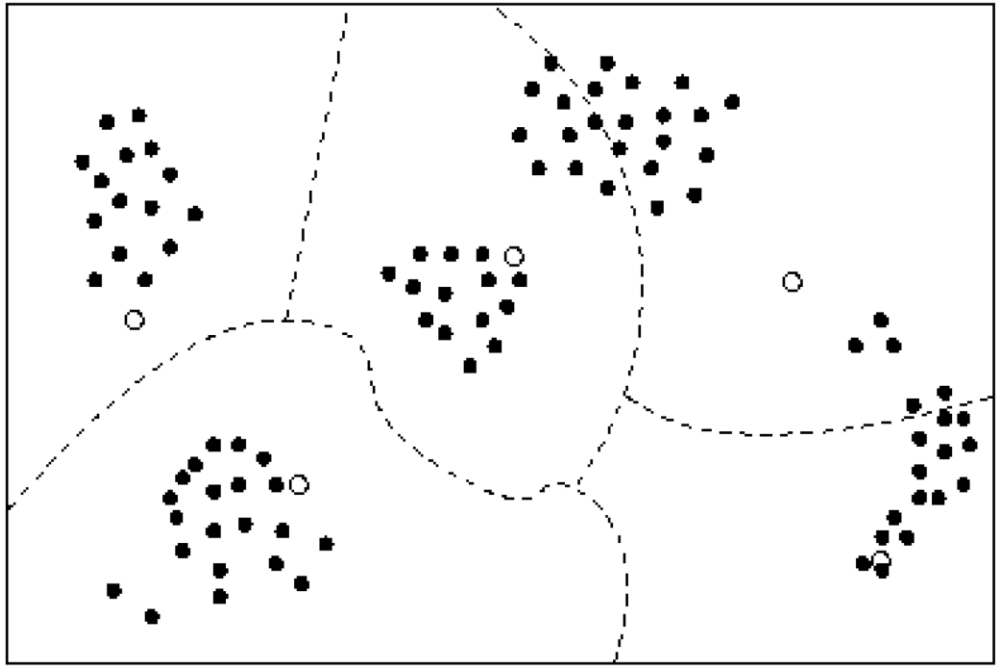
- Algoritmo popular. Variables numéricas y distancia euclídea
  - $d(x_i, x_{i'}) = \sum_{j=1}^{p}(xij - x_{i'j})^2 = \|x_i - xi_{,}\|^2$
- Por lo tanto, la distancia intracluster puede escribirse como:
  - $\mathrm{W}(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i:C(i)=k}\|x_i - \bar{x}_k\|^2$
  - Donde $\bar{x}_k$ es el centroide del grupo *k*
- Por lo tanto, tendremos un problema de optimización:
  - $\min_{\mathrm{C}}\sum_{k=1}^{K}\sum_{i:C(i)=k}\|x_i - \bar{x}_k\|^2$
- Algoritmo que intenta resolver el problema siguiendo ascensión de colinas
  - Después de cada iteración, no puede volver atrás y probar otra cosa

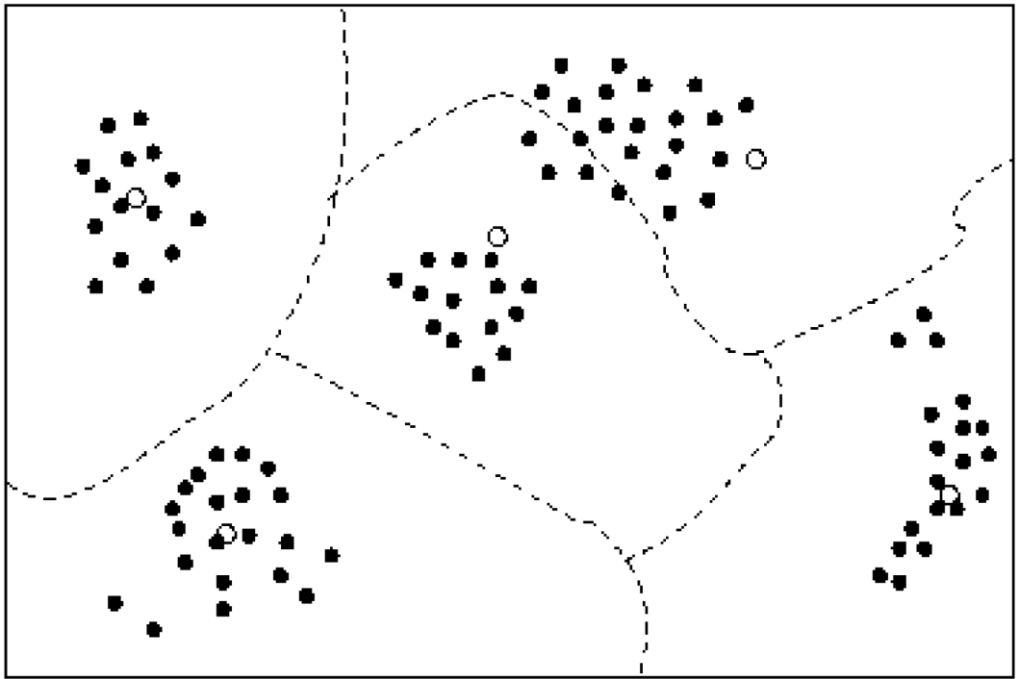
1. Dos posibles configuraciones iniciales del algoritmo:
   - Inicializar aleatoriamente los centroides, m_i que representan cada grupo *k*, ir al paso 2
   - Partir de una partición aleatoria de las observaciones a cada grupo, ir al paso 3
2. Distribuye los elementos entre los k grupos de acuerdo al representante:
   - $C(i) = arg\,\min_{1 \le k \le K} \|x_i - mk\|^2$
3. Calcula los nuevos centroides, m_k en base a todos los puntos donde $C(x) = k$
4. Si los m_k han cambiado, volver al paso 2. Si no, fin

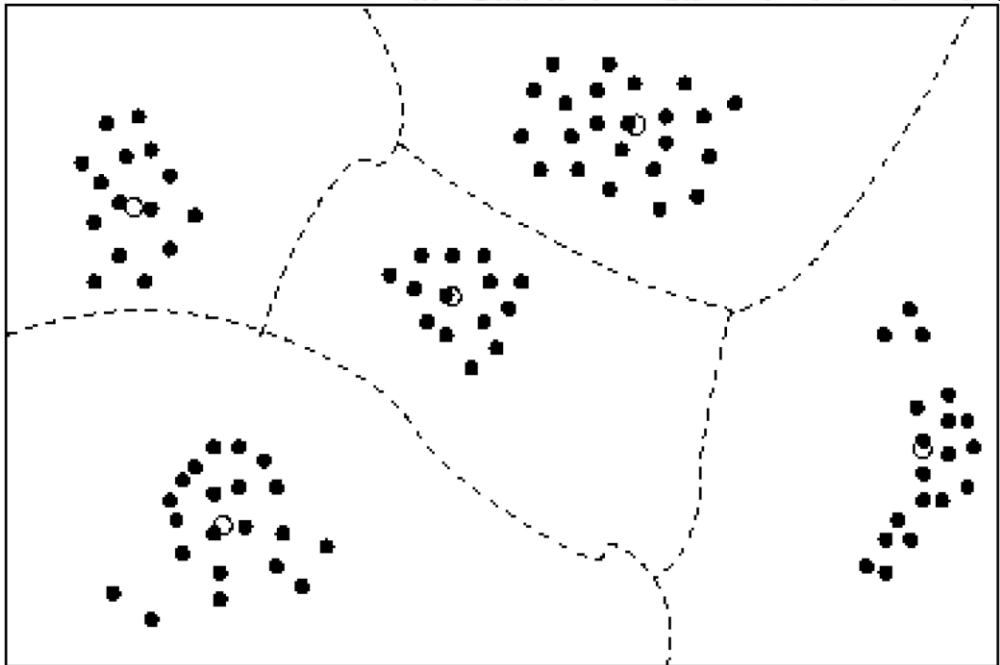
Unidad 3. Aprendizaje computacional
Algoritmo K-medias ejemplo

UNIVERSIDAD DE MURCIA



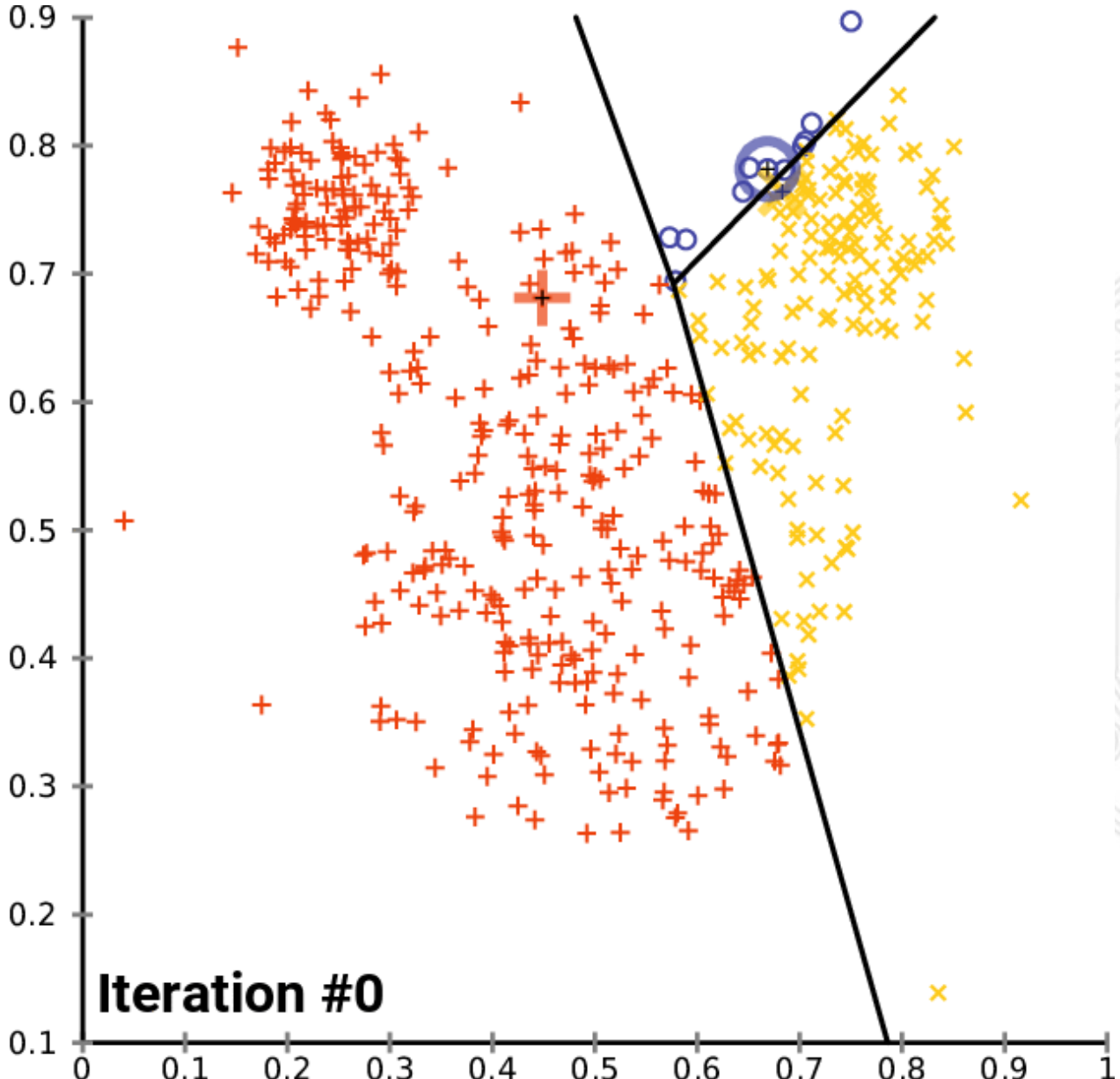






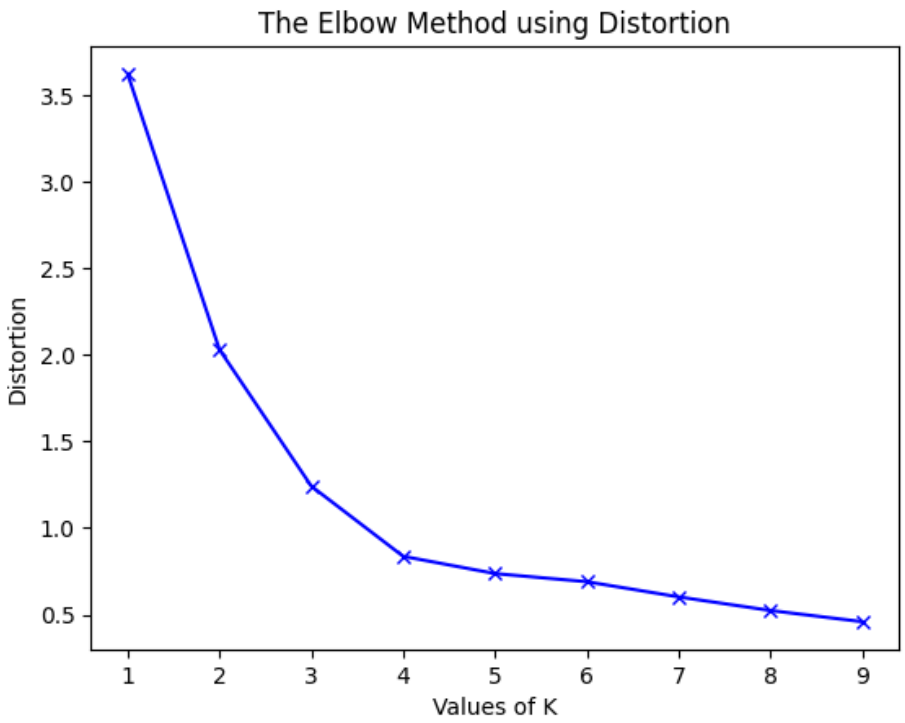
UNIVERSIDAD DE MURCIA

- Ejemplo online (https://www.youtube.com/watch?v=R2e3Ls9H_fc)



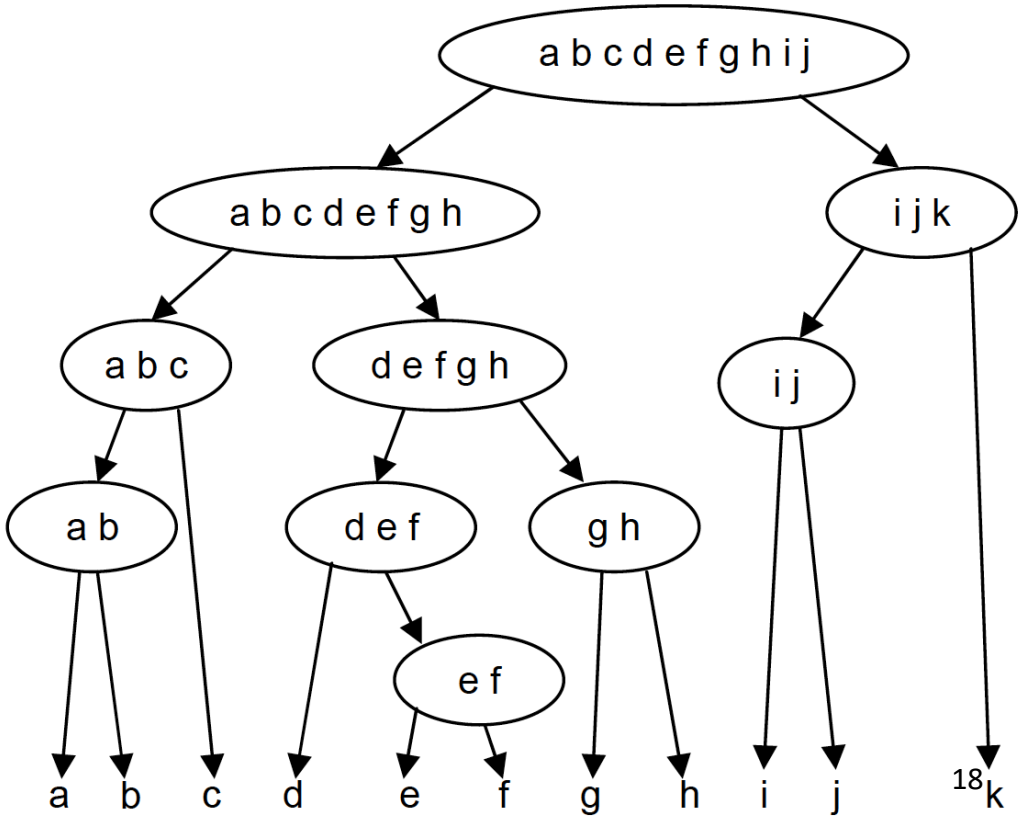
- Todos los atributos deben de ser reales
  - Debido al uso de esa métrica de distancia determinada
- Se ha usado la distancia euclídea
  - Los centroides se ubicarán en la media aritmética de los elementos que pertenecen al clúster
  - Por lo que se minimiza la distancia dentro del clúster (intracluster)
  - Amplifica el efecto de los atípicos debido al valor cuadrático
- Es posible utilizar otras medidas de distancia
- También se pueden aplicar medidas de similitud, pero entonces habrá que maximizarlas en vez de minimizar la distancia

- Extremadamente sensible a la elección inicial aleatoria de los centroides
  - Si se realizan múltiples ejecuciones, se puede comparar los resultados
  - Escoger algún caso para inicio
  - Usar vector de medias y añadir ruido aleatorio
- Muy sensible al número *k* de clústeres, elegidos previamente
  - Se suele usar el método del codo (*elbow method*)
  - También se puede aplicar agrupamiento jerárquico



- Generan descomposición jerárquica de un conjunto de objetos
- Aglomerativa: Aproximación *bottom-up*, cada objeto forma un grupo separado. Va uniendo objetos/grupos cercanos hasta que se unen todos o se alcanza condición terminación
- Divisiva: Aproximación *top-down*, empieza con todos los objetos en un grupo. En cada iteración cada grupo se divide, hasta ser grupos unitarios o se alcanza terminación
- Pueden ser basados en distancias, densidad o continuidad
- Sufren de que una vez se hace un paso (separación o unión), no se puede deshacer
- Esta rigidez simplifica la complejidad computacional, limitando el número de posibilidades a considerar

- Construirán un árbol en el que las hojas son elementos del conjunto de ejemplos, y el resto son subconjuntos
  - Se usa una representación visual llamada dendrograma
  - Se genera jerarquía de nodos, que permite obtener diferentes soluciones
  - Aglomerativa vs divisiva

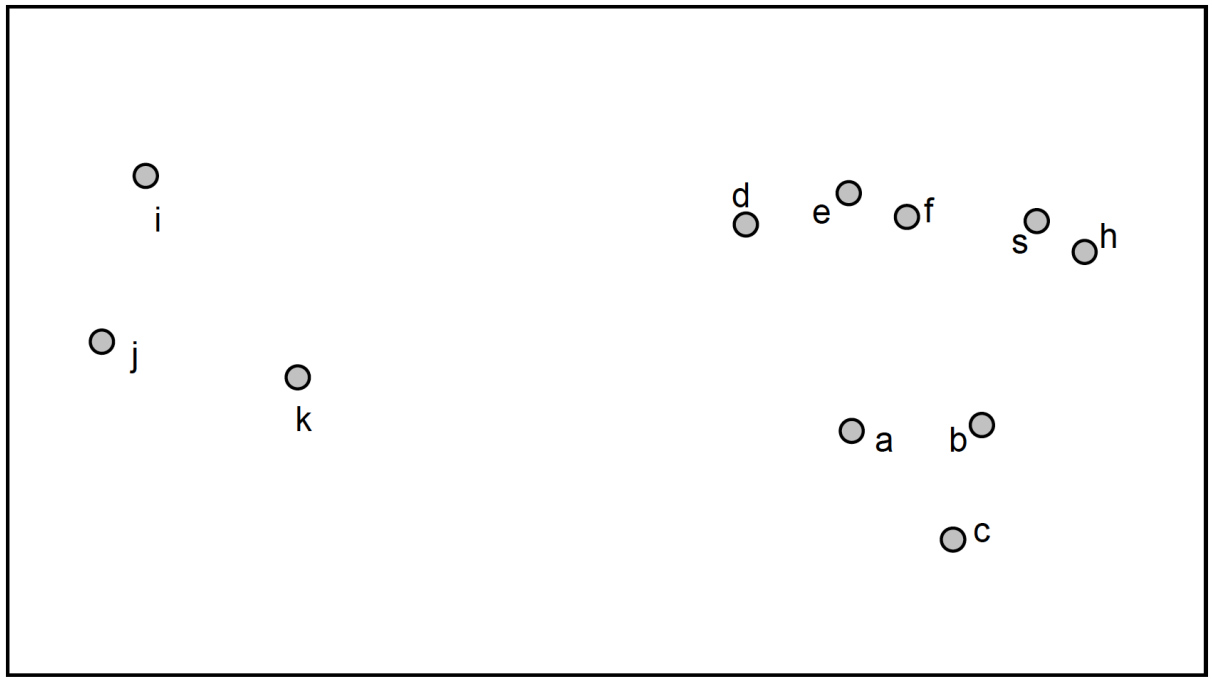


- Hay varios métodos para la elección de los grupos, pero el más común es que la distancia del enlace sea la menor
  1. Cada caso es el representante del grupo si sólo contiene dicho punto
  2. Se calculan las distancias dos a dos
  3. Elegir los dos grupos cuya distancia es menor
  4. Mezclar los grupos elegidos en el paso anterior
  5. Si hay más de un grupo ir a 2
- Las diferencias residen en la forma de calcular la distancia entre los grupos
  - **Enlace simple**: no se usan los representantes, se calculan distancias entre todos los puntos de ambos grupos y se coge la menor
  - **Enlace completo**: igual que el anterior, pero se toma la mayor
  - **Enlace en la media**: se toma como distancia la existente entre los centroides de los grupos

Unidad 3. Aprendizaje computacional
Método jerárquico: Ejemplo

UNIVERSIDAD DE MURCIA

- En primer lugar, se calculan las distancias dos a dos
- La distancia más corta es entre los puntos e y f
- Se genera el nuevo grupo unificado ef



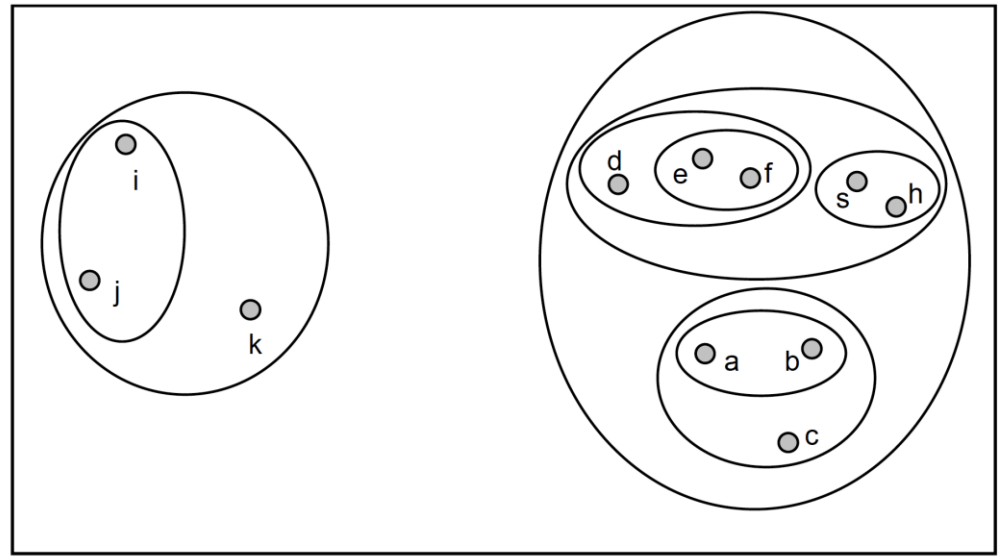
	a	b	c	d	e	f	g	h	i	j	k
a	-	2,4	4	9	7,8	7	8,9	8,7	27,6	28,3	21,4
b	-	-	4,3	11,4	9,1	7,2	7,3	7,4	30,8	32,6	26,8
c	-	-	-	13,6	12,2	11	12,6	10,5	31	31,9	25,2
d	-	-	-	-	4,8	6,4	10,9	12	21	23,7	17,6
e	-	-	-	-	-	1,7	6	7,4	26,3	28,7	22,5
f	-	-	-	-	-	-	4,5	5,6	27,3	30	23,6
g	-	-	-	-	-	-	-	1,8	31,9	34,3	28,8
h	-	-	-	-	-	-	-	-	33,4	35,4	28,8
i	-	-	-	-	-	-	-	-	-	4	6,8
j	-	-	-	-	-	-	-	-	-	-	6,9
k	-	-	-	-	-	-	-	-	-	-	-

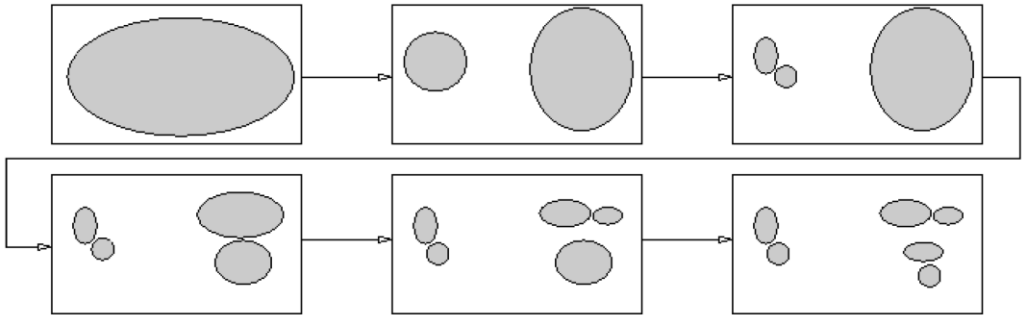
Unidad 3. Aprendizaje computacional
Método jerárquico: Ejemplo

UNIVERSIDAD DE MURCIA

- Ahora la distancia más corta es entre g y h, que también formarían grupo
- El árbol podría representarse por esta nube de puntos
- Se puede elegir el nivel para obtener la diferencia más clara

	a	b	c	d	ef	g	h	i	j	k
a	-	2,4	4	9	7,4	8,9	8,7	27,6	28,3	21,4
b	-	-	4,3	11,4	8,5	7,3	7,4	30,8	32,6	26,8
c	-	-	-	13,6	11,6	12,6	10,5	31	31,9	25,2
d	-	-	-	-	5,6	10,9	12	21	23,7	17,6
ef	-	-	-	-	-	5,2	6,4	26,8	29,1	22,9
g	-	-	-	-	-	-	1,8	31,9	34,3	28,8
h	-	-	-	-	-	-	-	33,4	35,4	28,8
i	-	-	-	-	-	-	-	-	4	6,8
j	-	-	-	-	-	-	-	-	-	6,9
k	-	-	-	-	-	-	-	-	-	-

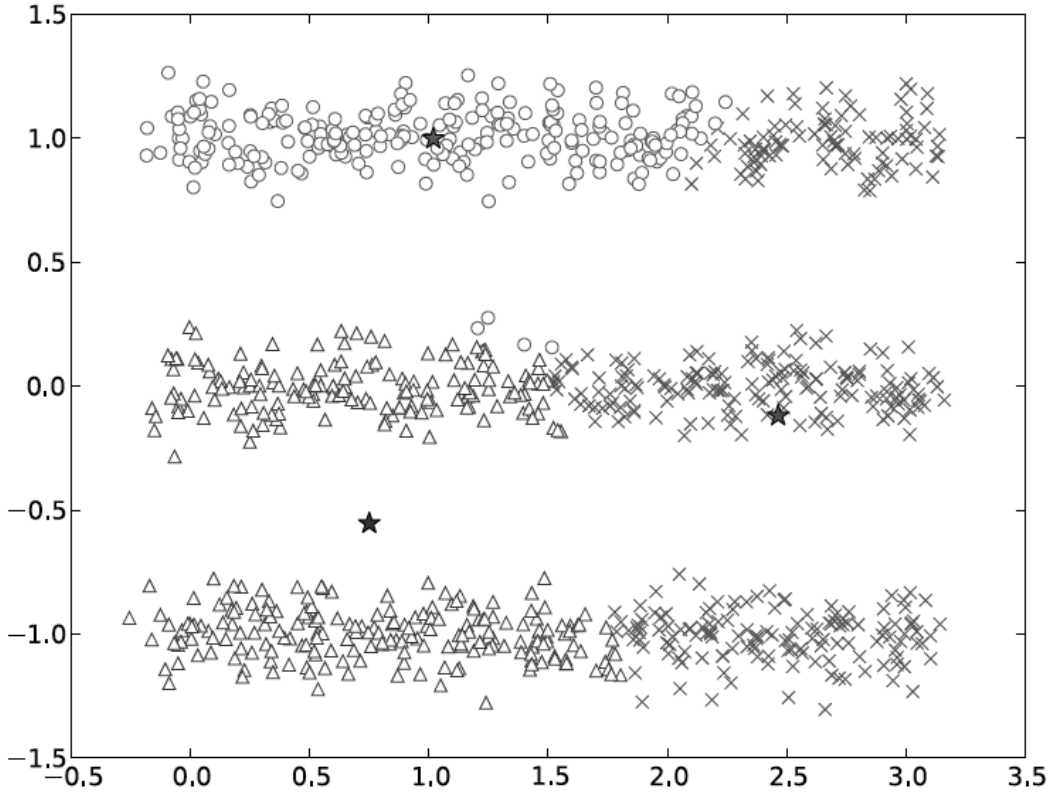




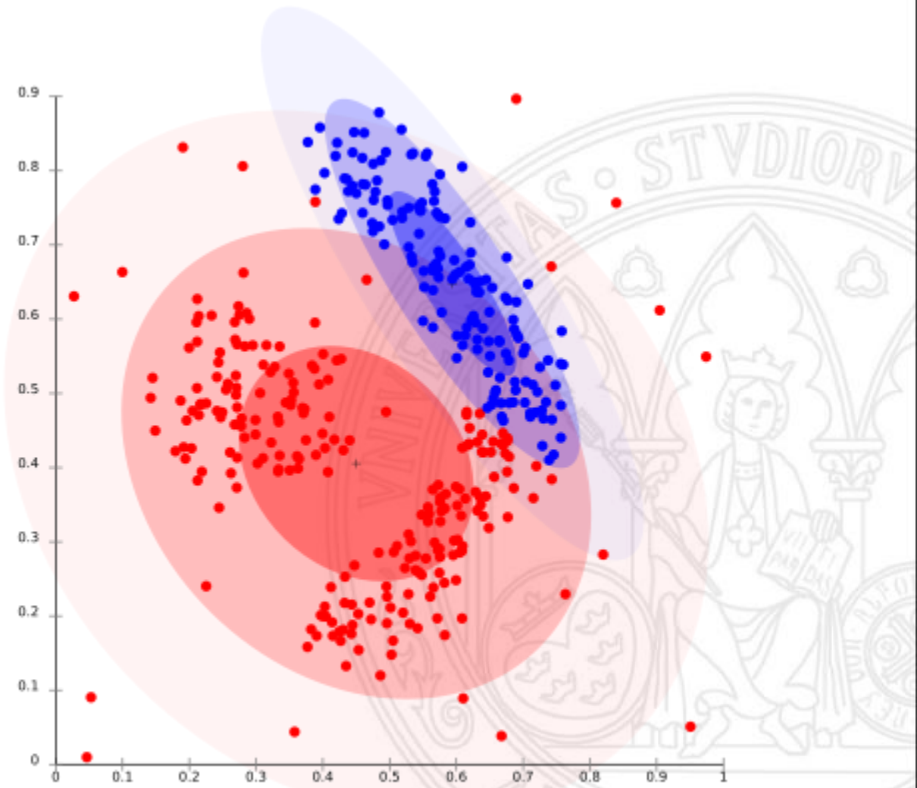
- Los métodos particionales se basan en distancias entre objetos
  - Llevan a encontrar grupos esféricos y complican encontrar otros tipos de formas
- Estos se basan en la noción de densidad
- Seguirán creciendo el grupo siempre y cuando la densidad (número de objetos) exceda un umbral
  - Son útiles para filtrar ruido y outliers, y descubrir grupos con formas arbitrarias
- Pueden ser usados para dividir grupos de forma particional o en jerarquías
  - Normalmente no usan aproximaciones difusas, pero también son posibles (lo veremos en NLP)

UNIVERSIDAD DE MURCIA

Problema con k-medias y distribuciones arbitrarias



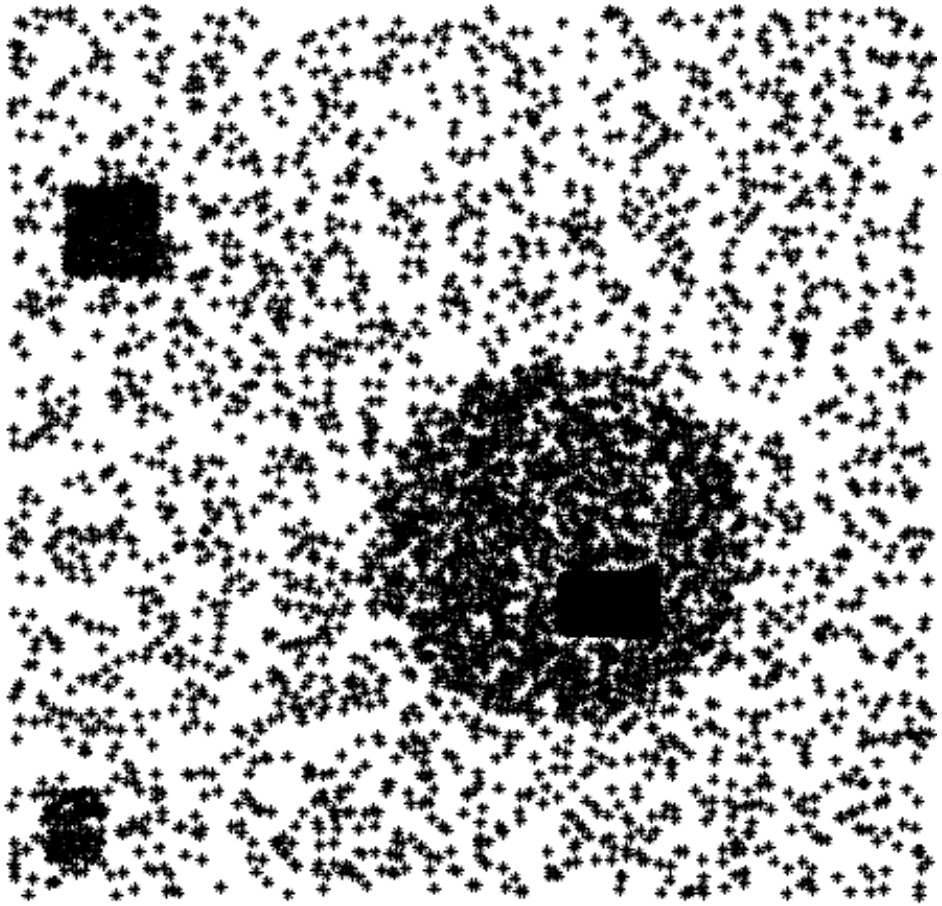
Beneficios de algoritmos basados en distribuciones



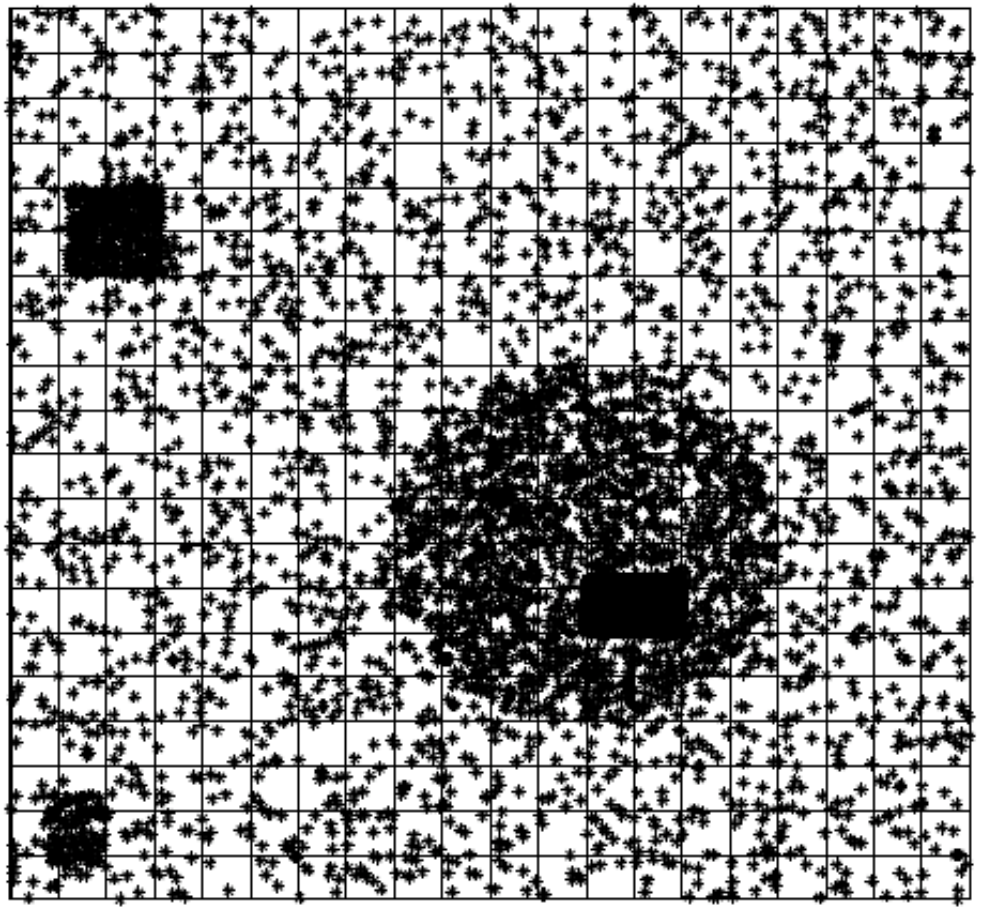
- Realizan una cuantización de un espacio en un número finito de celdas que forma una estructura en malla
- Todas las operaciones se realizan a nivel de malla (espacio cuantizado)
- La principal ventaja es el tiempo de procesado, que es independiente del número de objetos, y sólo depende de las celdas por dimensión
- Es una aproximación eficiente en problemas de minería espaciada

Ejemplo: métodos basados en malla

Datos iniciales



Malla uniforme cuadrada



Unidad 3. Aprendizaje computacional
Comparativa métodos

UNIVERSIDAD DE MURCIA

Método	Características generales
Métodos particionales	- Encuentran grupos mutuamente exclusivos con forma esférica - Basados en distancias - Pueden usar la media o medoid para representar el centro del grupo - Efectivos para juegos de datos pequeños/medianos
Métodos jerárquicos	- El agrupamiento es una descomposición jerárquica (múltiples niveles) - No puede corregir errores en uniones o divisiones - Puede incorporar otras técnicas como micro-agrupamiento o considerar objetos como “linkages” (uniones)
Métodos basados en densidades	- Pueden encontrar grupos con forma arbitraria - Los grupos son regiones densas separados por regiones poco densas - Densidad del grupo: Debe de haber un mínimo de puntos en el “vecindario” - Pueden filtrar los valores atípicos
Métodos basados en mallas	- Usan una estructura de datos en forma de malla - Tiempo de procesado rápido (no depende en el número de objetos, sino en el tamaño de la malla)

UNIVERSIDAD DE MURCIA

Unidad 3. Aprendizaje computacional

EVALUACIÓN DE RESULTADOS DE AGRUPAMIENTO

¿Qué resultado de agrupamiento es mejor?

- Dado un resultado de agrupamiento específico, ¿cómo analizamos su calidad y si se asemeja a la realidad?
- De forma similar al aprendizaje supervisado, en agrupamiento se prueban distintas alternativas (dos parámetros clave):
  - La medida de distancia o similitud utilizada
  - El número de clústeres que se le introduce al algoritmo
- Se debe elegir la mejor configuración de entre varias pruebas
  - Análogo a lo que hacíamos en aprendizaje supervisado
- Habrá que utilizar métricas específicas para medir la calidad de un agrupamiento en concreto
  - ¿Cómo de buena es la estructura revelada con respecto a la real que tienen los datos?
- El éxito de la métrica de calidad es dependiente de la técnica y las características de los datos

- Los grupos encontrados por una técnica de agrupamiento, deben satisfacer dos propiedades:
  - Compactación: indica como de cerca se encuentran los elementos asignados a cada clúster
    - Mayor varianza implica menor compactación
  - Separabilidad: indica como de diferentes son los clústeres entre sí
- Como se ha comentado previamente, una forma de medirlo son las distancias intra y intercluster, con pequeñas distancias intracluster y grandes intercluster

Índice de la silueta (Silhouette index)

- Para cada caso i, perteneciente al clúster C_j , se define el índice $s(i)$ mediante dos índices
- El índice $a(i)$ será la distancia media entre i y todos los elementos de mismo clúster
  - $a(i) = \frac{1}{|C_i|} \sum_{\forall j \in Ci \wedge j \neq i} d(i,j)$
  - Si el clúster sólo tuviera un elemento, entonces $s(i) = 0$
- El índice $b(i)$ será la distancia media entre i y todos los elementos del clúster más cercano
  - $\mathrm{b}(i) = \frac{1}{|C_k|} \sum_{\forall j \in C_k} d(i,j)$ $[con\ k = min_{k \neq i}\left(d(i, C_k)\right)$
- Entonces, se puede definir el índice de la silueta para un caso i como:
  - $s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$

- El índice varía entre [-1, 1]
  - Un valor cercano a 1 de $s(i)$ en una observación i, indica que está bien agrupada
  - Un valor cercano a 0 de $s(i)$ en una observación i, indica que está entre dos clústeres
  - Un valor negativo de $s(i)$ en una observación i, indica que está mal agrupada
- Hay otras diversas medidas de evaluación:
  - el índice de Dunn
  - el índice de separación
  - el gap
  - el índice Davies-Bouldin
  - Todas ellas basadas en conceptos similares relacionados con la distancia intracluster e intercluster