Clustering

Clustering

Objetivo

- Agrupar objetos similares entre sí que sean distintos a los objetos de otros agrupamientos (clusters).

Aprendizaje no supervisado

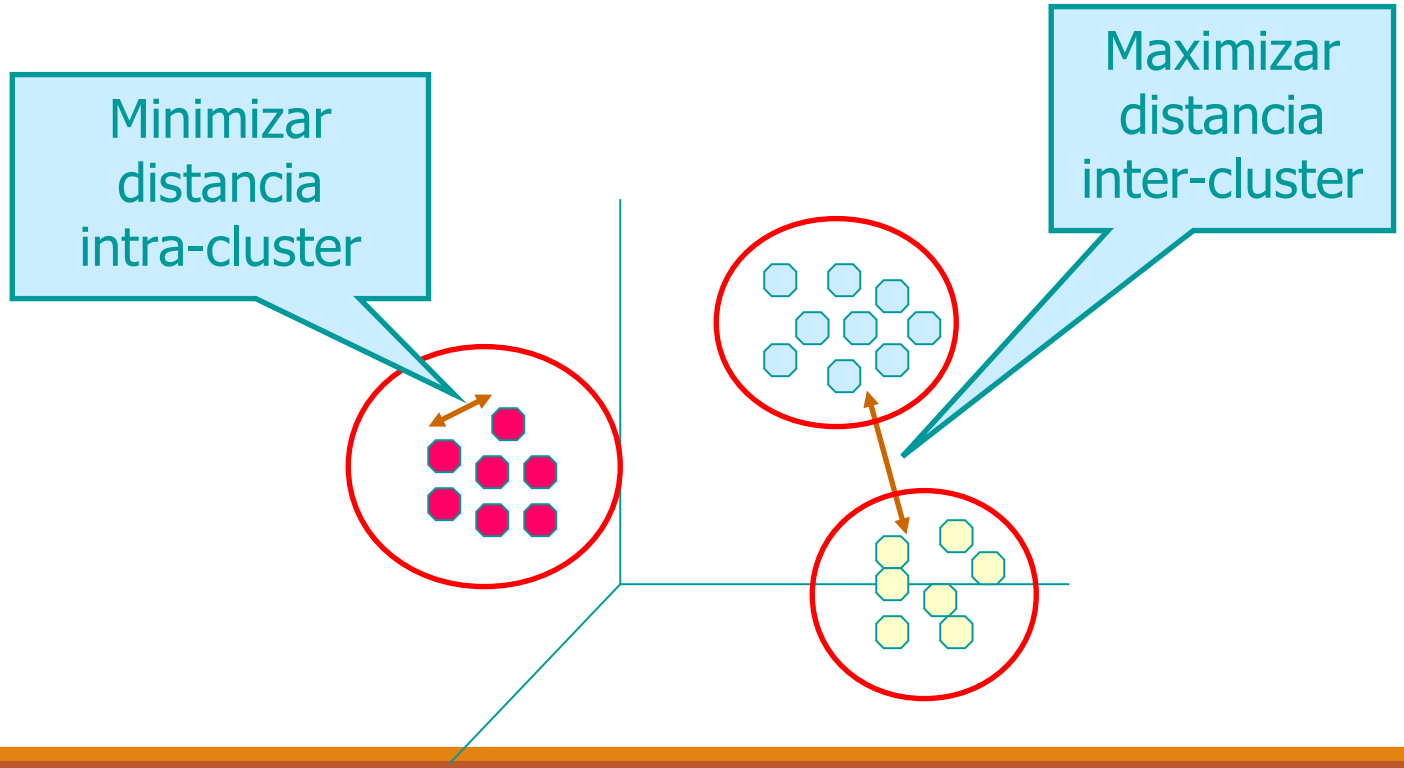
- No existen clases predefinidas

Los resultados obtenidos dependen de

- El algoritmo de agrupamiento seleccionado
- El conjunto de datos disponible
- La medida de similitud utilizada para comparar objetos

Clustering

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



Clustering

Medidas de Similitud

Usualmente, se expresan en términos de distancias:

d(i,j) > d(i,k) indica que el objeto i es más parecido a k que a j

La definición de la métrica de similitud/distancia dependerá del tipo de dato y de la interpretación semántica que se haga

En otras palabras, la similitud entre objetos es **subjetiva**

Clustering

Métodos de Particionamiento

Métodos Jerárquicos

Basados en Densidad

Basados en Grillas

Basados en Modelos

Clustering

Métodos de Particionamiento

El conjunto de datos es particionado en un número pre-especificado de agrupamientos K

Iterativamente se va reasignando las observaciones a los agrupamientos hasta que algún criterio de parada (función a optimizar) se satisface (suma de cuadrados dentro de los clusters sea la más pequeña)

Ejemplos: K-means, PAM, CLARA, SOM, Conglomerados basados en modelos de mezclas gausianas, Conglomerados difusos.

Clustering

Metodos Jerárquicos

El conjunto de datos es particionado en un número

En estos algoritmos se generan sucesiones ordenadas (jerarquias) de agrupamientos

Pueden unir agrupamientos pequeños en mas grandes o dividir grandes clusters en otros mas pequenos

La estructura jerárquica es representada en forma de un árbol y es llamada Dendograma

Clustering

Metodos Jerárquicos

Se dividen en dos tipos:

Algoritmos jerárquicos aglomerativos (bottom-up, inicialmente cada instancia es un cluster)

- AGNES (Agglomerative Nesting)

Algoritmos jerárquicos divisivos (top-down, inicialmente todas las instancias estan en un solo cluster

- DIANA (DIvisive ANAlysis Clustering)

K-Means

Algoritmo k-means (MacQueen, 1967)

Particionamiento

El objetivo es minimizar la distancia (dis-similaridad) de los elementos dentro de cada cluster y maximizar la distancia de los elementos que caen en diferentes clusters

K-Means

Dado un conjunto de datos S y k número de clusters a formar

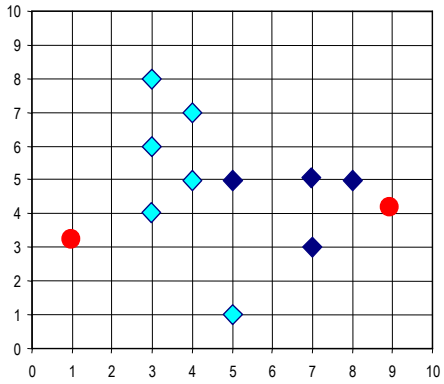
1. Seleccionar los centroides iniciales de los K clusters c_1, c_2, ..., c_K, al azar entre las observaciones
2. Asignar cada observación x_i de S al cluster C(i) cuyo centroide c_i está más cerca de x_i, es decir, $C(i)=\text{argmin}_{1<=k<=K}||x_i-c_k||$
3. Para cada uno de los clusters se recalcula su centroide basado en las observaciones que están contenidos en el cluster, minimizando la suma de cuadrados dentro del cluster, es decir:

$$WSS = \sum_{k=1}^{K} \sum_{C(i)=k} || \mathbf{x}_i - \mathbf{c}_k ||^2$$

Volver al paso 2 mientras que no hayan cambios

K-Means

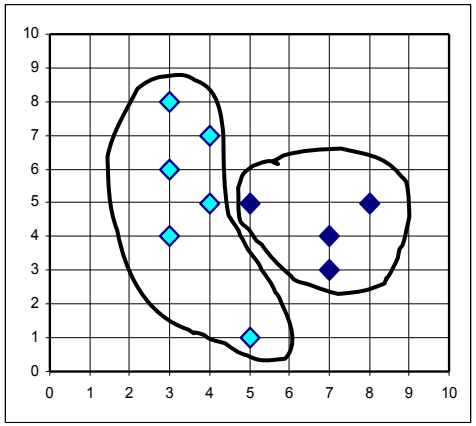
Ejemplo



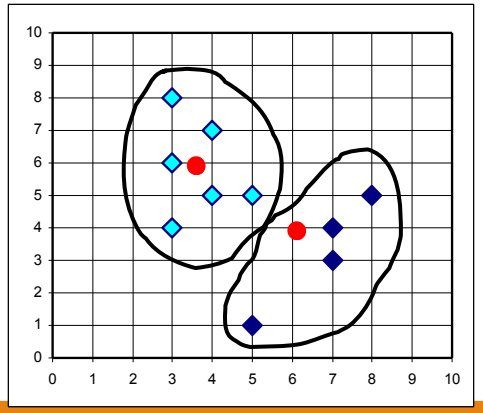
K=2

Escoger arbitrariamente K puntos como centroides

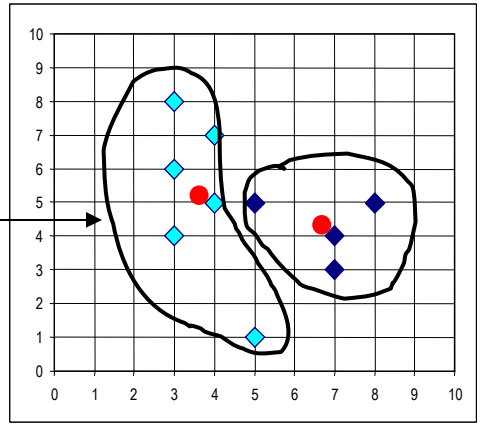
Asignar cada instancia al cluster más cercano



reasignar

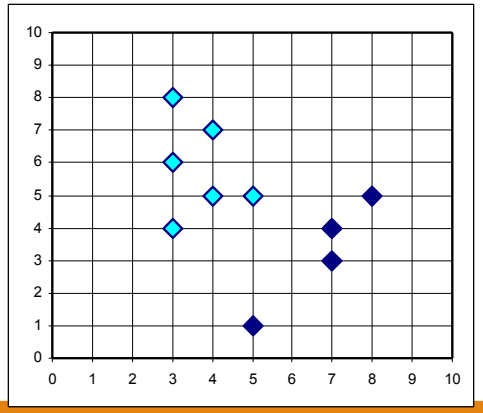


Actualizar los centroides



Actualizar los centroides

reasignar



K-Means

Alternativas para los k centroides iniciales

Usar las primeras k bservaciones

Elegir aleatoriamente k observaciones

Tomar cualquier partición al azar en k clusters y calculando sus centroides

K-Means

No se satisface el criterio de optimización global, ev. sólo produce un óptimo local

Rápido

Puede trabajar bien con datos faltantes

Es sensible a “outliers”

PAM - **Particionamiento alrededor de medoides**

Introducido por Kauffman y Rousseauw, 1987

MEDOIDES, son instancias representativas de los clusters que se quieren formar

Para un número especificado de clusters K, el procedimiento PAM está basado en la búsqueda de los K MEDOIDES, M = (m_1, . . . ,m_K) de todas las observaciones a clasificar

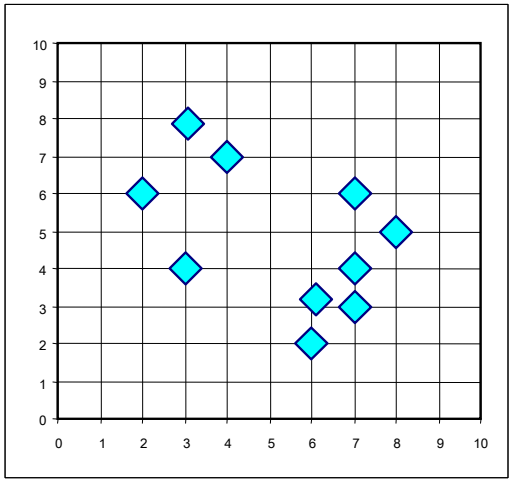
Para encontrar M hay que minimizar la suma de las distancias de las observaciones a su Medoide más cercana

$$M^* = \arg\min_M \sum_i \min_k d(x_i, m_k)$$

d es una medida de dissimilaridad

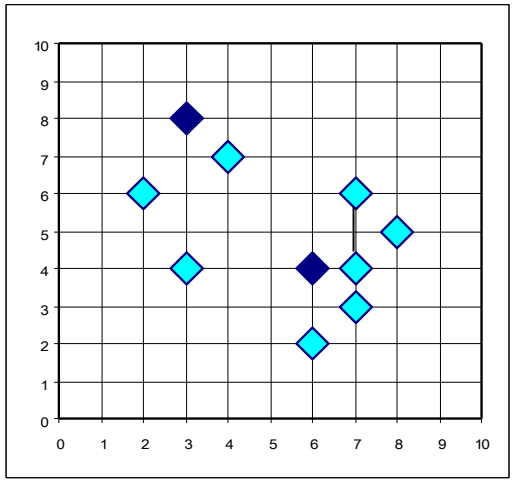
PAM

Ejemplo

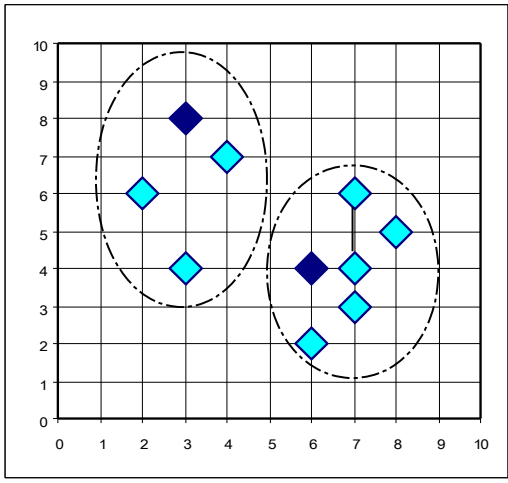


K=2

Elegir k datos como medoides



Asignar las instancias restantes a su medoide más cercano



Clustering

Medidas de Similitud

Con número de clases desconocido

- Método adaptativo
- Algoritmo de máxima distancia (Batchelor & Wilkins)

Con número de clases conocido

- Algoritmo de las k-means
- Algoritmo GRASP
- Algoritmo de agrupamiento secuencial
- Algoritmo ISODATA

Métodos basados en grafos

- Algoritmo basado en la matriz de similitud

Algoritmo Adaptativo

Ventajas

- Útil cuando no se conoce de antemano el número de clases del problema (número de clusters desconocido)
- Simplicidad y eficiencia

Desventajas

- Dependencia del orden de presentación (comportamiento sesgado por el orden de presentación de las observaciones)
- Presupone agrupamientos compactos separados claramente de los demás (puede no funcionar adecuadamente en presencia de ruido)

Algoritmo Adaptativo

Inicialización

- Se forma un agrupamiento con la primera observación del conjunto de datos

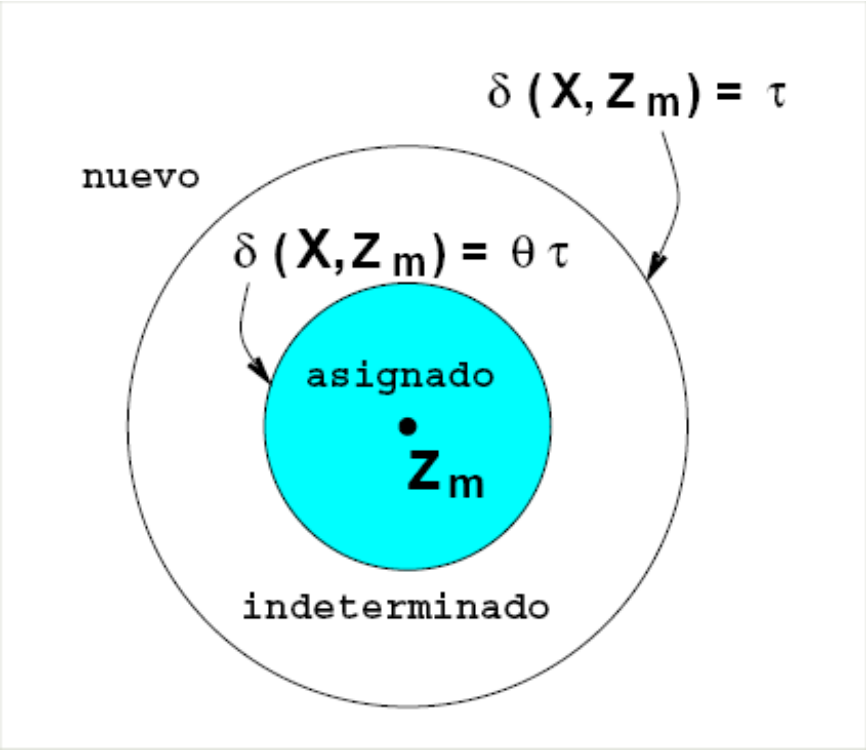
Mientras queden patrones por asignar

- La siguiente observación se asigna a un cluster si la distancia al centroide del cluster no supera un umbral 
- En caso contrario, se crea un nuevo agrupamiento si la distancia de la observación al cluster más cercano está por encima de 

Algoritmo Adaptativo

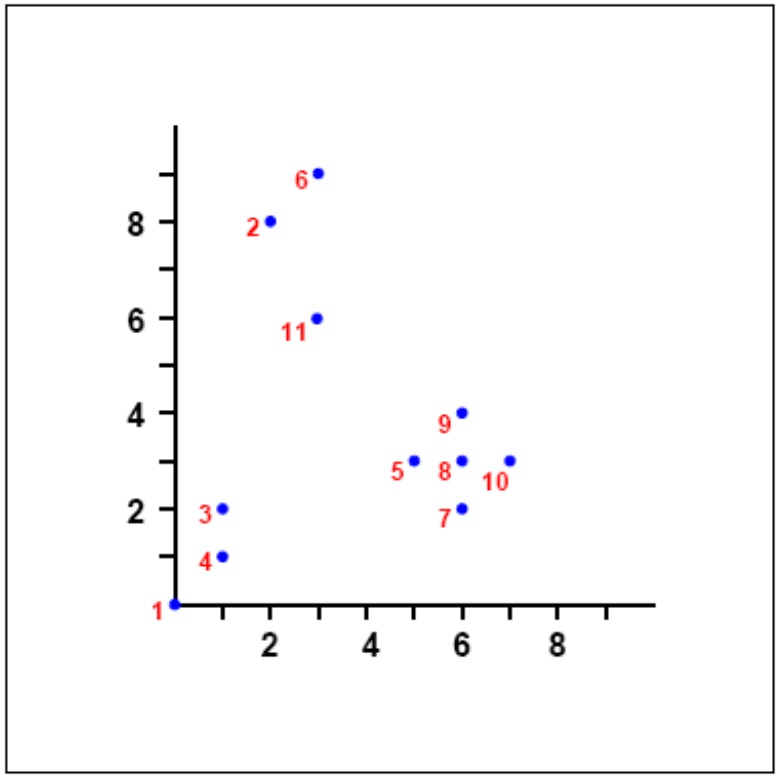
Este algoritmo incluye una clase de rechazo

- Algunas observaciones no son clasificadas



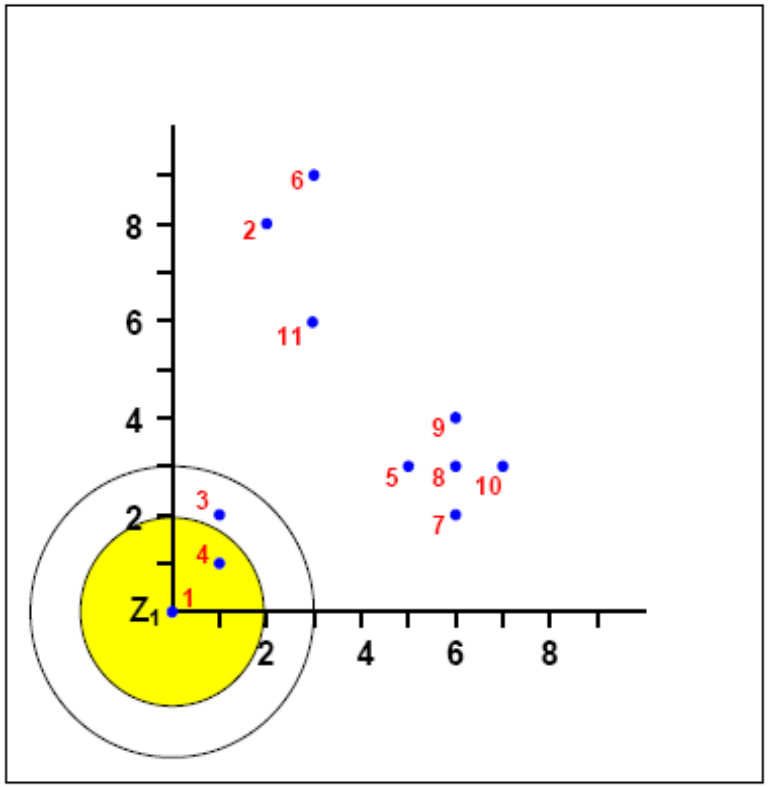
Algoritmo Adaptativo

Ejemplo



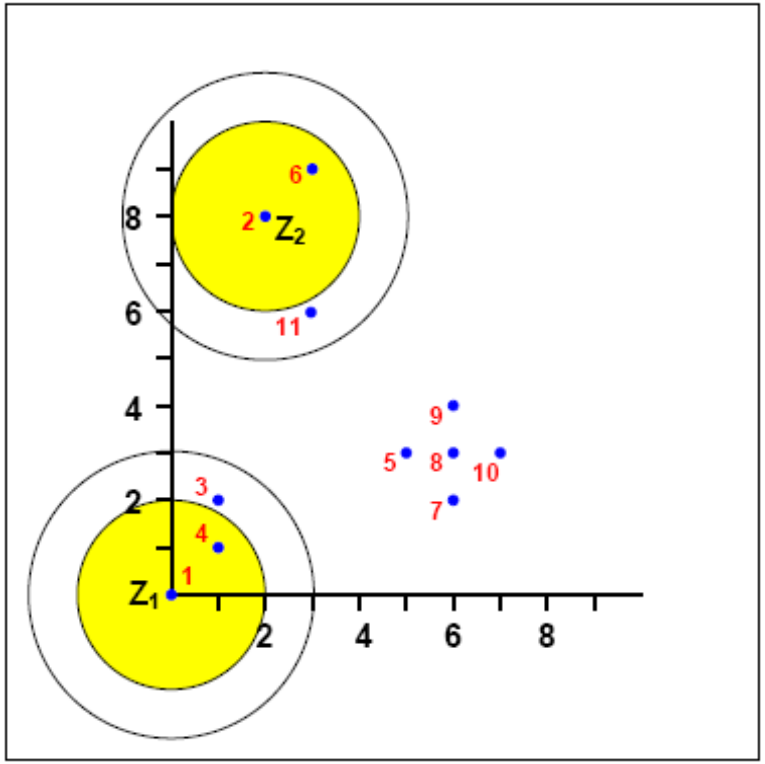
Algoritmo Adaptativo

Ejemplo



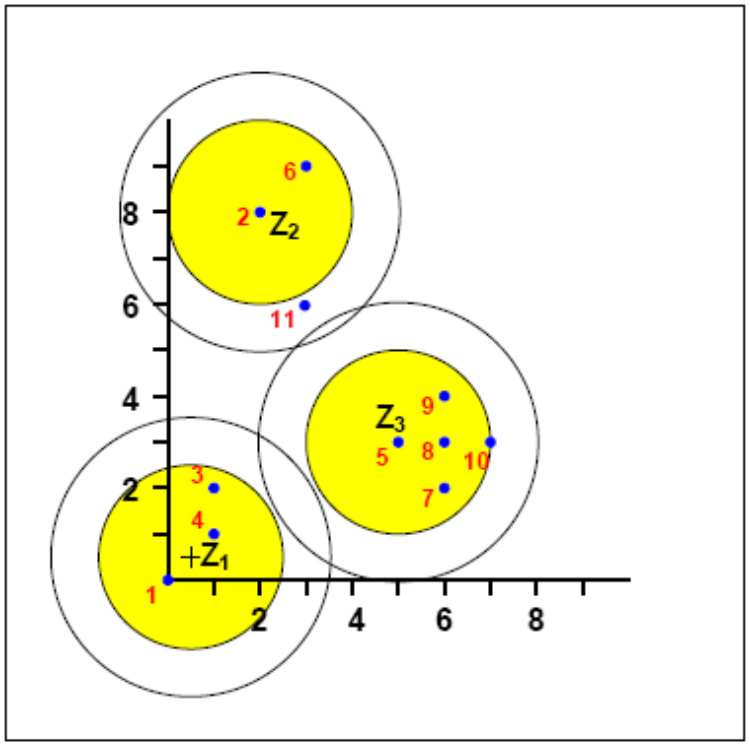
Algoritmo Adaptativo

Ejemplo



Algoritmo Adaptativo

Ejemplo



Algoritmo Adaptativo

Parámetros

Umbral de distancia 

Umbral de distancia utilizado para crear nuevos agrupamientos

Fracción ₡

Fracción del umbral de distancia que determina total confianza (utilizada para determinar si un patrón se le asigna a un cluster o no)

Batchelor & Wilkins

Características principales

Ventajas

Útil cuando no se conoce de antemano el número de clases del problema (número de clusters desconocido)

Un único parámetro

Desventajas

Sensibilidad al valor del parámetro

Batchelor & Wilkins

Primer agrupamiento:

- Observación elegida al azar

Segundo agrupamiento:

- Observación más alejada del primer cluster
- Mientras se creen nuevos clusters, obtener la observación más alejado de los clusters existentes (máximo de las distancias mínimas de las observaciones a los clusters)
- Si la distancia de la observación escogido al conjunto de clusters es mayor que una fracción ***f*** de la distancia media entre los clusters, crear un cluster con la observación seleccionado

Asignar cada observación a su agrupamiento más cercano

Batchelor & Wilkins

Primer agrupamiento:

- Observación elegida al azar

Segundo agrupamiento:

- Observación más alejada del primer agrupamiento
- Mientras se creen nuevos agrupamientos, obtener el observación más alejado de los agrupamientos existentes (máximo de las distancias mínimas de los patrones a los agrupamientos)
- Si la distancia de la observación escogido al conjunto de agrupamientos es mayor que una fracción ***f*** de la distancia media entre los agrupamientos, crear un agrupamiento con la observación seleccionado

Asignar cada observación a su agrupamiento más cercano