M2.2.2 Modelos Supervisados y No Supervisados

Programa Big Data y Business Intelligence

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/

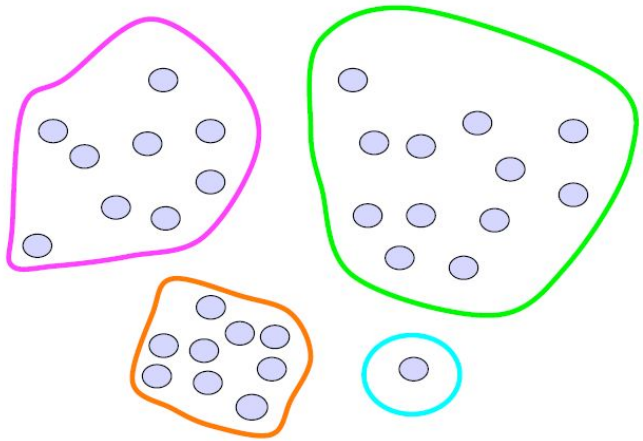
Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

Clustering

- Definición e idea del grupo
- Medidas de calidad del análisis cluster
- Clustering Jerárquico
  - Divisivo
  - Aglomerativo
- Midiendo distancias entre clusters
- El algoritmo k-means
- Estableciendo el número apropiado de clusters

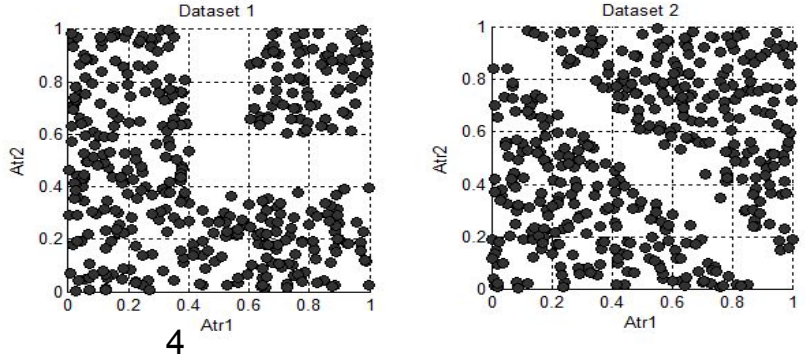
Definición

- Los métodos de clustering buscan agrupar un conjunto de datos en “clusters” (grupos), según cierta medida de distancia
  - Datos dentro del mismo cluster deben estar cerca los unos de los otros
  - Datos de clústeres diferentes deben estar lejos los unos de los otros



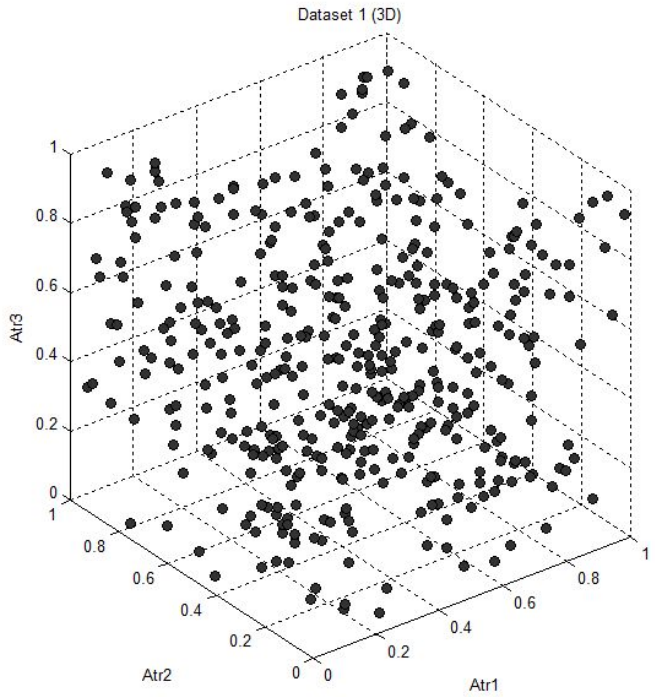
Una aclaración

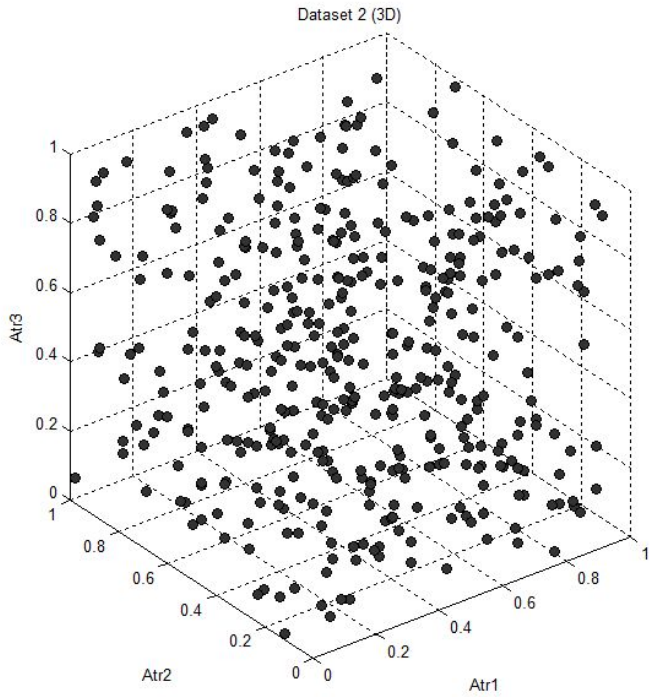
- Los clusters (en la mayoría de los casos) no se ven a simple vista
- Para “demostrarlo”, dos ejemplos
  - Genero 2 conjuntos de datos con 2 atributos - 500 puntos
    - Conjunto 1 -> elimino aquellos puntos en los que no se cumple que:
    - Conjunto 2 -> elimino aquellos puntos en los que no se cumple que:
  - Obtengo 2 datasets con 2 clusters “fácilmente diferenciables”



Una aclaración

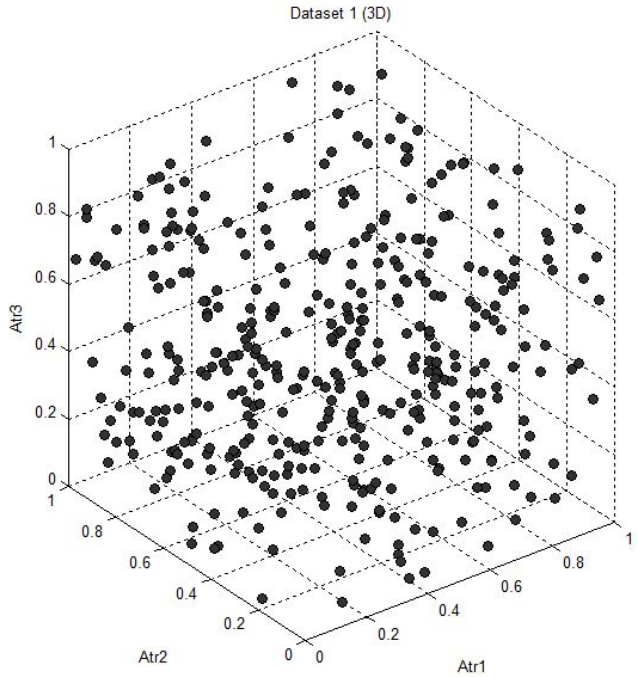
- Pero si hago dos datasets, siguiendo el mismo razonamiento, pero con 3 atributos...

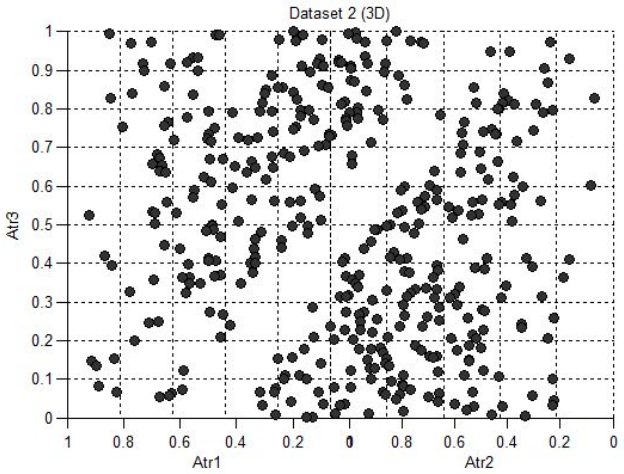




Una aclaración

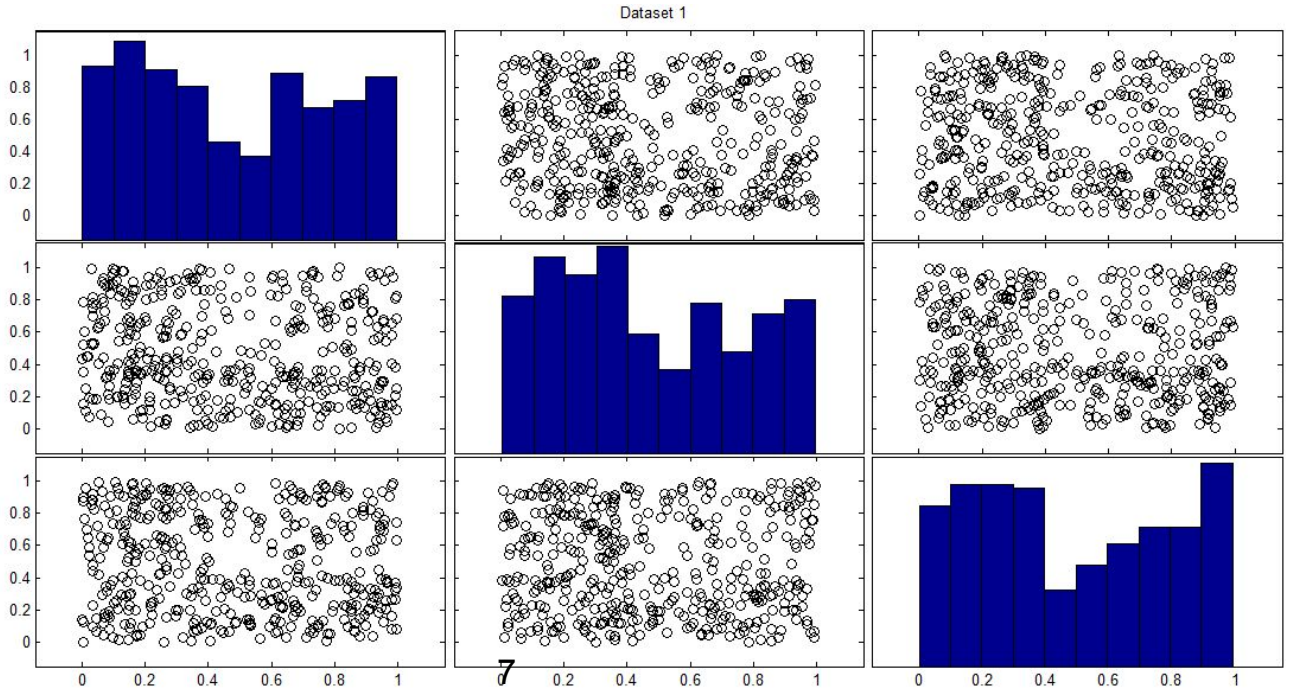
- En algunos (pocos) casos puedo “rotar” el dataset como si fuera un cubo hasta encontrar la separación → En otros casos, no





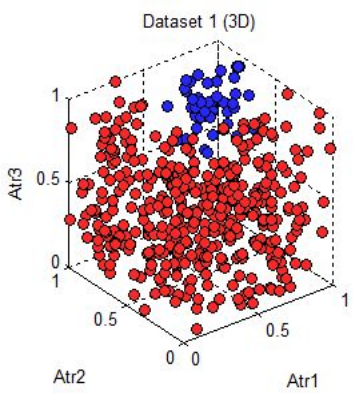
Una aclaración

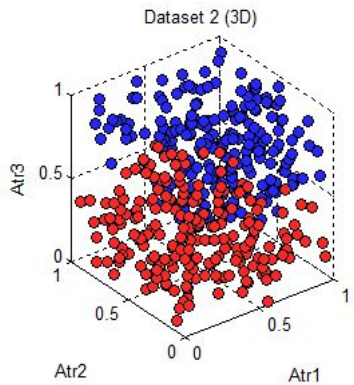
- También puedo mostrar los datos como un conjunto de dibujos de dos dimensiones
  - Es lo que se conoce como un “scatter plot”
- ¿Me ayuda?



Una aclaración

- Pero los dos clusters “están ahí”
  - (aunque no los veamos)
- ¿Y si tenemos 4 atributos?
  - Los ejemplos para ilustrar el análisis cluster trabajan en dos dimensiones para poder “ver” los clusters,
  - Y así comparar el resultado de un método con el resultado de un “experto”
- ¿Y si tenemos 5, 6, 7, ...?





¿Qué es el análisis cluster?

- Un cluster es una agrupación de muestras de datos
  - Similares a aquellas pertenecientes al grupo
  - Diferentes de aquellas no pertenecientes al grupo
- Objetivos
  - Comprender la estructura de los datos
  - Encontrar similitudes entre datos
  - Agrupar elementos
  - Es No-Supervisado → no hay “clases” definidas

¿Qué es el análisis cluster?

- Conjunto de técnicas que agrupan objetos en grupos o clusters
- ¿Cuántos grupos?
  - Grupos o clusters no definidos a priori.
  - Diferencia con los métodos supervisados.
- ¿Cómo buscarlos? Los objetos dentro de un cluster deben ser
  - Similares o cercanos entre sí
    - (gran similaridad **intra-clase**)
  - Diferentes o alejados a los objetos de otro cluster
    - (baja similaridad **inter-clase**)

¿Qué es el análisis cluster?

- Aplicaciones típicas
  - Comprender los datos
    - Nos permitirá ver si los datos están agrupados o no
    - Ganaremos conocimiento sobre nuestros datos
    - Nos permitirá tomar decisiones más acertadas
  - Etapa de preprocesamiento
    - Para beneficiar en otra tarea del ciclo de vida de los datos.
    - Calidad de datos: si hay un cluster de datos que son similares entre sí porque comparten cierto valor de atributo (asignación de valores perdidos)

Aplicaciones

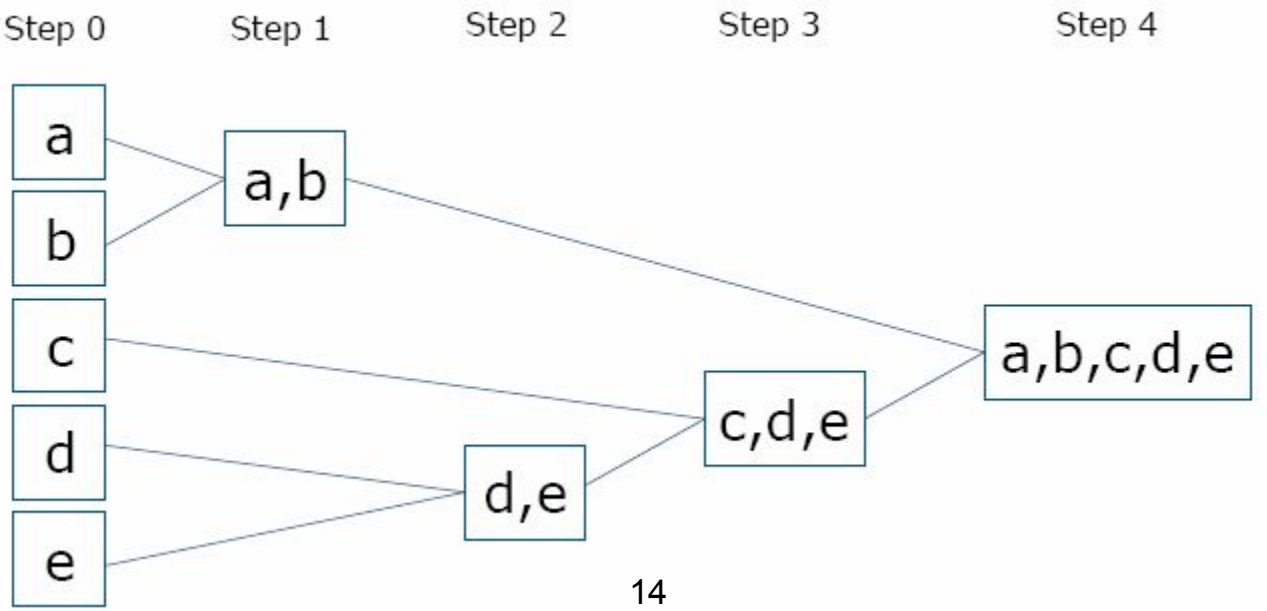
- Marketing
  - Ayudar a identificar grupos de clientes, para enfocar campañas específicas
- Seguros
  - Identificar grupos o características comunes dentro de los asegurados que reclaman costes altos
- Planificación urbana
  - Identificar grupos de viviendas similares en función del tipo, valor, área geográfica…

¿Cuándo es bueno?

- Una buena agrupación está compuesta por clusters con
  - Alta similitud intra-cluster
    - Elementos dentro del mismo cluster
  - Baja similitud inter-cluster
    - Elementos de clusters diferentes
- Métricas de evaluación de la calidad
  - Medidas de similitud inter/intra cluster
  - Inspección manual
  - Comparación con “etiquetas” pre-diseñadas
- Distancias
  - Por ello, es recomendable normalizar

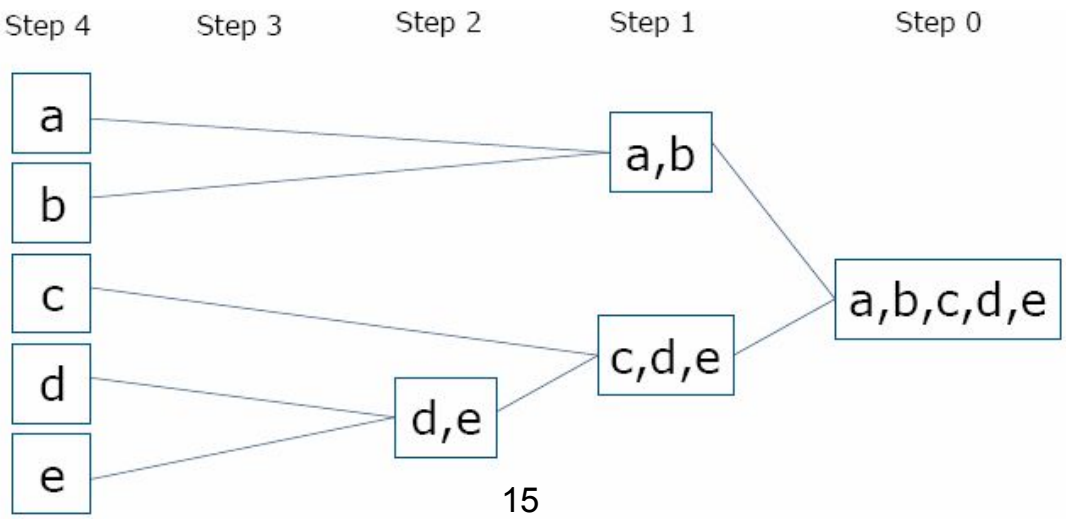
Clustering Jerárquico

- Supón que tienes 5 elementos {a,b,c,d,e}
  - Inicialmente, consideramos cada uno, por sí mismo, un cluster
  - Entonces, en cada paso, tomamos los clústeres más similares entre sí, y los agrupamos en un nuevo cluster



Clustering Jerárquico

- Supón que tienes 5 elementos {a,b,c,d,e}
  - Inicialmente, consideramos un único cluster con todos los elementos dentro
  - Entonces, en cada paso, partimos un cluster para mejorar la distancia intra-cluster, hasta que todos los elementos son clusters independientes



Clustering Jerárquico

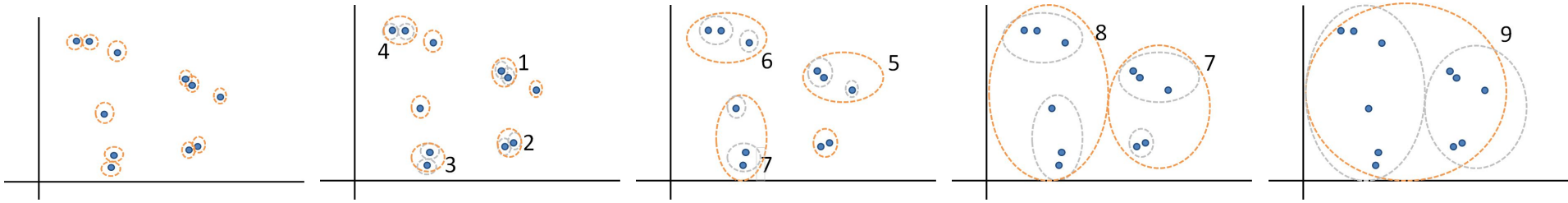
- Esos son los dos enfoques de clustering jerárquico, denominados
  - Aglomerativo → comenzamos con clústeres individuales, que vamos uniendo entre sí
  - Divisivo → comenzamos con un sólo cluster que vamos dividiendo hasta que todos los elementos pertenecen a clusters independientes

Clustering Jerárquico

- Fortalezas
  - No necesita que se establezca el número deseado de clústeres
  - Se pueden obtener divisiones en cualquier número de clústeres “cortando” el dendograma en el nivel apropiado
  - Pueden corresponderse con taxonomías reales
  - Utilizan una matriz de distancias o proximidad, para unir o dividir clústeres según ésta

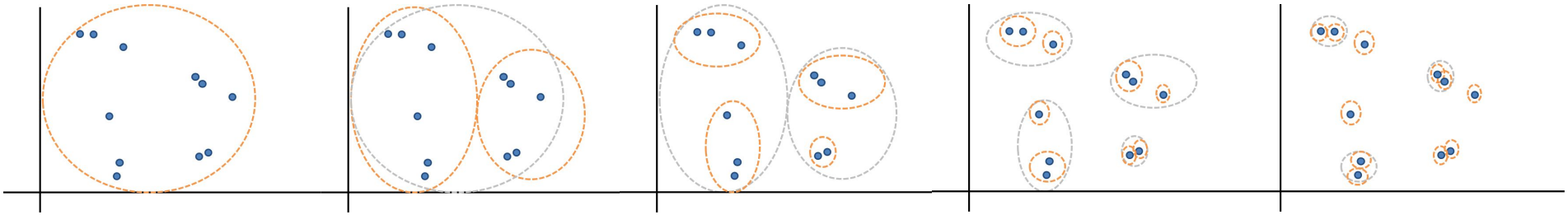
Clustering Aglomerativo

- Pasos
  - Comenzamos con cada elemento siendo un cluster independiente
  - Calcular la matriz de proximidad
  - Repetir hasta que sólo quede un cluster
    - Unir los dos clusters más cercanos
    - Actualizar la matriz de proximidad



Clustering Divisivo

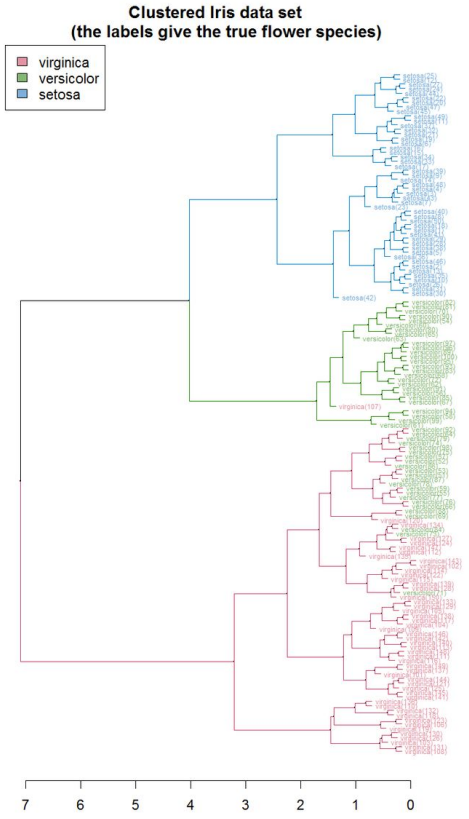
- Pasos
  - Comenzamos con todos los elementos en el mismo cluster
  - Repetir mientras se pueda
    - Utilizamos un método (otro) de clustering para dividir los elementos de cada cluster en 2 (o más) clusters



Clustering Jerárquico

- Con el clustering jerárquico podemos obtener tantas divisiones como elementos haya

- ¿Cuál escogemos?

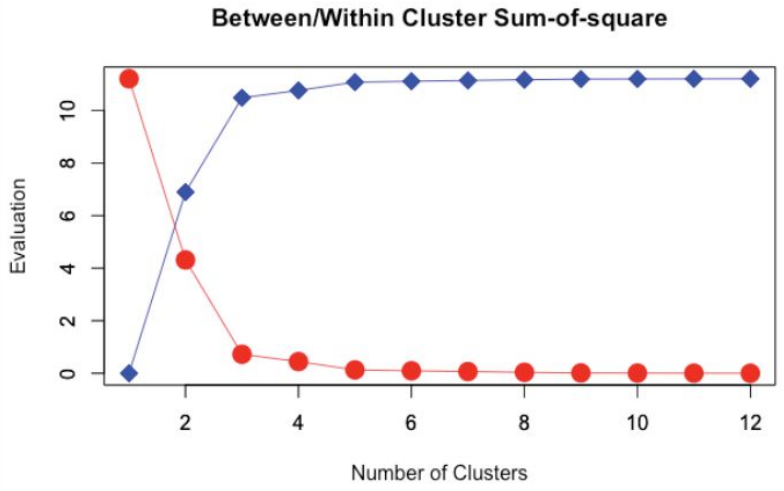


Eligiendo el número de clusters

- Withing-cluster sum of squares (WSS)
  - Distancia de los elementos de un cluster a su centroide
- Between-cluster sum of squares (BSS)
  - Distancia entre centroides de clusters
- Dibujar WSS y BSS y buscar el punto con cambio significativo

$$\mathrm{WSS}(C) \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \mu_i)^2$$

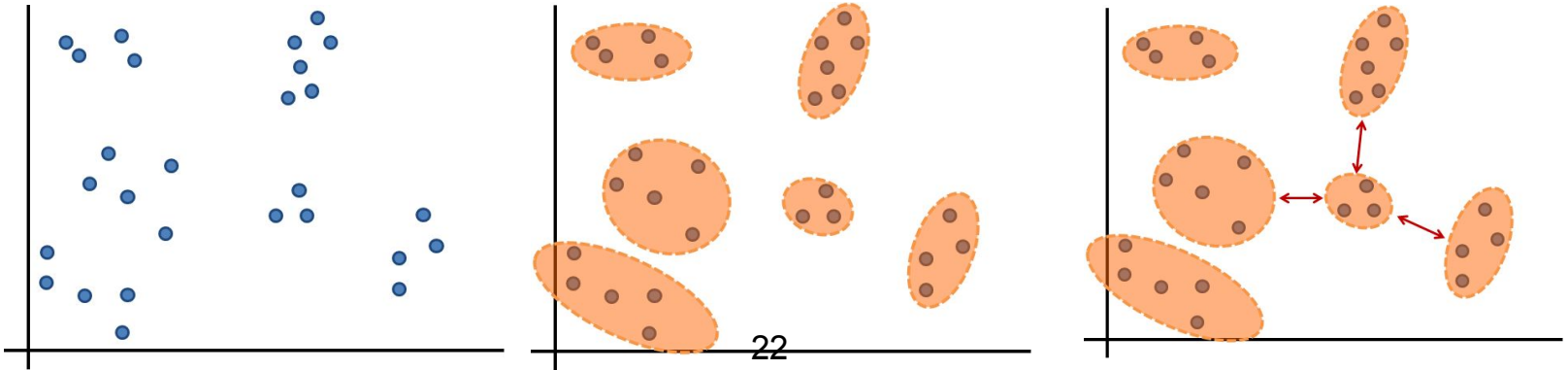
$$\mathrm{BSS}(C) \sum_{i=1}^{k} |C_i| d(\mu, \mu_i)^2$$



21

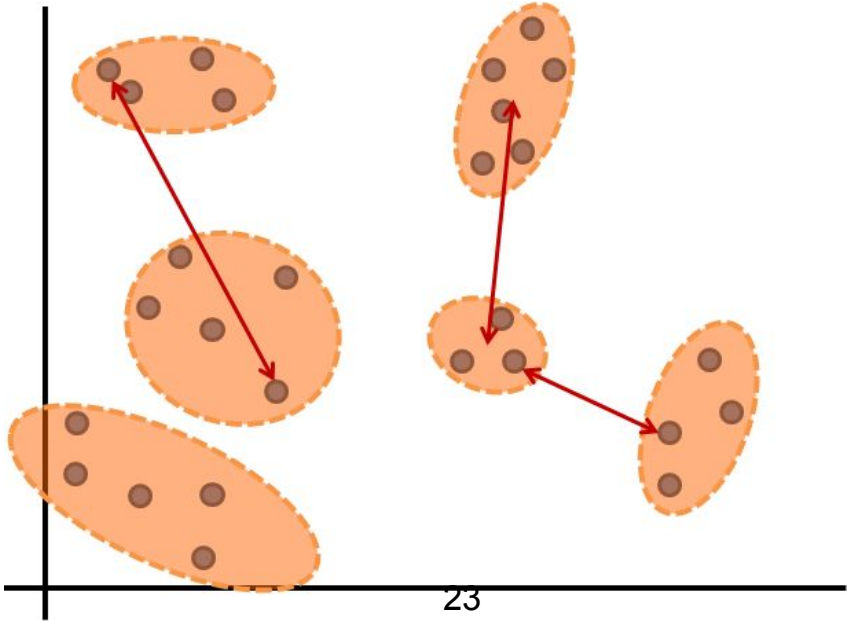
Midiendo Distancias (Cluster)

- Si comenzamos con N clusters
  - Tendremos una matriz de distancias entre puntos
  - Empezaremos a agrupar
  - Y a actualizar esa matriz de distancias
  - Y en algún punto, esa matriz de distancias contendrá distancias entre clusters,
  - ¿Cómo la calculamos?



Midiendo Distancias (Clusters)

- Medidas típicas:
  - Mínima distancia entre un elemento de un cluster y otro
  - Máxima distancia entre un elemento de un cluster y otro
  - Distancia promedio entre los elementos de los clusters
  - Distancia entre los **centroides** de los clusters
  - …

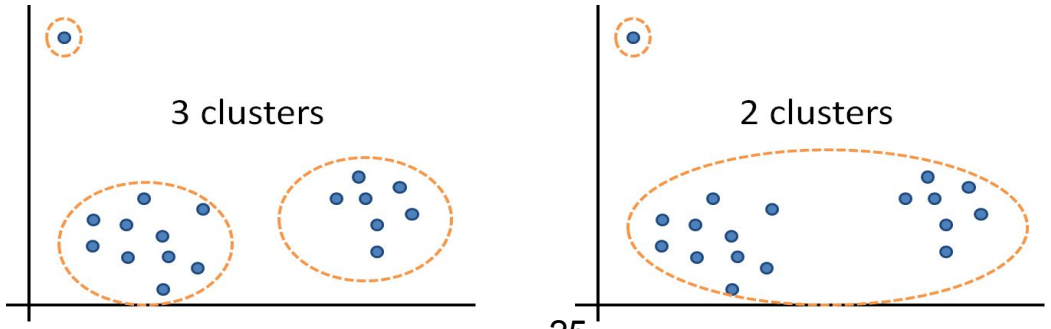


¿Cómo los representamos?

- Para datos numéricos
  - Podemos identificar un cluster por su centroide (Punto promedio)
  - De una manera alternativa, por su envolvente convexa
- Para datos no numéricos
  - Podemos utilizar “cualquier” distancia
  - No podemos establecer un centroide → clustoide
    - Es una instancia que se toma como representante del cluster
    - Puede ser el punto que minimiza la suma de las distancias con los otros puntos del cluster
      - O minimiza la distancia máxima a otro elemento
      - O minimiza la suma al cuadrado de las distancias con otros elementos
      - …

Problemas y Limitaciones

- Una vez que se combinan 2 clusters, la decisión no es reversible
- No hay una función objetivo a minimizar
- Problemas ante:
  - Presencia de outliers, y datos con ruido
  - Clusters con tamaños muy diferentes
  - Partición de clusters que agrupan muchos elementos



Clustering de “representación“

- Dado un dataset con N instancias, y un número de clusters a crear (k)
  - generan una partición de las N instancias en k clusters
- Para cada cluster, se define un punto que representa al conjunto
  - Lo más común es utilizar la media de los puntos del cluster

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$$

Clustering de “representación“

- El objetivo del método es encontrar la mejor partición según una función de scoring
  - La más común es la suma promedio de cuadrados
- Objetivo: encontrar la partición que minimiza SSE
  - Se podrían probar todas las combinaciones posibles…
  - Existen $k^N/k!$ Particiones posibles
    - Para dividir 100 datos en 5 clusters → 6.5738409e+67
    - (Un 6 seguido de 67 ceros, más o menos)

$$\mathrm{SSE}(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Algoritmo K-means

- Es el método más conocido
- Asume espacios numéricos, aunque se puede extender fácilmente a otros
- Utilizar una estrategia voraz iterativa para minimizar el SSE

Algoritmo K-means

K-MEANS $(\mathbf{D}, k, \epsilon)$:

$t = 0$

Randomly initialize k centroids: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$

repeat

  $t \leftarrow t + 1$

  $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$

  // Cluster Assignment Step

  foreach $\mathbf{x}_j \in \mathbf{D}$ **do**

    $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^t \right\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest centroid

    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

  // Centroid Update Step

  foreach $i = 1$ *to* k **do**

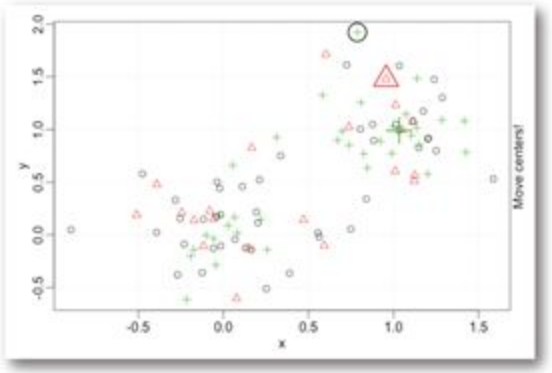
    $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

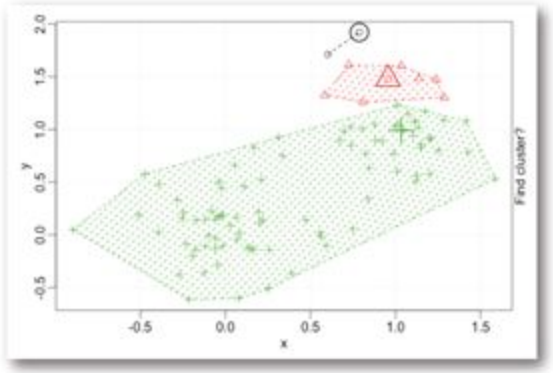
until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

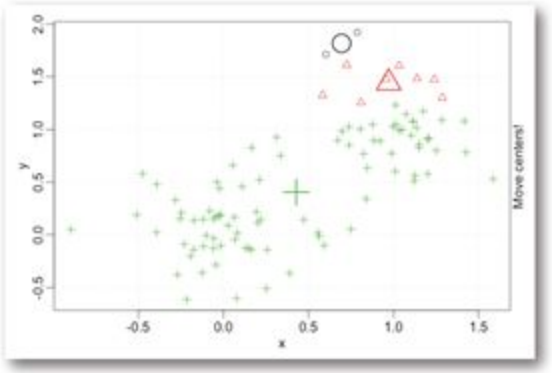
Algoritmo K-means

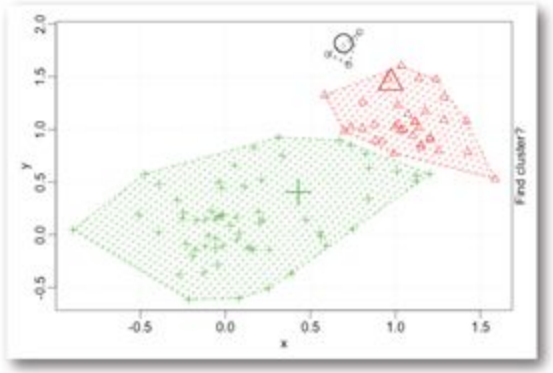
- ¿Cómo funciona el algoritmo?
  a. Elegir el valor de K (número de clusters).
  b. Elegir los centros de los k clusters, por ejemplo al azar (centroides)
  c. Asignar cada objeto al grupo más cercano (por ejemplo distancia euclídea)
  d. Re-estimar los centros de los k clusters, asumiendo que las asignaciones a los grupos están ok
  e. Repetir el paso c hasta que no haya más cambios
- Se puede cambiar el punto b, empezando con k centroides iniciales

Algoritmo K-means

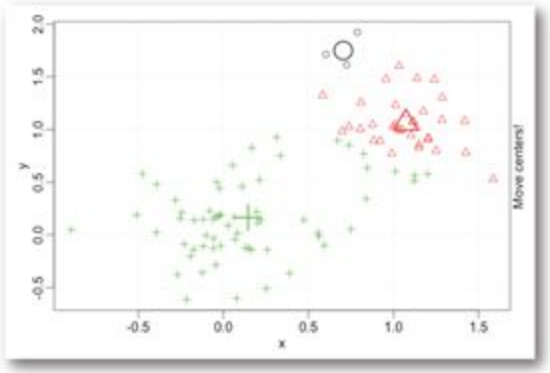


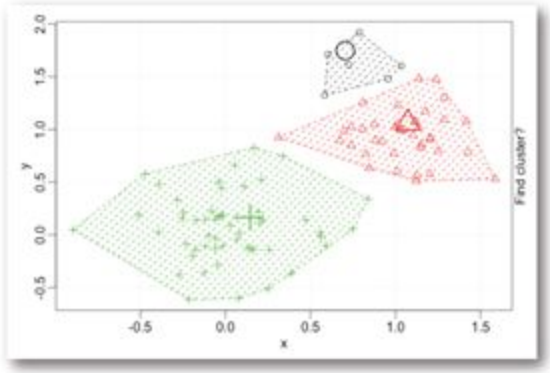


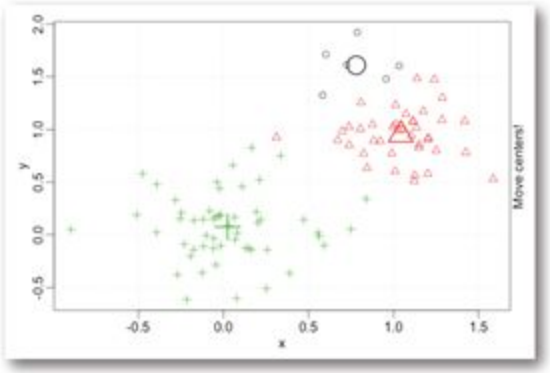


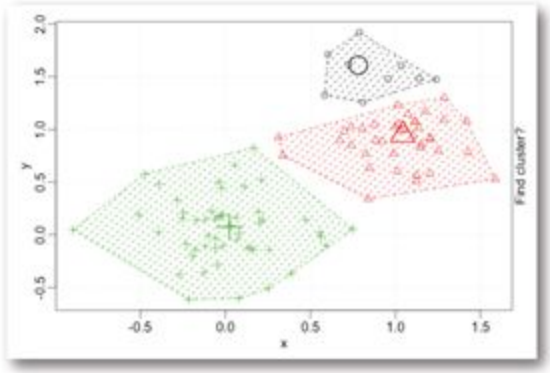


Algoritmo K-means

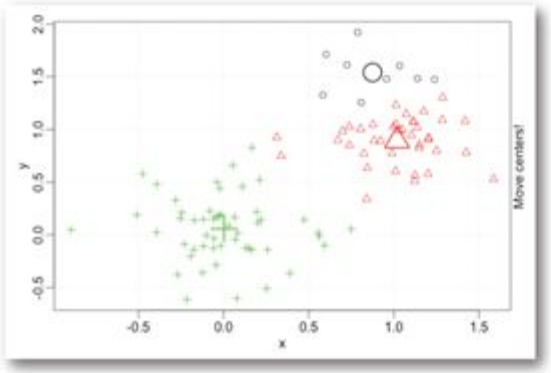


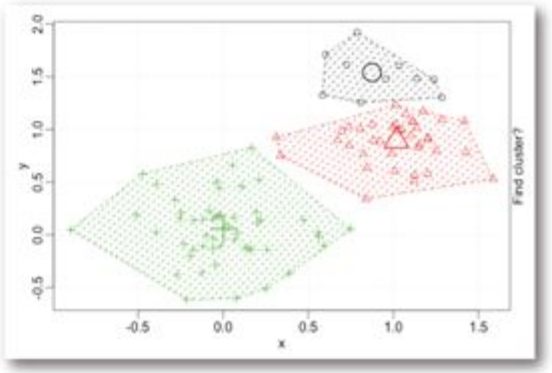


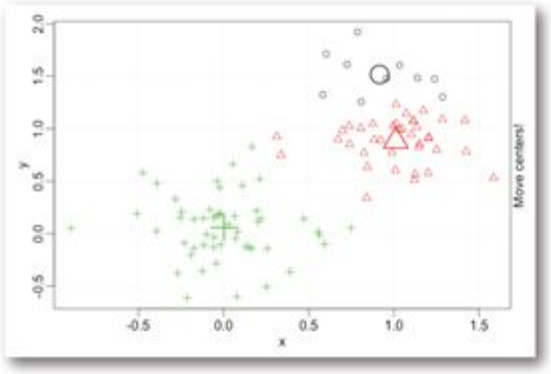


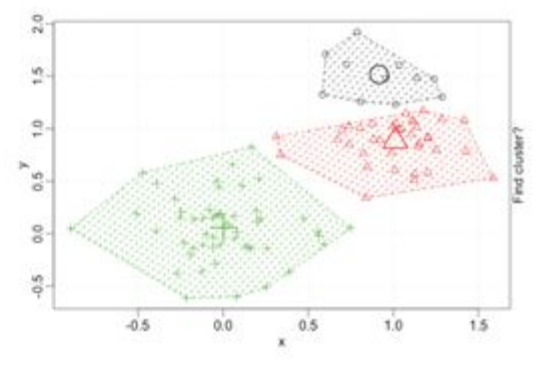


Algoritmo K-means



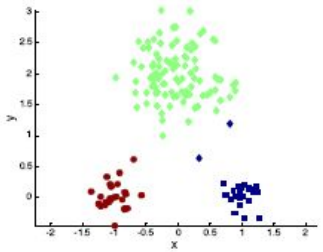


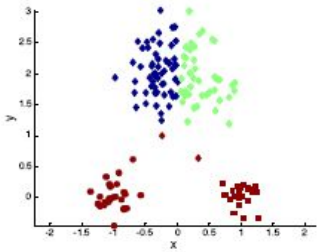




Algoritmo K-means

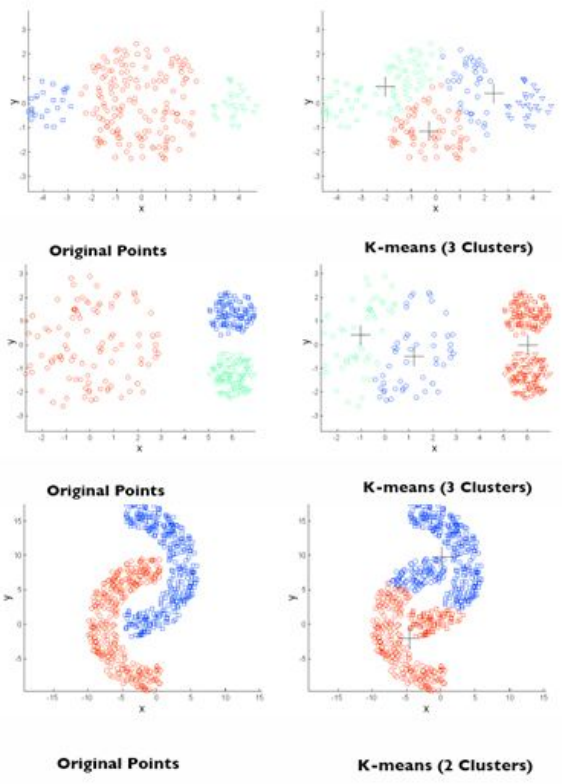
- Se trata de un método estocástico
  - Los puntos iniciales se escogen con cierto factor de aleatoriedad, por lo que el resultado obtenido NO siempre es el mismo
    - (Según implementación concreta)
  - El método seleccionado para elegir los centroides iniciales es Crítico para su desempeño
    - Usar otro método para determinarlos
    - Elegir un número mayor que k, y Seleccionar entre ellos
    - ...





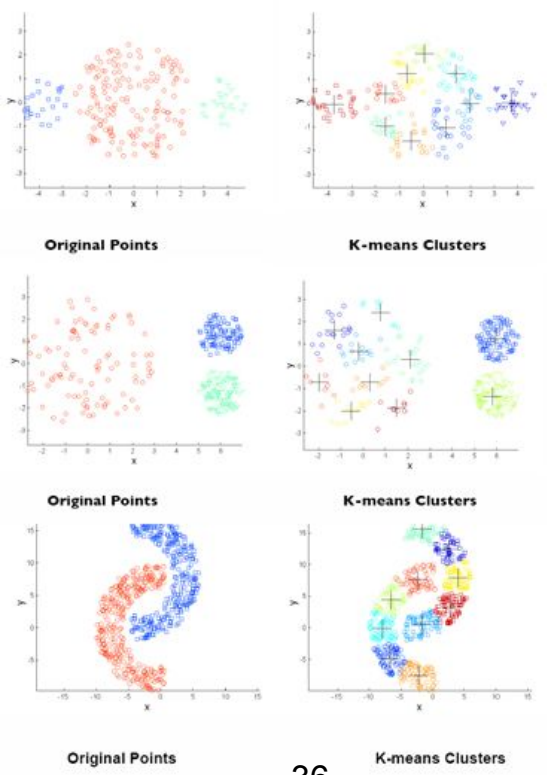
Limitaciones de K-means

- Principalmente, su desempeño se ve mermado cuando los clusters tienen
  - Diferentes tamaños
  - Diferentes densidades
  - Formas no globulares
- Presenta problemas cuando los datos contienen outliers
  - (Como casi todos los métodos)



Limitaciones de K-means

Una solución puede ser hacer un número superior de clusters, y luego “unir las partes”



Resúmen de K-means

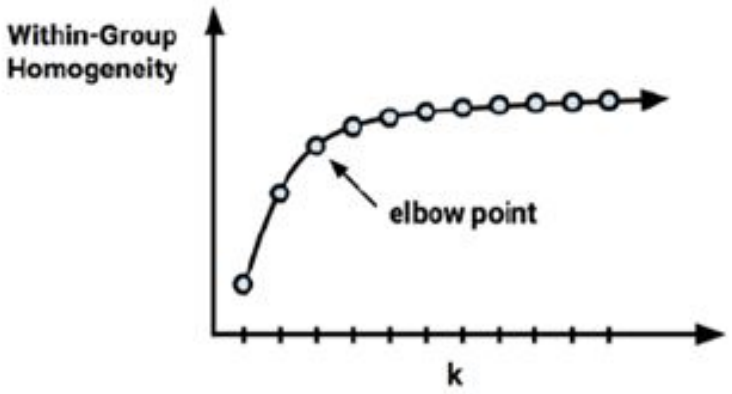
- Es un método simple
  - Se debe seleccionar el número de clusters a priori
  - Sensible a outliers
- Sus variantes giran en torno a:
  - Elección de los elementos iniciales
  - Cálculo de distancias
  - Diferentes definiciones de centroide (aparte de la media)
- Extensiones
  - Muchas… BFR es la más conocida:
    - Especialmente diseñada para lidiar con grandes volúmenes de datos
    - Mantiene un resúmen estadístico de los datos ya procesados

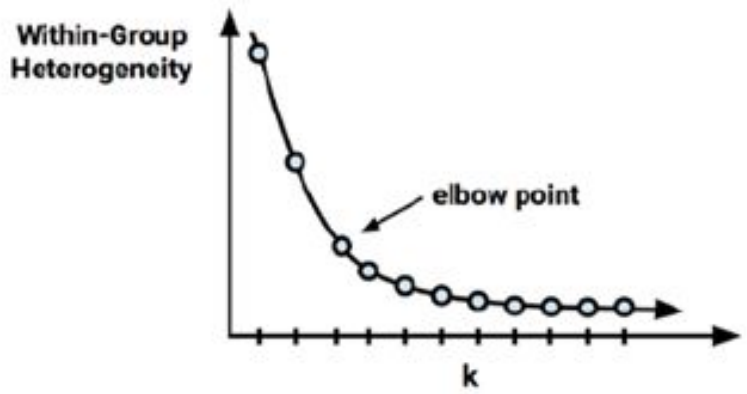
Resúmen de K-means

- Se debe de elegir a priori el número de clusters:
  - Conocimiento a priori: Por ejemplo, si clasificamos películas, k=nº de géneros
  - Dirigidos por el negocio: Por ejemplo, el departamento de Marketing sólo tiene recursos para hacer 3 campañas distintas de marketing
  - Sin nada de lo anterior: k=raíz(n/2)
    - Suele ser una buena elección

Resúmen de K-means

- Si el tiempo lo permite: Regla del codo
  - Realizar varias repeticiones del método, incrementando el valor de k
  - Tomar aquel valor a partir del cual, las diferencias en distancias inter e intra cluster no son significativas.





Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/