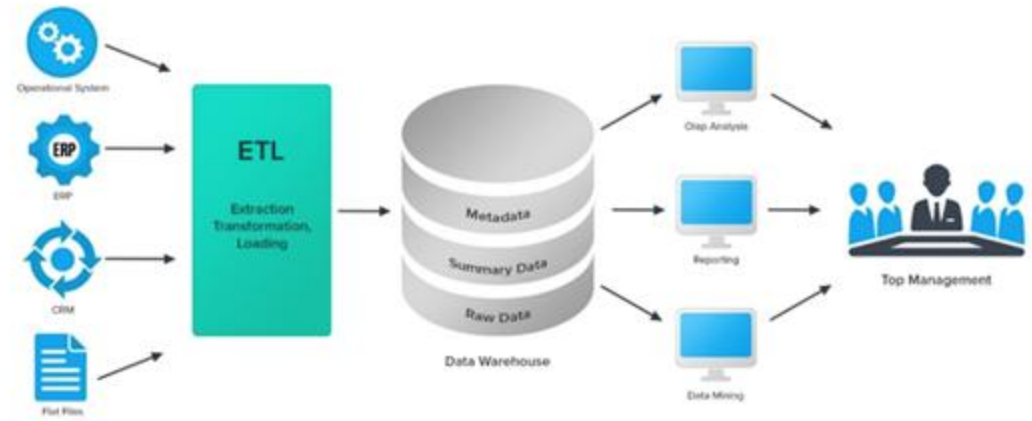
FACULTAD DE INGENIERÍA
ESCUELA DE COMPUTACIÓN



Datawarehouse y Minería de Datos

Guía #10: Agrupamiento con k-means

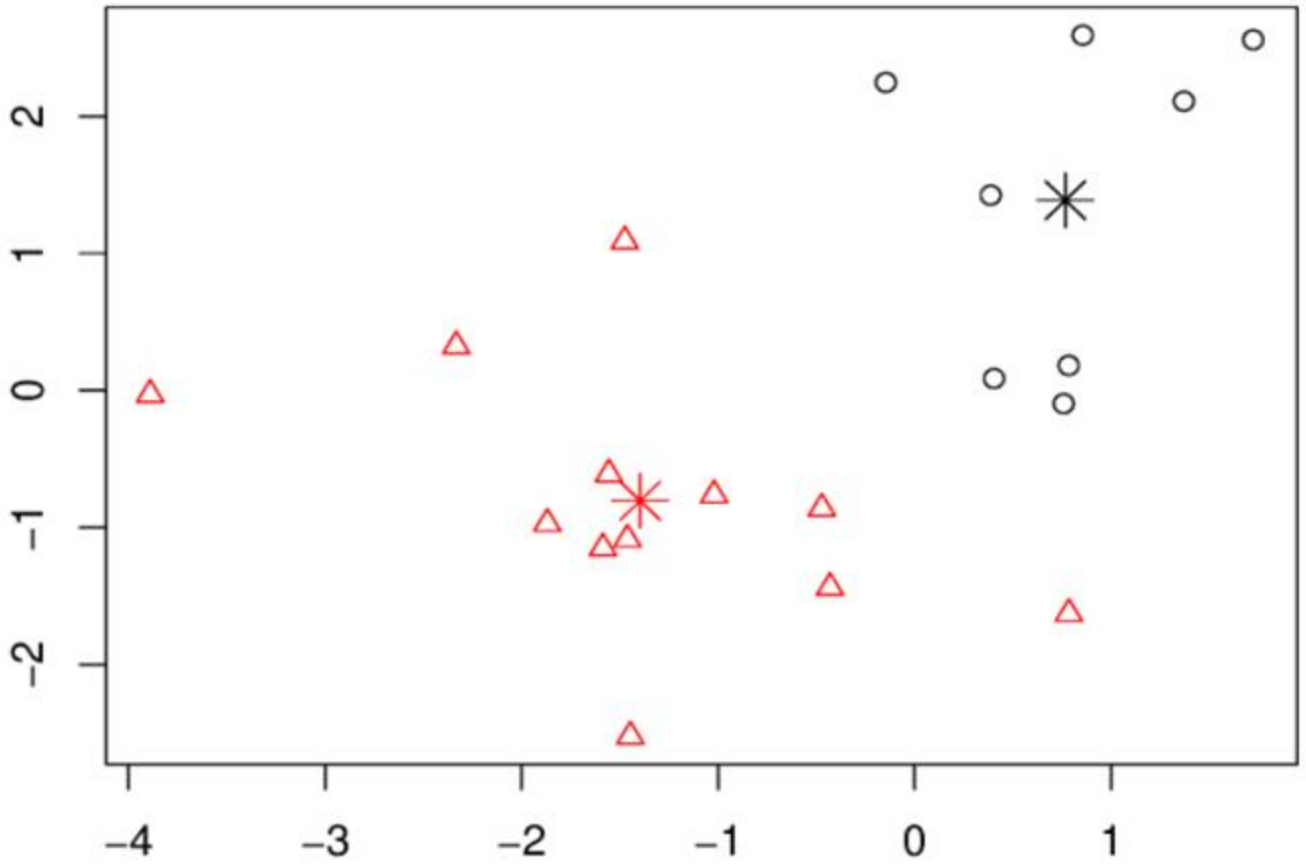
k-means o agrupamiento de k-medias

El algoritmo está basado en la participación de un conjunto de ***n*** observaciones en ***k*** grupos en el que cada observación pertenece al grupo cuya distancia es menor. Es un tipo de agrupamiento no supervisado, que se utiliza cuando tiene datos no etiquetados, es decir, datos sin categorías o grupos definidos. El objetivo de este algoritmo es encontrar grupos en los datos. Los puntos de datos se agrupan según la similitud de características.

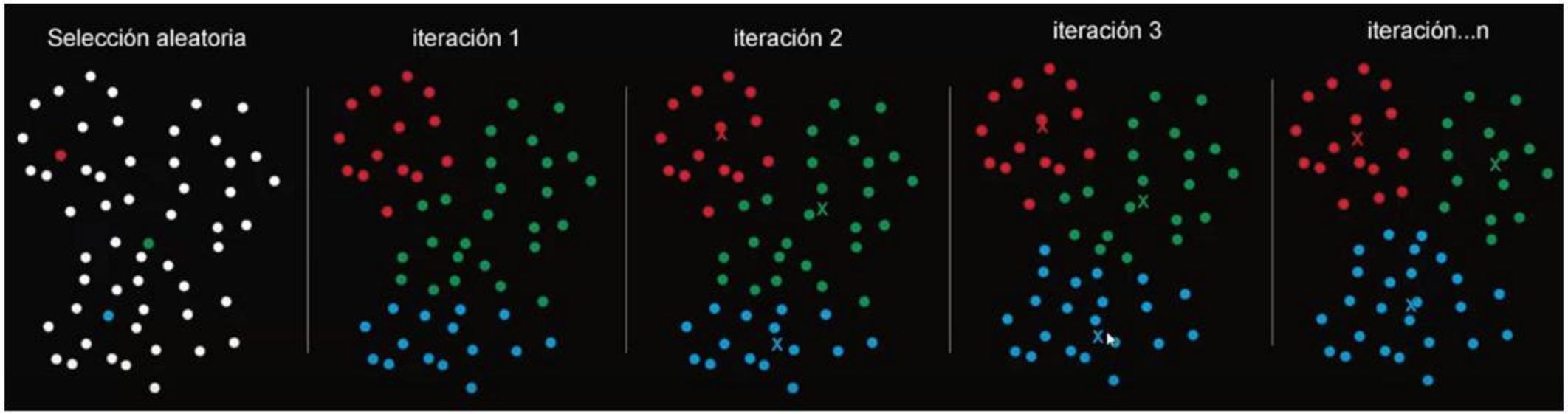
Algunas consideraciones importantes respecto al algoritmo k-means (k-medias) son:

- El número de grupos o clústeres, denotado como k, debe definirse antes de ejecutar el algoritmo.
- Cada grupo o clúster está definido por un punto, generalmente identificado como el centro y llamado centroide del clúster.

Ejemplo k-means



Ejemplo k-means



k-means o agrupamiento de k-medias

A grandes rasgos, el algoritmo funciona en dos fases principales:

En ***primer lugar***, la fase de inicialización, identificada con k puntos como centroides iniciales. Aunque no es necesario que sean puntos del conjunto de datos, si es importante que sean puntos dentro del mismo intervalo.

La ***segunda fase*** es iterativa, y consiste en:

a) Asignar a cada centroide los puntos del conjunto de datos más próximos, formando k grupos disjuntos,

b) y, a continuación, recalcular los centroides en base a los puntos que forman parte e su grupo o partición.

k-means o agrupamiento de k-medias

Inicialización de los centroides

Un primer enfoque, simple y ampliamente usado, consiste en inicializar los centroides con k puntos aleatorios del conjunto de datos. Es decir:

$$\zeta_i = rand(d_j) \mid d_j \in D, i = \{1, \ldots, k\}.$$

El algoritmo k-means utiliza la inicialización de centroides aleatoria, pero existen otras aproximaciones mas complejas que son utilizadas por métodos derivados de k-means. Por ejemplo, se pueden seleccionar los centroides iniciales a partir de la distribución de probabilidad de las instancias de D, intentando cubrir todo el rango de valores de los datos, especialmente las zonas con mayor concentración de instancias.

Cálculo de la distancia

El algoritmo k-means, generalmente, utiliza la distancia euclidea, la cual es calculada por medio del teorema de Pitágoras.

k-means o agrupamiento de k-medias

Recalculo de centroides

El valor de los centroides es calculado como la media (mean) de todos los puntos que pertenecen a este segmento (de aquí el nombre del algoritmo, k-means). Por lo tanto este algoritmo solo es aplicable a atributos continuos. En caso de tener atributos no continuos en el conjunto de datos , debemos aplicar una transformación previa.

Criterio de parada

Esta se alcanza cuando no hay cambios en el recalculo de los centroides durante una interacción completa, provocando que no haya alteraciones en la distribución de las instancias en las distintas particiones o clústeres. Es decir, se llega a una situacion de estabilidad en la distribución de las instancias. Esta condición se puede garantizar después de un finito de interacciones, dependiendo de la métrica empleada en el calculo de la distancia y el recalculo de los centroides.

k-means o agrupamiento de k-medias

Criterios para seleccionar un valor de K

Un criterio es minimizar la suma de residuos cuadrados, es decir, la suma de las distancias de cualquier vector o instancia a su centroide mas cercano. Este criterio busca la creación de segmentos lo mas compactos posibles.

Otro criterio podría ser el de maximizar la suma de distancias entre los segmentos, por ejemplo entres sus centros. En este caso estaríamos priorizando tener segmentos los mas alejados posible entre sí, es decir, tener segmentos lo mas diferenciados posible.

Los criterios anteriores (minimización de distancias intragrupo o maximización de distancia intergrupo) pueden usarse para establecer el valor adecuado para el parámetro k.

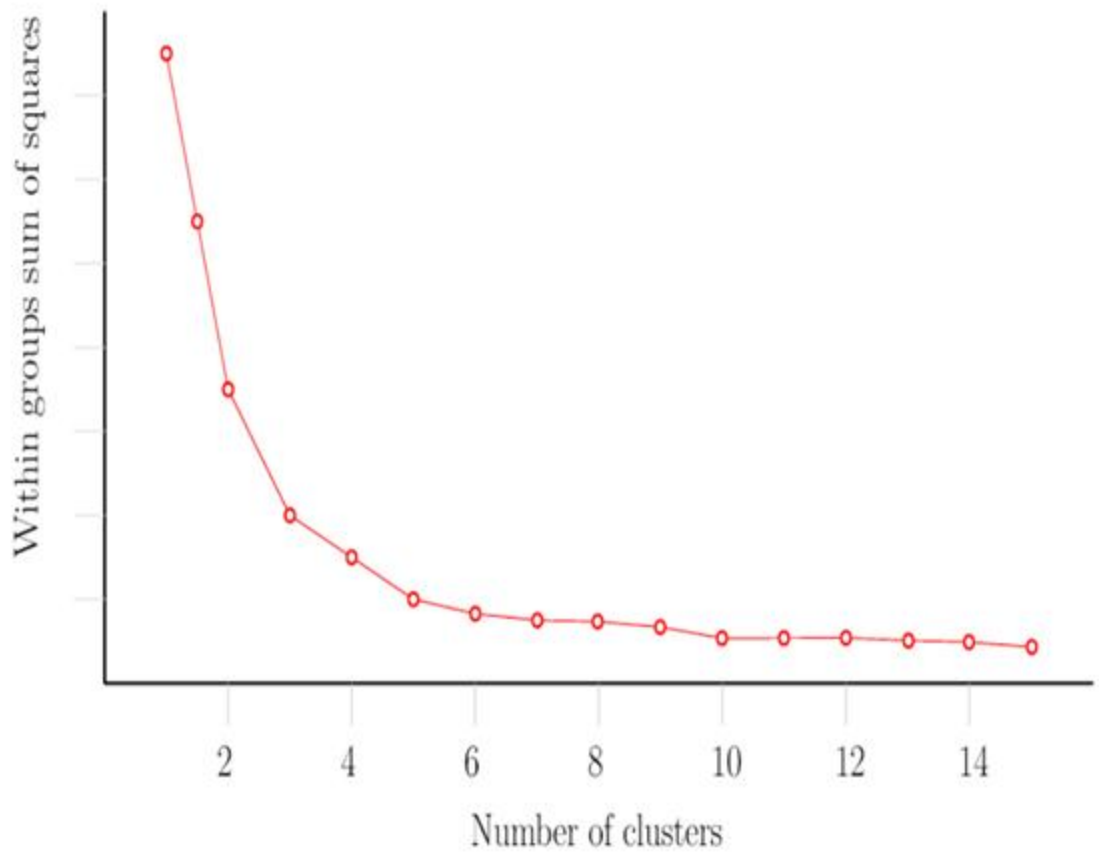
k-means o agrupamiento de k-medias

Criterios para seleccionar un valor de K

Por ejemplo en la siguiente figura podemos ver como evoluciona la métrica de calidad respecto al numero de particiones creadas.

En concreto, observemos como a partir de cinco segmentos, la mejora que se produce en la distancia interna de los segmentos ya es insignificante.

Este hecho debería ser indicativo de que cinco segmentos es un valor adecuado para k.



Métodos derivados de k-means

k-medians

El algoritmo k-mediana (k-medians) es una variación del método k-means, ***donde se sustituye el cálculo basado en el valor medio, por el valor de la mediana***. Aunque la obtención del valor de la mediana requiere mayor complejidad computacional, es preferible en determinados contextos. Al igual que en el caso del k-means, este método puede implementarse mediante distintas métricas de similaridad, aunque existen el riesgo de perder la garantía de convergencia.

Métodos derivados de k-means

k-medoids

El método k-medians presenta mejores resultados que el k-means cuando las distribuciones son asimétricas o existen valores extremos. Aun así, k-medians no garantiza que los centros sean similares a alguna instancia del conjunto de datos. Esto se debe a que, a menos que los atributos medianos pueden no parecerse a ninguna instancia existente del conjunto de datos ni del dominio completo.

El algoritmo k-medoids propone el ***recalculo de los centros a partir de instancias que presentan un valor de disimilitud (falta de semejanza) mínimo respecto a las demás instancias de la partición o clúster***. La identificación de los puntos, conocidos como medoids, es computacionalmente más costosa que las aproximaciones vistas anteriormente.

Comencemos con nuestra práctica