MAGISTER EN ESTADÍSTICA

CLUSTERING

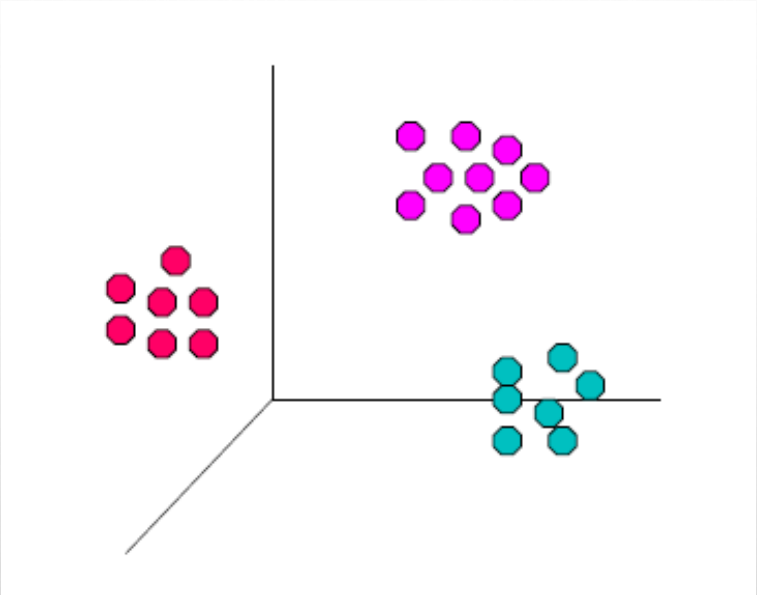
Agenda

Introducción

Los algoritmos que veremos a continuación tienen por objetivo descubrir cómo se agrupan naturalmente los datos (multivariados). Estos grupos típicamente se conocen como clusters, y la idea general es que todos los puntos que están en un mismo cluster son parecidos o similares entre sí, y a su vez diferentes a los puntos de los otros clusters.

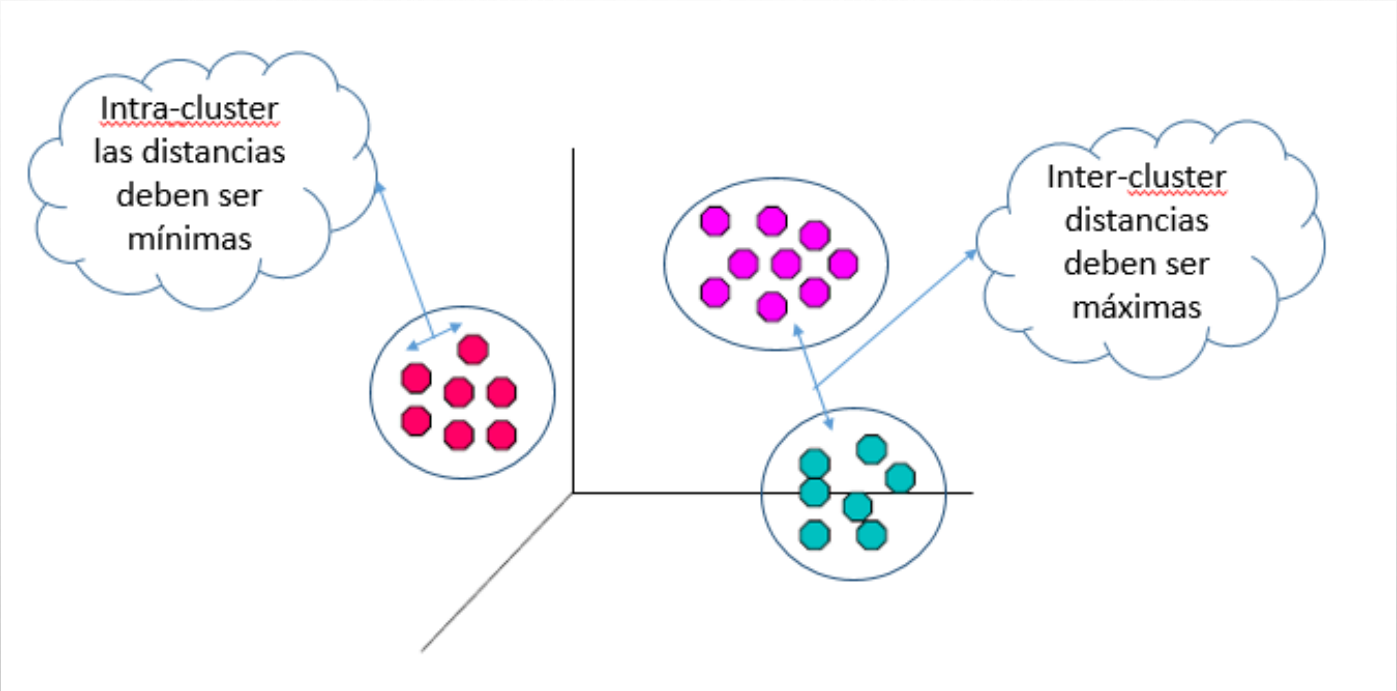
Definición de Clúster

Encontrar grupos de objetos, tal que estos objetos en el grupo sean similares entre si y distintos de los objetos en otros grupos.



Definición de Clúster

Encontrar grupos de objetos, tal que estos objetos en el grupo sean similares entre si y distintos de los objetos en otros grupos.



Medidas de distancias

Todos los métodos de clustering tienen una cosa en común, para poder llevar a cabo las agrupaciones necesitan definir y cuantificar la similitud entre las observaciones. El término distancia se emplea dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia.

Tipos de distancias

Distancia Euclidiana: La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y), la distancia euclídea entre p y q viene dada por la ecuación:

$$D_E(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}.$$

Generalizando la ecuación para $p = (p_1, p_2, \ldots, p_n)$ y $q = (q_1, q_2, \ldots, q_n)$, tenemos que:

$$D_E(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2},$$

$$D_E(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

Tipos de distancia

Distancia de Manhattan define la distancia entre dos puntos p y q como el sumatorio de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por outliers (es más robusta) que la distancia euclídea debido a que no eleva al cuadrado las diferencias.

$$d_{man}(p, q) = \sum_{i=1}^{n} |(p_i - q_i)|$$

Tipos de clustering

Un clustering es un conjunto de clúster. Distinción relevante entre conjuntos de clusters particional y jerárquico.

- **Particionamiento jerárquico**: Una división de objetos (datos) en subconjuntos que no se traslapan (clusters) tal que cada objeto pertenece exactamente a un subconjunto.
- **Clustering Jerárquico**: Un conjunto de clusters uno dentro de otro organizados como un árbol jerárquico.

K-means Clustering

El método K-means clustering (MacQueen, 1967) agrupa las observaciones en K clusters distintos, donde el número K lo determina el analista antes de ejecutar del algoritmo. K-means clustering encuentra los K mejores clusters, entendiendo como mejor cluster aquel cuya varianza interna (intra-cluster variation) sea lo más pequeña posible. Se trata por lo tanto de un problema de optimización, en el que se reparten las observaciones en K clusters de forma que la suma de las varianzas internas de todos ellos sea lo menor posible. Para poder solucionar este problema es necesario definir un modo de cuantificar la varianza interna.

K-means Clustering

Propiedades

Considérense $C_1, \ldots, C_K$ como los sets formados por los índices de las observaciones de cada uno de los clusters. Por ejemplo, el set C_1 contiene los índices de las observaciones agrupadas en el cluster 1. La nomenclatura empleada para indicar que la observación i pertenece al cluster k es: $i \in C_k$. Todos los sets satisfacen dos propiedades:

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. Significa que toda observación pertenece al menos a uno de los K clusters.
- $C_K \cap C_{K'} = \emptyset$ para todo $K \neq K'$. Implica que los clusters no solapan, ninguna observación pertenece a más de un cluster a la vez.

K-means Clustering

Medidas

Dos de las medidas más comúnmente empleadas definen la varianza interna de un cluster ($W(C_k)$) como:

- La suma de las distancias euclídeas al cuadrado entre cada observación (x_i) y el centroide (μ) de su cluster. Esto equivale a la suma de cuadrados internos del cluster.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- La suma de las distancias euclídeas al cuadrado entre todos los pares de observaciones que forman el cluster, dividida entre el número de observaciones del cluster.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

K-means Clustering

Medidas

Minimizar la suma total de varianza interna $\sum_{k=1}^{K} W(C_k)$ de forma exacta (encontrar el mínimo global) es un proceso muy complejo debido a la inmensa cantidad de formas en las que n observaciones se pueden dividir en K grupos. Sin embargo, es posible obtener una solución que, aun no siendo la mejor de entre todas las posibles, es muy buena (óptimo local).

K-means Clustering

Pasos para encontrar óptimo local

1. Asignar aleatoriamente un número entre 1 y K a cada observación. Esto sirve como asignación inicial aleatoria de las observaciones a los clusters.
2. Iterar los siguientes pasos hasta que la asignación de las observaciones a los clusters no cambie o se alcance un número máximo de iteraciones establecido por el usuario.

    - Para cada uno de los clusters calcular su centroide. Entendiendo por centroide la posición definida por la media de cada una de las dimensiones (variables) de las observaciones que forman el cluster. Aunque no es siempre equivalente, puede entenderse como el centro de gravedad.
    - Asignar cada observación al cluster cuyo centroide está más próximo.

K-means Clustering

Otra forma de encontrar el óptimo local

1. Especificar el número K de clusters que se quieren crear.
2. Seleccionar de forma aleatoria k observaciones del set de datos como centroides iniciales.
3. Asignar cada una de las observaciones al centroide más cercano.
4. Para cada uno de los K clusters recalcular su centroide.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones establecido.

K-means Clustering

Ventajas y desventajas

- Requiere que se indique de antemano el número de clusters que se van a crear. Esto puede ser complicado si no se dispone de información adicional sobre los datos con los que se trabaja.
- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. Para minimizar este problema se recomienda repetir el proceso de clustering entre 20 a 50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna.
- Presenta problemas de robustez frente a outliers. La única solución es excluirlos o recurrir a otros métodos de clustering más robustos como K-medoids (PAM).

Métricas de validación

La medida más común para validar clusters de k-means es la suma del error cuadrático (SSE): Para cada punto, el error es la distancia al clúster más cercano.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

x es un punto de dado en el cluster C_i y m_i es el punto representativo del clúster C_i. Una forma sencilla de reducir el SSE es incrementar el número de clústers K.

Clustering Jerárquico

Hierarchical clustering es una alternativa a los métodos de partitioning clustering que no requiere que se pre-especifique el número de clusters. Los métodos que engloba el hierarchical clustering se subdividen en dos tipos dependiendo de la estrategia seguida para crear los grupos:

- **Agglomerative clustering (bottom-up)**: el agrupamiento se inicia en la base del árbol, donde cada observación forma un cluster individual. Los clusters se van combinado a medida que la estructura crece hasta converger en una única "rama" central.
- **Divisive clustering (top-down)**: es la estrategia opuesta al agglomerative clustering, se inicia con todas las observaciones contenidas en un mismo cluster y se suceden divisiones hasta que cada observación forma un cluster individual.

Clustering Jerárquico

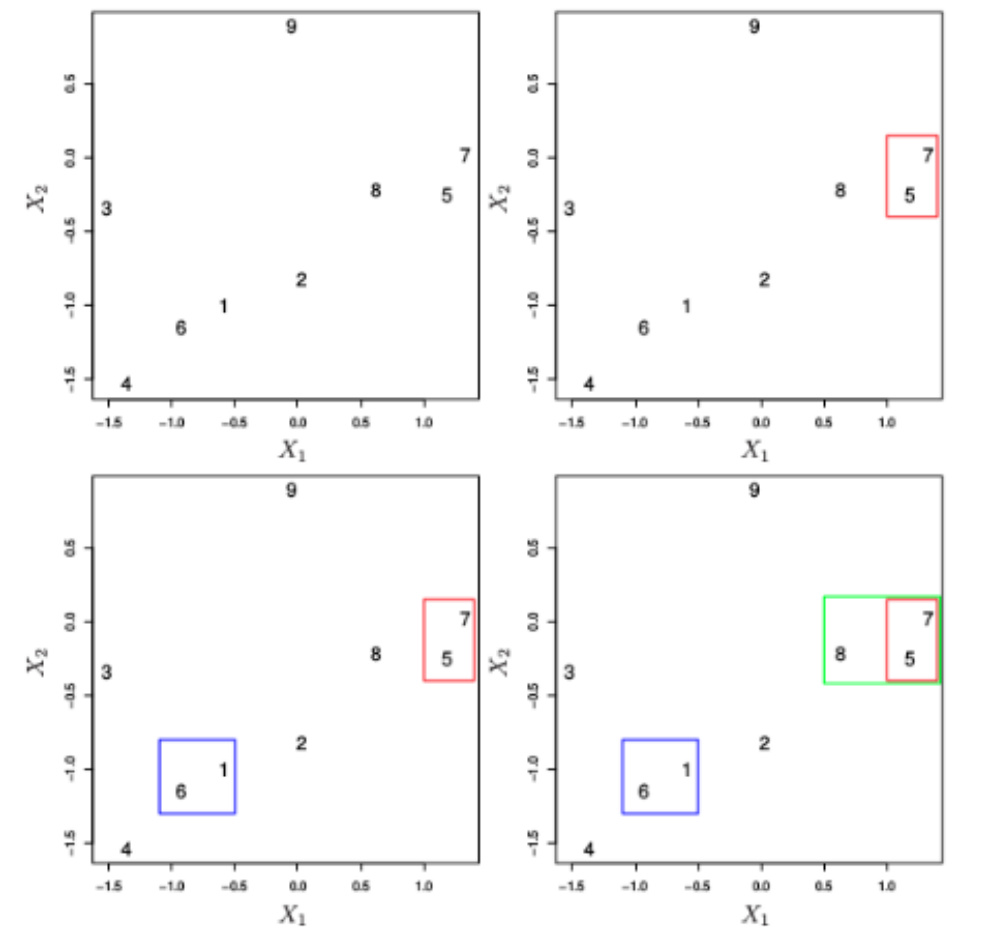
Aglomerativo

La estructura resultante de un agglomerative hierarchical clustering se obtiene mediante un algoritmo sencillo.

1. El proceso se inicia considerando cada una de las observaciones como un cluster individual, formando así la base del dendrograma (hojas).
2. Se inicia un proceso iterativo hasta que todas las observaciones pertenecen a un único cluster:
   - Se calcula la distancia entre cada posible par de los n clusters. El investigador debe determinar el tipo de medida emplea para cuantificar la similitud entre observaciones o grupos (distancia y linkage).
   - Los dos clusters más similares se fusionan, de forma que quedan $n-1$ clusters.
3. Determinar dónde cortar la estructura de árbol generada (dendrograma).

Aglomerativo

Figure: Cluster Aglomerativo



Clustering Jerárquico

Divisivo

Este algoritmo se inicia con un único cluster que contiene todas las observaciones, a continuación, se van sucediendo divisiones hasta que cada observación forma un cluster independiente.

1. Todas las n observaciones forman un único cluster.
2. Repetir hasta que hayan n clusters:
   - Calcular para cada cluster la mayor de las distancias entre pares de observaciones (diámetro del cluster).
   - Seleccionar el cluster con mayor diámetro.
   - Calcular la distancia media de cada observación respecto a las demas.
   - La observación más distante inicia un nuevo cluster.
   - Se reasignan las observaciones restantes al nuevo.
3. cluster o al viejo dependiendo de cual está más próximo.

Clustering Jerárquico

Medidas de disimilitud: La medida de disimilitud más común que se emplea es la distancia euclídea, pero existen otras (distancia de Mahalanobis, distancia de Minkowski, etc.). Por otro lado, se encuentra la disimilitud entre pares de grupos de observaciones, donde aparece el concepto de método de unión o linkage, que mide esta disimilitud.

Clustering Jerárquico

Medidas de disimilitud

Las medidas de disimilitud (linkage) más comunes son:

- **Completo**: Distancia máxima entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la máxima de las distancias.

$$D = max_{i,j} dist(X_i, Y_i)$$

- **promedio**: Distancia media entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la media de las distancias.

$$D = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{i} dist(X_i, Y_j)$$

Clustering Jerárquico

Medidas de disimilitud

- **simple**: Distancia mínima entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la mínima de estas medidas. Puede dar lugar a dendogramas donde las observaciones se fusionan una a una, obteniendo clústeres muy extendidos. Puede crear grupos muy homogéneos.

$$D = min_{i,j} dist(X_i, Y_j)$$

- **centroíde**: Distancia entre centros. Medida de disimilitud entre el centroide del clúster A y el centroide del clúster B. Suele utilizarse con frecuencia en genómica, pero puede dar lugar a inversiones indeseables que dificulten la visualización e interpretación.

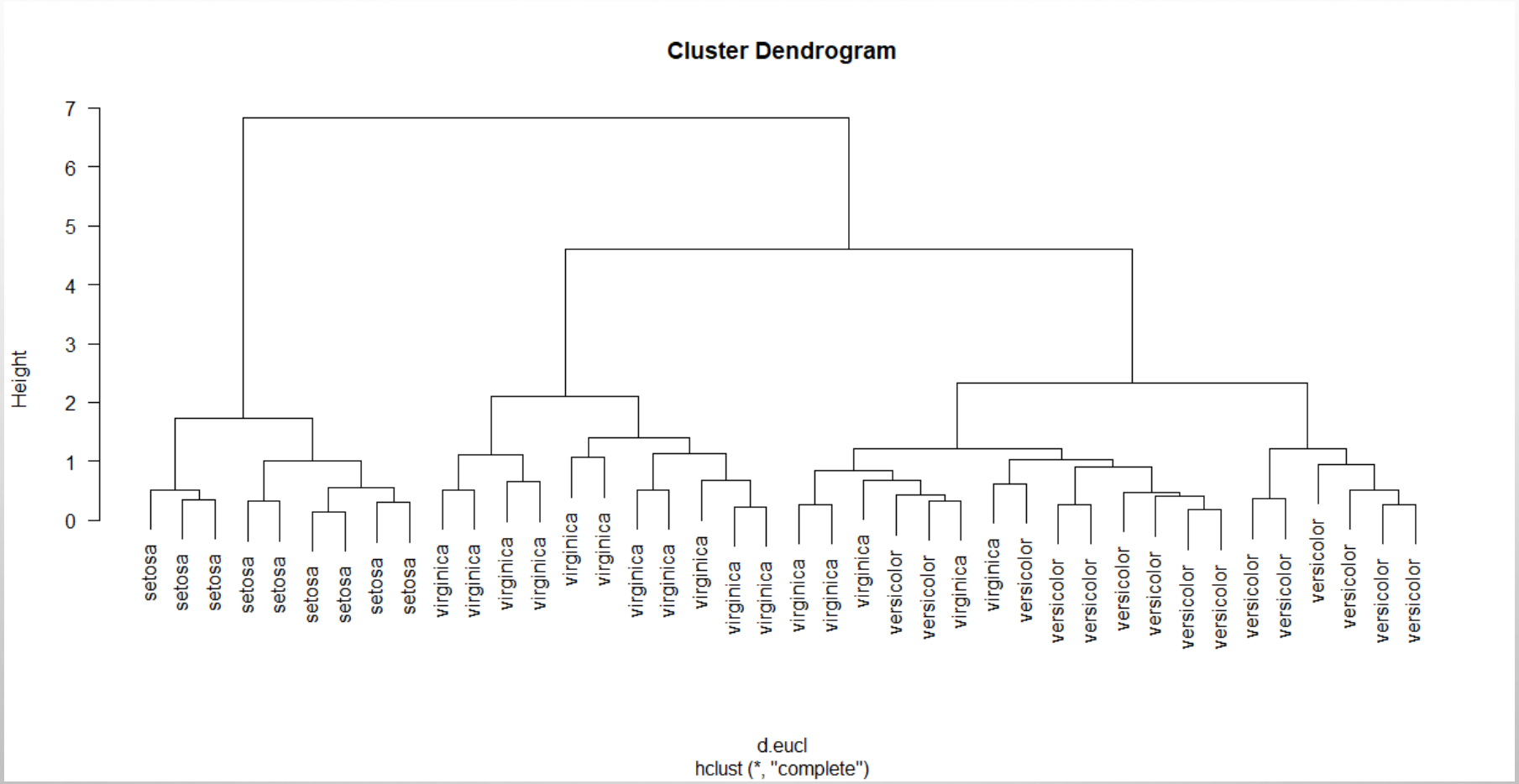
$$D = dist(\bar{x}, \bar{y})$$

Clustering Jerárquico

Dendograma

En la base del dendrograma, cada observación forma una terminación individual conocida como hoja del árbol. A medida que se asciende por la estructura, pares de hojas se fusionan formando las primeras ramas. Estas uniones (nodos) se corresponden con los pares de observaciones más similares. También ocurre que ramas se fusionan con otras ramas o con hojas. Cuanto más temprana (más próxima a la base del dendrograma) ocurre una fusión, mayor es la similitud.

Dendograma


Cluster Dendrogram
Height
0
1
2
3
4
5
6
7
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
setosa
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
virginica
versicolor
versicolor
virginica
virginica
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
versicolor
d.eucl
hclust (*, "complete")

Estimación mediante el método del codo

Estimación del valor del parámetro k con el método Elbow (codo):

- Se acumulan las sumas de diferencias al cuadrados de todos los grupos y se grafican para distintos valores del parámetro k.
- Se escoge visualmente aquel valor para el cual la caída en la suma total es marginal.

Métricas de validación

Índice de Dunn

Denótaremos como d_{min} la distancia mínima entre dos puntos de diferentes grupos y d_{max} como la mayor distancia dentro del grupo. La distancia entre los grupos C_k y $C_{k'}$ se mide por la distancia entre sus puntos más cercanos:

$$d_{kk'} = min_{i \in I_k,\ j \in I_{k'}} ||M_i^{\{k\}} - M_j^{\{k'\}}||,$$

donde d_{min} es la mínima distancia de las distancias $d_{kk'}$
Mientras que para cada grupo C_k, denotemos por D_k la distancia más grande que separa dos puntos distintos en el grupo:

$$D_k = max_{i,j \in I_k,\ i \neq j} ||M_i^{\{k\}} - M_j^{\{k'\}}||,$$

donde d_{max} es la máxima distancia de D_k. Finalmente el índice de Dunn viene dado por el cuociente entre d_{min} y d_{max}.

$$C = \frac{d_{min}}{d_{max}}$$

Métricas de validación

Índice de la Silueta

Es el promedio del valor silueta de cada observación. El valor silueta mide el grado de confianza en la asignación de una observación a un cluster, con observaciones bien clusterizadas obtienen valores cerca de 1, mientras que observaciones mal clusterizadas con valores cercanos $a - 1$. Para una observación i la silueta se define como:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)},$$

donde a_i es la distancia media entre i y todas las observaciones dentro del mismo clúster, mientras que b_i es la distancia media entre i y las observaciones en el grupo más cercano.

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i,j), \quad b_i = min_{C_k \in C} \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)}$$

Métricas de validación

Índice de la Silueta

Donde tenemos que $C(i)$ es el cluster que contiene a i , $dist(i,j)$ es la distancia entre las observaciones i y j, y $n(C)$ es la cardinalidad del cluster C. El índice de la Silueta se mueve entre los valores del íntervalo [-1,1]

Métricas de validación

Índice Davies-Bouldin

Se obtiene de la siguiente ecuación:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right),$$

donde n es el número de clústers, c_x es el centroide del clúster x, σ_x es la distancia media de todas las observaciones en el clúster x al centroide c_x y $d(c_i, c_j)$ es la distancia entre los centroídes c_i y c_j. Dado que algoritmos que producen clusters con distancias intra-cluster bajas (baja similitud entre clusters) y distancias inter-clusters altas (baja similitud inter-cluster) van a tener un índice Davies-Bouldin bajo.
Por lo tanto, el algoritmo de clustering que produce una colección de clusters con el valor más bajo de índice Davies-Bouldin es considerado el mejor ajuste de clustering, basados en este crítерio.