ANÁLISIS MULTIVARIANTE

Análisis de conglomerados o clúster.

Mg. Víctor Guevara P

Universidad Nacional José Faustino Sánchez Carrión

Febrero 2023

Normas de participación.

- Ingresar puntual
- Mantener silenciado el micrófono durante la sesión (salvo cuando se pida participación)
- Las preguntas se realizarán por el chat/en caso sea necesario se habilita el micrófono.
- Realizar las actividades encomendadas.

Introducción

Introducción a clustering

- El análisis cluster o clustering es un grupo de técnicas multivariadas cuyo objetivo principal es agrupar objetos en función de las características que poseen.
- El análisis de conglomerados se diferencia del análisis factorial exploratorio en que el análisis de conglomerados agrupa objetos, mientras que el análisis factorial exploratorio se ocupa principalmente de agrupar variables.
- El análisis factorial exploratorio hace que las agrupaciones se basen en patrones de variación (correlación) en los datos, mientras que el análisis de conglomerados hace agrupaciones en función de la distancia (proximidad).

Clustering

Clustering

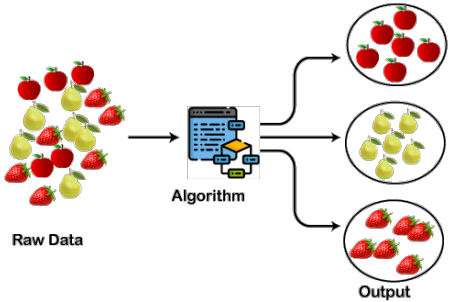
- El análisis cluster clasifica objetos (por ejemplo, encuestados, productos u otras entidades), en un conjunto de características seleccionadas por el usuario (variables clustering).
- Los clusters resultantes deben exhibir una alta homogeneidad interna (dentro del cluster) y una alta heterogeneidad externa (entre cluster).
- Si la clasificación tiene éxito, los objetos dentro de los conglomerados estarán muy juntos cuando se tracen geométricamente, y los diferentes conglomerados estarán muy separados.

Clustering

- Las técnicas de agrupamiento se aplican cuando no hay una clase que predecir sino más bien cuando las instancias se van a dividir en grupos naturales.
- Estos grupos presumiblemente reflejan algún mecanismo en el trabajo en el dominio del que se extraen las instancias, un mecanismo que hace que algunas instancias se parezcan más aentre sí que lo hacen con las instancias restantes.

Clustering

Clustering o agrupación, es la tarea de agrupar observaciones de tal manera que los miembros del mismo grupo sean muy similares entre si y los miembros de diferentes grupos sean muy diferentes entre sí.



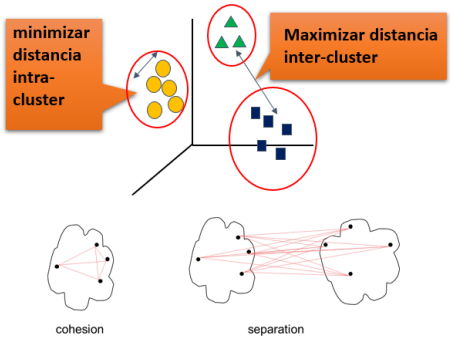
Clustering

Intra-cluster

Cada grupo (conglomerado o cluster) sea homogéneo respecto a las variables utilizadas para caracterizarlos

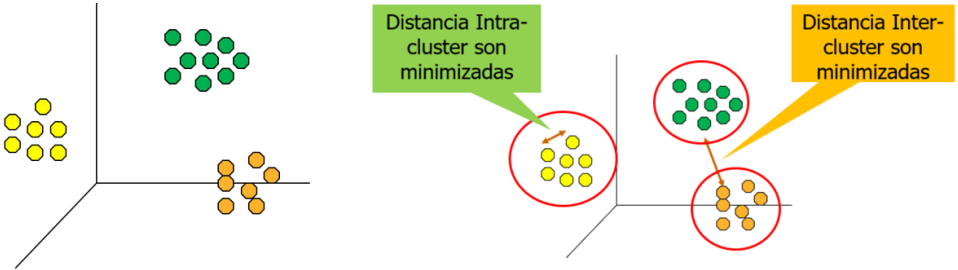
Inter-cluster

Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas.



Clustering

Proceso de examinar una colección de "puntos" y agrupar los puntos en "grupos" de acuerdo con alguna medida de distancia.



Clustering aplicaciones

- Paso intermedio para otros problemas fundamentales de minería de datos, machine learning entre otros
- Segmentación de clientes
- Resumen de datos
- Detección de tendencia dinámica
- Análisis de datos multimedia
- Análisis de datos biológicos
- Análisis de redes sociales

Clustering

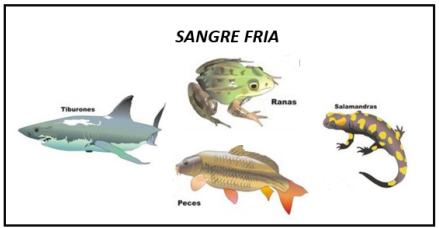
El objetivo principal del análisis de conglomerados es definir la estructura de los datos colocando las observaciones más similares en grupos.

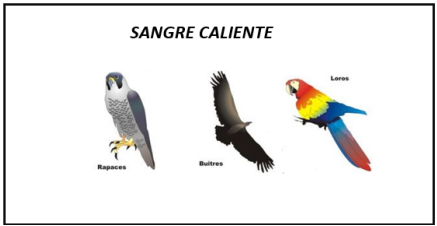
Para lograr esta tarea, debemos abordar tres preguntas básicas:

1. ¿Cómo formamos clústeres?
2. ¿Cuántos grupos formamos?
3. ¿Cómo medimos la similitud?

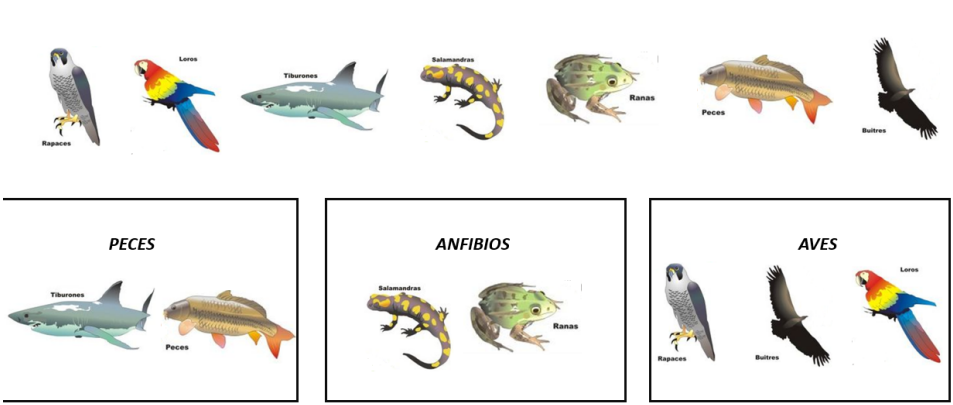
Clustering – ¿Cómo formamos grupos?

Clustering – formas de agrupar los datos

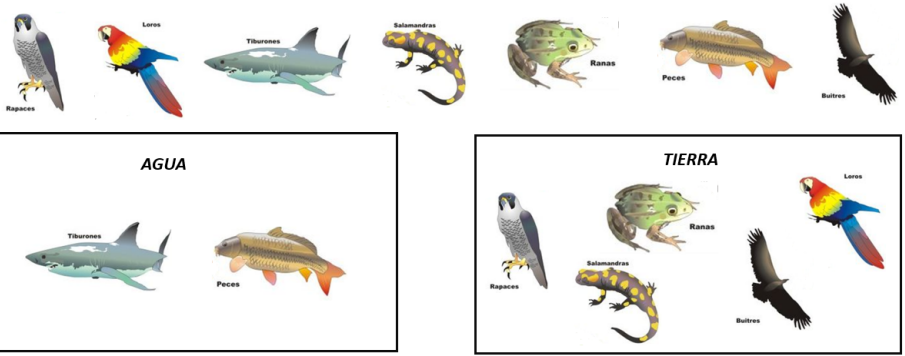




Clustering – formas de agrupar los datos



Clustering – formas de agrupar los datos



No hay una sola forma de agrupar los datos, el Clustering es **subjetivo**

Clustering – ¿Cuántos grupos formamos?

¿Cuántos clusters?

Clustering – ¿Cuántos grupos formamos?

¿Cuántos clusters?

Dos clusters

Clustering – ¿Cuántos grupos formamos?

¿Cuántos clusters?

Dos clusters

Cuatro Clusters

Clustering – ¿Cuántos grupos formamos?

¿Cuántos clusters?

Seis clusters

Dos clusters

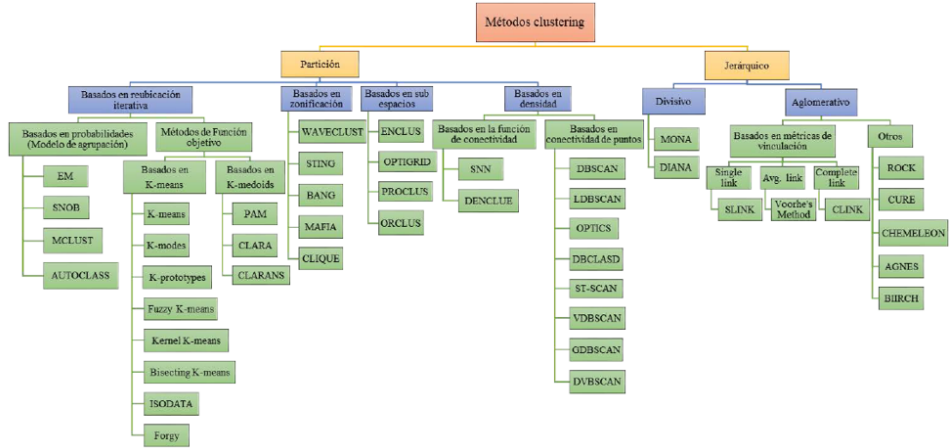
Cuatro Clusters

¿Cómo medimos la similitud? - Medidas de distancia

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. Una medida de distancia en este espacio es una función d(x; y) que toma dos puntos en el espacio como argumentos y produce un número real, y satisface los siguientes axiomas:

1. $d(x, y) \geq 0$ (distancias no negativas)
2. $d(x, y) = 0$ si y solo si $x = y$
3. $d(x, y) = d(y, x)$ (la distancia es simétrica)
4. $d(x, y) \leq d(x, z) + d(z, y)$ (La desigualdad de triángulo)

Técnicas de clustering



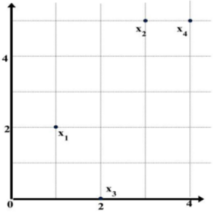
Medidas de distancia

Distancia Euclidiana

La distancia euclideana es la medida más popular d_{ij} que se puede utilizar para dos registros i, j y se define por:

$$d(\mathrm{x_j}, \mathrm{y_i}) = \sqrt{\sum_{i=1}^{d} (x_j - x_i)^2}$$

donde x_i es el valor del caso x en la variable i



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distancia de Manhattan

La Distancia de Manhattan es la suma de las diferencias absolutas entre puntos en todas las dimensiones. De una manera simple de decir que es la suma total de la diferencia entre las coordenadas x y las coordenadas y:

$$d(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

Distancia de Minkowski

La distancia de Minkowski es la forma generalizada de la distancia euclidiana y de Manhattan. Si p $= 2$ en la ecuación anterior tenemos la distancia euclidiana, mientras que para p $= 1$ tenemos lo que generalmente se conoce como la distancia de Manhattan.

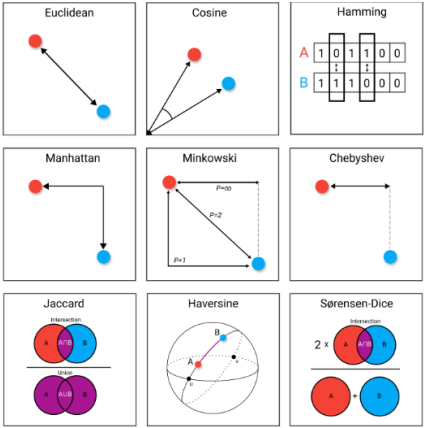
$$d(\mathrm{x}, \mathrm{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

Distancia de Chebyshev (Suprema)

La distancia de Chebyshev se define como la mayor diferencia entre dos vectores a lo largo de cualquier dimensión de coordenadas. En otras palabras, es simplemente la distancia máxima a lo largo de un eje.

$$d(i,j) = \max_{1 \leq k \leq M} |x_{ik} - x_{jk}|$$

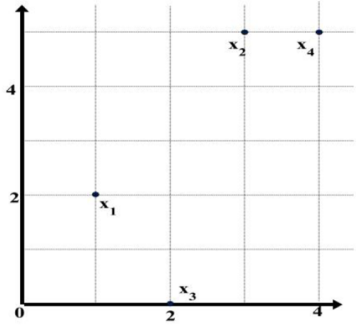
Otras medidas de distancia



- Revisar: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Comparación distancias

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Funciones en R

Function	Package	Option	Distance
`dist`	**stats**	`method = "euclidean"`	Euclidean
		`method = "minkowski"`	Minkowski
		`method = "manhattan"`	Minkowski with $q = 1$
		`method = "maximum"`	Minkowski with $q = +\infty$
`daisy`	**cluster**	`metric = "euclidean"`	Euclidean
		`metric = "manhattan"`	Minkowski with $q = 1$

Estandarización de los datos

¿Porqué estandarizar?
Las medidas de similaridad son muy sensibles a las unidades que estén medidas dichas variables. Para evitar esta influencia no deseable de una variable debida exclusivamente a la unidad en que viene medida, es necesario corregir el efecto de los datos recurriendo a un proceso de estandarización.

Observación	Hijos	Ingreso
A	2	1000
B	4	1050
C	12	1000

Proceso de decisión del análisis de clústeres

- Etapa 1: Objetivos de análisis clúster
- Etapa 2: Búsqueda del diseño en análisis clúster
- Etapa 3: Supuestos en el análisis clúster
- Etapa 4: Desarrollo de los clústers y evaluación del encaje general
- Etapa 5: Interpretación de los clústers
- Etapa 6: Valoración y caracterización de los clústeres

Críticas más comunes - clustering

El análisis de conglomerados es descriptivo, ateórico y no inferencial.

El análisis de conglomerados no tiene una base estadística sobre la cual extraer inferencias de una muestra a una población, y muchos sostienen que es solo una técnica exploratoria.

El análisis de conglomerados siempre creará conglomerados, independientemente de la existencia real de cualquier estructura en los datos.

Cuando se utiliza el análisis de conglomerados, el investigador asume alguna estructura entre los objetos.

La solución de conglomerados no se puede generalizar porque depende totalmente de las variables utilizadas como base para la medida de similitud.

Esta crítica puede hacerse contra cualquier técnica estadística, pero el análisis de conglomerados generalmente se considera más dependiente de las medidas utilizadas para caracterizar los objetos que otras técnicas multivariadas.

Clustering de Particionamiento

Clustering de Particionamiento

Son métodos clustering usados para clasificar observaciones de un conjunto de datos en múltiples grupos basado en su similaridad.

Los algoritmos conocidos requieren especificar el número de cluster a ser generados. Entre los principales métodos se tienen:

- k-means
- Partitioning Around Medoids, K-medoides (PAM)
- Clustering Large Applications (CLARA).

K-means

- K-means es una técnica para encontrar y clasificar K grupos de datos (clusters).
- Donde los elementos de similares características estarán juntos en un mismo grupo, separados de los otros grupos con los que no comparten características.
- Para saber si los datos son parecidos o diferentes el algoritmo K-means se utiliza una medida de distancia entre los datos.

K-means

- El objetivo de este algoritmo es encontrar grupos en los datos, con el número de grupos representados por la variable K.
- El algoritmo funciona de forma iterativa para asignar cada punto de datos a uno de los grupos K en función de las características que se proporcionan.
- Los puntos de datos se agrupan según la similitud de características.
- Los resultados del algoritmo de agrupación K-medias son:
  2. Los centroides de los grupos K , que se pueden usar para etiquetar nuevos datos.
  3. Etiquetas para los datos de entrenamiento (cada punto de datos se asigna a un solo grupo).

Principios del k-means

La idea básica del k-means consiste en definir clusters tal que la variación total intra-cluster sea mínima.

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

La variación total intra-cluster (total within-cluster)

$$tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Variación dentro del clúster

La variación dentro del clúster para el clúster C_k es una medida $W(C_k)$ de la cantidad en que las observaciones dentro del clúster difieren unas de otras. De aquí que se quiere resolver el problema.

$$\underset{C_1,\ldots,C_k}{Minimizar} \left\{ \sum_{k=1}^{K} W\left(C_k\right) \right\}$$

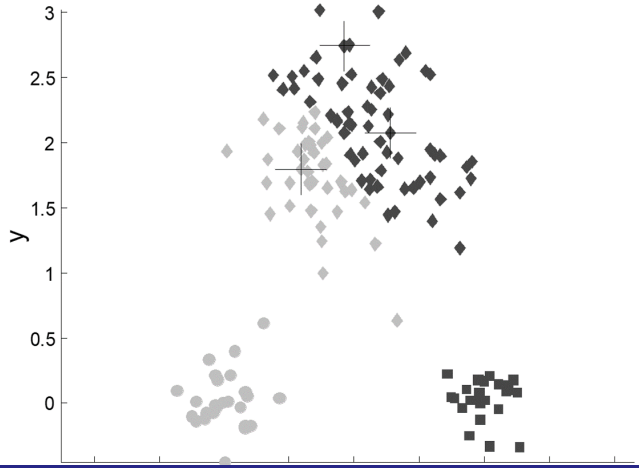
En otras palabras la fórmula indica que se desea particionar las observaciones en K clústers de tal forma que la variación total dentro del clúster sea lo más pequeña posible. Resolver la ecuación parece razonable, pero es necesario definir la variación dentro del clúster.

Algoritmo K-means

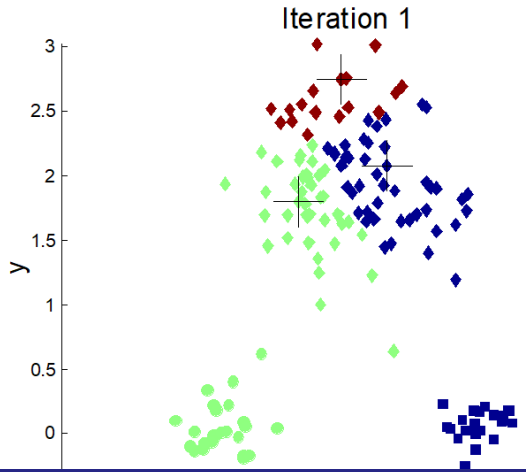
1. Selecciona K puntos como centroides iniciales.
2. Forma grupos K asignando cada punto a su centroide más cercano.
3. Recalcule el centroide de cada grupo.
4. repite

- Hasta que los Centroides no cambien.

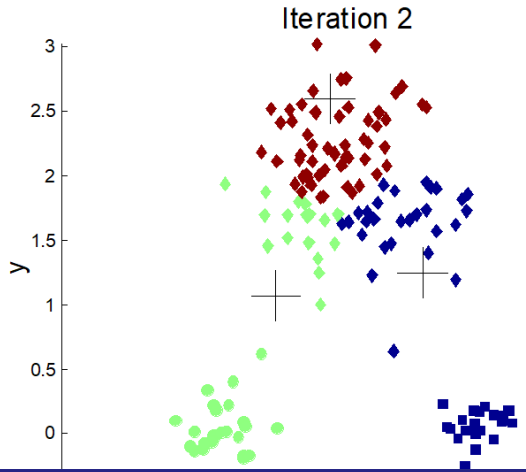
Ejemplo algoritmo K-means Clustering



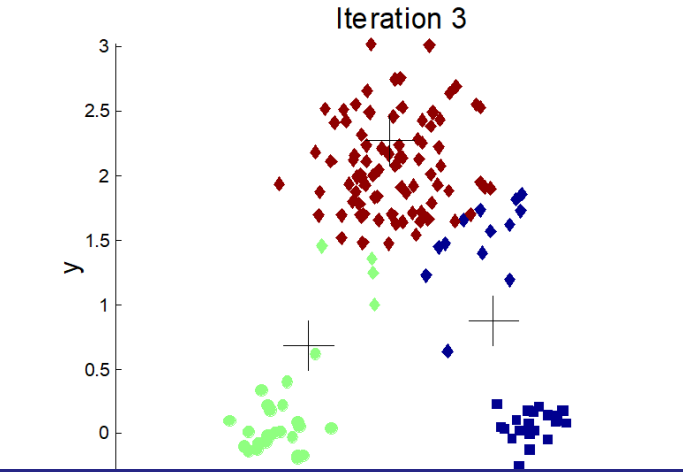
Ejemplo algoritmo K-means Clustering



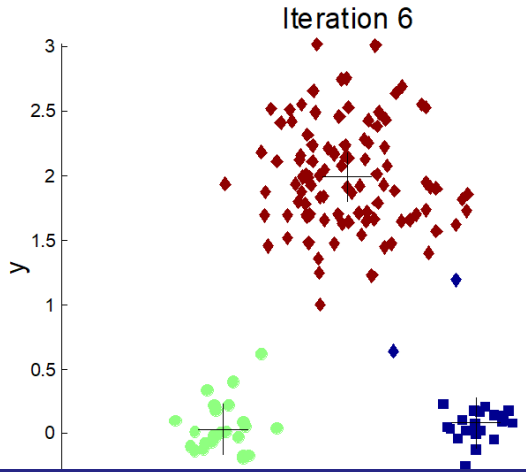
Ejemplo algoritmo K-means Clustering



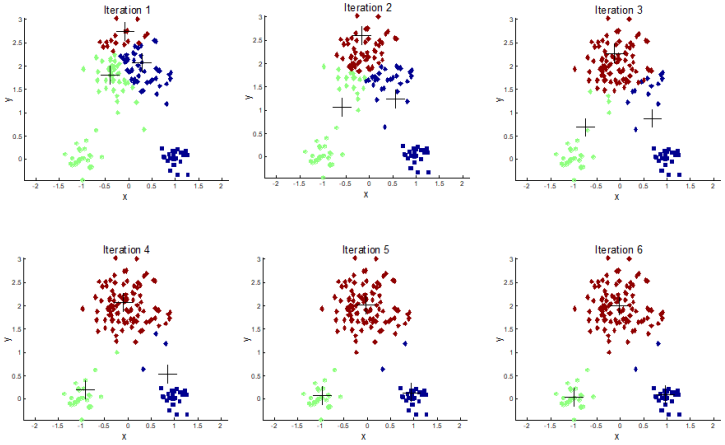
Ejemplo algoritmo K-means Clustering



Ejemplo algoritmo K-means Clustering



Ejemplo algoritmo K-means Clustering



Determinando el número de clusters

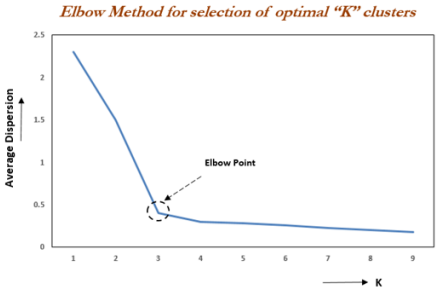
Criterio de la Suma de Cuadrados (SSE)

1. Aplicar el algoritmo cluster (por ejemplo, k-means) para diferentes valores de k. (por ejemplo, variar k de 1 a 10 clusters)
2. Para cada k, calcular la suma de cuadrados total dentro de clusters (wss).
3. Graficar los valores de wss de acuerdo con el número de clusters k.
4. La ubicación del punto de inflexión en el gráfico se considera generalmente como el indicador de la cantidad adecuada de clusters.

Criterio de la Suma de Cuadrados (SSE)

El método del codo (Elbow point):

Mediante este método se traza el valor de la función de costo producida por diferentes valores de k. Si k aumenta, la distorsión promedio disminuirá, cada grupo tendrá menos instancias constituyentes, y las instancias estarán más cerca de sus respectivos centroides. El valor de k en el cual la mejoría en la distorsión disminuye más se llama codo, en el cual deberíamos dejar de dividir los datos en grupos adicionales.



Evaluación de grupos con coeficiente de silueta

1. Aplicar el algoritmo cluster (por ejemplo, k-means) para diferentes valores de k. (por ejemplo, variar k de 1 a 10 clusters)
2. Para cada k, calcule el average silhouette de las observaciones (avg.sil).
3. Grafique los valores de avg.sil de acuerdo con el número de clusters k.
4. La ubicación del punto máximo se considera como el número apropiado de clusters.

Evaluación de grupos con coeficiente de silueta

El coeficiente de silueta es una medida de la compacidad y separación de los grupos.

- Los valores más altos representan una mejor calidad del clúster.
- El coeficiente de silueta es más alto para grupos compactos que están bien separados y más bajo para grupos superpuestos.
- Los valores del coeficiente de silueta cambian de -1 a $+1$, y cuanto mayor sea el valor, mejor.

El coeficiente de Silueta para una observación i se denota como s(i) y se define como:

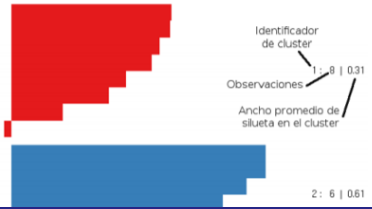
$$s(i) = \frac{b - a}{max(a, b)}$$

Dónde: * a es la distancia media entre el objeto y todos los otros objetos de la misma clase * b es la distancia media entre el objeto y todos los otros objetos del clúster más próximo

Evaluación de grupos con coeficiente de silueta

Medición para valores del índice:

- $0.71 - 1.0$, las estructuras encontradas son sólidas.
- $0.51 - 0.70$, las estructuras encontradas con razonables.
- $0.26 - 0.50$, las estructuras encontradas son débiles y tienden a ser artificiales. Se deberían intentar métodos alternativos para el análisis de los datos.
- < 0.25, no se encuentran estructuras



Criterio de Calinski-Harabasz

El índice de Calinski-Harabasz, también conocido como criterio de la razón de varianza, es la razón de la suma de dispersión entre grupos y de dispersión entre grupos para todos los grupos, cuanto mayor sea la puntuación, mejores serán los rendimientos.

$$\frac{(n-k) \times \text{Suma de cuadrados entre los grupos}}{(k-1) \times \text{Suma de cuadrados dentro de los grupos}}$$

Validación de los clusters

- La solución para encontrar el mejor algoritmo clustering y el número óptimo de conglomerados k se llama generalmente validez del cluster.
- Una vez que la partición se obtiene mediante un método de agrupación, la función de validez cuantifica la precisión de la estructura del conjunto de datos
- La validación de clusters hace referencia al procedimiento de evaluación de la bondad de los resultados del algoritmo de cluster.
- Esto se hace para evitar encontrar patrones en una data aleatoria, así como, para comparar dos algoritmos cluster.

Estadísticos para validación de clusters

Validación interna

usa la información interna del proceso de cluster para evaluar la bondad de una estructura de agrupamiento.

Validación externa

consiste en comparar los resultados de un análisis de clúster con un resultado conocido externamente, como las etiquetas de una clase proporcionada externamente.

Validación relativa

evalúa la estructura del cluster variando diferentes valores de parámetros para el mismo algoritmo.

Medidas internas para validación de clusters

Compactibilidad o cohesión del clúster

mide cuán cerca están los objetos dentro del mismo cluster. Una menor variación dentro del clúster es un indicador de una buena compactibilidad (es decir, una buena agrupación).

Separación

mide qué tan bien separado está un cluster de los otros clusters.

Conectividad

corresponde al grado en que los elementos se colocan en el mismo clúster como con sus vecinos más cercanos en el espacio de datos. La conectividad tiene un valor entre 0 e infinito y debe ser minimizado.

Comparando usando medidas internas

Usa información intrínseca de la data para evaluar la calidad del agrupamiento. Las medidas internas incluyen:

- El coeficiente de Silueta
- El índice de Davies-Bouldin
- El índice de Dunn

Índice de Validación de Davies-Bouldin

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

- Donde n es el número de clústeres, c_x denota el centro del clúster x, σ_x es la distancia media de todos los elementos del clúster x al centro c_x, y $d(c_i\ ,\ c_j)$ es la distancia entre los centroides c_i y c_j
- El máximo valor de este índice $máx_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$ representa el peor caso para el clúster i.
- La solución óptima es aquella que tiene el índice de Davies Bouldin más bajo

Índice de Dunn

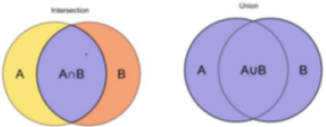
1. Para cada clúster se calcula la distancia entre cada uno de los objetos en el clúster y los objetos en los otros clusters.
2. Usa el mínimo de esta distancia por parejas como la separación entre clusters (mínima separación)
3. Para cada clúster se calcula la distancia entre los objetos en el mismo clúster.
4. Use el máximo de esta distancia dentro del clúster (es decir, el máximo diámetro) como medida de la compactibilidad.

$$D = \frac{(mínima\ Separación)}{(maximo\ Diámetro)}$$

$$ID = \frac{\min\left(\min_{1 \le i \le j \le k} d(c_i, c_j)\right)}{\max\left(\max_{1 \le i \le k} \Delta_k\right)}$$

Evaluación de Clusters - Bootstrap

- El algoritmo utiliza el coeficiente de Jaccard, una medida de similitud entre conjuntos. La similitud de Jaccard entre dos conjuntos A y B es la relación del número de elementos en la intersección de A y B sobre el número de elementos en la unión de A y B.



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Desventajas de K-means

- No se satisface el criterio de optimización globalmente, solo produce un optimo local.
- Es sensible a la elección de los centroides iniciales.
- Es sensible a “outliers”.