Tecnológico de Monterrey

Aprendizaje automático

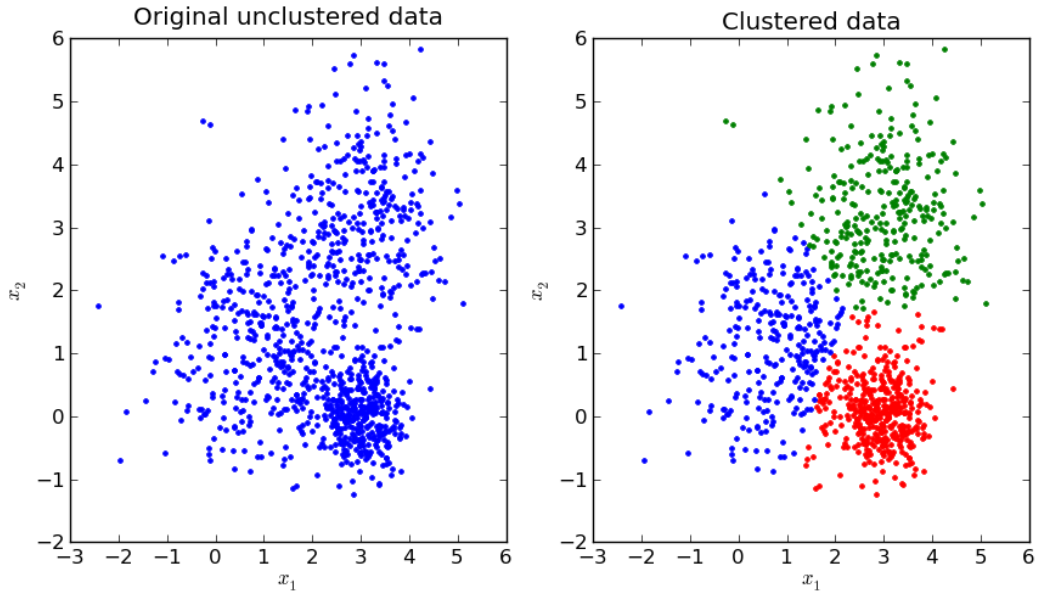
Agrupamiento

Dr. Omar Mendoza Montoya

04/09/2024

Agrupamiento

Análisis de grupos (cluster analysis), también llamado segmentación de datos, consiste en agrupar o separar una colección de objetos en subconjuntos (llamados clústeres), de tal forma que los elementos de cada grupo están más relacionados entre ellos que con elementos de otros grupos.



¿Por qué hacer agrupamiento?

- El análisis de grupos es utilizado como una forma de estadística descriptiva para mostrar si un conjunto de datos está conformado por subgrupos, donde cada uno de ellos representa objetos con propiedades substancialmente distintas a las de los otros subgrupos.

- Cuando se trabaja con conjuntos de datos grandes, una manera eficiente de analizarlos consiste primero en dividir los datos en grupos con observaciones con una relación lógica.

- Agrupamiento es una forma eficiente para el descubrimiento de conocimiento en forma de patrones recurrentes y reglas.

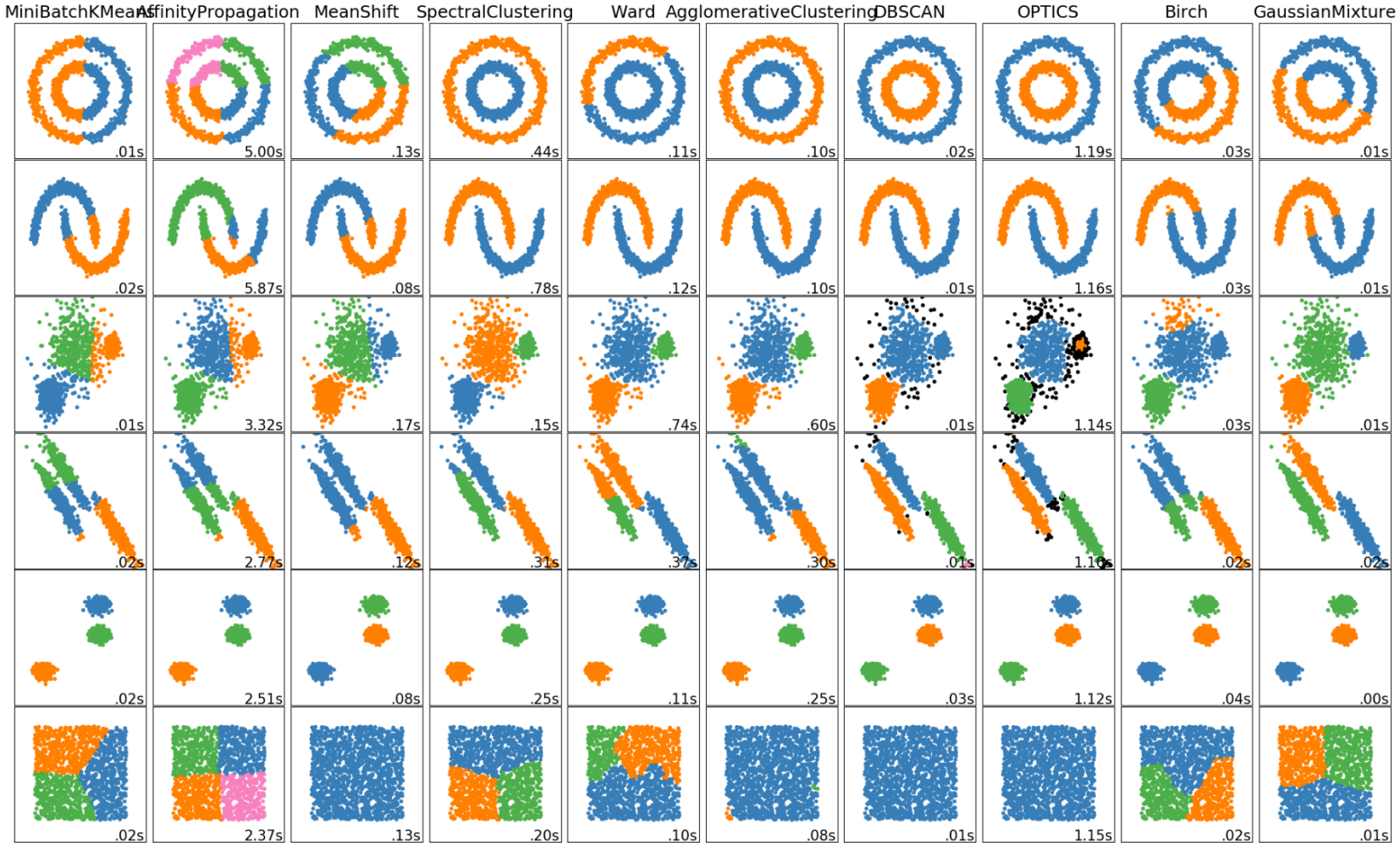
¿Cómo operan los algoritmos de agrupamiento?

Algoritmos de agrupamiento

- Existe una gran cantidad de algoritmos de agrupamiento, los cuales cada uno tiene un entendimiento propio sobre cómo generar grupos de observaciones similares.

- Los algoritmos de agrupamiento suelen trabajar con medidas de disimilitud (distancia) y medidas de similitud (semejanza). Con dichas medidas se establece qué objetos son similares o diferentes para generar grupos.

Algoritmos de agrupamiento



Medidas de distancia y de semejanza

Medidas de distancia

Ejemplos de medidas de disimilitud o distancia entre dos objetos $X = [x_1, x_2, x_3, \dots, x_p]$ y $Y = [y_1, y_2, y_3, \dots, y_p]$ con p atributos o variables.

- Distancia de Minkowski

$$D(X,Y) = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{1/m}$$

- Distancia Euclidiana

$$D(X,Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Medidas de distancia

- Distancia Manhattan (city block)

$$D(X,Y) = \sum_{i=1}^{p} |x_i - y_i|$$

- Distancia de Hamming

$$D(X,Y) = \sum_{i=1}^{p} I(x_i \neq y_i)$$

donde $I(a \neq b)$ es una función indicadora, la cual devuelve 1 si a y b son diferentes, y 0 en caso contrario.

- Distancia angular

$$D(X,Y) = \frac{1}{\pi} \arccos\left(\frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} {x_i}^2} \sqrt{\sum_{i=1}^{p} {y_i}^2}} \right)$$

Matriz de distancias

Para un conjunto de datos $\{X_1, X_2, X_3, \dots, X_n\}$, dada una función de distancia $D(X, Y)$, es posible calcular la matriz de distancias entre todos los objetos.

$$D = \begin{bmatrix} D(X_1, X_1) & D(X_1, X_2) & \cdots & D(X_1, X_n) \\ D(X_2, X_1) & D(X_2, X_2) & \cdots & D(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ D(X_n, X_1) & D(X_n, X_2) & \cdots & D(X_n, X_n) \end{bmatrix}$$

Típicamente, $D(X_i, X_i) = 0$ y $D(X_i, X_j) = D(X_j, X_i)$.

Medidas de semejanza

Ejemplos de medidas de semejanza entre dos objetos $X = [x_1, x_2, x_3, \dots, x_p]$ y $Y = [y_1, y_2, y_3, \dots, y_p]$ con p atributos o variables.

- Producto punto

$$S(X,Y) = \sum_{i=1}^{m} x_i y_i$$

- Semejanza Coseno

$$S(X,Y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} {x_i}^2}\sqrt{\sum_{i=1}^{m} {y_i}^2}}$$

Medidas de semejanza

- Semejanza angular

$$S(X,Y) = 1 - \frac{1}{\pi}\arccos\left(\frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} {x_i}^2}\sqrt{\sum_{i=1}^{m} {y_i}^2}}\right)$$

- Semejanza de Jaccard (X y Y son conjuntos, no vectores)

$$S(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

- Similaridad de Hamming

$$S(X,Y) = \sum_{i=1}^{m} I(x_i = y_i)$$

donde $I(a = b)$ es una función indicadora, la cual devuelve 1 si a y b son iguales, y 0 en caso contrario.

Medidas de semejanza

- Función de Shepard

$$S(X,Y) = e^{-D(X,Y)}$$

donde $D(X,Y)$ es una función de distancia.

La función de Shepard permite transformar cualquier medida de distancia en una medida de semejanza.

Matriz de semejanzas

Para un conjunto de datos $\{X_1, X_2, X_3, \ldots, X_n\}$, dada una función de distancia $D(X, Y)$, es posible calcular la matriz de semejanzas entre todos los objetos.

$$S = \begin{bmatrix} S(X_1, X_1) & S(X_1, X_2) & \cdots & S(X_1, X_n) \\ S(X_2, X_1) & S(X_2, X_2) & \cdots & S(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ S(X_n, X_1) & S(X_n, X_2) & \cdots & S(X_n, X_n) \end{bmatrix}$$

Típicamente, $S(X_i, X_j) = S(X_j, X_i)$.

Tipos de algoritmos de agrupamiento

Algoritmos de agrupamiento

- Modelos de conectividad. Agrupan las observaciones tomando en cuenta la proximidad entre ellas. Ejemplo: dendrogramas (agrupamiento jerárquico).

- Modelos de distribución. Están basados en la probabilidad de que las observaciones que pertenecen a un clúster pertenezcan a la misma distribución de probabilidad. Ejemplo: Mezcla de Gaussianas.

Algoritmos de agrupamiento

- Modelos de densidad. Estos modelos tratan de aislar las regiones del espacio en los que haya una gran densidad de observaciones. DBScan y Optics son ejemplos de algoritmos de este tipo.

- Modelos de centroides. Son modelos iterativos donde la semejanza de las observaciones se deriva por su cercanía con el centroide de cada clúster. K-medias es el ejemplo más conocido de esta categoría.

Ejemplos de algoritmos de agrupamiento

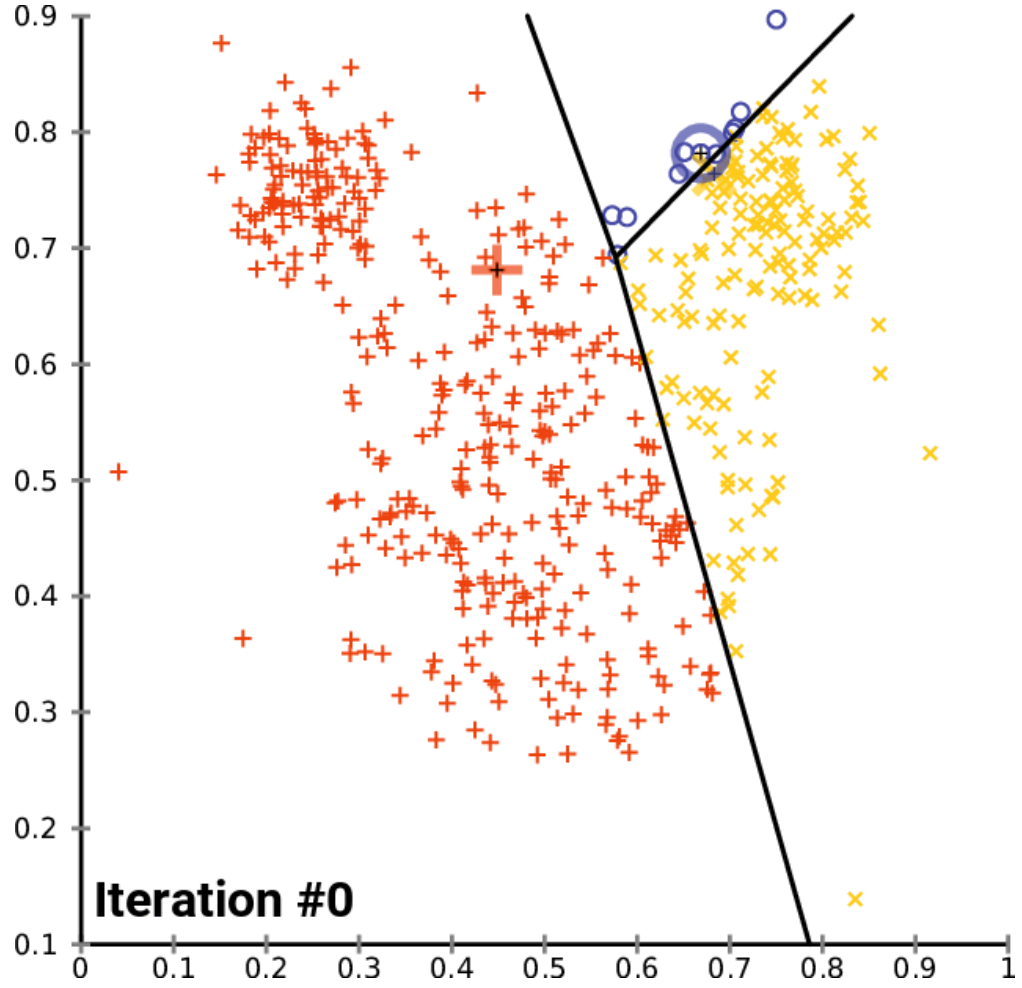
K-Medias

- Este algoritmo se utiliza en situaciones donde todas las variables o predictores de los datos son variables cuantitativas u ordinales. Este algoritmo se basa en la distancia euclidiana como función de distancia.

- Dado un conjunto inicial de k medias, el algoritmo repite los siguientes pasos hasta convergencia:

  - Paso de asignación. Cada observación es asignada a la media más cercana.
  - Paso de actualización. Las k medias son recalculadas de acuerdo al conjunto de observaciones de cada clúster.

K-medias



Wikipedia

El resultado de aplicar el algoritmo k-medias depende en la selección del número de clústeres y la configuración inicial.

Sin embargo, si los datos tienen una estructura de grupos, estos se revelan independientemente de la configuración del algoritmo.

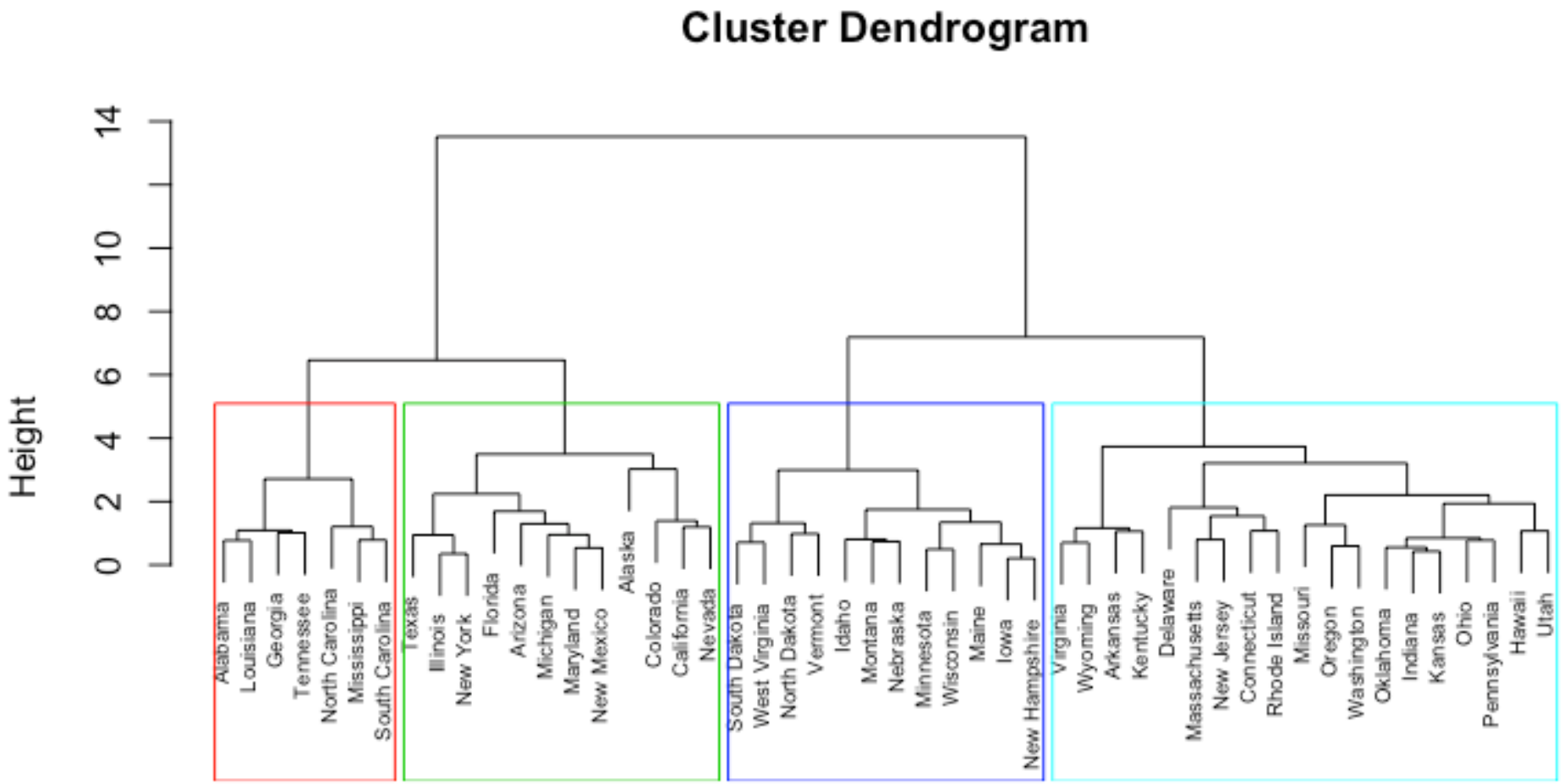
Agrupamiento jerárquico

- En el agrupamiento jerárquico se requiere una medida de distancia entre grupos de observaciones, basada en la distancia entre observaciones de dos grupos.

- La jerarquía de clústeres se suele representar como un árbol (dendrograma). La raíz del árbol es un único clúster que agrupa a todas las observaciones, y las hojas son grupos con una sola observación.

- Cada nivel de la jerarquía representa un agrupamiento particular de los datos en la que se separan dos grupos.

Agrupamiento jerárquico



Estrategias de agrupamiento jerárquico

- Aglomerativos (bottom-up). Se comienza desde abajo en la jerarquía, y recursivamente se unen pares de clústers en uno solo.

- Divisivos (top-down). Se comienza en la parte superior del árbol, y en cada nivel se divide cada uno de los clústeres en dos nuevos.

Agrupamiento aglomerativo

- En agrupamiento aglomerativo, en cada iteración, los dos clústeres más cercanos se unen en un solo clúster.

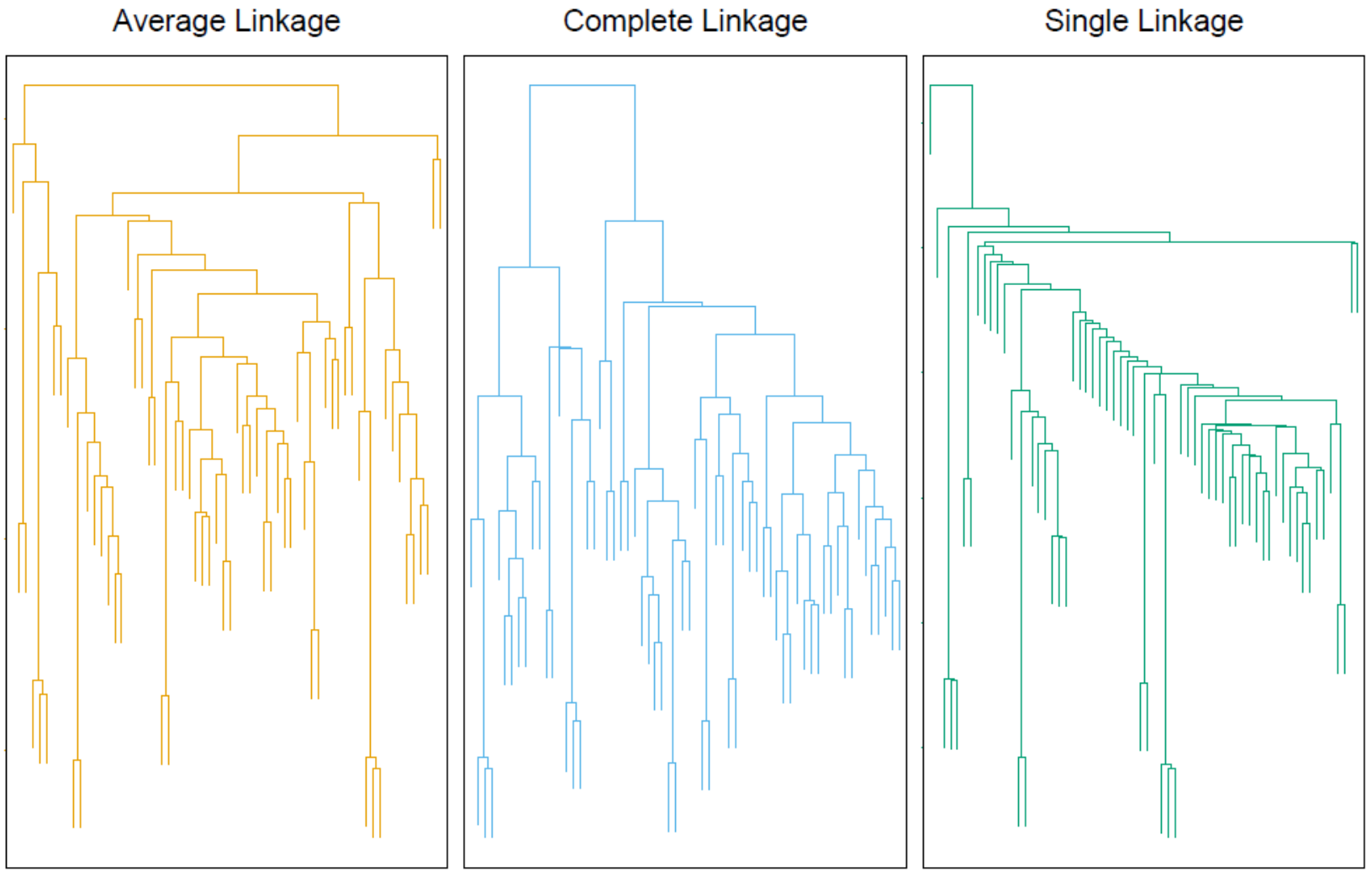
- Para ello, hay tres estrategias:

  - Single linkage, la cual toma la disimilaridad entre grupos como la distancia entre las parejas más cercanas entre ambos grupos.

  - Complete linkage, la cual toma la disimilaridad entre grupos como la distancia entre las parejas más lejanas entre ambos grupos.

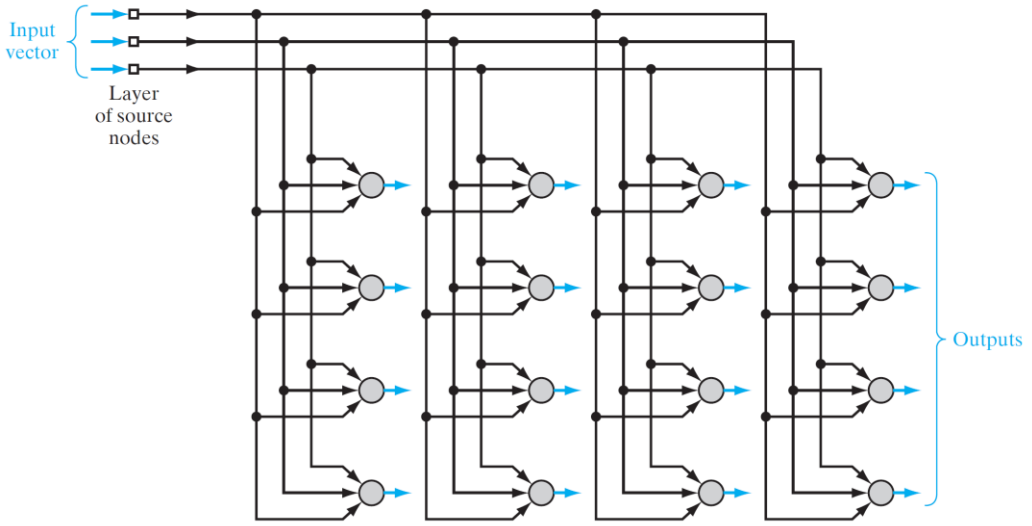
  - Group average, la cual toma la disimilaridad entre grupos al promedio de las distancias entre parejas de los grupos.

Agrupamiento jerárquico



Mapas autoorganizados

- Estas redes están basadas en el aprendizaje competitivo, de tal forma que las neuronas pelean por ser activadas para una entrada dada.
- Sólo una neurona logra la activación, la cual es llamada "winner-takes-all neuron" o simplemente neurona ganadora.
- En el entrenamiento, las neuronas son ajustadas para que respondan selectivamente a patrones (estímulos) o clases de patrones de los datos de entrada.



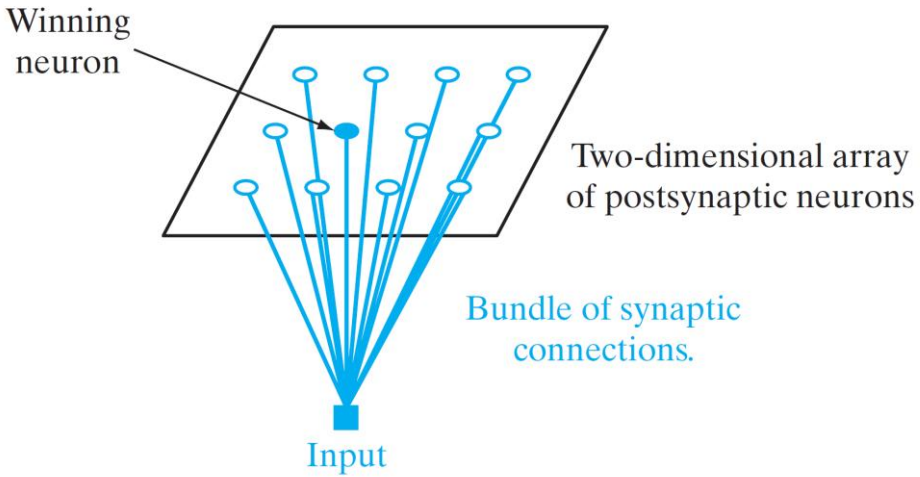
Haykin, 2009

Mapas autoorganizados

- La arquitectura de este tipo más conocida es la de Kohonen, y resultan muy útiles para tareas de agrupamiento.

- Las neuronas se colocan en una retícula o alguna distribución geométrica bidimensional o tridimensional, de tal forma que cada posición de la retícula corresponde a una neurona.

- La distribución geométrica de las neuronas permite que haya una relación vecinal entre ellas, de tal forma que neuronas vecinas reaccionan a patrones similares.



(b) Kohonen model

Haykin, 2009

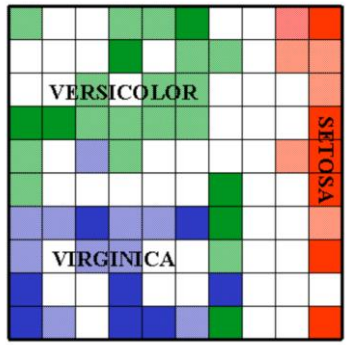
Mapas autoorganizados

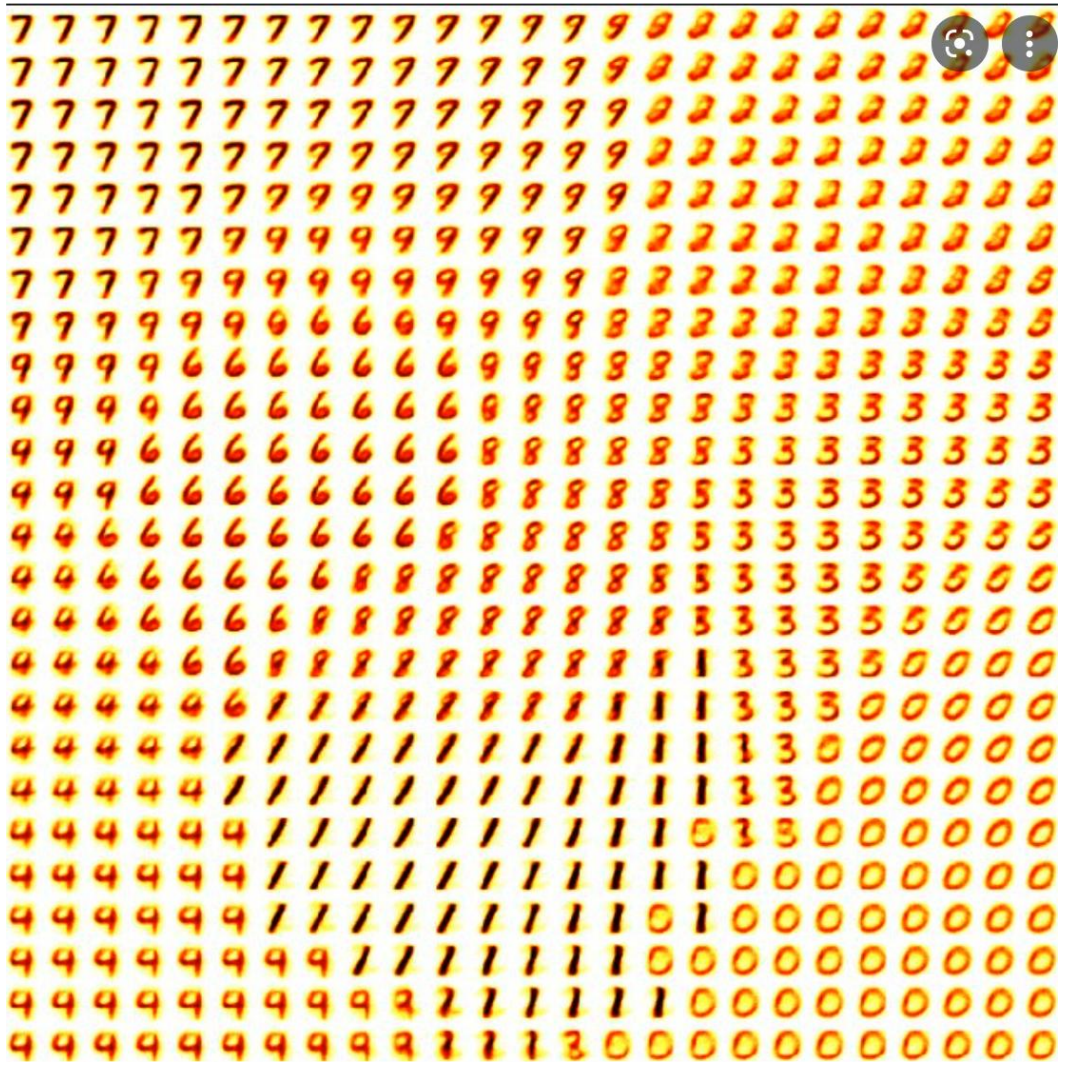
- El funcionamiento de esta red está basado en tres principios:
  - Competición: Para cada patrón de entrada, las neuronas de la red compiten de acuerdo a su función discriminante. Aquella con mayor valor se declara ganadora.
  - Cooperación. La neurona ganadora determina la posición espacial de la vecindad donde se localiza, proporcionando con ello las bases de una cooperación con los vecinos. Esto se inspira en el hecho que neuronas vecinas en el cerebro tienden a excitar a sus vecinos con mayor probabilidad que a neuronas lejanas.
  - Adaptación de pesos sinápticos. Tanto la neurona ganadora como sus vecinos son ajustadas para que incrementen su respuesta ante el mismo patrón o estímulo. Esto mejora la salida de la neurona ganadora y de sus vecinos.

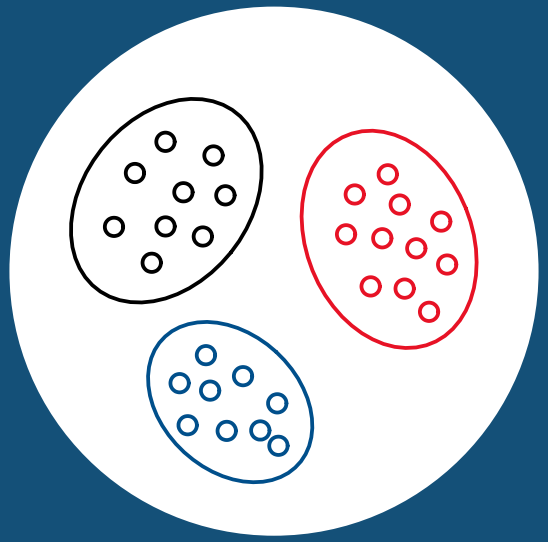
Mapas autoorganizados

- Posiblemente una de las características más atractivas de un mapa autoorganizado es la manera como se presentan los datos.

- Usualmente se muestran las observaciones en el mismo mapa de tal forma que se ve la manera como se organizaron las neuronas ante los datos.







¿Cómo encuentro el número de grupos que debo indicar a un algoritmo de agrupamiento?

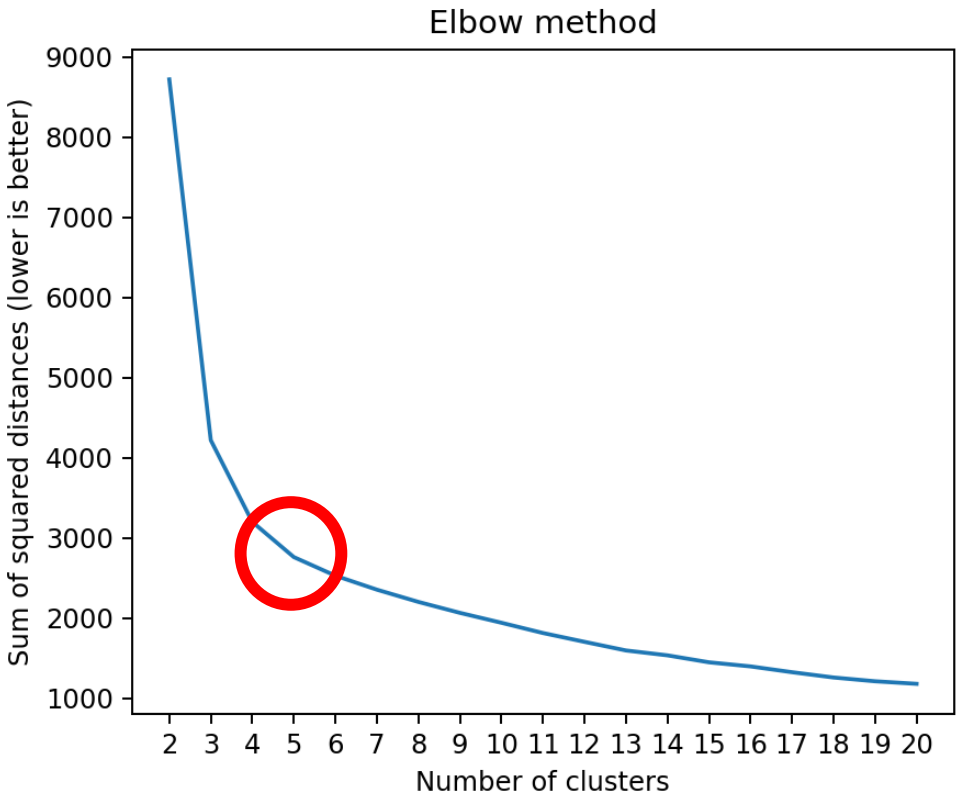
El método del codo

- Se calcula la suma de los errores cuadráticos para diferentes números de grupos:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde C_i es el conjunto de observaciones del grupo i, k es el número de grupos, y μ_i es el promedio de las observaciones del grupo i.

- Se verifica en qué parte de la gráfica la variación entre el valor SSE empieza en menor medida (codo de la gráfica).
- En el codo está el número óptimo de grupos.

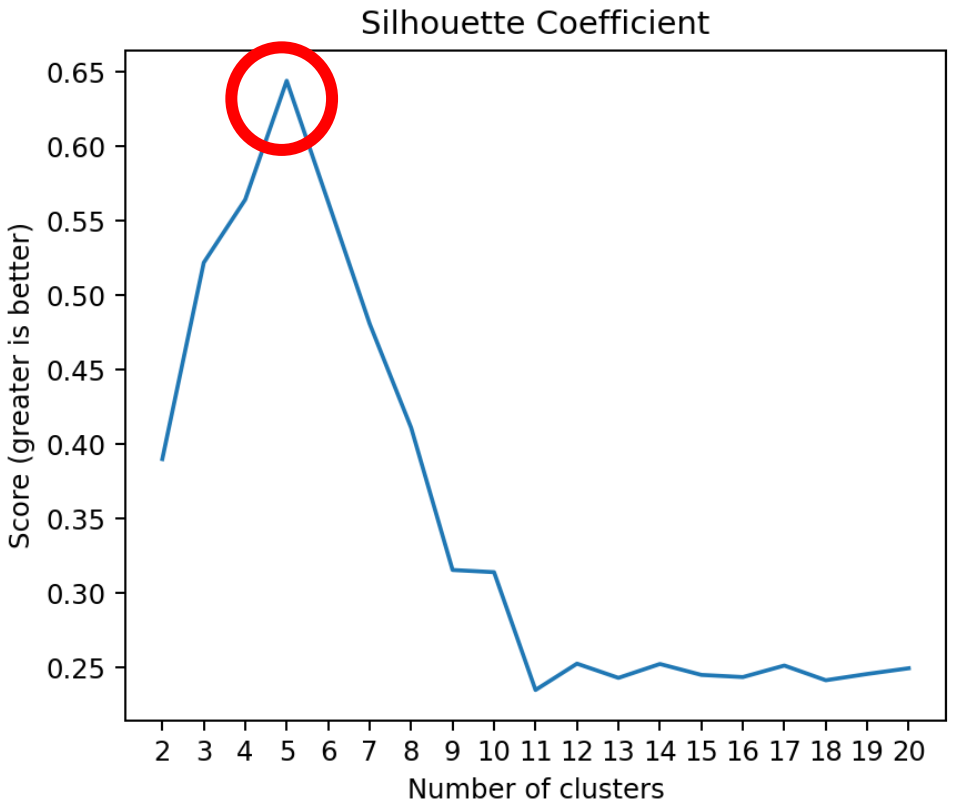


El coeficiente de Silhouette

- Se calcula el coeficiente de Silhouette promedio para diferentes números de grupos.
- El coeficiente de Silhouette para una observación i está dado por la siguiente relación:

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

donde $a(i)$ es el promedio de las distancias entre el punto i y el resto de los puntos dentro del mismo grupo, y $b(i)$ es el promedio de las distancias entre el punto i y los puntos del grupo más cercano al grupo de i.



- El coeficiente combina la cohesión y la separación entre grupos para saber qué tan bien están agrupados.

El índice de Calinski-Harabasz

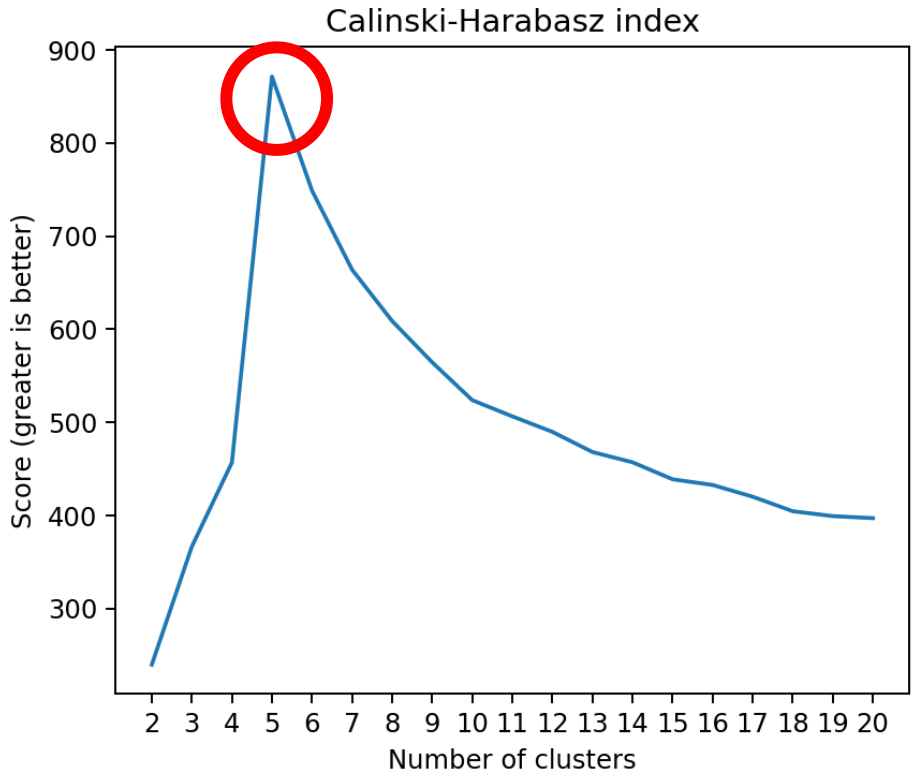
- El índice de Calisnki-Harabasz es una medida de la calidad de un agrupamiento (entre más grande mejor).
- Se calcula como :

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

$$SSB = \sum_{i=1}^{k} n_i \|\mu_i - \mu\|^2$$

$$CH = \frac{SSB/(k-1)}{SSE/(N-k)}$$

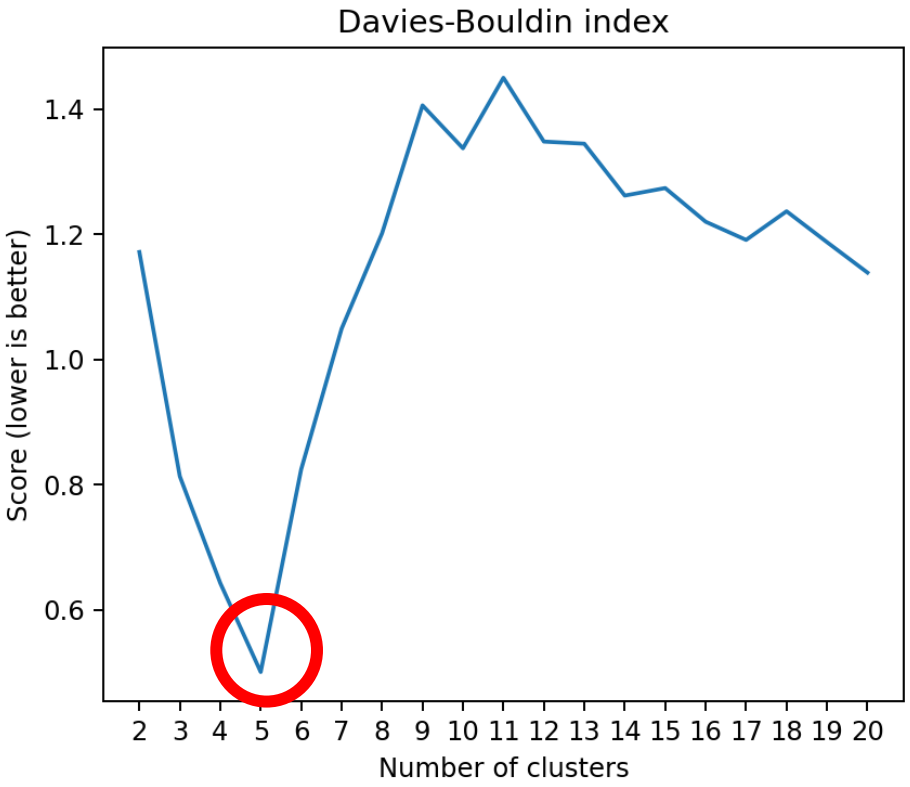
donde C_i es el conjunto de observaciones del grupo i, k es el número de grupos, μ_i es el promedio de las observaciones del grupo i, μ es el promedio de todas las observaciones, n_i es el número de observaciones del grupo i y N es el número total de observaciones.



Índice de Davies-Bouldin

- El índice de Davies-Bouldin es útil porque proporciona una medida de la relación entre la compacidad de los clústeres y su separación, considerando tanto la variación dentro de los clústeres como la distancia entre los clústeres.

- Al comparar diferentes resultados de clustering, se prefiere el modelo con el índice DBI más bajo.



No siempre es fácil determinar el número de grupos de una manera directa ya que depende de la necesidad del investigador.

Las medidas puedes ser útiles, pero no dan la última palabra.

Bibliografía

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 12

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 14

- Haykin , S. (2009). *Neural networks and learning machines* (3ra ed.). Person Education.
  - Capítulo 9, páginas 425-474

- *Clustering algorithms in machine learning:* https://www.mygreatlearning.com/blog/clustering-algorithms-in-machine-learning/