Aprendizaje no supervisado

Minería de Datos - Grado en Matemáticas

DSLAB

2025-10-13

Introducción

- Clustering es el proceso de agrupar objetos similares
- **Objetivo**: particionar el conjunto de datos en grupos de observaciones donde cada observación se parezca lo más posible a las observaciones de su mismo grupo y lo menos posible a las observaciones de los otros grupos
- Grupos se llaman conglomerados o clústeres
- Ejemplos: segmentación del mercado, visualización, detección de anomalías, imputación de valores faltantes, etc.
- Se encuadra dentro del aprendizaje basado en similitud o semejanza
- No se dispone de etiqueta, no hay clases a aprender

Dos tipos de inputs y dos objetivos

El input del algoritmo puede ser de dos tipos:

- Una matriz de desemejanza o matriz de distancia D de tamaño $n \times n$ que refleje la disimilitud entre las observaciones. Este clustering es conocido como clustering basado en similitud (similarity-based clustering)
- Una matriz X de tamaño $n \times m$ con los datos originales. n es el número de observaciones y m es el número de variables. Este clustering es conocido como clustering basado en características (feature-based clustering) y busca lograr agrupaciones homogéneas de variables.

Clasificación (similarity-based clustering)

- **Agrupamiento jerárquico.** Se trata de estructurar los elementos de un conjunto de forma jerárquica y en función de su similitud. Una clasificación jerárquica conlleva que los datos se ordenan en niveles y los niveles superiores contienen a los inferiores. Cuando una observación forma parte de un cluster, permanece en él. Se crea un árbol anidado de particiones
- **Agrupamiento no jerárquico (particiones de los datos)**. Se dividen los datos en un número de grupos de tal modo que cada elemento pertenezca a uno y sólo uno de los grupos, todo elemento quede clasificado y cada grupo sea internamente homogéneo. En los algoritmos de partición de datos todos los clusters se encuentran de forma simultánea

Aprendizaje basado en similitud

- Cada variable descriptiva de un conjunto de datos representa una dimensión en el espacio m-dimensional
- "A feature space is an abstract m-dimensional space that is created by making each descriptive feature in the dataset an axis of an m-dimensional coordinate system and mapping each observation in the dataset to a point in this coordinate system based on the values of its descriptive features" [J. Kelleher, 2015]
- Dos observaciones con los mismos valores en sus variables descriptivas serán el mismo punto en el espacio de características
- A medida que aumentan las diferencias entre los valores de las características descriptivas de dos observaciones, también lo hace la distancia entre los puntos correspondientes en el espacio de características
- La distancia (¿qué distancia?) entre dos puntos del espacio de características es una medida útil de la similitud de las características descriptivas de las dos observaciones

Aprendizaje basado en similitud

- Elegir un espacio de características adecuado es crucial en cualquier tarea de ML
- La representación de una observación es uno de los aspectos más relevantes a tener en cuenta cuando se entrena un modelo
- Para representar la relación entre observaciones es necesario definir una métrica en el espacio de características
- La tarea de clasificación esencialmente es una tarea de comparación
- Dadas dos observaciones en un espacio de características, no existe una forma única de compararlas
- ¿Cómo se realiza dicha comparación? –> Usando medidas de similitud (semejanza), disimilitudes (desemejanza) o distancias

Medida de similitud

- En el análisis de conglomerados las agrupaciones se hacen en función de la semejanza entre los individuos

Medida de desemejanza o disimilitud

La medida de desemejanza entre dos observaciones x e y es una función $\delta(x, y)$ que cumple

- $\delta(x, y) = 0 \Leftrightarrow x = y$
- $\delta(x, y) \geq 0$ (no negatividad)
- $\delta(x, y) = \delta(y, x)$ (simetría)

Si además verifica la desigualdad triangular se trataría de una distancia (métrica):

$$\delta(x, y) \leq \delta(x, z) + \delta(y, z)$$

La idea de la distancia y de la medida de desemejanza es que cuanto mayor sean, menos parecido habrá entre los dos puntos que se estudian

Distancia euclídea

- La distancia más conocida
- Calcula la longitud de la línea recta entre dos puntos
- La distancia euclídea entre dos instancias $\mathbf{x}, \mathbf{z} \in \mathbb{R}^m$ es

$$||\mathbf{x} - \mathbf{z}||_2 = d_{\text{Euclídea}}(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{j=1}^{m} \left(\mathbf{x}_j - \mathbf{z}_j\right)^2} = \left(\sum_{j=1}^{m} \left(\mathbf{x}_j - \mathbf{z}_j\right)^2\right)^{1/2}$$

- Las distancias al cuadrado enfatizan las diferencias grandes (porque las diferencias se elevan al cuadrado)

Distancias datos continuos

- La distancia de Manhattan (distancia l_1) entre dos instancias $\mathbf{x}, \mathbf{z} \in \mathbb{R}^m$ es

$$||\mathbf{x} - \mathbf{z}||_1 = d_{Manhattan}(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{m} |\mathbf{x}_j - \mathbf{z}_j|$$

  Se denomina así puesto que en 2D se calcula contando cuántas filas y columnas hay que moverse horizontal y verticalmente para llegar de $\mathbf{x}_i$ a $\mathbf{z}_i$.

- Distancia de Minkowski (distancia l_p) entre dos instancias $\mathbf{x}, \mathbf{z} \in \mathbb{R}^m$

$$||\mathbf{x} - \mathbf{z}||_p = d_{Minkowski}(\mathbf{x}, \mathbf{z}) = \left(\sum_{j=1}^{m} (\mathbf{x}_j - \mathbf{z}_j)^p \right)^{1/p},$$

  con $p = 1, 2, \ldots, \infty$.

Distancias datos continuos

- La distancia de Minkowski o distancia l_p es una generalización de las anteriores
- En función del valor de p se logran distintas métricas de distancia. En particular, cuando $p = 1$ obtenemos la distancia de Manhattan y cuando $p = 2$ obtenemos la distancia euclídea. –> Se pueden definir infinitas distancias
- Mayores valores de p dan más énfasis a las diferencias grandes que valores pequeños de p, ya que todas las diferencias se elevan a la potencia de p
- En el extremo, cuando $p = \infty$, la métrica devuelve la diferencia máxima entre cualquiera de las variables, es decir, $||\mathbf{x} - \mathbf{z}||_\infty = max_{1 \leq j \leq m} |\mathbf{x}_j - \mathbf{z}_j|$. Se conoce como la distancia de Chebyshev.

Distancias para datos binarios

		Obsservation i		
		1	0	
Observation j	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	p

El valor a es el número de variables binarias con valores simultáneos iguales a 1 para las observaciones i y j.

Distancias para datos binarios

Czekanowski, Sørensen, Dice:

$$\frac{2a}{2a+b+c}$$

Hamman:

$$\frac{(a+d)-(b+c)}{a+b+c+d}$$

Jaccard:

$$\frac{a}{a+b+c}$$

Pearson:

$$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Russell-Rao:

$$\frac{a}{a+b+c+d}$$

'Simple Matching Coefficient':

$$\frac{a+d}{a+b+c+d}$$

Sokal & Sneath:

$$\frac{a+d}{a+(b+c)/2+d}$$

Yule:

$$\frac{ad-bc}{ad+bc}$$

Distancia para datos categóricos

- Para variables categóricas nominales, una opción es asignar una distancia de 1 si los atributos son diferentes y una distancia de 0 en caso contrario. La distancia total será la suma de dichos valores. Dicha distancia se conoce como la distancia de Hamming:

$$d_{Hamming}(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{m} I(\mathbf{x}_j - \mathbf{z}_j)$$

siendo $I(\cdot)$ la función indicatriz.

Distancia para datos categóricos

- Cuando se tienen datos categóricos nominales, es común medir la similitud entre las observaciones en términos de las frecuencias observadas en las distintas categorías (tablas de contingencia). Un ejemplo sería la **distancia de Goodall.**
- Si las variables son ordinales, hay que tener en cuenta el orden de los datos. En estos casos hay distintas alternativas: utilizar la correlación de Spearman, pasar los datos a rangos y normalizarlos al intervalo $[0, 1]$ y usar distancias para datos de intervalo.

Distancia para datos mixtos

- De forma general, cuando tenemos datos mixtos, podemos definir una distancia que combine distintas distancias en función del tipo de variable f:

$$d_{Mixta}(\mathbf{x}, \mathbf{z}) = \frac{\sum_{j=1}^{m} w_j^{(f)} d_j^{(f)}}{w_j^{(f)}}$$

  En esta fórmula, $d_j^{(f)}$ dependerá de la tipología de la variable f.

- Un ejemplo clásico es la **distancia de Gower**.

Parámetros de un modelo de ML

- Un **parámetro** es un valor que el algoritmo del modelo de ML ajusta durante el proceso de entrenamiento para hacer que el modelo se adapte mejor a los datos de entrenamiento y, en última instancia, haga predicciones más precisas en datos no vistos.
- Los parámetros son esenciales para definir la estructura y el comportamiento del modelo
- Tipos de parámetros:
  - **Parámetros** del modelo: componentes internos del modelo, se obtienen de los datos. Ej: los coeficientes asociados a las variables en una regresión lineal o logística
  - **Hiperparámetros** del modelo: Los establece el científico de datos antes del entrenamiento y controlan aspectos más generales del modelo. Ej: el valor k en el modelo de los k vecinos (me fijo en $k = 2$ vecinos para determinar la clase de un punto).
- Ajuste de parámetros e hiperparámetros: clave para el desarrollo de modelos existosos

Clustering no jerárquico

- Asignar cada observación a un grupo o cluster en función de su desemejanza
- Se especifica un número de clusters $K < n$ y se etiqueta cada cluster con un entero $k \in \{1, \dots, K\}$
- Cada observación $i \in \{1, \dots, n\}$ se asigna a un único cluster
- Esto se caracteriza con $k = S(i)$ que asigna la observación i-ésima al k-ésimo cluster
- Se busca que esta asignación $k = S(i)$ cumpla ciertas condiciones en base a las desemejanzas entre puntos
- En particular, se quiere minimizar una función de pérdida que indique lo bueno que es el cluster

¿Qué función minimizar?

- Dado que el objetivo es asignar puntos cercanos al mismo cluster, una función natural de pérdida sería

$$W(S) = \sum_{k=1}^{K} \sum_{S(i)=k} \sum_{S(i')=k, i' \neq i} d(x_i, x_{i'})$$

Es decir, minimizar la distancia entre las observaciones asignadas al mismo cluster

Clustering no jerárquico

- $W(S)$ es la varianza intra-cluster (within-cluster variance) dado que podemos descomponer la dispersión total $T(X)$ como

$$T(X) = \sum_{i=1}^{n} \sum_{i'=1, i' \neq i}^{n} d(x_i, x_{i'}) =$$

$$= \sum_{k=1}^{K} \sum_{S(i)=k} \left(\sum_{S(i')=k, i' \neq i} d(x_i, x_{i'}) + \sum_{S(i') \neq k, i' \neq i} d(x_i, x_{i'}) \right)$$

$$T(X) = W(S) + B(S)$$

- Dado un conjunto de datos, $T(X)$ es fijo, no varía en función de la estructura cluster final asignada

- $B(S)$ es la varianza entre los clusters (between-cluster variation):

$$B(S) = \sum_{k=1}^{K} \sum_{S(i)=k} \sum_{S(i')\neq k, i'\neq i} d(x_i, x_{i'})$$

  Cuanto más lejanos sean los puntos asignados a distintos clusters, mayor será $B(S)$

- Como $T(X) = W(S) + B(S) \Rightarrow W(S) = T(X) - B(S)$
- Por tanto, minimizar $W(S)$ es equivalente a maximizar $B(S)$

- En base a $T(X) = W(S) + B(S)$, se puede definir la contribución de la partición a la dispersión de los datos

$$\text{Contribución} = \frac{B(S)}{T(X)}$$

Clustering no jerárquico

- Esta situación de clustering podría resolverse directamente con optimización combinatoria
- ¿Problema? Demasiado costoso computacionalmente
- ¿Solución? Los algoritmos de clustering no jerárquico sólo pueden examinar una fracción de todas las posibles combinaciones. Su objetivo es identificar un subconjunto que pueda contener el óptimo o al menos una buena solución subóptima
- La estrategia es comenzar con una partición, ir actualizando la asignación de modo que se minice la función objetivo y parar cuando no haya mejora
- En función del algoritmo de clustering, la actualización de la asignación de los puntos a los clusters variará

Partición de los datos: k-medias

- El algoritmo de las k-medias es el algoritmo de aprendizaje automático no supervisado más utilizado para agrupar un conjunto de observaciones en un conjunto de grupos o clústeres
- K representa el número de grupos pre-especificados por el científico de datos
- Se utiliza cuando todas las variables son cuantitativas y la medida de distancia utilizada es la distancia euclídea al cuadrado

Partición de los datos: k-medias

- **Idea**: definir clústeres de tal modo que se minimice la variabilidad total dentro de los clústeres (*within-cluster variation*):

$$W(S) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} (\mathbf{x}_i - \mathbf{c}_k)^2$$

siendo K el número total de clusters, S_k el cluster k, x_i cada una de las observaciones del conjunto de datos y $\mathbf{c}_k$ el centroide del cluster S_k (la media de los elementos de dicho cluster). Nótese que los clústeres resultantes $S = \{S_1, \dots, S_K\}$ son disjuntos dos a dos, es decir, $S_h \cap S_q = \emptyset, \forall h, q = 1, \dots, K, h \neq q$

Partición de los datos: k-medias

¿Por qué el centroide es la media de los elementos del cluster? Es el punto que minimiza las distancias

$$W(S) = \sum_{k=1}^{K} W(S_k) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} \sum_{j=1}^{m} (x_{ij} - c_{kj})^2$$

Como $W(S_k)$ es una función cuadrática de $\mathbf{c}_k$, para hallar el mínimo basta con derivar con respecto *a* c_{kj}

$$\frac{\partial W(S_k, \mathbf{c_k})}{\partial c_{kj}} = \frac{\partial \sum_{\mathbf{x_i} \in S_k} \sum_{j=1}^{m} (x_{ij} - c_{kj})^2}{\partial c_{kj}} = \frac{\partial \sum_{\mathbf{x_i} \in S_k} (x_{ij} - c_{kj})^2}{\partial c_{kj}} =$$

$$= -2 \sum_{\mathbf{x_i} \in S_k} (x_{ij} - c_{kj}) \qquad k = 1, ..., K, \ j \text{ fijo}$$

e igualar a 0: $\frac{\partial W(S_k, \mathbf{c_k})}{\partial c_{kj}} = 0 \iff c_{kj} = \frac{1}{|S_k|} \sum_{\mathbf{x_i} \in S_k} x_{ij}$

siendo $|S_k|$ el cardinal de S_k.

Los centroides óptimos para minimizar la función $W(S)$ es la media de cada cluster

Algoritmo de las k-medias

1. **Inicialización**. Se elige el número K de clusters y se escogen al azar K centroides del conjunto de datos
2. **Actualización de los clústeres**. Dados los K centroides $\mathbf{c}_k$, $k = 1, \dots, K$, cada punto $\mathbf{x}_i$ se asigna al centroide del que menos dista, es decir, al que minimice $(\mathbf{x}_i - \mathbf{c}_k)^2$. Los elementos atribuidos a cada centroide $\mathbf{c}_k$ forman el cluster S_k.
3. **Actualización de los centroides**. En cada cluster S_k, $k = 1, \dots, K$ se calcula la media, que se denota por $\mathbf{c}_k^{'}$ y se convierte en el nuevo centroide.
4. **Test de los centroides**. Se comparan los nuevos centroides $\mathbf{c}_k^{'}$ con los de la iteración previa ($\mathbf{c}_k$). Si $\mathbf{c}_k^{'} = \mathbf{c}_k \ \ \forall k = 1, \dots, K$, el algoritmo para y se devuelven los clústeres S_k y sus centroides $\mathbf{c}_k^{'}$, $\forall k = 1, \dots, K$. En caso contrario, se hace $\mathbf{c}_k = \mathbf{c}_k^{'}$ y se vuelve al paso 2.

Los clústeres están representados por su centroide, entendiendo éste como un punto de referencia

- Las k-medias no funcionan con datos categóricos
- Las k-medias son sensibles a outliers. ¿Por qué?

Alternativa: **k-medoides**

- Los centroides, en lugar de ser la media de las observaciones de cada cluster, será una de las propias observaciones del cluster
- Más robusto a outliers
- Definido para cualquier distancia arbitraria d

1. **Inicialización**. Se elige el número K de clusters y se escogen al azar K observaciones del conjunto de datos (medoides $\mathbf{m}_k$). Se elige una distancia d.
2. Se asigna cada observación $\mathbf{x}_i$ al medoide más cercano: $argmin_k d(\mathbf{x}_i, \mathbf{m}_k)$. Así se forman los K clusters $\{S_1, \dots, S_K\}$.
3. Una vez formado los clusters, se recalculan los medoides $\mathbf{m}_k = argmin_{\mathbf{x}_i \in S_k} \sum_{\mathbf{x'}_i \in S_k} d(\mathbf{x}_i, \mathbf{x'}_i), \forall k = 1, \dots, K$
4. Se repiten los pasos 2 y 3 hasta que no haya cambios

Los clústeres están representados por su medoide, entendiendo éste como un punto de referencia

Ventajas

- Simple y rápido
- Escalable: aplicable con facilidad a grandes conjuntos de datos
- Computacionalmente mejor que los algoritmos jerárquicos. Eficiente en tiempo y memoria
- Tiene su propia función objetivo, la cual se pretende minimizar, que permite hacerse una idea de cómo de buena es la solución
- Sólo depende de un parámetro k

Desventajas

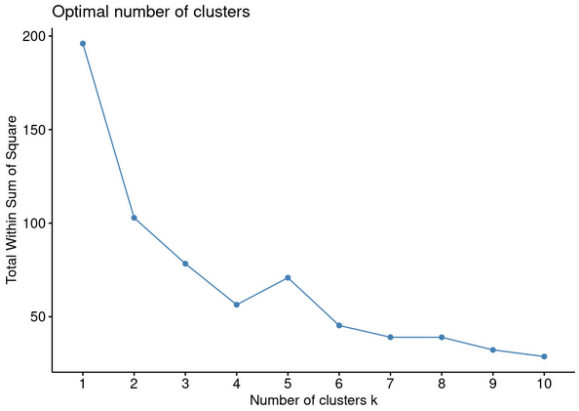
- Hay que seleccionar el valor de k y los centroides iniciales
- Puede converger a mínimos locales –> puede no ser la partición óptima
- Depende de la inicialización
  - Solución: replicar el algoritmo con distintas inicializaciones
- Tiende a crear grupos del mismo tamaño y con forma globular. Resultados pobres si los grupos son no convexos
- Usa la media –> se ve afectado por atípicos.
  - Solución: usar medoides en lugar de centroides. Los medoides son obligatoriamente puntos de la muestra

Métodos de elección de k

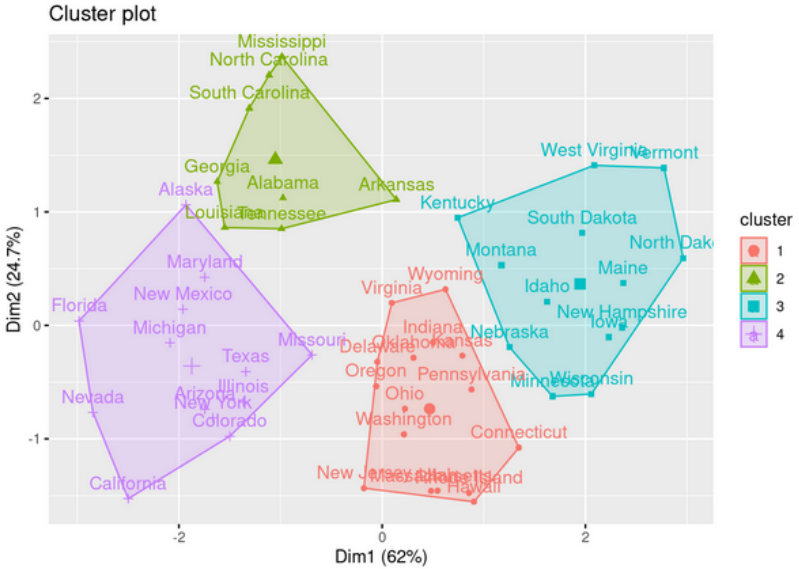
- Existen distintas aproximaciones para elegir el valor de k
  - Paquete NbClust de R contiene 30 criterios distintos
- Algunos de los más conocidos:
  - Método del codo
  - Método de la silueta (Silhouette)
  - Gap

Elección de k: método del codo

- **Objetivo de las k-medias**: construir grupos que minimicen la variación total dentro de los clústeres
- Calcular dicha variación para distintos valores de k y graficarla
- Un "codo" en el gráfico avisará del número apropiado de grupos, es decir, cuando el descenso de la variación de un k al siguiente no sea llamativa



Elección de k: método del codo



Elección de k: método de la silueta

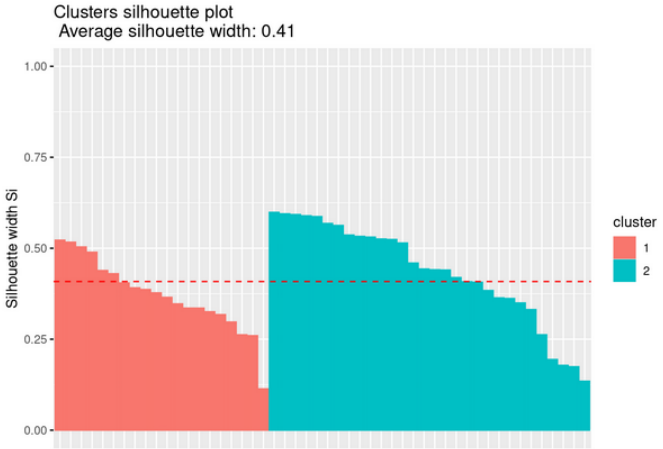
- Técnica no supervisada para valorar la coherencia o calidad de la partición
- La silueta determina hasta qué punto cada elemento se encuentra dentro de su agrupación.
- Para cada observación $\mathbf{x}_i$, la silueta se calcula como

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(a(\mathbf{x}_i), b(\mathbf{x}_i))}$$

  donde $a(\mathbf{x}_i)$ es la media de las distancias de la observación $\mathbf{x}_i$ a los puntos de su propio cluster y $b(\mathbf{x}_i)$ es la media de las distancias de $\mathbf{x}_i$ a los puntos de su cluster más cercano (excluyendo el suyo)

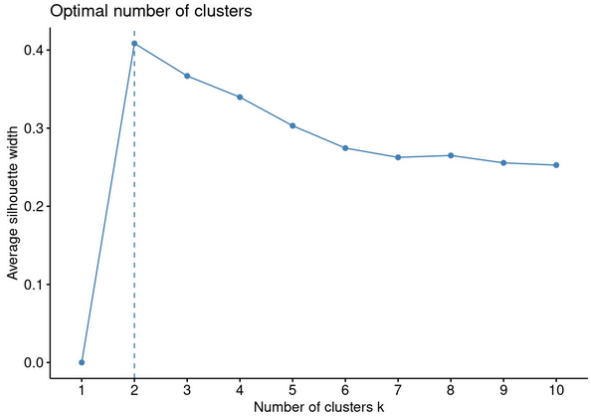
- Interpretación:
  - Valores altos: el punto encaja bien en su cluster
  - Valores bajos o negativos: el punto no encaja bien en su cluster

Elección de k: método de la silueta



Elección de k: método de la silueta

- La silueta media es la media de los valores silueta de todos los puntos.
- Se calcula para varios valores de k. El k adecuado es aquel que maximiza la silueta media.



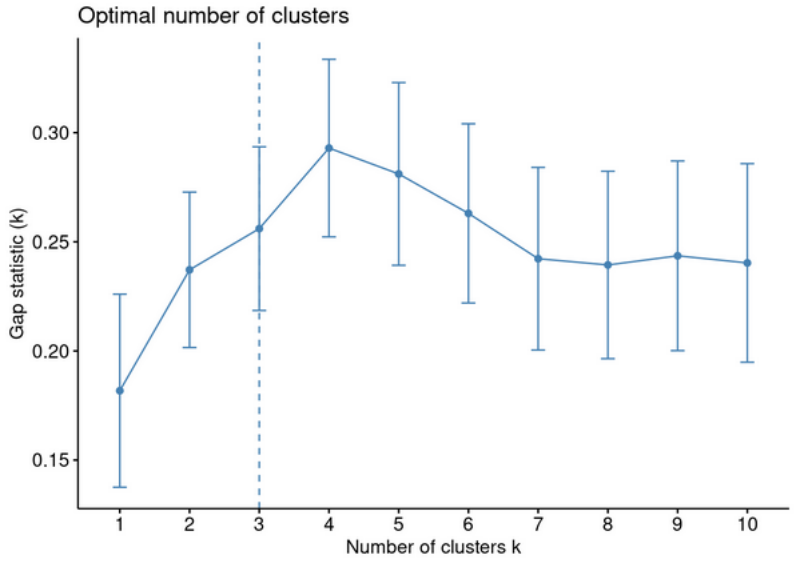
Elección de k: Gap

- Compara los resultados del clustering con datos aleatorios sin patrones de clustering (datos uniformes)
- Se realiza el clustering para distintos valores de k y se calcula la varianza total intracluster W_k
- Se simulan B conjuntos de datos aleatorios y se les aplica el mismo algoritmo de clustering con los distintos k y se calcula su varianza intracluster W_{kb}, $b = 1, \dots, B$
- Cálculo del estadístico Gap para cada k. Se comparan ambas varianzas intracluster. Cuanto mayor sea la diferencia, más sólida es la partición para dicho k.

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} log(W_{kb}) - log(W_k)$$

- Se escoge el k que maximiza dicha diferencia

Elección de k: Gap

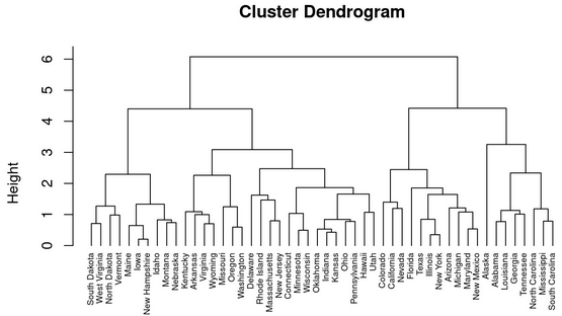


Cluster jerárquico

- Los modelos de clustering estudiados hasta ahora proporcionan una partición sin ningún tipo de información jerárquica
- Clustering jerárquico: obtiene agrupaciones jerárquicas de los datos, es decir, los clusters se alojan unos dentros de otros
- Este tipo de clustering no requiere conocer de antemano el número K de clusters

Cluster jerárquico

- Generan una clasificación iterativa de clústeres anidados mediante la unión o la separación de clústeres creados en etapas anteriores
- En el nivel más bajo, cada cluster contiene una única observación. En el nivel superior, sólo hay un cluster que contiene todos los datos



Cluster jerárquico: tipos

- Tipos:
  - **Aglomerativos (bottom-up)**. Cada observación comienza siendo un clúster, y en cada iteración se unen los dos clústeres más similares, hasta alcanzar una situación final en la que todos las observaciones pertenecen a un único clúster. Se conoce como **AGNES** ("AGglomerative NESting").
  - **Divisivos (top-down)**. Todas las observaciones comienzan en un único clúster y se va dividiendo hasta que cada observación forma un único clúster individual. Se conoce como **DIANA** ("DIvise ANAlysis").

Cluster jerárquico: tipos

- Ambos tipos de clustering jeráquico tienen como input una matriz de disimilitud entre objetos.
- Tanto la versión aglomerativa como la divisiva son heurísticas, es decir, no optimizan una función objetivo específica como en el caso de las k-medias.
- Siempre proporcionan una agrupación de los datos aunque no tengan ninguna estructura cluster (por ejemplo, si son datos uniformes).

Cluster jerárquico aglomerativo

El clustering aglomerativo comienza con n conglomerados (uno por cada dato) y, en cada paso, va fusionando los 2 grupos más similares hasta que hay un único grupo que contiene al total de datos.

Algoritmo

1. Input: matriz de disimilitud $D = (d_{ij}), i, j = 1, \dots, n$

   Inicializar los clusters: n clusters $S_i = \{i\}, i = 1, \dots, n$

   Inicializar el conjunto de clusters que faltan por unir: $Q = \{1, \dots, n\}$

2. Seleccionar los 2 clusters más similares S_j y S_k : $argmin_{j,k \in Q}\ d_{jk}$

   Con ellos, crear un nuevo cluster: $S_l \leftarrow \{S_j \cup S_k\}$

   Guardar dichos clusters como no disponibles: $Q \leftarrow Q \backslash \{j, k\}$

   Si $S_l \neq \{1, \dots, n\}$ entonces: $Q \leftarrow Q \cup \{l\}$

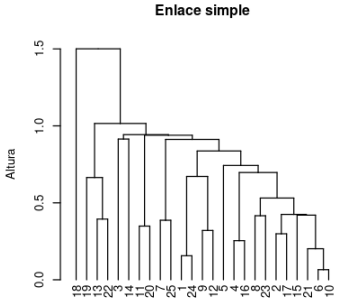
   Para cada $i \in Q$, actualizar la matriz de disimilitud $d(i, l)$

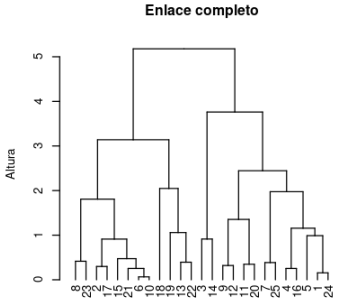
3. Repetir el paso 2 hasta que no queden cluster por unir

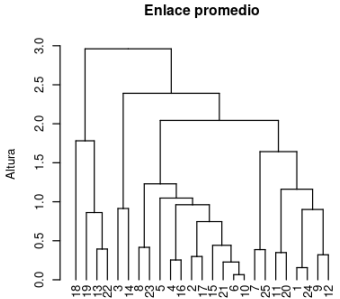
Cluster jerárquico aglomerativo: criterio de conexión

- Hay distintas variantes del clustering aglomerativo en función de cómo se define la disimilitud entre los grupos
- **Criterio de conexión** o "***linkage***" especifica cómo se determina el parecido (o la disimilitud) entre dos clústeres.
- Algunos de los criterios más comunes son:
  - Método de Ward
  - Agrupamiento de enlace completo
  - Agrupamiento de enlace promedio
  - Agrupamineto de enlace mínimo o simple
  - Agrupamiento de enlace de centroides
- El criterio de agrupación o conexión es un parámetro fundamental en el resultado final del clustering jerárquico

Cluster jerárquico aglomerativo: criterio de conexión







- **Agrupamiento de enlace mínimo o simple.** Minimiza las disimilitudes entre las observaciones más cercanas de dos clústeres. Es decir, calcula todas las disimilitudes por pares entre los elementos del conglomerado A y los elementos del conglomerado B. La disimilitud entre ambos conglomerados será la disimilitud de sus dos puntos más cercanos. Finalmente se unirán aquellos conglomerados con menor disimilitud

$$d_{ES}(A, B) = \min_{i \in A, i' \in B} d_{i,i'}$$

Al unir por la distancia mínima, tiende a producir clusters de mayor diámetro y más dispersos.

- **Agrupamiento de enlace completo.** Similar al anterior pero con la disimilitud máxima. Minimiza la disimilitud máxima entre las observaciones de dos clústeres:

$$d_{EC}(A, B) = \max_{i \in A, i' \in B} d_{i,i'}$$

Como uno los grupos buscando minimizar la distancia máxima, tiende a producir clusters más compactos, con menor diámetro.

- **Agrupamiento de enlace promedio.** Minimiza el promedio de las disimilitudes entre las observaciones de dos clústeres. Calcula todas las disimilitudes por pares entre los elementos del conglomerado A y los elementos del conglomerado B, y considera la media de estas disimilitudes como la distancia entre los dos conglomerados:

$$d_{EP}(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i' \in B} d_{i,i'}$$

siendo n_A y n_B el número de elementos en los grupos A y B, respectivamente.

Este tipo de enlace es una situación intermedia entre los dos previos, tiende a producir cluster relativamente compactos.

Como involucra hacer promedios de las distancias, cualquier cambio de escala alterará el agrupamiento final. En el caso del enlace simple y el enlace completo, las transformaciones que no alteren el orden (transformaciones monótonas) no cambiarán el resultado.

Cluster jerárquico aglomerativo: criterio de conexión

- **Método de Ward.** Minimiza la suma de las diferencias cuadradas dentro de los clústeres, es decir, la varianza intracluster. En cada paso, agrupa los clústeres que provocan el menor incremento de la varianza intracluster $W(S)$.
- **Agrupamiento de enlace de centroides.** La disimilitud entre los conglomerados A y B es la disimilitud entre sus centroides.

Cluster jerárquico divisivo: DIANA

- Comienza con todos los datos en un único cluster y, recursivamente, divide cada cluster en 2 cluster hijo
- En cada paso el grupo más *grande* se divide hasta que cada objeto es un único conglomerado (u otro criterio de parada)
- El clúster más *grande* será aquel de mayor diámetro, es decir, el que tiene mayor desemejanza entre dos de sus elementos o aquel con mayor desemejanza media. Es decir, dados los cluster $S = \{S_1, \dots, S_r\}$, el cluster más grande S_j es aquel que

$$S_j = argmax_{S_j \in S}\ d_{i,i'}, \forall i, i' \in S_j$$

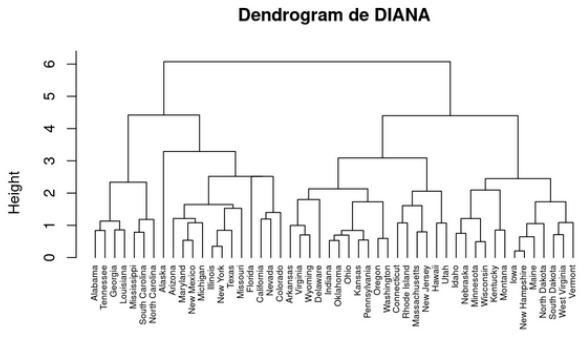
- La observación más lejana $i^* \in S_j$, $i^* = argmax_{i \in S_j}\ d_{i,i'}$, es la que inicia el nuevo cluster $S_{r+1} = \{i^*\}$
- Se asignan a este nuevo cluster S_{r+1} los puntos que sean más cercanos a él que al cluster del que provienen S_j.

Cluster jerárquico divisivo

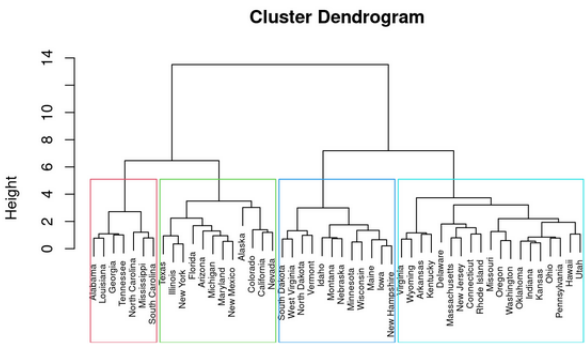
- Hay otras alternativas para realizar clustering jerárquico divisivo. Por ejemplo, bisecting k-means que divide el cluster de mayor diámetro en 2 cluster hijos haciendo uso de las k-medias o los k-medoides

Cluster jerárquico: dendrograma

Divisivo



Aglomerativo (Ward)

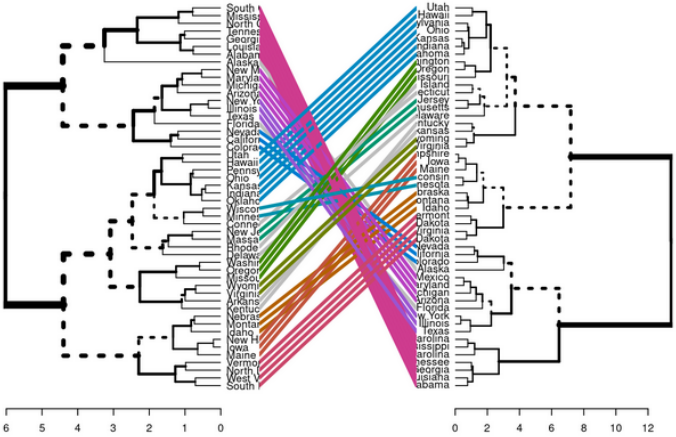


Cluster jerárquico: dendrograma

- Cada hoja corresponde a una observación
- La altura de la fusión indica la disimilitud entre las observaciones. Cuanto mayor es la altura de la fusión, menos parecidas son las observaciones
- **¡Ojo!** Cuando empleamos un dendrograma, las conclusiones sobre la proximidad de dos observaciones sólo pueden extraerse a partir de la altura a la que se fusionan las ramas que contienen primero esas dos observaciones. No podemos utilizar la proximidad de dos observaciones a lo largo del eje horizontal como criterio de su similitud.
- La **altura del corte del dendrograma controla el número de clusters** (similar al k en las k-medias)

Cluster jerárquico: comparación de dendrogramas

Comparación de enlace Ward y completo:



- Líneas discontinuas: elementos no presentes en el otro dendrograma
- Función de entrelazamiento: calidad de alineación entre los dos dendrogramas. 1 (entrelazamiento total) y 0 (sin entrelazamiento)

Cluster jerárquico: elección del número de grupos

- Al igual que en el cluster no jerárquico, se pueden aplicar métodos para elegir el número de conglomerados
- Los métodos vistos anteriormente (método del codo, silueta, estadístico Gap) son perfectamente aplicables

Cluster jerárquico: ventajas

- **Jerarquía de clusters**. Permite análisis a distintos niveles
- **Interpretación visual** de cómo se agrupan los datos y se relacionan
- **No requiere especificar previamente el número de grupos**
- **Identificación de subgrupos**. Permite detectar clases dentro de clases
- **Detección de outliers**
- **No sensible a la inicialización**
- **Análisis exploratorio de datos**. Visión general de cómo se agrupan naturalmente los datos sin necesidad de conocimiento previo.

Cluster jerárquico: desventajas

- **Requiere definir un criterio de corte** para convertir la jerarquía en clusters
- **No hay una única respuesta correcta.** Deben considerarse los diversos resultados con información interesante
- **No es óptimo para todo tipo de datos**. Funcionan mejor cuando los datos tienen una estructura jerárquica natural
- **No es adecuado para datos de alta dimensión**
- **Resultados no siempre reproducibles**
- **Sin capacidad de predicción**. No son útiles para predecir a qué clúster pertenece una nueva observación

Mapas auto-organizados de Kohonen

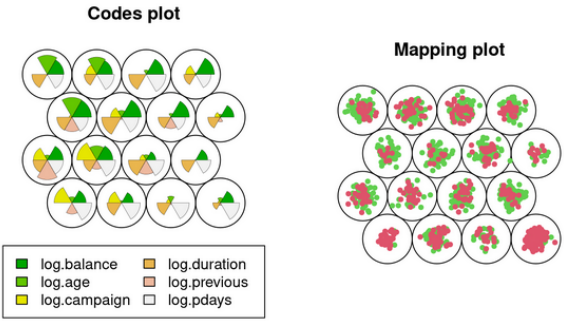
- SOM: Self-Organizing Maps
- Herramienta de **reducción de la dimensión**
- **Algoritmo de clustering**: organizar datos en grupos tal que los elementos dentro de un mismo grupo sean similares entre sí en función de ciertas características.
- **Red neuronal** bidimensional:
  - Capa de entrada con tantas neuronas como variables
  - Capa para representar en 2 dimensiones el total de observaciones

Mapas auto-organizados de Kohonen: algoritmo

1. **Inicialización**: Creación de la red de neuronas bidimensional, conocida como "mapa auto-organizativo (SOM)". Cada neurona representa una ubicación en el espacio SOM.
2. **Asignación de Pesos**: Cada neurona en el SOM tiene asociado un vector de pesos que es del mismo tamaño que los datos originales
3. **Entrenamiento**: Se presentan los datos al SOM, y cada dato se asigna a la neurona cuyos pesos son más similares a los atributos del dato. Las neuronas ganadoras (aquellas a las que se asigna un dato) y sus vecinas en el mapa SOM se ajustan para que se parezcan más al dato presentado. Este proceso de aprendizaje se repite varias veces.
4. **Agrupación en Clusters**: Después del entrenamiento, las neuronas en el mapa SOM que están cerca una de la otra representan clusters de datos. Los datos que se asignaron a estas neuronas durante el entrenamiento se consideran miembros de un mismo cluster.

Mapas auto-organizados de Kohonen

Mapa con estructura 4x4



- **Codes plot**: En cada neurona se muestra el peso de cada variable. Se aprecia la estrucutra de similitud: las variables dominan por zonas
- **Mapping plot:** Densidad de puntos por neurona. En este caso se colorean los puntos en función del target (estudio supervisado)

Mapas auto-organizados de Kohonen: ventajas

- **Topología Preservada**: Los clusters en el SOM reflejan la estructura de vecindad en los datos originales -> facilita la interpretación de los resultados
- **Escalabilidad**: Pueden manejar grandes conjuntos de datos y dimensiones elevadas
- **Visualización**
- **Exploración interactiva:** Los usuarios pueden navegar por el mapa para inspeccionar las regiones y sus contenidos
- **Reducción de ruido:** Pueden ayudar a reducir el ruido y la redundancia en los datos, lo que mejora la calidad del análisis

Mapas auto-organizados de Kohonen: desventajas

- **Sensibilidad a la inicialización** de los pesos de las neuronas
- **Determinación del tamaño del mapa**: Si el mapa es demasiado pequeño, puede no capturar la estructura de los datos correctamente, mientras que si es demasiado grande, puede sobreajustarse a los datos y perder la capacidad de generalización
- **Interpretación de los resultados**: Se dificulta en mapas de gran tamaño
- **Requiere ajuste de hiperparámetros** y el rendimiento depende de ellos
- **Puede converger a mínimos locales**
- **Requiere grandes conjuntos de datos**: Pueden no funcionar bien en conjuntos de datos pequeños o altamente desequilibrados, ya que su eficacia se basa en la capacidad de aprender patrones significativos a partir de una cantidad suficiente de datos

Referencias

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15-24.

Chiang, M. M. T., & Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, *27*, 3-40.

Gan, G., Ma, C., & Wu, J. (2020). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.

Mirkin, B. (2005). *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165-193.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.