Fundamentos de Aprendizaje Automático

Máster Universitario en Ciencia de Datos aplicada a las Ciencias Sociales
por la Universidad de Granada y la Universidad de Salamanca

Métodos no supervisados

01. Clustering

UNIVERSIDAD
DE GRANADA

Índice

Sesión 1

1. ¿Qué es el *clustering*?
2. Elementos básicos de un algoritmo de clustering
3. Métodos basados en centroides: k-means y k-medians
4. Clustering jerárquico: Aglomerativo y divisivo
5. Clustering basado en densidad: DBSCAN y OPTICS
6. Cómo elegir el número de clústeres apropiado
7. Otros algoritmos
8. Medidas de calidad del clustering

Sesión 2

1. Uso de algoritmos de clustering en Python
2. Ejercicio práctico

1. ¿Qué es el *clustering*?

Definición

Búsqueda de agrupaciones en datos

Proceso de agrupar un conjunto de objetos descritos mediante propiedades en grupos (o clústeres), de forma que un clúster contiene objetos similares entre sí y diferentes a los de otros clústeres

Es una técnica de aprendizaje no supervisado

No se dispone de clases predefinidas ni de ejemplo etiquetados; es decir, no se conocen las agrupaciones para ningún subconjunto de individuos

Suele realizarse en las primeras etapas del proceso de análisis de datos

Los clústeres sirven para resumir los datos, de forma que se pueden utilizar las agrupaciones como representación colectiva de los individuos

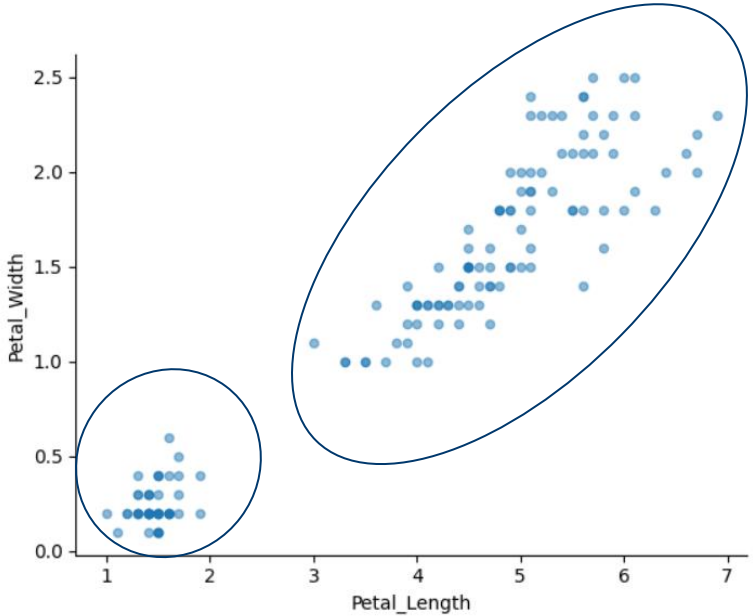
sklearn.cluster.

1. ¿Qué es el *clustering*?

Ejemplo

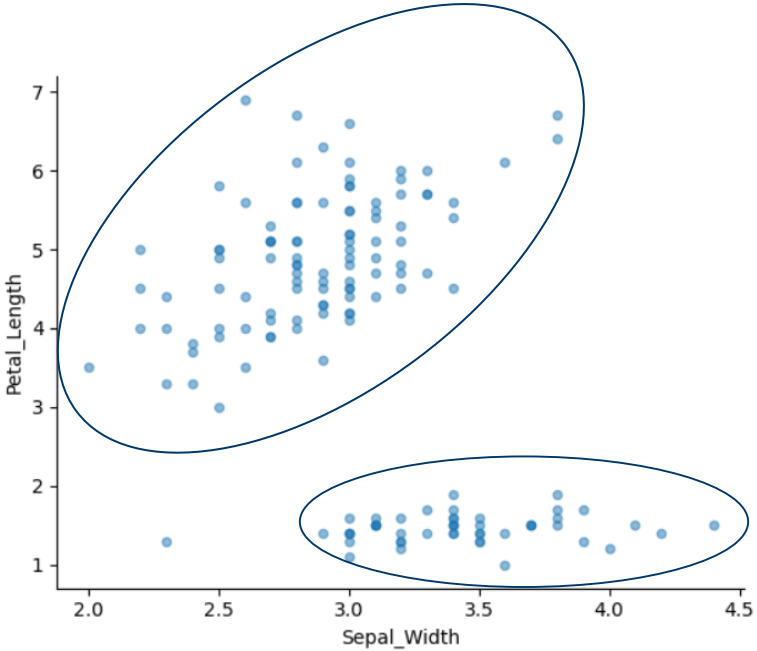
```
import pandas as pd
csv_url = 'https://archive.ics.uci.edu/ml/machine-
learning-databases/iris/iris.data'
col_names =
['Sepal_Length','Sepal_Width','Petal_Length','Petal_Width'
,'Class']
iris =  pd.read_csv(csv_url, names = col_names)
iris
```

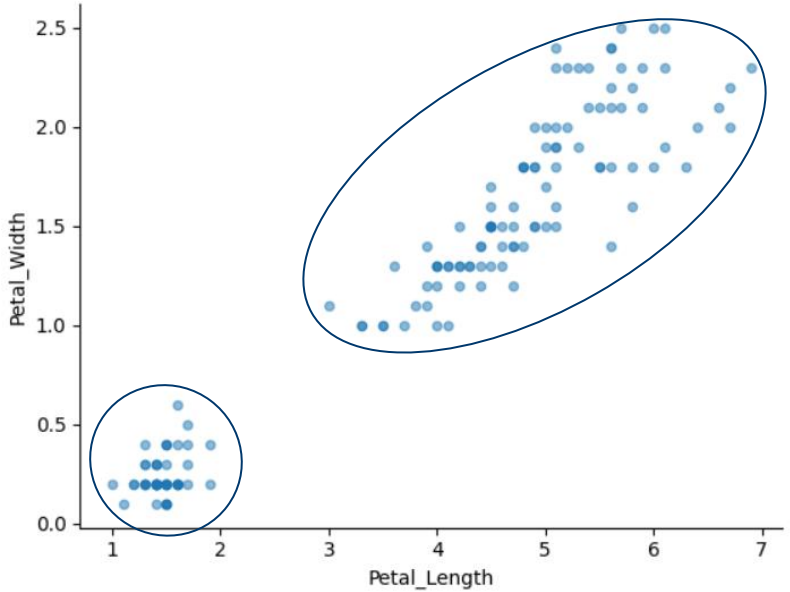
	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...



1. ¿Qué es el *clustering*?

Ejemplo





1. ¿Qué es el *clustering*?

Requisitos de un algoritmo de clustering

- Escalabilidad para procesar grandes conjuntos de datos
- Capacidad para trabajar con **diferentes tipos de atributos** (numéricos, categóricos)
- Obtención de *clústeres* con forma arbitraria, no solo esféricas
- **Mínimo conocimiento** del problema para definir los (hiper-)parámetros del procedimiento
- Capacidad para trabajar con datos imperfectos (perdidos, desconocidos, con ruido, erróneos, etc.)
- Independencia del orden en que se presentan los individuos
- Capacidad para trabajar con datos con **muchas dimensiones**
- Consideración de restricciones en los *clústeres*
- Interpretabilidad y usabilidad

1. ¿Qué es el *clustering*?

Aplicaciones

Identificación de individuos u objetos similares de forma no supervisada

- Segmentación de clientes
- Sistemas de recomendación
- Filtros de spam
- Análisis de redes sociales
- Caracterización de series temporales
- Etc.

2. Elementos básicos de un algoritmo de *clustering*

Medidas de distancia / similitud

Elemento fundamental, ya que cuantifica si dos individuos son similares o no

Aplicación en la **generación** de los *clústeres* o en la **evaluación** de la calidad de los *clústeres*

Uso directo o con alguna transformación; por ejemplo, elevada al cuadrado

Propiedades de una distancia:

- Identificadora $d(x,y)=0 \Leftrightarrow x=y$
- Positiva $d(x, y) \geq 0$
- Simétrica $d(x, y) = d(y, x)$
- Triangular $d(x, y) + d(y, z) \geq d(x, z)$

2. Elementos básicos de un algoritmo de *clustering*

Medidas de distancia

Distancia Euclídea

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Distancia de Minkowski

$$d(x,y) = (\sum_{k=1}^{n}|x_k - y_k|^r)^{\frac{1}{r}}$$

Distancia de Mahalanobis

$$d(x,y) = \sqrt{(x-y)\sigma^{-1}(x-y)^T}$$

Distancia coseno

$$cos(x,y) = \frac{(x \bullet y)}{||x||||y||}$$

Correlación de Pearson

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

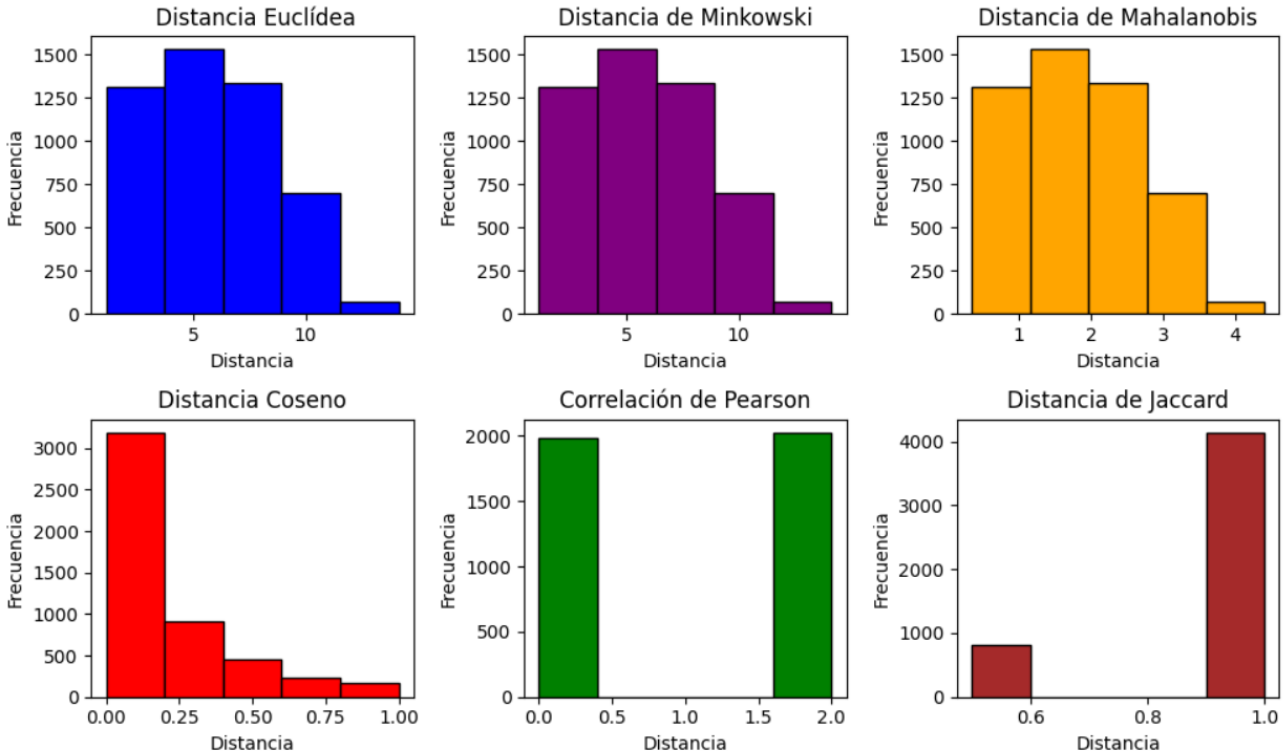
Distancia de Jaccard

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

2. Elementos básicos de un algoritmo de *clustering*

Medidas de distancia

Distancias sobre una malla de 100 puntos equidistribuidos 0x10



2. Elementos básicos de un algoritmo de *clustering*

Selección de *k*

Normalmente, el número de clústeres *k* se fija a priori

Debilidad: requiere cierto conocimiento del problema que se aborda (veremos algunas indicaciones para seleccionar y ajustar *k*)

Estudiaremos métricas para calcular cómo de bueno es un *clustering* sin necesidad de conocer las clases reales (<- manteniendo el problema como no supervisado)

2. Elementos básicos de un algoritmo de *clústering*

Tipos de algoritmos

Particionamiento

k-means

Se construyen *k* particiones de los datos, normalmente satisfaciendo: cada clúster tiene más de un objeto, cada objeto solo pertenece a un clúster

Jerárquicos

Se realiza una descomposición del conjunto de objetos de forma aglomerativa (objeto -> grupos) o divisiva (grupo -> objetos)

Basados en densidad

Aumentar un clúster de forma que, para todos los objetos dentro de un clúster, su vecindario a una distancia *d* debe contener un número mínimo de puntos

Otros: rejilla, basados en modelos

3. Métodos basados en centroides

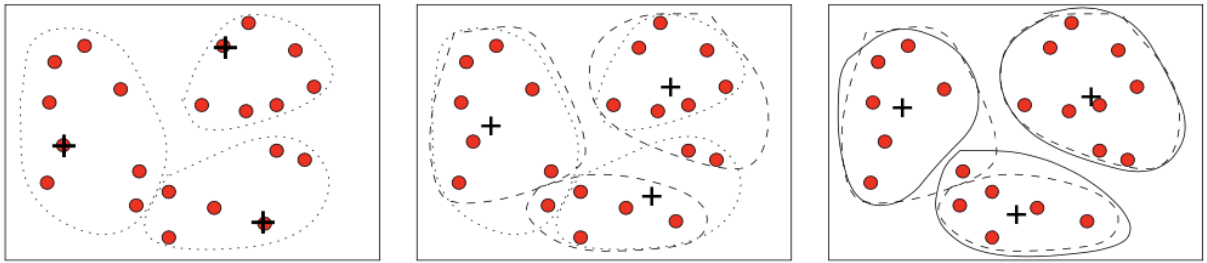
K-means

Aproximación básica (algoritmo de Lloyd)

Entrada: *k* (número de clústeres), *n* objetos

Procedimiento:

1. elegir aleatoriamente los centros de los clústeres (no necesariamente uno del conjunto)
2. repetir mientras haya cambios:
    2.1 (re)asignar cada objeto al clúster con centro más cercano
    2.2. recalcular los centros como el punto medio de cada clúster

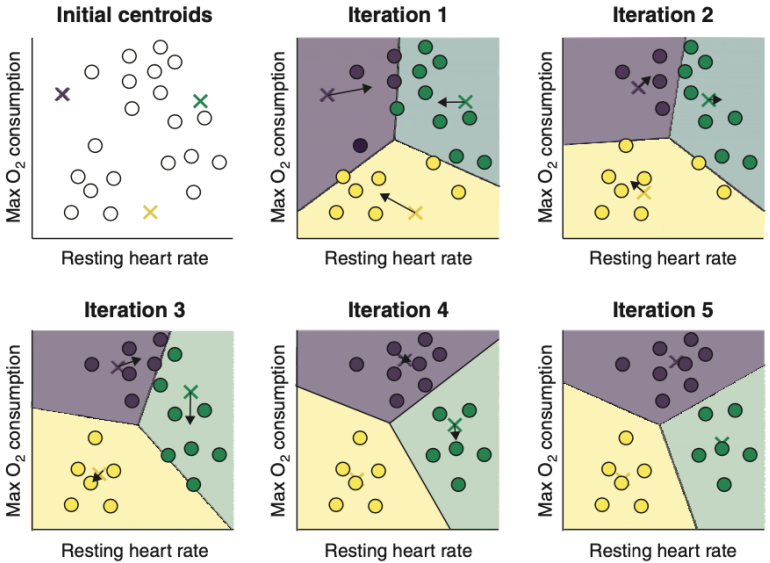


J. Han, M. Kamber, J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

3. Métodos basados en centroides

K-means

Aproximación básica (algoritmo de Lloyd)



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

3. Métodos basados en centroides

K-means

Aproximación básica (algoritmo de Lloyd)

https://www.youtube.com/watch?v=5I3Ei69I40s

¿Qué podemos observar?

3. Métodos basados en centroides

K-means

Aproximación básica (algoritmo de Lloyd)

https://www.youtube.com/watch?v=5I3Ei69I40s

¿Qué podemos observar?

- Los resultados son distintos dependiendo de la inicialización de los centros

3. Métodos basados en centroides

K-means

K-means es uno de los métodos de clustering más utilizados.

- Destaca por la sencillez y velocidad de su algoritmo.

Limitaciones:

- Requiere que se indique de antemano el número de clústers que se van a crear -> Estrategias para ayudar a identificar potenciales valores óptimos de K (elbow, shilouette)
- Dificultad para detectar clústers alargados o con formas irregulares.
- Las agrupaciones resultantes pueden variar dependiendo de la asignación inicial de los centroides.
  - Es recomendable repetir el proceso de clustering entre 25-50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna.
- Presenta problemas de robustez frente a outliers.

3. Métodos basados en centroides

K-medians

Diferencias con k-means

- Utiliza la **mediana** de cada dimensión de las observaciones dentro del clúster para definir el centro del clúster

- Más robusto al ruido y a los *outliers*.

- Medida más utilizada: la distancia de Manhattan (vs Euclídea usada normalmente en K-means), aunque puede usarse cualquier medida

Aproximación básica (Ejemplo: PAM Partitioning Around Medians)

Entrada: *k* (número de clústeres), *n* objetos

Procedimiento:

1. elegir aleatoriamente los centros de los clústeres (se utiliza uno del propio conjunto de datos)
2. repetir mientras haya cambios:
    2.1 (re)asignar cada objeto al clúster con centro más cercano
    2.2. recalcular los centros como la **mediana** de cada clúster en cada dimensión

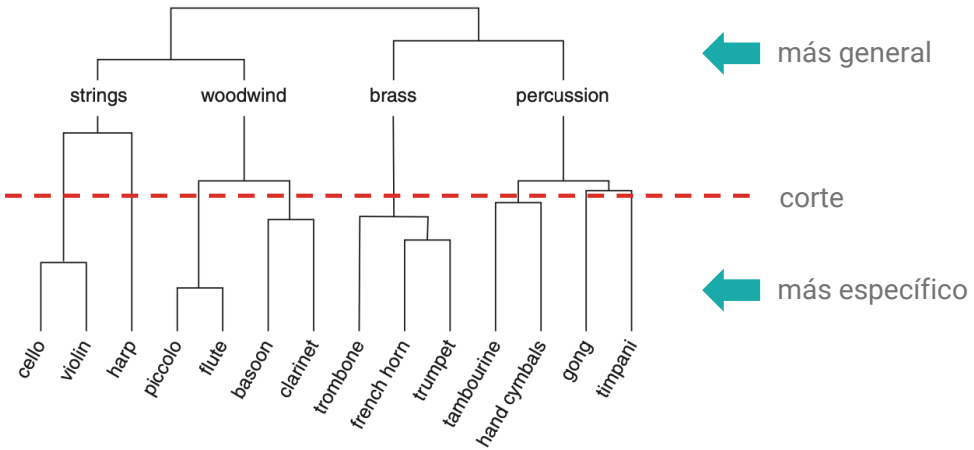
4. Clustering jerárquico

Concepto

En muchos problemas no es apropiado agrupar los objetos en divisiones “planas”: ítems de una tienda virtual, clientes de un supermercado, instrumentos de una banda, etc. << clases y subclases

Representación mediante **dendrograma**

- Woodwinds
  - Reeded
    - Clarinet
    - Bassoon
  - Non-reeded
    - Flute
    - Piccolo
- Strings
  - Plucked
    - Harp
  - Bowed
    - Violin
    - Cello

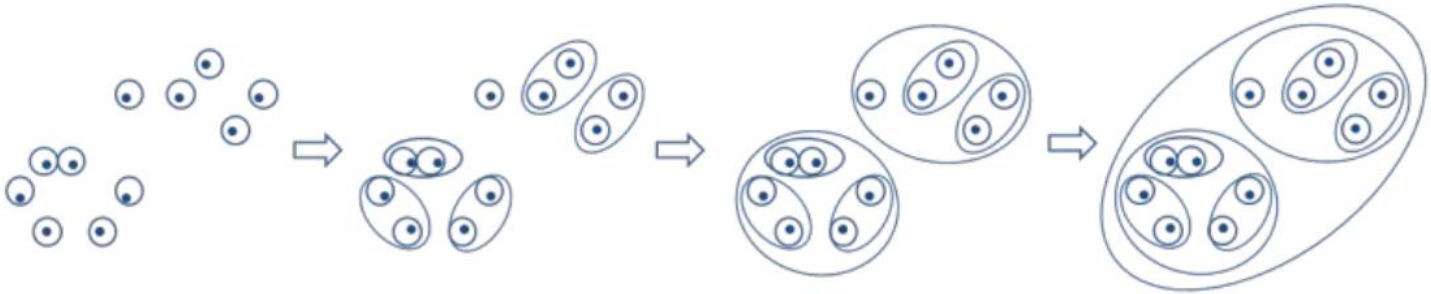


H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

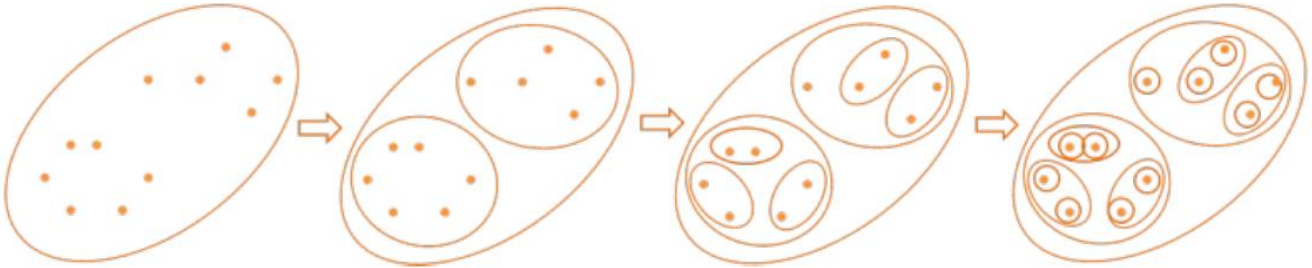
4. Clustering jerárquico

Tipos

Clustering jerárquico aglomerativo



Clustering jerárquico divisivo



4. Clustering jerárquico

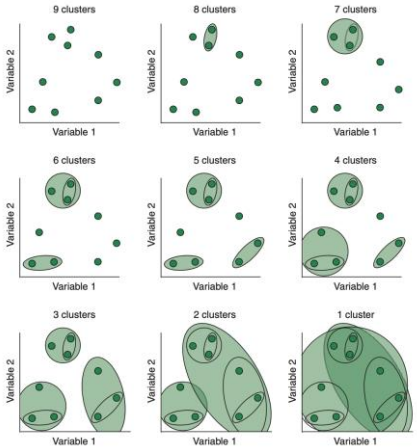
Clustering por agregación

Agglomerative clustering

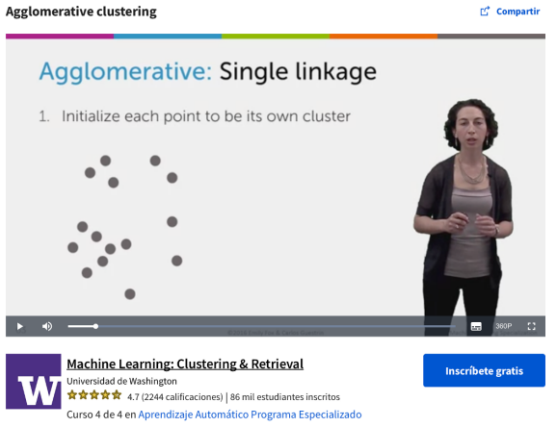
Comienza con clústeres muy pequeños (los propios ítems que serán las hojas) y va fusionando hasta llegar a una división adecuada

Cálculo: iterativamente, combina los dos clústeres más cercanos en uno solo

El criterio de parada puede establecerse de diferentes formas (1 solo clúster)



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr.* Manning, 2020.



https://es.coursera.org/lecture/ml-clustering-and-retrieval/agglomerative-clustering-bsFBT

4. Clustering jerárquico

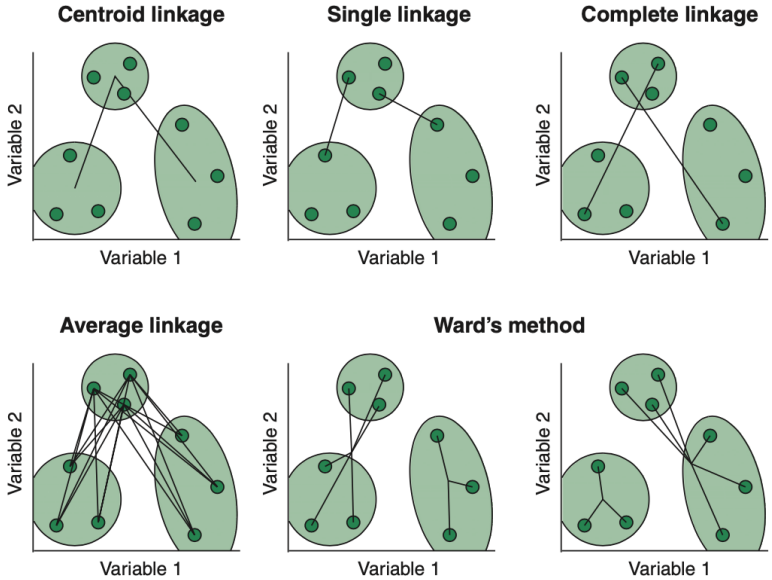
Clustering por agregación

Agglomerative clustering

Similitud entre 2 clústeres: métodos de proximidad (*linkage*) basados en una medida de distancia

- **centroid:** distancia entre los dos centroides (disponible en R)
- **single:** entre los dos objetos más cercanos
- **complete:** entre los dos objetos más lejanos de cada clúster
- **average:** media entre todos los objetos del clúster
- **Ward:** distancias entre cada par de clústeres usando la varianza intra-clúster

Método de Lance-Williams: Fórmula recursiva que permite calcular eficientemente las nuevas distancias sin tener que recalcular desde cero



4. Clustering jerárquico

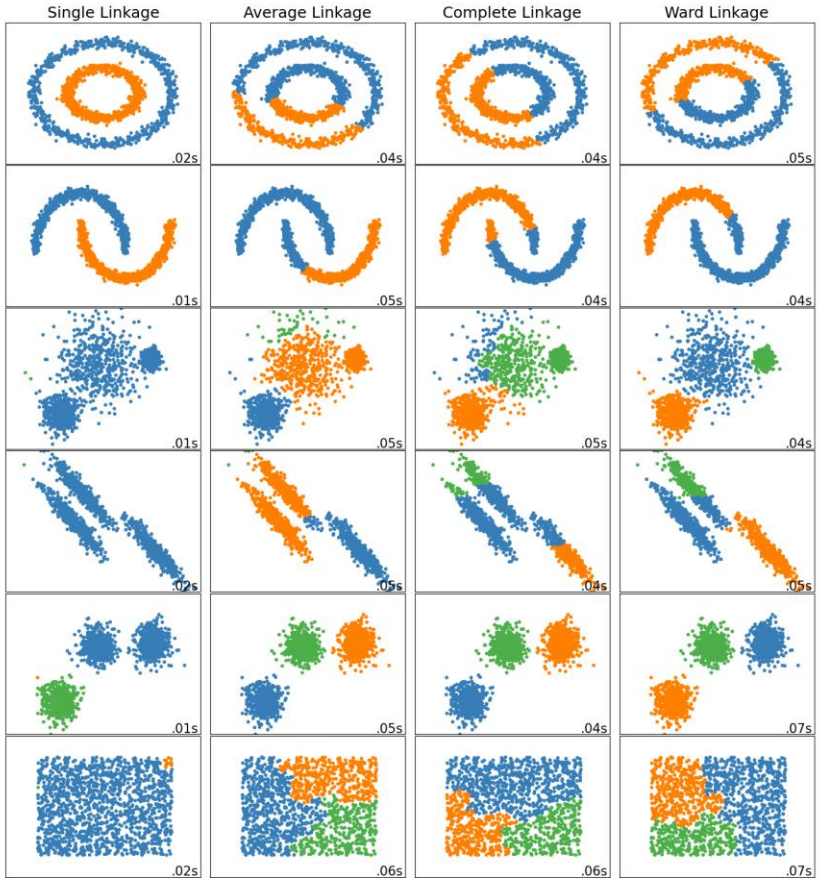
Clustering por agregación

Agglomerative clustering

Similitud entre 2 clústeres: métodos de proximidad (*linkage*) basados en una medida de distancia

- **centroid:** distancia entre los dos centroides (disponible en R)
- **single:** entre los dos objetos más cercanos
- **complete:** entre los dos objetos más lejanos de cada clúster
- **average:** media entre todos los objetos del clúster
- **Ward:** distancias entre cada par de clústeres usando la varianza intra-clúster

Método de Lance-Williams: Fórmula recursiva que permite calcular eficientemente las nuevas distancias sin tener que recalcular desde cero



4. Clustering jerárquico

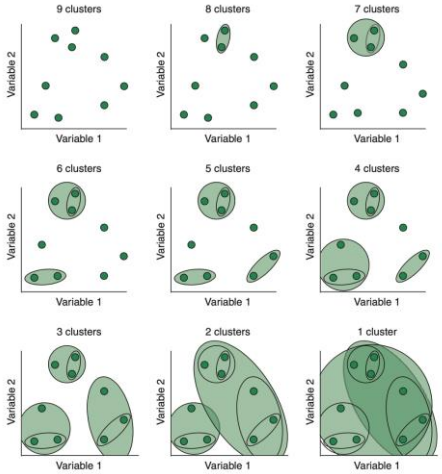
Clustering por división

Divisive clustering *(DIANA, DIvisive ANAlysis Clustering)*

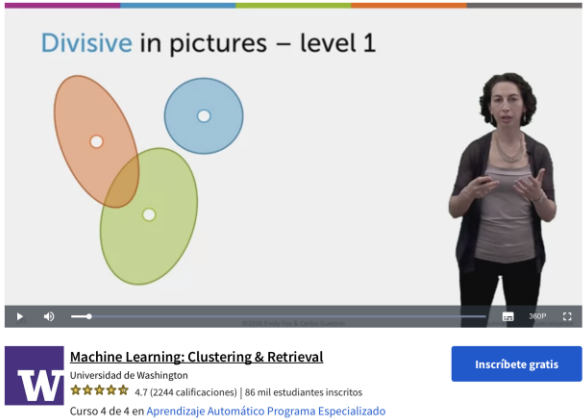
Comienza con clústeres muy grandes y va particionando hasta llegar a una división adecuada

Cálculo: iterativamente, divide un clúster en dos aplicando un criterio heurístico de distancia global

El criterio de parada puede establecerse de diferentes formas (hasta alcanzar *n* clústeres)



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr.* Manning, 2020.



https://es.coursera.org/lecture/ml-clustering-and-retrieval/divisive-clustering-gp6sl

4. Clustering jerárquico

Ventajas e inconvenientes

¿Cúando usarlo?

- Comprender resultados manera visual.
- Tengo un dataset pequeño.
- Desconozco la cantidad de clústeres por completo.
- Resultados rápidos.

Ventajas:

- No necesito el número de clústeres -> Ayuda visual del dendrograma.
- Resultados interpretables.
- Simple -> Única ejecución.

Inconvenientes:

- Tarda en datasets largos.
- Le afectan outliers drásticamente.
- Mayor necesidad de cómputo.

5. Clustering basado en densidad

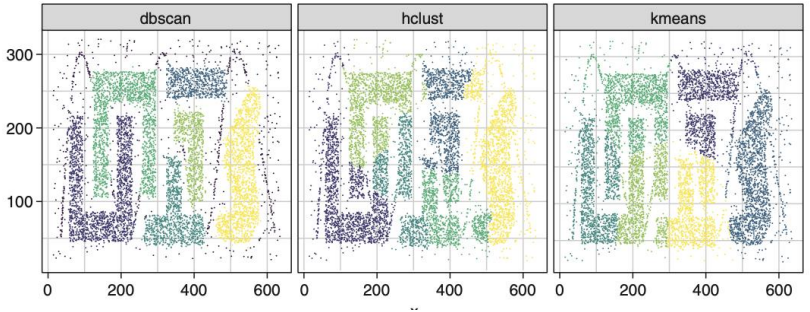
Concepto

Utiliza la ***densidad*** de objetos para asignarlos a un clúster, medida como "número de casos por unidad de volumen" en el espacio de características

Los clústeres se corresponden a regiones en el espacio con alta densidad, separados entre sí por áreas de poca densidad

Ventajas:

- No están sesgados para encontrar regiones esféricas
- No están sesgados para encontrar clústeres de diámetro similar
- Distingue los valores que no están en zonas de alta densidad (ruido, *outliers*)



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

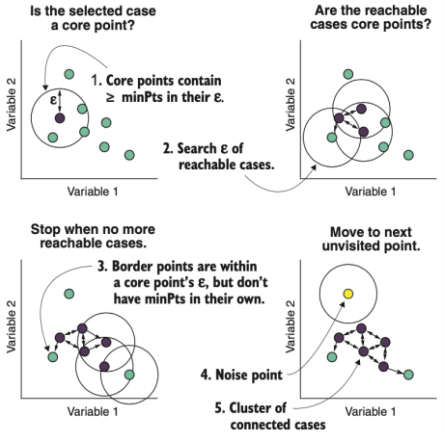
5. Clustering basado en densidad

DBSCAN

Busca regiones densas conectadas en el espacio de características

Cálculo: selecciona un objeto y determina si se trata de un punto central (tiene $\geq minPts$ objetos a distancia $\leq \epsilon$) o de un borde (no los tiene); si es central, repite para todos los puntos cercanos

El algoritmo finaliza cuando ha procesado todos los puntos. Los clústeres se forman con los vecindarios de los puntos centrales y los alcanzables por densidad a través de puntos centrales

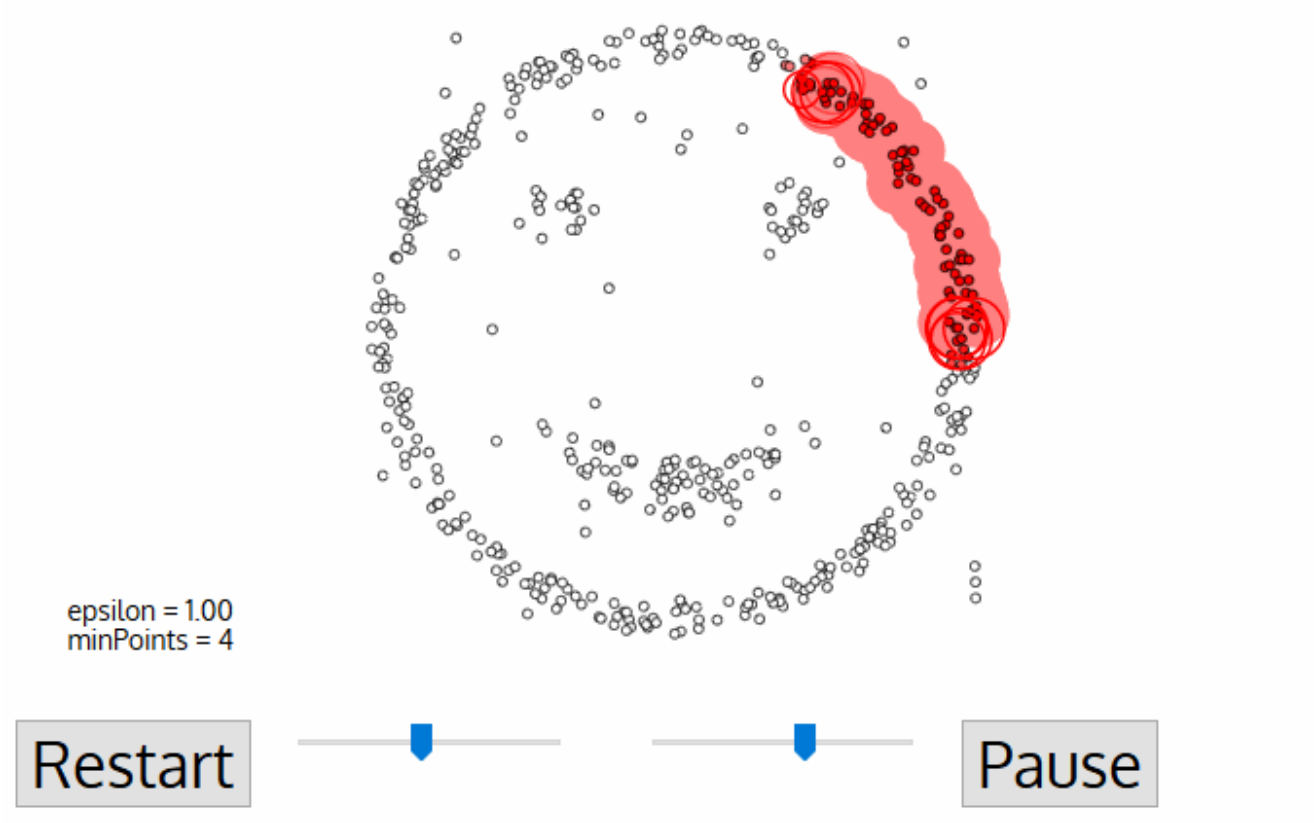


H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr.* Manning, 2020.

https://www.coursera.org/lecture/cluster-analysis/5-3-optics-ordering-points-to-identify-clustering-structure-JiYel

5. Clustering basado en densidad

DBSCAN



https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html.

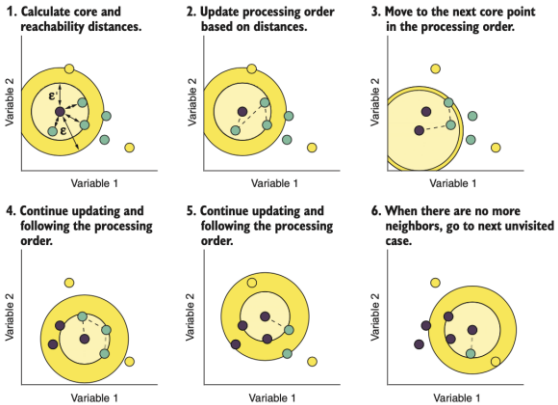
5. Clustering basado en densidad

OPTICS (Ordering Points To Identify the Clustering Structure)

Utiliza un valor de radio ϵ variable, que será mayor en las regiones menos densas (y viceversa), y la distancia al punto central más cercano ϵ'

Cálculo: calcula el valor ϵ' para todos los puntos, si $\epsilon' \leq \epsilon$ se trata de un punto central; se actualiza iterativamente una lista ordenada de puntos más cercanos y se calcula el valor de alcanzabilidad de cada uno; se pasa al próximo punto más cercano

El algoritmo finaliza cuando ha procesado todos los puntos. Como resultado, obtiene la lista de objetos ordenada por orden de visita y el valor de alcanzabilidad. Los clústeres se forman a partir de este orden



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

https://es.coursera.org/lecture/mineria-de-datos-introduccion/algoritmo-dbscan-raWvm

5. Clustering basado en densidad

Ventajas e inconvenientes

¿Cúando usarlo?

- Desconozco la cantidad de clústeres.
- No uso formas esféricas.
- Densidades similares entre clústeres

Ventajas:

- No requiere especificar el número de clústeres.
- Es capaz de detectar outliers o ruido.
- Puede encontrar clústeres en formas y tamaños arbitrarios.

Inconvenientes:

- Los hiper-parámetros son muy determinantes, algunas combinaciones no funcionan igual para todos los grupos con distintas densidades.
- Los puntos fronterizos a los que se puede acceder desde más de un clúster pueden formar parte de cualquier clúster.

6. Cómo elegir el número de clústeres apropiado

- **Métodos basados en centroides:**
  - Método del codo (**elbow):** se ejecuta K-means para un rango de valores de k y se calcula la suma de los errores cuadráticos (SSE) para cada k. El punto donde la gráfica de SSE vs. k forma un codo indica el número óptimo de clústeres
  - **Coeficiente de Silueta (Silhouette coefficient):** Se calcula para cada k y se selecciona el valor de k que maximiza el promedio de los coeficientes calculados.

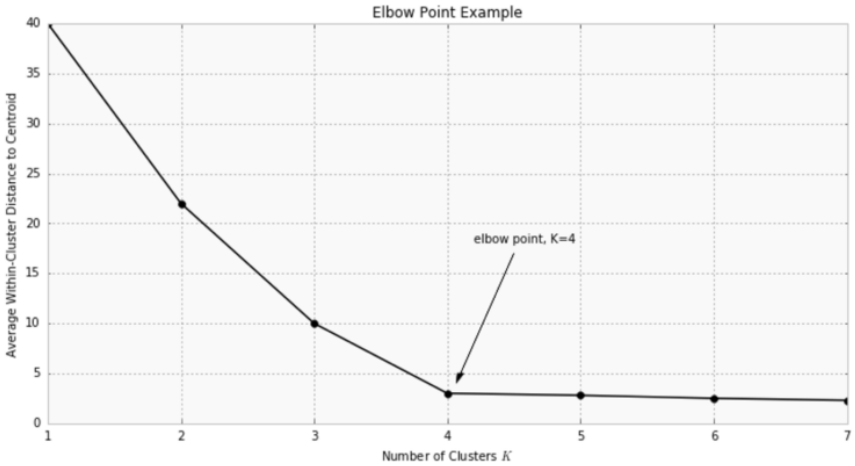
- **Métodos jerárquicos:**
  - **Dendograma:** puede ayudar a visualizar los clústeres. El corte del dendrograma en un nivel específico puede determinar el número óptimo de clústeres.

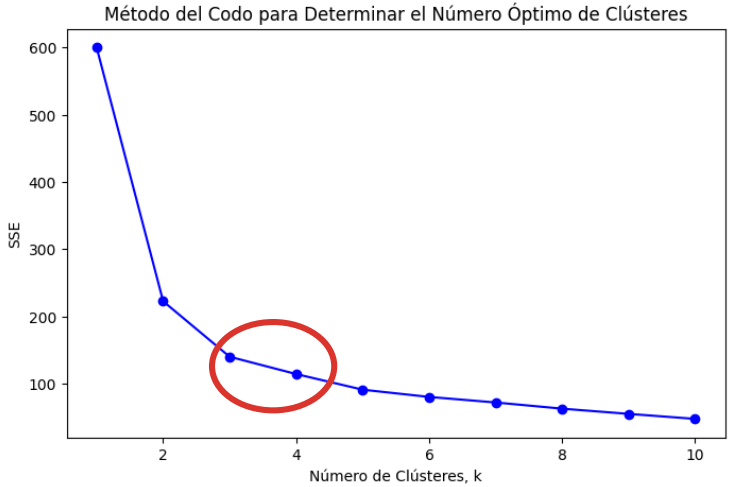
- **Métodos basados en densidad:**
  - No se estima el número de clústeres, en su lugar se estima ϵ y $MinPts$ utilizando el método k-neighbour.

6. Cómo elegir el número de clústeres apropiado

Método del codo

- **Métodos basados en centroides:**
  - Método del codo (**elbow):** se ejecuta K-means para un rango de valores de k y se calcula la suma de los errores cuadráticos (SSE) para cada k. El punto donde la gráfica de SSE vs. k forma un codo indica el número óptimo de clústeres, es decir, el punto después del cual la gráfica de SSE comienza a disminuir de forma lineal.



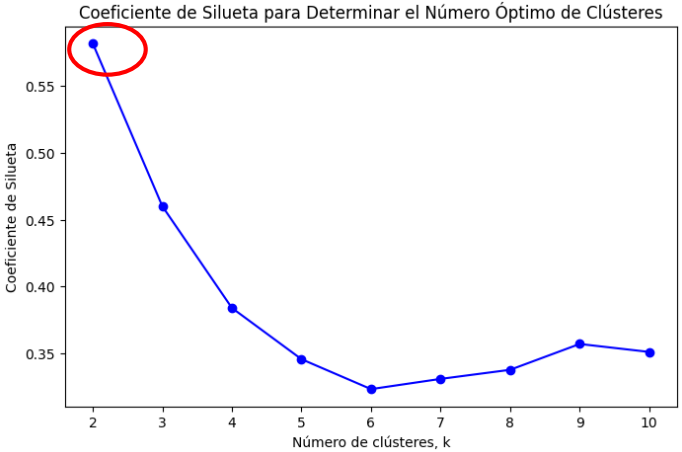


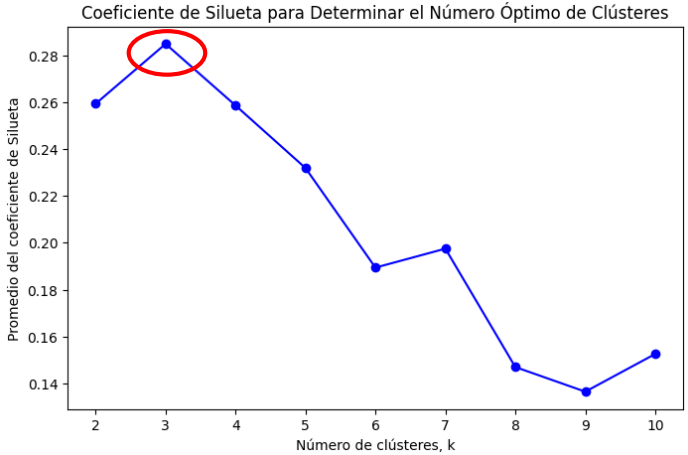
https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/

6. Cómo elegir el número de clústeres apropiado

Coeficiente de Silueta

- **Métodos basados en centroides:**
  - **Coeficiente de Silueta (Silhouette coefficient):** Se calcula para cada k y se selecciona el valor de k que maximiza el promedio de los coeficientes calculados.
  - Se elige el mayor valor. El valor más alto es 1 y el peor es -1. Valores cercanos a 0 indican que hay solapamiento en los clústeres. Valores negativos suelen indicar que un ejemplo se ha asignado a un clúster erróneo.

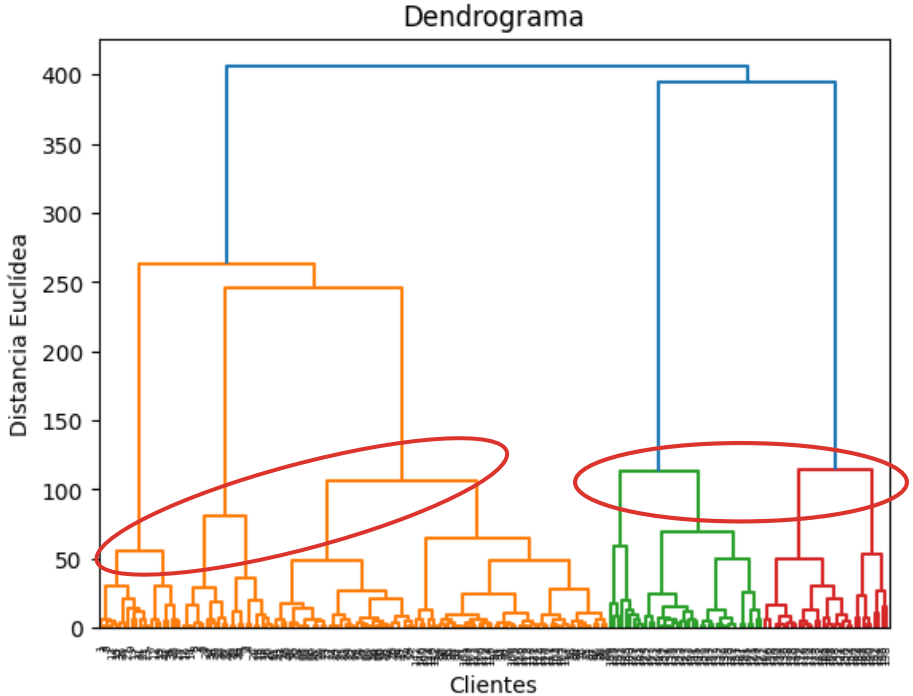




6. Cómo elegir el número de clústeres apropiado

Dendograma

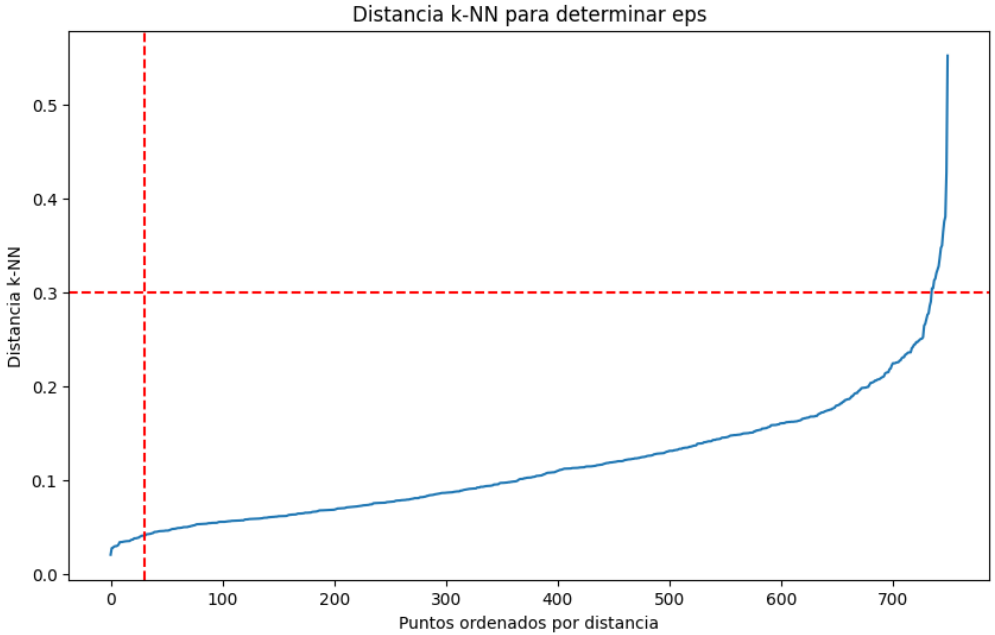
- **Métodos jerárquicos:**
  - **Dendograma:** puede ayudar a visualizar los clústeres. El corte del dendrograma en un nivel específico puede determinar el número óptimo de clústeres.



6. Cómo elegir el número de clústeres apropiado

Elección de ϵ y MinPts

- **Métodos basados en densidad:**
  - No se estima el número de clústeres, en su lugar se estima ϵ y $MinPts$ utilizando el método k-NN.



7. Otros algoritmos

Más...

Clustering difuso
Los objetos pertenecen a los clústeres con un grado [0, 1]

Modelos de mezcla
Se ajustan los parámetros de un modelo probabilístico (e.g., Gaussiano) con los datos

Métodos basados en rejillas
Dividen el espacio en regiones organizadas jerárquicamente y ordenadas según densidad

Métodos basados en grafos
Se forma un grafo con los objetos y se aplican técnicas de búsqueda de comunidades

Clustering con subespacios
Aplican reducción de dimensionalidad y trabajan en el espacio reducido

Métodos de búsqueda y optimización
Aplican técnicas de optimización (Tabu Search, Simmulated Annealing, Algoritmos Genéticos, etc.) para optimizar una métrica de calidad

Clustering de objetos complejos
Aplicaciones sobre series temporales, transacciones de datos, etc.

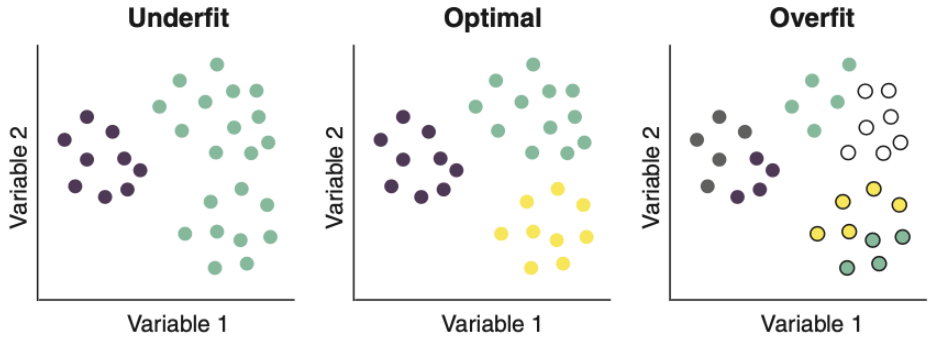
8. Medidas de calidad del clustering

Motivación

Para seleccionar *k* necesitamos discriminar entre agrupaciones buenas y no tan buenas

¿Cómo formalizar esta idea?

- Número de elementos, densidad de elementos, distancia entre las agrupaciones, etc. [métodos *intrínsecos*]
- Comparar con un *ground truth,* donde se conozcan las agrupaciones ideales [métodos *extrínsecos*]



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

8. Medidas de calidad del clustering

Coeficiente de Silueta

Cuantifica cómo de lejano es un objeto de un clúster a los objetos de los otros clústeres; esto es, si está separado del borde o no

Cálculo: para cada objeto *i* se mide la distancia media entre él y los otros puntos del mismo clúster (*a(i)*), y la distancia media entre el objeto *i* y los puntos del clúster más cercano (*b(i)*); finalmente, se calcula el ratio entre ambas

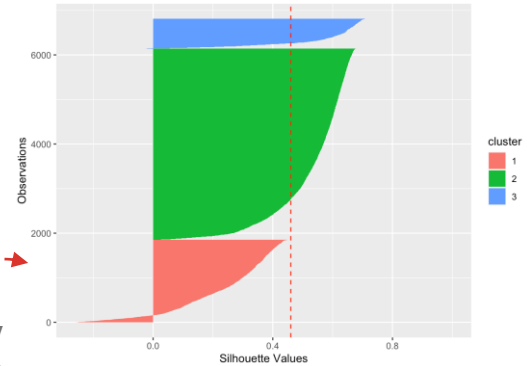
Rango: [−1, 1], mejor cuanto más próximo a 1

Interpretación: Valores cercanos a 1 indican que los puntos están bien agrupados; valores cercanos a 0 indican clústeres superpuestos; valores negativos indican que los puntos están mal clasificados.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$



Buscamos formas homogéneas y que estén por encima de la media

8. Medidas de calidad del clustering

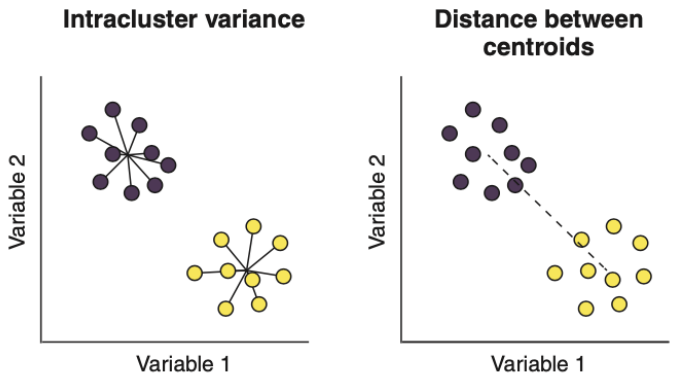
Índice de Davies-Boulding

Cuantifica la “separabilidad media” de un clúster frente a la del clúster más cercano

Cálculo: varianza dentro de los clústeres (dispersión o *scatter*) / separación entre centroides

Rango: 0 a ∞. Mejor cuanto más pequeño

Interpretación: Un índice más bajo indica que los clústeres son más compactos y bien separados.



$$\text{scatter}_k = \left(\frac{1}{n_k}\sum_{i \in k}^{n_k}(x_i - c_k)^2\right)^{1/2}$$

$$\text{separation}_{j,k} = \left(\sum_{1 \le j \le k}^{N}(c_j - c_k)^2\right)^{1/2}$$

$$\text{ratio}_{j,k} = \frac{\text{scatter}_j + \text{scatter}_k}{\text{separation}_{j,k}}$$

$$DB = \frac{1}{N}\sum_{k=1}^{N} R_k$$

R_k es el máximo $\text{ratio}_{j,k}$

H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

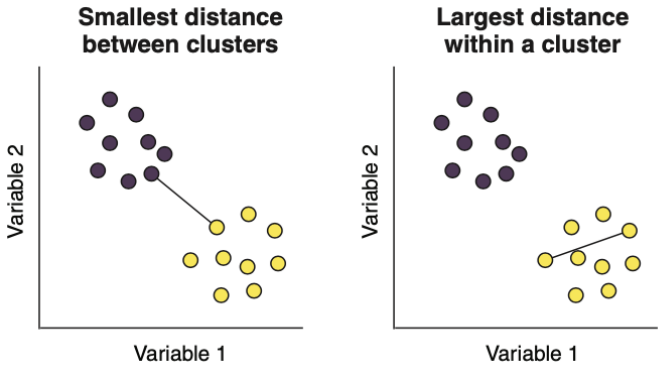
8. Medidas de calidad del clustering

Índice de Dunn

Cuantifica el ratio entre la distancia mínima entre puntos de diferentes clústeres y el diámetro máximo de los clústeres

Rango: 0 a ∞. Mejor cuanto más alto

Interpretación: Un índice más alto indica que los clústeres están bien separados y son internamente compactos



H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning, 2020.

$$\text{Dunn} = \min_{1 \le i \le k} \left\{ \min_{k} \left(\frac{\delta(c_i, c_j)}{\max_{1 \le i \neq i \le k} \Delta(c_k)} \right) \right\}$$

$\delta(c_i, c_j)$ es la distancia entre todos los pares de los clústeres *i, j*

$\Delta(c_k)$ es la distancia máxima entre objetos del clúster *k*

8. Medidas de calidad del clustering

Índice DBCV

Índice DBCV (Density-Based Clustering Validation) está específicamente diseñada para evaluar la calidad de los clústeres generados por algoritmos de clustering basados en densidad, como DBSCAN u OPTICS.

Rango: [-1, 1]. Mejor cuanto más próximo a 1

Interpretación:

- Valor cercano a 1: Indica que el clustering es bueno, con clústeres bien definidos y separados, y una alta densidad dentro de los clústeres.

- Valor cercano a -1: Indica que el clustering es pobre, con clústeres mal definidos o con mucha superposición entre ellos.

- Valor cercano a 0: Sugiere que el clustering no es claramente bueno ni malo, indicando una calidad de clustering intermedia.

8. Medidas de calidad del clustering

Índice de Rand y Jaccard

Los índices de Rand y Jaccard, son métodos extrínsecos, es decir, tenemos que saber el agrupamiento o la clasificación real de los datos.

Definición: El **Índice de Rand** se calcula como la proporción de decisiones correctas (pares de puntos que están en el mismo clúster en ambas clasificaciones o en diferentes clústeres en ambas clasificaciones) sobre el número total de decisiones (todos los pares de puntos posibles).

Rango: El Índice de Rand varía de 0 a 1, donde 1 indica una concordancia perfecta entre el clustering y la clasificación de referencia.

$$\textit{Índice Rand} = \frac{a+d}{a+b+c+d} \quad \text{donde:}$$

a es el número de pares de puntos que están en el mismo clúster en ambas clasificaciones.

d es el número de pares de puntos que están en diferentes clústeres en ambas clasificaciones.

b es el número de pares de puntos que están en el mismo clúster en la clasificación verdadera pero en diferentes clústeres en el clustering predicho.

c es el número de pares de puntos que están en diferentes clústeres en la clasificación verdadera pero en el mismo clúster en el clustering predicho.

8. Medidas de calidad del clustering

Índice de Rand y Jaccard

Definición: El **Índice de Jaccard** se calcula como la proporción de pares de puntos correctamente agrupados (tanto en el mismo clúster como en diferentes clústeres) sobre el total de pares de puntos que están en el mismo clúster en al menos una de las dos clasificaciones.

Rango: El Índice de Jaccard varía de 0 a 1, donde 1 indica una concordancia perfecta entre el clustering y la clasificación de referencia.

$$\textit{Índice Jaccard} = \frac{a}{a+b+c}$$ donde:

a es el número de pares de puntos que están en el mismo clúster en ambas clasificaciones.

b es el número de pares de puntos que están en el mismo clúster en la clasificación verdadera pero en diferentes clústeres en el clustering predicho.

c es el número de pares de puntos que están en diferentes clústeres en la clasificación verdadera pero en el mismo clúster en el clustering predicho.

Índice

Sesión 1

1. ¿Qué es el *clustering*?
2. Elementos básicos de un algoritmo de clustering
3. Métodos basados en centroides: k-means y k-medians
4. Clustering jerárquico: Aglomerativo y divisivo
5. Clustering basado en densidad: DBSCAN y OPTICS
6. Cómo elegir el número de clústeres apropiado
7. Otros algoritmos
8. Medidas de calidad del clustering

Sesión 2

1. Uso de algoritmos de clustering en Python
2. Ejercicio práctico