Aprendizaje automático

Clase 6

Maximiliano Beckel

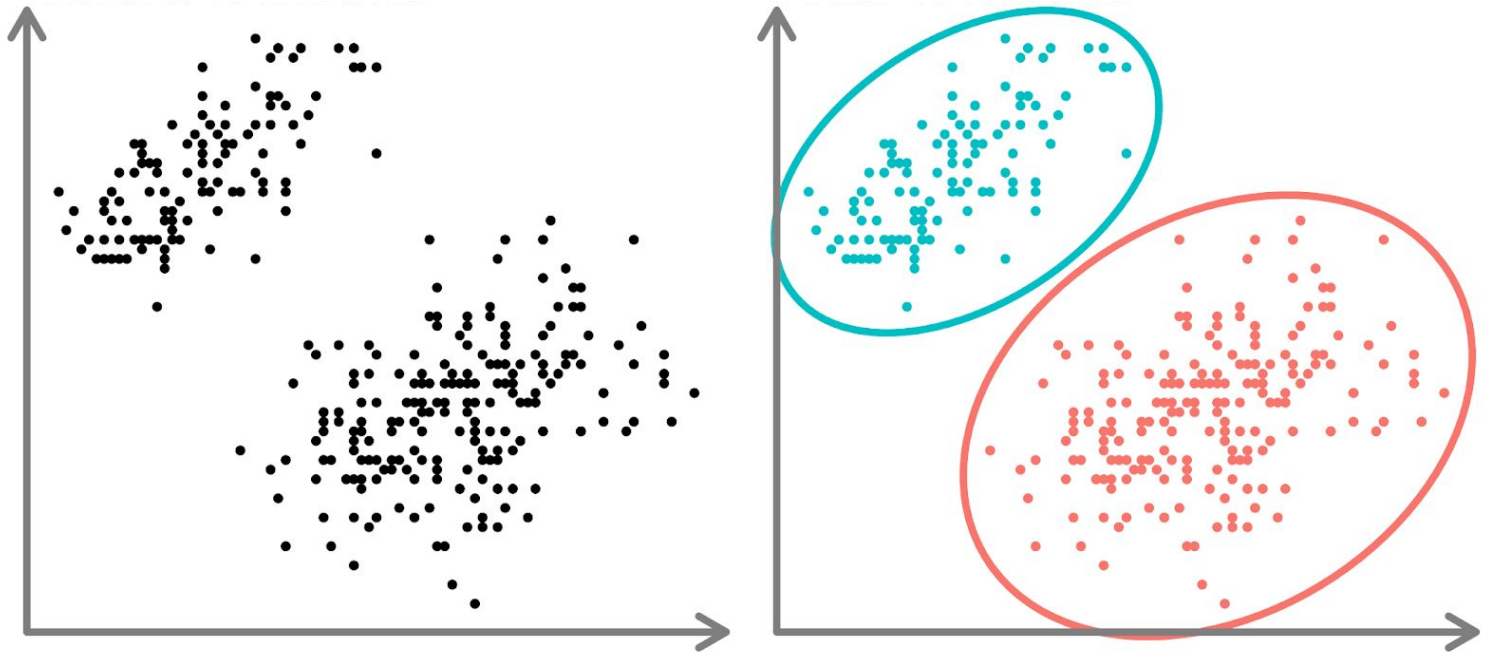
UNAHUR

Análisis de Clustering (grupos o conglomerados)

Objetivo
Separar las observaciones en distintos grupos que cumplan con:

1. Cada grupo (conglomerado o clúster) sea homogéneo respecto a las variable utilizadas para caracterizarlo, que cada observación contenida en él sea parecida a todas las que estén incluidas en ese grupo.
2. Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas.

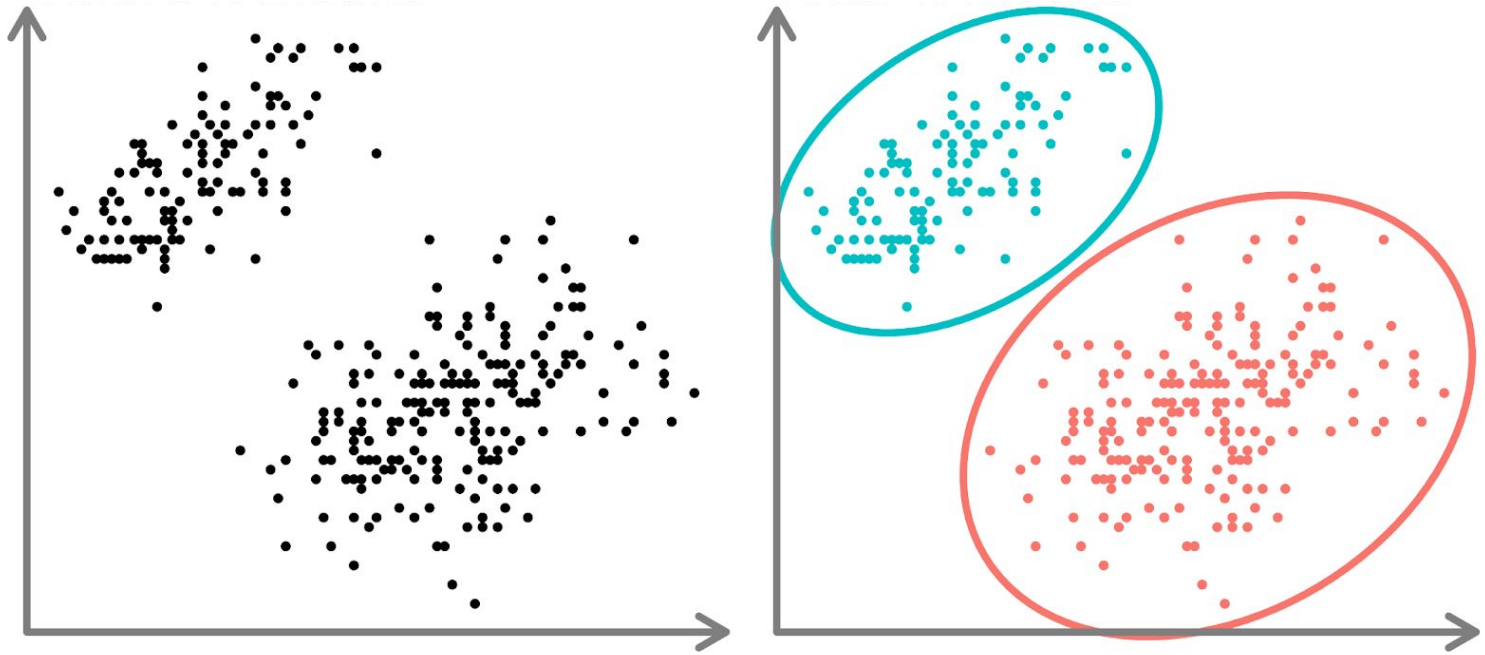
Problema: antes de realizar el análisis no sé cuántos grupos se pueden establecer a partir de mis datos.



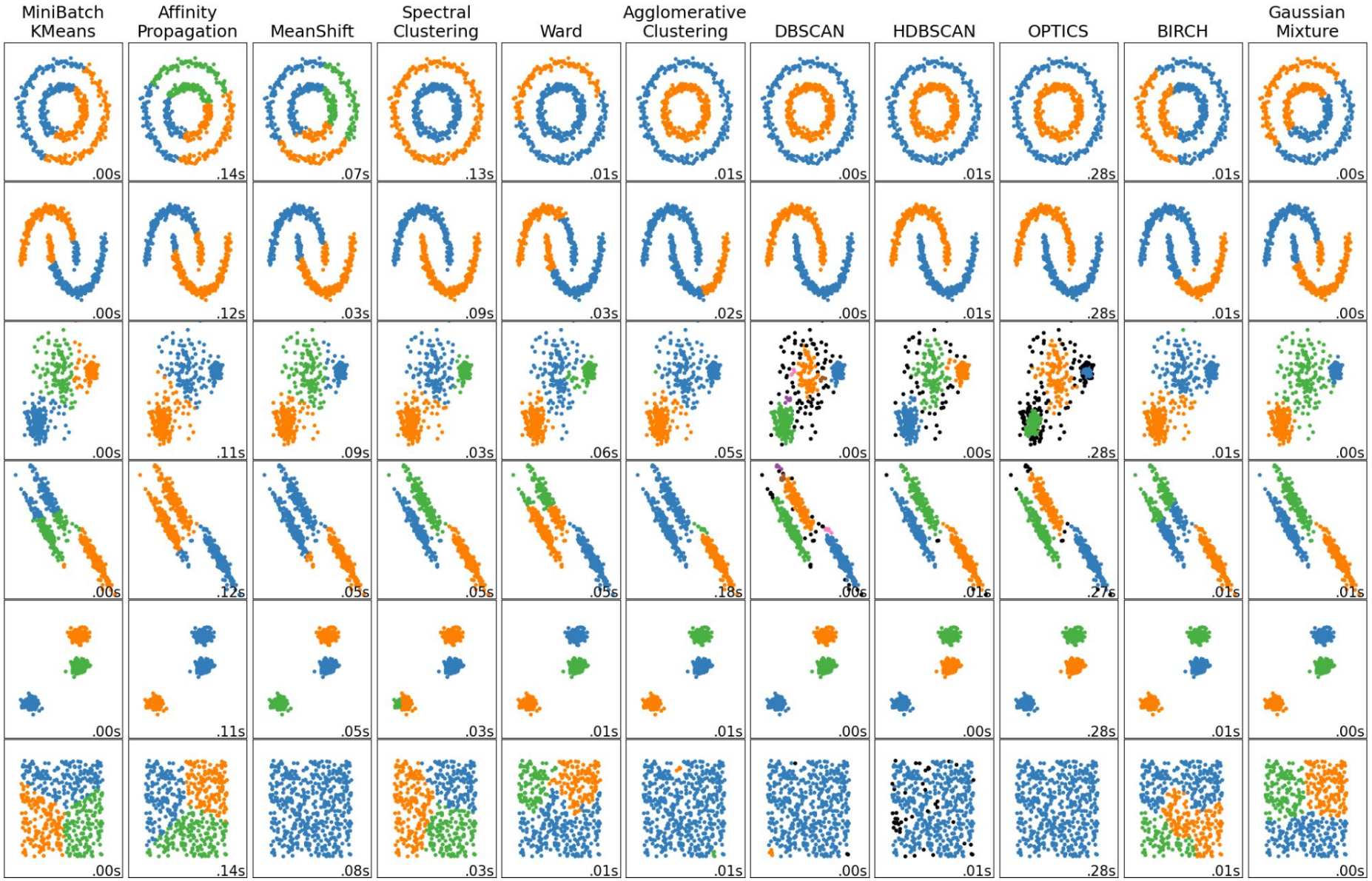
Análisis de Clustering (grupos o conglomerados)

Pasos a seguir:

1. Se parte de datos tabulares en las que tengo n observaciones, de las cuales se tiene información de k atributos
2. Estandarizar los atributos numéricos, transformar datos categóricos. Detectar outliers.
3. Establecer alguna medida que me permita saber qué observaciones se parecen más entre sí (medidas de similaridad) o qué tanto difieren (medidas de disimilaridad o de distancia)
4. Elegir el tipo de clustering que queremos llevar a cabo e implementarlo.
5. Tomar alguna medida de validación para comprobar la calidad de los grupos que se establecieron.



Dime cómo se distribuyen tus datos y te diré qué método de clustering elegir…



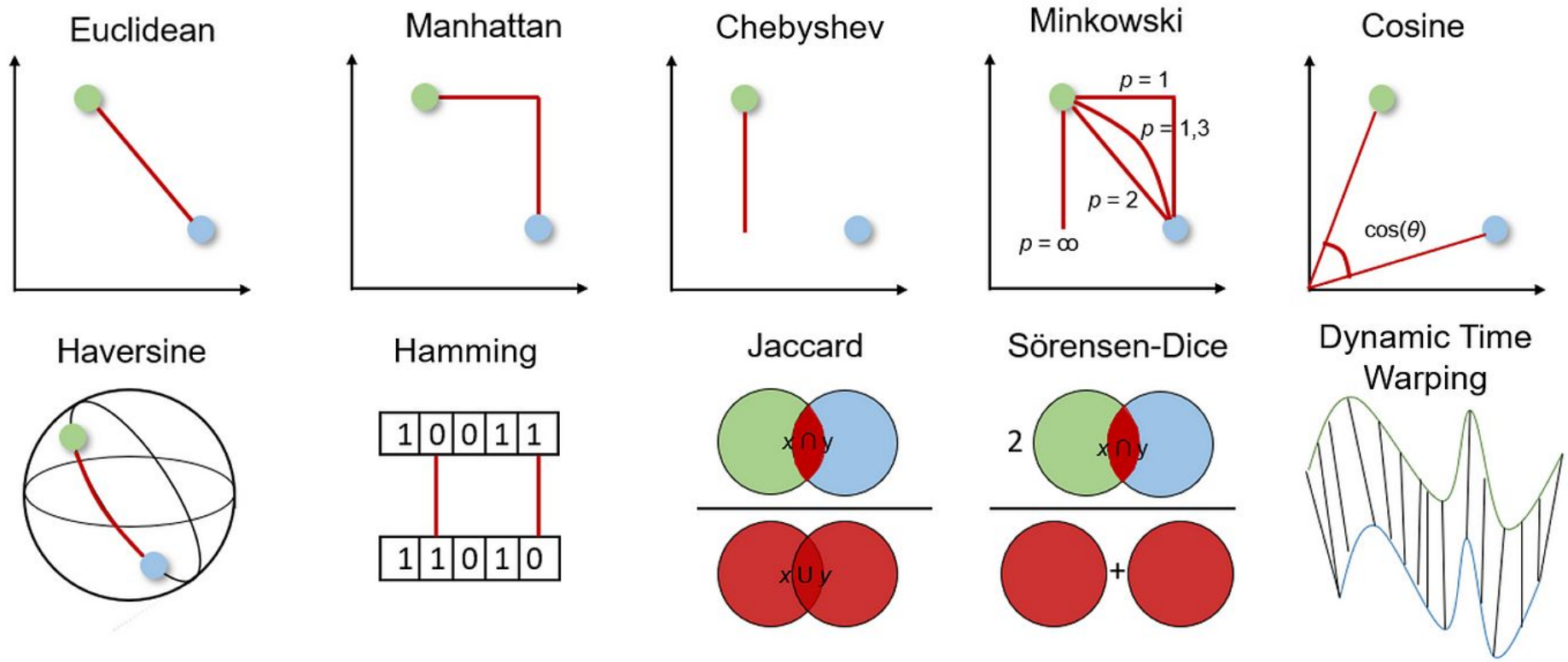
https://scikit-learn.org/stable/modules/clustering.html

Medidas de distancia

- La medida de distancia que elijamos va a depender del tipo de datos que tenemos. No es lo mismo calcular distancias entre dos puntos en una ciudad, entre dos barcos en alta mar o entre dos palabras.
- Hay algunas características que toda medida de distancia cumple:
  - d(A, A) = 0
  - d(A, B) >= 0
  - d(A, B) = d(B, A)
  - d(A,B) <= d(A,C) + d(C,B)

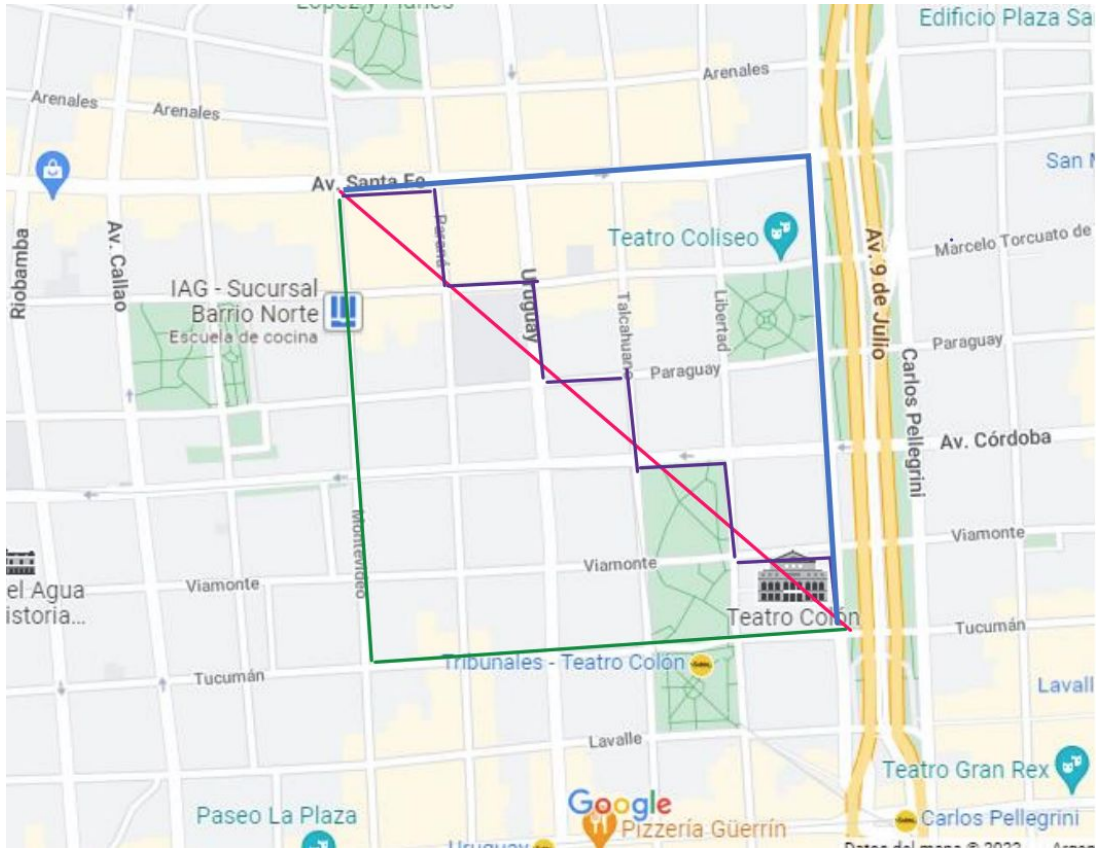
Medidas de distancia

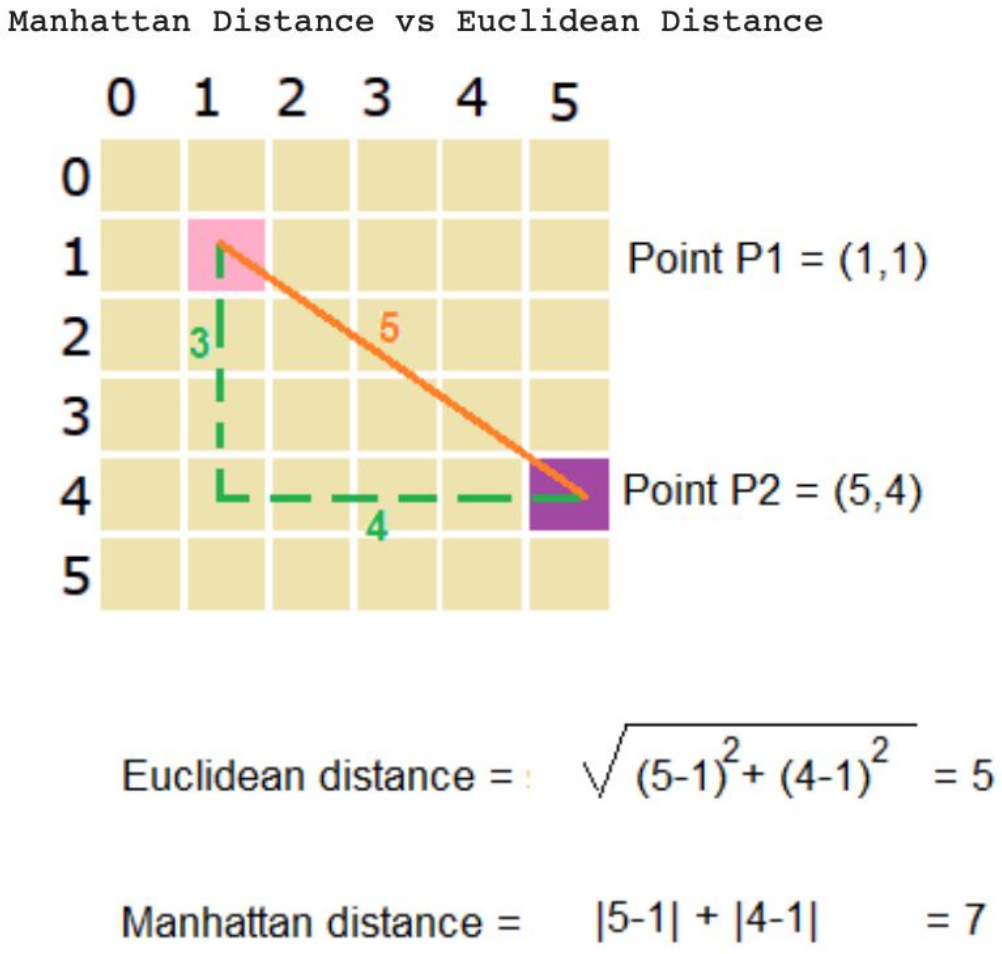
- La medida de distancia que elijamos va a depender del tipo de datos que tenemos. No es lo mismo calcular distancias entre dos puntos en una ciudad, entre dos barcos en alta mar o entre dos palabras.
- Hay algunas características que toda medida de distancia cumple:
  - d(A, A) = 0
  - d(A, B) >= 0
  - d(A, B) = d(B, A)
  - d(A,B) <= d(A,C) + d(C,B)



Medidas de distancia

- Distancia Euclídea VS Distancia Manhattan





Tipos de Clustering

1. **Jerárquico vs Partitivo (no jerárquico)**

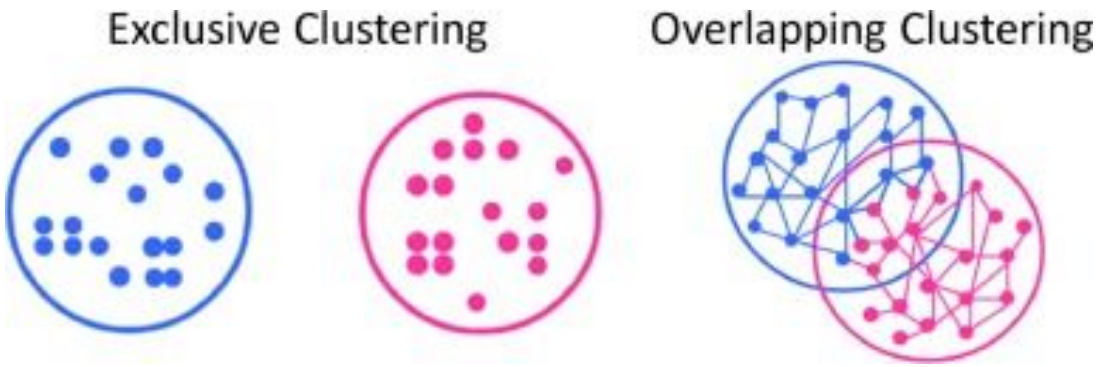
   En un clustering jerárquico cada cluster presenta una estructura de anidamiento, es decir, los clusters pueden tener a su vez sub-clusters, que a su vez puede tener sub-sub… Formando una estructura de árbol que llamamos dendrograma. En cambio, en los algoritmos de clustering partitivo cada observación pertenece a un grupo determinado.

1. **Exclusivo vs Superpuesto vs Difuso**

   ¿A qué cluster pertenece una dada observación? Si cada observación puede pertener a un único cluster, entonces es un clustering exclusivo. En cambio, si una dada observación puede pertenecer a más de un cluster tenemos un clustering superpuesto (overlapping). En el clustering difuso (Fuzzy) se asigna a cada observación una probabilidad de pertenecer a cada uno de los clusters sin asignarla a uno en particular.

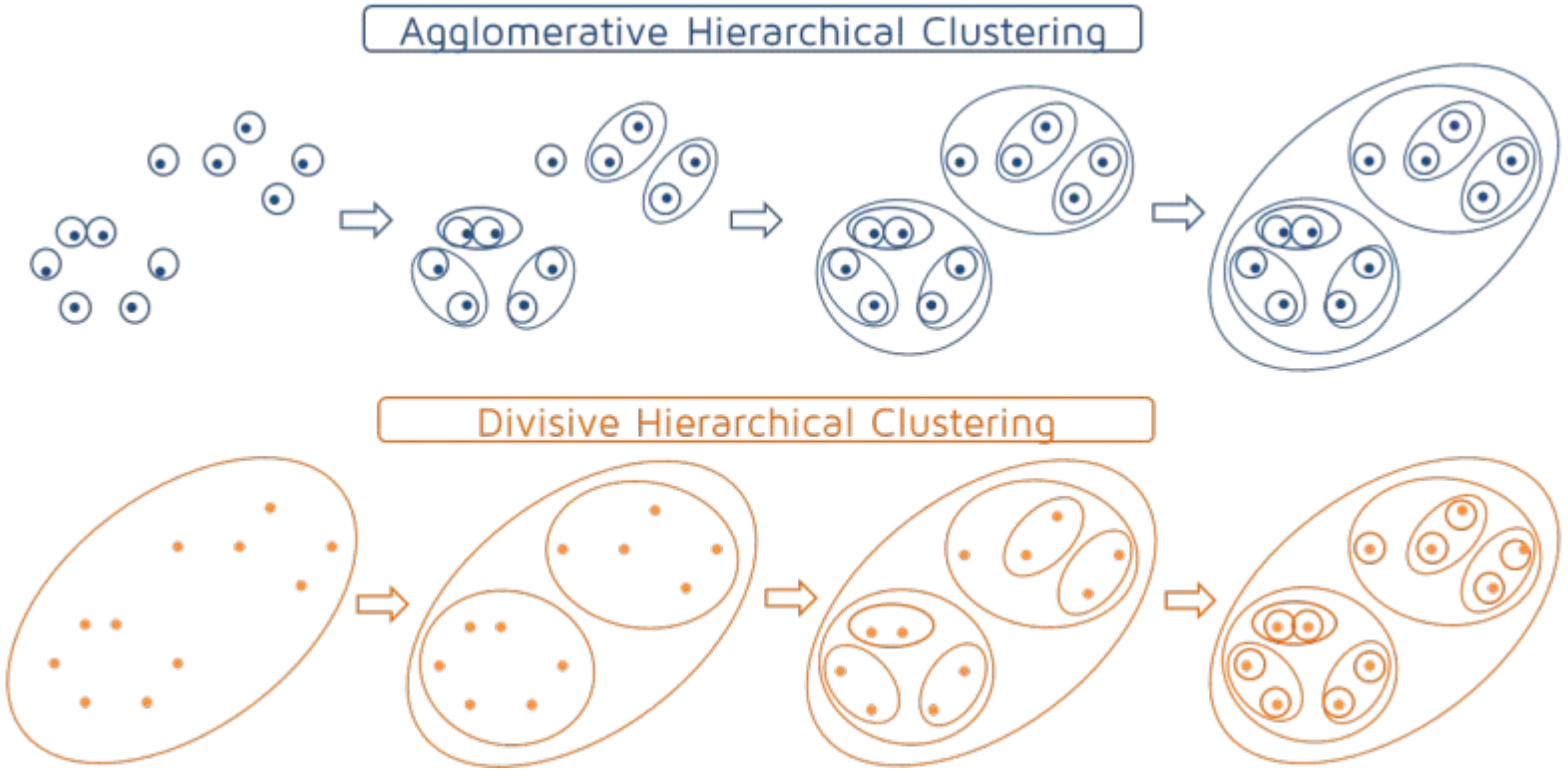
1. **Completo vs Parcial**

   ¿Puede haber observaciones que no pertenezcan a ningún cluster? Si la respuesta es no, entonces el clustering será completo. En caso contrario, tendremos un clustering parcial.



Clustering Jerárquico

- **Aglomerativo**: Se parte de clusters individuales (singleton, hojas) y se van uniendo los más cercanos.
- **Divisivo**: Se parte de un sólo cluster (root, raíz) y se van separando hasta quedarse sólo con los clusters individuales.



Clustering Jerárquico

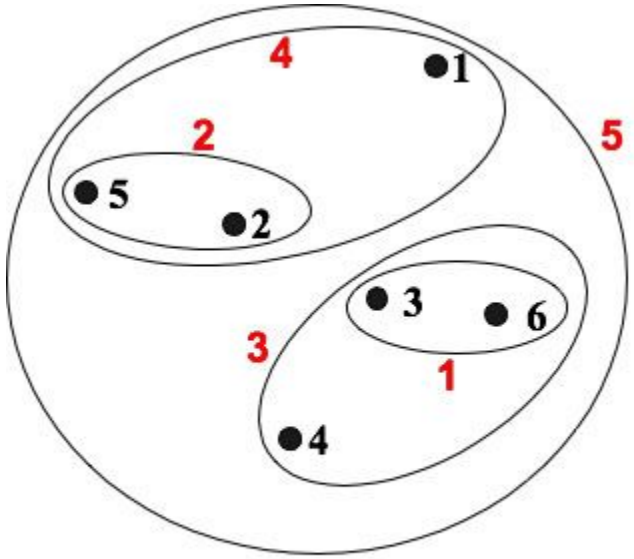
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

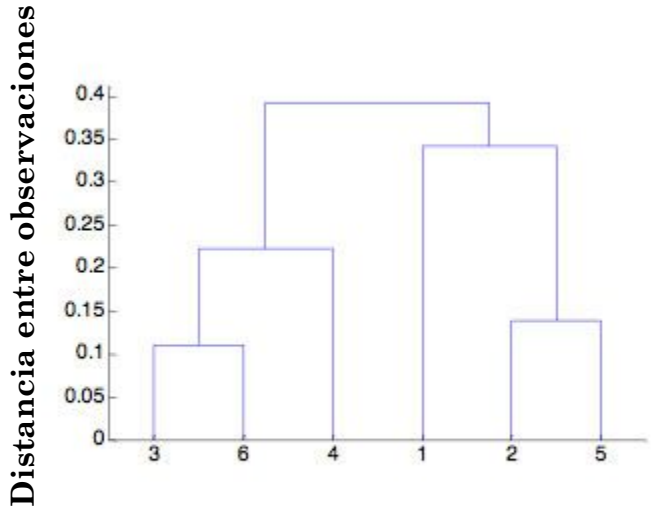
Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters



Dendrogram

Clustering Jerárquico

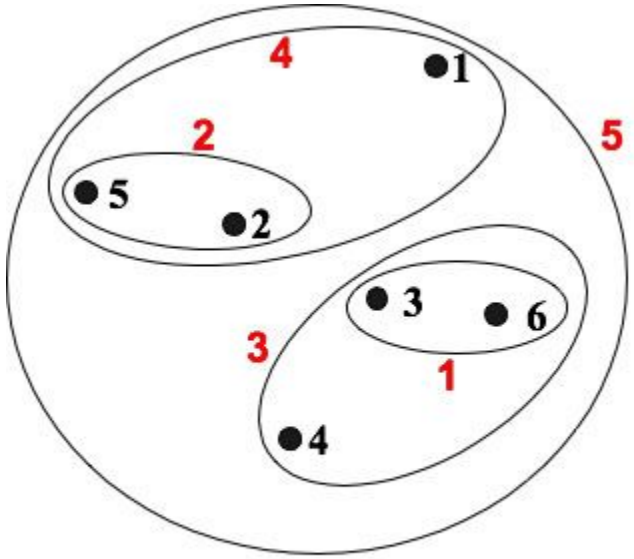
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

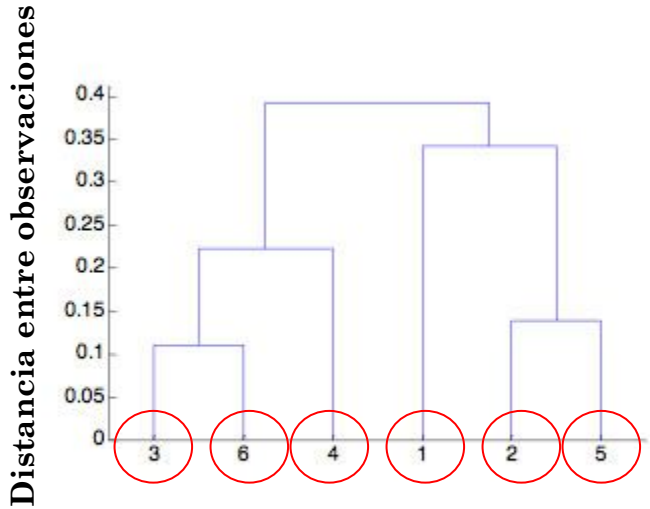
Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters



Dendrogram

Clustering Jerárquico

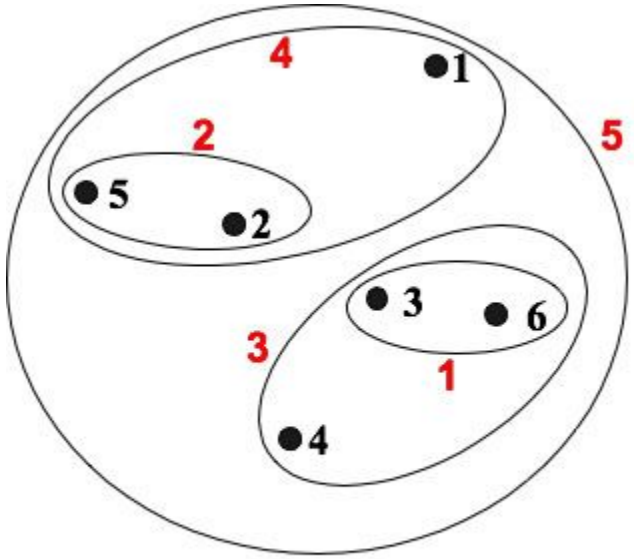
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

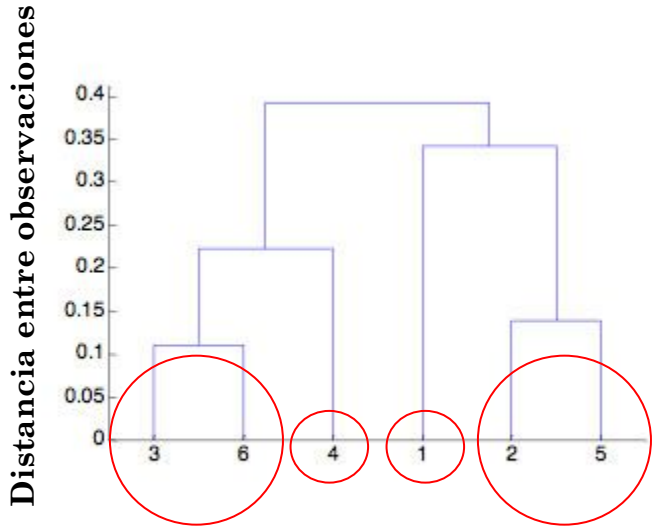
Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters



Dendrogram

Clustering Jerárquico

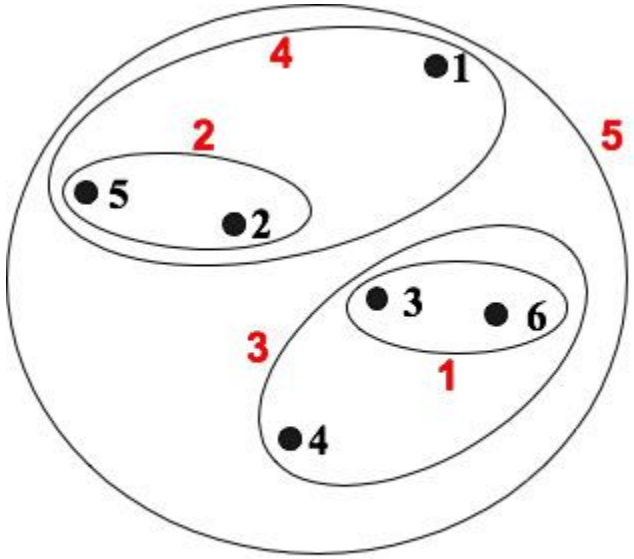
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

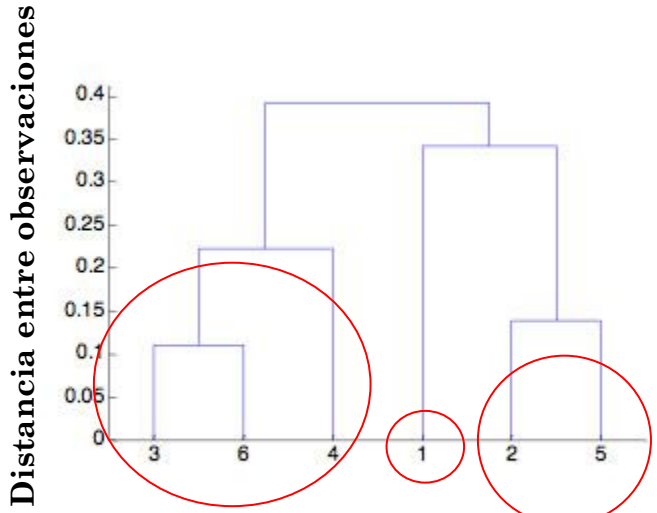
Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters



Dendrogram

Clustering Jerárquico

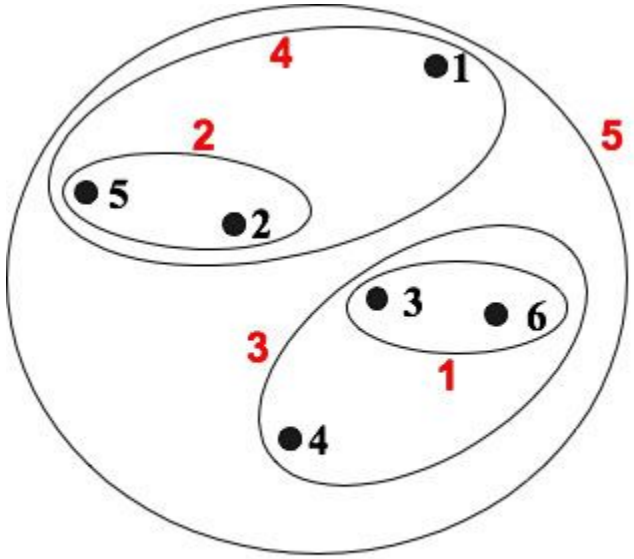
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

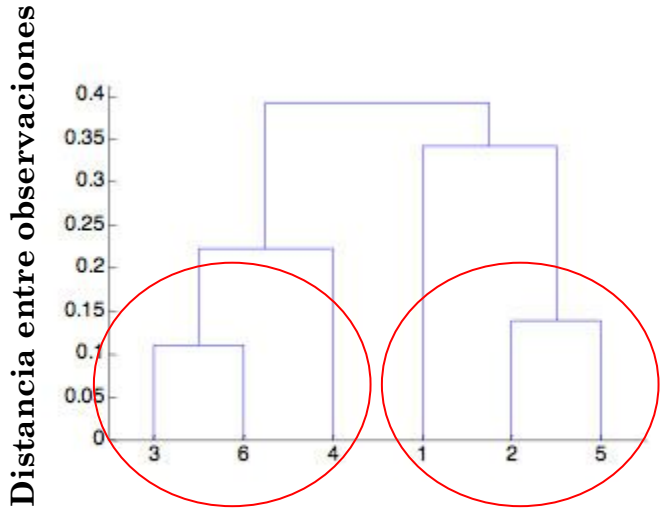
Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters



Dendrogram

Clustering Jerárquico

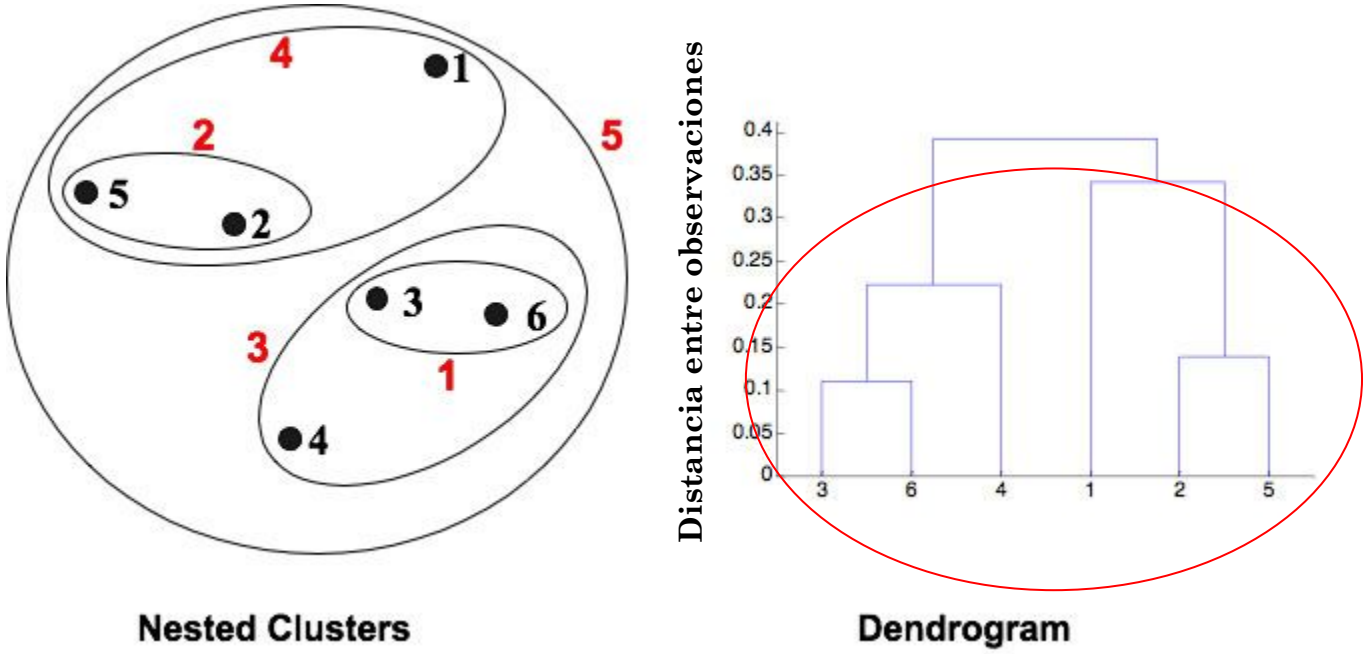
Clustering jerárquico aglomerativo

0. Computar la matriz de similaridad.

Repetir:

1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Clustering Jerárquico

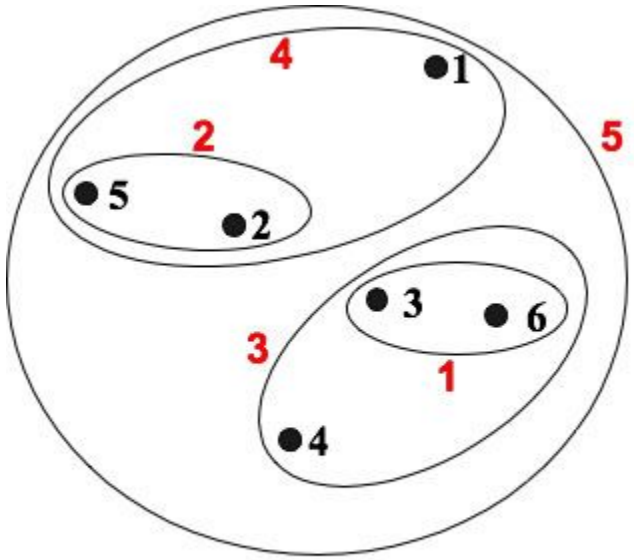
<u>Clustering jerárquico aglomerativo</u>

0. Computar la matriz de similaridad.

Repetir:

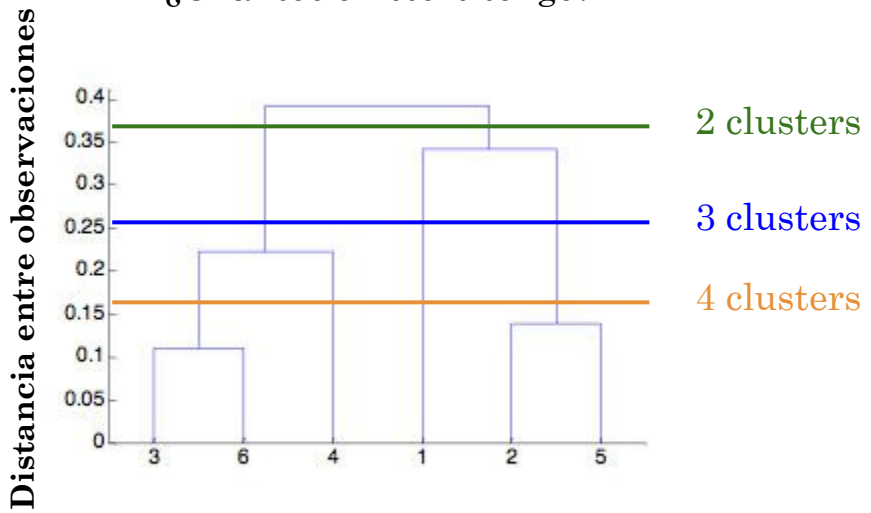
1. Juntar los dos más cercanos.
2. Actualizar la matriz de similaridad utilizando el nuevo cluster.

Hasta que: Sólo quede un cluster



Nested Clusters

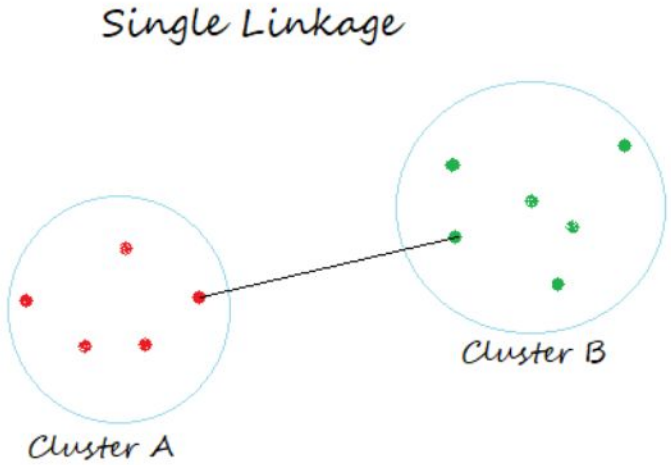
¿Cuántos clusters tengo?

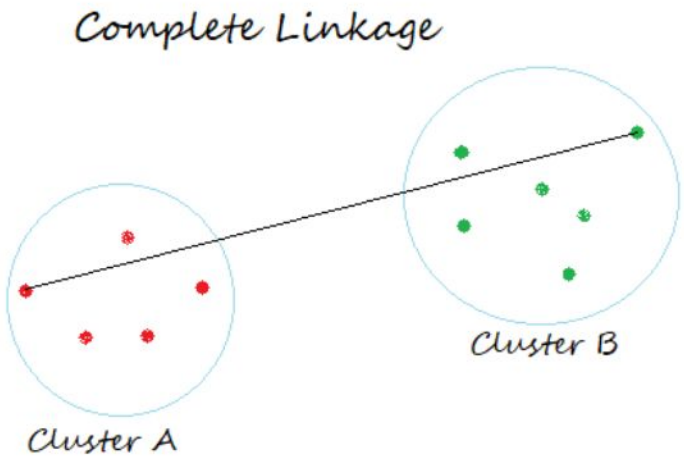


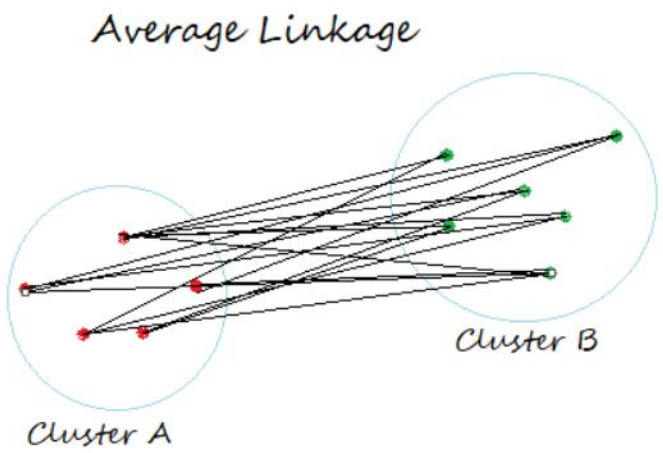
Dendrogram

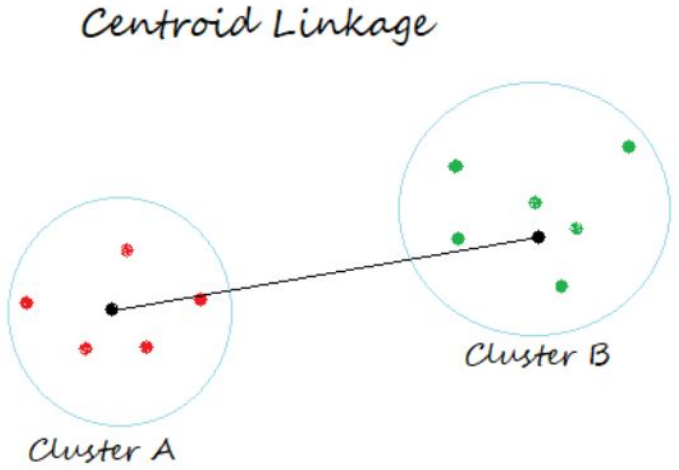
Clustering Jerárquico

¿Cómo calculamos la similaridad entre los clusters?









Todas estas formas de calcular la similaridad entre dos clusters tiene sus pros y contras, el mejor dependerá de los datos que tengas (y de nuestra experimentación)

Clustering No Jerárquico: **k-means**

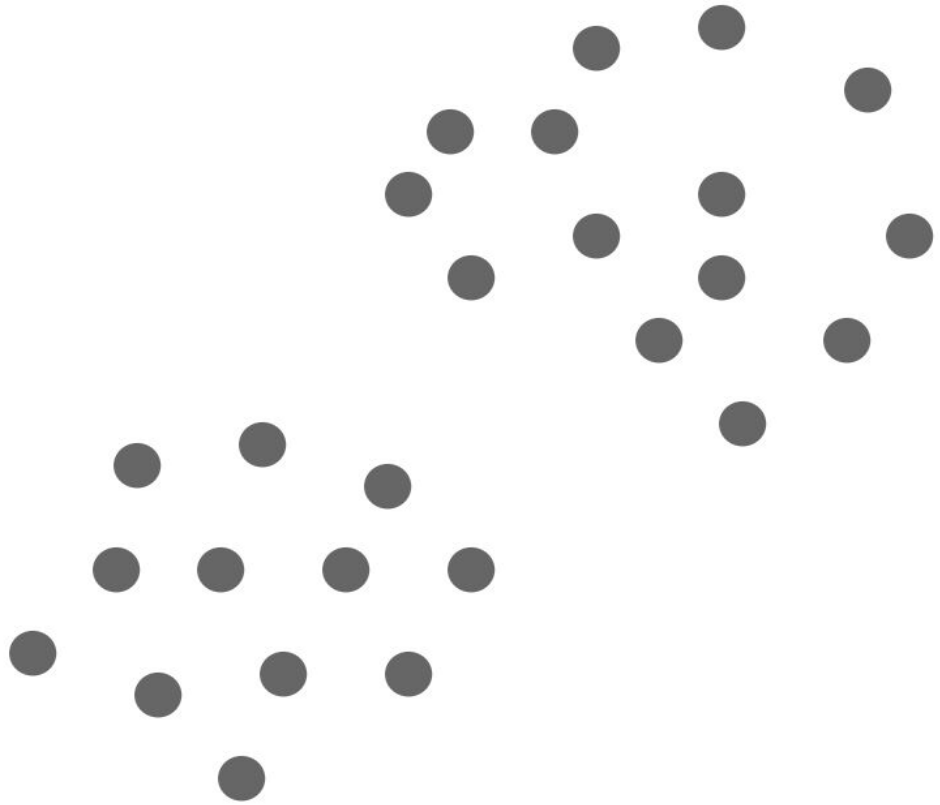
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K. K = 2
   1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

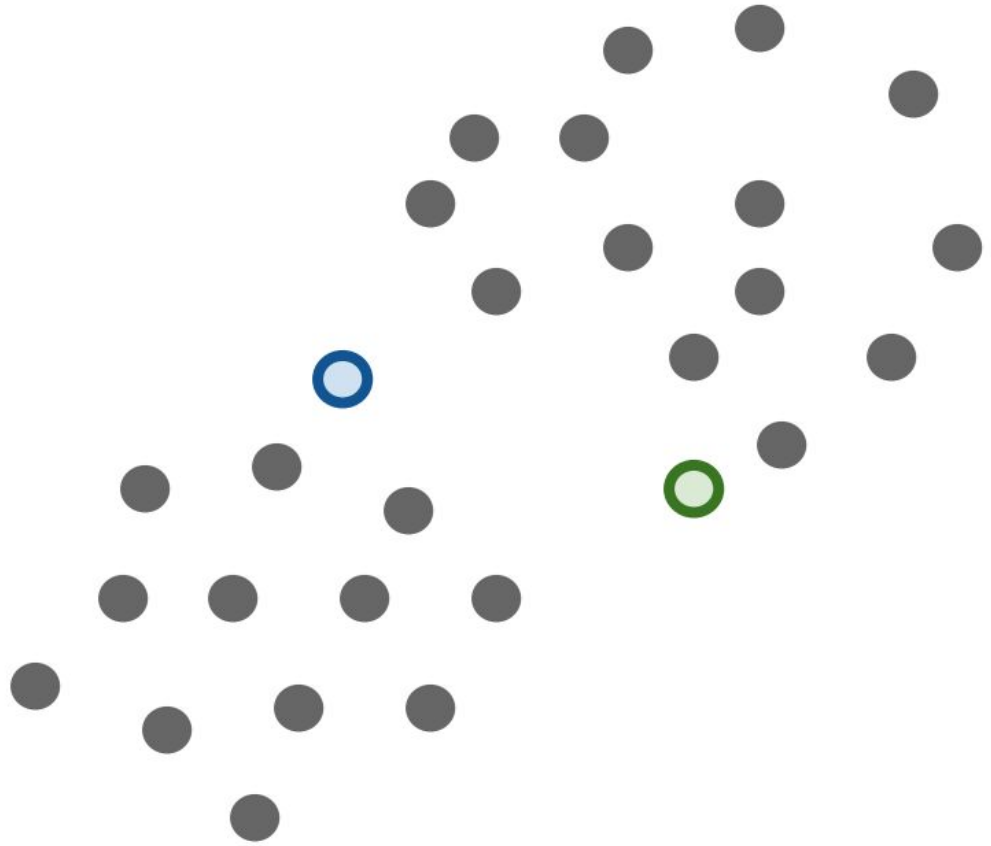
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

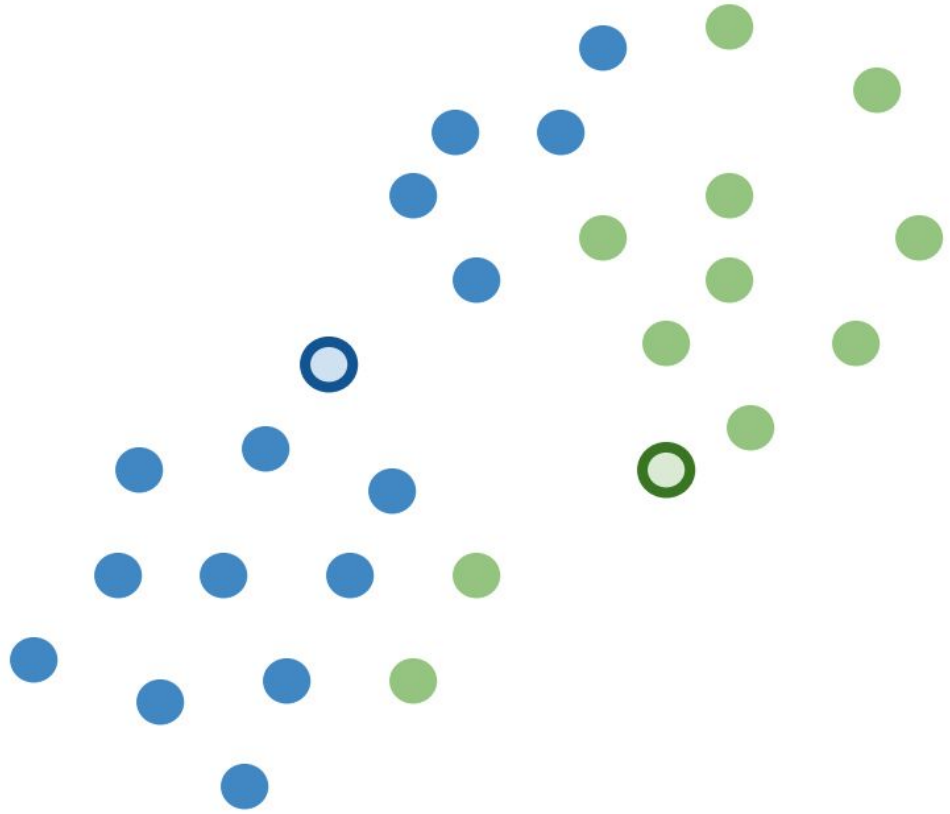
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

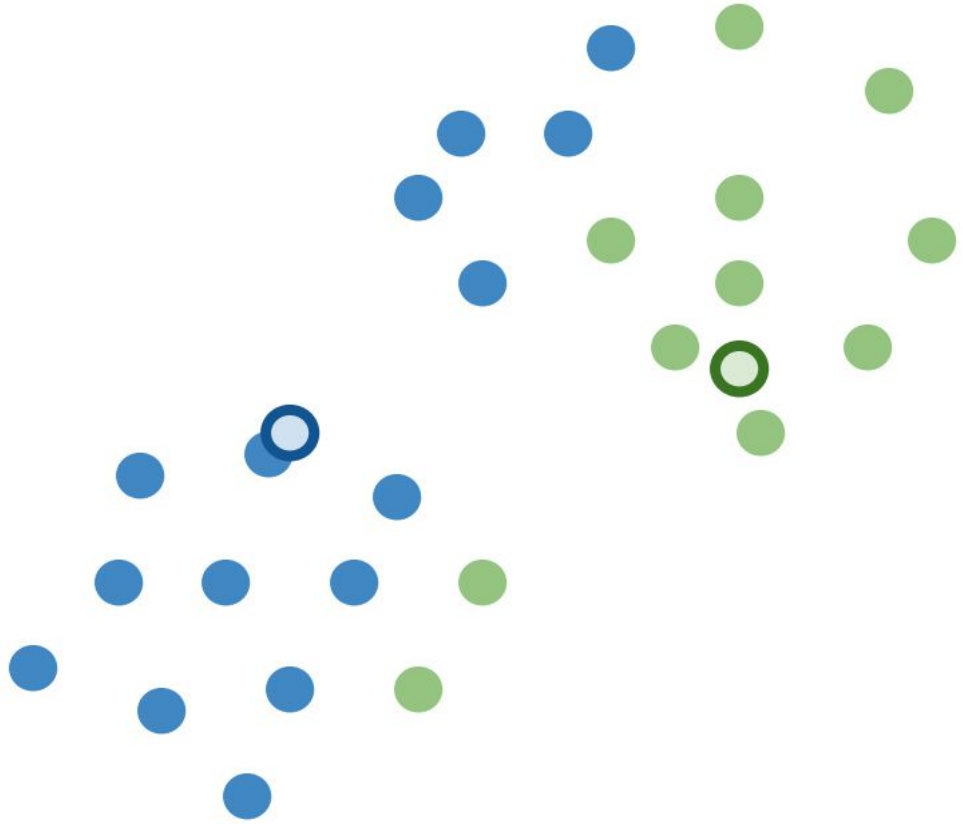
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

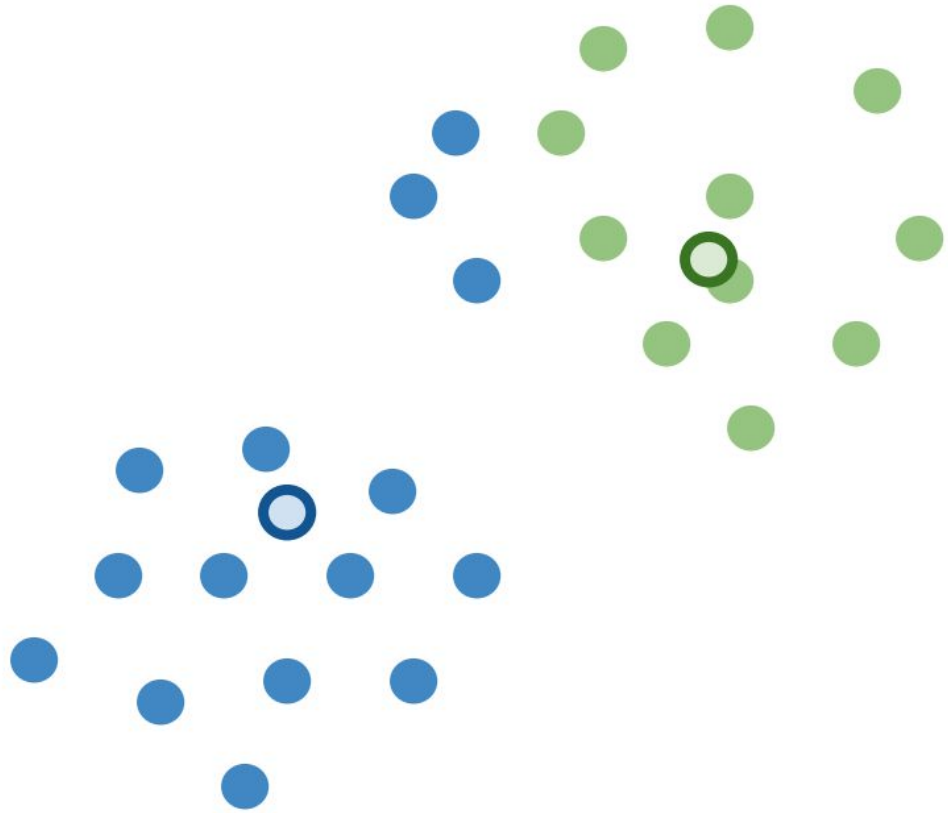
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

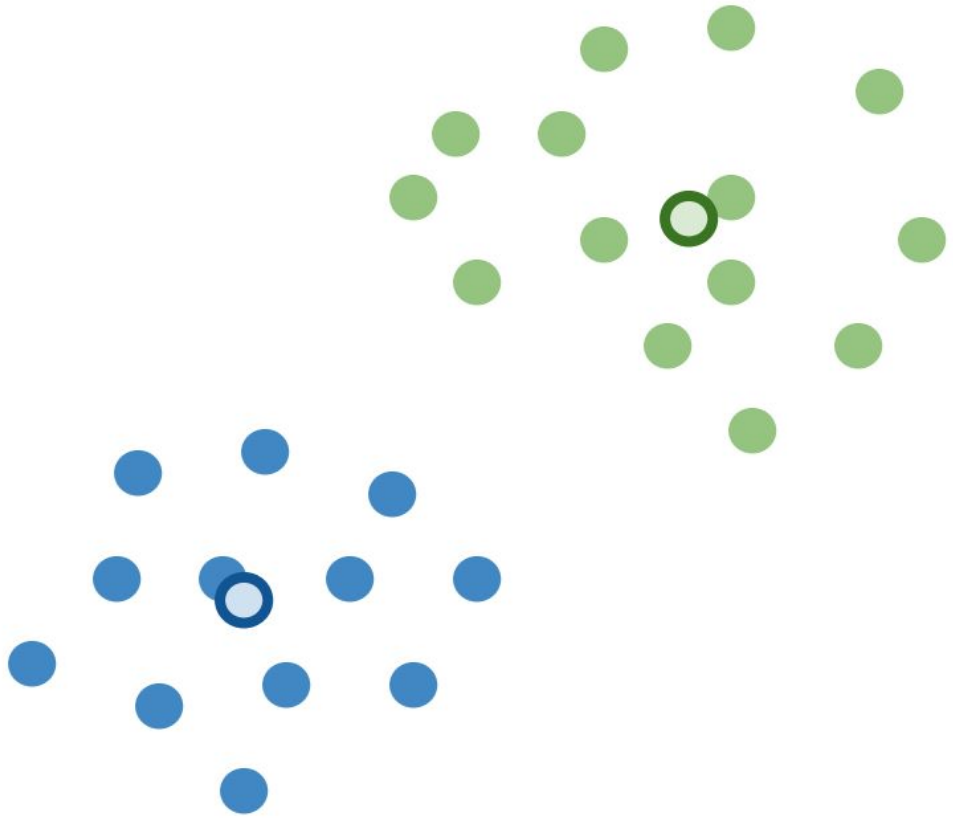
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien



Clustering No Jerárquico: **k-means**

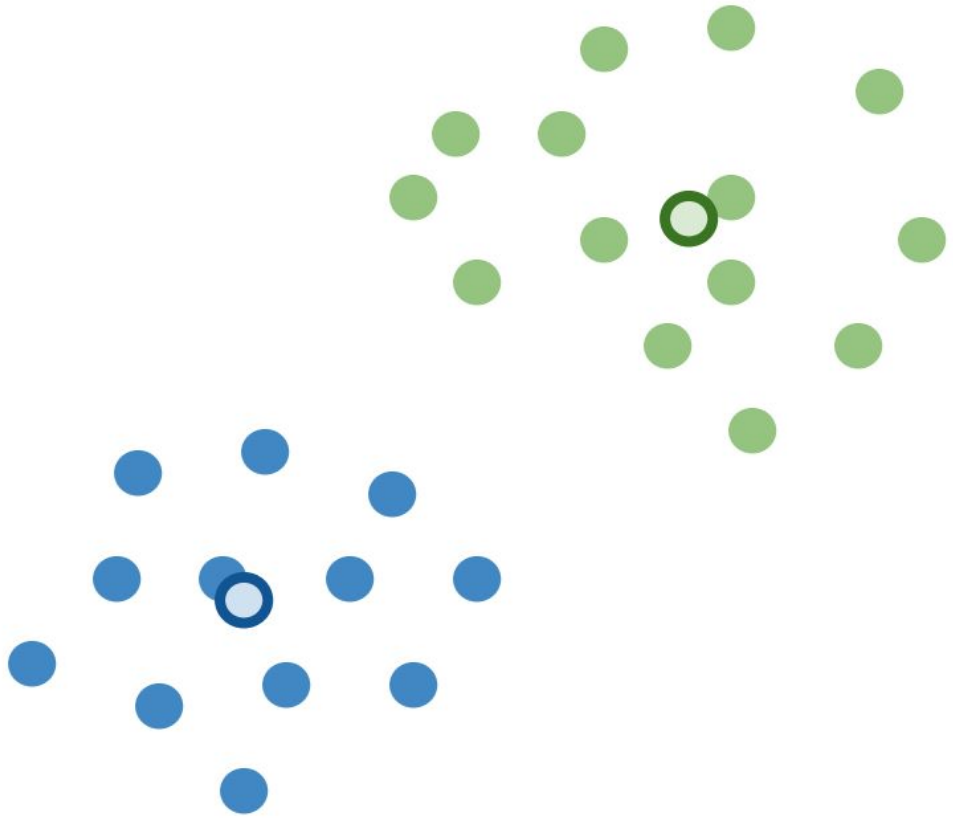
K-means es un método iterativo que busca generar un agrupamiento de k clusters. Veamos cómo funciona…

0. Seleccionar K.
1. Seleccionar K puntos como centroides iniciales.

Repetir:

1. Asignar cada punto a uno de los K clusters.
2. Recomputar los centroides de cada clusters.

Hasta que: los centroides no cambien

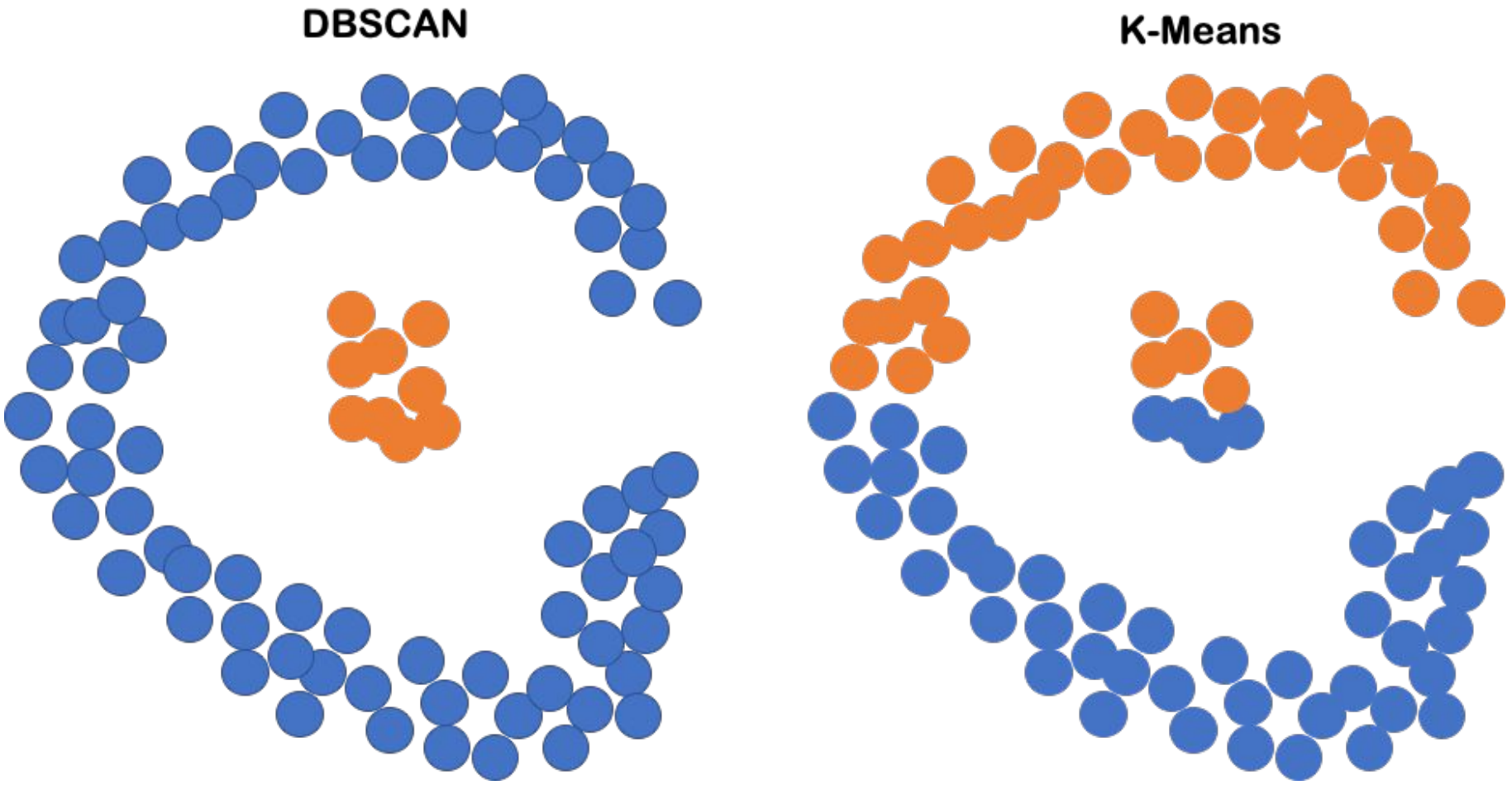


Algunos problemitas…

- Los resultados pueden ser dependientes de la inicialización
- Método sensible a outliers.
- No es la mejor opción si tenemos clusters que no sean compactos
- ¿Cómo elegimos el k?

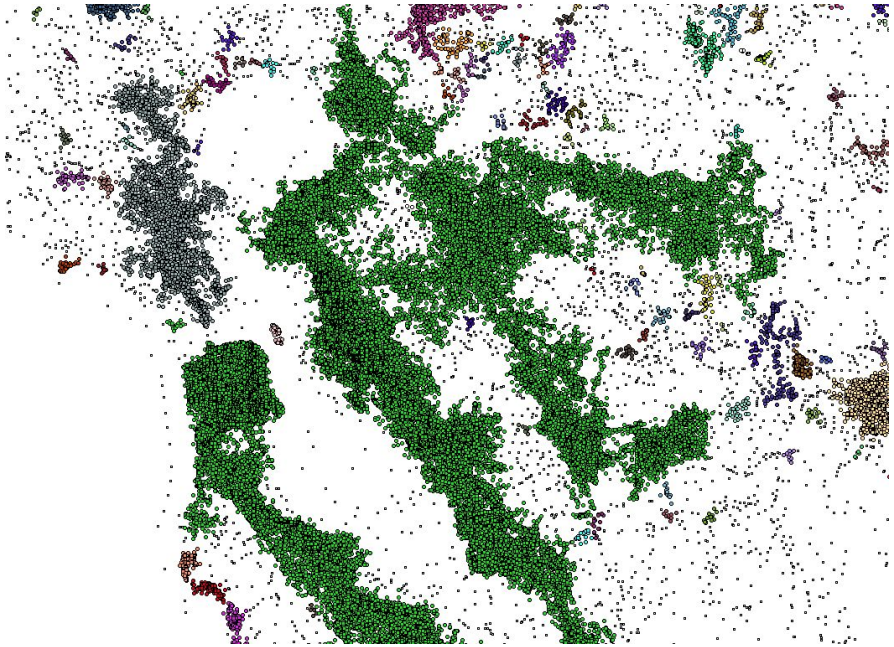
DBSCAN (Density-based Spatial Clustering of Applications with Noise)

Muchos métodos de clustering se especializan en encontrar clusters con formas compactas pero entran en problemas para identificar clusters con formas arbitrarias, como por ejemplo alargadas o en forma circular. DBSCAN es un método de clustering que resuelve este problema al identificar zonas del espacio con alta densidad de puntos, independientemente de cuál sea su forma.



DBSCAN (Density-based Spatial Clustering of Applications with Noise)

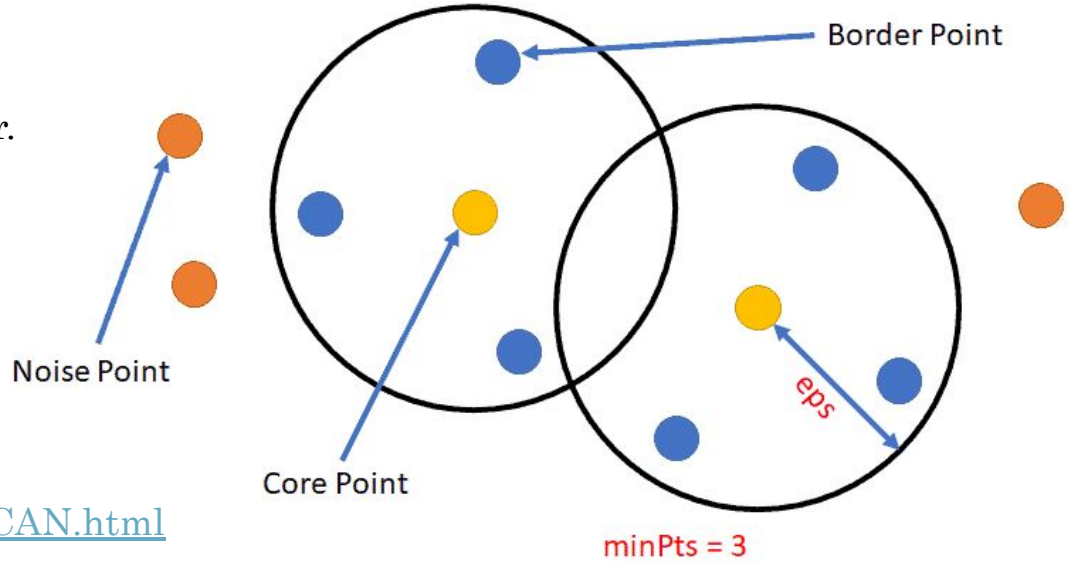
Muchos métodos de clustering se especializan en encontrar clusters con formas compactas pero entran en problemas para identificar clusters con formas arbitrarias, como por ejemplo alargadas o en forma circular. DBSCAN es un método de clustering que resuelve este problema al identificar zonas del espacio con alta densidad de puntos, independientemente de cuál sea su forma.



DBSCAN (Density-based Spatial Clustering of Applications with Noise)

Muchos métodos de clustering se especializan en encontrar clusters con formas compactas pero entran en problemas para identificar clusters con formas arbitrarias, como por ejemplo alargadas o en forma circular. DBSCAN es un método de clustering que resuelve este problema al identificar zonas del espacio con alta densidad de puntos, independientemente de cuál sea su forma.
El algoritmo consiste en establecer un círculo de un radio (**Eps**) alrededor de cada punto y contar la cantidad de puntos que entra en cada uno de estos círculos. Si esa cantidad supera un cierto valor mínimo (**MinPts**) entonces el punto será considerado semilla, en caso contrario será un punto borde (si entra dentro del radio de un punto semilla) o un punto ruido (si está fuera del radio de un punto semilla)

Pasos del Algoritmo:
0. Elegir valores para los parámetros Eps y MinPts.
1. Identificar todos los elementos como Semilla, Borde o Ruido.
2. Eliminar Ruido.
3. Unir todos los elementos Semilla que se encuentren a menos de distancia Eps.
4. Cada grupo de elementos Semilla conectados entre sí es un cluster.
5. Asignar los elementos Borde a algún cluster.

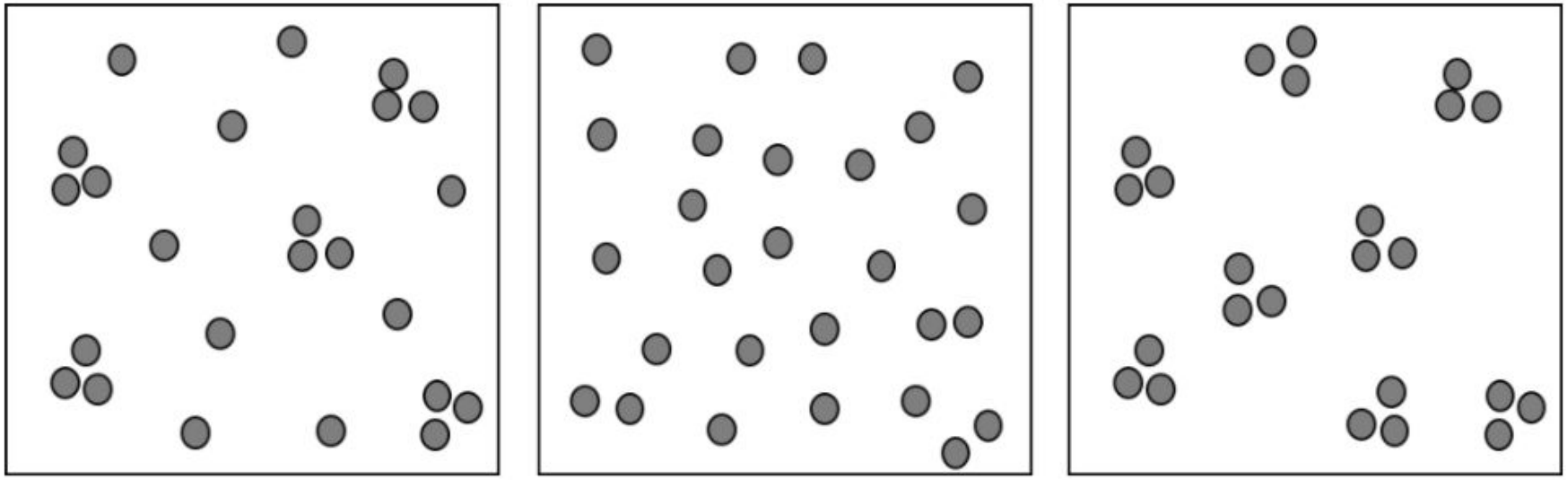


https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Validación

El problema de los métodos de clustering es que siempre encuentran clusters… Pero son buenos esos clusters?

Comparemos las siguientes 3 situaciones…



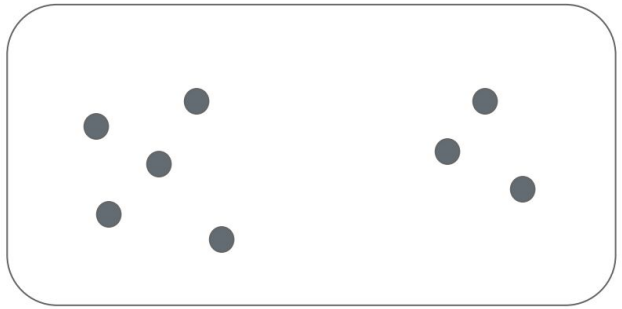
Validación

El problema de los métodos de clustering es que siempre encuentran clusters… Pero son buenos esos clusters?

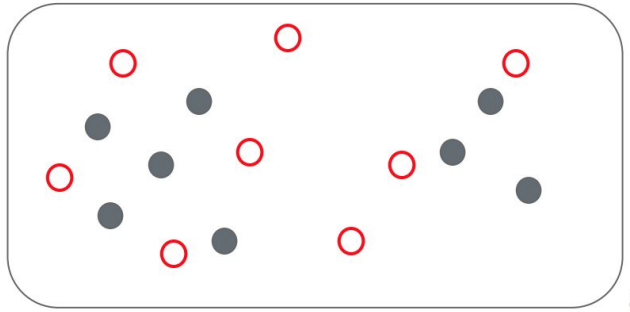
Estadístico de Hopkins
Es una medida que nos va a hablar sobre la tendencia que tienen los datos a agruparse (formar clusters). Mide si los datos que tenemos están distribuidos espacialmente de forma aleatoria (no hay agrupamientos). Si H tiene un valor cercano a 0,5 entonces la distribución de los datos será azarosa y no tendrá sentido buscar agrupamientos.

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

u_i: distancia entre un punto del dataset i elegido al azar y su vecino más cercano.
w_i: distancia entre un punto i agregado al azar y su vecino más cercano en el dataset



elementos del dataset



elementos del dataset

elementos agregados al azar

Validación

El problema de los métodos de clustering es que siempre encuentran clusters… Pero son buenos esos clusters?

Hay dos preguntas que podemos hacernos:

- ¿Cuál es el mejor método de clustering para mis datos?
- ¿Qué número de clusters representa mejor a mis datos?

Y habrá dos formas de contestar estas preguntas en función de dos posibles escenarios. Si existe algún tipo de agrupamiento externo de mis datos puedo compararlo con mi clustering (**Validación Externa**) pero en caso de no existir siempre puedo intentar entender qué tan bien definidos están mis clusters, comparando la similitud de los elementos que están dentro de cada cluster con los que están afuera (**Validación Interna**).

Veamos un ejemplo de cada tipo …

Validación Externa

Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI)

Si sabemos a qué grupo pertenece cada una de las observaciones, podemos comparar esa etiqueta con la agrupación que obtuvimos con el método de clustering. El índice Rand compara la cantidad de veces que un par de observaciones son agrupadas de la misma manera en ambas clasificaciones con el número total pares posibles.

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.
b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.
c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.
d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.
n = número total de elementos.

¿Se acuerdan de TRUE POSITIVE, TRUE NEGATIVE, …?

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

Validación Externa

Si sabemos a qué grupo pertenece cada una de las observaciones, podemos comparar esa etiqueta con la agrupación que obtuvimos con el método de clustering.

Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI)

El índice Rand compara la cantidad de veces que un par de observaciones son agrupadas de la misma manera en ambas clasificaciones con el número total pares posibles.

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.
b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.
c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.
d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.
n = número total de elementos.

Validación Externa

Si sabemos a qué grupo pertenece cada una de las observaciones, podemos comparar esa etiqueta con la agrupación que obtuvimos con el método de clustering.

Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI)

El índice Rand compara la cantidad de veces que un par de observaciones son agrupadas de la misma manera en ambas clasificaciones con el número total pares posibles.

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.
b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.
c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.
d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.
n = número total de elementos.

El índice Rand ajustado (ARI) es una modificación del índice Rand que compara su valor con el esperado por azar.

$$ARI = (R - E(R))/(max(R)-E(R))$$

E(R) = valor esperado de R si se distribuyen al azar.
max(R) = valor máximo posible de R para los datos.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

Validación Interna

En la validación interna nos interesa evaluar dos cosas:

- **Cohesión:** es una medida de las proximidad de los miembros de un clúster entre sí o con respecto al prototipo.
- **Separación:** es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

Davies-Bouldin index (DB)

Es un índice que compara la distancia de los elementos de un cluster con el centroide del mismo y la distancia que hay entre los centroides de distintos clusters

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

k: número de clústeres
σi: la distancia promedio entre cada punto en el clúster i y el centroide del clúster
σj: la distancia promedio entre cada punto del clúster j y el centroide del clúster
d(ci , cj): la distancia entre los centroides de los 2 clústeres .

Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros estan bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

Validación Interna

En la validación interna nos interesa evaluar dos cosas:

- **Cohesión:** es una medida de las proximidad de los miembros de un clúster entre sí o con respecto al prototipo.
- **Separación:** es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

Coeficiente Silhouette

1. Para cada elemento i se calcula su distancia promedio a todos los otros elemento de su clúster (a_i).
2. Para el elemento i y todos los otros clústeres que no lo contienen, se calcula las distancias promedio a todos los elementos de cada clúster. Se buscar el mínimo de esas distancias promedio a cada clúster (b_i).
3. Se calcula el coeficiente Silhouette (s_i) del elemento i.
4. Luego se puede calcular el promedio para cada cluster o el promedio global.

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

El coeficiente de Silhouette para todo el agrupamiento es:

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Validación Interna

En la validación interna nos interesa evaluar dos cosas:

- **Cohesión:** es una medida de las proximidad de los miembros de un clúster entre sí o con respecto al prototipo.
- **Separación:** es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

Coeficiente Silhouette

El coeficiente de Silhouette se suele representar en un gráfico que permite ver rápidamente si existen observaciones que podrán estar mal agrupadas (con valores negativos).