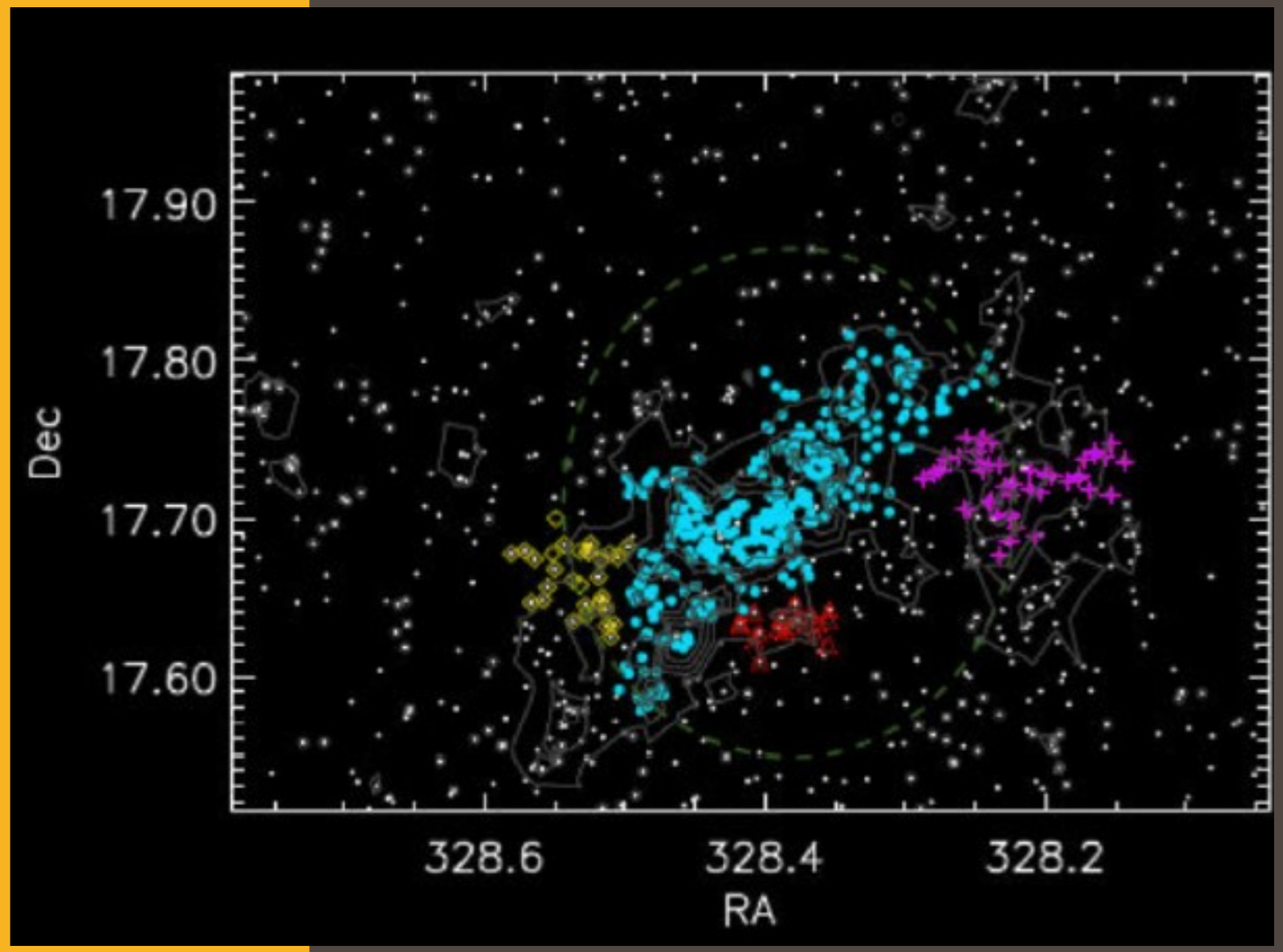
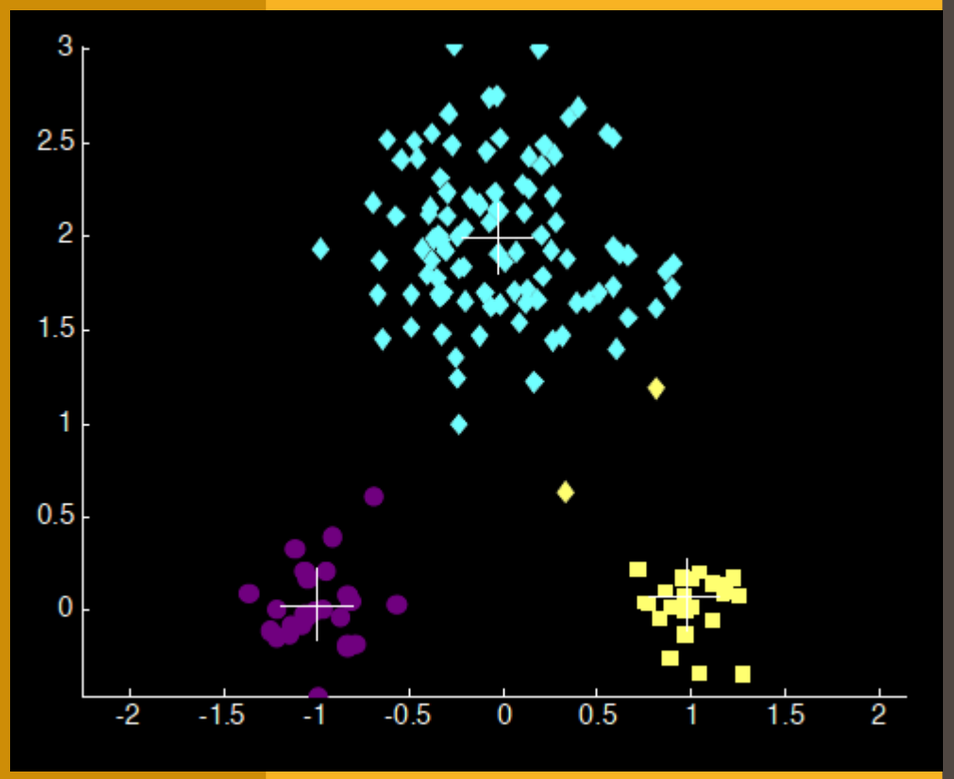
TÉCNICA DE ANÁLISIS EXPLORATORIO Y NO SUPERVISADO

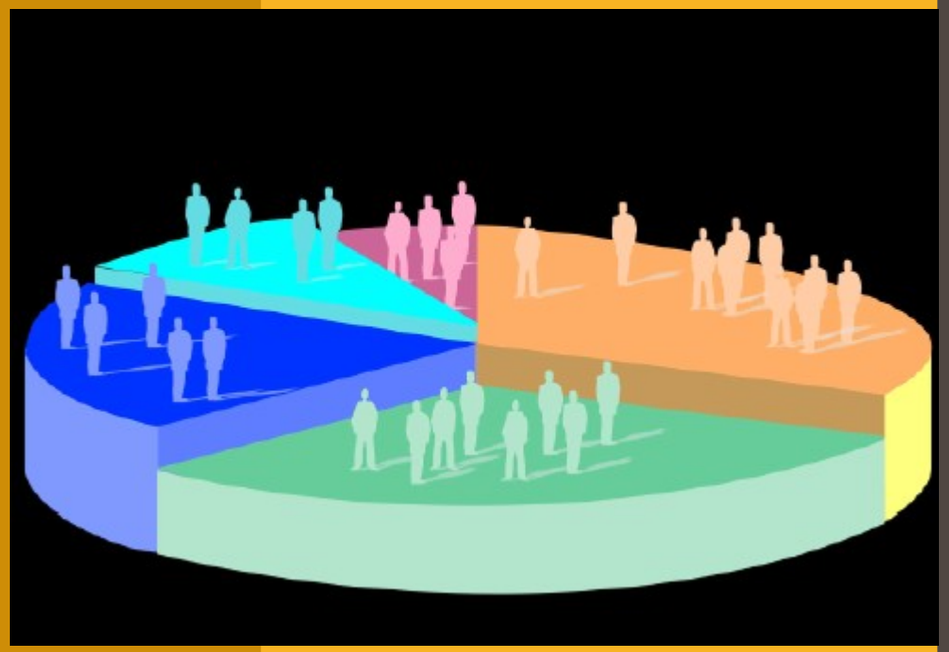


"CLUSTERING"

MUNDO NO SUPERVISADO

INTRODUCCIÓN





- El término clustering hace referencia a un amplio abanico de técnicas unsupervised cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones. Las particiones se establecen de forma que las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método unsupervised ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable). Esta característica diferencia al clustering de las técnicas estadísticas conocidas como análisis discriminante, que emplean un set de entrenamiento en el que se conoce la verdadera clasificación.

¿QUÉ SON LAS TÉCNICAS DE SEGMENTACIÓN?

Son técnicas que permiten resumir los datos. Se puede realizar un resumen global o específico de ciertas variables.

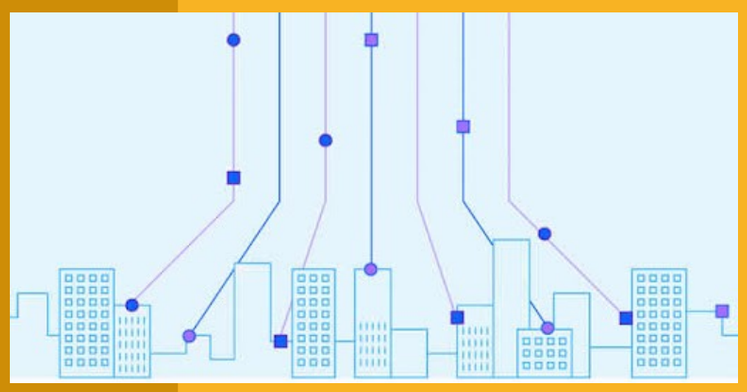
En general existen tres tipos de enfoques:

- Estimación de densidad: determina una representación compacta de la distribución de probabilidad sobre todos los datos P(X)=P(X1,X2, …,Xp).

- Búsqueda de patrones: busca una asociación descriptiva entre las variables.

- Clustering: separa las instancias en grupos de datos con características similares.

¿PARA QUE CLUSTERIZAR?



- BI - Hacer inteligencia de negocio dando uso inteligente de TODA la información disponible.
- Generar grupos o Clusters de Sujetos X idénticos o similares entre ellos (“gemelos estadísticos”), para compararlos para distintos fines.
- Caracterización.
- Minería de datos.
- Para Transitar desde el Mundo No Supervisado al Mundo Supervisado.

¿PARA QUE CLUSTERIZAR?

DEJAR QUE LOS DATOS HABLEN

ENTENDER

NO SESGAR

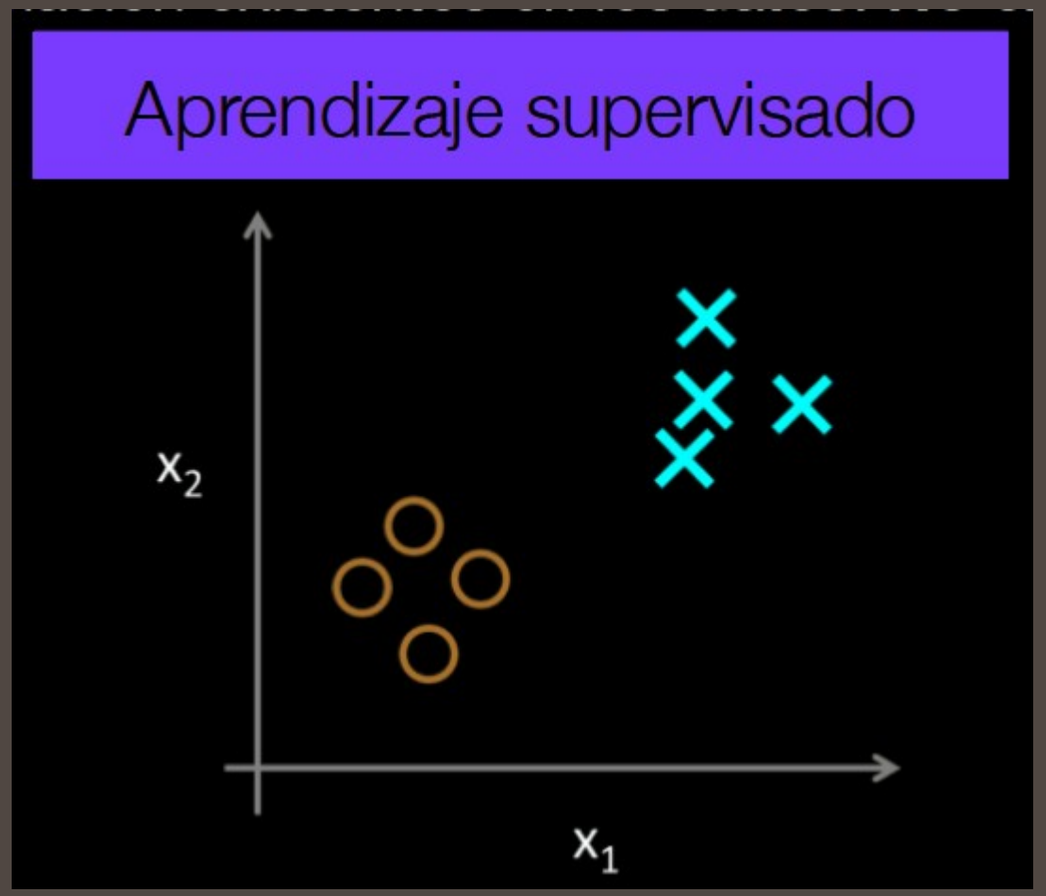
SESGOS COGNITIVOS

NO SESGAR

En el aprendizaje supervisado existe una variable objetivo a predecir.

- Si el objetivo son clases, se llama problema de clasificación o análisis discriminate.

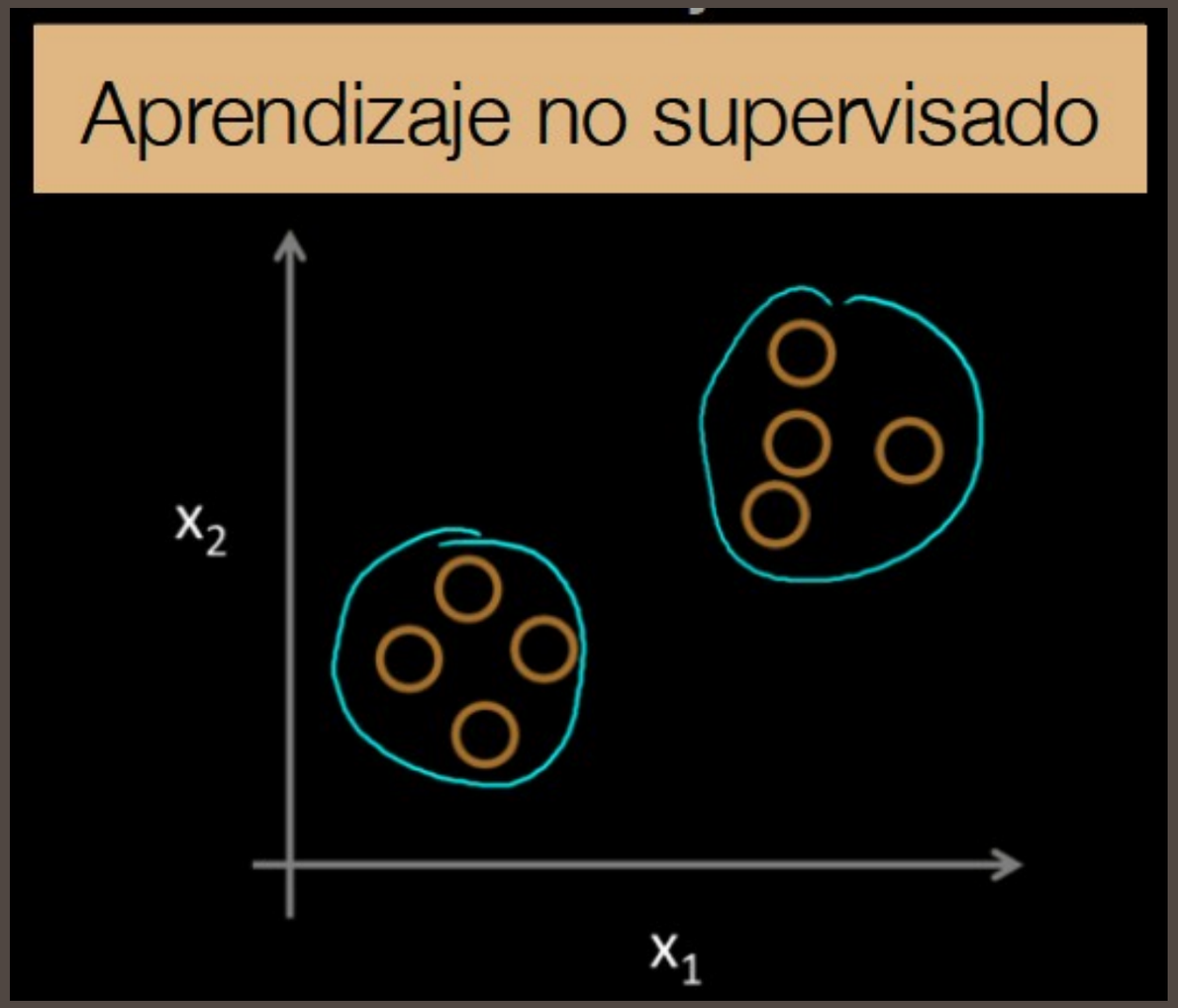
- Si los datos son continuos, se denomina problema de regresión.



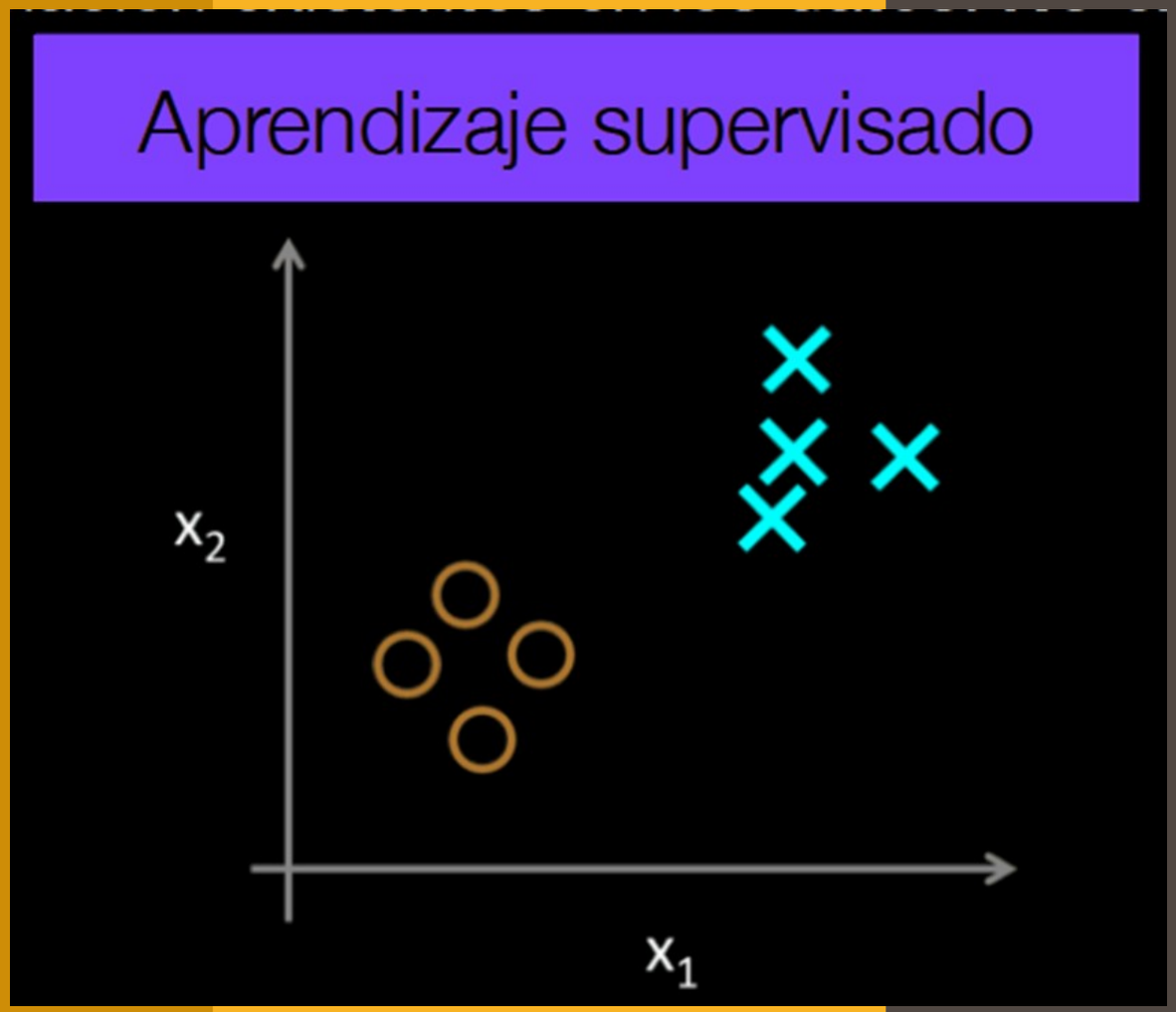
NO SESGAR

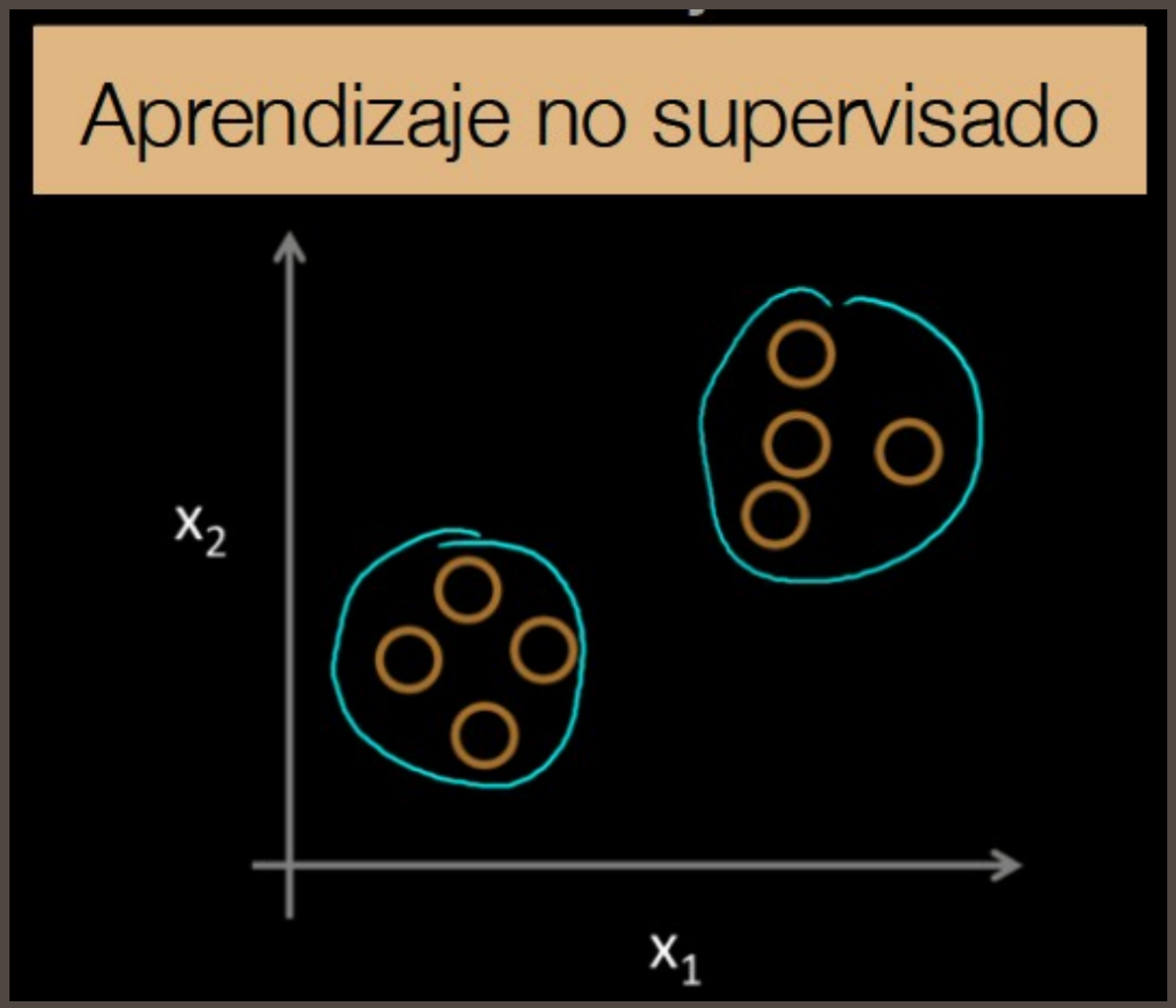
En el aprendizaje NO supervisado se busca observar, describir y también aprender la estructura o relación existentes en los datos.

No existe una variable objetivo.



NO SESGAR


Aprendizaje supervisado
x_2
x_1


Aprendizaje no supervisado
x_2
x_1

TIPOS DE PROBLEMAS

El tipo de problema o enfoque S o NS esta basado en el objetivo de la persona que analiza la tarea. Ejemplos:

• De un set de datos con etiqueta, cree un modelo capaz de predecir si un corredor de bolsa realizará algún fraude en el futuro cercano.

• Dado un set de datos sin etiqueta, agrupe a los corredores de bolsas en

grupos de personas homogéneas basado en su información demográfica

IDEA INTUITIVA

QUE BUSCAN LAS DISTINTAS TÉNICAS,MODELOS Y/O ALGORITMOSDE CLUSTERING?

MEDIDAS DE DISTANCIA

Todos los métodos de clustering tienen una denominador común; para poder llevar a cabo las agrupaciones necesitan definir “una Medida de Distancia”, y luego cuantificarla, con el objetivo de cuantificar la similitud o diferencia entre las observaciones (puntos en el plano cartesiano).

MEDIDAS DE DISTANCIA

El término distancia se emplea entonces dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión.

¿Podemos Utilizar Variables Dummy?

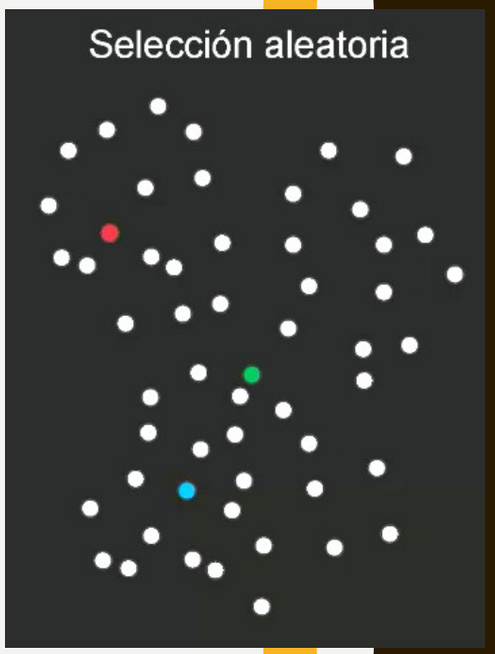
OBJETIVO DE TODO MODELO DE Clustering

MEDIDAS DE DISTANCIA

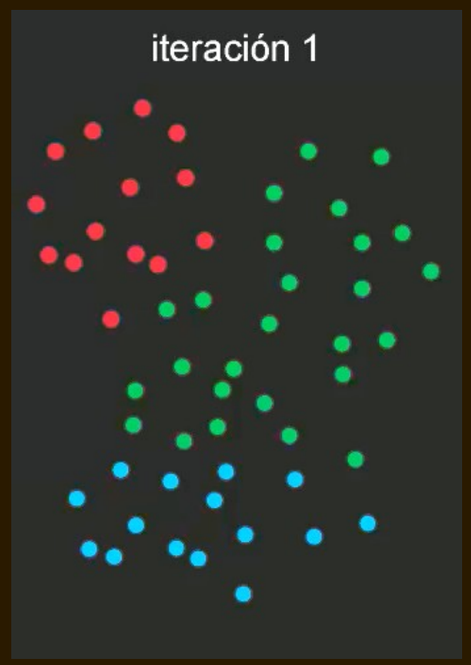
El objetivo del algoritmo es minimizar la distancia de los puntos dentro de cada cluster y maximizar la distancia entre Clusters.

Gráficamente

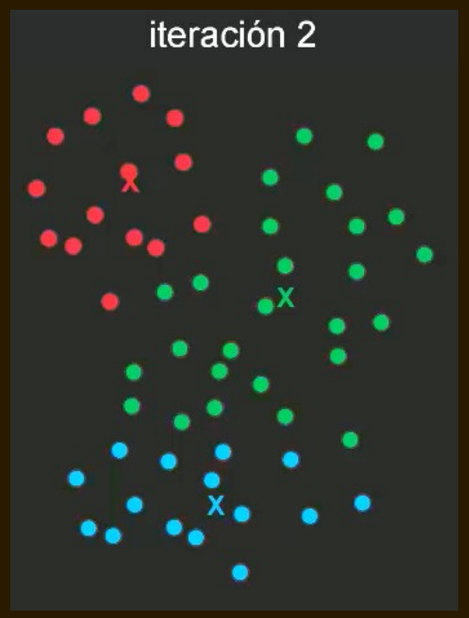
1. Si mi óptimo de Cluster (K) fue 3, el algoritmo seleccionará de manera aleatoria 3 observaciones.



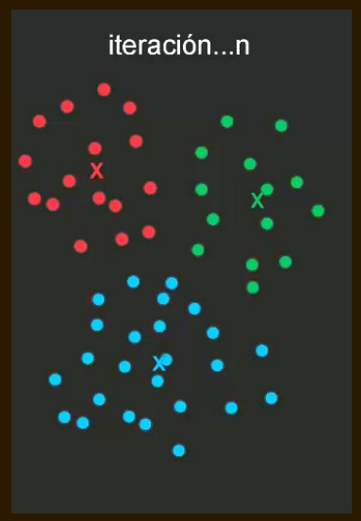
2. Luego, en una primera iteración, tomará las siguientes observaciones y las asignará al punto más cercano de las 3 observaciones anteriores (proto centroides).



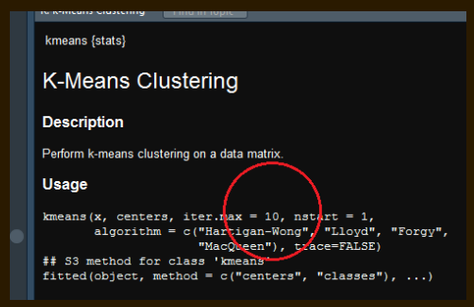
3. En una segunda iteración vuelve a calculará los centroides definitivos.



4. En una tercera iteración, vuelve a asignar las observaciones a los centroides definidos en el paso 3.



5. El algoritmo en Rstudio itera 10 veces.


kmeans {stats}

K-Means Clustering

Description

Perform k-means clustering on a data matrix.

Usage

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"), trace=FALSE)
## S3 method for class 'kmeans'
fitted(object, method = c("centers", "classes"), ...)
```

ALGORITMOS

Dada la popularidad del clustering en disciplinas muy distintas (genómica, marketing...), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos. Pueden diferenciarse tres grupos:

- Partitioning Clustering: Este tipo de algoritmos requieren que el usuario especifique de antemano el número de clusters que se van a crear (K-means, K-medoids, CLARA).

- Hierarchical Clustering (Jerárquico): Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clusters. (agglomerative clustering, divisive clusterig).

- Métodos que combinan o modifican los anteriores (hierarchical K-means, fuzzy clustering, model based clustering y density based clustering).

CONSIDERACIONES PARA LA CORRELACION PARA

- Mundo Supervisado. 2 Caminos o 2 Objetivos:

  Predecir o

  Explicar

- Mundo No Supervisado. 2 Caminos o 2 Objetivos:

  Es Relevante que Influyan.

  No es Relevante que Influyan.

CONSIDERACIONES DE LA CORRELACION PARA...

MUNDO O UNIVERSO	CONSIDERACION I	CONSIDERACION II	¿ES RELEVANTE LA CORRELACIÓN?
SUPERVISADO	Explicar. Causalidad	¿Es Relevante que Influyan las Variables al objetivo del Análisis?	SI
SUPERVISADO	Predecir. Comprobar. Determinar	¿No es Relevante que Influyan las Variables al objetivo del Análisis: Muy Bien Justificado?	NO
NO SUPERVISADO	Conocer. Apreciar. No Sesgar	¿Es Relevante que Influyan las Variables al objetivo del Análisis?	SI
NO SUPERVISADO	Conocer. Apreciar. No Sesgar.	¿No es Relevante que Influyan las Variables al objetivo del Análisis: Muy Bien Justificado?	NO

TODO DEPENDERÁ DE LOS OBJETIVOS DEL ESTUDIO DE MINERÍA DE DATOS VERSUS LOS OBJETIVOS DE NEGOCIO

VENTAJAS Y DESVENTAJAS DE K-MEANS

K-means es uno de los métodos de clustering más utilizados. Destaca por la sencillez y velocidad de su algoritmo, sin embargo, presenta una serie de limitaciones que se deben tener en cuenta:

- Requiere que se indique de antemano el número de clusters que se van a crear. Esto puede ser complicado si no se dispone de información adicional sobre los datos con los que se trabaja o de una técnica para determinar el mejor número de K. Una posible solución es aplicar el algoritmo para un rango de valores k y evaluar con cual se consiguen mejores resultados, por ejemplo, menor suma total de varianza interna.

- Presenta problemas de robustez frente a outliers. La única solución es excluirlos o recurrir a otros métodos de clustering más robustos como K-medoids (PAM).

- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. Para minimizar este problema se recomienda repetir el proceso de clustering entre 20 - 50 veces (dependiendo del tamaño de la matriz) y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna. Aun así, no se garantiza que para un mismo set de datos los resultados sean exactamente iguales.

VENTAJAS Y DESVENTAJAS DE K-MEDOIS (PAM)

K-medoids es un método de clustering muy similar a K-means en cuanto a que ambos agrupan las observaciones en K clusters, donde K es un valor preestablecido por el analista. La diferencia es que en K-medoids cada cluster está representado por una observación presente en el cluster (medoid), mientras que en K-means cada cluster está representado por su centroide que se corresponde con el promedio de todas las observaciones del cluster pero con ninguna en particular.

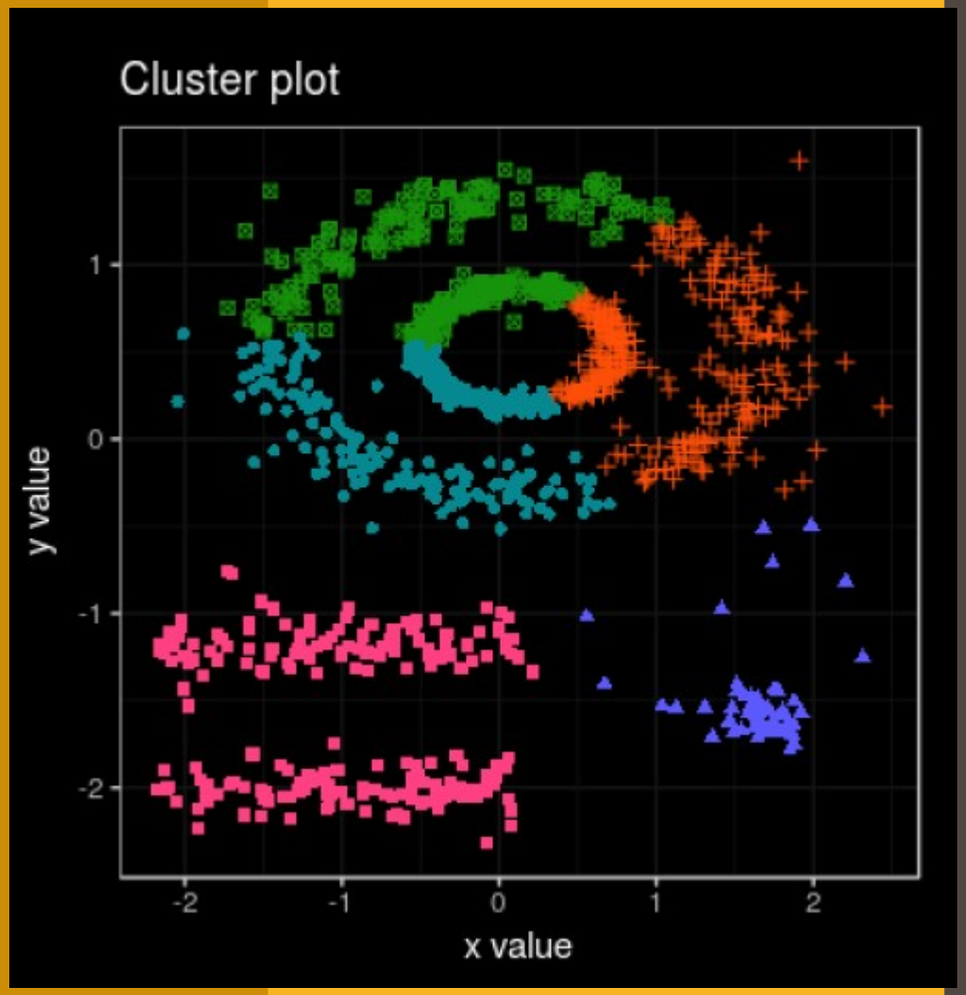
Una definición más exacta del término medoid es: elemento dentro de un cluster cuya distancia (diferencia) promedio entre él y todos los demás elementos del mismo cluster es lo menor posible. Se corresponde con el elemento más central del cluster y por lo tanto puede considerarse como el más representativo. El hecho de utilizar medoids en lugar de centroides hace de K-medoids un método más robusto que K-means, viéndose menos afectado por outliers o ruido. A modo de idea intuitiva puede considerarse como la analogía entre media y diferencia entre el valor mínimo y máximo.

VENTAJAS Y DESVENTAJAS DE K-MEDOIS (PAM)

- Al igual que K-means, necesita que se especifique de antemano el número de clusters que se van a crear. Esto puede ser complicado de determinar si no se dispone de información adicional sobre los datos o de una técnica para determinar el mejor número de K.

- Para sets de datos grandes necesita muchos recursos computacionales.

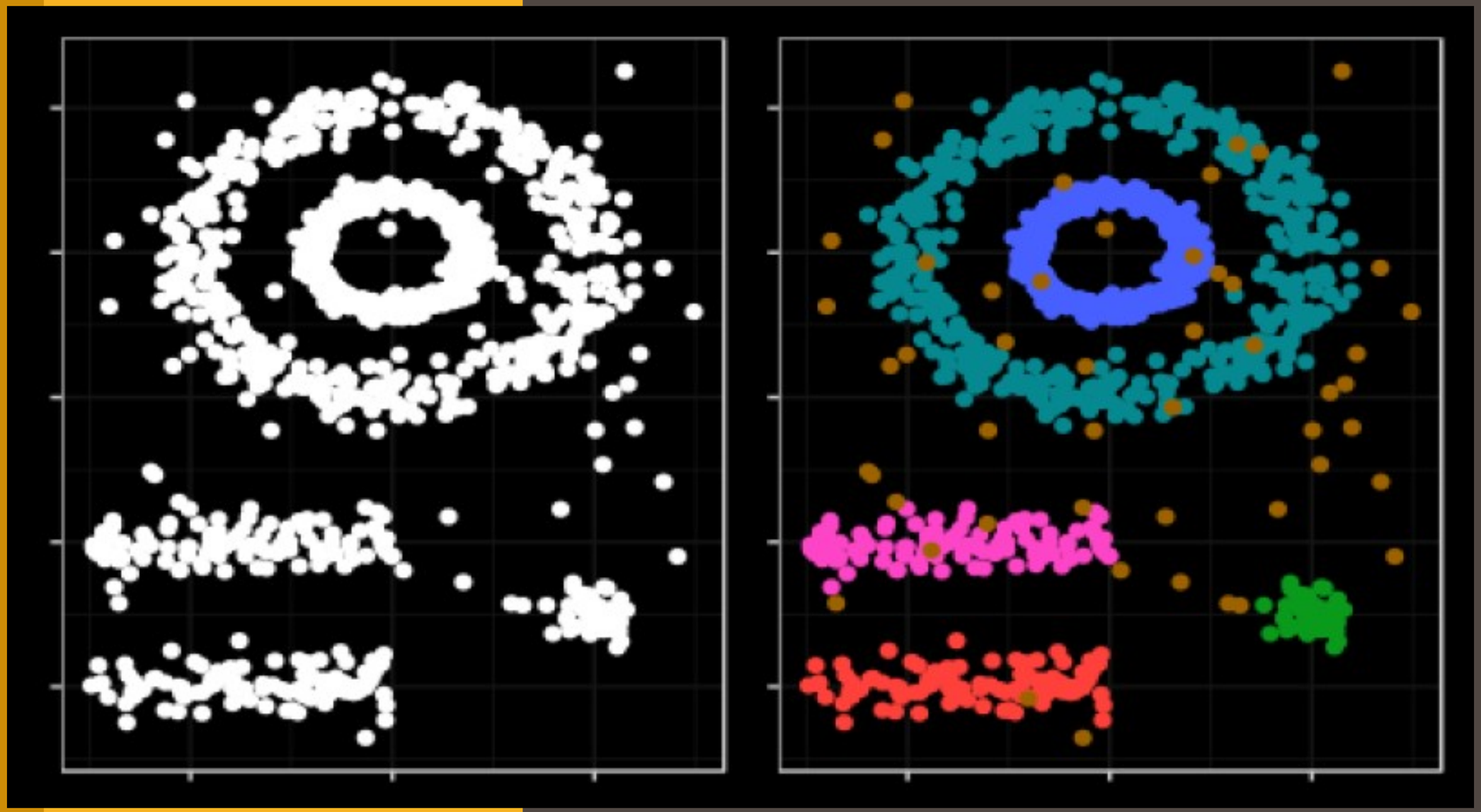
DENSITY BASED CLUSTERING (DBSCAN)



Density-based spatial clustering of applications with noise (DBSCAN) fue presentado en 1996 por Ester et al. como una forma de identificar clusters siguiendo el modo intuitivo en el que lo hace el cerebro humano, identificando regiones con alta densidad de observaciones separadas por regiones de baja densidad.

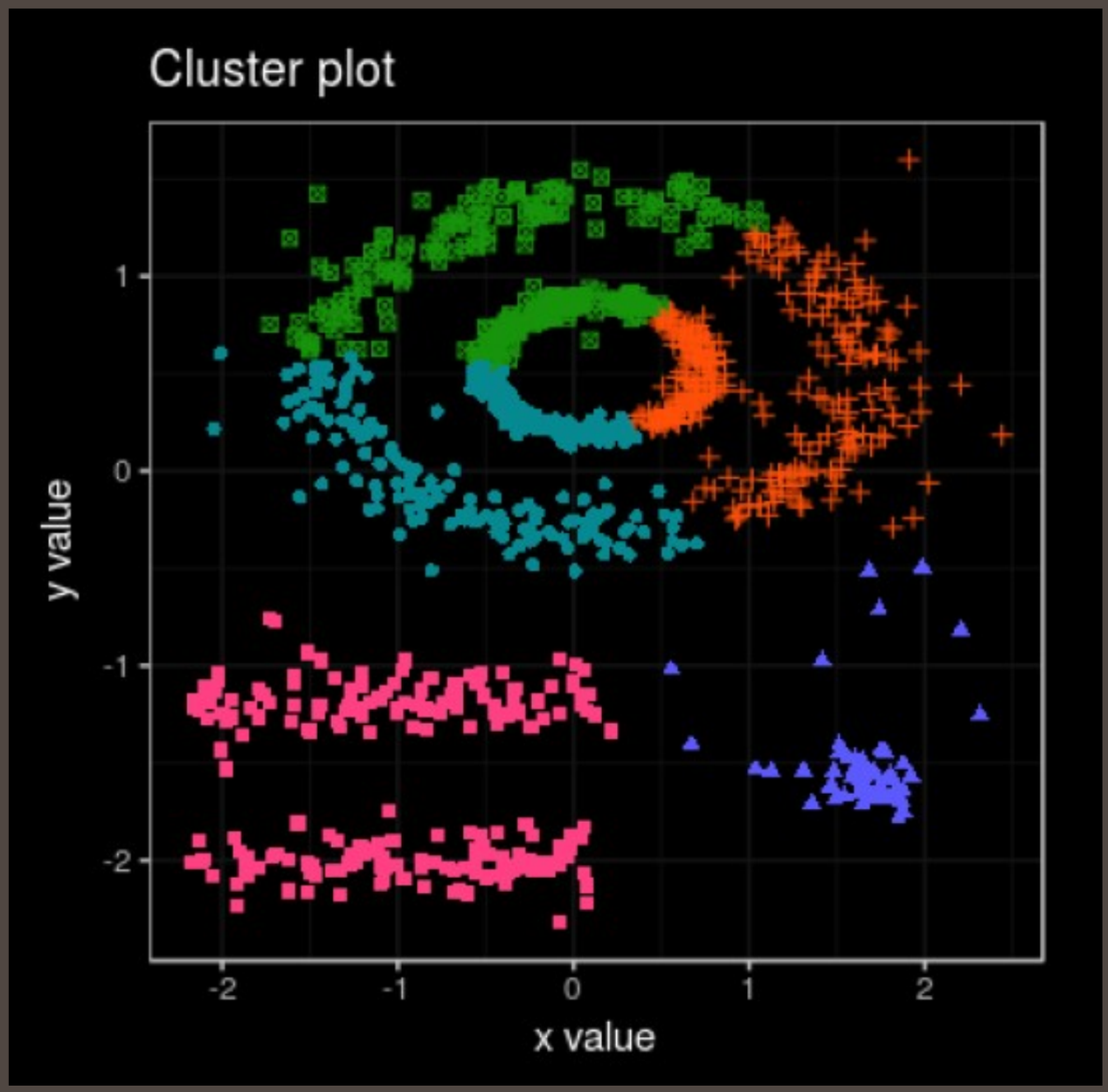
DENSITY BASED CLUSTERING (DBSCAN)

El cerebro humano identifica fácilmente 5 agrupaciones y algunas observaciones aisladas (ruido).



DENSITY BASED CLUSTERING (DBSCAN)

Véanse ahora los clusters que se obtienen si se aplica, por ejemplo, K-means Clustering.



VENTAJAS Y DESVENTAJAS DE BASED CLUSTERING (DBSCAN)

Los clusters generados distan mucho de representar las verdaderas agrupaciones. Esto es así porque los métodos de partitioining clustering como k-means, hierarchical, k-medoids, c-means... son buenos encontrando agrupaciones con forma esférica o convexa que no contengan un exceso de outliers o ruido, pero fallan al tratar de identificar formas arbitrarias.

De ahí que el único cluster que se corresponde con un grupo real sea el azul con el verde.

DBSCAN evita este problema siguiendo la idea de que, para que una observación forme parte de un cluster, tiene que haber un mínimo de observaciones vecinas dentro de un radio de proximidad y de que los clusters están separados por regiones vacías o con pocas observaciones.

VENTAJAS Y DESVENTAJAS DE BASED CLUSTERING (DBSCAN)

La siguiente imagen muestra las conexiones existentes entre un conjunto de observaciones si se emplea $minPts$=4. La observación A y el resto de observaciones marcadas en rojo son core points, ya que todas ellas contienen al menos 4 observaciones vecinas (incluyéndose a ellas mismas) en su ϵ-neighborhood. Como todas son alcanzables entre ellas, forman un cluster. Las observaciones B y C no son core points pero son alcanzables desde A a través de otros core points, por lo tanto, pertenecen al mismo cluster que A. La observación N no es ni un core point ni es directamente alcanzable, por lo que se considera como ruido.

VENTAJAS Y DESVENTAJAS DE BASED CLUSTERING (DBSCAN)

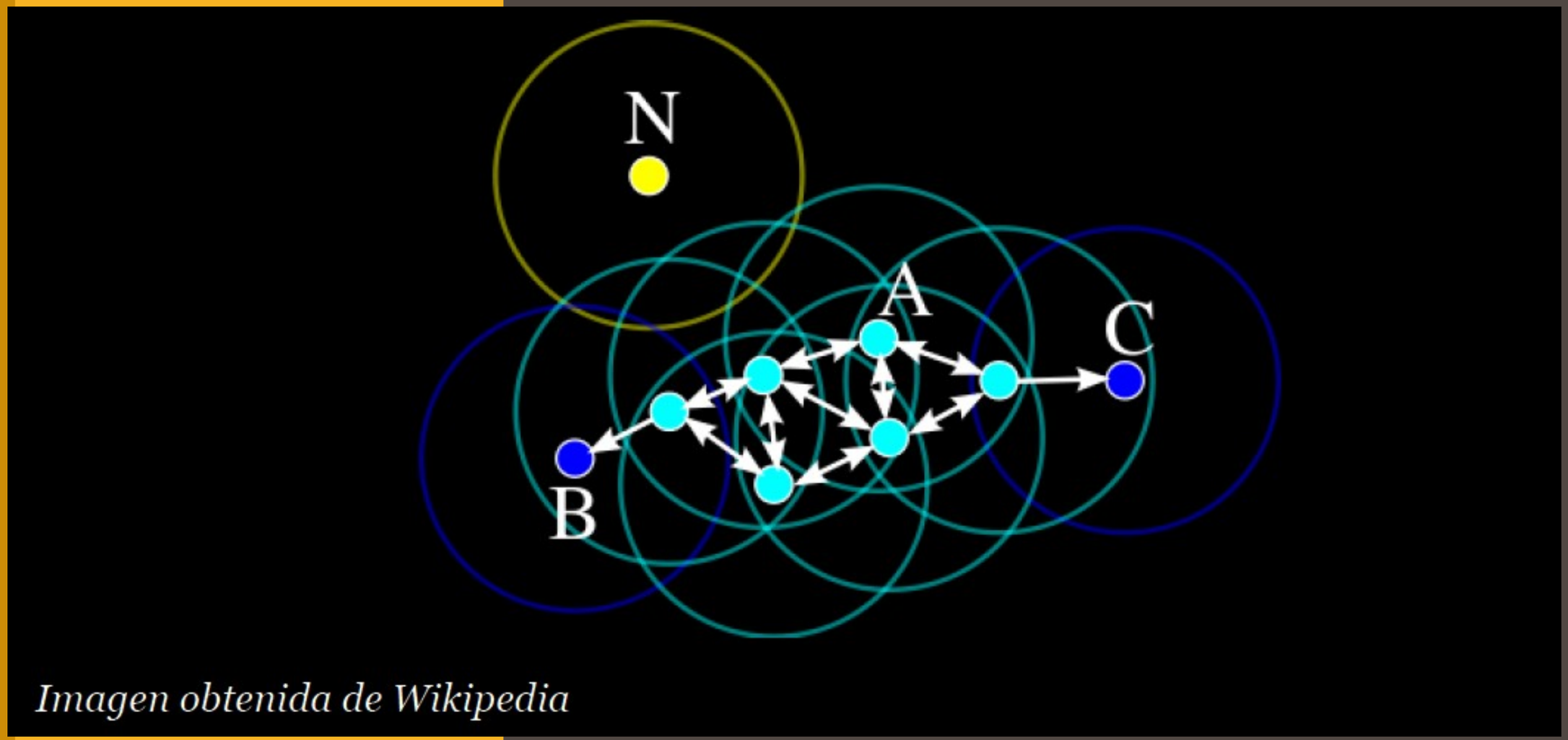


Imagen obtenida de Wikipedia

VENTAJAS Y DESVENTAJAS DE BASED CLUSTERING (DBSCAN)

Ventajas de DBSCAN

- No requiere que el usuario especifique el número de clusters.
- Es independiente de la forma que tengan los clusters. No tienen por qué ser circulares.
- Puede identificar outliers, por lo que los clusters generados no se influenciados por ellos.

Desventajas de DBSCAN

- No es un método totalmente determinístico no supervisado: los border points que son alcanzables desde más de un cluster pueden asignarse a uno u otro dependiendo del orden en el que se procesen los datos.
- No genera buenos resultados cuando la densidad de los grupos es muy distinta, ya que no es posible encontrar los parámetros ϵ y minPts que sirvan para todos a la vez.

TÉCNICA DE ANÁLISIS EXPLORATORIO Y NO SUPERVISADO "CLUSTERING“

Desarrollado por Adrián Armando Araneda Toro
Y Alex Sebastián Meléndez Suazo
Julio 2022
Versión 0.2

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

- Todos los métodos de clustering tienen una elemento en común, para poder llevar a cabo las agrupaciones necesitan definir “una Medida de Distancia”, y luego cuantificarla, con el objetivo de cuantificar la similitud o diferencia entre las observaciones (puntos en el plano cartesiano).

- El término distancia se emplea entonces dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión.

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

A. **Distancia Euclidiana:** La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y), la distancia euclídea entre p y q viene dada por la ecuación:

$$d_{euc}(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

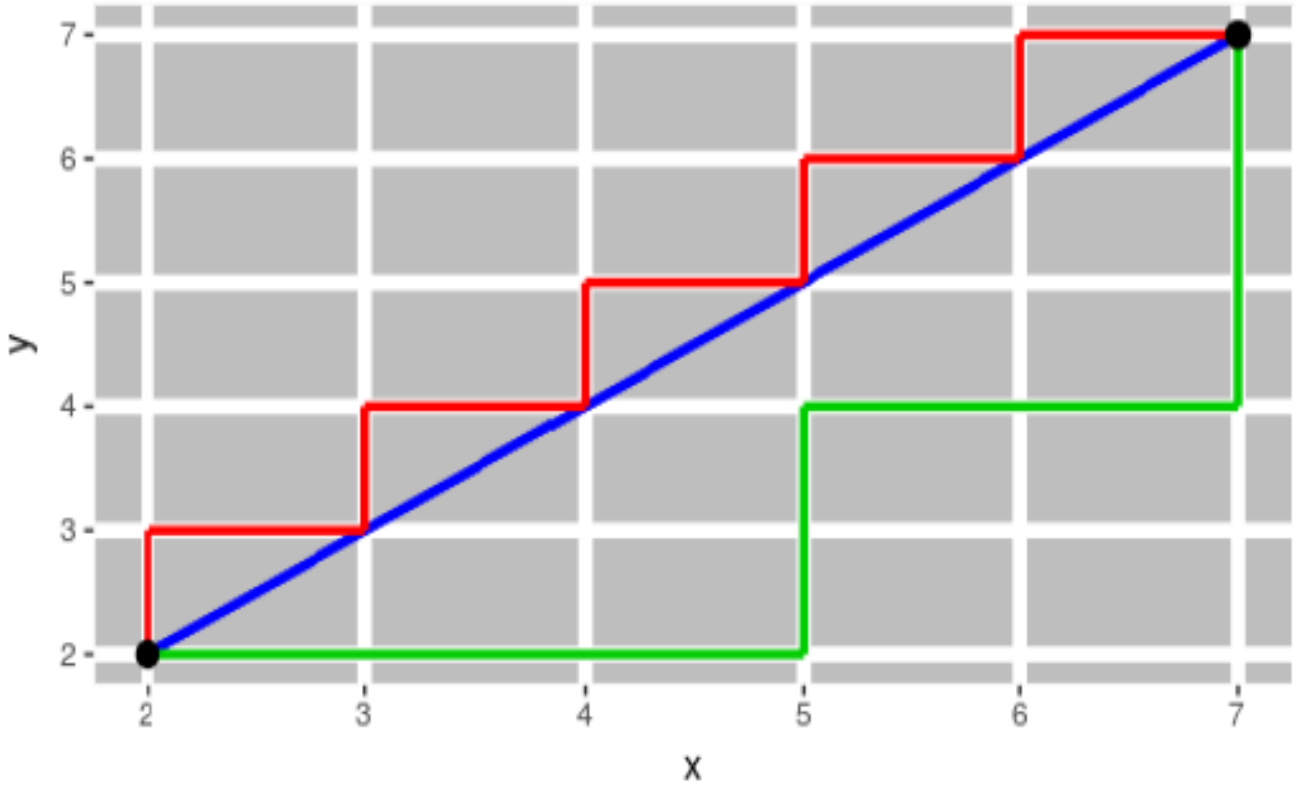
B. **Distancia de Manhattan:** La distancia de Manhattan, también conocida como taxicab metric, rectilinear distance o L1 distance, define la distancia entre dos puntos p y q como la sumatoria de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por outliers (es más robusta) que la distancia euclídea debido **a que no eleva al cuadrado las diferencias.**

$$d_{man}(p,q) = \sum_{i=1}^{n} |(p_i - q_i)|$$

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

La siguiente imagen muestra una comparación entre la distancia euclídea (segmento azul) y la distancia de manhattan (segmento rojo y verde) en un espacio bidimensional. Existen múltiples caminos para unir dos puntos con el mismo valor de distancia de manhattan, ya que su valor es igual al desplazamiento total en cada una de las dimensiones.



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Una vez seleccionado el número adecuado de clusters y aplicado el algoritmo de clustering pertinente se tiene que evaluar la calidad de los de los mismos, de lo contrario, podrían derivarse conclusiones de agrupación que no se corresponden con la realidad. Pueden diferenciarse tres tipos de estadísticos empleados con este fin:

• Validación interna de los clusters: Emplean únicamente información interna del proceso de clustering para evaluar la bondad de las agrupaciones generadas. Se trata de un proceso totalmente unsupervised ya que no se incluye ningún tipo de información que no estuviese ya incluida en el clustering.

• Validación externa de los clusters (ground truth): Combinan los resultados del clustering (unsupervised) con información externa (supervised), como puede ser un set de validación en el que se conoce el verdadero grupo al que pertenece cada observación. Permiten evaluar hasta qué punto el clustering es capaz de agrupar correctamente las observaciones. Se emplea principalmente para seleccionar el algoritmo de clustering más adecuado, aunque su uso está limitado a escenarios en los que se dispone de un set de datos de validación.

• Significancia de los clusters: Calculan la probabilidad (p-value) de que los clusters generados se deban únicamente al azar.

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

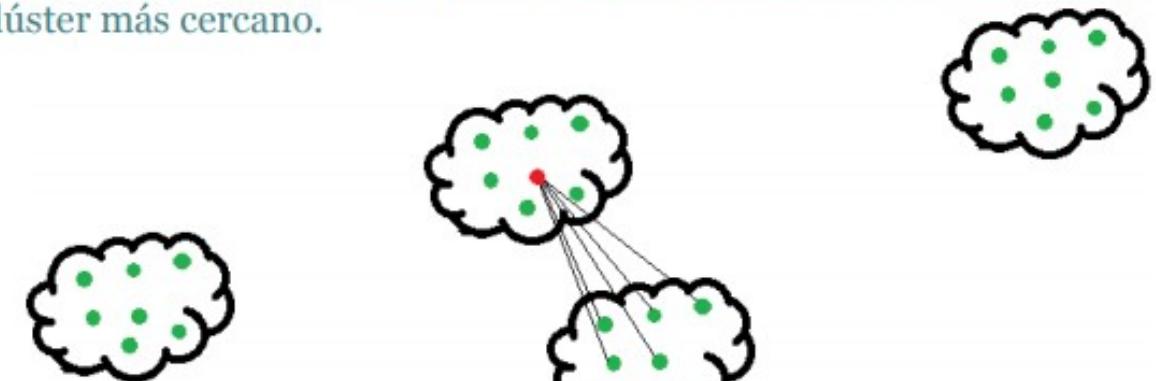
La idea principal detrás del clustering es agrupar las observaciones de forma que sean similares a aquellas que están dentro de un mismo cluster y distintas a las de otros clusters, es decir, que la homogeneidad (también llamada compactness o cohesion) se lo mayor posible a la vez que lo es la separación entre clusters.

• Cohesión: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.

• Separación: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

- Cohesión a(x) : distancia promedio de x a todos los demás puntos en el mismo clúster.

- Separación b(x) : distancia promedio de x a todos los demás puntos en el clúster más cercano.



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

Sum of Squared Within (SSW)

Suma de cuadrados para la distancia intra Clusters.

Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster C_i y m_i el centroide del clúster C_i.

Sum of Squared Between (SSB)

Suma de cuadrados para la distancia entre Clusters.

Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación)

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

Siendo k el número de clústeres, n_j el número de elementos en el clúster j, c_j el centroide del clúster j y $\bar{x}$ es la media del data set.

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

- Cohesión se mide como **within cluster sum of squares (SSE)**

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

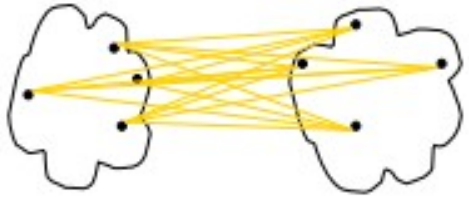
- Separación se mide como **between cluster sum of squares (BSS)**

$$BSS = \sum_i |C_i|(m - m_i)^2$$

Medidas internas: Cohesión y separación

- Enfoque basado en grafos de proximidad
- Cohesión: suma de los pesos de todos los arcos en un cluster
- Separación: suma de los pesos entre nodos del cluster y de otros clusters

cohesion



separation

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Silhouette width

El coeficiente de silueta contrasta la distancia media a elementos en el mismo grupo con la distancia media a elementos en otros grupos. Los objetos con un valor de silueta alto están considerados bien agrupados, los objetos con un valor bajo pueden ser ruido o anomalías. Estos índices trabajan bien con k-means, y es también utilizado para determinar el número óptimo de grupos:

El valor de la silueta es una medida de cuán similar es un objeto a su propia agrupación (cohesión) en comparación con otras agrupaciones (separación). La silueta va de -1 a +1, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con las agrupaciones vecinas. Si la mayoría de los objetos tienen un valor alto, entonces la composición de la agrupación es apropiada.

• Calcula la media de las distancias (llámese ai) entre la observación i y el resto de observaciones que pertenecen al mismo cluster. Cuanto menor sea ai mayor la similitud que tiene con el resto de observaciones de su cluster.

• Calcula la distancia promedio entre la observación i y el resto de clusters. Entendiendo por distancia promedio entre i y un determinado cluster C como la media de las distancias entre i y las observaciones del cluster C.

• Identifica como bi a la menor de las distancias promedio entre i y el resto de clusters, es decir, la distancia al cluster más próximo (neighbouring cluster).

• Calcula el valor de silhouette como:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

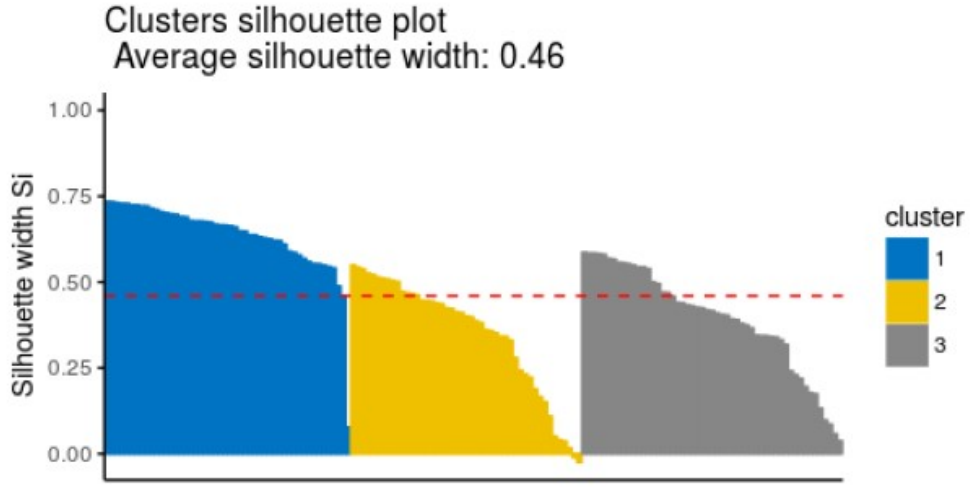
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Su valor puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al cluster correcto. Cuando su valor es próximo a cero significa que la observación se encuentra en un punto intermedio entre dos clusters. Valores negativos apuntan a una posible asignación incorrecta de la observación. Se trata por lo tanto de un método que permite evaluar el resultado del clustering a múltiples niveles:

- La calidad de asignación de cada observación por separado. Permitiendo identificar potenciales asignaciones erróneas (valores negativos de silhouette).
- La calidad de cada cluster a partir del promedio de los índices silhouette de todas las observaciones que lo forman. Si por ejemplo se han introducido demasiados clusters, es muy probable que algunos de ellos tengan un valor promedio mucho menor que el resto.
- La calidad de la estructura de clusters en su conjunto a partir del promedio de todos los índices silhouette.

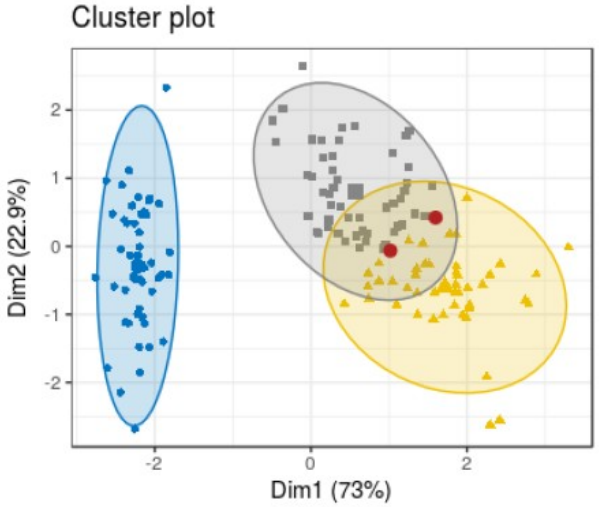
```
##   cluster size ave.sil.width
## 1       1   50          0.64
## 2       2   47          0.35
## 3       3   53          0.39
```

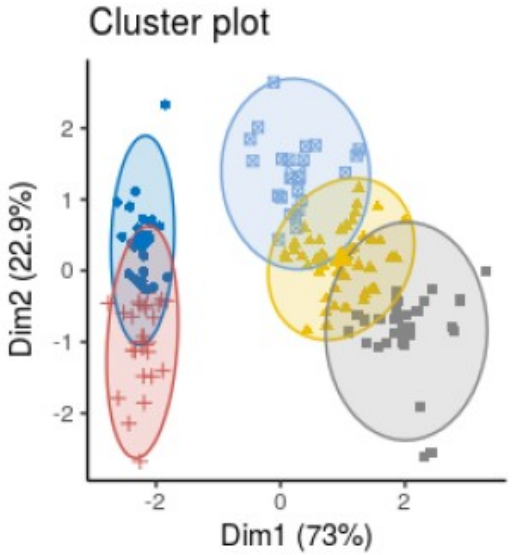


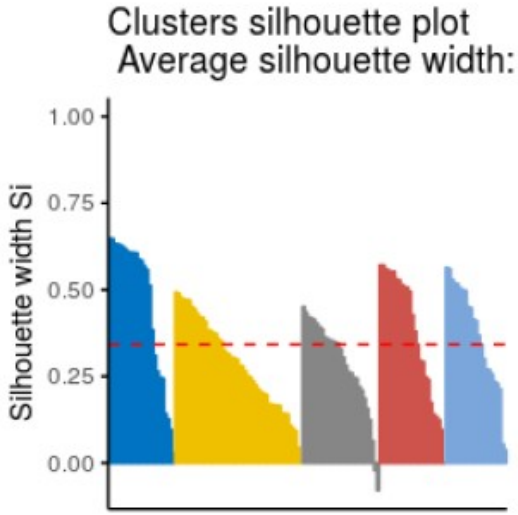
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

El cluster número 2 (amarillo) tiene observaciones con valores de silhouette próximos a 0 e incluso negativos, lo que indica que esas observaciones podrían estar mal clasificadas. Viendo la representación gráfica del clustering, cabe esperar que sean observaciones que están situadas en la frontera entre los clusters 2 y 3 ya que solapan.







Mundo No Supervisado
b.¿Qué Buscan?

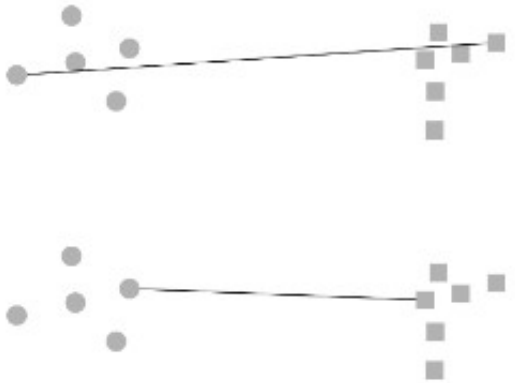
Calidad de los clusters

Índice Dunn

El objetivo de este índice es identificar un conjunto de clústeres que sean compactos, con una varianza pequeña entre los miembros del clúster, y que éstos estén bien separados de los miembros de otros clústeres. Un valor más alto del índice de Dunn indica un mejor rendimiento del algoritmo de clustering.

El índice de Dunn tiene un valor entre cero e infinito, y debe ser lo más alto posible. Por lo tanto, la distancia entre los miembros de un clúster debe ser los más baja posible, y la distancia entre los clústeres lo más alta posible.

Por ejemplo, en el caso de la distancia entre clústeres puede utilizarse la distancia más corta entre dos puntos de diferentes clústeres, o la distancia más larga, o la distancia entre los centroides. También pueden utilizarse diferentes indicadores de la distancia dentro de un clúster.



$$D = \frac{separacion\ minima\ interclusters}{separacion\ maxima\ intracluster}$$

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad

Las medidas de estabilidad son un tipo particular de validación interna que cuantifican el grado en que varían los resultados de un clustering como consecuencia de eliminar, de forma iterativa, una columna del set de datos. Todas ellas son relativamente costosas desde el punto de vista computacional ya que requieren repetir el clustering tantas veces como columnas tenga el set de datos. Dentro de esta familia de medidas se encuentran:

- Average proportion of non-overlap (APN): mide la proporción media de observaciones que no se asignan al mismo cluster cuando se elimina una columna del set de datos en comparación a cuando se incluyen todas.

- Average distance (AD): mide la media de las distancias promedio intra-cluster empleando todos los datos y eliminando una columna a la vez.

- Average distance between means (ADM): mide la media de las distancias entre centroides empleando todos los datos y eliminando una columna a la vez.

- Figure of merit (FOM): mide la media de la varianza intra-cluster de la columna eliminada, empleando la estructura del clustering, calcula con las columnas no eliminadas.

Los valores de APN, ADM, y FOM pueden ir desde 0 a 1, siendo valores pequeños un indicativo de alta estabilidad. En el caso de AD ocurre lo mismo pero sus valores pueden ir de 0 hasta infinito.

TÉCNICA DE ANÁLISIS EXPLORATORIO Y NO SUPERVISADO "CLUSTERING“

Desarrollado por Adrián Armando Araneda Toro
Y Alex Sebastián Meléndez Suazo
Julio 2022
Versión 0.2

MUNDO NO SUPERVISADO

El término clustering hace referencia a un amplio abanico de técnicas unsupervised cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones.

Las particiones se establecen de forma que las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método unsupervised ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable).

Esta característica diferencia al clustering de las técnicas estadísticas conocidas como análisis discriminante, que emplean un set de entrenamiento en el que se conoce la verdadera clasificación.

1 – ¿Para que Clusterizar?

Mundo No Supervisado
a.¿Para que Clusterizar?

Teoría

Dada la popularidad del clustering en disciplinas muy distintas (genómica, marketing…), se han desarrollado multitud de variantes y adaptaciones de sus métodos y algoritmos. Pueden diferenciarse tres grupos:

- Partitioning Clustering: Este tipo de algoritmos requieren que el usuario especifique de antemano el número de clusters que se van a crear (K-means, K-medoids, CLARA).
- Hierarchical Clustering: Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de clusters. (agglomerative clustering, divisive clusterig).
- Métodos que combinan o modifican los anteriores (hierarchical K-means, fuzzy clustering, model based clustering y density based clustering).
- BI - Hacer inteligencia de negocio dando uso inteligente de TODA la información disponible.
- Generar grupos o Clusters de Sujetos X idénticos o similares entre ellos (“gemelos estadísticos”), para compararlos para distintos fines.
- Caracterización.
- Minería de datos.
- Para Transitar desde el Mundo No Supervisado al Mundo Supervisado.

Mundo No Supervisado
a.¿Para que Clusterizar?

- DEJAR QUE LOS DATOS HABLEN
- ENTENDER
- NO SESGAR
- IDEA INTITUIVA

2– ¿Qué buscan las distintas Técnicas, Modelos y/o Algoritmos de Clustering?

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

- Todos los métodos de clustering tienen una cosa en común, para poder llevar a cabo las agrupaciones necesitan definir “una Medida de Distancia”, y luego cuantificarla, con el objetivo de cuantificar la similitud o diferencia entre las observaciones (puntos en el plano cartesiano).

- El término distancia se emplea entonces dentro del contexto del clustering como cuantificación de la similitud o diferencia entre observaciones. Si se representan las observaciones en un espacio p dimensional, siendo p el número de variables asociadas a cada observación, cuando más se asemejen dos observaciones más próximas estarán, de ahí que se emplee el término distancia. La característica que hace del clustering un método adaptable a escenarios muy diversos es que puede emplear cualquier tipo de distancia, lo que permite al investigador escoger la más adecuada para el estudio en cuestión.

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

A. **Distancia Euclidiana:** La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras. Por ejemplo, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x,y), la distancia euclídea entre p y q viene dada por la ecuación:

$$d_{euc}(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

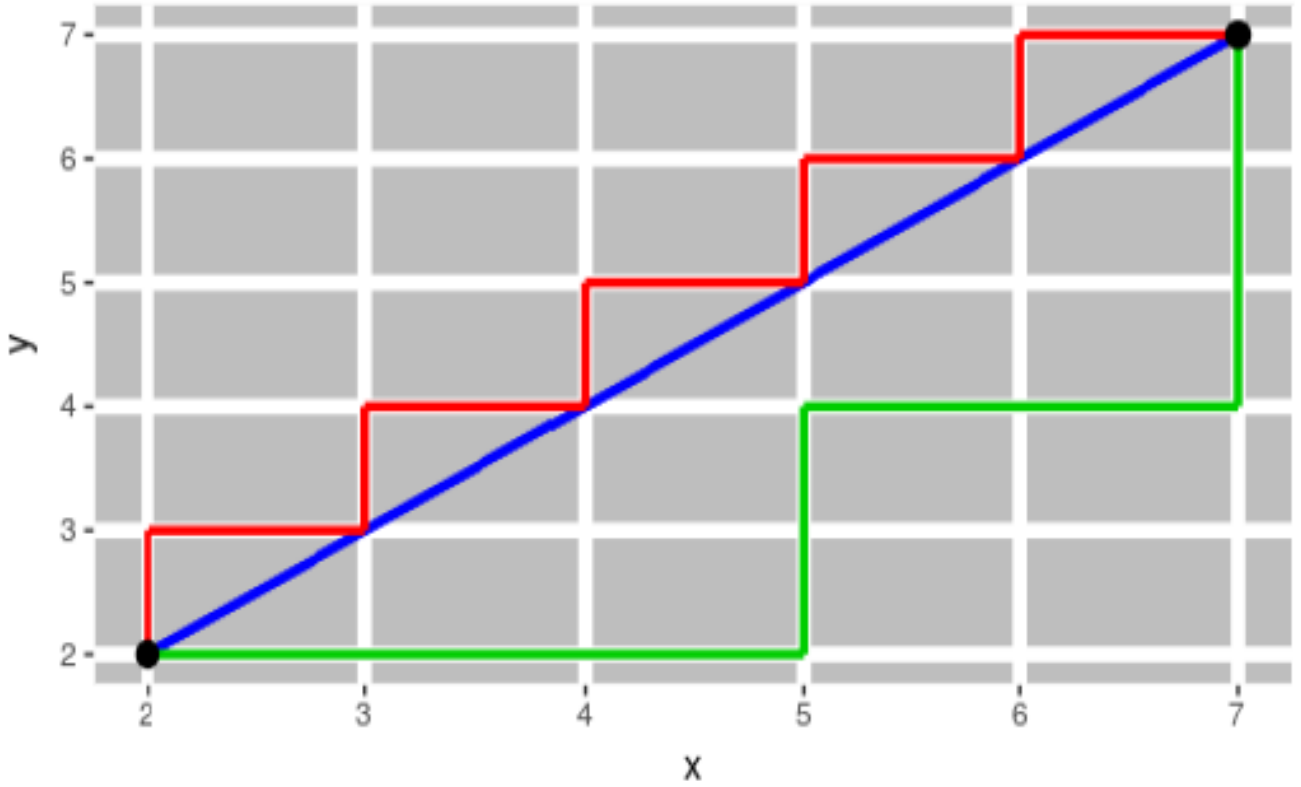
B. **Distancia de Manhattan:** La distancia de Manhattan, también conocida como taxicab metric, rectilinear distance o L1 distance, define la distancia entre dos puntos p y q como el sumatorio de las diferencias absolutas entre cada dimensión. Esta medida se ve menos afectada por outliers (es más robusta) que la distancia euclídea debido a que no eleva al cuadrado las diferencias.

$$d_{man}(p,q) = \sum_{i=1}^{n} |(p_i - q_i)|$$

Mundo No Supervisado
b.¿Qué Buscan?

Medidas de Distancia

La siguiente imagen muestra una comparación entre la distancia euclídea (segmento azul) y la distancia de manhattan (segmento rojo y verde) en un espacio bidimensional. Existen múltiples caminos para unir dos puntos con el mismo valor de distancia de manhattan, ya que su valor es igual al desplazamiento total en cada una de las dimensiones.



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Una vez seleccionado el número adecuado de clusters y aplicado el algoritmo de clustering pertinente se tiene que evaluar la calidad de los de los mismos, de lo contrario, podrían derivarse conclusiones de agrupación que no se corresponden con la realidad. Pueden diferenciarse tres tipos de estadísticos empleados con este fin:

• Validación interna de los clusters: Emplean únicamente información interna del proceso de clustering para evaluar la bondad de las agrupaciones generadas. Se trata de un proceso totalmente unsupervised ya que no se incluye ningún tipo de información que no estuviese ya incluida en el clustering.

• Validación externa de los clusters (ground truth): Combinan los resultados del clustering (unsupervised) con información externa (supervised), como puede ser un set de validación en el que se conoce el verdadero grupo al que pertenece cada observación. Permiten evaluar hasta qué punto el clustering es capaz de agrupar correctamente las observaciones. Se emplea principalmente para seleccionar el algoritmo de clustering más adecuado, aunque su uso está limitado a escenarios en los que se dispone de un set de datos de validación.

• Significancia de los clusters: Calculan la probabilidad (p-value) de que los clusters generados se deban únicamente al azar.

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

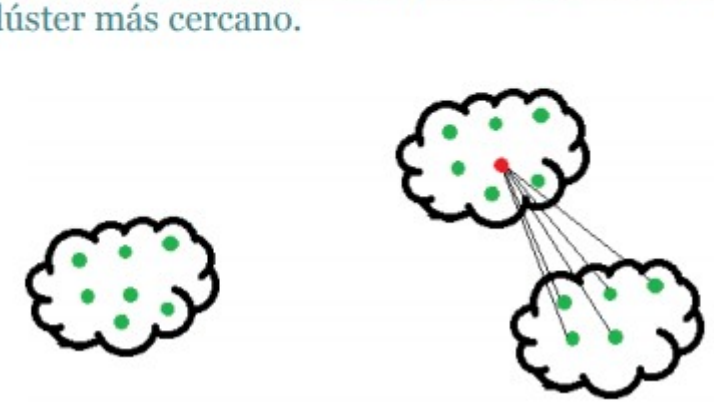
La idea principal detrás del clustering es agrupar las observaciones de forma que sean similares a aquellas que están dentro de un mismo cluster y distintas a las de otros clusters, es decir, que la homogeneidad (también llamada compactness o cohesion) se lo mayor posible a la vez que lo es la separación entre clusters.

• Cohesión: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.

• Separación: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

- Cohesión a(x) : distancia promedio de x a todos los demás puntos en el mismo clúster.

- Separación b(x) : distancia promedio de x a todos los demás puntos en el clúster más cercano.



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

Sum of Squared Within (SSW)

Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster C_i y m_i el centroide del clúster C_i.

Sum of Squared Between (SSB)

Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación)

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

Siendo k el número de clústeres, n_j el número de elementos en el clúster j, c_j el centroide del clúster j y $\bar{x}$ es la media del data set.

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Validación interna de los clusters: estabilidad, silhouette y Dunn

- Cohesión se mide como **within cluster sum of squares (SSE)**

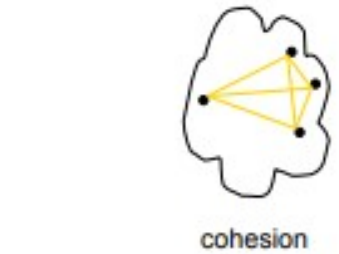
$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

- Separación se mide como **between cluster sum of squares (BSS)**

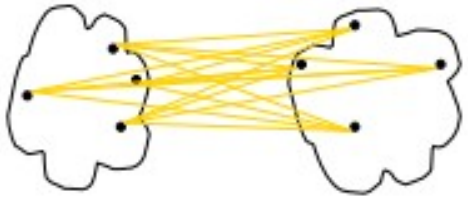
$$BSS = \sum_i |C_i|(m - m_i)^2$$

Medidas internas: Cohesión y separación

- Enfoque basado en grafos de proximidad
- Cohesión: suma de los pesos de todos los arcos en un cluster
- Separación: suma de los pesos entre nodos del cluster y de otros clusters



cohesion



separation

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Silhouette width

El coeficiente de silueta contrasta la distancia media a elementos en el mismo grupo con la distancia media a elementos en otros grupos. Los objetos con un valor de silueta alto están considerados bien agrupados, los objetos con un valor bajo pueden ser ruido o anomalías. Estos índices trabajan bien con k-means, y es también utilizado para determinar el número óptimo de grupos:

El valor de la silueta es una medida de cuán similar es un objeto a su propio cúmulo (cohesión) en comparación con otros cúmulos (separación). La silueta va de -1 a +1, donde un valor alto indica que el objeto está bien emparejado con su propio cúmulo y mal emparejado con los cúmulos vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración del cúmulo es apropiada.

• Calcular la media de las distancias (llámese ai) entre la observación i y el resto de observaciones que pertenecen al mismo cluster. Cuanto menor sea ai mayor la similitud que tiene con el resto de observaciones de su cluster.

• Identificar como bi a la menor de las distancias promedio entre i y el resto de clusters, es decir, la distancia al cluster más próximo (neighbouring cluster).

• Calcular el valor de silhouette como:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

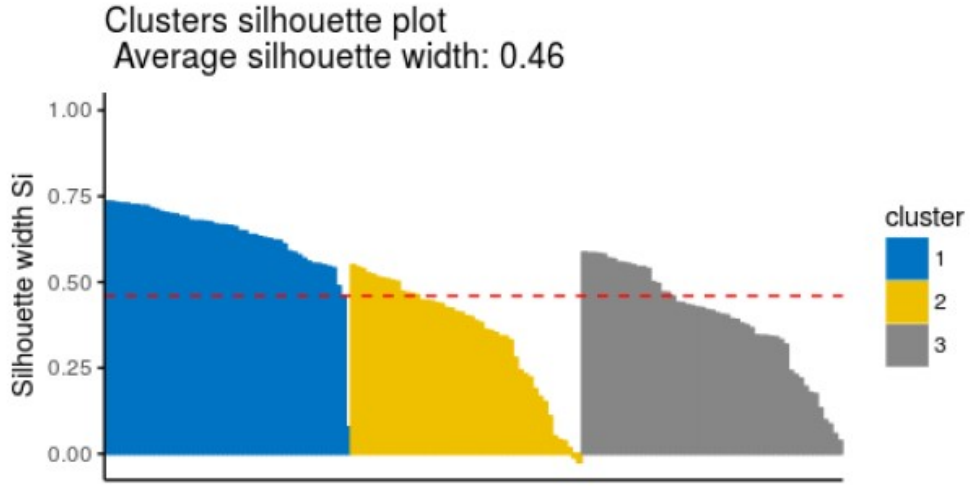
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Su valor puede estar entre -1 y 1, siendo valores altos un indicativo de que la observación se ha asignado al cluster correcto. Cuando su valor es próximo a cero significa que la observación se encuentra en un punto intermedio entre dos clusters. Valores negativos apuntan a una posible asignación incorrecta de la observación. Se trata por lo tanto de un método que permite evaluar el resultado del clustering a múltiples niveles:

- La calidad de asignación de cada observación por separado. Permitiendo identificar potenciales asignaciones erróneas (valores negativos de silhouette).
- La calidad de cada cluster a partir del promedio de los índices silhouette de todas las observaciones que lo forman. Si por ejemplo se han introducido demasiados clusters, es muy probable que algunos de ellos tengan un valor promedio mucho menor que el resto.
- La calidad de la estructura de clusters en su conjunto a partir del promedio de todos los índices silhouette.

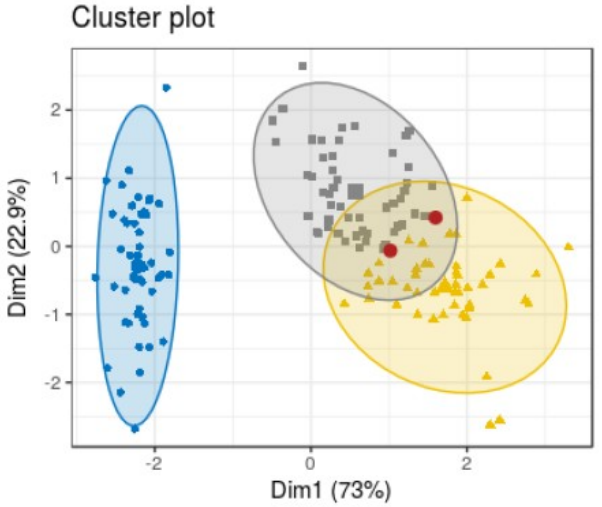
```
##   cluster size ave.sil.width
## 1       1   50          0.64
## 2       2   47          0.35
## 3       3   53          0.39
```

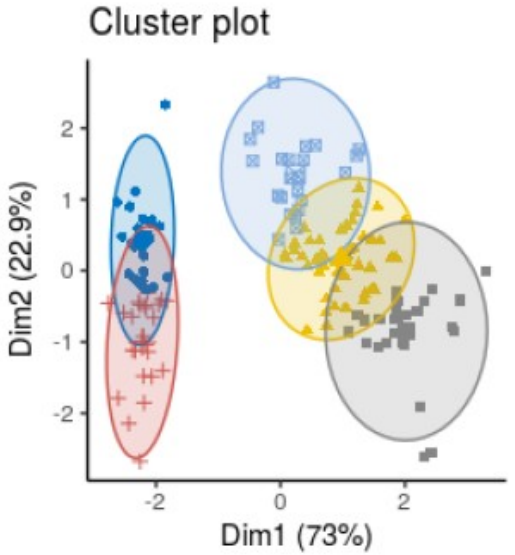


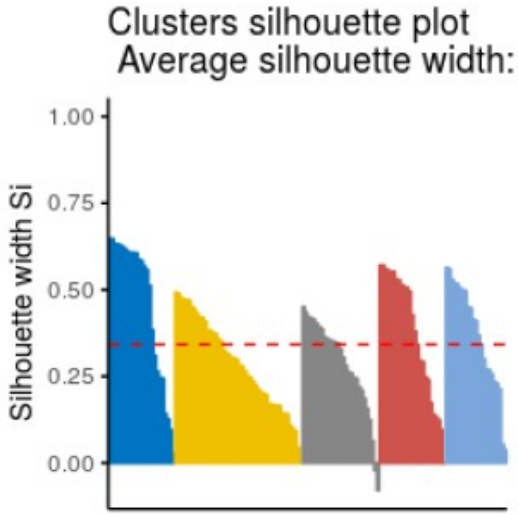
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

El cluster número 2 (amarillo) tiene observaciones con valores de silhouette próximos a 0 e incluso negativos, lo que indica que esas observaciones podrían estar mal clasificadas. Viendo la representación gráfica del clustering, cabe esperar que sean observaciones que están situadas en la frontera entre los clusters 2 y 3 ya que solapan.







Mundo No Supervisado
b.¿Qué Buscan?

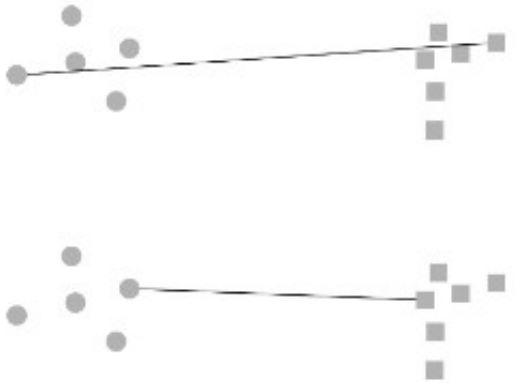
Calidad de los clusters

Índice Dunn

El objetivo de este índice es identificar un conjunto de clústeres que sean compactos, con una varianza pequeña entre los miembros del clúster, y que éstos estén bien separados de los miembros de otros clústeres. Un valor más alto del índice de Dunn indica un mejor rendimiento del algoritmo de clustering.

El índice de Dunn tiene un valor entre cero y infinito, y debe ser lo más alto posible. Por lo tanto, la distancia entre los miembros de un clúster debe ser los más baja posible, y la distancia entre los clústeres lo más alta posible.

Por ejemplo, en el caso de la distancia entre clústeres puede utilizarse la distancia más corta entre dos puntos de diferentes clústeres, o la distancia más larga, o la distancia entre los centroides. También pueden utilizarse diferentes indicadores de la distancia dentro de un clúster.



$$D = \frac{separacion\ minima\ interclusters}{separacion\ maxima\ intracluster}$$

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad

Las medidas de estabilidad son un tipo particular de validación interna que cuantifican el grado en que varían los resultados de un clustering como consecuencia de eliminar, de forma iterativa, una columna del set de datos. Todas ellas son relativamente costosas desde el punto de vista computacional ya que requieren repetir el clustering tantas veces como columnas tenga el set de datos. Dentro de esta familia de medidas se encuentran:

• Average proportion of non-overlap (APN): mide la proporción media de observaciones que no se asignan al mismo cluster cuando se elimina una columna del set de datos en comparación a cuando se incluyen todas.

• Average distance (AD): mide la media de las distancias promedio intra-cluster empleando todos los datos y eliminando una columna a la vez.

• Average distance between means (ADM): mide la media de las distancias entre centroides empleando todos los datos y eliminando una columna a la vez.

• Figure of merit (FOM): mide media de la varianza intra-cluster de la columna eliminada, empleando la estructura del clustering calcula con las columnas no eliminadas.

Los valores de APN, ADM, y FOM pueden ir desde 0 a 1, siendo valores pequeños un indicativo de alta estabilidad. En el caso de AD ocurre lo mismo pero sus valores pueden ir de 0 hasta infinito.

Mundo No Supervisado
b.¿Qué Buscan?

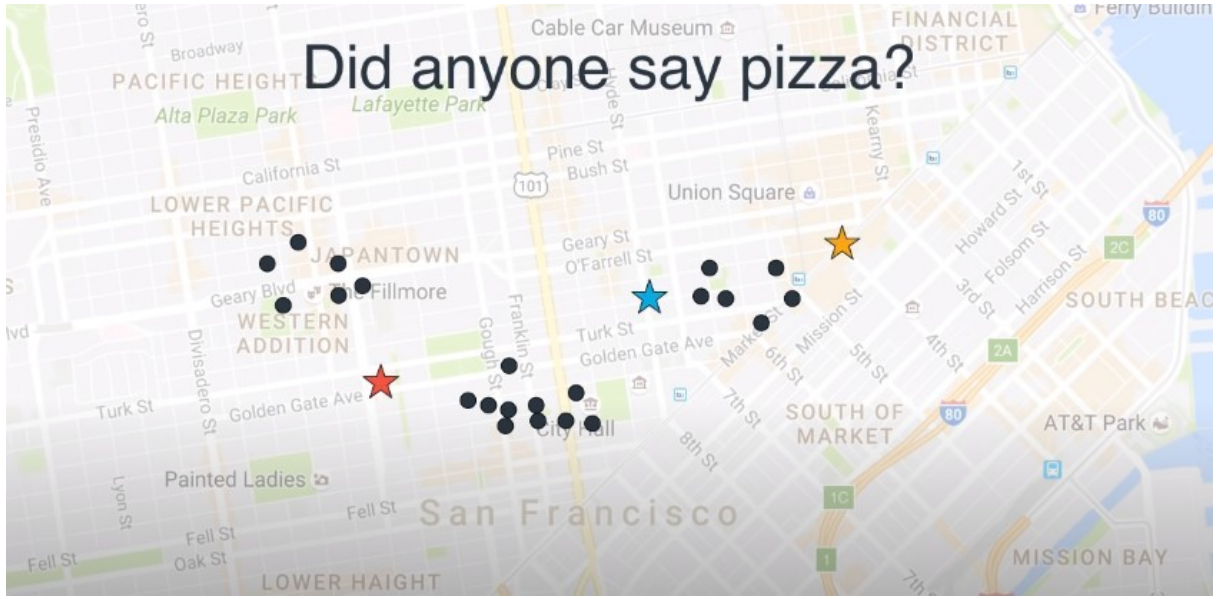
Calidad de los clusters

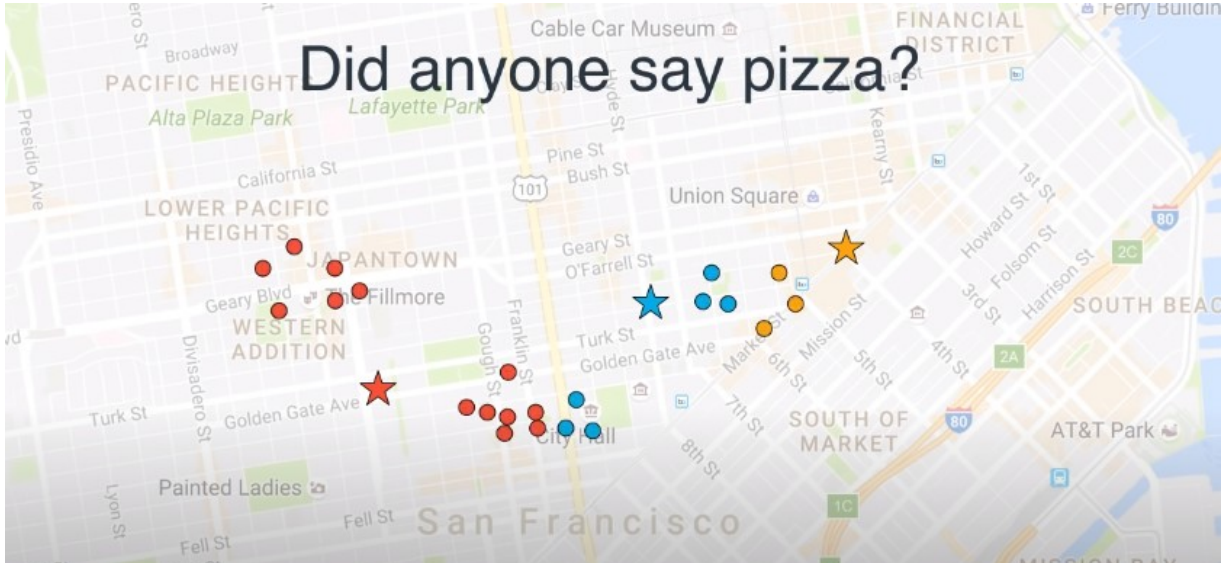
Medidas de estabilidad

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad





Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad

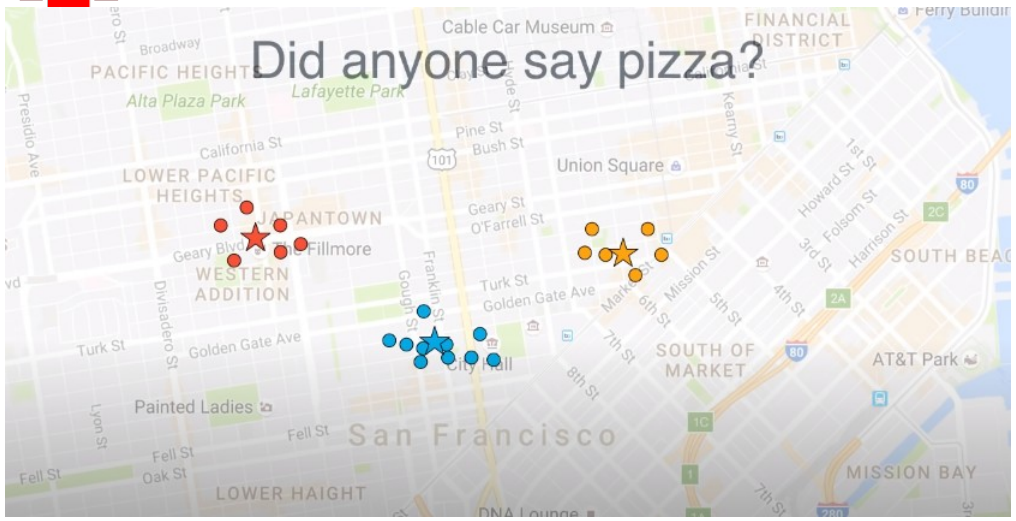
Mundo No Supervisado
b.¿Qué Buscan?

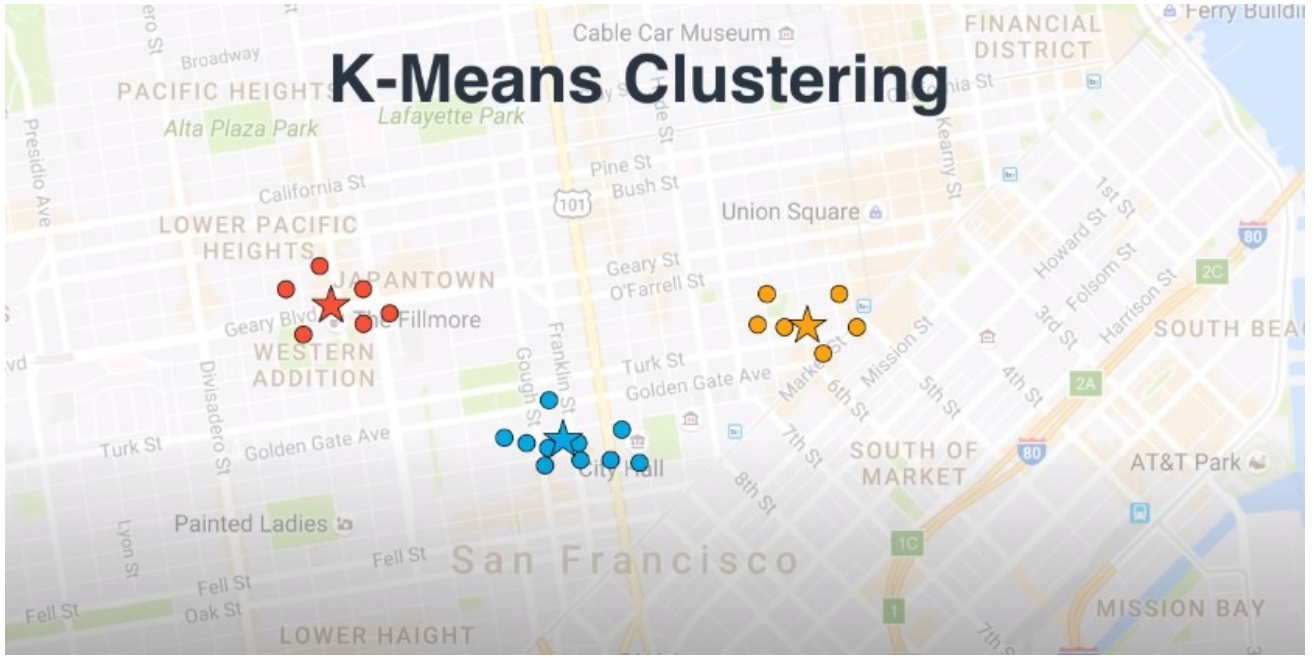
Calidad de los clusters

Medidas de estabilidad

Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

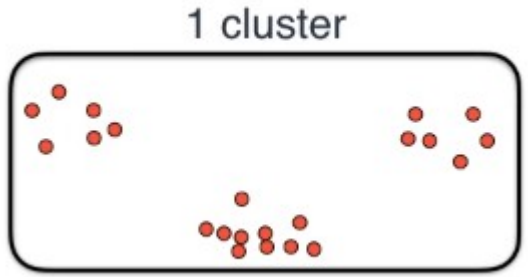


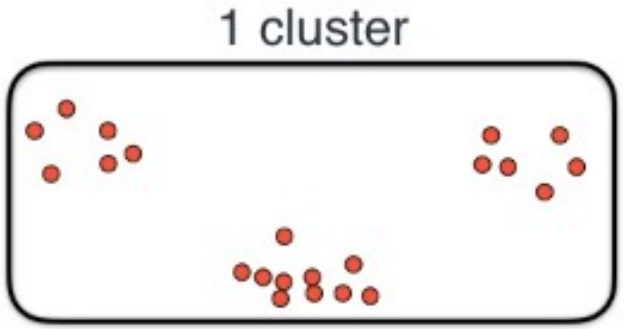


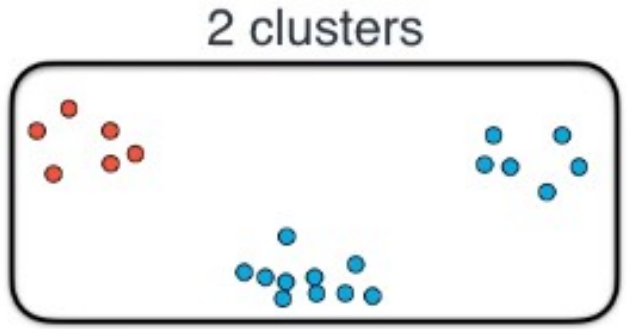
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad

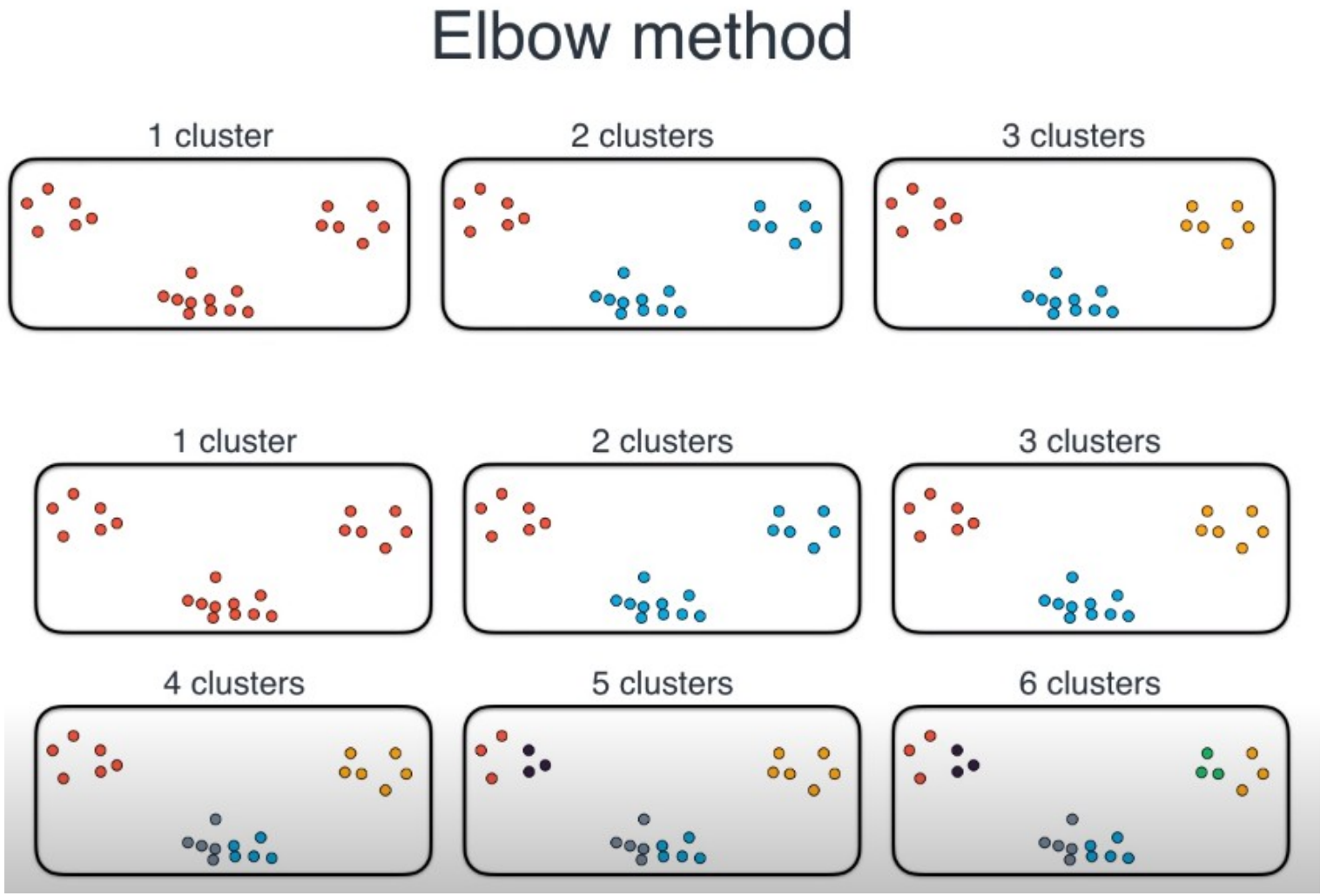






Mundo No Supervisado
b.¿Qué Buscan?

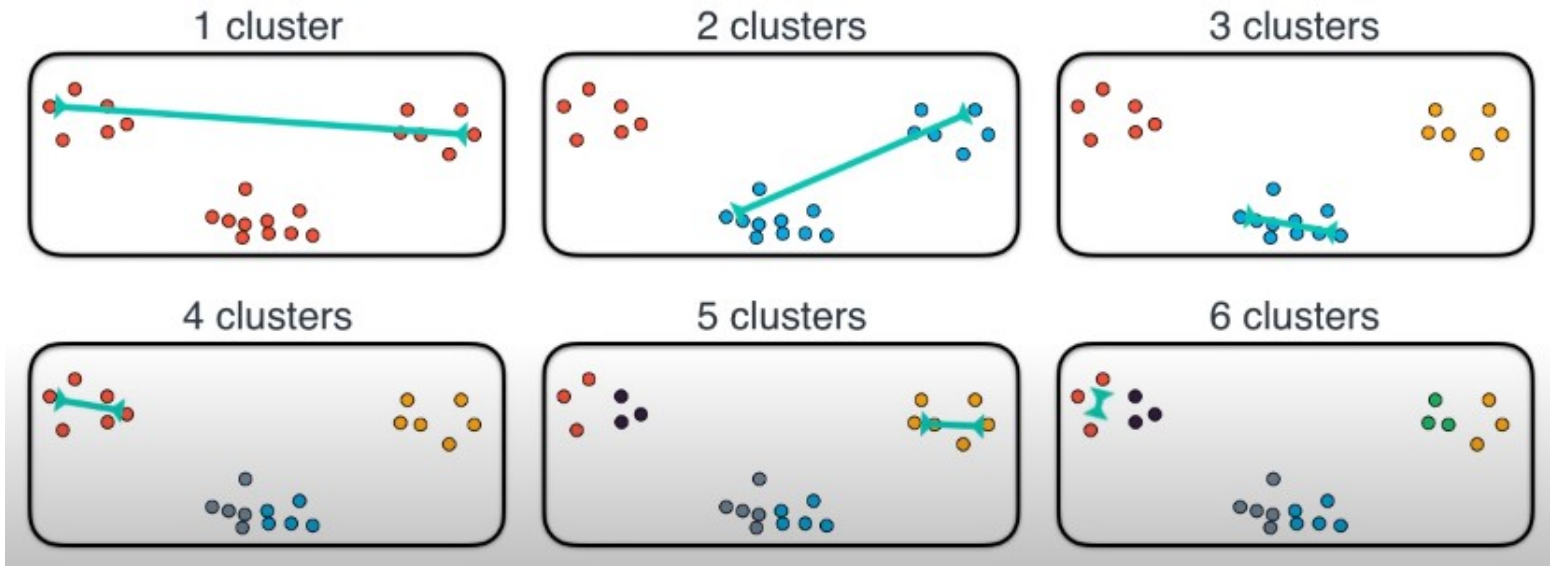
Calidad de los clusters



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

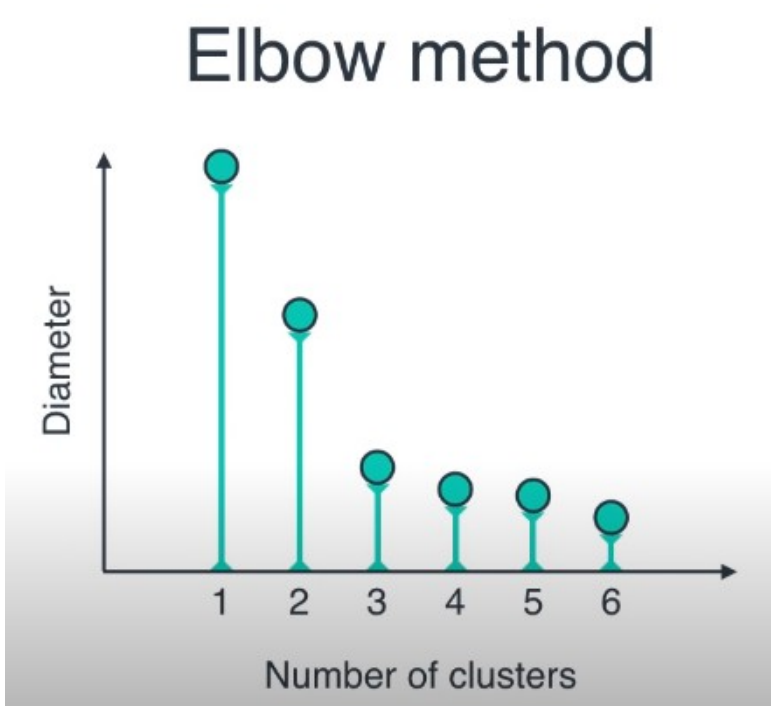
Medidas de estabilidad

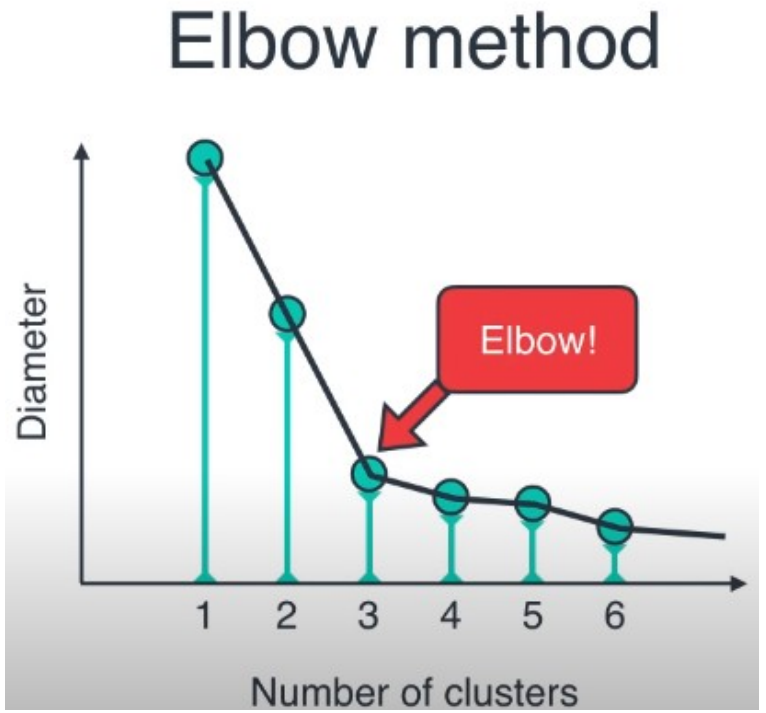


Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

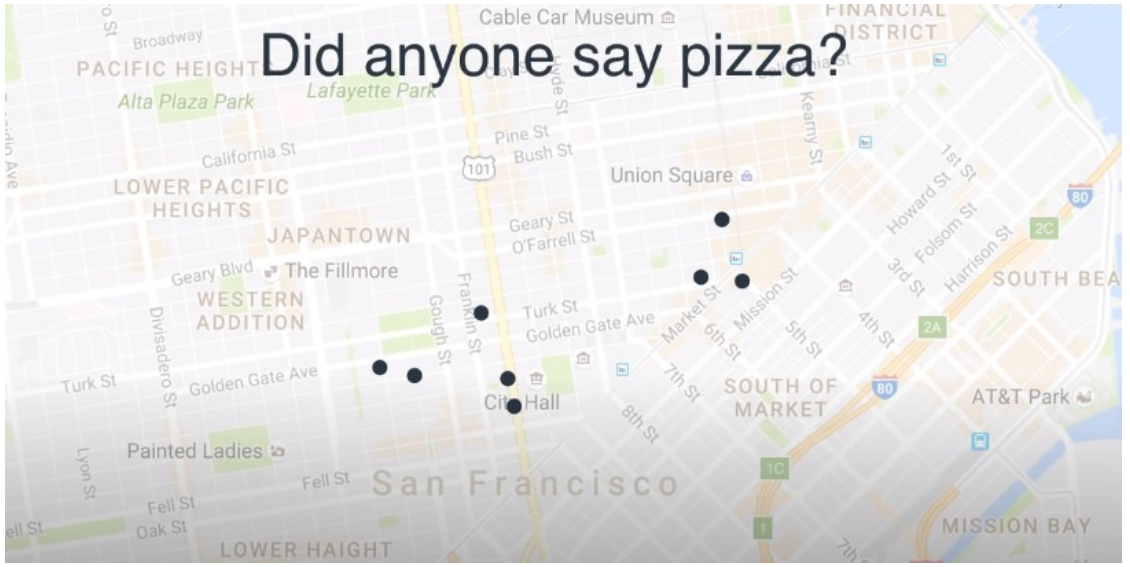
Medidas de estabilidad





Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters



Mundo No Supervisado
b.¿Qué Buscan?

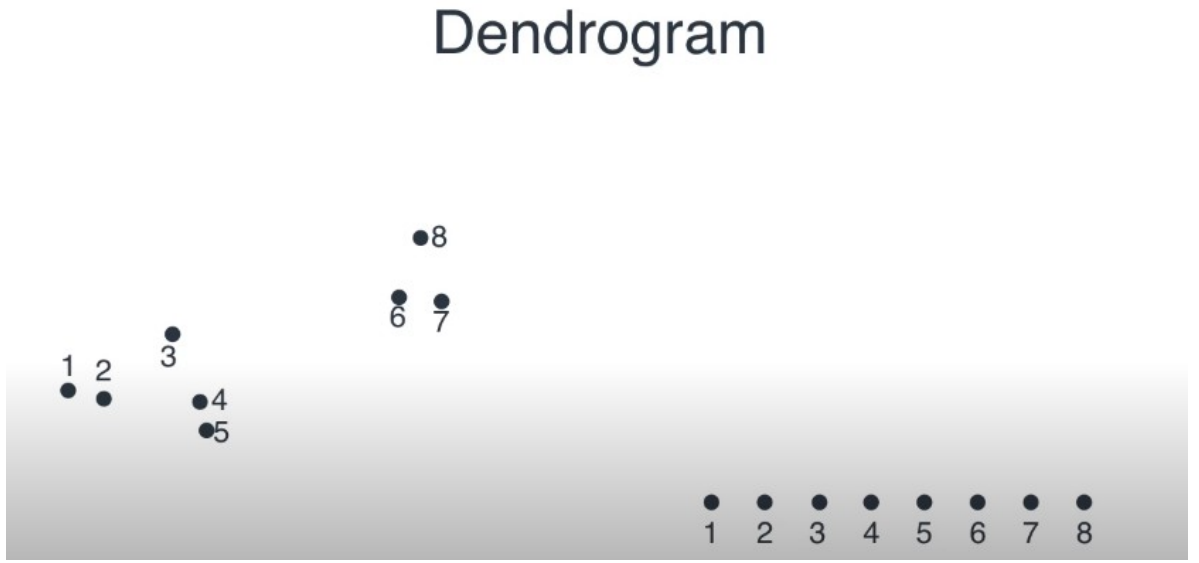
Calidad de los clusters

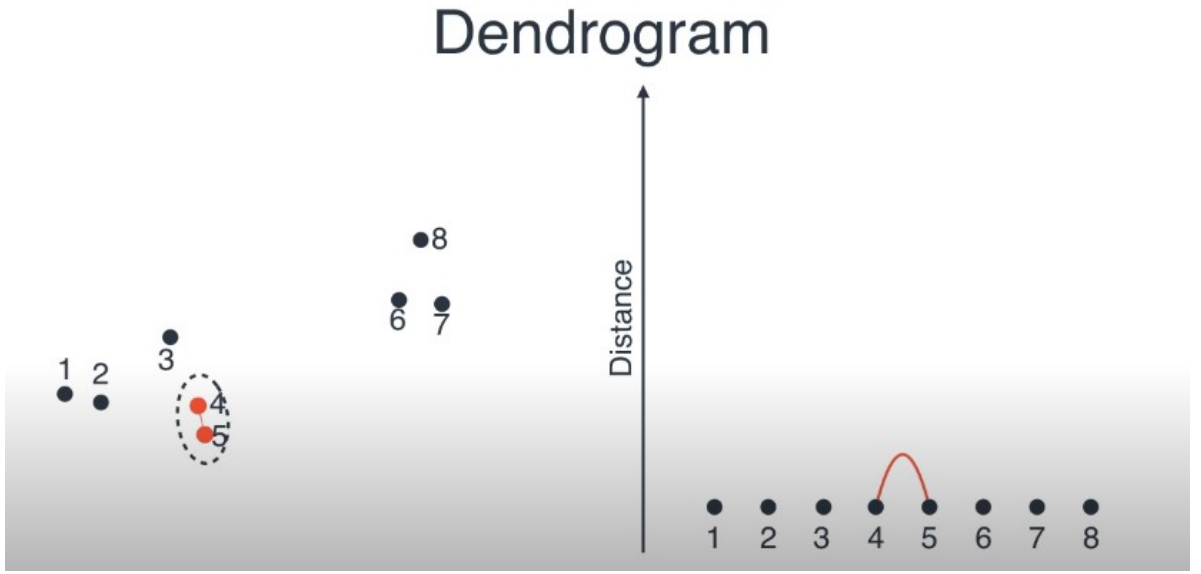
Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Mundo No Supervisado
b.¿Qué Buscan?

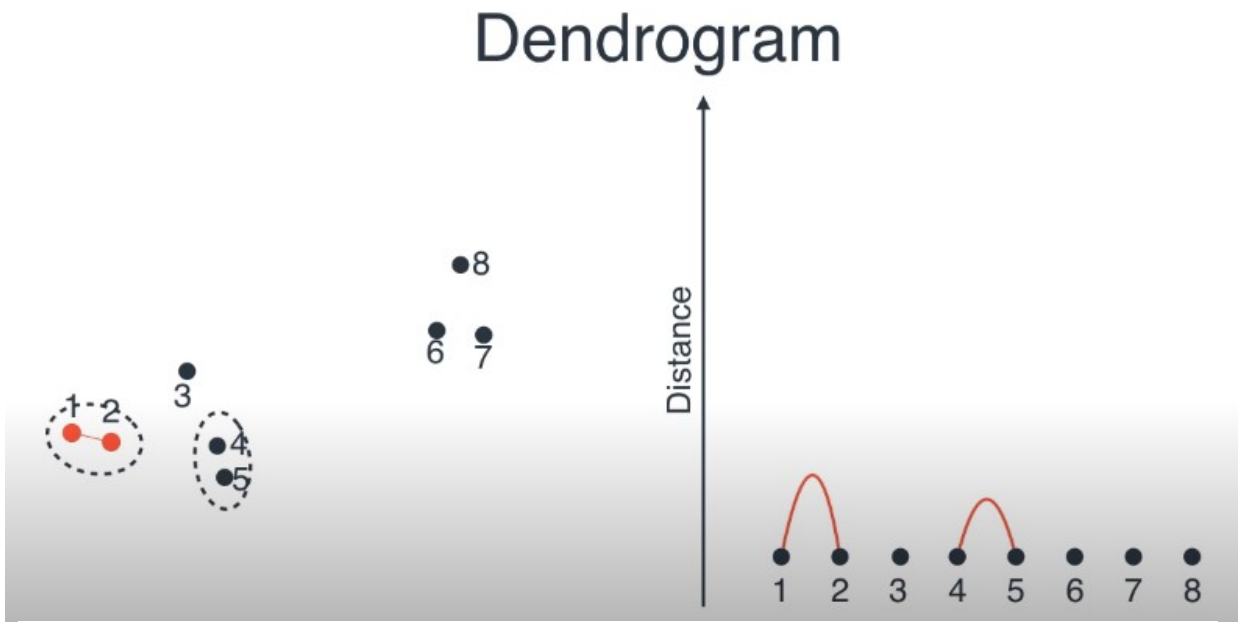
Calidad de los clusters

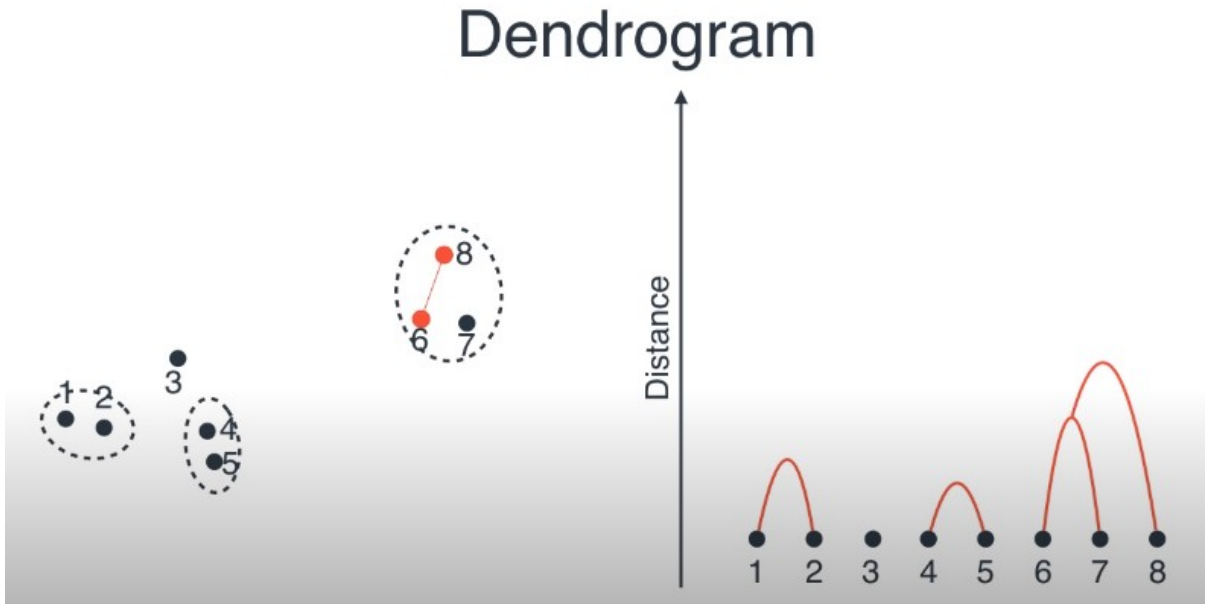




Mundo No Supervisado
b.¿Qué Buscan?

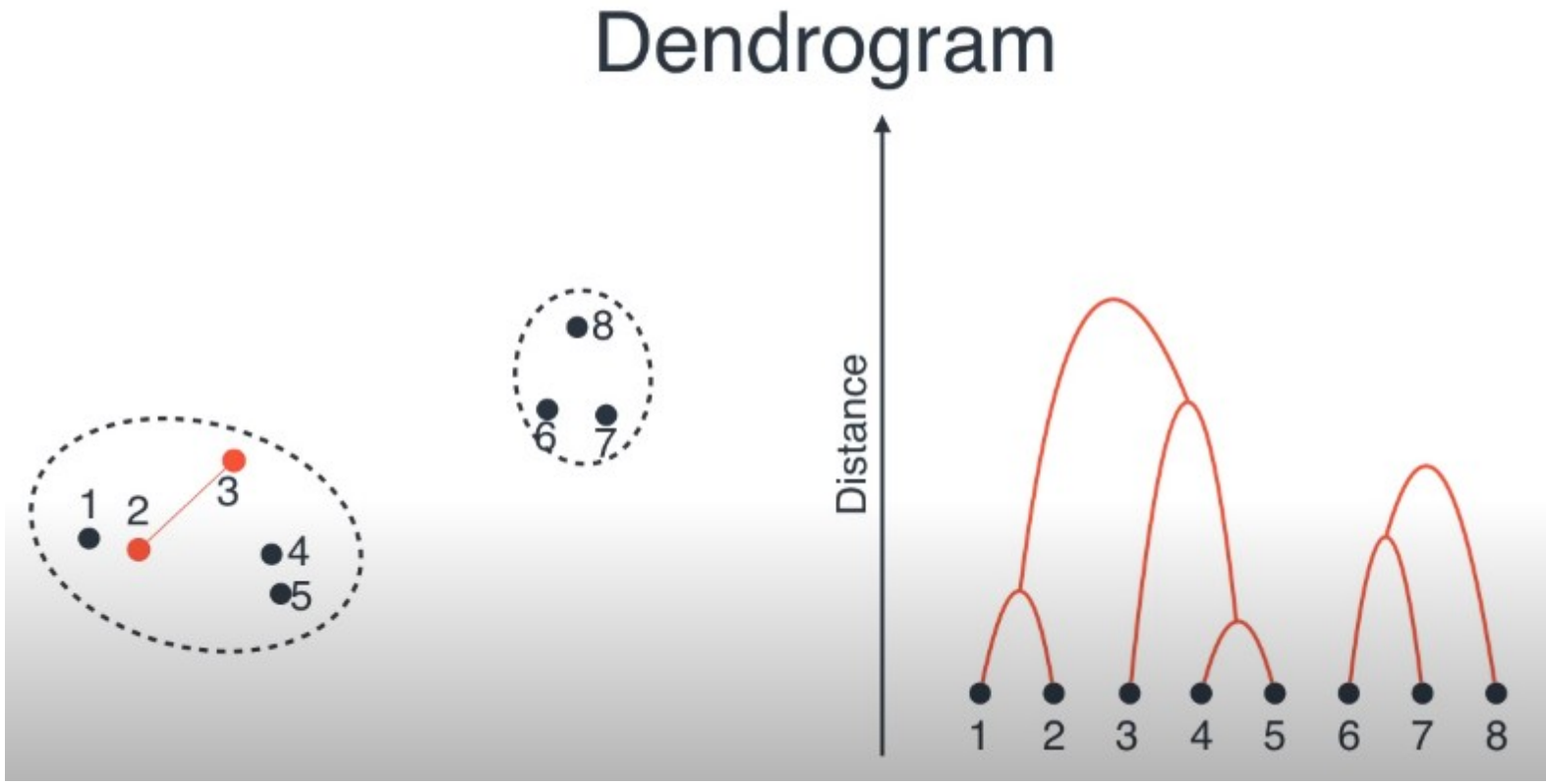
Calidad de los clusters





Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters



Mundo No Supervisado
b.¿Qué Buscan?

Calidad de los clusters

Medidas de estabilidad