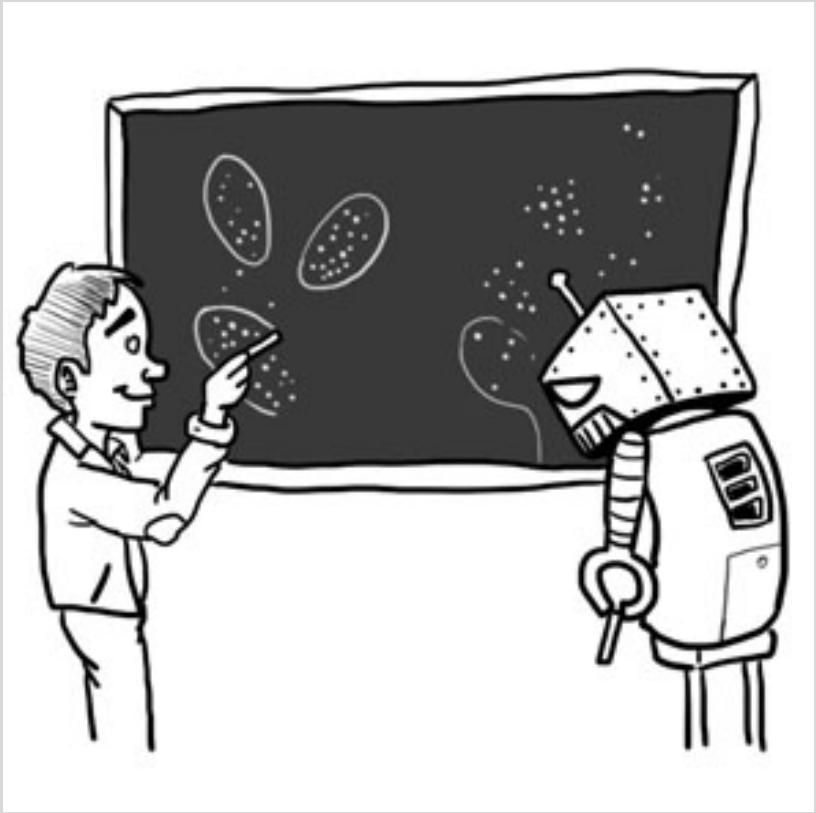


DigitalHouse > Coding School

DATA SCIENCE

Unidad: 4
Modulo: 7

Introducción al proceso de Clustering

Noviembre
2017
Buenos Aires

OBJETIVOS DE LA CLASE

1

SOBRE LOS DATOS

Analizar formato y preprocesamiento para aplicar algoritmos de clusterización

2

SOBRE ALGORITMOS

Conocer los algoritmos más usados.
Realizar un análisis de cluster con K-Means

3

SOBRE EVALUACIÓN DE AJUSTE

Conocer conceptos de evaluación para el ajuste de clusterización

APRENDIZAJE NO SUPERVISADO

- El grueso del curso trabajó sobre el marco del Aprendizaje Supervisado
- En ese marco buscábamos predecir una variable target (**Y**) a partir de una matriz de predictores (**X**)
- En este módulo trabajaremos sobre problemas de Aprendizaje No Supervisado: solamente dispondremos de la matriz de predictores **X** => No hay una variable target.
- El objetivo en un problema de Aprendizaje No Supervisado es descubrir alguna estructura-patrón interesante en los datos.
  - ¿hay alguna forma informativa de visualizar los datos?
  - ¿se pueden identificar grupos de casos o variables similares?

- El Aprendizaje No Supervisado tiene un carácter un poco más subjetivo que el Supervisado, dado que no hay un objetivo “simple” en el análisis.

- No obstante, las técnicas de Aprendizaje No Supervisado están creciendo bastante en varios campos:
  - identificación de subgrupos de pacientes con cáncer, agrupados según mediciones de sus expresiones genéticas
  - identificación de grupos de compradores caracterizados por sus historias de compra previas
  - grupos de películas agrupadas por las calificaciones asignadas por los espectadores

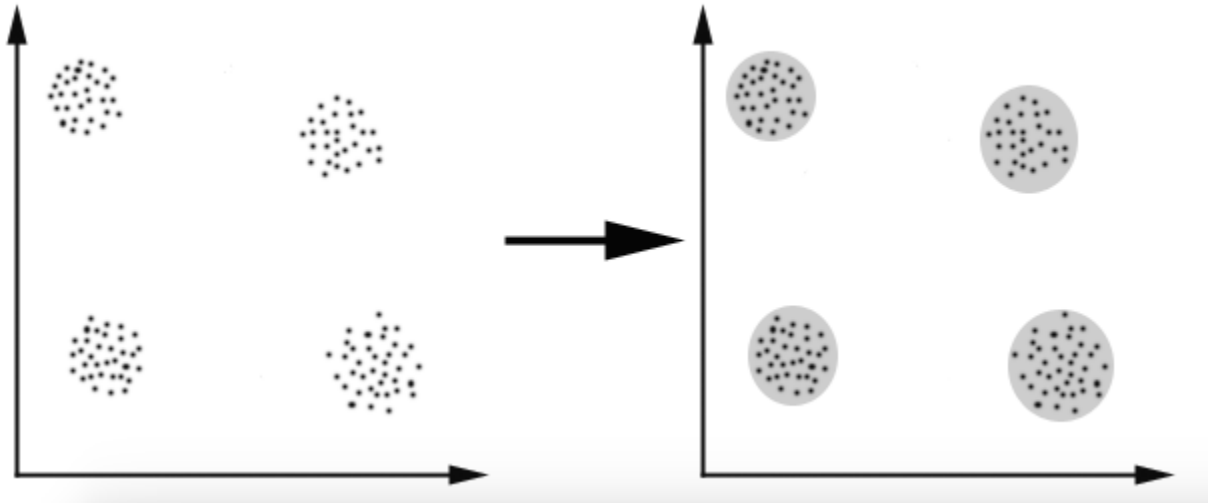
INTRODUCCIÓN AL PROCESO DE CLUSTERING

- El proceso de **Clusterización** o **Clustering** es uno de los métodos más usados para llevar comprender cierta estructura en un conjunto de datos. Probablemente sea uno de los métodos de aprendizaje no supervisado más importantes

- Una definición informal de Clustering podría ser "**el proceso de organizar los objetos en grupos cuyos miembros son similares de alguna manera**".

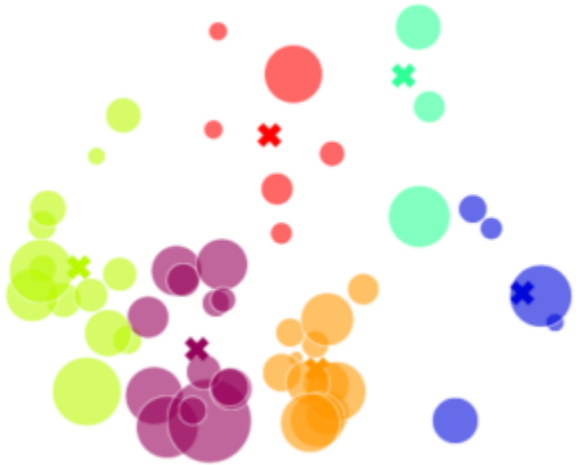
- Un **cluster** es por lo tanto una colección de objetos que son "similares" entre ellos y son "disímiles" a los objetos pertenecientes a otros clusters.

- **El criterio de similitud es la distancia:** dos o más objetos pertenecen al mismo grupo si están "cercanos" según una distancia dada (en este caso distancia geométrica).



- **Marketing:** encontrar grupos de clientes con un comportamiento similar dado una gran base de datos de clientes que contienen sus propiedades y registros de compras.
- **Biología:** clasificación de plantas y animales dados sus características.
- **Seguros:** identificación de grupos de titulares de pólizas de seguros de automóviles con un costo alto en promedio en reclamos.
- **Urbanismo:** identificación de grupos de casas según su tipo de vivienda, valor y ubicación geográfica.
- **Estudios de terremotos:** agrupar los epicentros de terremotos observados para identificar zonas peligrosas
- **WWW:** clasificación de documentos; Agrupación de datos de weblog para descubrir grupos de patrones de acceso similares.

¿Qué diferencia un Proceso de Clusterización de uno de Clasificación?



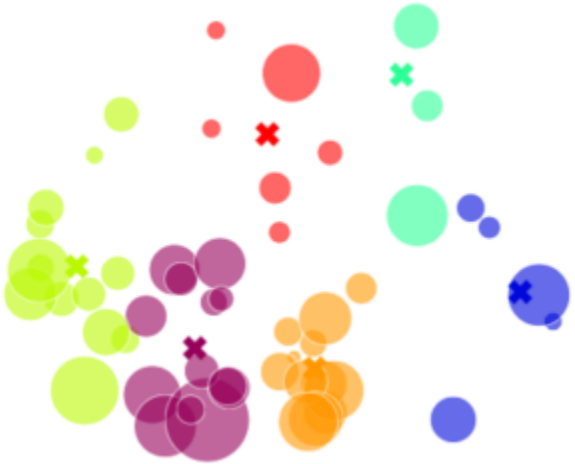
Si sólo estamos creando grupos, ¿no son los dos el mismo proceso?

- Existe una importante distinción entre clasificación y agrupación: En la clasificación, estamos agrupando los datos de acuerdo con un conjunto de grupos predefinidos. *(Supervisado)*

"*Sabemos cuáles son las características de un mamífero, y los humanos tienen las características de ese grupo predefinido.*"

- En la agrupación, sin embargo, nos propusimos averiguar si los puntos de nuestro conjunto de datos tienen relaciones entre sí, y agrupamos aquellos con características similares en un grupo. En otras palabras, tenemos que descubrir las propias clases. *(No Supervisado)*

¿Cómo funciona Clustering?



Los algoritmos de clustering deberían satisfacer alguno(s) de los siguientes requerimientos:

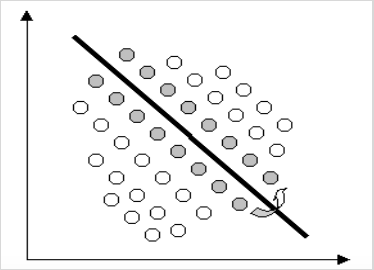
- Escalabilidad
- Tratar con diferentes tipos de atributos
- Capacidad de descubrir clusters con forma arbitraria
- Capacidad para hacer frente al ruido y los valores extremos
- Insensibilidad al orden de los registros de entrada
- Alta dimensionalidad
- Interpretabilidad y usabilidad

Hay una serie de problemas en Clustering. Entre ellos:

- Las técnicas actuales de Clustering no abordan todos los requisitos adecuadamente (y simultáneamente).
- Tratar con un gran número de dimensiones y un gran número de elementos de datos puede ser problemático debido a la complejidad del tiempo de cómputo.
- La eficacia del método depende de la definición de "distancia" (para el agrupamiento basado en la distancia).
- Si no existe una medida obvia de distancia, debemos "definirla", lo que no siempre es fácil, especialmente en espacios multidimensionales.
- El resultado del algoritmo de agrupamiento (que en muchos casos puede ser arbitrario) puede interpretarse de diferentes maneras.

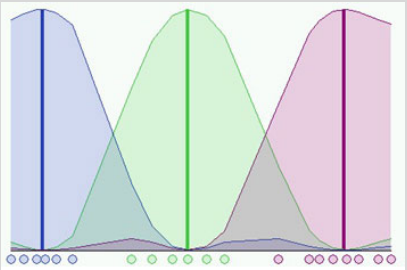
Los algoritmos de Clustering pueden clasificarse como:

Si un determinado dato pertenece a un grupo definido, entonces no puede ser incluido en otro.



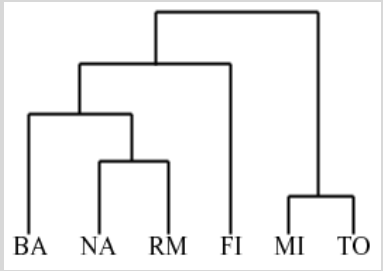
Exclusive Clustering (ej: K-means)

Utilizan conjuntos difusos para agrupar datos, de modo que cada punto puede pertenecer a dos o más grupos con diferentes grados de pertenencia.



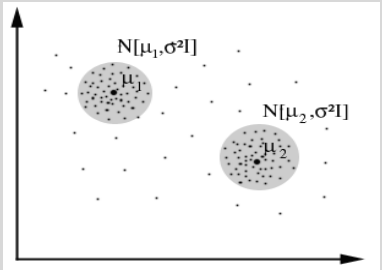
Overlapping Clustering (ej: Fuzzy C-means)

Se basan en la unión entre los dos clusters más cercanos. La condición inicial se realiza estableciendo cada datum como un cluster. Después de algunas iteraciones llega a los clusters finales deseados.



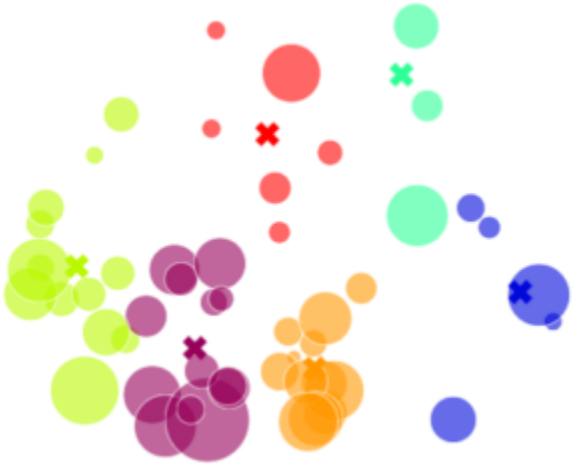
Hierarchical Clustering

Utilizan un enfoque completamente Probabilístico



Probabilistic Clustering (ej: Mixture of Gaussians)

Métricas de Distancia



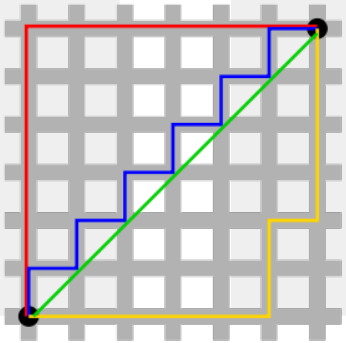
Para alta Dimensionalidad una medida muy popular es la **Métrica de Minkowski**:

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$$

donde *d* es la dimensionalidad de los datos.

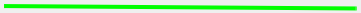
Si p=1 obtenemos la **distancia de Manhatt**

$$d_1(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$



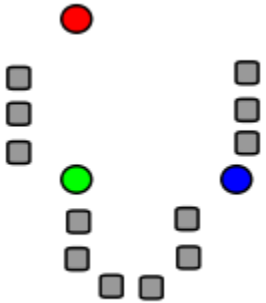
Si p=2 obtenemos la **distancia de**

$$d_2(x_i, x_j) = \sqrt{\sum_{k=1}^{d} |x_{ik} - x_{jk}|^2}$$

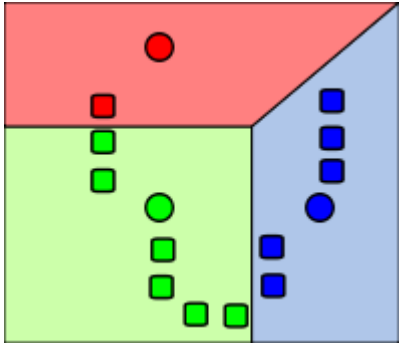


ALGORITMO K-MEANS
(MacQueen,1967)

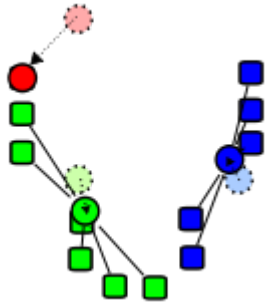
- El procedimiento sigue una manera sencilla y fácil de clasificar un determinado conjunto de datos a través de un cierto número de clusters (suponga k clusters) **fijado a priori**.
- La idea principal es **definir k centroides**, uno para cada cluster.



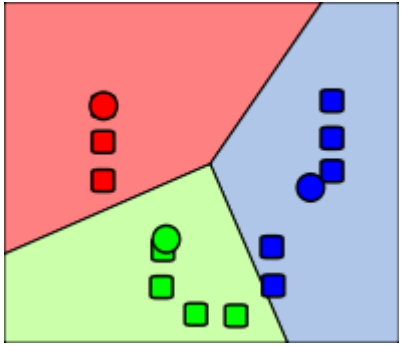
Paso 1: *Se define un número K de cluster y una ubicación (al azar) de los centroides.*



Paso 2: *Se forman K cluster asociando los puntos a los centroides más cercanos. Las particiones aquí representan el diagrama de Voronoi generado por los centroides.*



Paso 3: *Se recalcula el centroide de cada uno de los K-cluster.*



Paso 4: *Se repite el paso 2 y 3 hasta que la convergencia ha sido alcanzada.*

- Finalmente, este algoritmo tiene como objetivo **minimizar una función objetivo**, en este caso una función de error cuadrático. La función objetivo:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

donde $\left\| x_i^{(j)} - c_j \right\|^2$ es la distancia del *punto-i* al centroide del *cluster-j.*

Práctica Guiada

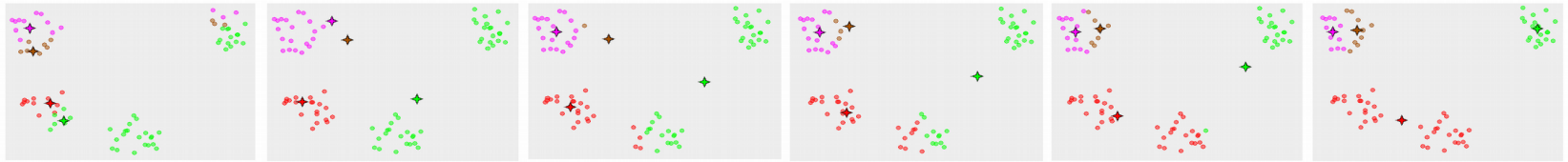
(Implementando K-MEANS)

— Algunos problemas del algoritmo K-Means

Dado que comienza con un asignación de clusters aleatoria puede suceder que los resultados varíen en diferentes corridas

Es necesario “tunear” en parámetro K (cantidad de clusters)

Por las características del algoritmo, el mínimo global puede no ser logrado



K-means solo logra captar separaciones lineales entre los datos.

VARIABLES CATEGÓRICAS
ALGORITMO K-MODES
(Huang, 1999)

- K-means funciona muy bien para variables cuantitativas. ¿Pero qué pasa si queremos clusterizar variables categóricas?

- Una opción directa sería clusterizar en base a las variables dummies para cada categoría de las variables originales y luego utilizar k-means.
  Problema: para datasets de alta dimensionalidad este enfoque se hace impracticable

- Una opción es el algoritmo K-modes: este enfoque reemplaza las métricas de distancia euclidiana por una medida de “disimilaridad” y usa las modas de cada variable para representar los centros de los clusters.

- La distancia es definida de la siguiente forma:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \qquad \delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j . \end{cases}$$

  - Es decir, se computa la cantidad de “diferencias” o “no coincidencias” que existen entre dos elementos Xi e Yi.
  - En el caso específico del algoritmo k-modes, las distancias se calcularán para cada individuo que presenta un vector de categorías de cada una de las variables y el vector de centroides de un cluster.

- Algoritmo

  Paso 1. Se seleccionan aleatoriamente k casos únicos como los centros de los clusters (modas)

  Paso 2. Se calculan las distancias entre cada objeto y entre los centros. El objeto es asignado al centro con menor distancia. Se repite este proceso hasta que todos los objetos han sido asignados.

  Se selecciona una nueva moda para cada cluster y se compara con la moda anterior. Si es diferente se vuelve al paso 2. Si es igual, se termina el proceso.

- El proceso de c[illegible] función

$$F(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}u_{i,l}d(x_{i,j},z_{l,j})$$

- donde U es la matriz de partición (el dataset). Z es el set de vectores de modas.

	Locus									
Indiv	1	2	3	4	5	6	7	8	9	10
1	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB
2	AB	BB	BB	AB	BB	BB	AB	AB	BB	AB
3	BB	BB	AA	AA	AB	AB	AA	AB	BB	AB
4	AB	BB	AB	AB	AB	AB	BB	AB	AA	AB
5	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB
6	BB	AB	AB	AB	BB	BB	AB	AA	AB	AB
7	BB	BB	BB	BB	AB	AB	AA	AB	BB	AB
8	AB	BB	AB	AB	AA	AA	AB	BB	AB	BB
9	BB	AA	AB	AB	BB	AB	AB	AA	AB	AB
10	AB	BB	AB	BB	AB	AB	BB	AB	AB	AA
11	AA	BB	AA	AA	AA	AB	AA	AB	AB	AB
12	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB
13	AB	BB	AB	AA	AB	AB	BB	AB	AA	AA
14	BB	AA	AB	AB	BB	BB	AB	AA	AB	AB
15	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA

- Dataset de secuencias genotípicas.
  - Cada fila es un individuo
  - Cada columna es un secuencia
- Objetivo: encontrar grupos de individuos con secuencias genotípicas similares => K-modes

(a)

	Locus									
Cluster	1	2	3	4	5	6	7	8	9	10
1 (1)	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB
2 (5)	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB
3 (12)	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB
4 (15)	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA

- Se seleccionan los individuos 1, 5, 12 y 15 como los centros de los clusters
- Se calculan las distancias de cada individuo al cluster cercano (como el número de “no coincidencias” respecto al centroide del cluster)
- Se asigna a cada individuo al cluster con más similaridad

(b)

	Locus										Cluster Distance			
Indiv	1	2	3	4	5	6	7	8	9	10	1	2	3	4
1	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB	0	6	5	5
2	AB	BB	BB	AB	BB	BB	AB	AB	BB	AB	8	8	6	6
3	BB	BB	AA	AA	AB	AB	AA	AB	BB	AB	8	5	7	8
4	AB	BB	AB	AB	AB	AB	BB	AB	AA	AB	7	6	7	7
5	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB	6	0	5	10
6	BB	AB	AB	AB	BB	BB	AB	AA	AB	AB	4	7	5	8
7	BB	BB	BB	BB	AB	AB	AA	AB	BB	AB	8	6	5	6
8	AB	BB	AB	AB	AA	AA	AB	BB	AB	BB	4	7	8	4
9	BB	AA	AB	AB	BB	AB	AB	AA	AB	AB	5	7	8	8
10	AB	BB	AB	BB	AB	AB	BB	AB	AB	AA	7	8	8	4
11	AA	BB	AA	AA	AA	AB	AA	AB	AB	AB	9	5	9	8
12	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB	5	7	0	5
13	AB	BB	AB	AA	AB	AB	BB	AB	AA	AA	8	7	8	6
14	BB	AA	AB	AB	BB	BB	AB	AA	AB	AB	5	7	6	8
15	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA	5	10	5	0

(c)

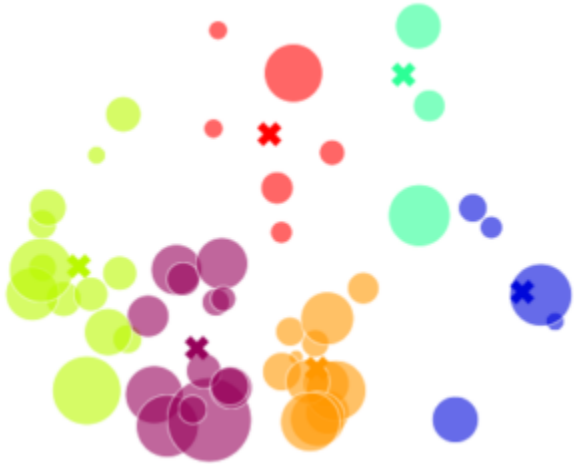
	Locus									
Cluster	1	2	3	4	5	6	7	8	9	10
1	BB	**AA**	AB	AB	**BB**	AA	AB	**AA**	AB	**AB**
2	BB	**BB**	AA	**AA**	AA	AB	AA	AB	AA	**AB**
3	BB	**BB**	BB	BB	AB	BB	AB	**AB**	**BB**	AB
4	AB	BB	**AB**	BB	AB	**AB**	**BB**	**AB**	AB	AA

- El siguiente paso es actualizar los centroides del clúster en función de las personas ahora asignadas a los clústeres.
- El genotipo modal entre los individuos asignados a un grupo se convierte en el genotipo centroide en ese locus. Los genotipos que cambiaron de la inicialización a la actualización se muestran en negrita.
- Se vuelven a calcular las disimilaridades de cada individuo con los nuevos centroides y se vuelven a asignar. El único caso que se reasigna es el 4.
- Se repite hasta que ningún individuo cambia la pertenencia a su cluster.

(d)

	Locus										Cluster Distance			
Indiv	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**
1	BB	AB	AB	AB	AB	AA	AB	BB	AB	BB	**4**	9	7	7
2	AB	BB	BB	AB	BB	BB	AB	AB	BB	AB	6	7	**3**	7
3	BB	BB	AA	AA	AB	AB	AA	AB	BB	AB	8	**3**	4	6
4	AB	BB	AB	AB	AB	AB	BB	AB	AA	AB	7	6	6	**3**
5	BB	AB	AA	AB	AA	AB	AA	AB	AA	BB	8	**3**	8	8
6	BB	AB	AB	AB	BB	BB	AB	AA	AB	AB	**2**	8	6	8
7	BB	BB	BB	BB	AB	AB	AA	AB	BB	AB	8	4	**2**	5
8	AB	BB	AB	AB	AA	AA	AB	BB	AB	BB	**5**	8	8	6
9	BB	AA	AB	AB	BB	AB	AB	AA	AB	AB	**1**	7	7	7
10	AB	BB	AB	BB	AB	AB	BB	AB	AB	AA	8	7	6	**0**
11	AA	BB	AA	AA	AA	AB	AA	AB	AB	AB	8	**2**	7	6
12	BB	AB	BB	BB	AB	BB	AB	BB	AA	AB	7	7	**3**	8
13	AB	BB	AB	AA	AB	AB	BB	AB	AA	AA	9	5	7	**2**
14	BB	AA	AB	AB	BB	BB	AB	AA	AB	AB	**1**	8	6	8
15	AB	BB	BB	BB	AB	AA	AB	BB	AB	AA	7	9	5	**4**

CONCLUSIONES



- El proceso de Clustering puede ser muy útil para intentar entender un set de datos no clasificado (no supervisado).
- Según el algoritmo utilizado para el proceso se puede obtener una clasificación categórica de la clase (K-Means) o probabilística (Fuzzy C-means)
- El resultado de K-Means depende de las condiciones iniciales para los centroides y la cantidad de k clusters elegidos (convergencia local).
- En el caso de encontrarse con variables categóricas puede recurrirse al algoritmo K-Modes
- Graficar los datos de a pares de variables nos ayuda a elegir en número de k clusters.
- Silhouette score es un buena métrica para medir la calidad de los clusters obtenidos.