CURSOS PIT

CTIC UNI

Fundamentos de Inteligencia Artificial

CURSOS PIT

CTIC UNI

PhD Wester Edison Zela Moraya

PhD en Computer Science – Inteligencia Artificial por la Universidad Politécnica de Madrid. Master en Ingeniería de Software por la Universidad de Oxford. Master en Análisis Financiero y Económico por la Universidad Complutense de Madrid. Ingeniero de Sistemas de la UNI.

Amplia experiencia profesional en Transformación Digital, Machine Learning, RPAs, Data Science, Metodologías Ágiles, Microservices, gestión económica de proyectos. Docente de Inteligencia Artificial en la Universidad Nacional de Ingeniería.

Director de TI en empresas en Peru y Europa

Consultor de IA y Datos en la SGTD en la PCM

Miembro del AI Connect Program (US Department y Atlantic Council)

Creador de Troomes.com

Temas – Sesion 6

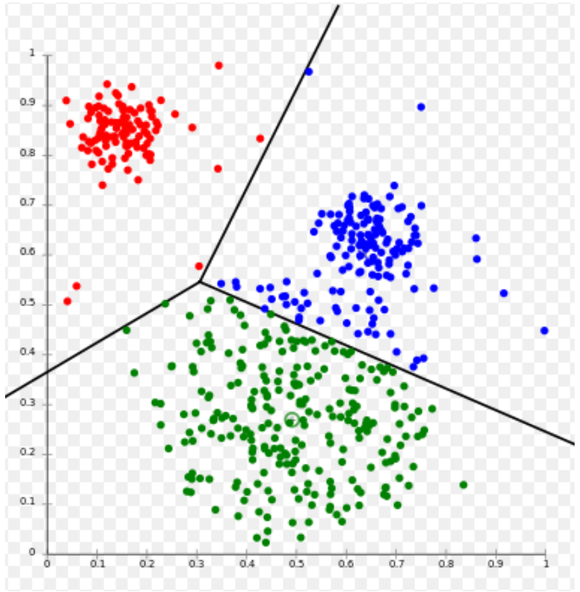
- Aprendizaje No Supervisado
  - K-MEANS

Aprendizaje No Supervisado

- En el aprendizaje no supervisado, no hay una etiqueta ni un valor objetivo para los datos.
- El problema del aprendizaje no supervisado es intentar encontrar una estructura oculta en datos sin etiquetar.
- No hay señal de error o de recompensa para evaluar una posible solución.
- Algunas tareas:
  - **Agrupación:** una tarea en la que agrupamos elementos similares o dividimos un gran conjunto de datos en conjuntos de datos más pequeños de cierta similitud

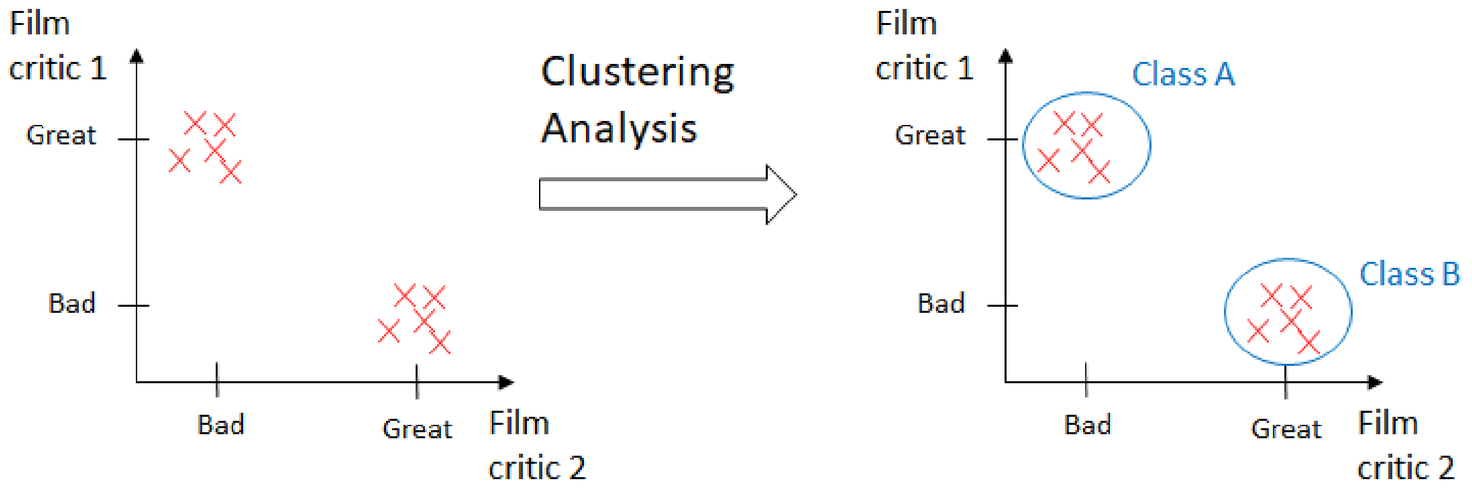
Agrupamiento

- Es la tarea de agrupar un conjunto de objetos de tal manera que los objetos de un mismo grupo (llamado cluster) sean más similares (en un sentido u otro) entre sí que a los de otros grupos (clusters).
- Está clasificado como un método de aprendizaje no supervisado, lo que significa que no hay datos de entrenamiento previo de los que aprender.
- No hay atributo "Clase"



Clustering o Agrupamiento

De una manera visual: Imagina que tenemos un conjunto de datos de películas y queremos clasificarlas. Tenemos las siguientes reseñas de películas de los critico de peliculas 1 y 2, quienes evaluar las películas como buenas o malas.



El modelo de aprendizaje automático podrá inferir que hay dos clases diferentes sin saber nada más de los datos.

Algoritmos de Agrupamiento

Algunos de los algoritmos de agrupación más comunes son:

- KMeans
- EM (Expectation Maximization)
- XMEANS
- Redes SOM o Mapas Autoorganizados
- Clusterización Jerárquica
- Density Based Scan Clustering (DBSCAN)
- Modelo de Agrupamiento Gaussiano

Casos de Uso de Métodos de Agrupamiento Sector Privado

Negocios y marketing

- Alcance de mercado para dividir la población general de clientes en segmentos de mercado para comprender las relaciones entre grupos de clientes / clientes potenciales. Ejemplo: Agrupar a clientes de un empresa de retail en base a su patrón de compras para realizar promociones dirigidas: clientes “jóvenes deportistas”, “clientes con hijos niños”, etc.

- Agrupación de artículos de compras disponibles en la web en un conjunto de productos únicos. Ejemplo: Sistemas de recomendación para que cuando un cliente compre (o busque) un artículo se le recomiende otros artículos que los clientes suelen comprar junto con el primer artículo.

Casos de Uso de Métodos de Agrupamiento Sector Privado

En Redes Sociales

- En el estudio del análisis de redes sociales, la agrupación puede reconocer comunidades dentro de grandes grupos de personas con características similares. Esto puede ser muy útil para las empresas o organizaciones como instituciones públicas que quieran realizar campañas de difusión o de marketing específicas.
- Ejemplo: Identificación de comunidades de personas en una red social que tienen ciertas preferencias musicales.

Casos de Uso de Métodos de Agrupamiento Sector Público

Lucha contra el crimen

El análisis de clusters se puede utilizar para identificar áreas donde hay una mayor incidencia de un tipo particular de delito, identificando las distintas áreas donde se han cometido delitos similares durante un período de tiempo.
El objetivo es orientar los recursos necesarios para algunos tipos de delitos.

Casos de Uso de Métodos de Agrupamiento Sector Público

En el Sector Educativo

Identificar grupos de escuelas o estudiantes con propiedades similares para realizar alguna actuación en este grupo para ayudarlos en su desempeño escolar.

Por ejemplo:

- Identificar grupos de estudiantes que van a tener un problema de desempeño durante el año escolar
- Identificar grupos de estudiantes con problemas de dejar la escuela

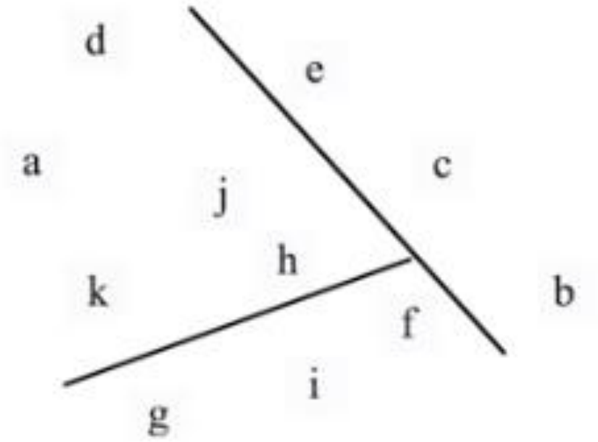
Casos de Uso de Métodos de Agrupamiento
Sector Público

En la Agricultura

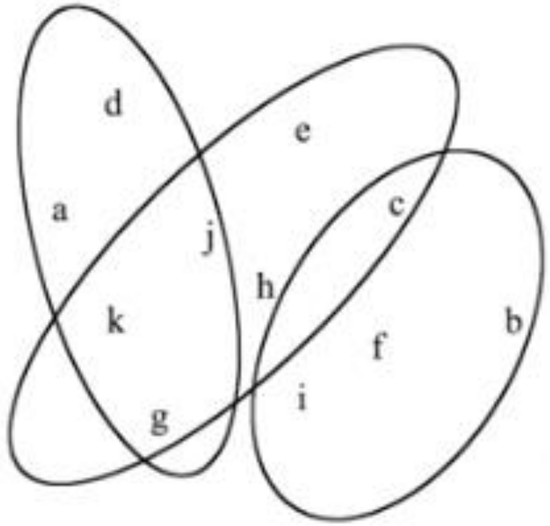
- Identificar grupos de suelos para mejorar el cultivo de ciertos vegetales como la papa o el tomate, y que no puedan ser impactados por la bacteria Ralstonia Solanacearum. El crecimiento de la población de esta bacteria depende del valor del pH del suelo, humedad, la concentración de fósforo, soluble y el contenido total de boro, cadmio y aluminio y otros.

Tipos de Clusters

1. Disjoint sets



2. Overlapping sets

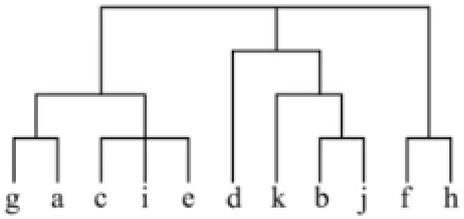


Tipos de Clusters

3. Probabilistic clusters

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
…			

4. Hierarchical clusters



Calculo de Número de Clusters

- El método de la regla del pulgar

K = raíz cuadrada (objetos / 2)
es decir: 120 objetos -> número de clusters: 7

- Otros metodos:
  - El método del codo, basado en el porcentaje de datos que cubren los clústeres
  - Método de validación cruzada, dividiendo el conjunto de datos en diferentes particiones y comparar errores de suma de cuadrados.

Kmeans (Macqueen, 1967)

- Dado un conjunto de observaciones (x_1, x_2,…, x_n), donde cada observación es un vector real d-dimensional, la agrupación de k-medias tiene como objetivo dividir las n observaciones en k (≤ n) conjuntos S = {S_1, S_2, …, S_k} para minimizar la suma de cuadrados dentro del grupo (WCSS) (suma de las funciones de distancia de cada punto del grupo al centro K). En otras palabras, su objetivo es encontrar:

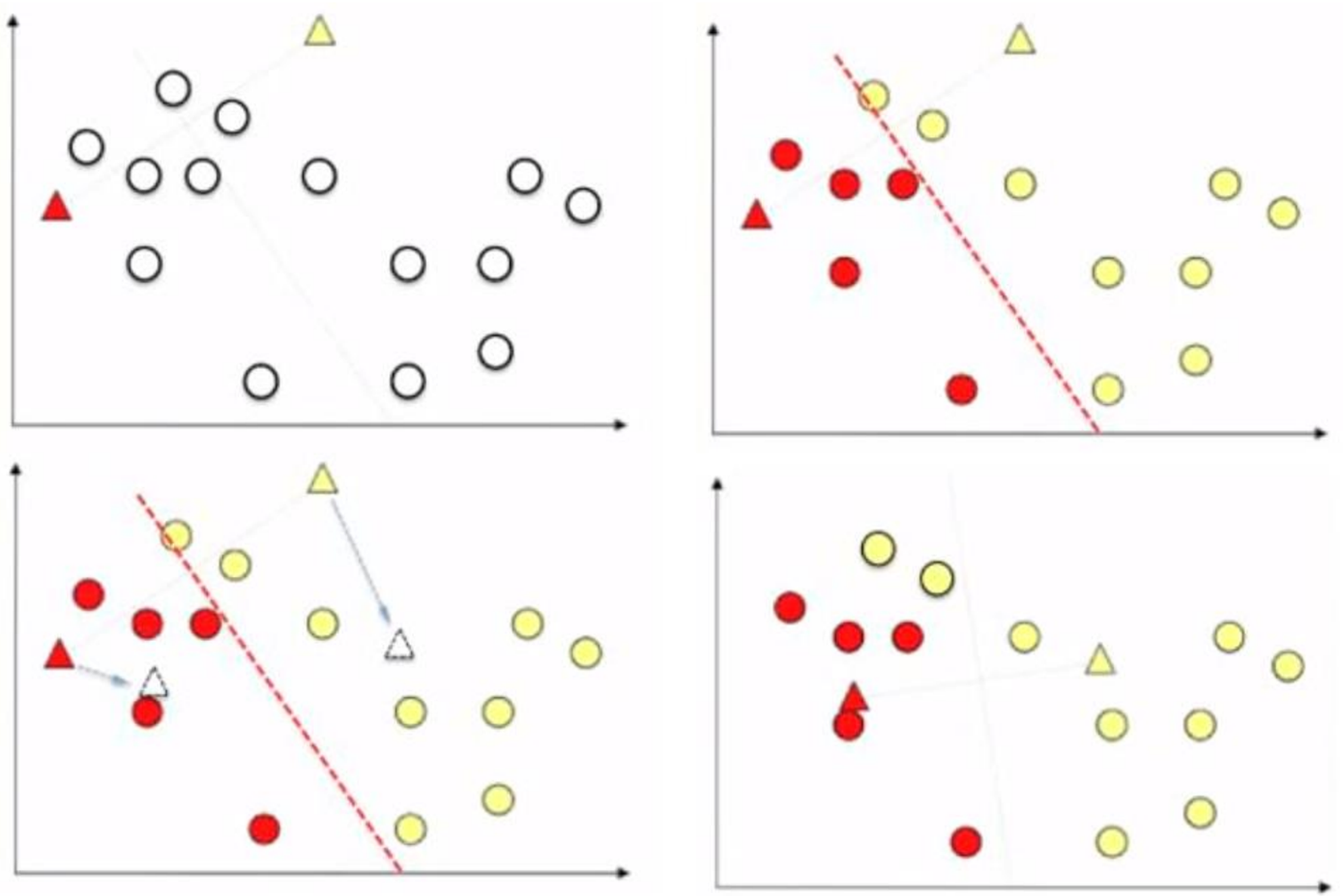
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

Donde $\boldsymbol{\mu}_i$ es la media de los puntos en S_i.

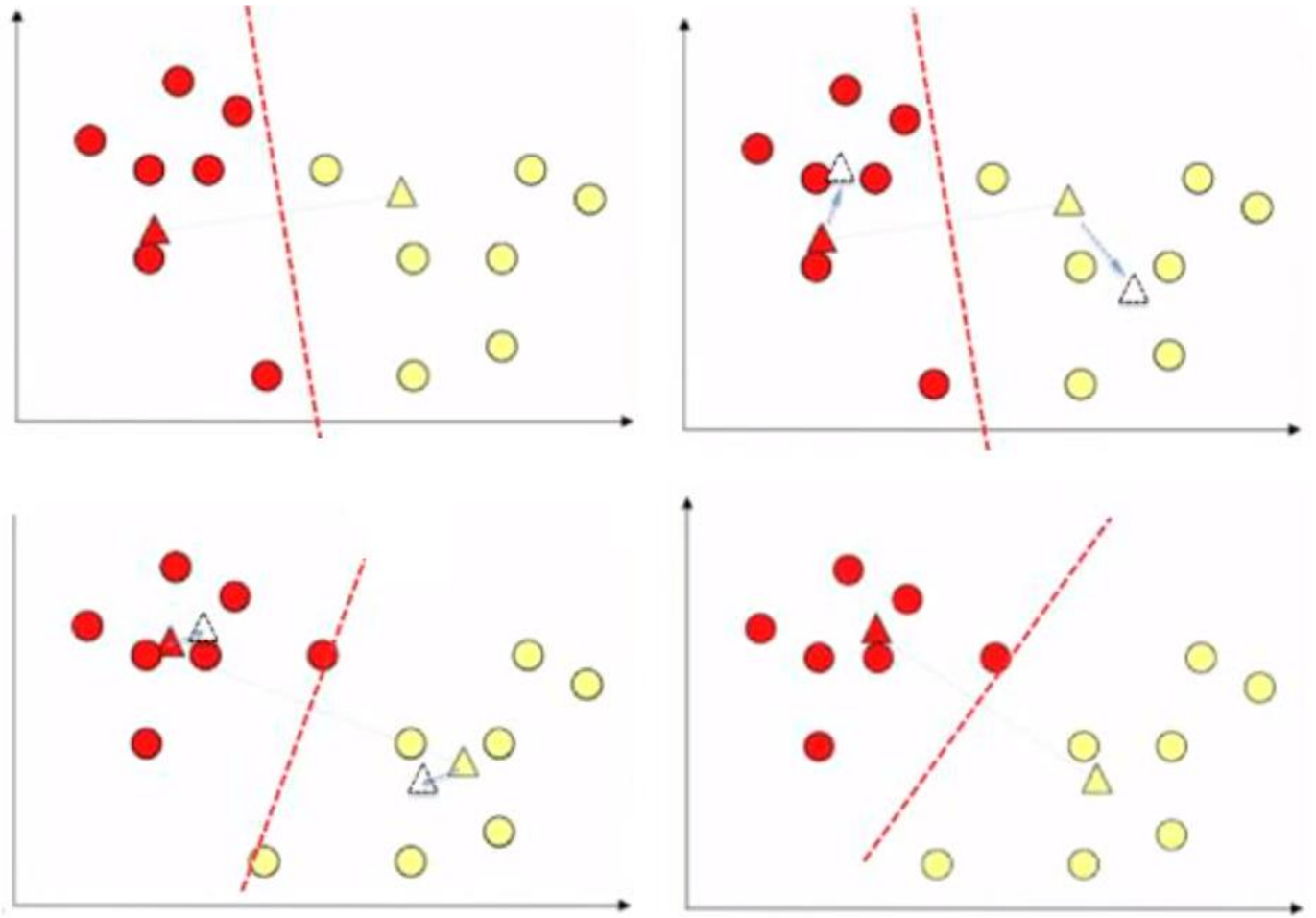
Algoritmo Kmeans

- Input: K, conjunto de puntos $x_1 \dots x_n$
- Poner los centroides $c_1 \dots c_k$ en localizaciones aleatoreas
- Repetir hasta converger:
  - Por cada punto x_i:
    - Buscar el centroide mas cercano c_j (distancia .i.e distancia euclideana)
    - Asignar el punto x_i al cluster c_j
  - Por cada cluster j = 1 … k:
    - Nuevo centroide c_j = promedio de todos lo puntos xi asignados al cluster j en los pasos previos
- Parar cuando los clusters asignados no cambian

Algoritmo Kmeans



Algoritmo Kmeans



Clustering in Weka

Lineas de Comandos

- Convert .CSV file .ARFF
  - java –cp “path of weka/weka.jar” weka.core.converters.CSVLoader kmeansdata.csv > kmeansdata.arff
- First Run
  - Java –cp “path of weka/weka.jar” weka.clusterers.SimpleKMeans –t kmeansdata.arff
  - Analyze the results
- Refining based on results
- Calculate the number of clusters
  - Java –cp “path of weka/weka.jar” weka.clusterers.SimpleKMeans –t kmeansdata.arff –N 6 –S 42
  - Analyze results

Clustering in Weka

- Abra Weka y cargue diabetes.arff (elimine el atributo Clase)

- Ir a la pestaña Clúster
- Seleccione SimpleKMeans:
  - -N determina la cantidad de clústeres que SimpleKMeans va a crear
  - -A es la función de distancia utilizada. Tiene como valor predeterminado la distancia euclidiana y utiliza todo el rango de valores como su rango para actuar (-R primero-último).
  - La bandera –l define el número de iteraciones que hace k-means para definir el clúster.
  - -S es una semilla de número aleatorio. Puede ser cualquier valor que desee
  - Seleccione el número de clústeres 2,3,4

Clustering in Weka

- Name de Cluster
  - Java –cp “path of weka/weka.jar” weka.clusterers.SimpleKMeans –t kmeansdata.arff –N 6 –S 42 –p 0
  - P=0 row of the cluster
  - P=1 position x
  - P=2 position y

Método del Codo

El método del codo considera el WSS total como una función del número de clústeres: se debe elegir un número de clústeres para que agregar otro clúster no mejore mucho mejor el WSS total.
El número óptimo de clústeres se puede definir de la siguiente manera:

- Calcule el algoritmo de agrupación en clústeres (p. Ej., Agrupación de k-medias) para diferentes valores de k. Por ejemplo, variando k de 1 a 10 grupos.
- Para cada k, calcule la suma total del cuadrado dentro del grupo (wss).
- Trace la curva de wss de acuerdo con el número de grupos k.
- La ubicación de una curva (rodilla) en la parcela se considera generalmente como un indicador del número apropiado de agrupaciones.

Elbow Method

cluster 1			cluster 2		
point	x	y	point	x	y
1	7	3	2	4	5
5	9	7	3	2	4
6	6	8	4	0	1
mean	7.33	6		2	3.33

$$WSS[1] = (7 - 7.33)^2 + (9 - 7.33)^2 + (6 - 7.33)^2 + (3 - 6)^2 + (7 - 6)^2 + (8 - 6)^2 = 18.67$$

$$WSS[2] = (4 - 2)^2 + (2 - 2)^2 + (0 - 2)^2 + (5 - 3.33)^2 + (4 - 3.33)^2 + (1 - 3.33)^2 = 16.67$$

$$\text{WSS} = \text{WSS}[1] + \text{WSS}[2] = 18.67 + 16.67 = 35.34$$

R y R-Studio

Install R
https://cran.r-project.org/mirrors.html

https://www.rstudio.com/products/rstudio/download/

R – Numero de Clases

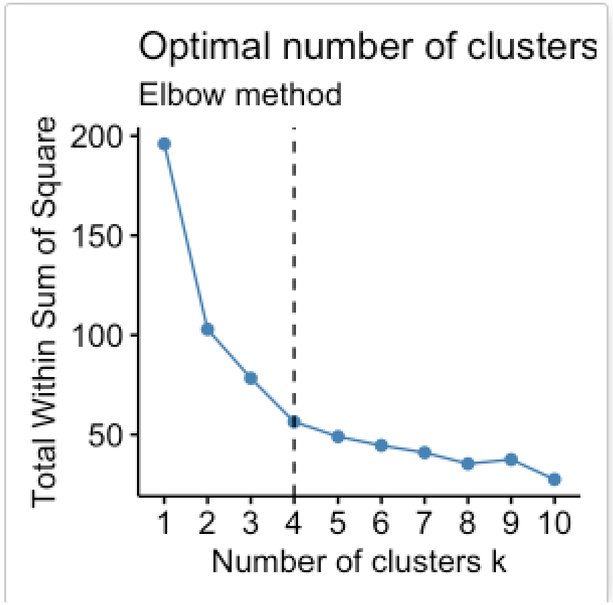
```
pkgs <- c("factoextra", "NbClust")
install.packages(pkgs)
library(factoextra)
library(NbClust)
# Standardize the data
df <- scale(USArrests)
head(df)
##            Murder Assault UrbanPop     Rape
## Alabama    1.2426   0.783   -0.521 -0.00342
```

R – Number of Clusters

 # Elbow method

fviz_nbclust(df, kmeans, method = "wss") + geom_vline(xintercept = 4, linetype = 2)+ labs(subtitle = "Elbow method")

R – Number of Clusters



Ejercicio Grupar

Archivo: debiates.csv

- **Preg: Número de embarazos**
- **Plas: Concentración de glucosa en plasma a 2 horas en una prueba de tolerancia a la glucosa oral**
- **Pres: Presión arterial diastólica (mm Hg):** Cuando su corazón está en reposo, entre latidos, su presión arterial baja
- Skin: Espesor del pliegue cutáneo del tríceps (mm)
- **Insu: Insulina sérica de 2 horas (mu U / ml):** Es una prueba que mide cuánta insulina tiene en la sangre.
- **mass: Índice de masa corporal (peso en kg** / (altura en m) ^ 2)
- **Pedi: Función pedigrí de la diabetes:** Una función que califica la probabilidad de diabetes según los antecedentes familiares.
- **Age: Edad** (años)

R – Number of Clusters

```
# Ejemplo diabetes
# Import diabetes.csv (file -> Import Dataset)
train <- diabetes
train[which(train=="?",arr.ind=TRUE)]<-NA   # Not Available for train==“?”
train <- data.frame(lapply(train,as.numeric))  # lapply apply to a list
train <- as.matrix(train[-length(train)]) # eliminate las column
df <- scale(train)
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +  geom_vline(xintercept = 4, linetype = 2)+ labs(subtitle = "Elbow method")
```

Algoritmo EM (Expectation Maximization)

El algoritmo de maximización de expectativas, o algoritmo EM, es un enfoque para la estimación de máxima verosimilitud en presencia de variables latentes.

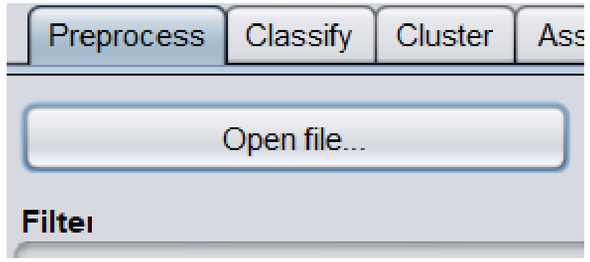
El algoritmo EM es un enfoque iterativo que alterna entre dos modos. El primer modo intenta estimar las variables latentes o faltantes, llamado paso de estimación o paso E. El segundo modo intenta optimizar los parámetros del modelo para explicar mejor los datos, llamado paso de maximización o paso M.

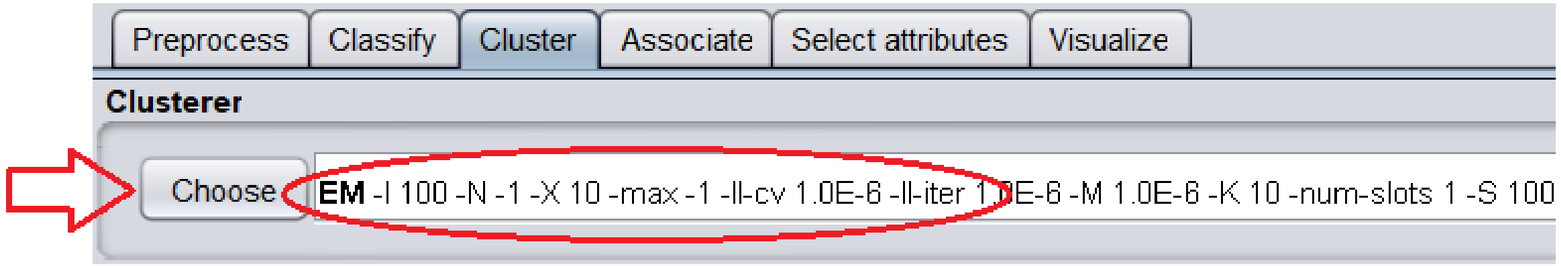
Paso E: Estima las variables que faltan en el conjunto de datos.
Paso M: Maximizar los parámetros del modelo en presencia de los datos.

Permite encontrar el número de clusters óptimo según el dataset.

Algoritmo EM (Expectation Maximization)

Entrar a Weka y cargar el archivo diabetes.csv (el dataset no tiene la variable clase)
Ir al tab **Cluster** selecciona **EM (Expectation Maximization)**:





Opciones:

maximunNumbersOfClusters : El número máximo de clústeres a considerar durante la validación cruzada para seleccionar el mejor número de clústeres. -1 para que el algoritmo busque el número de clusters.

numClusters: Establecer el número de agrupaciones. -1 para seleccionar el número de clústeres automáticamente mediante validación cruzada.

Algoritmo EM (Expectation Maximization)

Attribute	0 (0.05)	1 (0.22)	2 (0.35)	3 (0.39)
preg				
mean	3.3354	1.881	6.9455	2.2714
std. dev.	3.6224	1.7663	2.9726	1.7547
plas				
mean	121.5761	138.0569	129.5271	104.6522
std. dev.	29.3353	30.2865	32.8379	24.8844
pres				
mean	0.6367	72.1958	77.5049	68.3249
std. dev.	5.0425	14.2219	11.3011	11.1689
skin				
mean	1.7564	33.5342	16.6382	17.5289
std. dev.	7.4348	10.2412	17.6306	13.0582
insu				
mean	1.261	196.8335	57.5172	41.0342
std. dev.	11.2271	154.668	104.8937	50.5543
mass				
mean	26.0311	36.395	32.7032	29.6414
std. dev.	16.683	8.0773	7.1002	5.7784
pedi				
mean	0.4029	0.6558	0.4683	0.3889
std. dev.	0.2455	0.4636	0.2986	0.2194
age				
mean	31.9307	29.0218	45.8007	25.3952
std. dev.	11.0512	7.794	9.8927	4.3893

```
Time taken to build model (percentage split) : 1.63 seconds

Clustered Instances

0       11 (  5%)
1       43 ( 19%)
2       76 ( 33%)
3      101 ( 44%)

Log likelihood: -28.09559
```

Resultados:

- El número de clusters óptimo sugerido por EM son 4 clusters para el dataset diabetes.csv
- En el cluster 0 caen el 5% de los datos con los que fueron probados.
- En el cluster 3 caen el 44% de los datos con los que fueron probados.

Clustering in Weka

- Analize Resultado
  - Información de la ejecución
  - Clusters Finales
  - Instancias en los clusters

- Visualizar datos

```
Final cluster centroids:
                           Cluster#
Attribute    Full Data           0           1
               (537.0)     (195.0)     (342.0)
=============================================
preg             3.8901           7       2.117
plas           121.4227    130.8769    116.0322
pres            69.3296     76.0462        65.5
skin            19.9441     17.1231     21.5526
insu             78.622     66.8051     85.3596
mass            32.0158     32.6415     31.6591
pedi             0.4751      0.4754       0.475
age             33.6965     46.4256     26.4386



Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0       66 ( 29%)
1      165 ( 71%)
```

Ejercicio

- Con el archivo de diabettes.csv
- Crear clusters para k=2, k=3, k=4
- Ejecutar el algoritmo EM
- Responder las siguientes preguntas:
- Cual es la K que representa una mejor los grupos personas con diabetes?

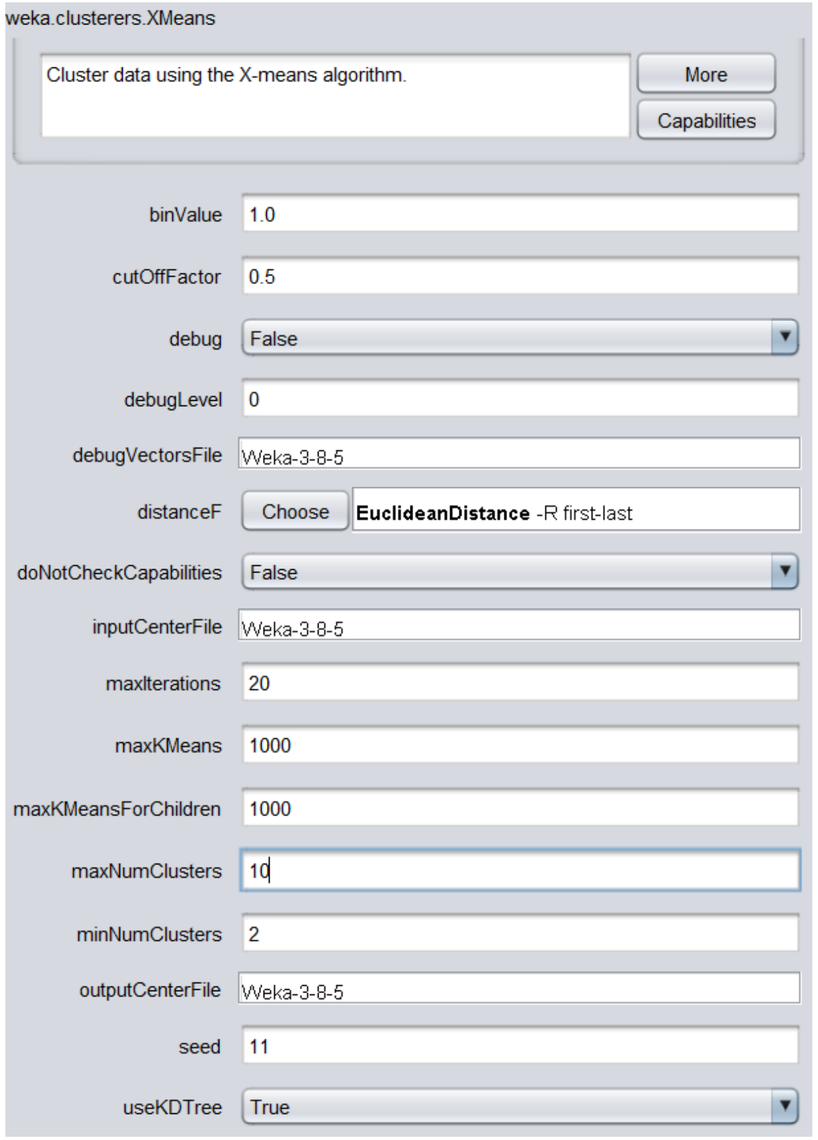
XMEANS

- Pero en esta situación, esta determinación del punto central del grupo inicial es la debilidad del algoritmo K-Means. Esto se debe a que no se utiliza ningún enfoque para seleccionar y determinar el punto central del clúster. El punto central del grupo se elige de forma arbitraria o aleatoria de un conjunto de datos. Los resultados de agrupamiento del algoritmo K-Means a menudo no son óptimos ni óptimos en todos los experimentos realizados. Por tanto, se puede decir que los buenos y malos resultados del agrupamiento dependen del punto central del cúmulo o centroide inicial.
- Holísticamente, K-means sufre de las siguientes limitaciones:
  - K-means es lento y se escala mal con respecto al tiempo que lleva completar cada iteración.
  - El número de grupos 'K' debe ser predeterminado y proporcionado por el usuario.
  - Cuando se limita a ejecutarse con un valor fijo de K, encuentra empíricamente peores óptimos locales que cuando puede alterar dinámicamente K.
  - La solución para los dos primeros problemas y un remedio parcial para el tercero es X-means.

XMEANS

- X-means entra en acción después de cada ejecución de K-means, tomando decisiones locales sobre qué subconjunto de los centroides actuales deben dividirse para lograr un mejor ajuste. La decisión de división se realiza calculando el criterio de información bayesiano (BIC).
- X-Means funciona aplicando alternativamente dos operaciones: el algoritmo K-Means (mejorar parámetros) para detectar de manera óptima los grupos para un valor elegido de k y la división de grupos (mejorar estructura) para optimizar el valor de k de acuerdo con el criterio de información . En este método, el valor real de K se estima de una manera que no se controla y solo se basa en el conjunto de datos. Kmax y Kmin como los límites superior e inferior para los valores posibles de X. En el primer paso de la agrupación de medias X, sepa que actualmente X = Xmin, X significa encontrar la estructura inicial y el centroide. En el siguiente paso, cada clúster de la estructura estimada se trata como el clúster principal, que se puede dividir en dos grupos.

XMEANS - Weka


weka.clusterers.XMeans
Cluster data using the X-means algorithm.
More
Capabilities
binValue
1.0
cutOffFactor
0.5
debug
False
debugLevel
0
debugVectorsFile
Weka-3-8-5
distanceF
Choose
EuclideanDistance -R first-last
doNotCheckCapabilities
False
inputCenterFile
Weka-3-8-5
maxIterations
20
maxKMeans
1000
maxKMeansForChildren
1000
maxNumClusters
10
minNumClusters
2
outputCenterFile
Weka-3-8-5
seed
11
useKDTree
True

XMEANS

```
XMeans
======
Requested iterations             : 20
Iterations performed             : 1
Splits prepared                  : 2
Splits performed                 : 1
Cutoff factor                    : 0.5
Percentage of splits accepted
by cutoff factor                 : 100 %
------
Cutoff factor                    : 0.5
------

Cluster centers                  : 2 centers

Cluster 0
            2.0855457227138645 115.93805309734513 65.46017699115045 21.430678466076696 85.06489675516224 31.632448377581113 0.4749823008849557 26.37168141592
Cluster 1
            6.97979797979798 130.81313131313132 75.95454545454545 17.3989898989899 67.5909090909091 32.672222222222224 0.47532828282828266 46.23737373737374

Distortion: 206.510396
BIC-Value : 878.371163


Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0      165 ( 71%)
1       66 ( 29%)
```