Tema 4. Agrupación o Clustering.

4.1. K-means y otros métodos directos
4.2. Métodos jerárquicos
4.3. Mapas autoorganizados y métodos de malla.
4.4. Otros métodos.
4.5. Aplicaciones empresariales.

---

La **agrupación o clustering** es una tarea descriptiva que pretende como su nombre indica agrupar los individuos disponibles en grupos o clusters de comportamiento similar. A diferencia de la tarea de ***clasificación*** y las distintas técnicas utilizadas en ella, en el análisis cluster ( en sentido estricto) y en otras técnicas de agrupación, los grupos son desconocidos a priori y son precisamente lo que queremos determinar.

Así pues, el objetivo es obtener agrupaciones (clusterings, o clusters), teniendo, por lo tanto, el análisis un marcado carácter exploratorio.

Se trata, fundamentalmente, de resolver el siguiente problema: Dado un conjunto de **N individuos** caracterizados por la información de **n variables** $X_j$ , (j = 1,2,..., n), nos planteamos el reto de ser capaces de clasificarlos de manera que los individuos pertenecientes a un grupo (cluster) (y siempre con respecto a la información disponible) sean tan similares entre sí como sea posible, siendo los distintos grupos entre ellos tan disimilares como sea posible.

Como puede comprenderse fácilmente el análisis cluster tiene una extraordinaria importancia en la investigación científica, en cualquier rama del saber. Téngase presente que la clasificación, en el sentido de taxonomía, es uno de los objetivos fundamentales de la ciencia .Y en la medida en que el análisis cluster nos proporciona los medios técnicos para realizarla, se nos hará imprescindible en cualquier investigación.

Ya desde Linneo, las clasificaciones y taxonomías fueron piezas clave en las investigaciones biológicas, y, en consecuencia, no puede resultarnos extraño que haya sido en los entornos de este tipo de ciencias donde hayan surgido las técnicas estadísticas tradicionales del análisis cluster. Los trabajos de Sokal y Sneath , marcan el inicio de las técnicas de clusterización, que, poco a poco, han ido extendiendo sus aplicaciones a todos los ámbitos científicos.

Pretendemos encontrar un conjunto de grupos a los que ir asignando los distintos individuos por algún criterio de **homogeneidad**. Por lo tanto, se hace imprescindible definir una medida de **similitud** o bien de **divergencia (o distancia)** para ir clasificando a los individuos en unos u otros grupos.

También hay que hacer algunas consideraciones sobre el proceso de agrupación. Podemos estar interesados en obtener **un número predeterminado de grupos** (digamos, 4 grupos, por ejemplo) o una solución consistente en **un rango de grupos** ( por ejemplo entre 2 y 8 grupos). Si optamos por un rango de grupso como solución puede que pretendamos una agrupación **jerárquica** que conserva los grupos determinados con anterioridad. Y dentro del cluster jerárquico podemos optar por un proceso **aglomerativo** ( partimos de más grupos y vamos reduciendo el número) o **divisivos** ( partimos de un solo grupo y vamos dividiendo.

Así pues el análisis cluster tradicional suele distinguir fundamentalmente entre métodos directos con un número predeterminado de grupos y métodos jerárquicos. De entre los jerárquicos los más habituales son los aglomerativos que suponen un proceso que parte de tantos grupos como individuos que va secuencialmente agrupando hasta la obtención de un único grupo. De entre todas las posibles soluciones elegiremos una concreta o un rango de ellas.

Los procesos jerárquicos completos puede estructurarse de acuerdo con el siguiente esquema:

- Partimos de un conjunto de N individuos de los que se dispone de una información cifrada por un conjunto de n variables (una matriz de datos de N individuos × n variables).
- Establecemos un criterio de similaridad o de distancia para poder determinar: Una matriz de similaridades que nos permita relacionar la semejanza de los individuos entre sí (matriz de N individuos x N individuos).
- Escogemos un algoritmo de clasificación para determinar la estructura de agrupación de los individuos.
- Especificamos esa estructura mediante diagramas arbóreos o dendogramas u otros gráficos.

Los no jerárquicos también debe fijar un criterio de similaridad o distancia , un número fijo de grupos a obtener y un procedimiento de asignación/reasignación de los individuos a los grupos.

Además de análisis cluster jerárquico y no jerárquico ( que son los métodos estadísticos tradicionales) en los últimos decenios han surgido otras técnicas inspiradas en distintos procesos de aprendizaje no supervisado ( que en definitiva es lo que es una clusterización) . Destacan los mapas auto-organizados de Kohonen (SOM) y otros métodos de malla o basados en ejemplares.

**Distancias y similitudes**.

De una u otra forma todos los métodos de agrupación implican alguna medida de similitud entre individuos y habitualmente entre grupos ( y entre individuos y grupos, que puede verse como caso particular de "entre grupos").Lo normal es la utilización de una medida de distancia ( concepto opuesto a similaridad que nos informe de los lejos que están los individuos o los grupos en función de lo diferentes que resulten sus valores en las variables consideradas.

Dados dos puntos del espacio, ( dados dos indiviuod de los que se dispone información de n variables ) ,toda distancia debe verificar, al menos, las siguientes propiedades:

(P.1) d(x,y) > 0 (no negatividad)
(P.2) d(x,y) = 0
(P.3) d(x,y) = d(x,y) (simetría)

Entre las distancias más habituales tenemos:

D. Euclídea $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

D.Manhattan $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$

D.Chebyshev $d(x,y) = \max_{i=1,2...,n} |x_i - y_i|$

D.Coseno $d(x,y) = arc\cos\left(\frac{x^T y}{\|x\|.\|y\|}\right)$

D. Euclidea normalizada $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$

siendo $S^{-1}$una matriz con (1/desv.tipicas) en la diagonal

D. Mahalanobis $d(x,y) = \sqrt{(x-y)^T V^{-1}(x-y)}$

siendo V la matriz de varianzas y covarianzas

También , en el caso de que estemos en presencia de atributos nominales pueden usarse otras distancias específicas como la distancia delta :

$$d(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i, y_i)$$

donde delta se define como $\delta$(a,b)=1 si y sólo si a=b y es cero en caso contrario

**Distancias entre grupos.**

En algunos métodos de clustering es necesario contar con algún criterio para definir la distancia ( o la similitud ) , no sólo entre individuos, sino también entre grupos o clusters de individuos . ( La distancia entre un individuo y un grupo puede considerarse un caso particular de distancia entre grupos.

DISTANCIA MINIMA (NEAREST NEIGHBOUR DISTANCE)

Podemos definir la distancia entre un grupo y un individuo como la menor de las distancias entre los individuos del grupo y el individuo exterior considerado.

Si llamamos I al grupo formado por los individuos ($i_1$ , $i_2$ ,... $i_i$ ) y j al individuo exterior, definiremos, entonces, la distancia entre I y j como: D (I,j) = min D (i , j)

Análogamente, siguiendo este criterio, puede definirse la distancia entre dos grupos I = { $i_1$ , $i_2$ ,... $i_i$ } y J = { $j_1$ , $j_2$ ,... $j_j$ }, como la mínima de las distancias entre un individuo de I y otro de J: D (I,J) = min D (i ,j )

Como veremos, la distancia mínima será la utilizada en el algoritmo jerárquico de clasificación conocido como método de la distancia mínima o single linkage.

DISTANCIA MAXIMA (FURTHEST NEIGHBOUR DISTANCE)

También podemos definir la distancia entre un grupo I y un individuo j como el valor máximo de las distancias entre j y los individuos de I; esto es:D (I,j) = max D (i , j)

Y, la distancia entre dos grupos, I y J, análogamente, sería : D (I,J) = max D (i , j )

Esta distancia ser la utilizada en el método o algoritmo de la distancia máxima o complete linkage.

DISTANCIA ENTRE CENTROIDES

También se puede definir la distancia entre el grupo I y el individuo j como la distancia entre el centroide o centro de gravedad de I y j. Si ***i*** es el centro de gravedad de I, tendremos que: D(I,j) = D (***i***, j)

Y de la misma manera la distancia entre dos grupos I y J nos vendrá dada por la distancia entre sus centroides:

D (I,J) = D (***i***, ***j*** )

Estas y otras definiciones de distancias entre grupos serán utilizadas como criterios a emplear en los distintos algoritmos de clasificación

**4.1.K-Means y otros métodos directos**

El método de K-means ( o K-medias) es un procedimiento iterativo mediante el cual cada observación se asigna al cluster más cercano. Inicialmente se establecen de forma aleatoria las ubicaciones centrales de los grupos o *centroides* .Se calculan las distancias de los distintos individuos a estos centroides y se agrupa cada uno en el cluster más cercano.Una vez las observaciones han sido clasificadas en el grupo cuyo centroide es más cercano ,se recalculan los centroides y se repite el proceso de reagrupación. Hasta que los centroides se estabilizan.

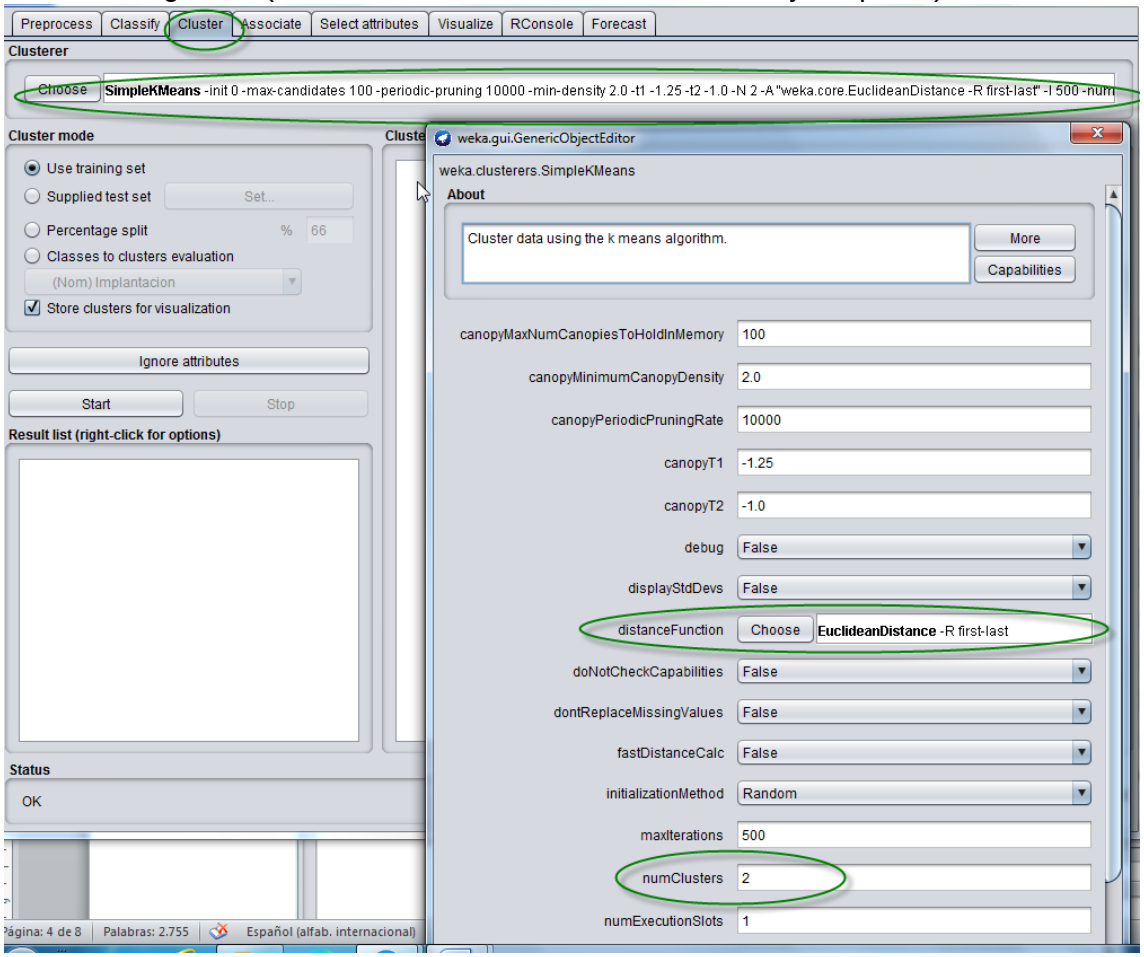
En R la función *KMclus* nos permite llevar a cabo a un análisis cluster por K-medias 0 K-means. El argumento de la función es una matriz formada por las variables que se van utilizar . Aquí se presenta un script aplicado a los datos amValencia1.xlsx.

( https://www.uv.es/mlejarza/datamine/kmeans.R )

```
amValencia1 <- read_excel("datamining/datos/amValencia1.xlsx")
view(amValencia1) ; names(amValencia1)
 Z=model.matrix( ~-1+nortesur+Idependencia+Tnatalidad+Tmortalidad+Tcrecimiento+Tmigracion+
envejecimiento  +nlineastelef + nvehi+ indimas  +segvivienda+vivdesocu+ta  +taf + iacomercial + iafinan
+ iaindus +cenerdomest+cenergindus +formAcad + IparLabSer+ IparLabIndustria+ IparLabConstr,
amValencia1)
 row.names(Z)=amValencia1$muni
  KMclus <-  kmeans(Z, centers = 3,
               iter.max = 10 )
KMclus$size # Cluster Sizes
KMclus$centers # Cluster Centroids
KMclus$withinss # Within Cluster Sum of Squares
KMclus$tot.withinss # Total Within Sum of Squares
KMclus$betweenss # Between Cluster Sum of Squares
KMclus$cluster
amv<-cbind(KMclus$cluster,amValencia1)
```

En weka, en la pestaña de cluster, nos encontramos con el procedimiento SimpleKmeans que permite ejecutar este análisis. Entre sus principales opciones tenemos la determinación del número de cluster a construir ( por defecto 2) y la elección de la función de distancia a considerar . Si se elige la euclidea ( puede editarse) por defecto está normalizada ( descuenta el efecto de las unidades pero no el de la correlación)

No es posible utilizar la distancia de Mahalanobis, por lo que para descontar el efecto de la correlación es necesario trabajar con las componentes principales de las variables originales.( Esto también ocurrirá con los métodos jerárquicos)



**Otros métodos directos**

**Canopy**: se fija de manera aleatoria el primer centroide y, seguidamente, realiza una primera partición utilizando una métrica más sencilla, a nivel computacional, en la que se generan grupos en superposición que reciben el nombre de canopies. Esta inicialización divide el proceso de agrupación en dos etapas:

- Primera etapa: en este paso se utiliza la medida de distancia sencilla con el fin de crear canopies. Un canopy es simplemente un subgrupo de elementos que, según la métrica de similitud empleada en este paso, están dentro de un umbral de distancia desde un punto central. Significativamente, un elemento puede pertenecer a más de un canopy y cada elemento debe aparecer en al menos uno de ellos. Los canopies se crean con la intención de que los elementos que no aparezcan en ningún canopy en común estén lo suficientemente separados de manera que no podrán estar en el mismo cluster y, por ende, puedan ser candidatos a centroides. Puesto que la medida de distancia utilizada para crearlos es aproximada, puede que no se garantice esto, sino que permita que estos subgrupos se superpongan entre sí, eligiendo un umbral de distancia suficientemente grande y comprendiendo las propiedades de la distancia aproximada medida, podemos tener una garantía en algunos casos.
- Segunda etapa: en este punto se ejecuta el paso de clasificación del algoritmo K-Means, usando una métrica de distancia más exigente, pero con la restricción de que no se calcula la distancia entre aquellas instancias que no pertenecen al mismo canopy, esto es equivalente a suponer entre ellos una distancia infinita.

**Farthest first**: escoge un elemento de los datos al azar como primer centroide. A continuación, calcula la distancia entre cada uno de los elementos restantes y dicho centroide. Por último, define como nuevo centroide a aquel elemento que esté más alejado. Repite los cálculos de distancia y realiza las asignaciones hasta tener los k representantes de los clusters. Gracias al procedimiento que sigue farthest first son necesarias menos operaciones de reajuste de clusters y reasignación de centroides, esto brinda una mejora en cuanto a la velocidad de agrupamiento con respecto a la inicialización aleatoria.

## 4.2.Métodos jerárquicos ( aglomerativos)

En los métodos jerárquicos los individuos no se particionan en clusters de una sola vez, sino que se van haciendo particiones sucesivas a " distintos niveles de agregación o agrupamiento ".

Fundamentalmente, los métodos jerárquicos suelen subdividirse en métodos aglomerativos (ascendentes), que van sucesivamente fusionando grupos en cada paso; y métodos divisivos (descendentes), que van desglosando en grupos cada vez más pequeños el conjunto total de datos.

Nosotros desarrollaremos sólo métodos aglomerativos

Cabe concluir, por tanto, que la clusterización jerárquica produce taxones o clusters de diferentes niveles y estructurados de forma ordenada, para ser exactos, estableciendo una "jerarquía"; de ahí su nombre.

Establecer una clasificación jerárquica supone poder realizar una serie de particiones del conjunto de individuos total
$W = \{ i_1 , i_2 , ...,i_N \}$ ; de forma que existan particiones a distintos niveles que vayan agregando (o desagregando, si se trata de un método divisivo) a las particiones de los niveles inferiores .

La representación de la jerarquía de clusters obtenida suele llevarse a cabo por medio de un diagrama en forma de árbol invertido llamado "dendograma", en el que las sucesivas fusiones de las ramas a los distintos niveles nos informan de las sucesivas fusiones de los grupos en grupos de superior nivel (mayor tamaño, menor homogeneidad) sucesivamente:

El nivel de agrupamiento para cada fusión viene dado por un indicador llamado "valor cofenético" que debe ser proporcional a la distancia o disimilaridad considerada en la fusión (distancia de agrupamiento). Esta distancia o disimilaridad considerada en cada fusión estará definida, a veces, entre individuos y, otras, entre clusters; razón por la cual, será necesario ampliar el concepto de distancia o disimilaridad de acuerdo con algún criterio que nos permita realizar el algoritmo de clasificación.

Una vez completamente definida la distancia para individuos, clusters y cluster-individuo, la clasificación jerárquica se puede llevar a cabo mediante un sencillo algoritmo general :

PASO 1 Formamos la partición inicial:

$P = \{ i_1 \},\{ i_2 \},...\{ i_N \}$

considerando cada individuo como un cluster.

PASO 2 Determinamos los dos clusters más próximos (de menor distancia) $i_i$ ,$i_j$ , y los agrupamos en uno solo.

PASO 3 Formamos la partición:

$P = \{ i_1 \},\{ i_2 \},...\{ i_i \cup i_j \},...,\{ i_N \}$

PASO 4 Repetimos los pasos 2 y 3 hasta obtener la partición final Pr= {W}

Este algoritmo será esencialmente el mismo para todos los métodos de clasificación jerárquica (ascendente); las diferencias residirán , como ya hemos apuntado, en el criterio de definición de la distancia entre clusters.

## Encadenamiento simple .Método de la distancia mínima (nearest neighbour o single linkage)

En este método se procede de acuerdo con el algoritmo general considerando la distancia ENTRE CLUSTERS como la distancia mínima entre los individuos más próximos
Este método es espacio-contractivo, esto es, tiende a aproximar los individuos más de lo que indicarían sus disimilaridades o distancias iniciales.

El método del mínimo ha sido reivindicado "matemáticamente preferible" por sus propiedades por Jardine y Sibson . Sin embargo, ha sido muy criticado por ser muy **sensible** en aquellos casos en los que existen individuos perturbadores entre clusters bien diferenciados **individuos intermedios**) (casos con "ruido").

## Endaenamiento completo.Método de la distancia máxima (furthest neighbour o complete linkage)

Este método, debido a Johnson, utiliza el algoritmo general para la obtención de la clasificación jerárquica ascendente, pero considerando la distancia entre clusters como la distancia entre los individuos más alejados.
Por modificar la métrica en sentido inverso que el método anterior, este método es espacio-dilatante, en el sentido en que tiende a separar a los individuos en mayor medida que la indicada por sus disimilaridades iniciales.
El método de la distancia máxima se encuentra, como el anterior, en franca decadencia, ya que presenta los inconvenientes de **alargar mucho el proceso y dar como resultado agrupaciones encadenadas.**
Mientras el método de la distancia mínima asegura que la distancia entre los individuos más próximos de un cluster será siempre menor que la distancia entre elementos de distintos clusters, el de la distancia máxima va a asegurar que la distancia máxima dentro de un cluster será menor que la distancia entre cualquiera de sus elementos y los elementos más alejados de los demás clusters.

## Método de la media (u.p.g.m.a.)

Los dos métodos anteriores, a pesar de poseer buenas propiedades teóricas tienen el inconveniente de distorsionar las medidas iniciales de disimilaridad, constringiendo o dilantando, respectivamente, la métrica. Una solución al problema fue el método ideado por Sokal y Michener, conocido como Group Average.
Sokal y Michener propusieron utilizar como distancia entre un grupo I y un individuo j la media de las distancias entre los individuos del grupo I y el individuo j:

$$D(I,j) = 1/N_I \sum D(i, j)$$

Posteriormente, Lance y Williams extendieron la definición a la distancia entre dos grupos como la media de todas las distancias entre todos los pares de individuos de los dos grupos.
Este método es espacio-conservativo, ésto es, no hace variar considerablemente la métrica inicial, y resulta ser uno de los más utilizados, resolviendo de forma más aceptable la presencia de ruido.

## Método del centroide

Fue propuesto originalmente, también, por Sokal y Michener, y utiliza como distancia entre grupos la distancia entre los centroides de cada grupo. Este método es, también, espacio-conservativo, pero presenta el inconveniente de dejarse influir excesivamente por los grupos de mayor tamaño. Esto hace que sea menos utilizado que el anterior.

## Método de la mediana

La mayor desventaja del método del centroide es que si se fusionan dos grupos de diferente tamaño, el centroide del nuevo grupo queda más cerca del grupo de mayor tamaño y más alejado del de menor tamaño en proporción a sus diferencias de tamaño. Esto trae como consecuencia que durante el proceso aglomerativo de fusión se van perdiendo paulatinamente las propiedades de los grupos pequeños.

Para evitar esto, puede suponerse, con independencia del tamaño que tengan los grupos en realidad, que los grupos son de igual tamaño.
Llevando a cabo esta estrategia, la distancia entre un individuo o grupo K de centroide k y el grupo formado por la fusión de los grupos I y J de centroides i y j viene dada por la mediana del triángulo i,j, k. Razón por la cual Gower propuso el nombre de método (distancia) de la mediana.
Este método es, como el del centroide, espacio-conservativo, aunque también como él no resulta ser invariante ante transformaciones monótonas de la distancia empleada, cosa que sí ocurría con los tres primeros métodos.

## Método de Ward

Ward propuso que la pérdida de información que se produce al integrar los distintos individuos en clusters puede medirse a través de la suma total de los cuadrados de las desviaciones entre cada punto (individuo) y la media del cluster en el que se integra. Para que el proceso de clusterización resulte óptimo, en el sentido de que los grupos formados no distorsionen los datos originales, proponía la siguiente estrategia: En cada paso del análisis, considerar la posibilidad de la unión de cada par de grupos y optar por la fusión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse.
El método de Ward es uno de los más utilizados en la práctica; posee casi todas las ventajas del método de la media y suele ser más discriminativo en la determinación de los niveles de agrupación. Sin embargo, y a pesar de que en cada paso del proceso obtiene una agrupación óptima dad la agrupación anterior no garantiza que siempre vaya a obtener una solución óptima de un número prefijado de grupos. Con todo, una investigación llevada a cabo por Kuiper y Fisher probó que este método era capaz de acertar mejor con la clasificación óptima que otros métodos (mínimo, máximo, media y centroide).
La implementación en weka de los métodos jerárquicos obece al esquema habitual y en cuanto a cómo desarrollar estos método , mostramos un script de ejemplo (https://www.uv.es/mlejarza/datamine/hclustAMV.R )

```
library(readxl)
amValencia1 <- read_excel("datamining/datos/amValencia1.xlsx")
View(amValencia1)
names(amValencia1)
Z=model.matrix( ~-1+nortesur+
                  +Idependencia+Tnatalidad+Tmortalidad
                +Tcrecimiento+Tmigracion+envejecimiento
                +nlineastelef+ nvehi+ indimas
                +segvivienda+vivdesocu+ta
                +taf + iacomercial + iafinan +
                  + iaindus +cenerdomest+cenergindus
                +formAcad + IparLabSer+ IparLabIndustria
                + IparLabConstr, amValencia1)
row.names(Z)=amValencia1$muni
#Cluster jerárquico ( WARD, Distanc. Euclidea cuadrado)
clus1 <- hclust(dist(Z)^2,  method= "ward.D")
plot(clus1,labels=row.names(Z),hang=-1,main= "Dendrograma agrupación municipios A M valencia",
xlab=
      "municipios ",
    sub="Metodo de ward Distancia euclidea-cuadrado")
#consideramos 4 grupos
# variable de adscripción a uno de los clusters, agrupación de 4 grupos
clusWard4=cutree(clus1,k=4)
```

```
clusWard4
nuevosdatos<-cbind(clusWard4,amValencia1)
#cluster jerárquico (encadenamiento completo, distancia euclidea)
clus2 <- hclust(dist(Z),  method= "complete")
plot(clus2,labels=row.names(Z),hang=-1,main= "Dendrograma agrupación municipios A M valencia",
xlab=
     "municipios ",
   sub="Encadenamiento completo Distancia euclidea")

cluscompleto4=cutree(clus2,k=4)
cluscompleto4
nuevosdatos2<-cbind(cluscompleto4,nuevosdatos)
dendograma<-as.dendrogram(clus2)
#podemos ampliar el dendograma por zonas
plot(cut(dendograma, h = 250)$upper, main = "Arbol superior al corte  h=250")
plot(cut(dendograma, h = 250)$lower[[2]],
   main = "Segunda rama del árbol inferior al corte  at h=250")
plot(cut(dendograma, h = 250)$lower[[3]],
   main = "Tercera rama del árbol inferior al corte  at h=250")
plot(cut(dendograma, h = 350)$lower[[3]],
   main = "Tercera rama del árbol inferior al corte  at h=350")
```

### 4.3.Mapas autoorganizados y métodos de malla.

Los métodos de malla ( o de grid) se basan en la consideración de un espacio de menor dimensión que el número de variables , normalmente 1, 2 o 3 y en la división de este espacio en un número de celdas  constituyendo una rejilla o malla. Entre los procedimiento de este tipo están el STING ( Statistical Information Grid) y las redes neuronales de Kohonen de  1 y de 2 dimensiones ( LVQ y SOM, respectivamente)

EL MODELO NEURONAL DE KOHONEN

Los mapas auto-organizados de Kohonen se inspiran en la conocida tendencia del cerebro biológico de adecuar por zonas  sus funciones sensoriales, motoras y cognitivas, de forma que, al parecer, ante estímulos similares se suele producir la activación de grupos de neuronas próximas entre sí. Esto sugiere que el cerebro podría tener cierta capacidad para formar *mapas topológicos* de las informaciones recibidas del exterior e incluso ser capaz de *auto-organizarse* para procesar funciones cognitivas de importancia como pueden ser las asociadas al procesamiento de elementos semánticos.

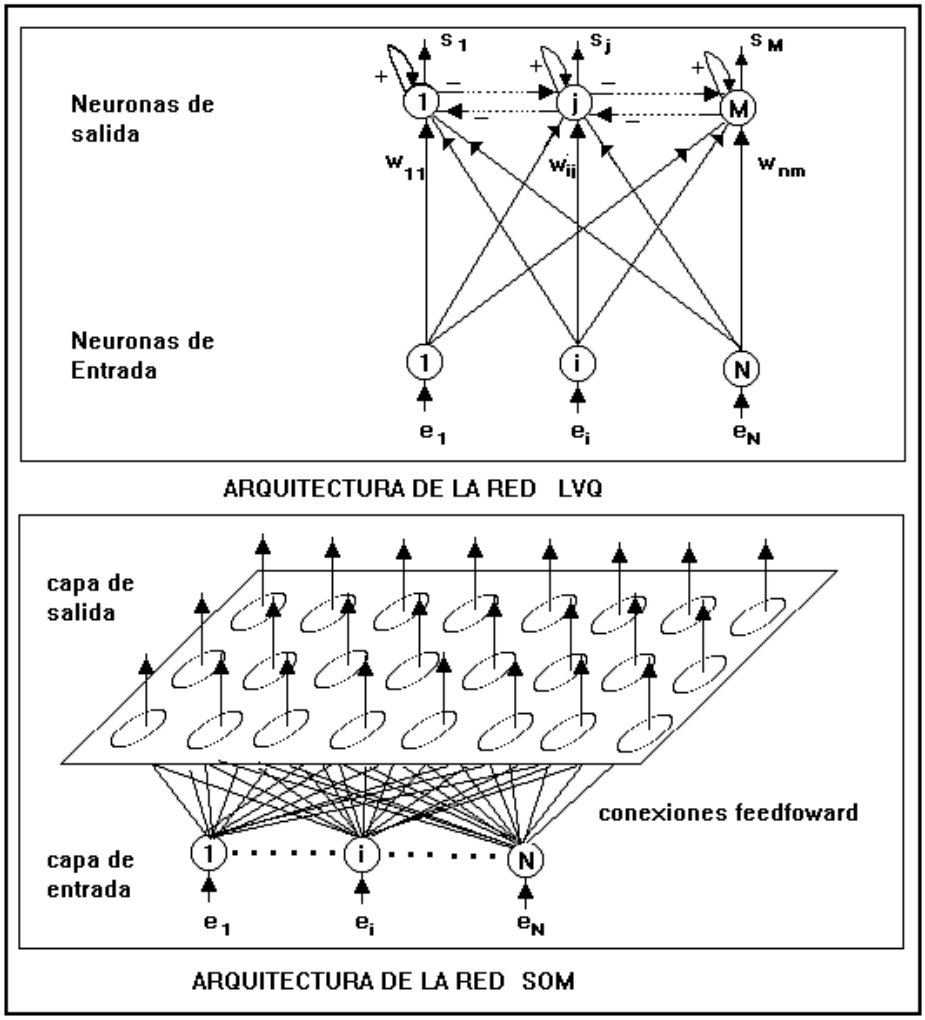
El modelo presentado por Teuvo Kohonen (1982), pretendía, en esa línea, imitar artificialmente esta capacidad para formar **mapas de características.** La idea era demostrar cómo un estímulo, por sí solo, suponiendo una determinada estructura y una descripción funcional del comportamiento de la red, era suficiente para forzar la formación de estos mapas.

A pesar de sus limitaciones en cuanto a la duración de su aprendizaje y a la imposibilidad de introducir nuevos patrones de aprendizaje sin tener que repetir el proceso completo, el modelo de Kohonen es uno de los más útiles en computación neuronal. Con el tiempo, el modelo se ha mostrado eficaz en tareas relacionadas con la clasificación, la reducción de la dimensionalidad, la extracción de rasgos básicos de un conjunto de informaciones, y en el reconocimiento del habla. En campos mucho más cercanos a nosotros se ha empezado a utilizar fundamentalmente vinculado a procedimientos alternativos de agrupación o clustering de empresas, regiones o países en función de la información económica disponible.

## Arquitectura de los modelos de Kohonen.

El modelo de Kohonen toma por objetivo fundamental el de crear una imagen de un espacio multidimensional de entrada en un espacio de salida de menor dimensionalidad. Las dos versiones del modelo ( LVQ y SOM ) constan de dos capas de neuronas; una de entrada constituida por n neuronas de entrada que se limita a recoger y canalizar la información n-dimensional de entrada, y una capa de salida que procesa la información de entrada y da como resultado la representación reducida de la información introducida. Como ya se ha apuntado, la diferencia entre ambas versiones consiste en que mientras en el Learning Vector Quantization (LVQ) la capa de salida está constituida por un conjunto de M neuronas de salida dispuestas en una dimensión ,el modelo SOM ( mapa auto-organizado, propiamente dicho ) dispone de una capa de salida constituida por $M=m_X \times m_Y$ neuronas dispuestas en un plano ( dos dimensiones).De esta forma, mientras el proceso llevado a cabo por el modelo LVQ puede asemejarse al análisis (estadístico) cluster tradicional, como veremos; el resultado obtenido con un mapa auto-organizado de dos dimensiones resulta ser una clasificación de los vectores (información) de entrada de diferente naturaleza que no tiene un símil preciso en los métodos de análisis de datos tradicionales. Con todo, es obvio que el modelo LVQ puede considerarse como un caso unidimensional de mapa auto-organizado.

En este gráfico se presenta un esquema de la arquitectura de las dos redes auto organizadas:



En ambos casos existirán conexiones laterales entre las neuronas de la capa de salida, dependiendo la influencia de cada neurona sobre las demás de una función de la distancia con ellas, habitualmente una función tipo sombrero mejicano ; produciendo la excitación de las neuronas cercanas , la inhibición de las lejanas y siendo imperceptible en las muy lejanas.Precisamente estas conexiones laterales determinarán, en el proceso de aprendizaje, los pesos sinápticos que acabarán conectando las neuronas de la capa de entrada con las de salida.

## Funcionamiento de la red.

El funcionamiento de la red es relativamente simple. Cada vez que se presenta a la red una información de entrada; esto es, cada vez que se introduce un vector N-dimensional de datos a través de la capa de entrada, la señal llega ponderada por los pesos $w_{ji}$ a las neuronas de la capa de salida por medio de las conexiones feedforward.Al mismo tiempo, estas neuronas reciben las correspondientes entradas del resto de las neuronas de salida a través de las conexiones laterales, y cuya influencia dependerá de la distancia a la que se encuentren:

Así ,la salida generada en una neurona de salida $u_j$ ante el vector de entrada $E_k$ dado por $E_k = ( e_{1(k)},... ,e_{N(k)})$ vendrá dada por:

$$Sj\ (t+1)= f\ \left( \sum w_{ji}\ e_{i(k)} + \sum Int_{pj}\ S_p(t) \right)$$

Donde $Int_{pj}$ es una función ( tipo sombrero mejicano ) que da cuenta de la influencia lateral de la neurona de salida $u_p$ sobre la $u_j$ .Y siendo la función de salida una función continua.

Al tratarse de una red competitiva la red evolucionará hasta alcanzar una situación estable en la que, ante una determinada entrada, $E_k = ( e_{1(k)},... ,e_{N(k)})$ , sólo se active una neurona: la neurona ganadora.De forma que la formulación de su funcionamiento puede simplicarse representando la activación final de las M neuronas de salida como:

$$S_j = 1 \quad si \quad \| E_k - W_j \| = \min_j \{ \| E_k - W_j \| \}$$

$$S_j = 0 \quad \text{en caso contrario}$$

Siendo $W_j$ el vector de pesos ( $w_{j1}, w_{j2}$ , ... ,$w_{jN}$ ) y siendo $\| E_k - W_j \|$ una medida de la diferencia o discrepancia entre el vector de entrada y el vector de pesos ; habitualmente la distancia euclidea .

Así pues, en la fase de entrenamiento, se produce la adopción de los pesos para que éstos acaben registrando los datos aprendidos y, de esta forma, en la fase posterior de funcionamiento poder encontrar el dato "aprendido" al que más se parece la información de entrada suministrada.En definitiva, como vemos, la red lleva a cabo un proceso de clasificación, representando, la neurona activada, la clase a la que pertenece la información de entrada: Ante otra entrada parecida se activará la misma neurona, o quizá otra muy cercana ( en el modelo bidimensional ) garantizándose que las neuronas topológicamente próximas den cuenta de informaciones físicamente similares.

## Aprendizaje de la red

El aprendizaje en los dos modelos de Kohonen es básicamente igual difiriendo obviamente en la dimensionalidad.Ambas aprenden según un procedimiento OFF

LINE, secuencialmente separado del funcionamiento ( y previo a él), utilizando, como ya se ha comentado un aprendizaje no supervisado de tipo competitivo.

En la fase de entrenamiento, se facilitan a la red, un conjunto de vectores N-dimensionales de entrada ( vectores de entrenamiento), para que ésta establezca, en función de su similitud, las categorías o clusters, ( una por cada neurona de salida ), que en el proceso de funcionamiento servirán para clasificar eventualmente nuevos datos.

Se observa, por lo tanto, el enorme parecido entre el proceso de aprendizaje y el análisis cluster tradicional y se comprende, cómo el caso del aprendizaje en un mapa auto-organizado unidimensional (LVQ) pueda ser considerado , sin más, como un método de clusterización (tradicional ) alternativo.

El algoritmo de aprendizaje empleado requiere de la iteración en la presentación de todos los patrones de aprendizaje, siendo necesario presentar varias veces ( habitualmente, cientos o miles ) el conjunto de datos de entrada para refinar el mapa topológico de salida e incluso conseguir la convergencia.

En el caso del modelo LVQ, el algoritmo, es básicamente, el siguiente:

1. En primer lugar, se inicializan los pesos $w_{ji}$ con valores aleatorios pequeños, aunque también pueden establecerse a priori valores iniciales.Y se fija igualmente la zona inicial de vecindad entre las neuronas de salida.
2. A continuación, se presenta la información de entrada que la red debe aprender; el vector de entrada $E_k = ( e_{1(k)},... ,e_{N(k)})$ .
3. Se determina la neurona ganadora $u_j$ ,cuyo vector de pesos,$W_j=(w_{j1},w_{j2},...,w_{jN})$, sea el más parecido al patrón de entrada introducido, $E_k$ .Habitualmente la determinación se realiza a partir de la distancia euclidea al cuadrado:

   $$d_j = \sum (e_{i(k)} - w_{ji})^2 \qquad \text{para } 1 \leq j \leq M.$$

   Ello suele exigir, en las aplicaciones de clasificación la normalización (tipificación) previa para evitar la sobrevaloración de algunas variables en detrimento de otras.
4. Una vez determinada la neurona ganadora ( j* ) se actualizan los pesos de las conexiones entre las neuronas de entrada y las neuronas de salida cercanas a la ganadora consiguiéndose con ello asociar la información de entrada con cierta zona de la capa de salida. La actualización se lleva a cabo de acuerdo con la regla:

   $$w_{ji}(t+1) = w_{ji}(t) + \alpha(t)\,(e_{i(k)} - w_{j*i}(t)) \quad \text{para las neuronas } u_j \in \Gamma_{j*}(t).$$

   Donde $\Gamma_{j*}(t)$ es una zona próxima alrededor de la neurona ganadora en la que se encuentran las neuronas cuyas pesos serán actualizados en el momento t del proceso de aprendizaje.Esta zona, en ocasiones va reduciéndose con el proceso de aprendizaje y en otros casos se mantiene constante durante todo el proceso.Cuando se reduce esta zona suele lo hacerlo de forma lineal o exponencial de manera que, si $\Gamma_{j*}(t)$ viene dada por aquellas neuronas que distan de la ganadora una distancia menor a R(t) ( radio de aprendizaje ), éste va variando según:

   $$R(t) = R_0\,(1 - t/\alpha_2) \quad \text{o bien :} \quad R(t) = R_0\,(R_f/R_0)^{t/\alpha_2}$$

   siendo $R_0$ el radio inicial de aprendizaje, $R_f$ el radio final de aprendizaje y $\alpha_2$ el número de iteraciones.

   El término $\alpha(t)$ es el llamado parámetro de ganancia, o factor de aprendizaje, toma valores comprendidos entre 0 y 1 y es habitual que decrezca con el número de iteraciones ( con el tiempo de aprendizaje ), de forma que al final del proceso sea prácticamente nulo.Habitualmente el factor de aprendizaje decrece lineal o

exponencialmente, siendo lo más frecuente que dependa su decrecimiento del factor de aprendizaje inicial ,$\alpha_1$, y del número de iteraciones,$\alpha_2$ ,que suelen convertirse así en los dos parámetros básicos del proceso:Las expresiones más habituales para el factor de aprendizaje son :
$\alpha$ (t) = 1/t ; $\alpha$ (t) = $\alpha_1$ ( 1 - t/$\alpha_2$ ) ; o bien $\alpha$ (t) =$\alpha_1$ ($\alpha_f/\alpha_1$)$^{t/\alpha_2}$

5. Finalmente, el proceso, como se ha comentado se reitera un gran número de veces: Quinientas, mil o incluso más veces son presentadas a la red todos los patrones de entrenamiento: {$E_k$} ( k= 1,2,..., P )

El paquete Kohonen de R permite la realización de agrupamientos según estos métodos. Así mismo, con weka se puede disponer de dos paquete adicional SOM y LVQ para realizar los agrupamientos según estos esquemas. Sin embargo weka no permite sacar partido de alguna de las virtualidades topológicas de estos métodos ofreciendo únicamente una agrupación final de un determinado número de clusters.

El script somdiscricochesR10.R ( www.uv.es/mlejarza/somdiscricochesR10.R) muestra una aplicación en R para la base de datosdiscricochesR10.sav (una versión de la base de datos con todas la variables numéricas) .

### 4.4.Otros métodos de clustering

Algunos otros métodos de clustering no encajan en ninguno de los esquemas anteriores tal es el caso del DBSCAN (Density-based spatial clustering of applications with noise ) agrupamiento espacial basado en densidad, o del EM clustering ( Expectation-Maximization ) o cluster por esperanza –maximización.
El primero de ellos intenta buscar “regiones densas” en las que un conjunto de puntos son alcanzables secuencialmente ( según su proximidad) y a partir de estas regiones organiza la agrupación.
El segundo, que es el método por defecto que ofrece weka, utiliza un algoritmo en dos fases (fase esperanza, fase maximización) pretendiendo estimar la esperanza de la verosimilitud de que un elemento pertenezca a cada posible grupo y , a partir de ahí se agrupar de forma que se maximice la esperanza de pertenencia a cada grupo candidato.