# Tema 6

Cluster: es un grupo de objetos, similares o distintos. El clustering es la acción de hacer un cluster.

Los principales inconvenientes del clustering son los **outliers**¸ que puede implicar que haya clusters con un solo valor o que se incluyan en otros clusters y bajen de calidad. Si se hace en BBDD dinámicas, la pertenencia a los clusters varía con el tiempo.

Los algoritmos **sufren** cuantos más atributos tengan ya que es más difíciles agruparlos.

## Bondad de un análisis cluster

Un buen cluster debe producir:

* Maximizar la similaridad intra-cluster
* Minimizar la similaridad inter-clustes

Es decir: separar lo mejor posible. Por lo tanto, la calidad depende de la medida de similaridad tanto como de la implementación.

## Propiedades deseables de un método de clustering

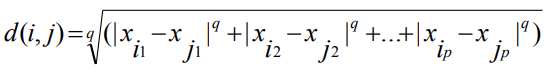
## Medidas de distancia y similaridad

Lo ideal es que haya menor distancia entre las instancias inter-cluster con respecto a las que están fuera.

La medida realmente depende del tipo de variable: continua, booleana,etc.

**Medida de Minkowski**

Es una distancia en un espacio euclidiano que puede considerarse una generalización de la distancia Euclidea como de la distancia de Manhattan.



Si q=2 -> Euclidea

Si q=1 -> Manhattan

IMPORANTÍSIMO NORMALIZAR

## Distintas aproximaciones al clustering

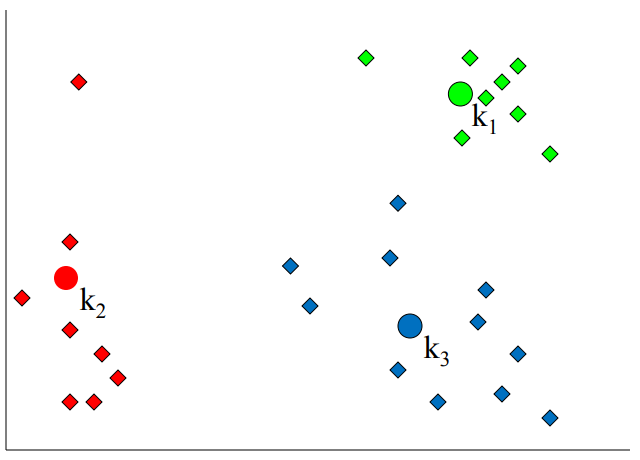
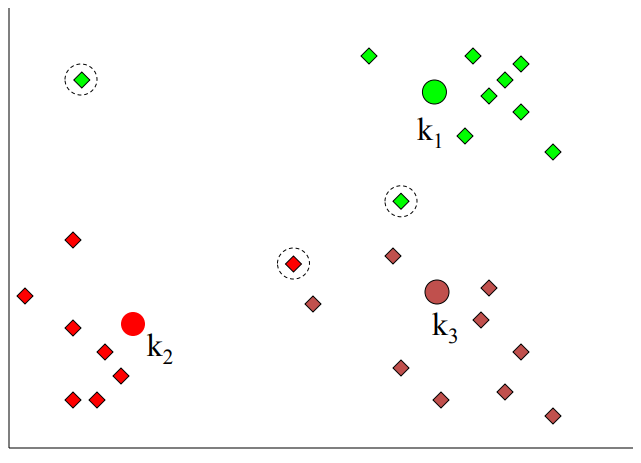
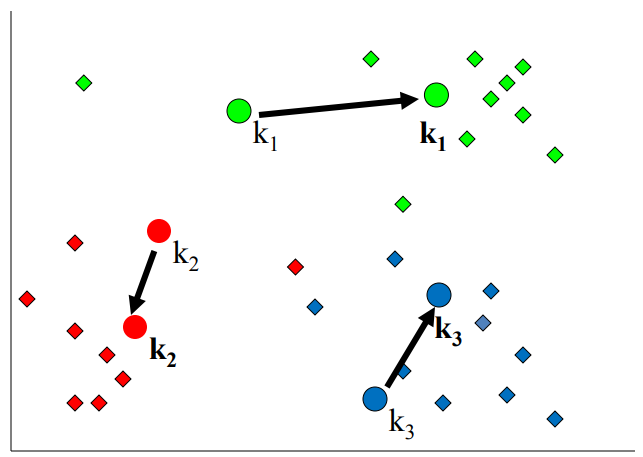
Existen dos tipos de métodos para hacer los clusters:

### Métodos de particionamiento

Se crean particiones y se evalúan con algún criterio. Entonces, se parte la base de datos en K clusters. Los métodos heurísticos son:

#### K-means

Lo difícil es detectar el K, que es número de clusters. Se cogen 3 centros aleatoriamente, y se asignan los puntos al cluster más cercano. A continuación, mueve cada centro del cluster a la media del cluster. Después hace otra reasignación.



Cuando después de mover los centros, todos los puntos pertenecen al mismo centro, se para.

**Ventajas**

Es relativamente eficiente: O(tkn) y con frecuencia finaliza en un óptimo local.

**Desventajas**

* Solo es aplicable cuando el concepto de media es definible (no vale para nominales)
* Se necesita definir K inicialmente. Otra opción es iterar con valores distintos de K y elegir la mejor
* Es débil a los datos ruidosos y/o outliers
* Está indicado para clusters convexos (esféricos)

#### Mean Shift

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteSe coge una circunferencia inicial con radio r (se puede estimar con K-nn). Se va desplanzando la ventana deslizante hacia áreas más densas, desplanzando el punto central a la media de los puntos dentro de la ventana . Cuando ya no se puede desplazar más (la densidad es la máxima), se para. Esto se hace para cada cluster. Habrá puntos que no entren en ninguna esfera por lo que este algoritmo es más robusto a ruido y outliers. El problema es la esfera. Los clusters esféricos no se adaptan a la perfección.

A diferencia de K-means, no es necesario saber el número de clusters. Sin embargo, si que tenemos que seleccionar el tamaño de la ventana (el radio).

#### DBSCAN

Es una mejora del Mean Shift. Inicia desde un punto arbitrario y toma una distancia. Todos los que están dentro de ella, pertenecen al **vecindario**. Si hay suficientes puntos para empezar (minpts), se crea toma el punto inicial como primer punto del grupo. SI no, se cataloga como outliers. Al final puede ser que se incluya.

Se van visitando todos los puntos del vecindario y añadiendo al cluster. Cuando ha acabado, visita otro punto de fuera del vecindario y se reinicia el proceso.

Imagen que contiene Forma

Descripción generada automáticamente

**Ventajas**

No hace falta pasarle un número de clusters.

Es robusto frente a datos ruidosos y outliers

Encuentra clusters de cualquier forma.

**Desventajas**

No funciona bien cuando la densidad de los clusters es variable. Es decir: no hay la misma densidad en todos los clusters. Esto es porque épsilon no se puede ir ajustando durante el lanzamiento del algoritmo.

Depende de la noción de distancia.

#### BIRCH (clustering incremental)

Va agrupando conforme recibe puntos. Es el peor de la familia. Cada vez que llega un nuevo objeto, desciende por el árbol escogiendo en cada nodo el CF más cercano. Si cuando llega a una nueva hoja puede ser absorbido por algún CF existente se agrega y si no, se crea un CF nuevo.

#### Medidas de rendimiento para los métodos basados en particionamiento

**Coeficiente silhouette**

Mide cómo de similares son los objetos de un mismo cluster (cohesión) en comparación con otros clusters (separación). Va de -1 a 1. Saca las medias de todos los objetos con el cluster y luego hace la media de eso. Tiene varios problemas porque si aunque el cluster se diferencie bien, si estan lejos del centro la longitud será grande. Cuando haces la medida te saldrá baja.

Si se acerca a 0 quiere decir que está en el borde. La media de todos es el coeficiente de silhouette.

**Coeficiente Calisnki-Harabasz**

Es la razón entre la dispersión intra-clusters y la dispersión inter-clusters. Cuanto mayor, mejor agrupamiento. Sirve para comparar distintos algoritmos de clustering entre ellos.

### Métodos jerárquicos

Producen una jerarquía entre clusters. No requieren como parámetro el número de clusters. Dependiendo del nivel de corte, obtendremos un clustering distinto (cuanto más profundo lleguemos). Existen dos métodos:

#### Métodos aglomerativos

Se basan en medir la distancia entre clusters. En cada paso se fusionan los dos clusters más cercanos. Inicialmente, hay un cluster por cada objeto. DE MAYOR A MENOR (de todos clusters a menos clusters)

# Tema 7

En este tema tratamos en análisis categórico no supervisado, que son las **reglas de asociación.**

Buscamos patrones frecuentes, asociaciones, correlaciones, etc. El objetivo de los algoritmos de extracción de reglas de asociación es:

* Dada una base de **datos de transacciones** donde cada transacción es una lista de artículos.
* Encontrar todas las reglas que co-relacionen la presencia de un conjunto de artículos con otro conjunto de artículos.

**Definición de transacción**

Item (o artículo): elemento individual

Itemset (o conjunto): conjunto de ítems

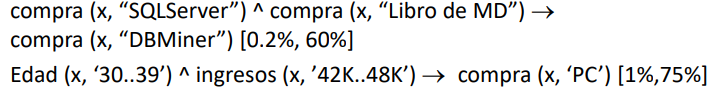
Soporte de un conjunto: nº de transacciones

Soporte mínimo ms: umbral de soporte

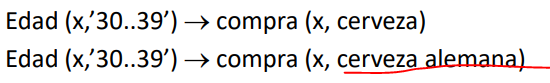
Conjunto frecuente: con soporte >= ms. Representan conjuntos de artículos que están correlacionados positivamente.

## Tipos de reglas de asociación

1. Asociaciones booleanas vs cuantitativas



1. Asociaciones unidimensionales vs multidimensionales



Sin embargo, la asociación **no implica** necesariamente correlación o causalidad.

## Medidas de soporte y confianza

Se trata de encontrar todas las reglas que cumplan cierta característica (un umbral).

**Soporte:** es la probabilidad de que una transacción contenga lo que pedimos.

**Confianza:** probabilidad condicional.

Se trata de sumar en los itemsets. Después dividimos entre todos. Si cumple un porcentaje mínimo se añade.

## Proceso de extracción

El proceso se descompone en dos fases:

1. Encontrar conjuntos de artículos frecuentes (que cumplan un umbral mínimo)
2. Generar reglas de asociación “fuerte” (deben satisfacer tanto el soporte mínimo como soporte de confianza)

## Algoritmo A priori

Este algoritmo encuentra las asociaciones más frecuentes. Itera sobre las base de datos buscando las asociaciones que cumplen el soporte mínimo y usa los conjuntos frecuentes para generar reglas de asociación. Es simple pero robusto. Los requisitos del algoritmo son:

* No necesita fijar los atributos consecuentes ni antecedentes (derecho e izquierdo)
* Existen variedades para poder tratar todos los tipos de datos
* Especificar soporte mínimo
* Especificar máximo número de reglas

La regla de este algoritmo es “*cualquier subconjunto de un conjunto de artículos frecuente debe ser frecuente*”. Esto permite definir un principio de poda (si un conjunto contiene un subconjunto infrecuente, no es frecuente).

PREUGNTAR pagina 27

**Soporte mínimo algoritmo**

Si es alto, tendremos muy pocas reglas, pero todas válidas.

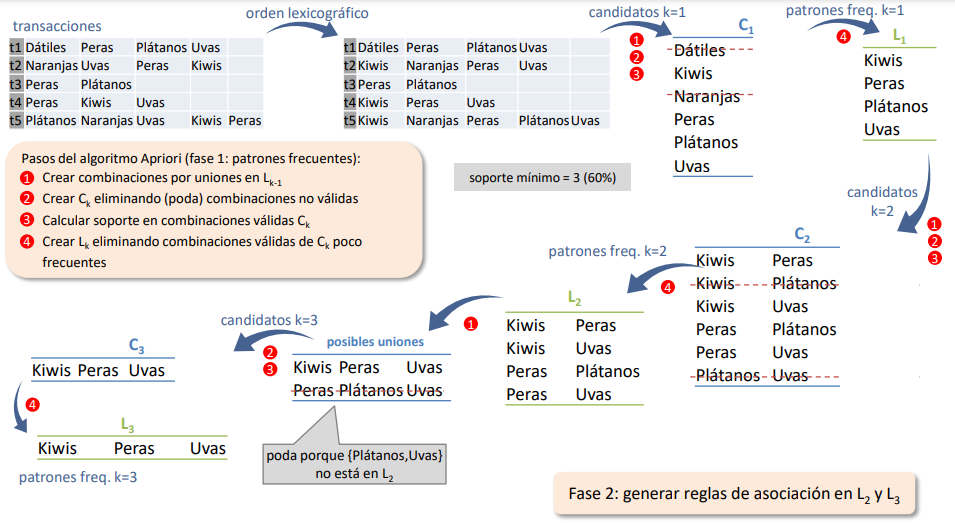
Si es bajo, tendremos muchas reglas que ocurren con raramente.

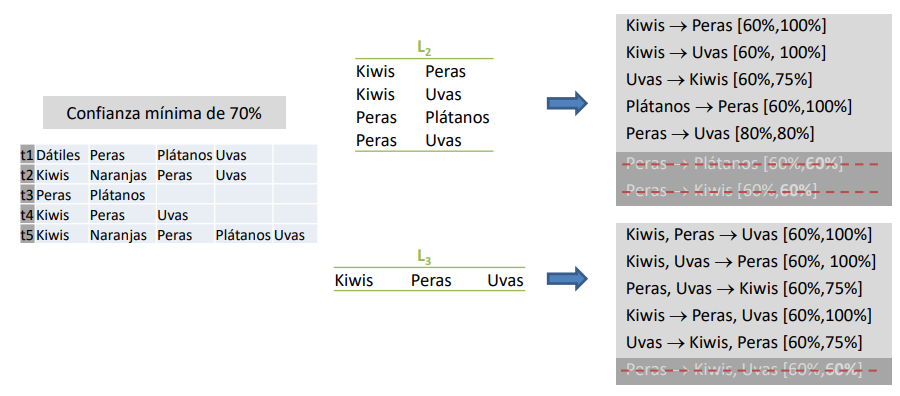
**Confianza mínima algoritmo**

Si es alta, pocas reglas pero todas “casi ciertas lógicamente”

Si es baja, muchas reglas pero muchas de ellas inciertas

Valores típicos soporte -> 2-10% y confianza 70-90%





## Medidas de interés

Las reglas son interesantes si son **inesperadas o útiles.**

El problema de la medida de confianza es que cuando hay reglas que se limitan a conjuntos, la confianza puede variar. La confianza se calcula sólo sobre el subconjunto de datos implicados en la regla.

Existe otra medida que se llama **Interés** o empuje que arregla lo anterior.

## Otros algoritmos

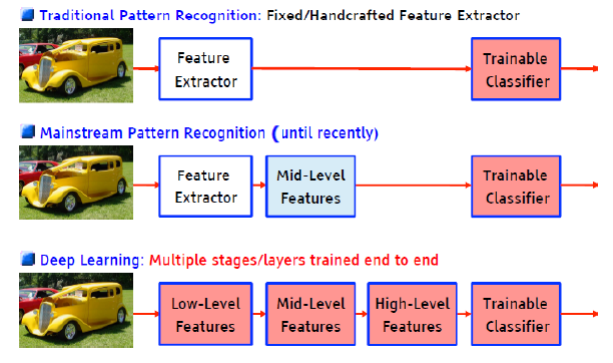
**FP-Growth:** basado en FP-Tree. Descubre frecuentes sin generar los itemset candidatos. Se construye una estructura FP-Tree y de ahí se extraen los frecuentes.

# Tema 8. Deep learning

Deep Learning significa utilizar una red neuronal con varias capas de nodos entre entrada y salida. Estas capas identifican características. Siempre hjemos tenido algoritmos con una capa oculta. Ahora tenemos más de una.

Lo que se hace ahora es que se programa la entrada hasta la salida y luego se retropropaga el error para corregir pesos en cada capa.

La nueva forma de entrenar RNA es entrenando capa a capa y no tdas de una vez. Cada capa está entrenada para ser un codificador automático (auto-encoder). Es como que se subdivide el aprendizaje.



## Red reuronal convolucional (CNN)

Es una red neuronal donde las capas son como neuronas que corresponden a campos receptivos. Se realizan en matrices bidimensionales por lo que son efectivas para tareas de visión artificial (imágenes)

# Tema 9. Problemas singulares

Las clases que solemos buscar a veces representan una minoría entre todos los datos. La clase positiva representa un alto interés en comparación con la negativa. Esto produce que el desbalanceo entre ambas sea un serio problema.

La evaluación se basa en las medidas:

*Sensivity (*positive true ratio) =

*Specificity* (negative true ratio) =

## Medidas

## F-Measure