Minería de patrones frecuentes y reglas de asociación

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y análitica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Tipos de patrones

- Un **patrón** se define como subsecuencias, subestructuras o itemsets que representan cualquier tipo de homogeneidad y regularidad en los datos
- Los patrones representan propiedades intrínsecas e importantes de un conjunto de datos
- La cuantificación del interés de los patrones descubiertos se relaciona con diferentes métricas que están estrechamente asociadas al propósito para el que se realiza la tarea

Se estiman 8.000 granos de arena por cm^3 ~ 10^{36} granos en un cubo de playa

Con solo 150 ítems, $2^{150} - 1$ posibles patrones ~ $1.42 \times 10^{45} \gg 10^{36}$

Nuestra cesta de la compra como ejemplo ilustrativo

TID	Items
1	Plátano, Pan, Manteca, Naranja
2	Plátano, Pan, Leche
3	Plátano, Manteca, Leche
4	Pan, Manteca, Leche, Naranja
5	Plátano, Manteca, Leche, Naranja

1-itemsets

- Plátano: 4
- Pan: 3
- Manteca: 4
- Leche: 4
- Naranja: 3

2-itemsets

- Plátano, Pan: 2
- Plátano, Manteca: 3
- Plátano, Leche: 3
- Plátano, Naranja: 2
- Pan, Manteca: 2
- Pan, Leche: 2
- Pan, Naranja: 2
- Manteca, Leche: 3
- Manteca, Naranja: 3
- Leche, Naranja: 2

3-itemsets

- Plátano, Pan, Manteca: 1
- Plátano, Pan, Leche: 1
- Plátano, Pan, Naranja: 1
- Plátano, Manteca, Leche: 2
- Plátano, Manteca, Naranja: 2
- Plátano, Leche, Naranja: 1
- Pan, Manteca, Leche: 1
- Pan, Manteca, Naranja: 2
- Pan, Leche, Naranja: 1
- Manteca, Leche, Naranja: 2

4-itemsets

- Plátano, Pan, Manteca, Leche: 0
- Plátano, Manteca, Leche, Naranja: 1
- Plátano, Pan, Leche, Naranja: 0
- Plátano, Pan, Manteca, Naranja: 1
- Pan, Manteca, Leche, Naranja: 1

TID	*Items*
1	Plátano, Pan, Manteca, Naranja
2	Plátano, Pan, Leche
3	Plátano, Manteca, Leche
4	Pan, Manteca, Leche, Naranja
5	Plátano, Manteca, Leche, Naranja

Tipos de patrones

Patrones frecuentes

Definición

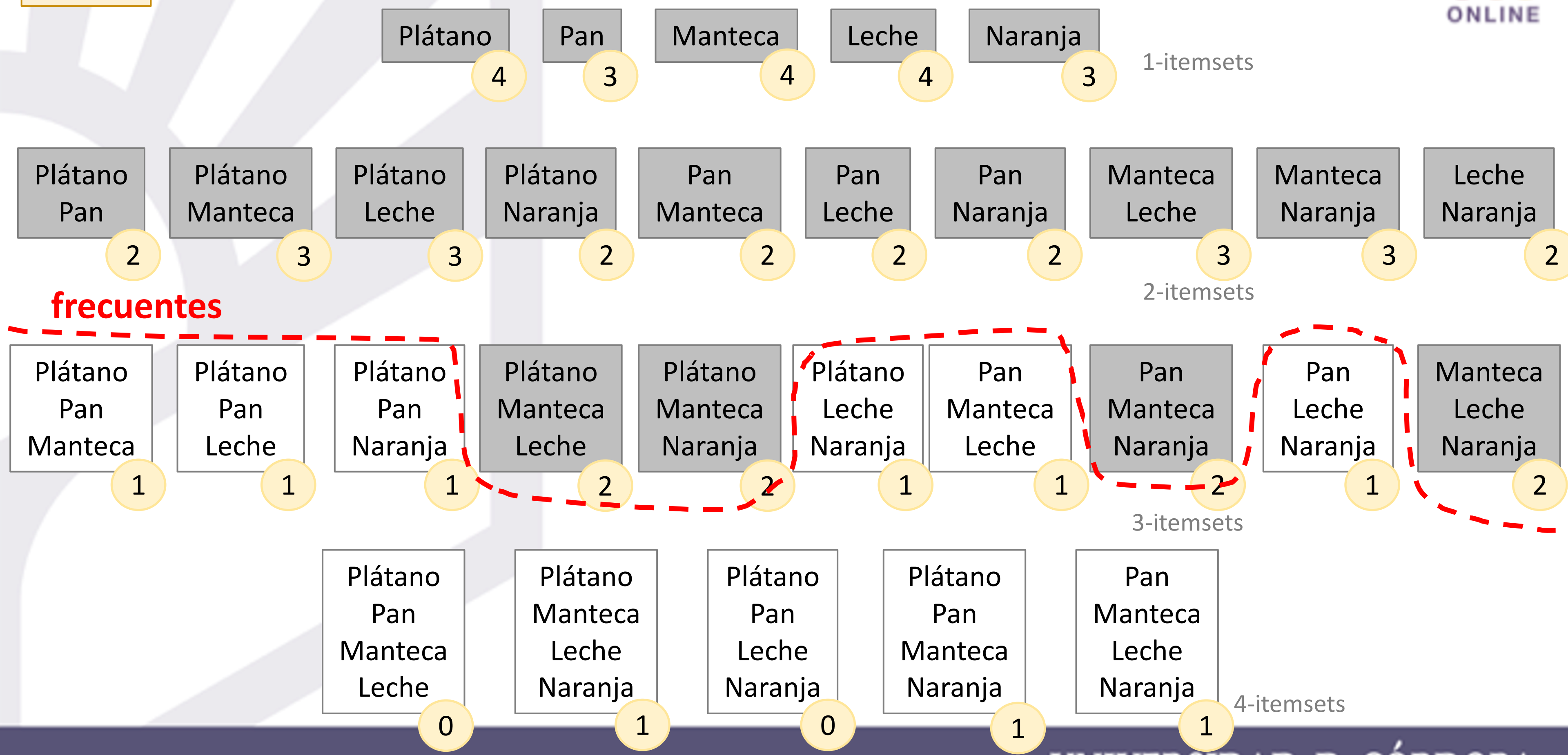
Sea $I = \{i_1, i_2, \ldots, i_n\}$ el conjunto de ítems en el conjunto de datos, y sea $T = \{t_1, t_2, \ldots, t_m\}$ el conjunto de todas las transacciones de la base de datos:

- Cada transacción t_j es un conjunto de ítems, tal que $t_j \subseteq I$
- Sea P un patrón conteniendo un conjunto de ítems $P \subseteq I$, se dice que el **patrón satisface la transacción** t_j *sii* $P \subseteq t_j$ y la frecuencia del patrón $f(P) = |\{\forall t_j \in T : P \subseteq t_j\}|$
- Un **patrón es frecuente** *sii* el número de transacciones que satisface es mayor o igual que un valor mínimo f_{min}, $f(P) \geq f_{min}$

Complejidad

- Un conjunto de datos con **n** ítems contiene **2^n – 1** *itemsets* diferentes, y el número de *itemsets* de tamaño k es igual a $\binom{n}{k}$ para cualquier k ≤ n
- La cantidad de computaciones necesarias para cualquier candidato es ***O*(k)**, y la **complejidad global** del proceso de minería es
$$O(\sum_{k=1}^{n} k \, x \binom{n}{k}) = O(2^{n-1} \, x \, n)$$
- Se trata de una complejidad de orden exponencial, incluso mayor si se calcula la frecuencia f de cada *itemset*
- La **complejidad de computar f** teniendo n ítems y m transacciones es igual a $O(2^{n-1} \, x \, m \, x \, n)$

$f_{min} = 2$


Plátano 4
Pan 3
Manteca 4
Leche 4
Naranja 3
1-itemsets
Plátano Pan 2
Plátano Manteca 3
Plátano Leche 3
Plátano Naranja 2
Pan Manteca 2
Pan Leche 2
Pan Naranja 2
Manteca Leche 3
Manteca Naranja 3
Leche Naranja 2
2-itemsets
frecuentes
Plátano Pan Manteca 1
Plátano Pan Leche 1
Plátano Pan Naranja 1
Plátano Manteca Leche 2
Plátano Manteca Naranja 2
Plátano Leche Naranja 1
Pan Manteca Leche 1
Pan Manteca Naranja 2
Pan Leche Naranja 1
Manteca Leche Naranja 2
3-itemsets
Plátano Pan Manteca Leche 0
Plátano Manteca Leche Naranja 1
Plátano Pan Leche Naranja 0
Plátano Pan Manteca Naranja 1
Pan Manteca Leche Naranja 1
4-itemsets

Tipos de patrones

Patrones infrecuentes (o raros)

UCO ONLINE

Definición

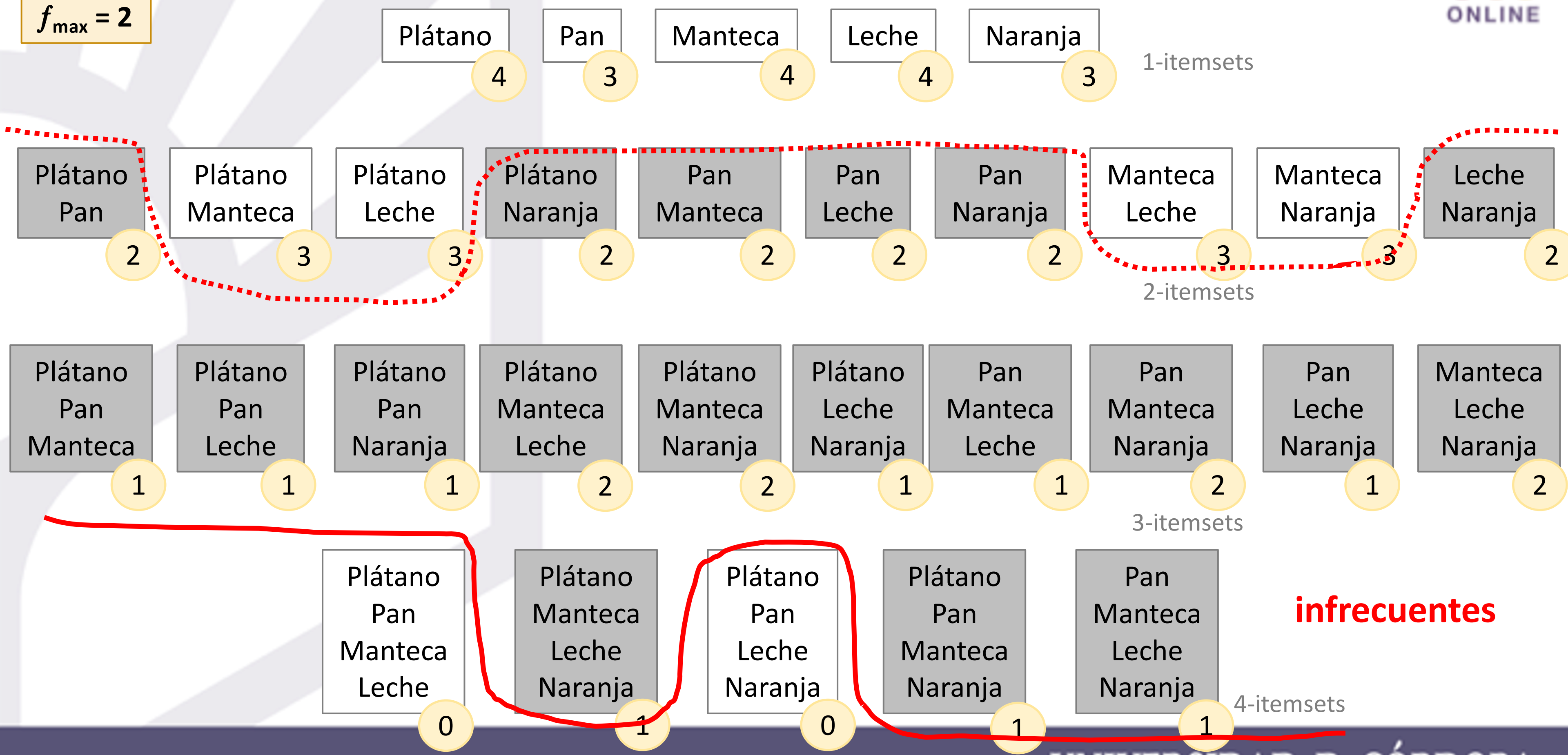
- Apropiados para descubrir comportamientos anormales o inusuales en la base de datos, es decir, no siguen una tendencia de otros

  Sea P un patrón conteniendo un conjunto de ítems $P \subseteq I$, se dice que el **patrón es infrecuente** (o raro) *sii* el número de transacciones que satisface es menor que un valor máximo predefinido f_{max}, esto es $\boldsymbol{f(P) = |\{\forall t_j \in T : P \subseteq t_j\}| \leq f_{max}}$

- **¡CUIDADO!** Esta definición implicaría que cualquier *itemset* que no aparezca en la base de datos sería considerado infrecuente

  Un patrón P **es infrecuente** *sii* el número de transacciones que satisface es menor que un valor máximo predefinido f_{max} y mayor que un valor mínimo f_{min}, esto es: $\boldsymbol{f_{min} \leq f(P) \leq f_{max}}$

$f_{min} = 1$
$f_{max} = 2$


Plátano 4
Pan 3
Manteca 4
Leche 4
Naranja 3
1-itemsets
Plátano Pan 2
Plátano Manteca 3
Plátano Leche 3
Plátano Naranja 2
Pan Manteca 2
Pan Leche 2
Pan Naranja 2
Manteca Leche 3
Manteca Naranja 3
Leche Naranja 2
2-itemsets
Plátano Pan Manteca 1
Plátano Pan Leche 1
Plátano Pan Naranja 1
Plátano Manteca Leche 2
Plátano Manteca Naranja 2
Plátano Leche Naranja 1
Pan Manteca Leche 1
Pan Manteca Naranja 2
Pan Leche Naranja 1
Manteca Leche Naranja 2
3-itemsets
Plátano Pan Manteca Leche 0
Plátano Manteca Leche Naranja 1
Plátano Pan Leche Naranja 0
Plátano Pan Manteca Naranja 1
Pan Manteca Leche Naranja 1
4-itemsets
infrecuentes

Tipos de patrones

Patrones máximos (*maximal freq patterns*) y cerrados (*closed freq patterns*)

Definición

En el mundo real, es habitual que los patrones tengan una longitud demasiado elevada (¡un supermercado vende más de 5 productos!)

- La extracción de patrones frecuentes sobre *itemsets* de gran extensión supone un elevado coste computacional
- ¡**RECORDEMOS**! Cualquier subconjunto de un patrón frecuente es también frecuente
- Para un patrón $|P| = 5$, contiene $2^5 - 2 = 30$ subpatrones también frecuentes
- El descubrimiento de patrones frecuentes con representaciones más condensadas supone un alivio en términos computacionales y de almacenamiento

Definición

Sea $\mathbf{P^F} = \{P_1, P_2, \ldots, P_n\}$ el **conjunto de patrones frecuentes** de una base de datos, un patrón frecuente $P_i \in P^F$ se define como **patrón frecuente máximo** *sii* no tiene superconjuntos frecuentes:

$$\{P_i : \nexists P_j \supset P_i,\ P_j \in P^F \wedge P_i \in P^F\}$$

El número de patrones frecuentes máximos <u>es considerablemente menor</u> que el número de todos los patrones frecuentes

$f_{min} = 2$

1-itemsets

Itemset	Frecuencia
Plátano	4
Pan	3
Manteca	4
Leche	4
Naranja	3

2-itemsets

Itemset	Frecuencia
Plátano, Pan	2
Plátano, Manteca	3
Plátano, Leche	3
Plátano, Naranja	2
Pan, Manteca	2
Pan, Leche	2
Pan, Naranja	2
Manteca, Leche	3
Manteca, Naranja	3
Leche, Naranja	2

3-itemsets

Itemset	Frecuencia
Plátano, Pan, Manteca	1
Plátano, Pan, Leche	1
Plátano, Pan, Naranja	1
Plátano, Manteca, Leche	2
Plátano, Manteca, Naranja	2
Plátano, Leche, Naranja	1
Pan, Manteca, Leche	1
Pan, Manteca, Naranja	2
Pan, Leche, Naranja	1
Manteca, Leche, Naranja	2

4-itemsets

Itemset	Frecuencia
Plátano, Pan, Manteca, Leche	0
Plátano, Manteca, Leche, Naranja	1
Plátano, Pan, Leche, Naranja	0
Plátano, Pan, Manteca, Naranja	1
Pan, Manteca, Leche, Naranja	1

Patrones frecuentes máximos

Definición

Las representaciones más condensadas de patrones frecuentes se obtienen habitualmente con los **patrones frecuentes cerrados**

Sea **P^F** = $\{P_1, P_2, \ldots, P_n\}$ el **conjunto de patrones frecuentes** de una base de datos, un patrón frecuente $P_i \in P^F$ se define como **patrón frecuente cerrado** *sii* no tiene superconjuntos frecuentes con la misma frecuencia que él mismo:

$$\{\nexists P_j \supset P_i : f(P_i) \geq f_{min} \wedge f(P_j) = f(P_i)\}$$

Los patrones frecuentes máximos **<u>son</u>** patrones frecuentes cerrados

Los patrones frecuentes máximos pierden información sobre las frecuencias de los subconjuntos que los patrones frecuentes cerrados no pierden

$f_{min} = 2$

UCO ONLINE

Plátano	Pan	Manteca	Leche	Naranja
4	3	4	4	3

1-itemsets

Plátano Pan	Plátano Manteca	Plátano Leche	Plátano Naranja	Pan Manteca	Pan Leche	Pan Naranja	Manteca Leche	Manteca Naranja	Leche Naranja
2	3	3	2	2	2	2	3	3	2

2-itemsets

Plátano Pan Manteca	Plátano Pan Leche	Plátano Pan Naranja	Plátano Manteca Leche	Plátano Manteca Naranja	Plátano Leche Naranja	Pan Manteca Leche	Pan Manteca Naranja	Pan Leche Naranja	Manteca Leche Naranja
1	1	1	2	2	1	1	2	1	2

3-itemsets

Plátano Pan Manteca Leche	Plátano Manteca Leche Naranja	Plátano Pan Leche Naranja	Plátano Pan Manteca Naranja	Pan Manteca Leche Naranja
0	1	0	1	1

4-itemsets

Patrones frecuentes cerrados

Tipos de patrones

Patrones positivos y negativos

Definición

Los **patrones pueden ser representados en formato binario**, en el que una columna de **valor 1** indica un ítem positivo (presente en la transacción) y un **valor 0** indica un ítem negativo (no presente en la transacción)

En el cálculo de patrones frecuentes, la presencia de valores 1 es más importante que los valores 0

Para el cálculo de patrones infrecuentes, consideraremos las relaciones negativas entre ítems (valores 0)

Transc.	Plátano	¬Plátano	Pan	¬Pan	Manteca	¬Manteca	Leche	¬Leche	Naranja	¬Naranja
ID1	1	0	1	0	1	0	0	1	1	0
ID2	1	0	1	0	0	1	1	0	0	1
ID3	1	0	0	1	1	0	1	0	0	1
ID4	0	1	1	0	1	0	1	0	1	0
ID5	1	0	0	1	1	0	1	0	1	0

Es fácil determinar que el patrón P_i = [Plátano, Pan] es frecuente ya que **$f(P_i) \geq f_{min} = 2$**

El patrón P_i se describe como una relación positiva entre la compra de ambos ítems

El patrón negativo P_j = [Plátano, ¬Pan] (*el cliente compra plátanos y no compra pan*) describe la relación negativa entre la compra de ambos ítems con **$f(P_j) = 2$**

En muchos dominios puede ser interesante considerar la no ocurrencia de un ítem en la transacción (*distinto a que no aparezca en el itemset*)

Transc.	Plátano	¬Plátano	Pan	¬Pan	Manteca	¬Manteca	Leche	¬Leche	Naranja	¬Naranja
ID1	1	0	1	0	1	0	0	1	1	0
ID2	1	0	1	0	0	1	1	0	0	1
ID3	1	0	0	1	1	0	1	0	0	1
ID4	0	1	1	0	1	0	1	0	1	0
ID5	1	0	0	1	1	0	1	0	1	0

¡Precaución!

El **uso de patrones negativos** debe ser estudiado cuidadosamente ya que puede implicar la presencia de ítems con una probabilidad muy alta, por lo que el conocimiento extraído puede ser insignificativo

En estas circunstancias, **se recomienda la extracción de patrones infrecuentes**

Transc.	Plátano	¬Plátano	Pan	¬Pan	Manteca	¬Manteca	Leche	¬Leche	Naranja	¬Naranja
ID1	1	0	1	0	1	0	0	1	1	0
ID2	1	0	1	0	0	1	1	0	0	1
ID3	1	0	0	1	1	0	1	0	0	1
ID4	0	1	1	0	1	0	1	0	1	0
ID5	1	0	0	1	1	0	1	0	1	0

Tipos de patrones

Patrones continuos

UCO ONLINE

Definición

- Es habitual recopilar información continua (más rica), como edad o salario, que no puede representarse de forma binaria
  - Los **patrones continuos** son aquellos que se pueden representar como un rango de valores en términos de un límite inferior y superior

  Sea un conjunto de ítems $I = \{i_1, i_2, \ldots, i_n\}$, donde al menos un ítem $i_j \in \mathbf{I}$ se define en un dominio continuo $\Re$, $i_j \in \Re$. A i_j se le denomina **ítem numérico**.

  Un ítem numérico se representa por el **rango $i_j = [x_l, x_u]$**, donde $\mathbf{x_l}$ es el límite inferior y $\mathbf{x_u}$ es el límite superior.

  Un **patrón P se dice continuo** si contiene al menos un ítem numérico, esto es, $\{P \subseteq I : \exists i_j \in P \wedge i_j \in \Re\}$

Definición

Sea $T = \{t_1, t_2, \ldots, t_m\}$ el conjunto de todas las transacciones de la base de datos, y cada transacción t_j un conjunto de ítems tal que $t_j \subseteq I$, un patrón continuo **P satisface la transacción t_j** ***sii*** $P \in t_j$ y la frecuencia del patrón $f(P)$ se define como el número de transacciones diferentes que lo satisfacen, esto es: $\boldsymbol{f(P) = \{t_j : P \subseteq t_j, t_j \in T\}}$

Ejemplo: Consideremos el patrón **P = {Plátano, Factura=[5,3, 5,9]}** tiene ***f*(P) = 2**

La minería de patrones continuos **es un reto** debido a que el número de patrones numéricos es infinito en relación a un rango mínimo/máximo.
Se debe limitar el espacio de búsqueda mediante la discretización de los valores

Transc.	Plátano	Pan	Manteca	Leche	Naranja	Factura
ID1	1	1	1	0	1	6,5
ID2	1	1	0	1	0	5,4
ID3	1	0	1	1	0	5,8
ID4	0	1	1	1	1	6,6
ID5	1	0	1	1	1	7,4

Transc.	Plátano	Pan	Manteca	Leche	Naranja	Factura
ID1	1	1	1	0	1	6,5
ID2	1	1	0	1	0	5,4
ID3	1	0	1	1	0	5,8
ID4	0	1	1	1	1	6,6
ID5	1	0	1	1	1	7,4

Discretizamos para convertir los patrones continuos en patrones binarios o discretos

Transc.	Plátano	Pan	Manteca	Leche	Naranja	Factura [5,4, 6,0]	Factura [6,1, 6,7]	Factura [6,8, 7,4]
ID1	1	1	1	0	1	0	1	0
ID2	1	1	0	1	0	1	0	0
ID3	1	0	1	1	0	1	0	0
ID4	0	1	1	1	1	0	1	0
ID5	1	0	1	1	1	0	0	1

Tipos de patrones

Patrones colosales

UCO ONLINE

Definición

Un **patrón colosal** es un patrón de gran tamaño (p.ej., los extraídos en el campo de la bioinformática a partir de expresiones de genes)

Dado un conjunto de ítems $I = \{i_1, i_2, \ldots, i_n\}$, un **patrón colocal P** se define como un subconjunto de I, $\{P = \{i_j, \ldots, i_k\}\ I, 1 \leq j, k \leq n\}\}$, cuya longitud **|P|** es demasiado elevada, y cualquiera de sus sub-patrones tiene una frecuencia similar, esto es, no hay una caída significativa en la frecuencia de P al añadir nuevos ítems.

Para obtener los patrones colosales es importante el concepto de **patrón nuclear** (*core pattern*):

Dado un patrón colosal P, *con un conjunto de subpatrones de frecuencia similar*, **P' es un patrón nuclear** si es un sub-patrón de P, $\mathbf{P' \subset P}$, y $\boldsymbol{f(P') \approx f(P)}$.

Definición

Se entiende como frecuencia similar aquella definida en términos de un **ratio r**, que determina que **$f(P)/f(P')$ = r**, $0 \leq r \leq 1$

Debido al gran número de patrones nucleares que contiene un patrón colosal, el cómputo de todos los subpatrones frecuentes de un patrón colosal es computationalmente fuerte (*computationally hard*), esto es, no puede ser calculado eficientemente en un tiempo polinomial $O(n^k)$

El **uso de patrones frecuentes cerrados y máximos** puede parcialmente aliviar el problema computacional:

- Un **patrón colosal** es un **patrón frecuente cerrado** pero un **patrón frecuente cerrado** no es necesariamente un **patrón colosal**
- Un **patrón colosal** no es necesariamente un **patrón frecuente máximo**

Tipos de patrones

Patrones secuenciales y espacio-temporales

UCO ONLINE

Patrones secuenciales

En ocasiones, un patrón no solo es interesante por mostrar una relación fuerte entre ítems, sino porque también expresa una **relación secuencial entre ítems**

Una secuencia **S** se describe como un conjunto de eventos **S = <$e_1 \rightarrow \ldots \rightarrow e_n$>**, siendo e_i un *itemset* $\{i_i, \ldots, i_j\}$ I, donde los distintos eventos aparecen en **orden temporal de ocurrencia**.

Dada una secuencia S, otra secuencia **S' es sub-secuencia de S** *sii* hay un evento en S' que es un subconjunto de S, y los eventos se mantienen ordenados.

> S = ({Naranja}, {Pan, Manteca}, {Leche, Plátano}) y S' = ({Pan},{Leche, Plátano}) es una sub-secuencia de S, **S' $\subset$ S**
>
> **Sin embargo**: S''=({Naranja}, {Pan, Plátano}) no es una sub-secuencia de S, **S' $\not\subset$ S**, ya que {Pan, Plátano} $\not\subset$ {Pan, Manteca} $\wedge$ {Pan, Plátano} $\not\subset$ {Leche, Plátano}

Patrones espacio-temporales

Gracias al avance de los sistemas de posicionamiento y sensórica, el uso de marcas temporales (*time stamp*) puede tratarse en conjunción con características espaciales, dando lugar a **patrones espacio-temporales**

La captura de datos de movimiento es interesante para descubrir caminos frecuentes, de modo que estudiar **los movimientos pasados puede servir para entender las trayectorias futuras** (**minería de patrones espacio-temporales**)

> Dado un conjunto de **n** localizaciones descritas por sus coordenadas (x_i, y_i), y una marca de tiempo ti, se define una secuencia espacio-temporal S como
> $S = (\{(x_1, y_1), t_1\}, \{(x_2, y_2), t_2\}, \ldots, \{(x_n, y_n), t_n\})$

¡Gracias!