

# UNIDAD 2: MINERÍA DE ELEMENTOS FRECUENTES

## MODELO MERCADO CANASTA

---

Blanca Vázquez y Gibran Fuentes-Pineda

Octubre 2020

## Los principales bloques en minería de datos

- Clasificación
- Clustering
- Detección de valor atípicos
- Minería de patrones

*¿Por qué son especiales estos bloques?*

Consideremos una matriz  $D$  de tamaño  $n \times d$

	d_1	d_2	d_3	...	d_n
n_1	0	0	1	0	1
n_2	0	1	0	0	0
n_3	0	0	0	1	1
....	...	...	...	...	...
doc_n	0	0	1	0	0

- Relaciones entre columnas
  - Predecir el valor de una columna en especial (*clasificación*)
- Relaciones entre filas

- Relaciones entre columnas
- Relaciones entre filas
  - Determinar subconjuntos de filas con características similares (*clustering*)
  - Identificar valores anómalos entre filas (*detección de valores atípicos*)
  - Identificar asociaciones frecuentes entre los elementos (*minería de patrones*)

De manera general, la minería de patrones se define en el contexto de las matrices binarias dispersas.

	Pan	Mantequilla	Leche	Huevos	Yogurt
Cliente_1	1	1	1	0	0
Cliente_2	0	0	1	1	1
Cliente_3	0	1	0	1	1
Cliente_4	1	0	1	0	0

Uno de los estudios más comunes es la minería de elementos frecuentes.

- La minería de elementos frecuentes es el proceso de descubrimiento de tendencias o patrones a partir de grandes conjuntos de datos con el objetivo de guiar futuras decisiones.
- Surgió en el contexto de los datos de los supermercados

# ¡VÁMONOS DE COMPRAS!



Figura 1: Imagen tomada de Puig [1]

*¿Qué productos compran los clientes?, ¿Cuáles productos compran juntos?*



## ¡VÁMONOS DE COMPRAS!



Figura 2: Imagen tomada de Debashis Borgohain

# ¡VÁMONOS DE COMPRAS!

La minería de elementos frecuentes ha sido ampliamente usada en los supermercados\*.

## Comprados juntos habitualmente



+



+



Precio total: **\$75.05**

Agregar los tres al carrito

Describe una relación de muchos a muchos entre dos clases:

- **Elementos** (*items*): son cada uno de los eventos o elementos en una transacción
- **Transacciones** (*baskets*): es una colección de items (*itemset*)

Objetivo: identificar el conjunto de elementos que son adquiridos en conjunto.

$$\{\text{fideos, queso rallado}\} \Rightarrow \{\text{salsa}\}$$

Se asume lo siguiente:

- El número de elementos en una transacción es más pequeño que el número total de elementos.
- El número de transacciones puede ser tan grande, que podemos llegar a saturar el espacio de almacenamiento.

Son aquellos conjuntos de elementos que pueden aparecer en muchas transacciones, y estos tendrán un umbral llamado soporte.

**Definición:** sea  $I$  un conjunto de elementos, el soporte de  $I$  es la fracción de transacciones para lo cual  $I$  es un subconjunto.

Base de datos de transacciones de una tiendita

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, galletas, pan}
5	{queso, galletas, yogurt}

- Recordemos, cada transacción es un conjunto de elementos comprados al mismo tiempo.
- El total de transacciones en la bd = 5
- El soporte de un elemento  $I$  se denota por  $sup(I)$
- El  $sup(\{pan\}) = 3/5 = 0.6$
- El  $sup(\{mantequilla\}) = 1/5 = 0.2$
- El  $sup(\{yogurt\}) = 2/5 = 0.4$

- Cuando se calcula el soporte de un elemento, se dice que es un conjunto único.
- También podemos calcular el soporte para dos o más elementos:  $sup(\{galletas, leche\}) = 1/5 = 0.2$
- Decimos que  $I$  es frecuente, si su soporte es igual o mayor al umbral definido para  $sup$



Supongamos que el soporte mínimo  $minsup = 0.5$ , ¿qué pares de elementos, cumplen este umbral?

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, galletas, pan}
5	{queso, galletas, yogurt}

Las reglas de asociación son los elementos más importantes en el modelo mercado canasta.

$$\{X\} \Rightarrow \{Y\}$$

- \* dónde X y Y son elementos individuales o conjuntos de elementos,
- \* X se conoce como **antecedente** y Y como **consecuente**

$$\{\text{fideos, queso rallado}\} \Rightarrow \{\text{salsa}\}$$

Ejemplos de asociaciones comunes:

- $\{\text{pañales}\} \Rightarrow \{\text{cerveza}\}$
- $\{\text{leche}\} \Rightarrow \{\text{pan}\}$
- $\{\text{pan para hot dogs}\} \Rightarrow \{\text{mostaza}\}$

Sea  $X$  y  $Y$  dos conjuntos de elementos, la confianza  $conf(X \cup Y)$  de la regla  $X \cup Y$  es la probabilidad condicional de  $X \cup Y$  que ocurre en una transacción dado que la transacción contiene  $X$ .

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$$

T	Elementos
1	{pan, mantequilla, leche}
2	{huevos, leche, yogurt}
3	{pan, queso, huevos, leche}
4	{huevos, leche, yogurt}
5	{queso, leche, yogurt}

$$\text{conf}(\{\text{huevos, leche}\} \Rightarrow \{\text{yogurt}\}) = \frac{\text{sup}(\{\text{huevos, leche, yogurt}\})}{\text{sup}(\{\text{huevos, leche}\})}$$

$$\text{conf}(\{\text{huevos, leche}\} \Rightarrow \{\text{yogurt}\}) = \frac{\text{sup}(2/5)}{\text{sup}(3/5)}$$

$$\text{conf}(\{\text{huevos, leche}\} \Rightarrow \{\text{yogurt}\}) = 2/3$$

Las reglas de asociación, se definen por su soporte y su nivel de confianza.

- Semejante al soporte, en donde se define un soporte mínimo (*minsup*), también se define un nivel de confianza mínimo (*minconf*)
- El *minconf* puede ser usado para generar reglas de asociación relevantes
- La confianza (*conf*) es la fuerza de la asociación entre elementos

T	Elementos
1	{pan, gelatina, mantequilla-maní}
2	{pan, mantequilla-maní}
3	{pan, leche, mantequilla-maní}
4	{cerveza, pan}
5	{cerveza, leche}

Calcula la confianza de las siguientes reglas de asociación:

$conf(\{pan\} \Rightarrow \{mantequilla - mani\})$

$conf(\{cerveza\} \Rightarrow \{pan\})$

$conf(\{mantequilla - mani\} \Rightarrow \{gelatina\})$

$conf(\{gelatina\} \Rightarrow \{leche\})$



Calcula la confianza de las siguientes reglas de asociación:

$$\text{conf}(\{pan\} \Rightarrow \{mantequilla - mani\}) = (3/5)/(4/5) = 0.75$$

$$\text{conf}(\{cerveza\} \Rightarrow \{pan\}) = (1/5)/(2/5) = 0.5$$

$$\text{conf}(\{mantequilla - mani\} \Rightarrow \{gelatina\}) = (1/5)/(3/5) = 0.3$$

$$\text{conf}(\{gelatina\} \Rightarrow \{leche\}) = (0)/(2/5) = 0$$

- Reglas con bajo soporte: pueden haber aparecido por casualidad
- Reglas con baja confianza: es probable que no existe relación entre el antecedente y el consecuente
- $\{pepsi\} \Rightarrow \{coca - cola\}$

- Binarias vs cuantitativas
- Unidimensionales vs multidimensionales
- De un nivel vs multinivel

Se basan en los tipos de datos:

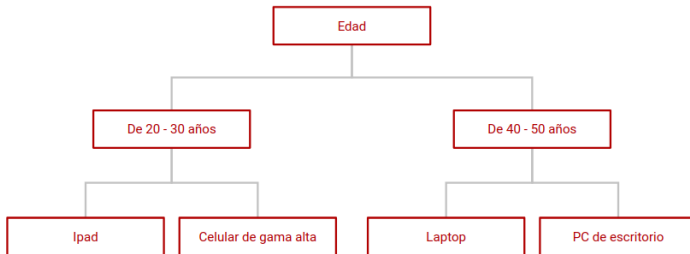
- $\text{compra}\{\text{laptop}\} \Rightarrow \text{compra}\{\text{impresora}\}$
- $\text{edad}\{> 30\} \wedge \text{sueldo}\{> 30,000\} \Rightarrow \text{compra}\{\text{SmartTV}\}$

Se basan en las dimensiones de los datos involucrados en la regla

- $\text{compra}\{\text{laptop}\} \Rightarrow \text{compra}\{\text{impresora}\}$
- $\text{edad}\{> 30\} \wedge \text{sueldo}\{> 30,000\} \Rightarrow \text{compra}\{\text{SmartTV}\}$
- $\text{compra}\{\text{traje, camisa}\} \Rightarrow \text{compra}\{\text{zapatos, corbata}\}$

# REGLAS DE UN NIVEL VS MULTINIVEL

Se basan en el nivel de abstracción involucrado:



# REGLAS DE UN NIVEL VS MULTINIVEL

Se basan en el nivel de abstracción involucrado:

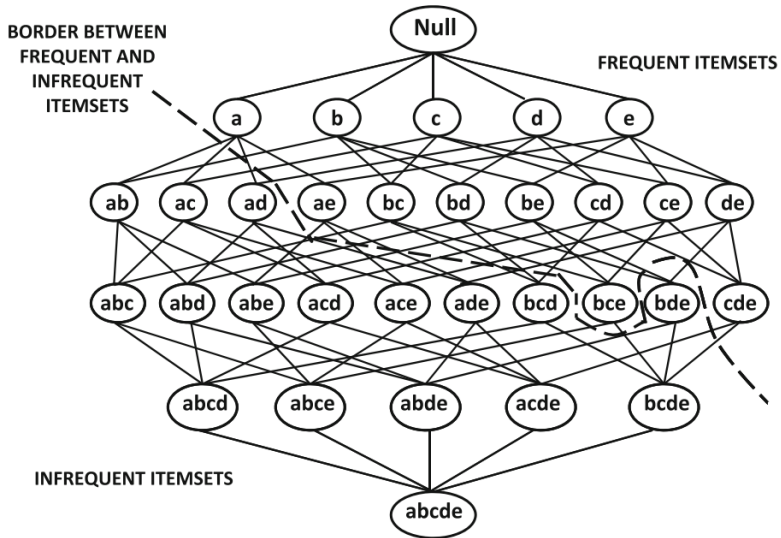


Imagen tomada de Aggarwal, 2015.

- Casas inteligentes:  
 $\{temp\_baja\} \Rightarrow \{encender\_calefaccion\}$   
 $\{despertador\_activo\} \Rightarrow \{encender\_cafetera\}$
- Sistemas de recomendación:  
 $\{El\_senor\_anillos\} \Rightarrow \{El\_hobbit\}$
- Dispositivos móviles:  
 $\{mensaje\_noimportate, reunion\} \Rightarrow \{no\_notificar\}$



- Identificación de patrones de compra (asociaciones)
- Minería de texto
- Detección de eventos espacio-temporales
- Detección de errores de programación
- Biomarcadores
- Identificación de plagio
- Identificación de patrones en bibliotecas, librerías.

1. Introduction to machine learning, Albert Orriols-Puig. URL:  
*<https://www.slideshare.net/aorriols/lecture13-association-rules>*
2. Inteligencia Artificial, Ariel Monteserin,