

Reglas de Asociación y Dependencia

Creado por

M^a Carmen Pegalajar Jiménez

Dpto. Ciencias de la Computación e I. A.

Profesor

Juan Gómez Romero

jgomez@decsai.ugr.es

Profesor

Juan Gómez Romero

Ciencias de la Computación e Inteligencia Artificial

E-mail: jgomez@decsai.ugr.es

Web: <https://ccia.ugr.es/~jgomez/>

Sesiones (x4):

L 26-abril (12h), X 28-abril (10h)

[TEORÍA]

L 3-mayo (12h), X 5-mayo (10h)

[PRÁCTICA]

PRADO: <https://pradogrado2021.ugr.es/course/view.php?id=8766#section-5>

Reglas de Asociación y Dependencia

- ¿Por qué?
 - Las reglas de asociación siguen un enfoque estadístico (se basan en los conceptos de correlación y de probabilidad condicionada)
 - Surgen en el ámbito de la Computación
 - Su cálculo requiere algoritmos y estructuras de datos especializadas

Reglas de Asociación y Dependencia

Punto de partida:

- Tratan con **atributos nominales / categóricos**

Expresan:

- **Patrones de comportamiento entre los datos** en función de la aparición conjunta de ciertos valores
- **Combinaciones** de valores de *items* que suceden más **frecuentemente**.

Reglas de Asociación y Dependencia

Aplicaciones

- **Análisis de la cesta de compra** de un supermercado.
Podemos conocer que productos suelen comprarse conjuntamente y así mejorar la distribución de los productos en estanterías

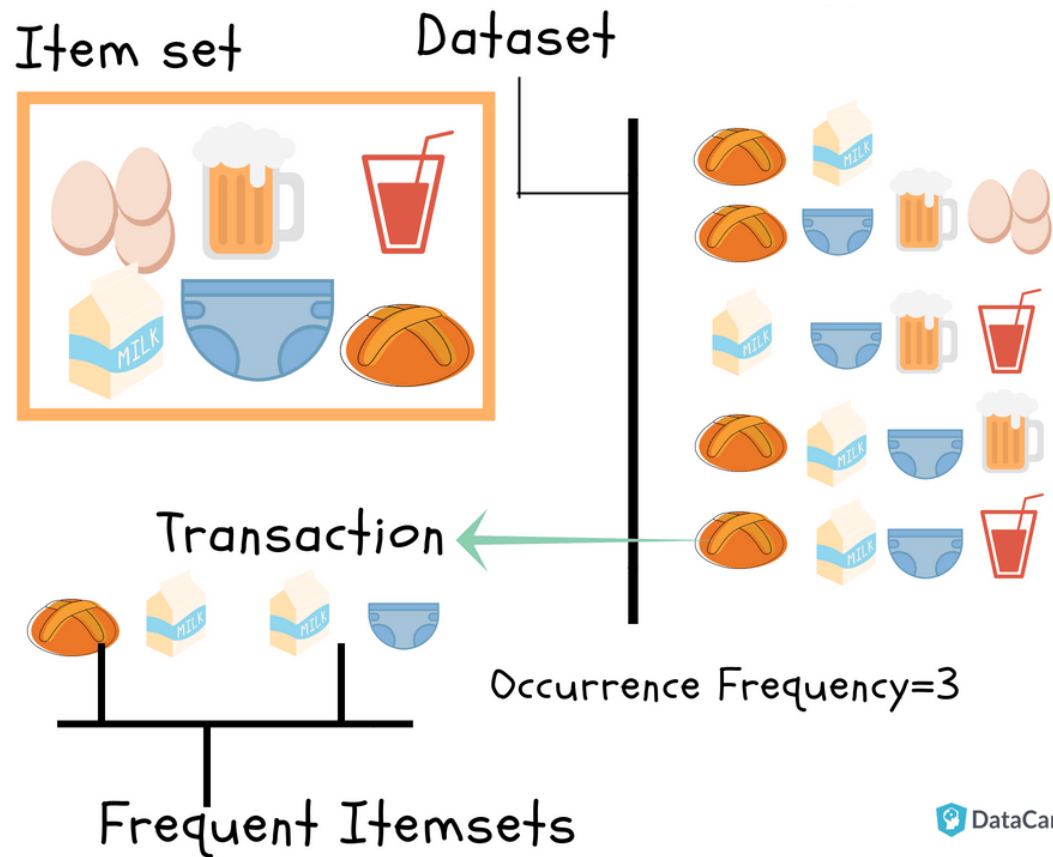


Reglas de Asociación y Dependencia


- **Estudio de textos.** Qué palabras con más frecuencia van asociadas
- **Búsqueda de patrones en páginas web.** Cuáles son los itinerarios más seguidos por los visitantes a las páginas web y utilizar esta información para estructurar las páginas web en el servidor

Suelen ser aplicaciones que llevan asociadas un gran volumen de datos por lo que la **eficiencia es un factor clave.**

Reglas de Asociación y Dependencia



Reglas de Asociación y Dependencia



	Vino "El cabezón"	Gaseosa "Chispa"	Vino "Tío Paco"	Horchata "Xufer"	Bizcochos "Goloso"	Galletas "Trigo"	Chocolate "La vaca"
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

SI bizcochos "Goloso" **Y** horchata "Xufer" **ENTONCES** galletas "Trigo"

Reglas de Asociación y Dependencia

- Una regla de asociación puede ser vista como reglas de la forma:

SI α ENTONCES β

El conjunto α es el predecesor (antecedente, A)
y el conjunto β es el sucesor (consecuente, C)

- Formalmente:

- Sea I un conjunto de ítems
- T un conjunto de transacciones con ítems en I
- La regla de asociación

$$\alpha \Rightarrow \beta \quad \text{con } \alpha, \beta \subseteq I, \alpha, \beta \neq \emptyset \text{ y } \alpha \cap \beta = \emptyset$$

significa que cada transacción de T que contiene a α contiene a β

Reglas de Asociación y Dependencia

- Medidas para la calidad (o importancia) de una regla:
 - **Soporte o cobertura:** porcentaje de asociaciones que la regla predice correctamente

Sobre un conjunto de ítems X (*itemset*): frecuencia con la que ocurre en la base de datos (también puede expresarse como recuento total)

$$\text{Soporte}(X) = \frac{\text{Nº de ocurrencias de } X}{\text{Nº total de transacciones en la BD}}$$

Considerando una regla: número de transacciones que contienen la unión de los ítems de la regla

$$\text{Soporte}(A \Rightarrow C) = \text{Soporte}(A \cup C) = \frac{\text{Nº de ocurrencias de } A \cup C}{\text{Nº total de transacciones en la BD}}$$

Reglas de Asociación y Dependencia

- Medidas para la calidad (o importancia) de una regla:
 - **Confianza o precisión:** porcentaje de veces que la regla se cumple cuando se puede aplicar

$$\text{Confianza } (A \Rightarrow C) = \frac{\text{Soporte } (A \cup C)}{\text{Soporte}(A)}$$

Puede verse como una probabilidad condicionada:

$$\text{prob}(C|A) = \frac{\text{prob}(A \wedge C)}{\text{prob}(A)}$$

Reglas de Asociación y Dependencia

	Vino "El cabezón"	Gaseosa "Chispa"	Vino "Tío Paco"	Horchata "Xufer"	Bizcochos "Goloso"	Galletas "Trigo"	Chocolate "La vaca"
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

SI bizcochos "Goloso" **Y** horchata "Xufer" **ENTONCES** galletas "Trigo"

- Soporte igual a 3 (o $3/10$)
- Confianza del 75% ($3/4$): se aplica correctamente 3 de las 4 veces en que se cumple el antecedente

Reglas de Asociación y Dependencia

SI bizcochos "Goloso" Y horchata "Xufer" **ENTONCES** galletas "Trigo"

Reglas derivadas del itemset {Goloso, Xufer, Trigo}

$\{\text{Goloso}\} \Rightarrow \{\text{Xufer}, \text{Trigo}\}$

$\{\text{Goloso}, \text{Xufer}\} \Rightarrow \{\text{Trigo}\}$

$\{\text{Xufer}\} \Rightarrow \{\text{Goloso}, \text{Trigo}\}$

$\{\text{Goloso}, \text{Trigo}\} \Rightarrow \{\text{Xufer}\}$

$\{\text{Trigo}\} \Rightarrow \{\text{Xufer}, \text{Goloso}\}$

$\{\text{Xufer}, \text{Trigo}\} \Rightarrow \{\text{Goloso}\}$

Observaciones

- Todas las reglas posibles se obtienen dividiendo en dos partes el itemset
- Todas las reglas tienen el mismo soporte, aunque su confianza pueda variar
- Interesan las reglas que tienen un **soporte por encima de un umbral**

Reglas de Asociación y Dependencia

Generación de reglas

Fuerza bruta 

1. Obtener todas las combinaciones de items y crear reglas
2. Calcular el soporte y la confianza de cada regla
3. Eliminar las reglas que no superan los umbrales de soporte y confianza

Solución en dos etapas

1. Generar itemsets frecuentes (por encima del umbral de soporte)
2. Generar reglas de asociación con los itemsets del paso 1. (por encima del umbral de confianza)

Algoritmo Apriori

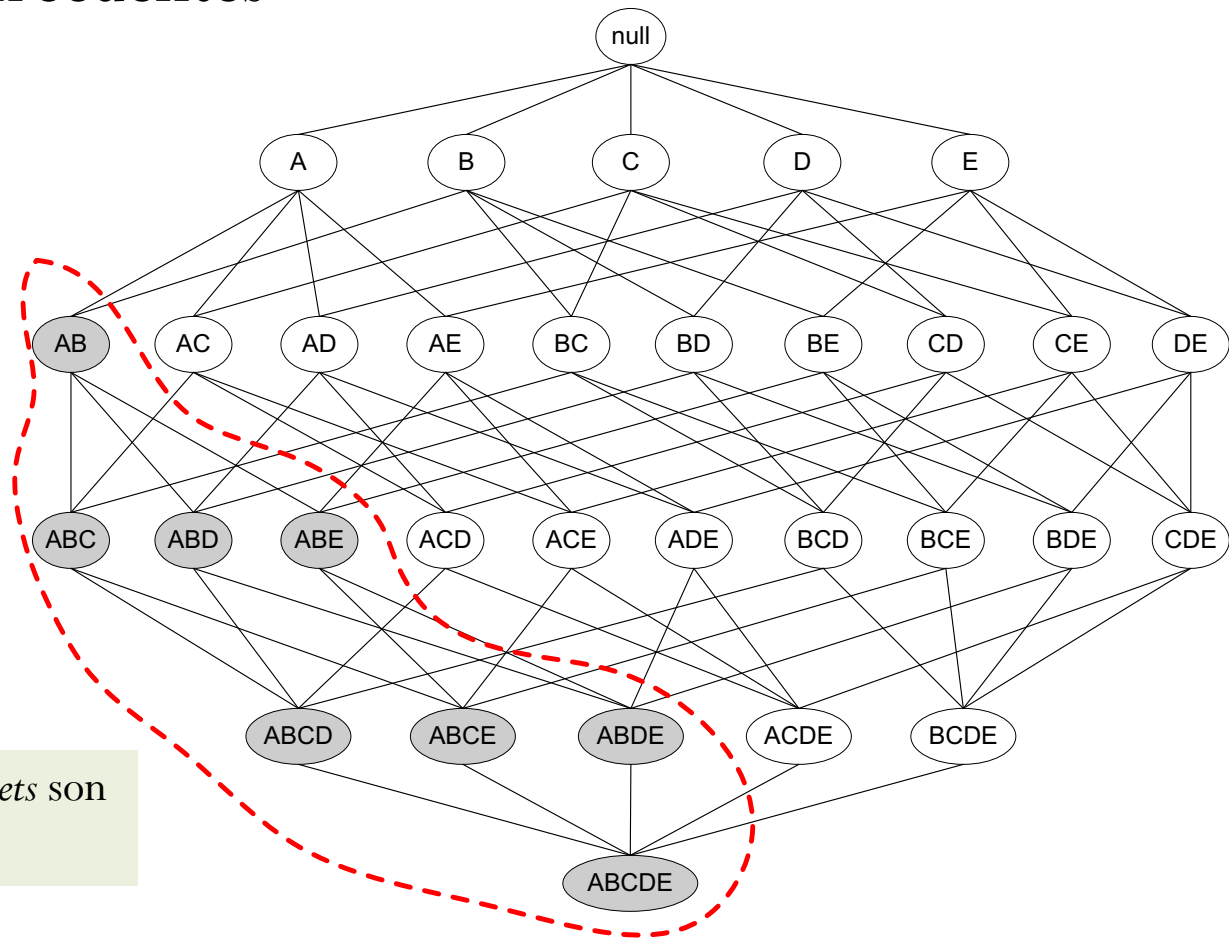
- Se basa en la búsqueda de los conjuntos de ítems con un determinado soporte
 1. En primer lugar, se construyen conjuntos formados por **un solo ítem ($n=1$)** que supera el soporte mínimo
 2. Este conjunto de conjuntos se utiliza para construir el conjunto de conjuntos **de dos ítems ($n=2$)** que superan el soporte mínimo
 3. **El procedimiento se repite** sucesivamente hasta que se llegue a un **tamaño n** en el cual **no existan** conjuntos de ítems **que superen el soporte mínimo exigido**
 4. Finalmente, se construyen **reglas que superen un nivel de confianza mínimo**

Algoritmo Apriori

- Reducción del número de candidatos (propiedad Apriori)
 - Si un itemset es frecuente, también lo son todos sus subconjuntos
 - ¿Por qué? Porque el soporte de un itemset nunca puede ser mayor que el de cualquiera de sus subconjuntos:
$$\forall X, Y \text{ itemsets}, X \subseteq Y \Rightarrow \text{Soporte}(X) \geq \text{Soporte}(Y)$$
 - Formalmente, esta propiedad se conoce con el nombre de anti-monotonía del soporte.

Algoritmo Apriori

Itemsets frecuentes



AB no frecuente

Todos estos *itemsets* son no frecuentes

	Vino "El cabezón"	Gaseosa "Chispa"	Vino "Tío Paco"	Horchata "Xufer"	Bizcochos "Goloso"	Galletas "Trigo"	Chocolate "La vaca"
T1	1	1	0	0	0	1	0
T2	0	1	1	0	0	0	0
T3	0	0	0	1	1	1	0
T4	1	1	0	1	1	1	1
T5	0	0	0	0	0	1	0
T6	1	0	0	0	0	1	1
T7	0	1	1	1	1	0	0
T8	0	0	0	1	1	1	1
T9	1	1	0	0	1	0	1
T10	0	1	0	0	1	0	0

Escogemos un soporte mínimo = 2

- 1 • Tomar ítems que aparecen ≥ 2 veces (todos)
- 2 • Pasamos a los conjuntos de dos items que aparecen ≥ 2 (15) :
 {Cabezón, Chispa}, {Cabezón, Goloso}, {Cabezón, Trigo},
 {Cabezón, La Vaca}, {Chispa, Paco}, {Chispa, Xufer}, {Chispa,
 Goloso}, {Chispa, Trigo}, {Chispa, La vaca}, {Xufer, Goloso},
 {Xufer, Trigo}, {Xufer, La Vaca}, {Goloso, Trigo}, {Goloso, La
 Vaca}, {Trigo, La Vaca}
- 3 • Tendremos once conjuntos de tres ítems
- Dos conjuntos de cuatro ítems

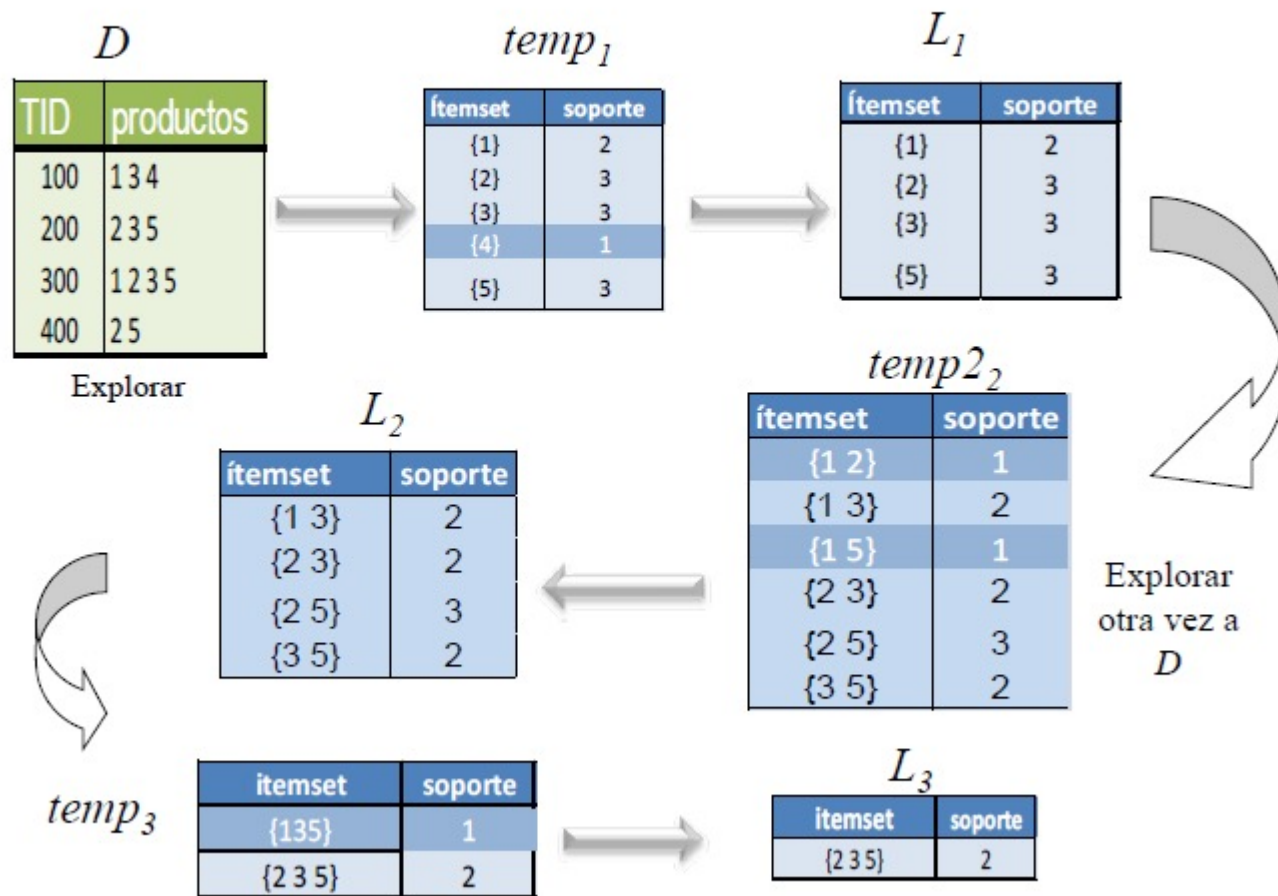
4

- Una vez se han seleccionado los conjuntos de ítems que cumplen con el soporte mínimo, el siguiente paso consiste en extraer de estos conjuntos de reglas las que tengan un nivel de confianza mínimo.

Por ejemplo, de {Xufer, Goloso, Trigo}

SI bizcochos "Goloso" Y horchata "Xufer" ENTONCES galletas "Trigo"	Cb=4, Cf=3/4
SI bizcochos "Goloso" Y galletas "Trigo" ENTONCES horchata "Xufer"	Cb=3, Cf=3/3
SI galletas "Trigo" Y horchata "Xufer" ENTONCES bizcochos "Goloso"	Cb=3, Cf=3/3
SI galletas "Trigo" ENTONCES bizcochos "Goloso" Y horchata "Xufer"	Cb=6, Cf=3/6
SI bizcochos "Goloso" ENTONCES horchata "Xufer" Y galletas "Trigo"	Cb=6, Cf=3/6
SI horchata "Xufer" ENTONCES bizcochos "Goloso" Y galletas "Trigo"	Cb=4, Cf=3/6
SI \emptyset ENTONCES bizcochos "Goloso" Y galletas "Trigo" Y horchata "Xufer"	Cb=10, Cf=3/10

Algoritmo Apriori : Ejemplo con minsup = 2



Algoritmo Apriori

- Extracción de reglas

- ¿Es la confianza anti-monótona como el soporte?

NO: La confianza de $ABC \Rightarrow D$ puede ser mayor o menor que la confianza de $AB \Rightarrow D$.

- Pero la confianza de las reglas generadas de un mismo itemset tienen una propiedad antimonótona:

$$\text{confianza}(ABC \Rightarrow D) \geq \text{confianza}(AB \Rightarrow CD) \geq \text{confianza}(A \Rightarrow BCD)$$

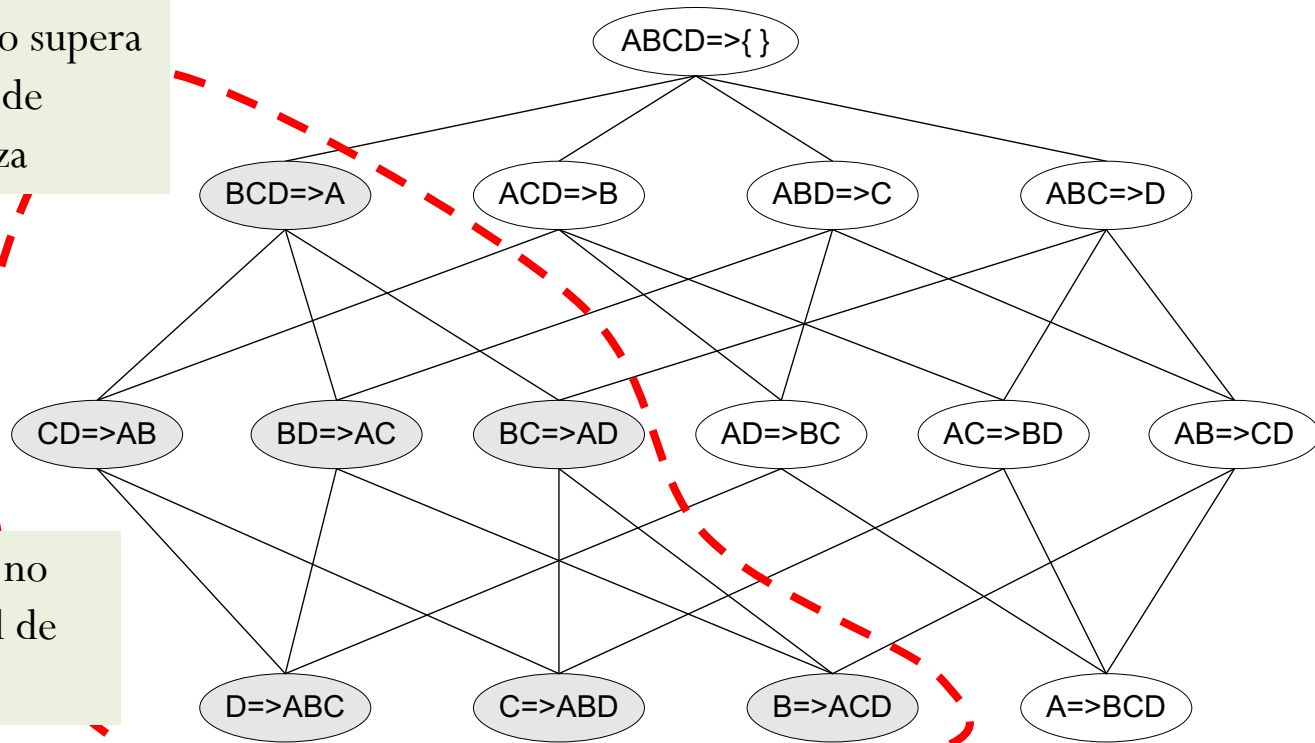
- La confianza es antimonótona con respecto al número de ítems en la parte derecha de la regla

Algoritmo Apriori

Reglas con alta confianza (a partir de un itemset)

Regla no supera
umbral de
confianza

Todas estas reglas no
superan el umbral de
confianza



Algoritmo Apriori

- Establecemos un soporte mínimo=3
- Tomemos el siguiente ejemplo:

Transacciones

Pan, leche, pañales

Pan, pañales, cerveza, huevos

Leche, pañales, cerveza, refresco, café

Pan, leche, pañales, cerveza

Pan, refresco, leche, pañales

- Calculamos la frecuencia para cada ítem, con un solo elemento:

Itemset n=1	Frecuencia
Cerveza	3
Pan	4
Refresco	2
Pañales	5
Leche	4
Huevos	1
Café	1

Puesto que los ítems refresco, huevos y café, no superan el soporte = 3, los eliminamos del grupo, quedándonos con el resto

Itemset freq. n=1	Frecuencia
Cerveza	3
Pan	4
Pañales	5
Leche	3

- Generamos combinaciones de 2 ítems con Cerveza, Pan, Pañales, Leche

Transacciones

Pan, leche, pañales

Pan, pañales, cerveza, huevos

Leche, pañales, cerveza, refresco, café

Pan, leche, pañales, cerveza

Pan, refresco, leche, pañales

Itemset n=2	Frecuencia
Cerveza, pan	2
Cerveza, pañales	3
Cerveza, leche	2
Pan, pañales	4
Pan, leche	3
Pañales, leche	4

- Eliminamos los pares $\{\text{Cerveza, Pan}\}$ $\{\text{Cerveza, Leche}\}$ que no cumplen el soporte mínimo y ya tenemos nuestro conjunto de ítems frecuentes de tamaño 2

Itemset freq. n=2	Frecuencia
Cerveza, pañales	3
Pan, pañales	4
Pan, leche	3
Pañales, leche	4

- Ahora a partir de nuestro conjunto de ítems, pasamos a formar conjuntos con tres ítems:

Itemset freq. n=2
Cerveza, pañales
Pan, pañales
Pan, leche
Pañales, leche

Transacciones
Pan, leche, pañales
Pan, pañales, cerveza, huevos
Leche, pañales, cerveza, refresco, café
Pan, leche, pañales, cerveza
Pan, refresco, leche, pañales

Itemset n=3	Frecuencia
Cerveza, pañales, pan	2
Cerveza, Pañales, Leche	2
Pan, pañales, leche	3
Pan, Leche, Cerveza	1

- Si observamos el soporte, vemos que únicamente supera la tripleta pan, pañales, leche. El siguiente paso sería:

Itemset frecuentes $n=3$	Frecuencia
Pan, pañales, leche	3

- Ahora tendríamos que construir los de cuatro items.
Partiendo de la anterior {pan, pañales y leche}. Pero como no hay, hemos acabado.
- A partir de ahora nos queda construir las reglas a partir de los conjuntos anteriores.

Itemset frecuentes n=3

Pan, pañales, leche

Itemset frecuentes n=2

Cerveza, pañales

Pan, pañales

Pan, leche

Pañales, leche

Transacciones

Pan, leche, pañales

Pan, pañales, cerveza, huevos

Leche, pañales, cerveza, refresco, café

Pan, leche, pañales, cerveza

Pan, refresco, leche, pañales

Itemset frecuentes n=3

Frecuencia

Pan, pañales, leche

3

$$\text{Confianza } (A \Rightarrow C) = \frac{\text{Soporte } (A \cup C)}{\text{Soporte}(A)}$$

Confianza mínima = 75%

- {pan, pañales} \Rightarrow leche 3/4 (75%) ✓
 - 3 repeticiones {pan, pañales y leche}
 - 4 repeticiones {pan, pañales}
- {pan} \Rightarrow {pañales, leche} 3/4 (75%) ✓
- {pañales} \Rightarrow {pan, leche} 3/5 (60%) ✗
- {pan, leche} \Rightarrow {pañales} 3/5 (60%) ✗
- {pañales, leche} \Rightarrow {pan} 3/4 (75%) ✓
 - {pañales} \Rightarrow {pan, leche} 3/5 (60%) ✗ (ya explorada)
 - {leche} \Rightarrow {pan, pañales} 3/4 (75%) ✓

Transacciones

Pan, leche, pañales

Pan, pañales, cerveza, huevos

Leche, pañales, cerveza, refresco, café

Pan, leche, pañales, cerveza

Pan, refresco, leche, pañales

Itemset freq. n=2	Frecuencia
Cerveza, pañales	3
Pan, pañales	4
Pan, leche	3
Pañales, leche	4

$$\text{Confianza } (A \Rightarrow C) = \frac{\text{Soporte } (A \cup C)}{\text{Soporte}(A)}$$

Confianza mínima = 75%

- $\{\text{cerveza}\} \Rightarrow \{\text{pañales}\}$
- ...

Ejemplo

L1	L2	L3	L4	L5
1	1			1
	1		1	
	1	1		
1	1		1	
1		1		
	1	1		
1		1		
1	1	1		1
1	1	1		

C1

ITEMSET	SUP-COUNT
{L1}	
{L2}	
{L3}	
{L4}	
{L5}	

C1

SOPORTE = 2

ITEMSET	SUP-COUNT
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

C1

SOPORTE = 2




PASAN TODOS


ITEMSET	SUP-COUNT
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

C2 SOPORTE = 2

ITEMSET	SUP-COUNT
{L1,L2}	4
{L1,L3}	4
{L1,L4}	1
{L1,L5}	2
{L2,L3}	4
{L2,L4}	2
{L2,L5}	2
{L3,L4}	0
{L3,L5}	1
{L4,L5}	0

C2 SOPORTE = 2  Pasan 6

ITEMSET	SUP-COUNT
{L1,L2}	4
{L1,L3}	4
{L1,L4}	1
{L1,L5}	2
{L2,L3}	4
{L2,L4}	2
{L2,L5}	2
{L3,L4}	0
{L3,L5}	1
{L4,L5}	0

C2 SOPORTE = 2  Pasan 6

ITEMSET	SUP-COUNT
{L1,L2}	4
{L1,L3}	4
{L1,L5}	2
{L2,L3}	4
{L2,L4}	2
{L2,L5}	2

C3 SOPORTE = 2  Pasan 2

ITEMSET	SUP-COUNT
{L1,L2,L3}	2
{L1,L2,L5}	2

C4 SOPORTE = 2  Pasan 0

ITEMSET	SUP-COUNT
{L1,L2,L3,L5}	1

No es frecuente

- Consideremos el item frecuente $\{L1, L2, L5\}$
- Las reglas que se pueden generar son:

SI L1 AND L2 THEN L5,	CONFIANZA=2/4 = 50%
SI L1 AND L5 THEN L2	CONFIANZA=2/2=100%
SI L2 AND L5 THEN L1	CONFIANZA=2/2=100%
IF L1 THEN L2 AND L5	CONFIANZA=2/6=33%
IF L2 THEN L1 AND L5	CONFIANZA=2/7=29%
IF L5 THEN L1 AND L2	CONFIANZA=2/2=100%

Medidas de Interés para la evaluación

Definimos:

- a : número de transacciones en la BD que contiene el antecedente y consecuente a la vez.
- b : número que contienen el antecedente pero no el consecuente
- c : número que no contienen al antecedente pero si el consecuente
- d : número de transacciones que no contienen ni el antecedente ni el consecuente.

Medidas de Interés para la evaluación.

Soporte $\frac{a}{a + b + c + d}$ Confianza $\frac{a}{a + b}$

Interés o Lift $\frac{a(a + b + c + d)}{(a + b)(a + c)}$

Lift = soporte observado / soporte esperado
Capacidad de "predicción" respecto a una asociación
tomada de forma aleatoria
Lift > 1 denota regla interesante

Medidas de Interés para la evaluación.

Confianza Centrada $\frac{ad - bc}{(a + b + c + d) + (a + b)}$

PIATETSKY-SHAPIRO $\frac{ad - bc}{(a + b + c + d)}$

Mínima Contradicción $\frac{a - b}{a + c}$

Problemas

- Soporte demasiado bajo
 - Cuando los valores de las transacciones son muy específicos puede llegarse a no obtener itemsets frecuentes (producto a nivel de marca, proveedor a nivel de nombre etc.)
=> Reglas difusas, reglas jerárquicas
- Demasiadas reglas y difíciles de interpretar
 - Cuando hay demasiados valores en las dimensiones se obtienen demasiadas reglas (explosión combinatoria)
 - Es necesario controlar la especificidad de las reglas
=> Visualización de reglas
- Asociaciones no útiles por BD con sesgo
 - Por ejemplo: Iphone \Rightarrow Auriculares vs Auriculares \Rightarrow Iphone