



ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

EDUCACIÓN  
PROFESIONAL

# Programación en R para ciencia de datos

DBDC-2022

Educación Profesional  
Escuela de Ingeniería

Profesor:

**Miguel Jorquera Viguera**





## REGLAS DE ASOCIACIÓN

# Objetivo

- Generar “reglas” que asocien productos.
- Estas reglas deben ser:
  - Frecuentes
  - Razonables.



# Definiciones

{Zapatos, cartera}  $\longrightarrow$  {Traje de Baño}

Conceptos claves:

- Item
- Itemset
- Antecedente
- Consecuente
- Regla de asociación

Métricas claves:

- Support
- Confidence
- Lift

# Definiciones

Reglas basadas en probabilidades.

- $Supp(\{a, b\}) = \frac{\# \text{ Transacciones que contienen } a \text{ y } b}{\# \text{ Transacciones}}$
- $Conf(\{a, b\} \rightarrow \{c\}) = \frac{Supp(\{a, b, c\})}{Supp(\{a, b\})} = \hat{P}(\{c\} | \{a, b\})$



# Definiciones

¿Qué hace “buena” a una regla?

Debe ser común:

$$Supp(\{a, b\}) \geq \theta$$

Debe ser razonable:

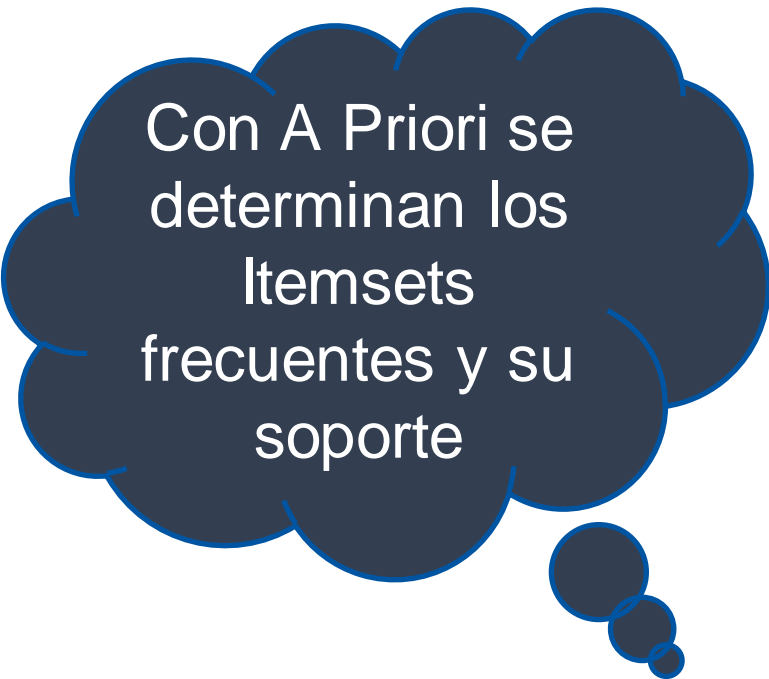
$$Conf(\{a, b\} \rightarrow \{c\}) \geq \text{minconf}$$

¿Cómo generar  
las reglas?

## Algoritmo apriori

Algoritmo:

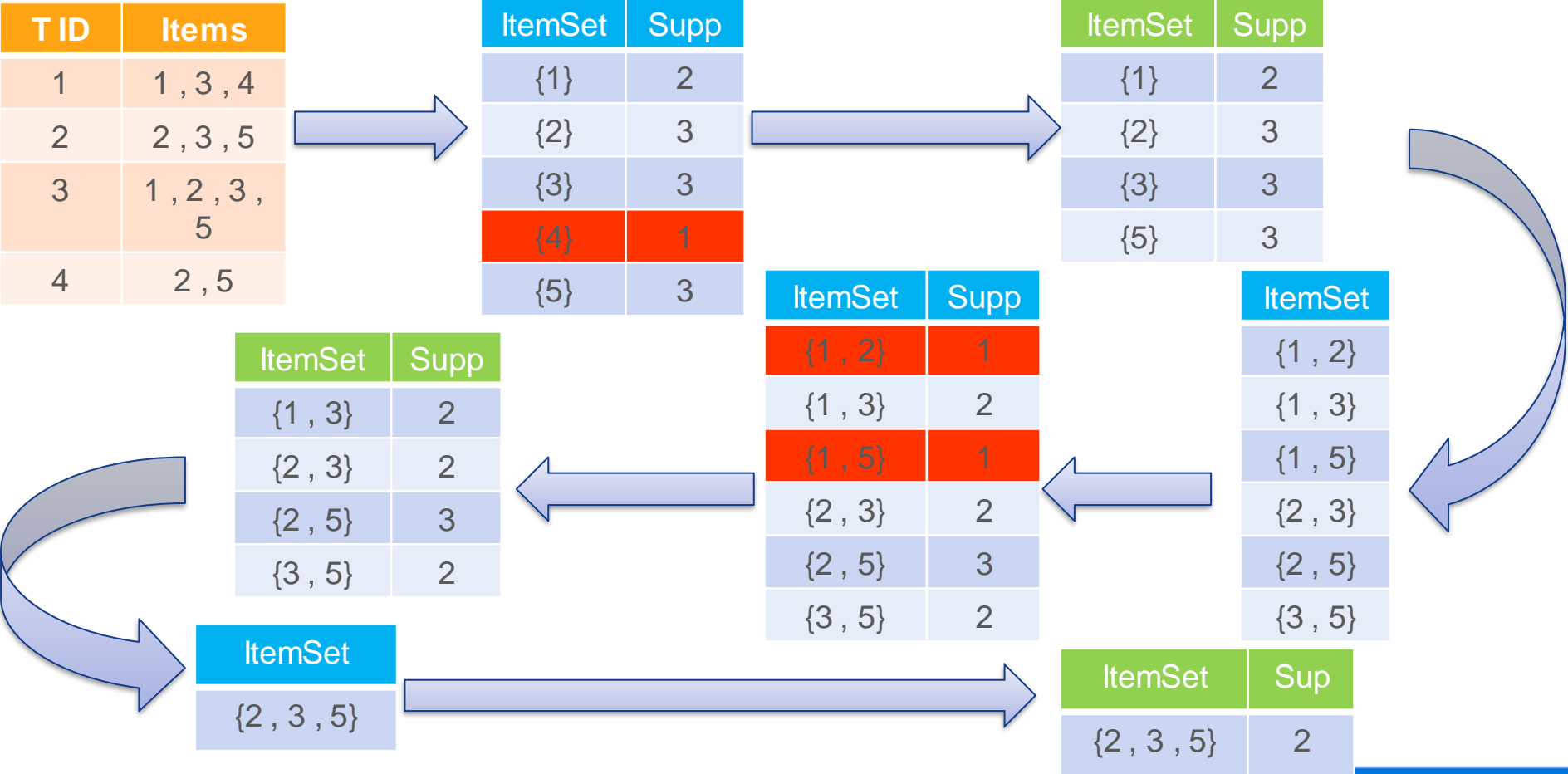
- Se buscan los itemset de un item y se filtran aquellos con soporte mayor o igual que  $\theta$
- Repetir hasta que no se puedan formar nuevos Itemsets:
  - Crea itemsets candidatos: Para cada par de itemsets ya listados con  $k$  elementos, combinarlos si comparten  $k-1$  elementos.
  - Poda: Retener candidato si tiene un soporte de al menos  $\theta$  para definir la lista con itemset con  $k+1$  elementos.
  - Fin: si la lista de itemsets con  $k+1$  elementos es vacía.



Con A Priori se  
determinan los  
Itemsets  
frecuentes y su  
soporte

# Algoritmo apriori

$$\theta = \frac{1}{4} = 0.25$$





## Algoritmo apriori

Itemsets

Itemset	Supp
{1}	2
{2}	3
{3}	3
{5}	3
{1, 3}	2
{2, 3}	3
{2, 5}	3
{3, 5}	2
{2, 3, 5}	2

¿Qué reglas escogemos?

Reglas de asociación

Regla	Confidence	Regla	Confidence
$1 \rightarrow 3$	$2/2 = 1$	$5 \rightarrow 3$	$2/3 = 0.66$
$2 \rightarrow 3$	$3/3 = 1$	$\{2,3\} \rightarrow 5$	$2/3 = 0.66$
$2 \rightarrow 5$	$3/3 = 1$	$\{3,5\} \rightarrow 2$	$2/2 = 1$
$3 \rightarrow 5$	$2/3 = 0.66$	$\{2,5\} \rightarrow 3$	$2/3 = 0.66$
$3 \rightarrow 1$	$2/3 = 0.66$	$5 \rightarrow \{2,3\}$	$2/3 = 0.66$
$3 \rightarrow 2$	$3/3 = 1$	$2 \rightarrow \{3,5\}$	$2/3 = 0.66$
$5 \rightarrow 2$	$3/3 = 1$	$3 \rightarrow \{2,5\}$	$2/3 = 0.66$

## Algoritmo apriori

¿Qué reglas son preferibles?

- Ordenar por confidence:

$$Conf(a \rightarrow b) = \hat{P}(b|a) = \frac{Supp(a \cup b)}{Supp(a)}$$

- Ordenar por lift:

$$Lift(a \rightarrow b) = \frac{Conf(a \rightarrow b)}{Supp(b)} = \frac{\hat{P}(a \cup b)}{\hat{P}(a)\hat{P}(b)}$$

## Algoritmo apriori

¿Qué reglas son preferibles?

- Ordenar por confidence:

$$Conf(a \rightarrow b) = \hat{P}(b|a) = \frac{Supp(a \cup b)}{Supp(a)}$$

- Ordenar por lift:

$$Lift(a \rightarrow b) = \frac{Conf(a \rightarrow b)}{Supp(b)} = \frac{\hat{P}(a \cup b)}{\hat{P}(a)\hat{P}(b)}$$

## Algoritmo apriori

Wikipedia:

“Lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values:”

$$Lift = \frac{\text{target response}}{\text{average response}}$$

## Algoritmo apriori

Wikipedia:

Por ejemplo,

En una población la tasa de respuesta es de un 5%, pero cierto modelo (o regla) logra identificar un segmento con una tasa de respuesta de un 20%. Entonces dicho segmento tiene un lift de 4.0 ( $20\%/5\%$ ).



**Vamos!**

