# TDVI: Inteligencia Artificial

## Arules

UTDT - LTD

# Reglas de asociación

MERCADONA S.A.

AVDA. CADIZ, 93 - C/ RONDA
CORDOBA
Teléfono 957481935
CIF: A-46103834
P.V.P. I.V.A. INCLUIDO
01/10/2013 09:35 OP: 078910
N: 969951 CAJA:010 SUP: 2145

| | | | |
|---|---|---|---|
| 1 | 1 | ROLLOS COCINA | 2,35 |
| 2 | 1 | C.POLLO/PAVO/L | 4,25 |
| 3 | 1 | PAN BLANCO 2B | 1,00 |
| 4 | 1 | ESPINACAS | 1,19 |
| 5 | 1 | LECHUGA CORAZN | 1,55 |
| 6 | 1 | QUESO RALLADO | 1,15 |
| 7 | 1 | NATA COCINAR | 1,05 |

TOTAL.....EUROS 12,54
ENTREGA.....EUROS 20,00

DEVOLUCIÓN..EUROS 7,46

DETALLE

| IVA | Base Imponible | Cuota |
|---|---|---|
| 4 % | 3,74 | 0,15 |
| 10 % | 5,73 | 0,57 |
| 21 % | 1,94 | 0,41 |
| TOTAL: | 11,41 | 1,13 |

LE ATENDIÓ: TRINI

# Reglas de asociación

Dado un conjunto de transacciones se quieren encontrar reglas que predigan la ocurrencia de un(nos) ítem(s) a partir de otros ítems de la transacción

Transacciones de una cesta de compras
(Market-Basket transactions)

| *TID* | *Items* |
|---|---|
| 1 | Pan, Leche |
| 2 | Pan, Pañales, Cerveza, Huevos |
| 3 | Leche, Pañales, Cerveza, Coca |
| 4 | Pan, Leche, Pañales, Cerveza |
| 5 | Pan, Leche, Pañales, Coca |

Ejemplo de reglas de asociación:

{Pañales} $\rightarrow$ {Cerveza}
{Leche, Pan} $\rightarrow$ {Huevos, Coca}
{Cerveza, Pan} $\rightarrow$ {Leche}

# Reglas de asociación

- ***Itemset***
  - Colección de uno o más items
    - Ejemplo: {Leche, Pan, Pañales}
  - *k*-itemset
    - Itemset con *k* items
- ***Support count* (σ) de un itemset**
  - Cantidad de transacciones que contienen al itemset
  - Ejemplo σ({Leche, Pan, Pañales}) = 2
- ***Support* de un itemset**
  - Transacción que contienen al itemset / Total de transacciones
  - Ejemplo: s({Leche, Pan, Pañales}) = 2/5
- ***Frequent Itemsets* (dado un *minsup*)**
  - Itemsets cuyo soporte es mayor o igual que un umbral *minsup (minimo soporte)*

| *TID* | *Items* |
|---|---|
| 1 | Pan, Leche |
| 2 | Pan, Pañales, Cerveza, Huevos |
| 3 | Leche, Pañales, Cerveza, Coca |
| 4 | Pan, Leche, Pañales, Cerveza |
| 5 | Pan, Leche, Pañales, Coca |

# Reglas de asociación

***Regla de asociación***

- Implicación de la forma X → Y, donde X e Y son itemsets
- Ejemplo: {Leche, Pañales} → {Cerveza}

***Dos métricas asociadas a una regla***

- Soporte (*Support*) (s)
  - Porcentaje de transacciones que contienen a X e Y sobre el total de transacciones
- Confianza *(Confidence)* (c)
  - Cantidad de transacciones que contienen a X e Y sobre las que contienen a X. Mide la frecuencia de ocurrencia de los ítems de Y en las transacciones que contienen a X

| *TID* | *Items* |
|---|---|
| 1 | Pan, Leche |
| 2 | Pan, Pañales, Cerveza, Huevos |
| 3 | Leche, Pañales, Cerveza, Coca |
| 4 | Pan, Leche, Pañales, Cerveza |
| 5 | Pan, Leche, Pañales, Coca |

$$\{\text{Leche}, \text{Pañales}\} \Rightarrow \text{Cerveza}$$

$$s = \frac{\sigma(\text{Leche}, \text{Pañales}, \text{Cerveza})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Leche}, \text{Pañales}, \text{Cerveza})}{\sigma(\text{Leche}, \text{Pañales})} = \frac{2}{3} = 0.67$$

# Reglas de asociación

| *TID* | *Items* |
|---|---|
| 1 | Pan, Leche |
| 2 | Pan, Pañales, Cerveza, Huevos |
| 3 | Leche, Pañales, Cerveza, Coca |
| 4 | Pan, Leche, Pañales, Cerveza |
| 5 | Pan, Leche, Pañales, Coca |

**Ejemplos de reglas:**

{Leche, Pañales} → {Cerveza} (s=0.4, c=0.67)
{Leche, Cerveza} → {Pañales} (s=0.4, c=1.0)
{Pañales, Cerveza} → {Leche} (s=0.4, c=0.67)
{Cerveza} → {Leche, Pañales} (s=0.4, c=0.67)
{Pañales} → {Leche, Cerveza} (s=0.4, c=0.5)
{Leche} → {Pañales, Cerveza} (s=0.4, c=0.5)

**Observaciones:**

• Todas estas reglas son particiones binarias del mismo itemset: {Leche, Pañales, Cerveza}

• Las reglas que se originan del mismo itemset tienen el mismo soporte pero pueden tener diferente confianza

# Algoritmo apriori

Problema:

¿Cómo encontrar las reglas que cumplen un soporte mínimo y una confianza mínima?

¿Esto es aprendizaje supervisado o no supervisado?

Hacer todos las combinaciones posibles (fuerza bruta) es muy costoso (reglas posibles: $3^d-2^{d+1}+1$)

Vamos a dividir al problema en dos grandes pasos:

1. Descubrir todos los itemsets frecuentes
2. Descubrir las reglas con suficente confianza contenidas en esos itemsets

# 1. Descubrimiento de todos los itemsets frecuentes

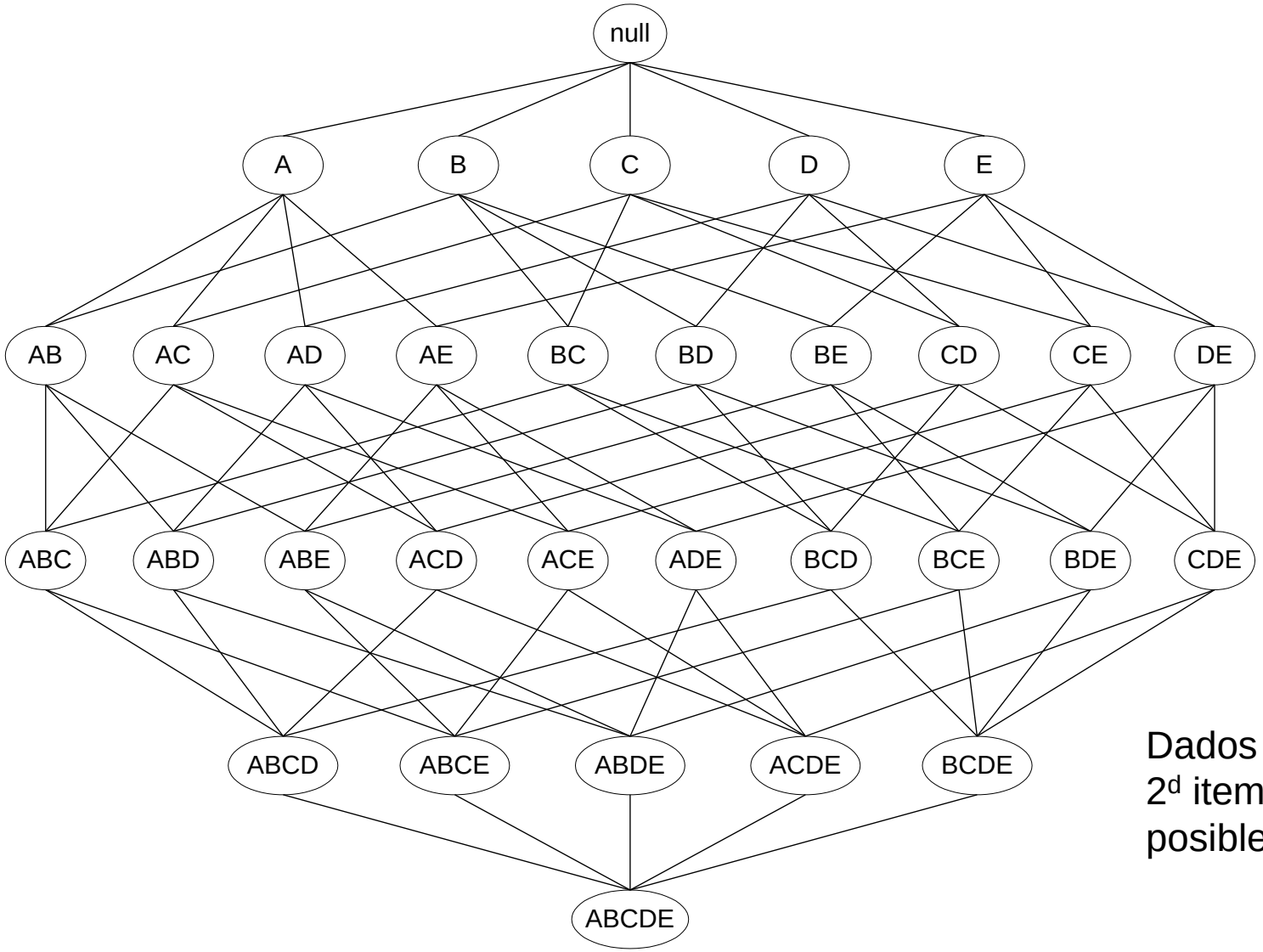
- Se manejan 2 conjuntos de itemsets:
  - Candidatos ($C_k$) y Frecuentes ($L_k$)

- Se aprovecha una propiedad que cumple el soporte:
  - Propiedad anti-monótona de f:

$$X \subseteq Y \rightarrow f(X) \geq f(Y)$$ (siendo X e Y conjuntos de ítems)

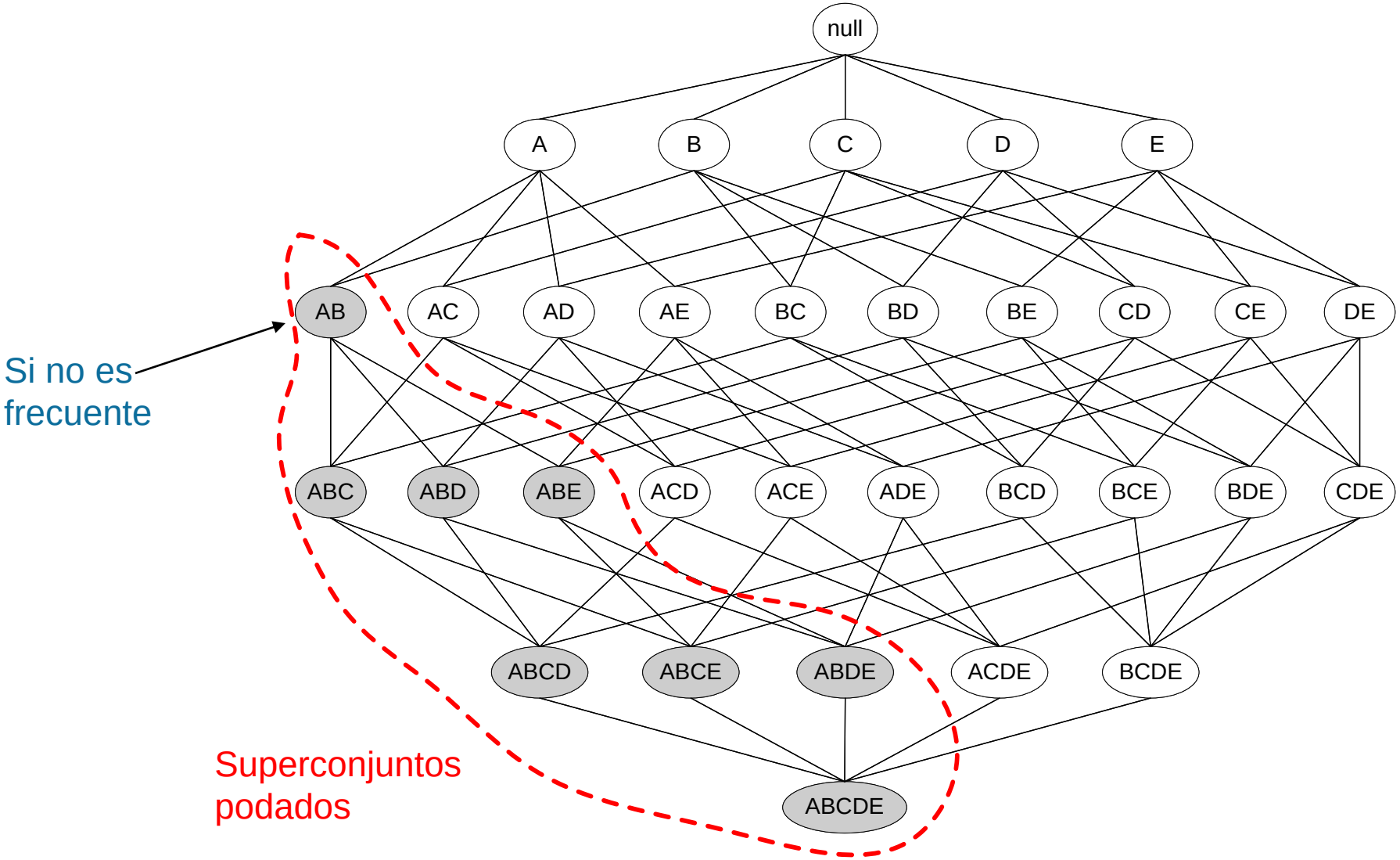
  - ¿Qué implica? Un itemset no puede ser frecuente si algún itemset contenido en él no lo es.

# 1. Descubrimiento de todos los itemsets frecuentes




Dados d items, existen $2^d$ itemsets como posibles candidatos

# 1. Descubrimiento de todos los itemsets frecuentes



# 1. Descubrimiento de todos los itemsets frecuentes

Algoritmo:

Create $L_1$ = set of supported itemsets of cardinality one
Set $k$ to 2
while ($L_{k-1} \neq \emptyset$) {
    Create $C_k$ from $L_{k-1}$
    Prune all the itemsets in $C_k$ that are not
        supported, to create $L_k$
    Increase $k$ by 1
}
The set of all supported itemsets with at least two members is $L_2 \cup \cdots \cup L_{k-2}$

# 1. Descubrimiento de todos los itemsets frecuentes

Ilustración:

| Item | Cantidad |
|---|---|
| **Cerveza** | **3** |
| **Coca** | **2** |
| **Huevos** | **1** |
| **Leche** | **4** |
| **Pan** | **4** |
| **Pañales** | **4** |

1-itemsets

| Itemset | Cantidad |
|---|---|
| **{Cerveza, Leche}** | **2** |
| **{Cerveza, Pan}** | **2** |
| **{Cerveza, Pañales}** | **3** |
| **{Leche, Pan}** | **3** |
| **{Leche, Pañales}** | **3** |
| **{Pan, Pañales}** | **3** |

2-itemsets

(No es necesario generar los candidatos que involucran Coca o Huevos)

conteo de soporte mínimo = 3

3-itemset

| Itemset | Cantidad |
|---|---|
| **{Leche, Pan, Pañales}** | **3** |

¿Por qué no {Cerveza, Leche, Pan}?

# 1. Descubrimiento de todos los itemsets frecuentes

Factores que influyen en la complejidad computacional del algoritmo:

- Elección del umbral mínimo de soporte
- Número de ítems en el conjunto de datos
- Cantidad de transacciones
- Cantidad promedio de ítems por transacción

# 2. Descubrimiento de las reglas con suficiente confianza

- Dado un itemset frecuente L, debemos encontrar todos los subconjuntos $f \subset L$ tales que $f \rightarrow L - f$ satisface el requerimiento mínimo de confianza

- Si {A,B,C,D} es un itemset frecuente, las reglas candidatas son:

| | | | |
|---|---|---|---|
| $ABC \rightarrow D$, | $ABD \rightarrow C$, | $ACD \rightarrow B$, | $BCD \rightarrow A$, |
| $A \rightarrow BCD$, | $B \rightarrow ACD$, | $C \rightarrow ABD$, | $D \rightarrow ABC$ |
| $AB \rightarrow CD$, | $AC \rightarrow BD$, | $AD \rightarrow BC$, | $BC \rightarrow AD$, |
| $BD \rightarrow AC$, | $CD \rightarrow AB$, | | |

- Si $|L| = k$, entonces existen $2^k - 2$ reglas de asociación candidatas (se ignoran $L \rightarrow \varnothing$ y $\varnothing \rightarrow L$)

# 2. Descubrimiento de las reglas con suficiente confianza

Propiedad:

$$\text{confianza}(AB \rightarrow C) \geq \text{confianza}(A \rightarrow BC)$$

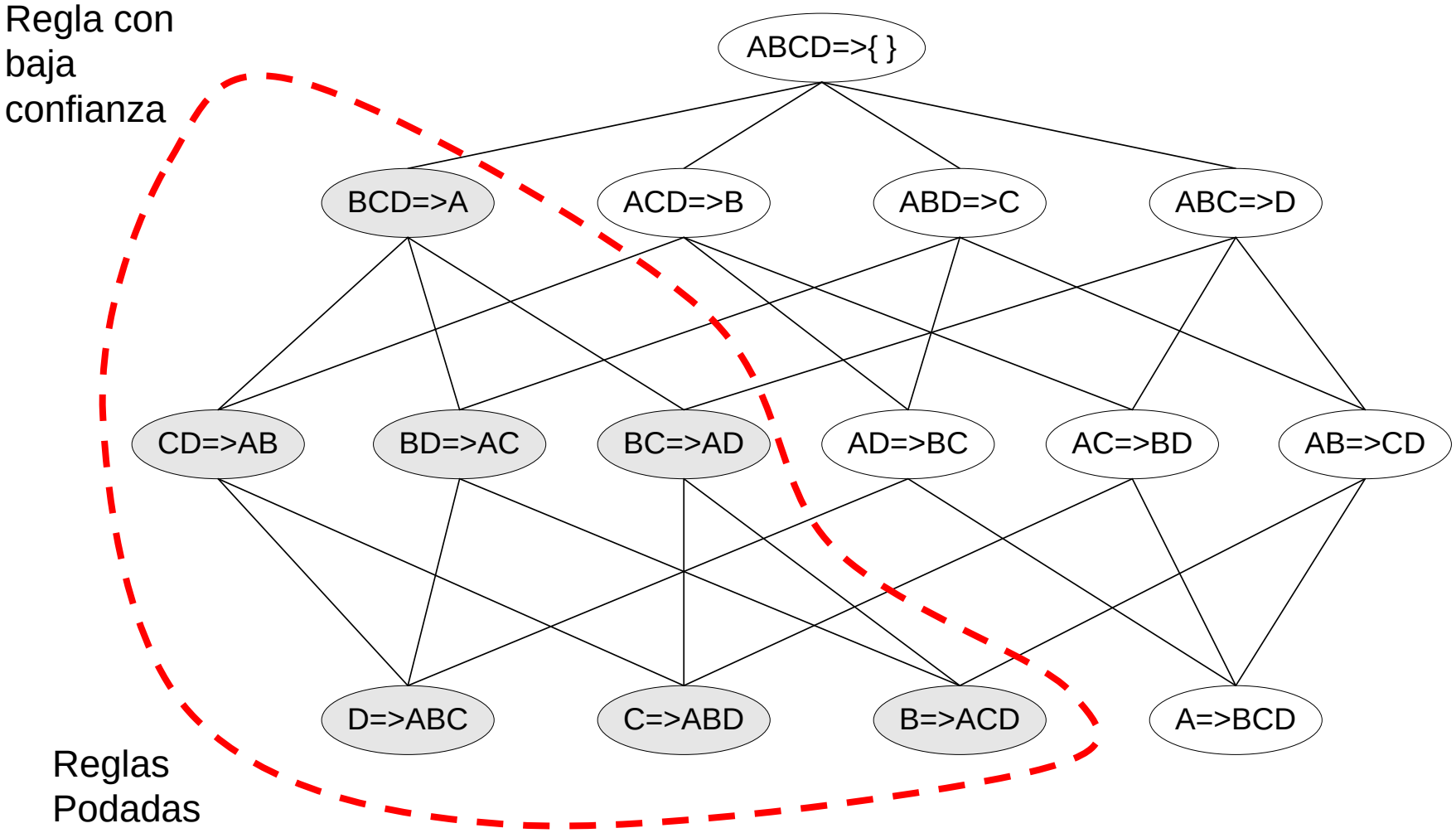
Ver que:

$$\text{confianza}(AB \rightarrow C) = \sigma(ABC) / \sigma(AB) = \text{soporte}(ABC) / \text{soporte}(AB)$$

$$\text{confianza}(A \rightarrow BC) = \sigma(ABC) / \sigma(A) = \text{soporte}(ABC) / \text{soporte}(A)$$

$$\text{soporte}(A) \geq \text{soporte}(AB)$$

# 2. Descubrimiento de las reglas con suficiente confianza



# 2. Descubrimiento de las reglas con suficiente confianza

Los algoritmos de reglas de asociación tienden a producir muchas reglas.

Se pueden utilizar medidas de interés para podar u ordenar las reglas. Las medidas más comunes son:

$$support(L \rightarrow R) = count(L \cup R) / |T|$$

$$confidence(L \rightarrow R) = count(L \cup R) / count(L) = support(L \cup R) / support(L)$$

$$lift(L \rightarrow R) = support(L \cup R) / (support(L) * support(R)) = lift(R \rightarrow L)$$

$$leverage(L \rightarrow R) = support(L \cup R) - support(L) * support(R)$$

$$coverage(L \rightarrow R) = support(L)$$

# 2. Descubrimiento de las reglas con suficiente confianza

¿Supongan que están analizando 2000 transacciones y que obtiene los conteos que muestran la tabla de abajo, cuánto valen support, confidence, lift, leverage y coverage?

| $\text{count}(L)$ | $\text{count}(R)$ | $\text{count}(L \cup R)$ |
|---|---|---|
| 220 | 250 | 190 |

¿Qué implica un lift menor a 1?

# Arules en Python

Probémoslo en Python!

# Bibliografía

- Reglas de asociación: Bramer (Cap 17)