

Minería de Datos

IIC2433

Reglas de asociación
Vicente Domínguez

¿Qué veremos esta clase?

- Reglas de asociación

Reglas de asociación

Base de datos de transacciones

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

¿Para qué compran huevos
nuestros clientes?

{huevos, mantequilla}

¿Para qué compran huevos
nuestros clientes?

{huevos, mantequilla}

quizás para hacer tostadas...

¿Para qué compran huevos nuestros clientes?

{huevos, mantequilla}  pan

quizás para hacer tostadas...

¿Para qué compran huevos
nuestros clientes?

{huevos, harina}

¿Para qué compran huevos
nuestros clientes?

{huevos, harina}

quizás para repostería...

¿Para qué compran huevos
nuestros clientes?

{huevos, harina} → azúcar

quizás para repostería...

Itemsets frecuentes

1. Listar **itemsets** posibles

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- {Pan}
- {Coca Cola}
- {Leche}
- {Pañales}
- {Cerveza}
- {Pan, Coca Cola}
- {Pan, Leche}
- {Pan, Pañales}
- {Pan, Cerveza}
- {Coca Cola, Leche}
- {Coca Cola, Pañales}
- {Coca Cola, Cerveza}
- {Leche, Pañales}
- {Leche, Cerveza}
- {Pañales, Cerveza}
- {Pan, Coca Cola, Leche}
- {Pan, Coca Cola, Pañales}
- {Pan, Coca Cola, Cerveza}
- {Pan, Leche, Pañales}
- {Pan, Leche, Cerveza}
- {Pan, Pañales, Cerveza}
- {Coca Cola, Leche, Pañales}
- {Coca Cola, Leche, Cerveza}
- {Coca Cola, Pañales, Cerveza}
- {Leche, Pañales, Cerveza}
- {Pan, Coca Cola, Leche, Pañales}
- {Pan, Coca Cola, Leche, Cerveza}
- {Pan, Coca Cola, Pañales, Cerveza}
- {Pan, Leche, Pañales, Cerveza}
- {Coca Cola, Leche, Pañales, Cerveza}
- {Pan, Coca Cola, Leche, Pañales, Cerveza}

Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

2. Contar la frecuencia de cada itemset

- $\sigma(\{\text{Pan}\}) = 4$
- $\sigma(\{\text{Coca Cola, Pañales}\}) = 2$
- $\sigma(\{\text{Pan, Leche, Cerveza}\}) = 1$

Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

3. Calcular el **soporte**

- $s(\{\text{Pan}\}) = \frac{4}{5}$
- $s(\{\text{Coca Cola, Pañales}\}) = \frac{2}{5}$
- $s(\{\text{Pan, Leche, Cerveza}\}) = \frac{1}{5}$

Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

4. Todos los itemsets cuyo soporte sea mayor a un **umbral** se consideran **frecuentes**

Por ejemplo, umbral = 0,6

- $s(\{\text{Pan}\}) = \frac{4}{5} = 0,8$
- $s(\{\text{Coca Cola, Pañales}\}) = \frac{2}{5} = 0,4$
- $s(\{\text{Pan, Leche, Cerveza}\}) = \frac{1}{5} = 0,2$

Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

4. Todos los itemsets cuyo soporte sea mayor a un **umbral** se consideran **frecuentes**

Por ejemplo, umbral = 0,6

- $s(\{\text{Pan}\}) = \frac{4}{5} = 0,8$ **frecuente**
- $s(\{\text{Coca Cola, Pañales}\}) = \frac{2}{5} = 0,4$ **no frecuente**
- $s(\{\text{Pan, Leche, Cerveza}\}) = \frac{1}{5} = 0,2$ **no frecuente**

Formalizando los conceptos:

- **Itemset**
 - Una colección de uno o más ítems
 - Ejemplo: {Leche, pan, cerveza}
- **Contador del soporte (σ)**
 - Frecuencia de ocurrencia de un itemset
 - Ej. $\sigma(\{\text{Leche, pan, cerveza}\}) = 1$
- **Soporte**
 - Fracción de las transacciones que contiene un itemset
 - Ej. $s(\{\text{Leche, pan, cerveza}\}) = \frac{1}{5}$
- **Itemset frecuente**
 - Un itemset cuyo soporte es mayor o igual a un determinado umbral.

- $s(\{\text{Pañales, Leche}\}) = 3/5 = 0,6$

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

¿Qué podemos decir sobre las siguientes **reglas de asociación**?

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

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

- $s(\{\text{Pañales}, \text{Leche}\}) = 3/5 = 0,6$

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

¿Qué podemos decir sobre las siguientes **reglas de asociación**?

$c(\{\text{Pañales}\} \Rightarrow \{\text{Leche}\}) =$

$c(\text{Leche} \Rightarrow \{\text{Pañales}\}) =$

Itemset frecuente y regla de asociación son conceptos diferentes

$\{A, B\}$ es frecuente $\neq \{A\} \Rightarrow \{B\} \neq \{B\} \Rightarrow \{A\}$

Formalizando más conceptos:

- **Regla de asociación**

- Una expresión de la forma $X \Rightarrow Y$, donde X e Y son itemsets
- Ejemplo: $\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}$

- **Soporte de una regla de asociación (s)**

- Soporte de la unión del antecedente y el consecuente
- $s(X \Rightarrow Y) = s(X \cup Y)$
- Ejemplo: $s(\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}) = s(\{\text{Leche, pañales, cerveza}\})$

- **Confianza de una regla de asociación:**

$$c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$$

- Ejemplo: $c(\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}) =$

$$\frac{s(\{\text{leche, paales, cerveza}\})}{s(\{\text{leche, paales}\})} = \frac{\sigma(\{\text{leche, paales, cerveza}\})}{\sigma(\{\text{leche, paales}\})}$$

Tenemos un pequeño problema...

- Con 5 items, obtuvimos $2^5 - 1$ posibles ítems
 - Es decir, 31

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- {Pan}
- {Coca Cola}
- {Leche}
- {Pañales}
- {Cerveza}
- {Pan, Coca Cola}
- {Pan, Leche}
- {Pan, Pañales}
- {Pan, Cerveza}
- {Coca Cola, Leche}
- {Coca Cola, Pañales}
- {Coca Cola, Cerveza}
- {Leche, Pañales}
- {Leche, Cerveza}
- {Pañales, Cerveza}
- {Pan, Coca Cola, Leche}
- {Pan, Coca Cola, Pañales}
- {Pan, Coca Cola, Cerveza}
- {Pan, Leche, Pañales}
- {Pan, Leche, Cerveza}
- {Pan, Pañales, Cerveza}
- {Coca Cola, Leche, Pañales}
- {Coca Cola, Leche, Cerveza}
- {Coca Cola, Pañales, Cerveza}
- {Pan, Coca Cola, Leche, Pañales}
- {Pan, Coca Cola, Leche, Cerveza}
- {Pan, Coca Cola, Pañales, Cerveza}
- {Pan, Leche, Pañales, Cerveza}

Tenemos un pequeño problema...

- Con 5 ítems, obtuvimos $2^5 - 1$ posibles ítems
 - Es decir, 31
- Con n ítems, tenemos $2^n - 1$ posibles ítems
- Una tienda suele tener varios ítems
 - Imaginemos $n = 100$

Tenemos un pequeño problema...

- Si $n = 100$
- $2^{100} - 1$ posibles ítems
- $2^{100} - 1 =$

1267650600228229401496703205376

Un computador actual puede hacer ~ 3 millones de operaciones por segundo (3 GHz)...

Entonces demoraríamos aproximadamente

64403322675823264816693248 segundos en sólo
encontrar los itemsets posibles

es decir **2040860051243962497** años

Solución: Algoritmo Apriori

Principio de Monotonidad:

Si un itemset es frecuente, entonces todos los subgrupos de éste también son frecuentes

- Si $\{pan, cerveza\}$ es frecuente, entonces $\{pan\}$ y $\{cerveza\}$ deben ser frecuentes.
- Si $\{pan, cerveza, pañales\}$, entonces...

Regla inversa (anti-monotonía):

Si un itemset no es frecuente, entonces todos sus supersets deben también ser infrecuentes

- Si $\{pan\}$ **no** es frecuente, entonces ningún conjunto que contenga panes será frecuente
- Si $\{pan, coca cola\}$, entonces...

Algoritmo Apriori

Buscando Itemsets frecuentes

1. Definimos el valor del umbral

Umbral = 0,6

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

Algoritmo Apriori

Buscando Itemsets frecuentes

2. Calculamos el soporte de los 1-itemsets (itemsets de tamaño 1)

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- $\sigma(\{\text{Pan}\}) =$
- $\sigma(\{\text{Coca Cola}\}) =$
- $\sigma(\{\text{Leche}\}) =$
- $\sigma(\{\text{Cerveza}\}) =$
- $\sigma(\{\text{Pañales}\}) =$

Algoritmo Apriori

Buscando Itemsets frecuentes

2. Calculamos el soporte de los 1-itemsets (itemsets de tamaño 1)

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- $\sigma(\{\text{Pan}\}) = 4$
- $\sigma(\{\text{Coca Cola}\}) = 3$
- $\sigma(\{\text{Leche}\}) = 4$
- $\sigma(\{\text{Cerveza}\}) = 2$
- $\sigma(\{\text{Pañales}\}) = 3$

Algoritmo Apriori

Buscando Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

3. Seleccionamos los datasets frecuentes

- $s(\{\text{Pan}\}) =$
- $s(\{\text{Coca Cola}\}) =$
- $s(\{\text{Leche}\}) =$
- $s(\{\text{Cerveza}\}) =$
- $s(\{\text{Pañales}\}) =$

$$C_1 = \{\{\text{Cerveza}\}, \{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$$

Ordenar alfabéticamente

$$L_1 = \{\{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$$

Algoritmo Apriori

Buscando Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

3. Seleccionamos los datasets frecuentes

- $s(\{\text{Pan}\}) = 0.8$
- $s(\{\text{Coca Cola}\}) = 0.6$
- $s(\{\text{Leche}\}) = 0.8$
- $s(\{\text{Cerveza}\}) = 0.4$
- $s(\{\text{Pañales}\}) = 0.6$

$$C_1 = \{\{\text{Cerveza}\}, \{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$$

Ordenar alfabéticamente

$$L_1 = \{\{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$$

Algoritmo Apriori

Buscando Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

K = 2

Construimos 2-itemsets (itemsets de tamaño 2) candidatos) a partir de los **1-itemsets** frecuentes, utilizando la operación join.

$$L_1 = \{ \{Coca Cola\}, \{Leche\}, \{Pan\}, \{Pañales\} \}$$

$$C_2 = L_1 \bowtie L_1$$

Operación Join (\bowtie)

Buscando Itemsets frecuentes

Conj 1 = { {A, B, C}, {E, F, G} }

Conj 2 = { {A, B, D}, {E, F, H}, {X, Y, Z} }

Conj 1 \bowtie Conj 2 = { {A, B, C, D}, {E, F, G, H} }

Algoritmo Apriori

Buscando Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

K = 2

Construimos 2-itemsets (itemsets de tamaño 2) candidatos) a partir de los **1-itemsets** frecuentes, utilizando la operación join.

$L_1 =$

$L_1 =$

$$C_2 = L_1 \bowtie L_1$$

=

Algoritmo Apriori

Buscando Itemsets frecuentes

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

K = 2

Construimos 2-itemsets (itemsets de tamaño 2) candidatos) a partir de los **1-itemsets** frecuentes, utilizando la operación join.

$L_1 = \{\{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$

$L_1 = \{\{\text{Coca Cola}\}, \{\text{Leche}\}, \{\text{Pan}\}, \{\text{Pañales}\}\}$

$$C_2 = L_1 \bowtie L_1$$

=

Algoritmo Apriori

Buscando Itemsets frecuentes

K = 2

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- $s(\{\text{Coca Cola, Leche}\}) =$
- $s(\{\text{Coca Cola, Pan}\}) =$
- $s(\{\text{Coca Cola, Pañales}\}) =$
- $s(\{\text{Leche, Pan}\}) =$
- $s(\{\text{Leche, Pañales}\}) =$
- $s(\{\text{Pan, Pañales}\}) =$

Algoritmo Apriori

Buscando Itemsets frecuentes

K = 2

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

- $s(\{\text{Coca Cola, Leche}\}) =$
- $s(\{\text{Coca Cola, Pan}\}) =$
- $s(\{\text{Coca Cola, Pañales}\}) =$
- $s(\{\text{Leche, Pan}\}) =$
- $s(\{\text{Leche, Pañales}\}) =$
- $s(\{\text{Pan, Pañales}\}) =$

Algoritmo Apriori

Buscando Itemsets frecuentes

K = 3

TID	Items
1	Pan, Coca Cola, Leche
2	Cerveza, Pan
3	Pan, Coca Cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca Cola, Pañales, Leche

En la pizarra

Algoritmo **Apriori**

Buscando reglas de asociación

Para cada itemset frecuente sacar sus subconjuntos y calcular la **confianza** entre ellas.

Recordando:

- **Regla de asociación**

- Una expresión de la forma $X \Rightarrow Y$, donde X e Y son itemsets
- Ejemplo: $\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}$

- **Soporte de una regla de asociación (s)**

- Soporte de la unión del antecedente y el consecuente
- $s(X \Rightarrow Y) = s(X \cup Y)$
- Ejemplo: $s(\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}) = s(\{\text{Leche, pañales, cerveza}\})$

- **Confianza de una regla de asociación:**

$$c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$$

- Ejemplo: $c(\{\text{Leche, pañales}\} \Rightarrow \{\text{cerveza}\}) =$

$$\frac{s(\{\text{leche, paales, cerveza}\})}{s(\{\text{leche, paales}\})} = \frac{\sigma(\{\text{leche, paales, cerveza}\})}{\sigma(\{\text{leche, paales}\})}$$

Formalizando más conceptos:

- **Lift**

- Mide qué tan correlacionados están X e Y

$$lift(X \Rightarrow Y) = \frac{c(X \Rightarrow Y)}{s(Y)} = \frac{s(X \cup Y)}{s(X)s(Y)}$$

$$Lift = \frac{c(X \rightarrow Y)}{s(Y)} = \begin{cases} < 1 \\ = 1 \\ > 1 \end{cases}$$

Negativamente correlacionadas
Independientes
Positivamente correlacionadas

Ejercicio

TID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3