Reglas de Asociación

Karim Pichara
Associate Professor
Computer Science Department
Pontificia Universidad Católica de Chile

Association Rules

- Association Rules are patterns like:

  A -> B

  Where A and B are sets of instantiated binary variables, ej:

  $\{V_1 = 1, V_3=1\}$ -> $V_5 = 1$

  {butter, cheese} -> bread

Association Rules

- Where A and B are sets of instantiated binary variables, ej:

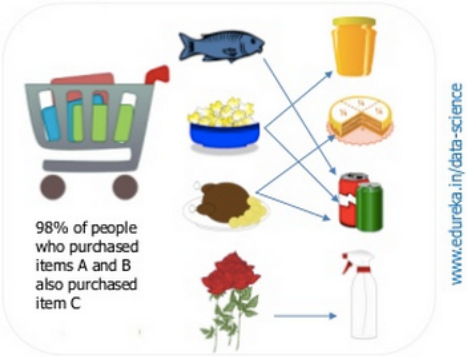
  $\{V_1 = 1, V_3=1\}$ -> $V_5 = 1$
  {butter, cheese} -> bread

We learn those patterns from a database of sets of instantiated binary variables

Market Basket Analysis

•Application of Association Rules to transactions in a store

•Will make us to understand better what association rules are



Market Basket Analysis

- Objetivos
  - Analizar los hábitos de compra de los clientes buscando asociaciones entre los diferentes items que los clientes agregan en sus "carros de compra".


Aplicaciones

- Market basket analysis
  - –Uso de info en Data Warehouse de las tiendas.
  - –Beneficios?
    - Ordenamiento de productos
    - Patrones de navegación en tienda
    - Sugerir ventas cruzadas, ej. Hamburguesas y ketchup
    - Promociones de productos cruzados
    - …

Algunas definiciones

- **Itemset**
  - Una colección de uno o más items Ejemplo: {Leche, pan, cerveza}
  - k-itemset
    - Un itemset que contiene k items
- **Contador del soporte (σ)**
  - Frecuencia de ocurrencia de un itemset
  - Ej. σ({Leche, pan, cerveza}) = 2
- **Soporte**
  - Fracción de las transacciones que contiene un itemset
  - Ej. s({Leche, pan, cerveza}) = 2/5
- **Itemset frecuente**
  - Un itemset cuyo soporte es mayor o igual a un determinado umbral

Algunas definiciones(Cont..)

- **Regla de asociación**
  - Una expresión de la forma
    X → Y, donde X e Y son itemsets
  - Ejemplo:
    {Leche, pañales} → {cerveza}
- **Métricas de evaluación de las reglas de asociación**
  - Soporte(s)
    Fracción de las transacciones que contiene a X e Y
  - Confianza(c)
    Fracción de veces que items en Y aparecen en transacciones que contienen X

Algunas definiciones(Cont..)

- **Ejemplo**

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

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

40% de las transacciones mostraron que leche, pañales cerveza se compraron juntos

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

67% de los consumidores que Compraron leche y pañales, También compraron cerveza

Reglas Significativas

- Dado un set de transacciones T, el objetivo es encontrar reglas de asociación que cumplan:
  - Soporte ≥ *min_sup*
  - Confianza ≥ *min_conf*
- Alto soporte = combinación es frecuente
  - Baja probabilidad que sea algo aleatorio
- Alta confianza = patrón significativo
  - Atributos están estrechamente relacionados

Ejemplo de Reglas.

T	Items
1	Pan, Leche
2	Pan, pañales, cerveza, huevos
3	Leche, pañales, cerveza, diario
4	Pan, leche, pañales, cerveza
5	Pan, leche, pañales, diario

Ejemplo 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)

- Todas las reglas son particiones binarias del mismo itemset: {leche, pañales, cerveza}
- Reglas originadas en el mismo itemset tienen idéntico soporte pero pueden tener un distinto nivel de confianza

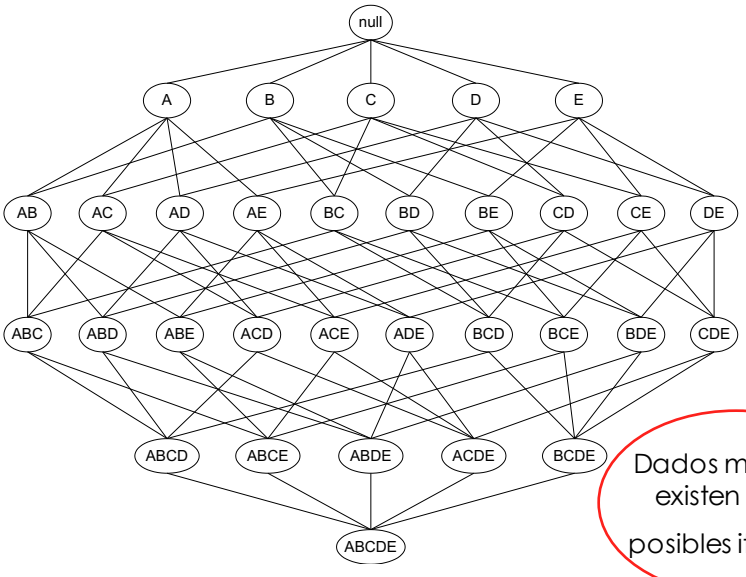
Encontrando Reglas de Asociación

- Objetivo: encontrar todas las reglas de asociación tales que:
  - soporte $\geq s$
  - confianza $\geq c$

Encontrando Reglas de Asociación

- Itemsets frecuentes
  - Encontrar todos los itemsets frecuentes X
  - Dado $X=\{A_1, \ldots, A_k\}$, generar todas las reglas $S \rightarrow X - S$ para todos los subconjuntos no vacíos S de X
  - Confianza = soporte(X)/soporte(S)
  - Soporte = soporte(X)
  - Excluir reglas cuya confianza es muy baja
- Encontrar los itemsets frecuentes es la parte complicada

Itemset Lattice



Dados m items, existen 2^m-1 posibles itemsets

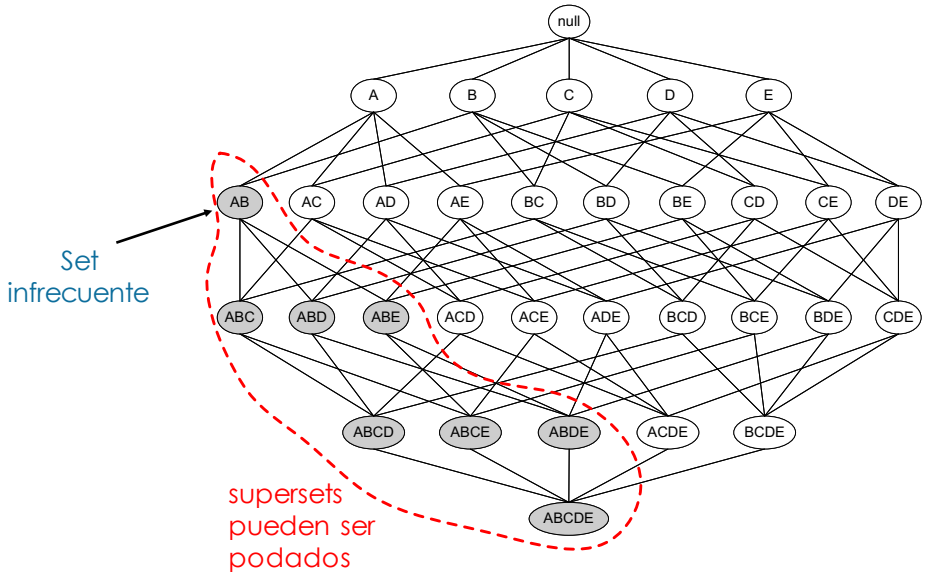
Idea: Podar Itemsets

- Principio de Monotonicidad:
  - Si un itemset es frecuente, entonces todos los subgrupos de éste también son frecuentes
- Monotonicidad y Soporte:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

- Regla inversa (antimonotonía):
  - Si un itemset no es frecuente, entonces todos sus supersets deben también ser infrecuentes

Usando Monotonicidad



Algoritmo A-Priori(Agrawal 94')

- En cada paso generar k-itemsets a partir de los (k-1)-itemsets frecuentes
- Utilizar propiedad de monotonicidad para eliminar los k-itemsets que no pueden ser frecuentes
- Evaluar la frecuencia de los k-itemsets posibles
- Generar Lk con los k-itemsets frecuentes según umbral *min_sup_count*.

A priori ejemplo:

Data

```
I1,I2,I5
I2,I4
I2,I3
I1,I2,I4
I1,I3
I2,I3
I1,I3
I1,I2,I3,I5
I1,I2,I3
```

Min sup = 2/9

A priori: para ejercitar

Data

```
A1,A3,A5
A2,A3
A1,A3
A1,A3,A4
A1,A3,A5
A2,A3,A4
A1,A2,A4
A2,A3,A5
A1,A2,A3
A3,A4,A5
A2,A3,A4,A5
A3,A4,A5
```

Min Sup: 3/12

Karim Pichara B.

Ejemplo 2

TABLE 14.1. *Inputs for the demographic data.*

Feature	Demographic	# values	Type
1	sex	2	categorical
2	marital status	5	categorical
3	age	7	ordinal
4	education	6	ordinal
5	occupation	9	categorical
6	income	9	ordinal
7	years in Bay Area	5	ordinal
8	dual incomes	3	categorical
9	number in household	9	ordinal
10	number of children	9	ordinal
11	householder status	3	categorical
12	type of home	5	categorical
13	ethnic classification	8	categorical
14	language in home	3	categorical

14 preguntas en 9409 questionarios a consumidores de un shopping mall en San Fco., Ca, USA

Karim Pichara B.

Ejemplo 2 (Cont..)

TABLE 14.1. *Inputs for the demographic data.*

Feature	Demographic	# values	Type
1	sex	2	categorical
2	marital status	5	categorical
3	age	7	ordinal
4	education	6	ordinal
5	occupation	9	categorical
6	income	9	ordinal
7	years in Bay Area	5	ordinal
8	dual incomes	3	categorical
9	number in household	9	ordinal
10	number of children	9	ordinal
11	householder status	3	categorical
12	type of home	5	categorical
13	ethnic classification	8	categorical
14	language in home	3	categorical

⇛

Transformación a 50 variables binarias

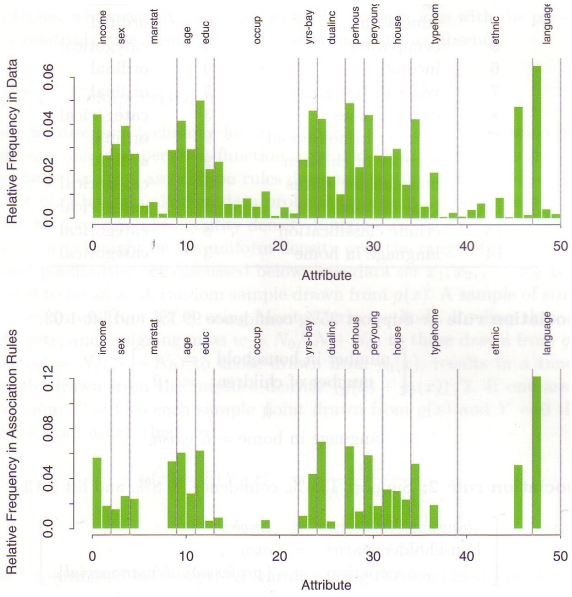
- Variables categóricas reemplazadas por k variables
- Variable ordinales binarizadas usando mediana

Base de Datos Resultante

- Matriz de 6876x50
- 6876 observaciones. Registros con datos faltante son eliminados
- 50 variables (atributos) binarias

Karim Pichara B.

Ejemplo 2 (Cont..)



Karim Pichara B.

PUC Chile

Ejemplo 2 (Cont..)

Association rule 1: Support 25%, confidence 99.7% and lift 1.03.

$$\begin{bmatrix} \text{number in household} & = & 1 \\ \text{number of children} & = & 0 \end{bmatrix}$$
$$\Downarrow$$
$$\text{language in home} = \textit{English}$$

Association rule 2: Support 13.4%, confidence 80.8%, and lift 2.13.

$$\begin{bmatrix} \text{language in home} & = & \textit{English} \\ \text{householder status} & = & \textit{own} \\ \text{occupation} & = & \{\textit{professional/managerial}\} \end{bmatrix}$$
$$\Downarrow$$
$$\text{income} \geq \$40{,}000$$

Association rule 3: Support 26.5%, confidence 82.8% and lift 2.15.

$$\begin{bmatrix} \text{language in home} & = & \textit{English} \\ \text{income} & < & \$40{,}000 \\ \text{marital status} & = & \textit{not married} \\ \text{number of children} & = & 0 \end{bmatrix}$$
$$\Downarrow$$
$$\text{education} \notin \{\textit{college graduate}, \textit{graduate study}\}$$