Reglas de Asociación

Reglas de Asociación

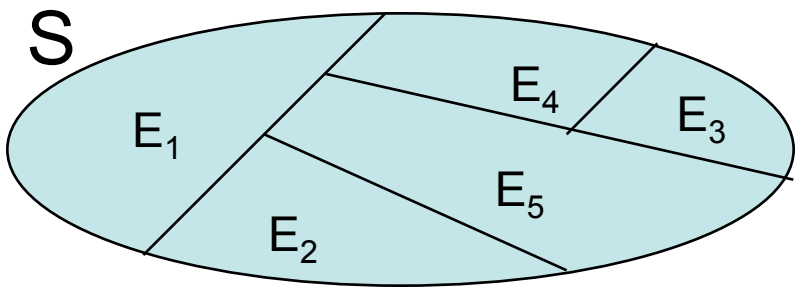
- Técnica no supervisada que busca hallar en una base de datos, conjuntos de atributos (variables binarias), cuya ocurrencia conjunta se observa con una alta probabilidad.

Objetivo

- Encontrar en una base de datos de “transacciones” (tabla de contingencia binaria) reglas de asociación entre items (variables).

Mini-repaso de Probabilidades

- Sea E un evento de un espacio muestral S



- Axioma 1: $0 \leq P(E) \leq 1$
- Axioma 2: $P(S) = 1$
- Axioma 3:

si E_1 y E_2 son mutuamente excluyentes

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Probabilidad condicional

- La probabilidad condicional de Y dado X

$$P(Y / X) = P(Y \cap X) / P(X)$$

es la probabilidad de que ocurra Y dado que "sabemos" que ya ocurrió X. Por lo tanto

$$P(Y \cap X) = P(Y / X) * P(X)$$

Independencia de eventos

- El evento X es independiente del evento Y si el conocimiento de la ocurrencia de X no modifica la probabilidad de ocurrencia de Y, mas formalmente

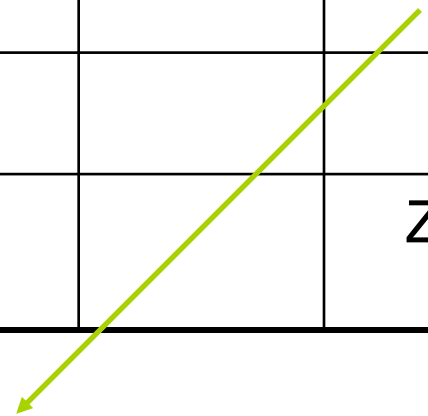
$$P(Y / X) = P(Y)$$

Por lo tanto, bajo independencia

$$P(Y \cap X) = P(Y) * P(X)$$

Los Datos

	Item 1	…	Item j	…	Item k
Trans1	$Z_{1,1}$		$Z_{1,j}$		$Z_{1,k}$
…					
Trans i	$Z_{i,1}$		$Z_{i,j}$		$Z_{i,k}$
…					
Trans N	$Z_{N,j}$		$Z_{N,j}$		$Z_{N,k}$



$Z_{i,j} = 1$ si el item j esta en la transaccion i

$Z_{i,j} = 0$ si el item j NO esta en la transaccion i

El problema

Cantidad de items

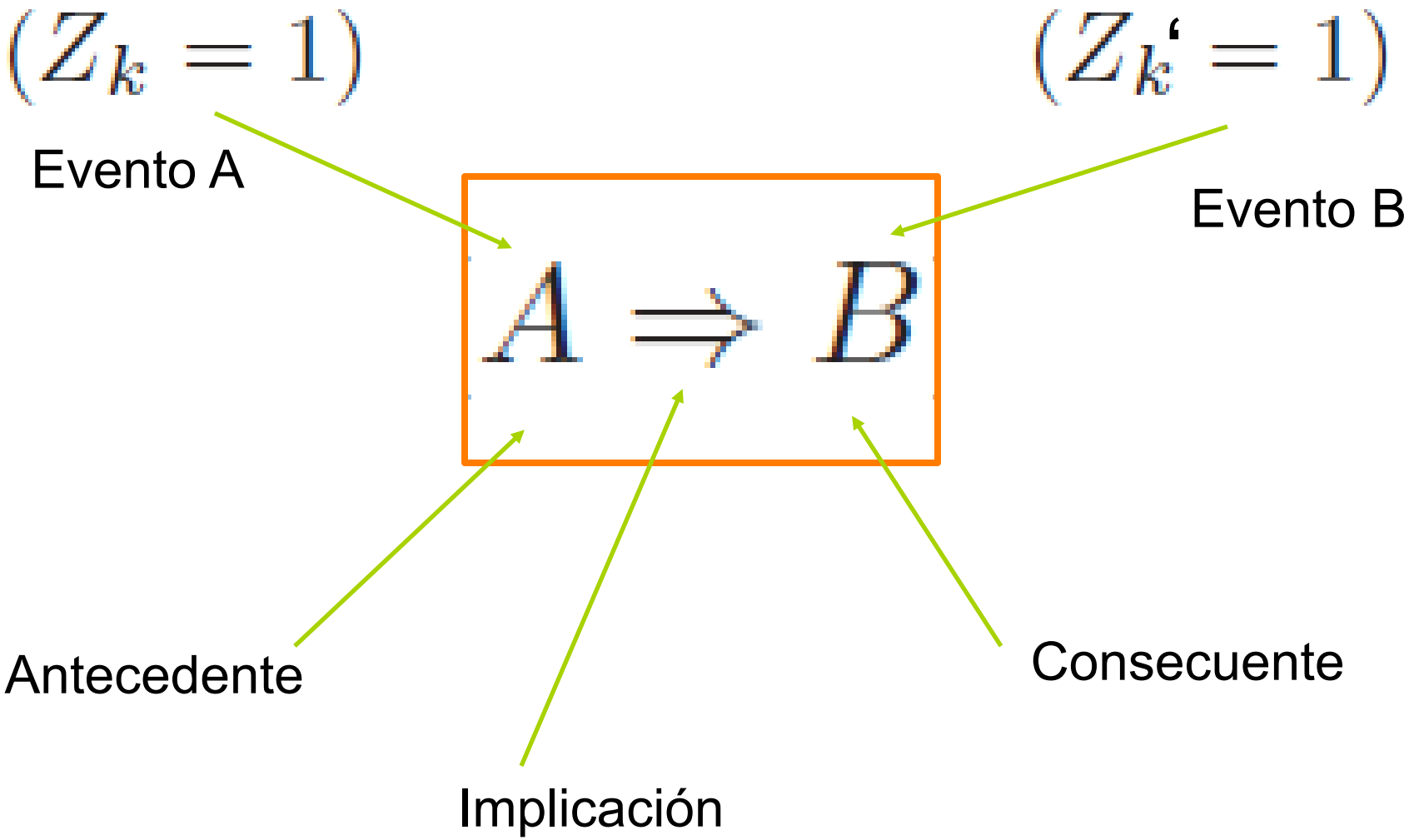
Buscar un subconjunto $\mathcal{K} \subset \{1, \ldots, K\}$

tal que $\Pr\left[\bigcap_{k \in \mathcal{K}} (Z_k = 1)\right] = \Pr\left[\prod_{k \in \mathcal{K}} Z_k = 1\right]$

sea alta.

Indice de item

La Regla de Asociación



Lift, Confianza y Soporte

$$T(A \Rightarrow B) = P(A \cap B)$$

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)} = \frac{T(A \Rightarrow B)}{T(A)\,T(B)}$$

Ejemplo 1: Lo no querido

- P(A) = 10/100
- P(B) = 90/100
- P(A,B) = 9/100

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$ = 9/10

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}$$ = 1

Ejemplo 2: Lo trivial

- P(A) = 90/100
- P(B) = 90/100
- P(A,B) = 81/100

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$ = 9/10

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}$$ = 1

Ejemplo 3: Lo buscado

- P(A) = 10/100
- P(B) = 10/100
- P(A,B) = 10/100

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$ = 1

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}$$ = 10

Estimacion del soporte

$$T(\mathcal{K}) = \widehat{\Pr}\left[\prod_{k\in\mathcal{K}} Z_k = 1\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}} z_{ik}$$

Vale 1, solo en las transacciones en las que estan los items de $\mathcal{K}$

Busqueda de un Soporte Mínimo

- La restricción mas importante a ser impuesta a las reglas de asociación halladas, es la de contar con un Soporte Mínimo, es decir

$$T(\mathcal{K}) > u$$

El Algoritmo “a priori”

Se basa en el siguiente principio:

Si $P(Z_{i1} = 1,\ldots, Z_{ia} = 1) < u$ entonces

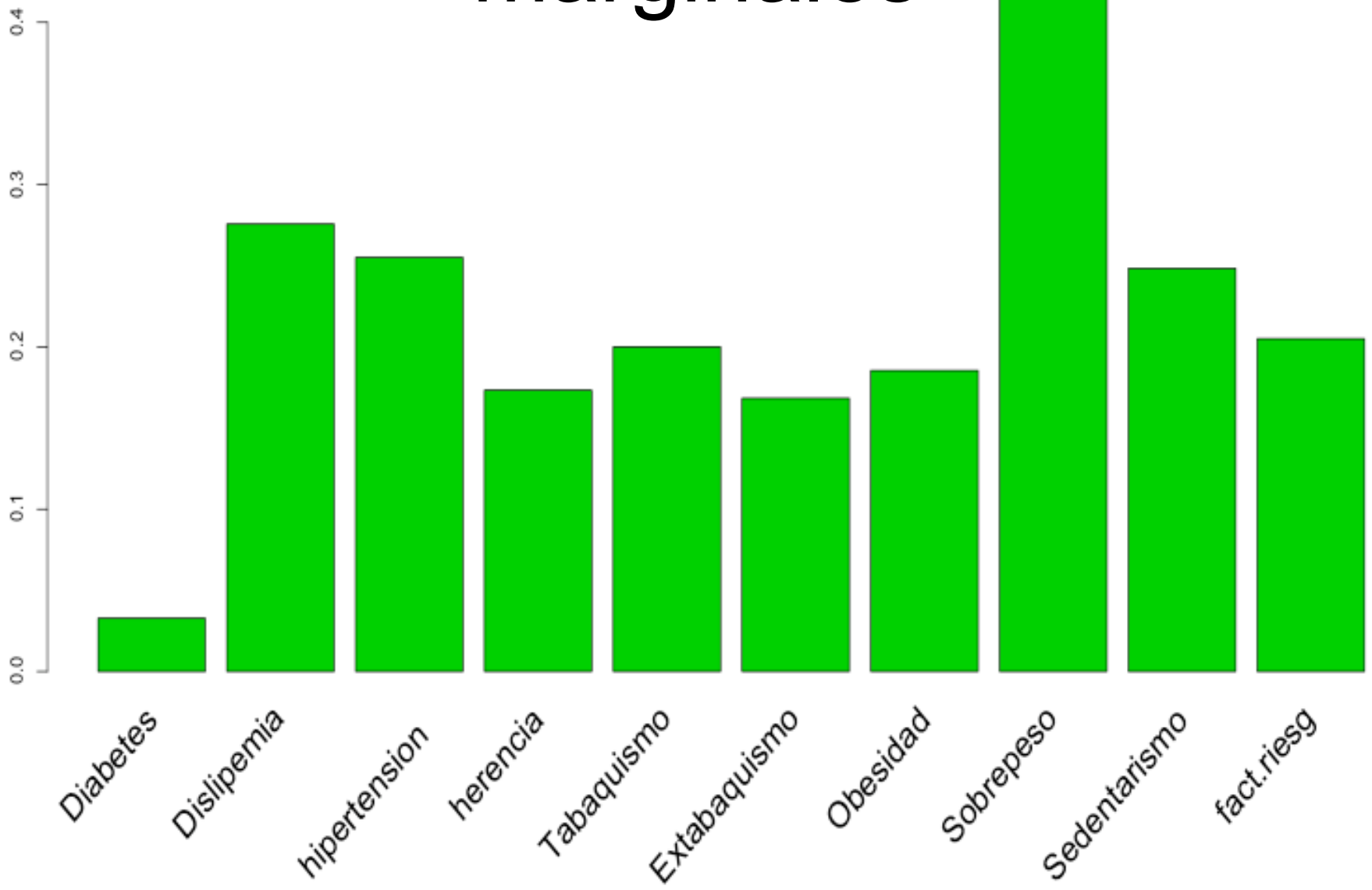
$P(Z_{i1} = 1,\ldots, Z_{ia} = 1, Z_{j1} = 1,\ldots, Z_{jb} = 1) < u$

1) Eliminar items con $p < u$
2) Eliminar pares de items (de los que dejo el paso 1) con $p<u$
3) Eliminar tripletes de items (de los que dejaron los pasos 1 y 2) con $p<u$
4) ….

Ejemplo 1: Medicina en el deporte

<>	A	C	D	E	F	G	H	I	J	K	L	M
1	NOMBRE	SEXO	Diabetes	Dislipemia	Hipertensión	Herencia	Tabaquismo	Extabaquismo	Obesidad	Sobrepeso	Sedentarismo	fact.riesg
2	PERGAÑE	1	0	0	1	0	0	0	0	0	0	0
3	DE PINA	0	0	0	0	0	0	0	0	0	0	1
4	TROUBOU	0	0	0	1	0	0	0	0	0	0	0
5	LIBSON T	1	0	1	0	0	0	0	0	1	0	0
6	BUSTOS	1	0	0	1	0	0	1	1	0	0	0
7	LASALVIA	1	0	1	0	1	0	1	0	0	1	0
8	BERGES	0	0	0	0	0	0	0	0	0	0	1
9	GARCÉS	1	0	1	0	0	0	0	0	0	0	0
10	ODDI INE	0	0	0	0	0	0	0	0	0	0	1
4031	LABOURD	1	0	0	0	0	1	0	0	1	1	0
4032	MORALES	1	0	0	0	0	1	0	0	1	1	0
4033	ARIAS OS	1	0	1	0	1	0	1	0	1	1	0
4034	HERRERA	0	0	0	0	0	0	0	0	0	0	1
4035	GALANEN	0	1	0	0	1	0	0	0	1	0	0
4036												

Frecuencias relativas marginales



Resumen de Reglas halladas

```
parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen
       0.25    0.1    1 none FALSE            TRUE    0.02      1      5
 target   ext
  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10 item(s), 4034 transaction(s)] done [0.00s].
sorting and recoding items ... [10 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [72 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Reglas halladas

```
> inspect(SORT(rules,by="confidence"))
    lhs                     rhs                    support confidence      lift
1   {Dislipemia,
     Extabaquismo}    => {Sobrepeso}       0.02999504  0.5062762 1.1825814
2   {Dislipemia}     => {Sobrepeso}       0.13758057  0.4991007 1.1658207
3   {Dislipemia,
     Sedentarismo}    => {Sobrepeso}       0.03594447  0.4982818 1.1639078
4   {Dislipemia,
     herencia}        => {Sobrepeso}       0.03197819  0.4961538 1.1589372
5   {Dislipemia,

hipertension} => {Sobrepeso}      0.04908280  0.4950000 1.1562420
6   {Dislipemia,
     Obesidad}        => {
hipertension} 0.02900347  0.4915966 1.9272117
7   {
hipertension} => {Sobrepeso}      0.12022806  0.4713314 1.1009559
8   {
hipertension,
     herencia}        => {Dislipemia}      0.02379772  0.4705882 1.7071519
9   {
hipertension,
     Tabaquismo}      => {Sobrepeso}       0.02057511  0.4585635 1.0711322
```

```
10 {
hipertension,
    Extabaquismo}    => {Sobrepeso}      0.02578086  0.4561404 1.0654720
11 {Extabaquismo}    => {Sobrepeso}      0.07635102  0.4536082 1.0595574
12 {Tabaquismo,
    Sedentarismo}    => {Sobrepeso}      0.03123451  0.4516129 1.0548966
13 {Dislipemia,
    Obesidad}        => {Sedentarismo}   0.02578086  0.4369748 1.7592378
14 {Sedentarismo}    => {Sobrepeso}      0.10832920  0.4361277 1.0187257
15 {herencia,
    Sobrepeso}       => {Dislipemia}     0.03197819  0.4314381 1.5651272
16 {Tabaquismo}      => {Sobrepeso}      0.08626673  0.4312268 1.0072778
17 {Dislipemia,
    Tabaquismo}      => {Sobrepeso}      0.02379772  0.4304933 1.0055645
18 {}                => {Sobrepeso}      0.42811106  0.4281111 1.0000000
19 {
hipertension,
    Sedentarismo}    => {Obesidad}       0.02949926  0.4280576 2.3085350
20 {herencia}        => {Sobrepeso}      0.07411998  0.4277539 0.9991658
21 {
hipertension,
    herencia}        => {Sobrepeso}      0.02131879  0.4215686 0.9847179
```

```
22 {
hipertension,
    Extabaquismo}    => {Dislipemia}     0.02379772  0.4210526 1.5274517
23 {
hipertension,
    Sedentarismo}    => {Sobrepeso}      0.02900347  0.4208633 0.9830704
24 {Obesidad}        => {Sedentarismo}   0.07610312  0.4104278 1.6523611
25 {
hipertension,
    Sobrepeso}       => {Dislipemia}     0.04908280  0.4082474 1.4809983
26 {
hipertension,
    Obesidad}        => {Sedentarismo}   0.02949926  0.4033898 1.6240265
27 {Dislipemia,
    Extabaquismo}    => {
hipertension} 0.02379772  0.4016736 1.5746856
28 {
hipertension,
    Obesidad}        => {Dislipemia}     0.02900347  0.3966102 1.4387819
29 {Obesidad}        => {
hipertension} 0.07312841  0.3943850 1.5461120
30 {Extabaquismo,
    Sobrepeso}       => {Dislipemia}     0.02999504  0.3928571 1.4251670
```