Reglas de asociación

ORLANDO TORRES ESTRADA

SELENE JAZMÍN RODRÍGUEZ GONZÁLEZ

Algoritmos de asociación

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, *items* o atributos que tienden a ocurrir de forma conjunta. Por ejemplo:

- La cesta de la compra en un supermercado.
- Los libros que compra un cliente en una librería.
- Las páginas web visitadas por un usuario.
- Las características que aparecen de forma conjunta.

- A cada uno de los eventos o elementos que forman parte de una transacción se le conoce como item y a un conjunto de ellos itemset.
- Una transacción puede estar formada por uno o varios items, en el caso de ser varios, cada posible subconjunto de ellos es un itemset distinto.
- Por ejemplo, la transacción T = {A,B,C} está formada por 3 items (A, B y C) y sus posibles itemsets son: {A,B,C}, {A,B}, {B,C}, {A,C}, {A}, {B} y {C}.

Base de datos transaccional

- Podemos representar una BDD Transaccional con las siguientes métricas de interés:
- -UNA LISTA
- -UNA REPRESENTACION VERTICAL
- -UNA REPRESENTACION HORIZONTAL

Lista

- Básicamente representa cada transacción como una fila
- Cada fila lista los artículos comprados por el consumidor
- Cada fila es una transacción por lo que cada fila puede tener un numero diferente de columnas.

	A	B	C	D
1	Tomate	lechuga	Mostaza	Jamon
2	Tomate	Pepino	Queso	
3	Agua	Periodico		
4	Agua	Coca		

Representación Vertical

- Es la forma mas eficiente de guardar lo datos de tamaño mas industrial o comercial, este ocupa solo 2 columnas
- Indica numero o ID de la transacción
- Indica el articulo

T_ID	Articulos
1	Tomate
2	Papa
3	Agua
4	Agua
5	Lechuga
2	Pepino
6	Periodico
7	Coca

Representación Horizontal

- Se representa como una matriz binara, cada fila de una matriz representa una transacción, y cada columna representa un articulo
- Si un articulo esta presente se representa como 1
- Si un articulo esta ausente se representa como 0

	Tomate	Lechuga	Mostaza	Jamon
1	1	1	1	1
2	1	0	0	0
3	1	1	0	0
4	0	1	0	0

Métricas de interés

MÉTODOS EFICIENTE PARA ENCONTRAR REGLAS

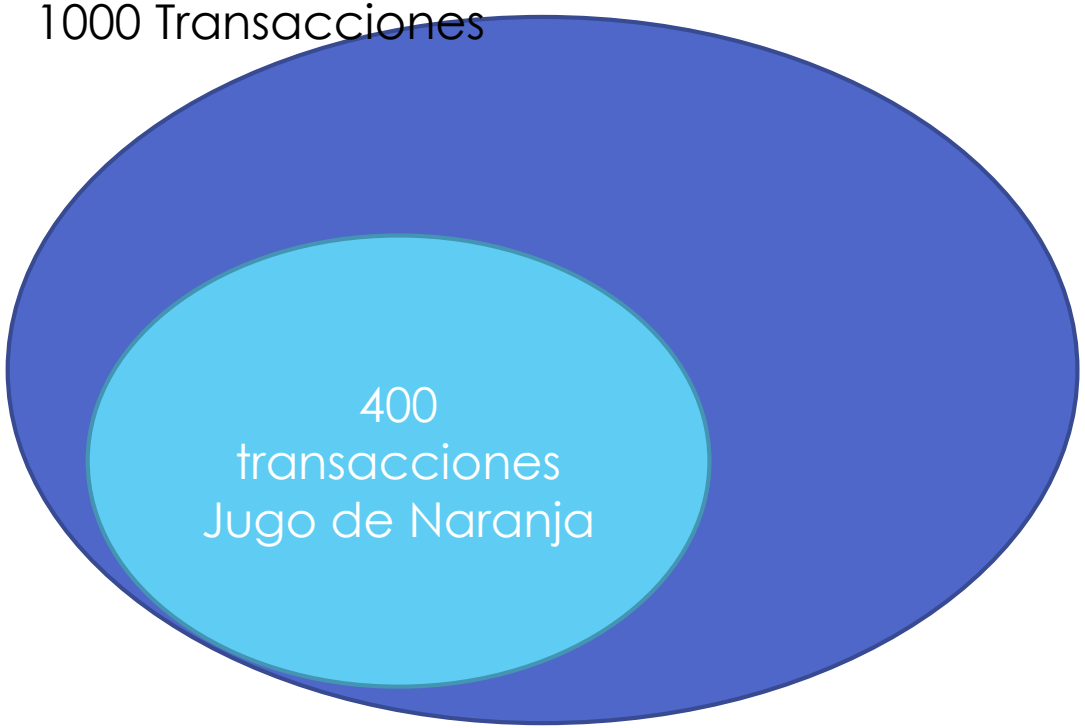
¿CÓMO HACEMOS PARA LIMITAR O REDUCIR EL NUMERO DE REGLAS?

Soporte(Frecuencia relativa)

- Dada regla si A=>B, el soporte de esta se define como el numero de veces o la frecuencia relativa con que A y B aparecen juntos en una BDD Transaccional
- Puede definirse para los artículos individualmente

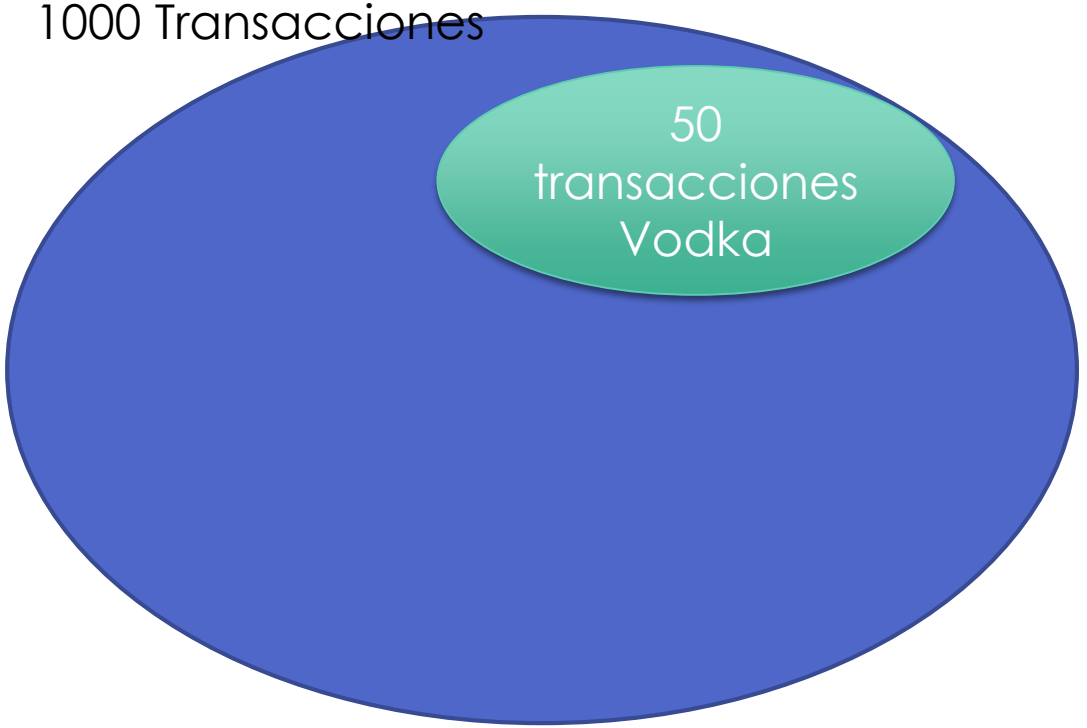
Ejemplo

1000 Transacciones



Soporte → (J/N) = 400/1000
40% → Probabilidad

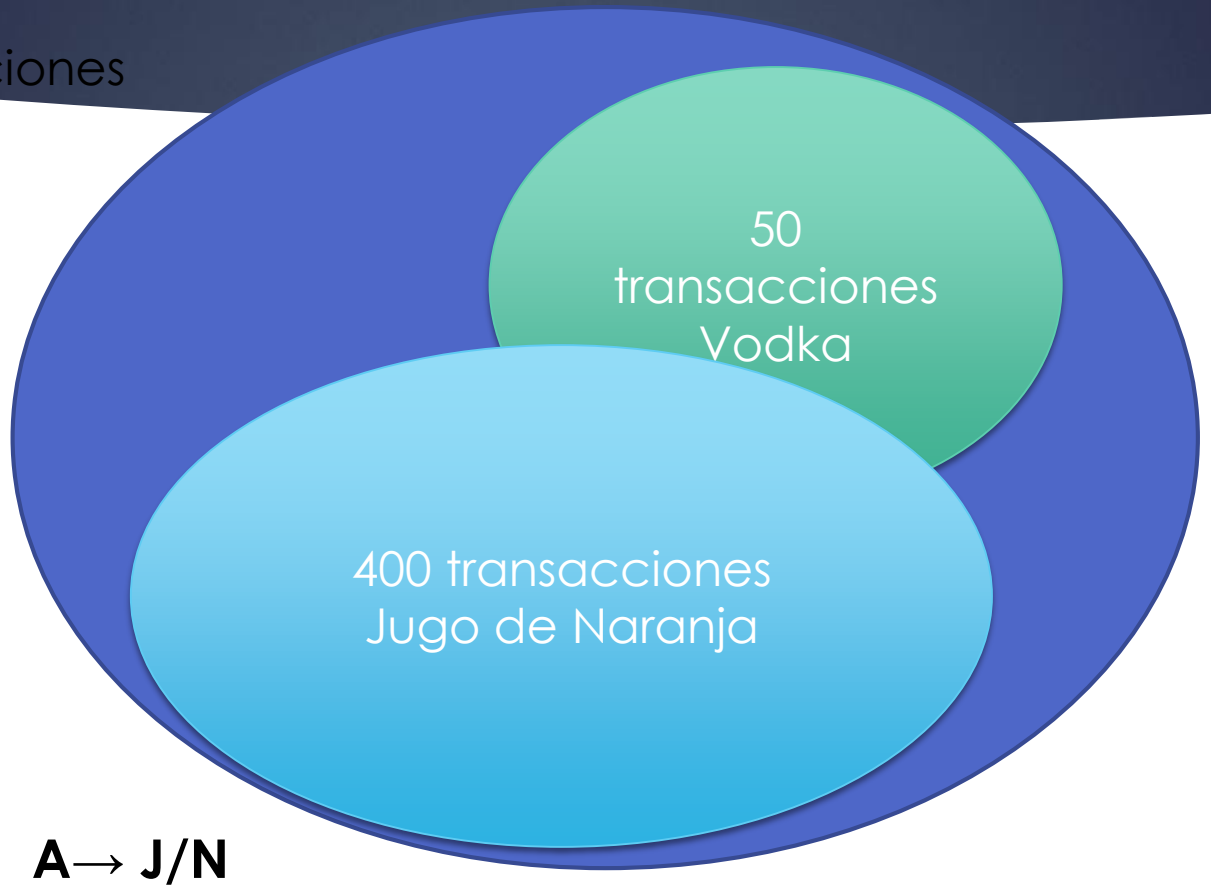
1000 Transacciones



Soporte → (Vodka) = 50/1000
5% → Probabilidad

Ejemplo

1000 Transacciones



A→ J/N
B→ Vodka
Soporte (J/N→Vodka) = 40/1000 = 4% → Probabilidad

Confianza(Probabilidad empírica)

- Dada una regla si A=>B, la confianza de esta regla es el cociente de soporte de la regla y soporte del antecedente solamente.
- **Confianza (A=>B)** = Soporte (A=>B)/Soporte (A)
- Si el soporte mide la frecuencia, Confianza mide la fortaleza de la regla.
- En lenguaje de probabilidad, confianza es una probabilidad condicional.
- Confianza (A=>B) = P(B/A)

Ejemplo

- ¿Cuál es la confianza del Vodka → J/N?

- **Confianza (Vodka → J/N)**
- Soporte (Vodka → J/N)/Soporte(Vodka)
- 40/50 = 0.8
- **80% de que el J/N este junto al Vodka en una transacción**

Ejemplo

- La inversa
- ¿Cuál es la confianza del J/N → Vodka?

- **Confianza (J/N→ Vodka)**
- Soporte (J/N → Vodka)/Soporte(J/N)
- 40/400 = 0.1
- **10% de que el Vodka este junto al J/N en una transacción**

Lift(Refleja probabilidad)

	PAN	PAN*	TOTAL
J/N	280	120	400
J/N*	420	180	600
TOTAL	700	300	1000

Proceso

- **Soporte(Pan)** = 700/1000 = 0.7
- **Soporte(J/N)** = 400/1000 = 0.4
- **Soporte(Pan→J/N)** = Soporte(J/N→Pan) = 280/1000 = 0.28
- **Confianza(Pan→J/N)** = 0.28/0.7 = 0.4
- **Confianza(J/N→Pan)** = 0.28/0.4 = 0.7

Definicion

- **Lift (A→B)** = Soporte(A→B)/(Soporte (A) * Soporte(B))
- Si Lift = 1 o muy cerca a 1, indica que la relación es producto del azar
- De lo contrario, indica que la relación es realmente fuerte(controlado por la frecuencia con que ambos ocurre)

- Lift <= 1, relación débil
- Lift > 1, relación fuerte

Ejemplo

- Lift (Pan→J/N) = Lift(J/N→Pan)
- Soporte (Pan→J/N)/(Soporte (Pan) * Soporte (J/N))
- **0.28/(0.7*0.4) = 1**
- **Relación producto del azar, dato que no nos proporciono el soporte, ni la confianza**

- **Soporte(Pan)** = 0.7
- **Soporte(J/N)** = 0.4
- **Soporte(Pan→J/N)** = 0.28
- **Confianza(Pan→J/N)** = 0.4
- **Confianza(J/N→Pan)** = 0.7

- **Lift (A→B)** = Soporte(A→B)/(Soporte (A) * Soporte(B))

Ejemplo

- Lift (Vodka→J/N) = Lift(J/N→Vodka)
- Soporte (Vodka →J/N)/(Soporte (Vodka) * Soporte (J/N))
- **0.04/(0.4*0.05) = 2**
- **Relación es producto al azar, aunque sea 2, esta muy cerca del 1, pero tiene mayor probabilidad en aparecer en una transacción**

- **Soporte(Vodka)** = 0.05
- **Soporte(J/N)** = 0.4
- **Soporte(Vodka→J/N)** = 0.04
- **Confianza(Vodka→J/N)** = 0.8
- **Confianza(J/N→Vodka)** = 0.1

- **Lift (A→B)** = Soporte(A→B)/(Soporte (A) * Soporte(B))

Apriori

Apriori fue uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación y sigue siendo uno de los más empleados, tiene dos etapas:

- Identificar todos los *itemsets* que ocurren con una frecuencia por encima de un determinado límite (*itemsets* frecuentes).
- Convertir esos *itemsets* frecuentes en reglas de asociación.

Ejemplo

- Una panadería llamada “El dorado” que vende productos como pan, leche, huevos, galletas, jugo, entro otros. Se quieren construir reglas de asociación en base a las siguientes transacciones:

Transacción ID	Conjunto-Items
1	Pan, huevos, leche
2	Pan, leche
3	Pan, galletas
4	Huevos, churro, jugo

Transacción ID	Conjunto-Items
1	Pan, huevos, leche
2	Pan, leche
3	Pan, galletas
4	Huevos, churro, jugo

2. Se eliminan los elementos infrecuentes, por lo tanto se eliminan del proceso

Conjunto-Items	Soporte
Pan	3
Huevos	2
Leche	2

1. Se crea una tabla que contenga conjuntos de ítems de un solo elemento y las veces que se repiten en las transacciones

Conjunto-Items	Soporte
Pan	3
Huevos	2
Leche	2
Galletas	1
Churro	1
Jugo	1

Transacción ID	Conjunto-Items
1	Pan, huevos, leche
2	Pan, leche
3	Pan, galletas
4	Huevos, churro, jugo

3. Tomando solamente en cuenta los elementos anteriores, se crea una nueva tabla en donde se registra cuantas veces se repiten los pares de transacciones.

Conjunto-Items	Soporte
{Pan, huevos}	1
{Huevos, leche}	1
{Pan, leche}	2

4. Se eliminan nuevamente los elementos infrecuentes, dejando solamente el siguiente conjunto:

Conjunto-Items	Soporte
{Pan, leche}	2

Transacción ID	Conjunto-Items
1	Pan, huevos, leche
2	Pan, leche
3	Pan, galletas
4	Huevos, churro, jugo

5. En base a la tabla resultante en el punto 4, se obtienen las siguientes reglas de asociación:

Regla de asociación	Soporte	Confianza
Pan -> Leche	2	2/3=0.66
Leche -> Pan	2	2/2=1