Reglas de Asociación

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Curso 2021/22

Ejemplo del problema

Supongamos los siguientes registros de transacciones en un supermercado:

ID	Leche	Pan	Galleta	Cereales	Mermelada	Pasta	Carne	Cacao	Café	Té	Pollo	Azúcar
1	1	1	1	0	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	1	0	1	0	0
4	1	1	0	0	1	1	0	0	0	0	0	0
5	0	0	1	0	0	1	0	0	0	1	0	0
6	0	1	0	0	0	0	0	1	0	1	0	0
7	0	0	0	1	0	1	0	0	0	1	0	0
8	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	1	1	0	0	0	0	1	0	1	0
10	0	0	1	1	0	0	0	0	1	0	1	0
11	0	0	0	0	0	0	0	1	1	0	0	1
12	0	1	0	0	0	0	0	0	1	0	1	0
13	0	1	1	0	0	0	0	0	0	0	0	1
14	0	0	0	1	0	0	0	0	1	0	0	1
15	0	1	0	0	0	0	0	0	1	0	0	1
16	0	1	0	0	0	0	0	0	1	0	0	1
17	1	0	0	1	0	0	0	0	1	1	0	0

Pretendemos conocer cuándo la compra de una combinación de artículos suele implicar la compra de otro/s artículos para de esta forma ofrecerlos al cliente o realizar promociones.

Ejemplo: Leche+Café suele conllevar la compra de Azúcar

Definición formal del problema

Sea:

- $I = \{i_1, i_2, ..., i_n\}$, el conjunto de ítems
- $T = \{t_1, t_2, ..., t_m\}$, el conjunto de transacciones,

donde cada registro de transacción es identificado mediante un identificador único y contiene un subconjunto de ítems de I. Buscamos una regla del tipo:

$$X \implies Y$$

, donde:

$$X, Y \subseteq I$$
$$X \cap Y = \emptyset$$

X se suele denominar 'Antecedente' e Y 'Consecuente'.

Medidas de confianza

- Soporte de un conjunto de items X en una base de datos T se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de ítems:

$$sop_T(X) = \frac{|X|}{|T|}$$

- La confianza de una regla X se define como:

$$conf_T(X \implies Y) = \frac{sop_T(X \cup Y)}{sop_T(X)} = \frac{|X \cup Y|}{|X|}$$

Lectura interesante

Los pañales y la cerveza:

https://web.archive.org/web/20100411063658/http://www.daedalus.es/mineria-de-datos/los-panales-y-la-cerveza/

Indicador 'lift'

El indicador *lift* expresa cuál es la proporción del soporte observado de un conjunto de productos respecto del soporte teórico y bajo el supuesto de independencia.

$$lift = \frac{sop_T(X \cup Y)}{sop_T(X) sop_T(Y)}$$

- lift <= 1, el conjunto aparece una cantidad de veces inferior o igual a lo esperado bajo condiciones de independencia (no parece existir una relación entre los productos por encima de lo normal)
- lift > 1, el conjunto aparece una cantidad de veces inferior a lo esperado bajo condiciones de independencia (puede existir una relación entre los productos por encima de lo normal)

Ejercicio 1

Dada la regla X: $\{Galleta, Cereales\} \Longrightarrow \{Café\}$, calcular el soporte del Antecedente y del Consecuente en la base de datos de Ejemplo, así como su confianza y su Lift. Interpreta dichos resultados.

Algoritmo APriori

Algorithm: APriori(T, ϵ)

```
1 k = 1
2 C_1 = {i ∈ I : sop_T(i) > ε}
3 while C_k ≠ ∅ do
4     foreach c ∈ C_k do
5         foreach i ∈ I do
6             if sop_T(c ∪ i) > ε then C_{k+1} = C_{k+1} ∪ {c ∩ i}
7     k=k+1
8 return ∪_k C_k
```

Ejercicio 3

Aplicar el algoritmo APriori con ε=3 sobre las siguientes transacciones:

ID	A	B	C	D
1	1	1	1	1
2	1	1	0	1
3	1	1	0	0
4	0	1	1	1
5	0	1	1	0
6	0	0	1	1
7	0	1	0	1

Ejercicio 3

Obtenemos las transacciones por encima del umbral ε=3 :

C1		#
	A	3
	B	6
	C	4
	D	5
C2		
	A, B	3
	A, C	1
	A, D	2
	B, C	3
	B, D	4
	C, D	3
C3		
	A, B, C	1
	A, B, D	2
	B, C, D	2

Solución: { (A, B), (B, C), (B, D), (C, D) }

Ejercicio 3

Obtenemos los indicadores de las reglas finales:

	Soporte		Confianza		Lift	
A->B	3/7	42,86 %	3/3	100,00 %	(6/7)/((3/7)*(4/7))	3,50
B->A	3/7	42,86 %	3/6	50,00 %	(6/7)/((3/7)*(4/7))	3,50
B->C	3/7	42,86 %	3/3	100,00 %	(7/7)/((3/7)*(4/7))	4,08
C->B	3/7	42,86 %	3/4	75,00 %	(7/7)/((3/7)*(4/7))	4,08
B->D	4/7	57,14 %	4/6	66,67 %	(7/7)/((3/7)*(4/7))	4,08
D->B	4/7	57,14 %	4/5	80,00 %	(7/7)/((3/7)*(4/7))	4,08
C->D	3/7	42,86 %	3/4	75,00 %	(7/7)/((3/7)*(4/7))	4,08
D->C	3/7	42,86 %	3/5	60,00 %	(3/7)/((3/7)*(4/7))	1,75

Otros algoritmos

Otros algoritmos que podemos citar son:

- **ECLAT**: la principal diferencia de este algoritmo con Apriori radica en el formato de la base de datos de entrada. En Eclat en cada registro figura las transacciones en las que aparece un ítem determinado.
- Partition
- FP-Growth