Probabilidad y Análisis de Datos

Daniel Fraiman

Maestría en Ciencia de Datos, Universidad de San Andrés

Aplicaciones en espacios grandes discretos

Supongamos que estamos interesados en entender:

"qué y cómo compra una persona cuando va al supermercado"

- En el supermercado hay N items
- Una compra es: $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, con $X_i = \{0, 1\}$ (no compró, compró).

Aplicaciones en espacios discretos grandes discretos

Supongamos que estamos interesados en entender:

"qué y cómo compra una persona cuando va al supermercado"

Concretamente

Objetivo: Identificar productos que tiendan a comprarse de forma conjunta.

Con el fin de:

- situarlos en posiciones cercanas dentro de la tienda y maximizar la probabilidad de que los clientes compren.
- presentar nuevos combos de productos al consumidor de manera de aumentar las ventas.
- Si es una compra web, sugerir otros productos.

Aplicaciones en espacios discretos grandes discretos

Reglas de Asociación e items frecuentes

El problema matemático es el siguiente:

Supongamos que sabemos que la persona tiene en su carrito de compras los productos A y B, ¿hay altas chances de que compre T si lo encuentra antes de llegar a la caja?

Llamemos S al conjunto que tiene las compras actuales.

$$S = \{A, B\}, \text{¿} S \to T \text{?}$$

¿Le recomendamos T?

Reglas de Asociación e items frecuentes

Problema matemático:

Dado S, ¿cuál es el producto T que maximiza la probalidad de compra? Llamemos Θ al conjunto que tiene a todos los productos del supermercado ($\Theta = \{A, B, \ldots, W\}$).

$$T = \underset{Y \in \Theta}{argmax}\, \mathbb{P}\left(Y|S\right)$$

- En realidad estaremos interesados en ordenar los productos según su probabilidad. Quizás ofrecemos los 5 primeros.

Quiero el producto que maximiza la probabilidad de que lo compre las personas que van a buscar otro producto distinto, y terminan agregando esta compra.

Reglas de Asociación e items frecuentes

Objetivo:

Estimar $\mathbb{P}(Y|S)$ para todo $Y \in \Theta$.

Recordemos:

$\mathbb{P}(Y|S) = \frac{\mathbb{P}(Y \cap S)}{\mathbb{P}(S)}$ con $\mathbb{P}(S) > 0$.

Estimación:

Basado en la historia de compras (tickets):

- $\mathbb{P}(S) = \frac{\#\{\text{tickets con } S\}}{\#\{\text{tickets}\}}$
- $\mathbb{P}(Y \cap S) = \frac{\#\{\text{tickets con } Y \text{ y } S\}}{\#\{\text{tickets}\}}$

Cual es la probabilidad de llevar leche en polvo y pañales
los tickets, deben tener si o si ambos productos y pueden tener mas tambien, pero no pueden faltar la leche en polvo y pañales

Reglas de Asociación e items frecuentes

tengo 2 alternativas, comprar o no comprar

Dificultad en la estimación:

El conjunto Θ es muy grande (N), por lo tanto el Espacio de Probabilidad de compras será gigante (2^N), no tendremos una muestra (historia de compras) suficientemente grande y entonces las estimaciones tendrán mucho error (o varianza).

- Pero independiente de lo anterior, no nos interesan realmente los productos Y con $\mathbb{P}(Y|S) << 1$.

producto que ya tengo en el changuito

la probabilidad de comprar Y, cuando tenga S, donde S es lo que tengo en el carrito e Y es lo que quiero hacer comprar

Nuevo planteo

Vamos a pedir:

- $\mathbb{P}(S \cap T) \geq s$, con s algún valor prefijado.
- $\mathbb{P}(T|S) \geq c$, con c algún valor prefijado.

producto que quiero que lleven

Reglas de Asociación e items frecuentes

Nuevo planteo

Vamos a pedir:

- $\mathbb{P}(S \cap T) \geq s$, con s algún valor prefijado.
- $\mathbb{P}(T|S) \geq c$, con c algún valor prefijado.

- *Soporte*$(S \to T)$=*Soporte*$(T \to S)$:=$\mathbb{P}(S \cap T) \geq s$, con s algún valor prefijado.
- *Confianza*$(S \to T)$:=$\mathbb{P}(T|S) \geq c$, con c algún valor prefijado.

vos podes tener coca y te ofrezco fernet, vale lo mismo que vos ya tengas el fernet y te ofrezca coca cola

La probabilidad de intercepcion de los productos, se lo llama SOPORTE

La confianza es cuanta “fe”, le tenes que cuando lleven el producto A, lleven tmb el producto B

Reglas de Asociación e items frecuentes

- *Soporte*$(S \to T)$:=$\mathbb{P}(S \cap T) \geq s$, con s algún valor prefijado.
- *Confianza*$(S \to T)$:=$\mathbb{P}(T|S) \geq c$, con c algún valor prefijado.

- Soporte = “cuántas ventas $S \cap T$ espero tener”
- Confianza = qué confianza (chances) tengo en que compren el producto T recomendado cuando tienen S.

Reglas de Asociación e items frecuentes

En la práctica:

Ponemos un límite al tamaño de S y T. Por ejemplo $|S| < 3$ y $|T| = 1$, Si compró salchichas y pan ($|S| = 2$), ¿qué le ofrezco? ¿Y si compró ojotas y protector solar?

- Fijamos s (Soporte:=$\mathbb{P}(S \cap T) \geq s$)
- Fijamos c (Confianza:=$\mathbb{P}(T|S) \geq c$)
- Y damos todas las recomendaciones (T) para todos los conjuntos S compatibles con tener un soporte y una confianza mayor a s y c.

ALGORITMO A PRIORI: $\mathbb{P}(S \cap T) \geq s$

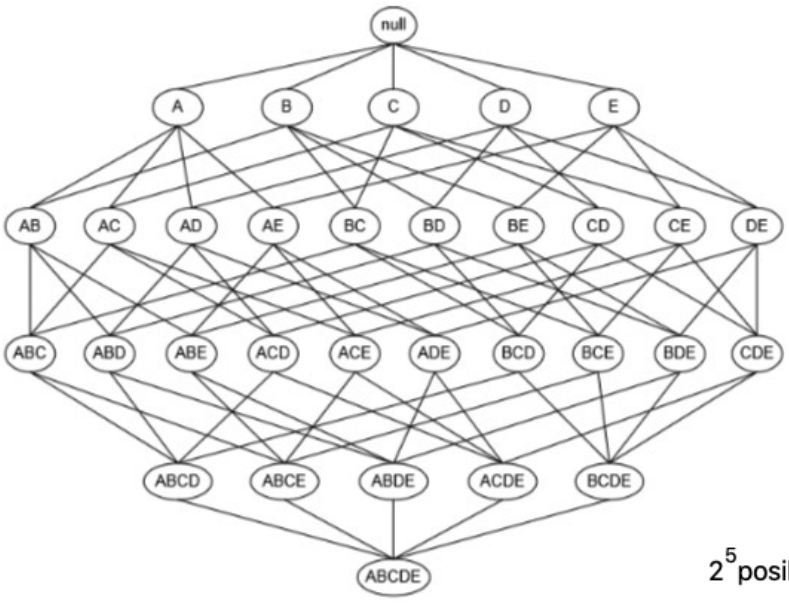
Algoritmo a Priori: $\mathbb{P}(S \cap T) =: \mathbb{P}(Z) \geq s$

Objetivo: Encontrar todos los conjuntos Z (con $|Z| < k$) que cumplen $\mathbb{P}(Z) \geq s$.

Algoritmo A Priori

1. Generar una lista con todos los conjuntos Z de tamaño 1.
2. $k = 1$.
3. Podar (prune) de la lista los candidatos Z de tamaño k que $\mathbb{P}(Z) < s$.
4. Generar todos los conjuntos Z' de tamaño $k+1$ que tienen como subconjunto a los elementos de la lista.
5. $k = k + 1$ y go to 3.

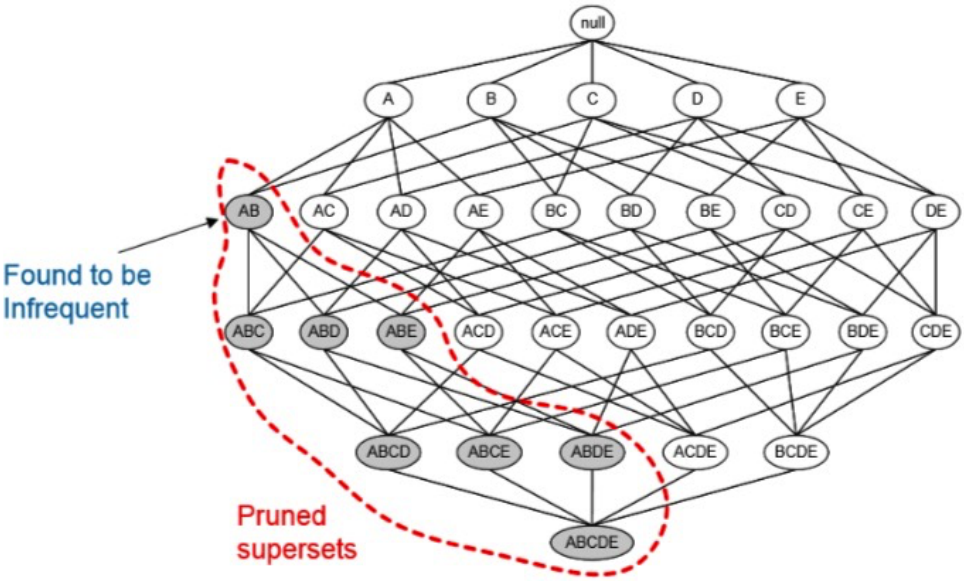
Algoritmo a Priori: $\mathbb{P}(Z) \geq s$



5 items

2^5 posibles compras

Algoritmo a Priori: $\mathbb{P}(Z) \geq s$



Algoritmo a Priori: $\mathbb{P}(Z) \geq s$

Ej :En cuantos tickets de todos los que tengo compraron Pan —> en 4 tickets

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Soporte s=3 ventas

que grupos podria armar con los productos que sobrevivieron, y luego, vuelo a los que tienen menos de 3 ventas

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk,Diaper}	3

Me voy a quedar con los productos que se vendieron mas de 3 veces, esa es la funcion del soporte

Observación:

Sea Z un conjunto de k items ($|Z| = k$) que tiene soporte c.

- Cualquier subconjunto, Z', no vacío de Z tiene soporte $\geq c$.

Proof

Supongamos que Z corresponde a los productos de las primeras k coordenadas del vector de compras.
$s = \mathbb{P}(Z) = \mathbb{P}((X_1, X_2, \ldots, X_k) = (1, 1, \ldots, 1))$
Z' es un subconjunto de Z, por ejemplo los dos primeros productos.

$$\begin{aligned}
\mathbb{P}(Z') =& \mathbb{P}((X_1, X_2) = (1, 1)) \\
=& \sum_{j_1, j_2, \ldots j_{k-2} \in \{0,1\}} \mathbb{P}((X_1, X_2, X_3, \ldots, X_k) = (1, 1, j_1, j_2, \ldots, j_{k-2})) \\
=& \mathbb{P}((X_1, X_2, \ldots, X_k) = (1, 1, \ldots, 1)) + \text{probabilidades} \\
=& s + \text{probabilidades} \geq s
\end{aligned}$$

REGLAS DE ASOCIACIÓN

Reglas de Asociación

Una vez que tenemos nuestro listado con los distintos Z con soporte $\geq s$.

1. Particionamos cada Z. $S \cup T = Z$ con $S \cap T = \emptyset$.
2. Calculamos la confianza de la regla $S \to T$

(si compraste S te recomiendo T)

Confianza —>Probabilidad de que compres T, dado que llevaste S

Reglas de Asociación

Supongamos que el conjunto de items $\{A, B, C\}$ es uno de los que tiene ~~confianza~~ SOPORTE $\geq s$

1. Particionamos $\{A, B, C\}$: $S \to T$
   - $\{A\} \to \{B, C\}$, $\{B\} \to \{A, C\}$, $\{C\} \to \{A, B\}$
     $\{A, B\} \to \{C\}$, $\{A, C\} \to \{B\}$, $\{B, C\} \to \{A\}$
2. Calculamos la confianza de cada una de estas reglas de asociación $S \to T$.
3. Presentamos las que tienen confianza mayor a c.

Reglas de Asociación

Confianza

$Confianza(S \to T)$=$\mathbb{P}(T|S) = \frac{\mathbb{P}(T \cap S)}{\mathbb{P}(S)} = \frac{Soporte(S \cup T)}{Soporte(S)}$. (raro, ¿no?)

La confianza de la "Regla S-T" es igual

Reglas de Asociación

Eventos

T =compra los artículos C y D. $\mathbf{X} = (X_1, X_2, 1, 1, X_5, \ldots, X_N)$
S =compra el artículos A y B. $\mathbf{X} = (1, 1, X_3, X_4, X_5, \ldots, X_N)$
$T \cap S$ = compra los A, B, C, y D. $\mathbf{X} = (1, 1, 1, 1, X_5, \ldots, X_N)$

$$confianza(S \to T) = \mathbb{P}(T|S) = \frac{\mathbb{P}(T \cap S)}{\mathbb{P}(S)}$$

Conjuntos de artículos

$T = \{C, D\}$.
$S = \{A, B\}$.
$T \cup S = \{A, B, C, D\}$.

$$Confianza(S \to T) = \frac{Soporte(S \cup T)}{Soporte(S)}$$

En programacion, para decirte que quieren las 2 cosas, utilizan UNION, en vez de intercepcion para decir que es la suman de las 2 cosas

Reglas de Asociación

Confianza

{Bread, Milk} ⇨ {Diaper} $\quad \frac{\text{Soporte}(\{\text{Bread, Milk,Diaper}\})}{\text{Soporte}(\{\text{Bread,Milk}\})} =$

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Soporte s=3 ventas

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk,Diaper}	3

Reglas de Asociación

Confianza

{Bread, Milk} ⇨ {Diaper} $\quad \frac{\text{Soporte}(\{\text{Bread, Milk,Diaper}\})}{\text{Soporte}(\{\text{Bread,Milk}\})} = \frac{3}{3} = 100\%$

Siempre que se compraron pan y leche, se compraron pañales, con un 100% de confianza

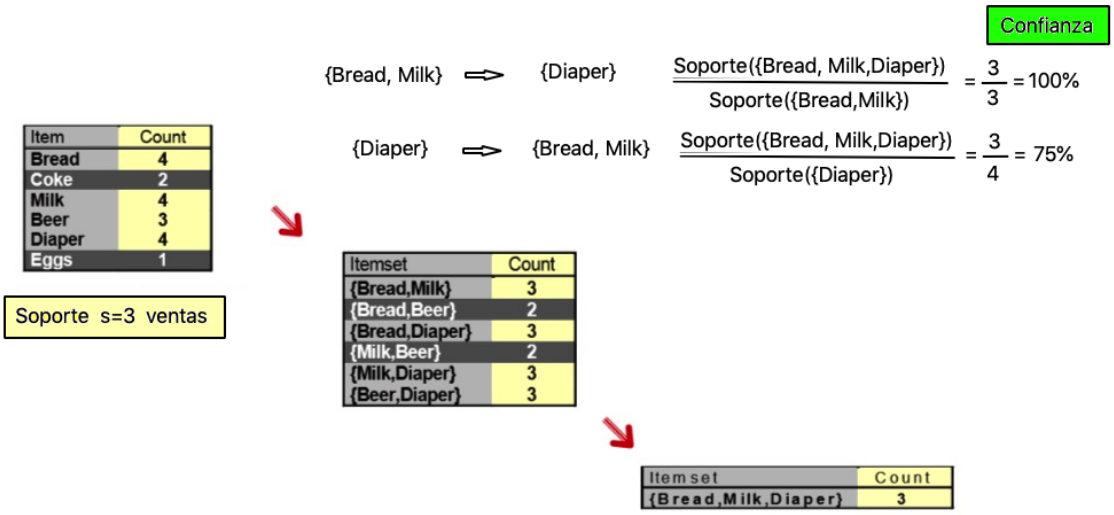
Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Soporte s=3 ventas

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk,Diaper}	3

Reglas de Asociación



Reglas de Asociación con número de items fijos

Una vez que tenemos nuestro listado con los distintos Z con soporte $\geq s$.

1. Particionamos cada Z. $S \cup T = Z$ con $S \cap T = \emptyset$.
2. Calculamos la confianza de la regla $S \to T$

(si compraste S te recomiendo T)

Al fijar en número de items en la regla de asociación en $|Z|$. Donde $S \cup T = Z$ podemos podar el arbol de reglas de decisión.

Reglas de Asociación con número de items fijos

Al fijar en número de items en la regla de asociación en $|Z|$. Donde $S \cup T = Z$ podemos podar el arbol de reglas de decisión.

Supongamos que $Z = \{A, B, C, D\}$. Las posibles reglas son:

- $\{A\} \leftrightarrows \{B, C, D\}$,$\{B\} \leftrightarrows \{A, C, D\}$, $\{C\} \leftrightarrows \{A, B, D\}$, $\{D\} \leftrightarrows \{A, B, C\}$
- $\{A, B\} \leftrightarrows \{C, D\}$,$\{A, C\} \leftrightarrows \{B, D\}$, $\{A, D\} \leftrightarrows \{B, C\}$

Propiedad:

Sea $S_1 \cup T_1 = Z$ y $S_2 \cup T_2 = Z$ con $dim(S_1) > dim(S_2)$. Se cumple

$$Confianza(S_1 \to T_1) \geq Confianza(S_2 \to T_2)$$

Reglas de Asociación con número de items fijos

Propiedad:

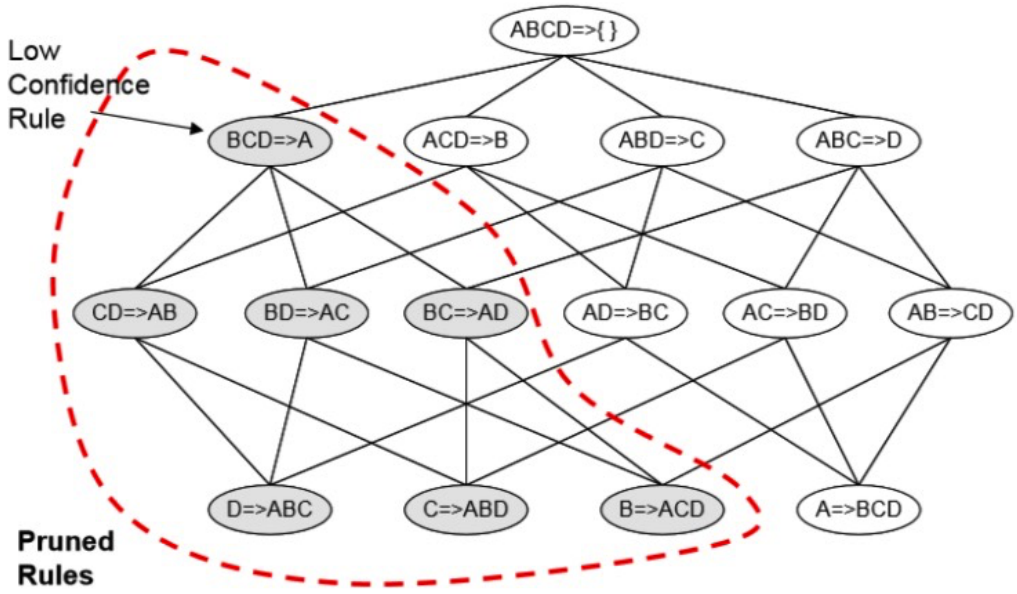
Sea $S_1 \cup T_1 = Z$ y $S_2 \cup T_2 = Z$ con $dim(S_1) > dim(S_2)$. Se cumple

$$Confianza(S_1 \to T_1) \geq Confianza(S_2 \to T_2)$$

Demostración.

$Confianza(S_k \to T_k) = \frac{Soporte(Z)}{Soporte(S_k)}$ y como
$Soporte(S_k) \geq Soporte(S_{k-1})$ □

Pruning de Reglas de Asociación



Alternativas a la Confianza: Lift, Leverage

Confianza vs Lift

Confianza:

$$Confianza(S \to T) = \mathbb{P}(T|S)$$

Problema: ¿ Qué pasa si T se compra casi siempre?

- Entonces $\mathbb{P}(T|S)$ probablemente sea alto.
- Peor aún si S y T son indep ($\mathbb{P}(T|S) = \mathbb{P}(T)$) S no predice nada.

independiente

Lift:

$$Lift(S \to T) = \frac{\mathbb{P}(T|S)}{\mathbb{P}(T)} = \frac{Confianza(S \to T)}{\mathbb{P}(T)}$$

- S y T son indep $\leftrightarrow Lift(S \to T) = 1$.
- Recomendamos $S \to T$ cuando Lift > 1.

Confianza vs Lift

Comparación

- Confianza: medida que dice las chances de que se compre T cuando compraste S.
- Lift: medida que compara el grado de dependencia entre S y T. Mide cuán buena es la regla respecto al azar.

Otras medidas

Tambien se debe mirar el soporte, la cantidad de ventas de S, el producto que lleva originalmente mi cliente, y que luego yo le voy a sugerir el producto T

Coverage:

$$Coverage(S \to T) = \mathbb{P}(S) = Support(S) = \frac{Soporte(S \to T)}{Confianza(S \to T)} = \frac{\mathbb{P}(S \cap T)}{\mathbb{P}(T|S)}$$

Leverage:

$$Leverage(S \to T) = \mathbb{P}(S \cap T) - \mathbb{P}(S)\,\mathbb{P}(T)$$

$$Leverage(S \to T) = Support(S \cup T) - Support(S)Support(T)$$

Es lo mismo que lift, pero en vez de ser mayor a 1, menor a 1, o 0, puede ser positivo, negativo o 0.

Reglas de Asociación con 

arules—> “reglas de asociacion”

Paquete *arules*

```
> transacciones = read.transactions(file = “datos_groceries.csv”,
format = “single”, sep = ”,”,header = TRUE, cols = c(“id_compra”,
ïtem”), rm.duplicates = TRUE)
> soporte=0.1; confianza=0.7
> reglas=apriori(data = transacciones, parameter = list(Support =
soporte,confidence = confianza)) #, minlen = 3, maxlen = 5
> inspect(reglas)
```

Resumen

- Confianza: medida que dice las chances de que se compre T cuando compraste S.
- Lift: medida que compara el grado de dependencia entre S y T. Mide que tan buena es la regla respecto al azar.
- Soporte de la regla: indica el impacto en término de ventas totales.