

Aprendizaje automático

Clase 3

Maximiliano Beckel

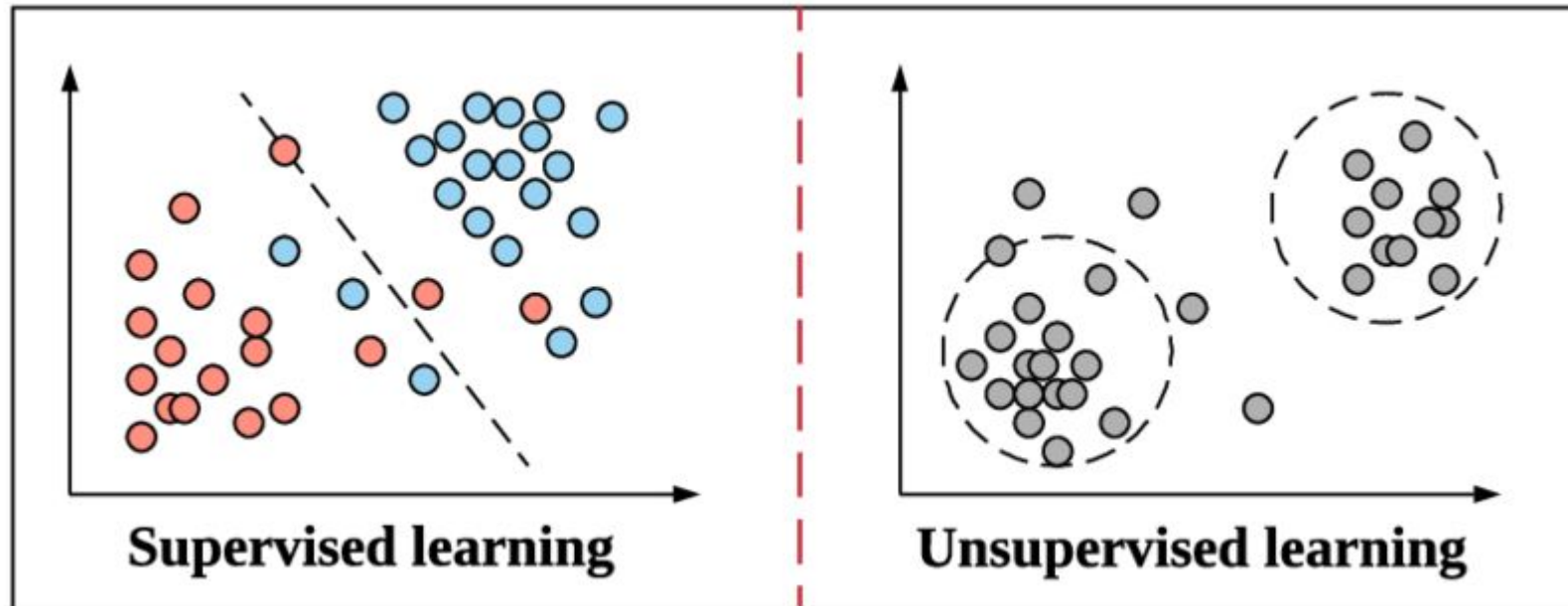


Aprendizaje no supervisado

El aprendizaje no supervisado se relaciona con el análisis de datos que no tienen ninguna etiqueta o variable a predecir. Su principal uso es el de realizar un análisis exploratorio de los datos, buscando aprender patrones en los mismos que permitan establecer relaciones y agrupamientos.

Algunos campos de aplicación:

- Creación de sistemas de recomendación
- Reconocimiento de patrones
- Agrupamiento de datos
- Procesamiento de imágenes (por ej. en medicina)
- Detección de anomalías.



Aprendizaje no supervisado

El aprendizaje no supervisado se relaciona con el análisis de datos que no tienen ninguna etiqueta o variable a predecir. Su principal uso es el de realizar un análisis exploratorio de los datos, buscando aprender patrones en los mismos que permitan establecer relaciones y agrupamientos.

Algunos campos de aplicación:

- Creación de sistemas de recomendación
- Reconocimiento de patrones
- Agrupamiento de datos
- Procesamiento de imágenes (por ej. en medicina)
- Detección de anomalías.

vs Aprendizaje Supervisado

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i$$

Y : Dependent variable

β_0 : Intercept

β_i : Slope for X_i

X = Independent variable


Aprendizaje no supervisado

El aprendizaje no supervisado se relaciona con el análisis de datos que no tienen ninguna etiqueta o variable a predecir. Su principal uso es el de realizar un análisis exploratorio de los datos, buscando aprender patrones en los mismos que permitan establecer relaciones y agrupamientos.

Algunos campos de aplicación:

- Creación de sistemas de recomendación
- Reconocimiento de patrones
- Agrupamiento de datos
- Procesamiento de imágenes (por ej. en medicina)
- Detección de anomalías.

vs Aprendizaje Supervisado


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i$$

Y : Dependent variable

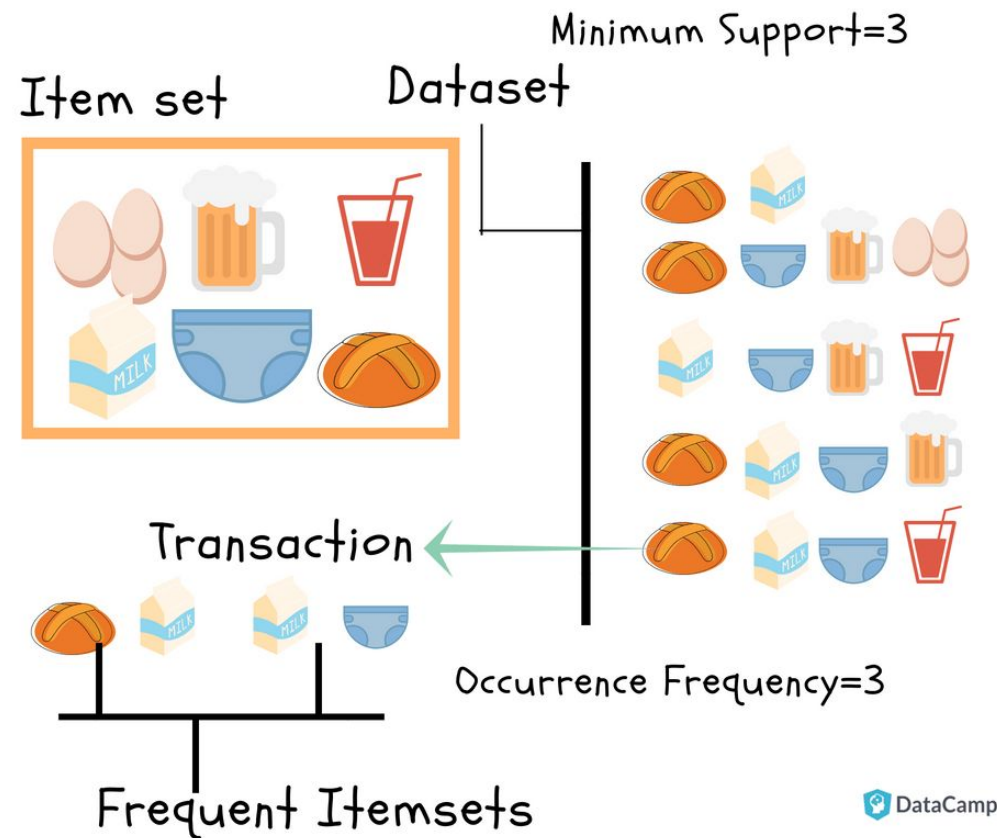
β_0 : Intercept

β_i : Slope for X_i

X = Independent variable

Reglas de asociación

Las Reglas de asociación son un método de análisis descriptivo que permite encontrar asociaciones entre conjuntos de ítems en grandes conjuntos de datos. Sus principales usos se relacionan con estudios de mercado en los que se busque determinar que tipo de productos se consumen de forma conjunta.




Reglas de asociación

Introducción a las reglas de asociación

Las reglas de asociación generalmente se escriben de la forma:

$$\{Antecedente\} \Rightarrow \{Consecuente\}$$

Esto indica que existe una relación entre los clientes que compran el antecedente y el consecuente en la misma transacción. La fuerza y el sentido de la relación se mide con diferentes indicadores: el soporte, la confianza y la mejora de la confianza.

Transaction 1	
Transaction 2	
Transaction 3	
Transaction 4	
Transaction 5	
Transaction 6	

Reglas de asociación

Introducción a las reglas de asociación

Las reglas de asociación generalmente se escriben de la forma:

$$\{Antecedente\} \Rightarrow \{Consecuente\}$$

Esto indica que existe una relación entre los clientes que compran el antecedente y el consecuente en la misma transacción. La fuerza y el sentido de la relación se mide con diferentes indicadores: el soporte, la confianza y la mejora de la confianza.

Transaction 1	
Transaction 2	
Transaction 3	
Transaction 4	
Transaction 5	
Transaction 6	

Matriz de Incidencia

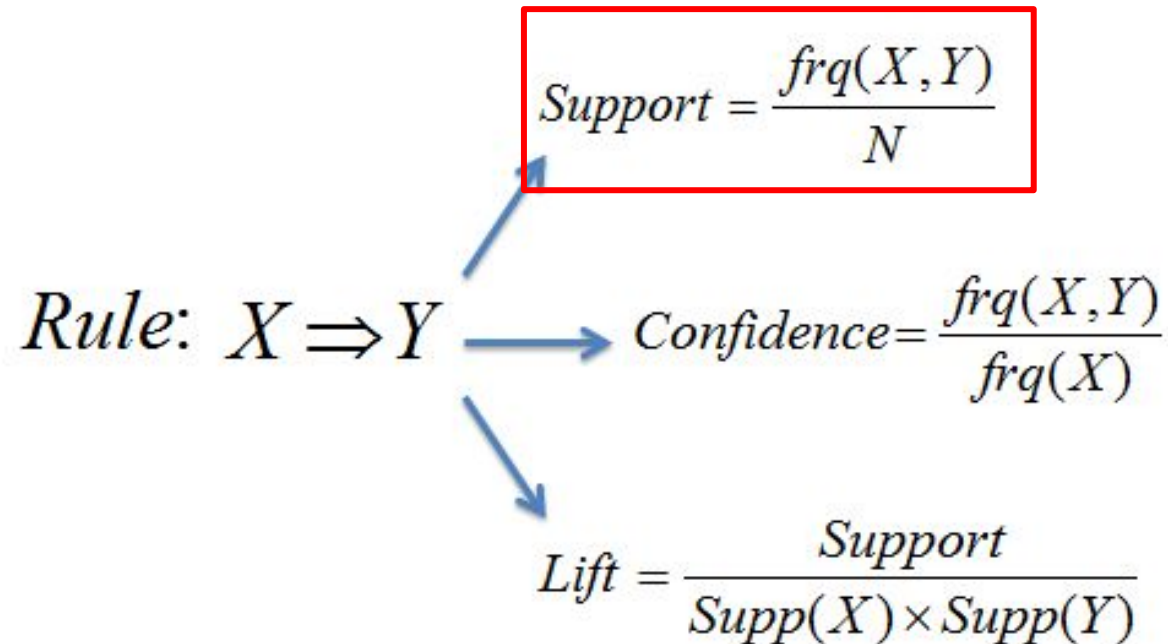
Pan	Leche	Papas fritas	Mostaza	Cerveza	Pañales	Huevos	Gaseosas
1	1	1	0	0	0	0	0
1	0	1	0	1	1	1	0
0	1	0	0	1	1	0	1
1	1	0	0	1	1	0	0
1	1	0	1	0	1	0	1
1	1	0	0	1	1	0	0
1	1	1	0	0	1	0	1

Ejemplos:

{Pan}  {Leche}
{Pan, Leche}  {Huevos}

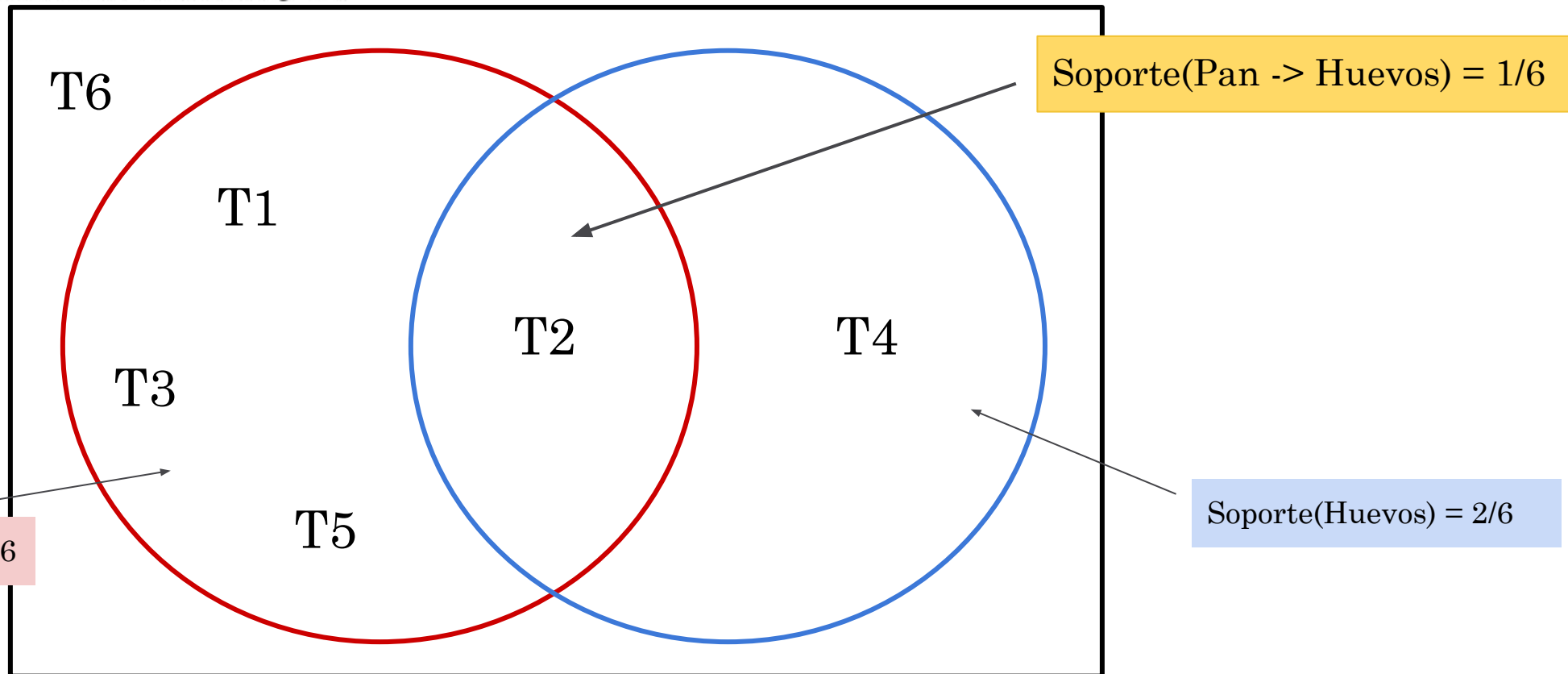
Soporte

- El soporte es la frecuencia relativa con la que se observa la regla. Es decir, un soporte de 0.15 indica que el antecedente y el consecuente se observan a la vez en el 15% de las transacciones. Este indicador mide la fuerza de la regla. Al ser un porcentaje, los posibles valores del soporte se encuentran entre 0 y 1.



Soporte

- El soporte es la frecuencia relativa con la que se observa la regla. Es decir, un soporte de 0.15 indica que el antecedente y el consecuente se observan a la vez en el 15% de las transacciones. Este indicador mide la fuerza de la regla. Al ser un porcentaje, los posibles valores del soporte se encuentran entre 0 y 1.



Confianza

La confianza es el porcentaje de las transacciones en las que aparece el antecedente en la que también aparece el consecuente. Lo que mide este indicador es la fiabilidad de la regla.

Matemáticamente se puede obtener utilizando la expresión

$$\text{conf}(\{\text{Antecedente}\} \Rightarrow \{\text{Consecuente}\}) = \frac{\text{soporte}(\{\text{Antecedente}, \text{Consecuente}\})}{\text{soporte}(\{\text{Antecedente}\})}$$

En donde $\text{conf}(\{\text{Antecedente}\} \Rightarrow \{\text{Consecuente}\})$ es la confianza de los registros en los que aparece a la vez el antecedente y consecuente.

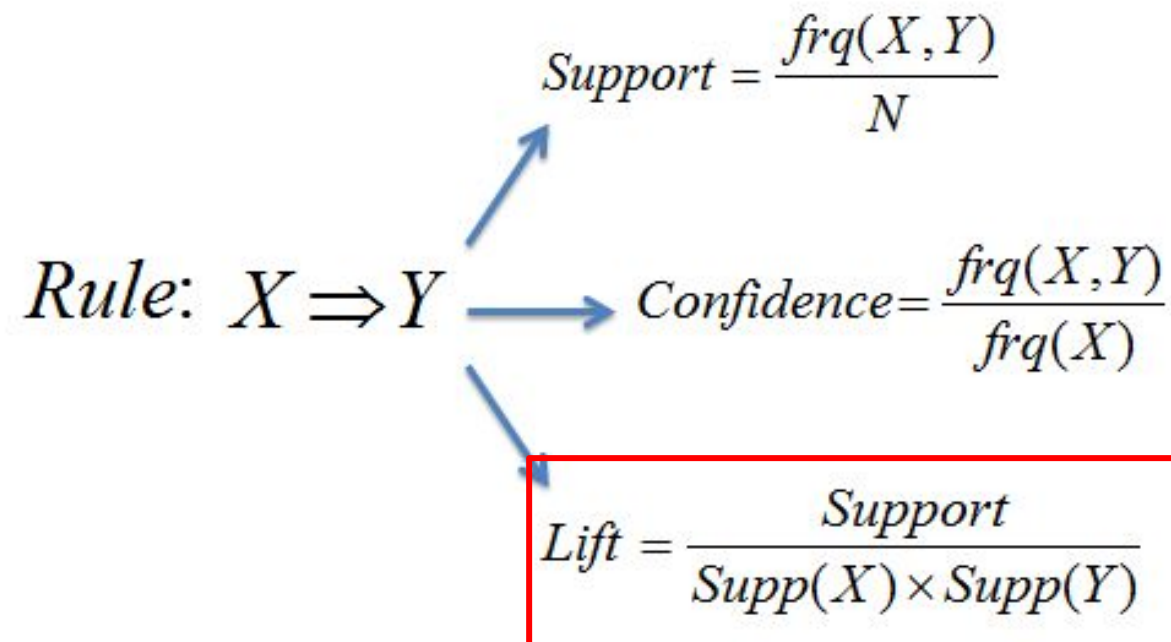
The diagram illustrates the relationship between a rule and its associated metrics. A central rule, $\text{Rule: } X \Rightarrow Y$, is shown on the left. Three blue arrows originate from this rule and point to three different metrics on the right:

- An arrow pointing up and to the right to the formula for Support: $\text{Support} = \frac{\text{frq}(X, Y)}{N}$
- An arrow pointing straight to the right to the formula for Confidence: $\text{Confidence} = \frac{\text{frq}(X, Y)}{\text{frq}(X)}$. This formula is enclosed in a red rectangular box.
- An arrow pointing down and to the right to the formula for Lift: $\text{Lift} = \frac{\text{Support}}{\text{Supp}(X) \times \text{Supp}(Y)}$

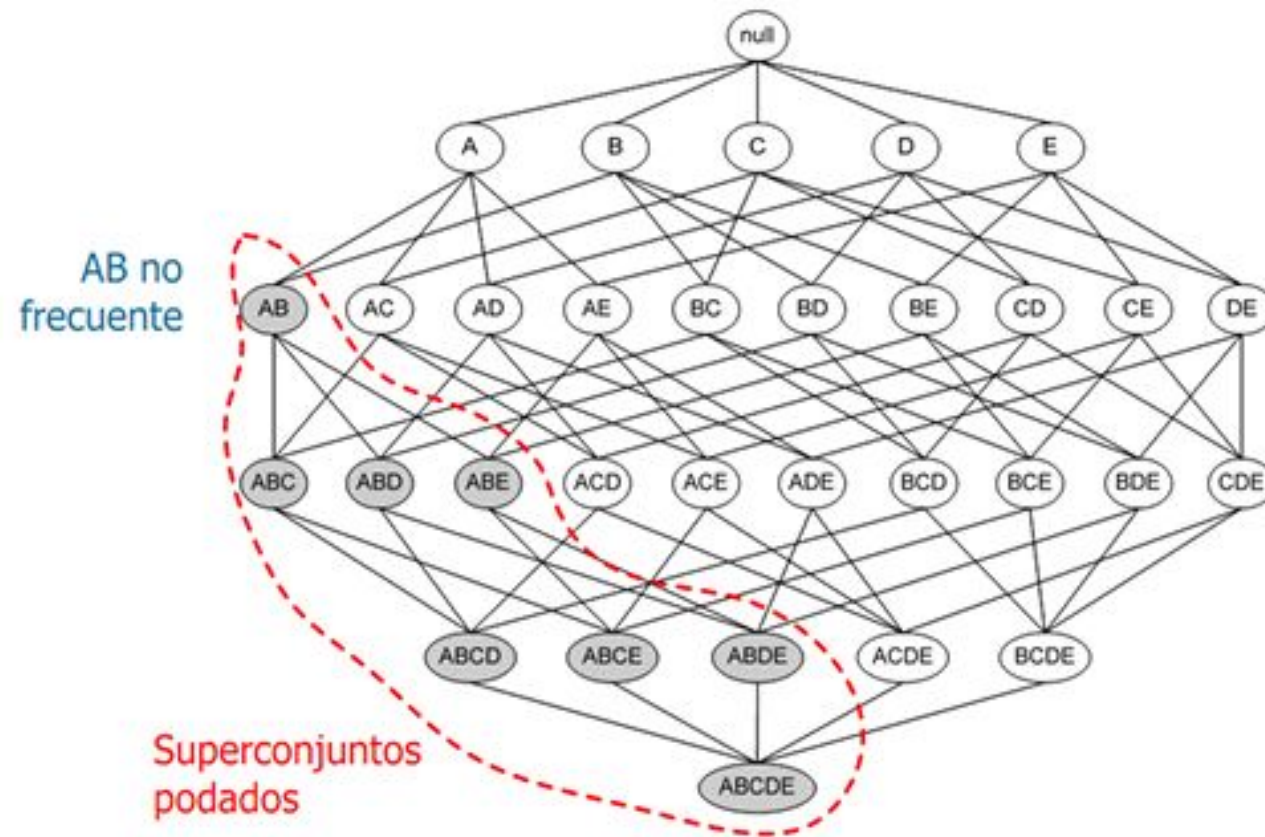
Lift

Compara la frecuencia observada de una regla con la frecuencia esperada simplemente por azar (si la regla no existe realmente)

Cuanto más se aleje el valor de lift de 1, más evidencias de que la regla no se debe a un artefacto aleatorio, es decir, mayor la evidencia de que la regla representa un patrón real.



¿Cuántas reglas puedo crear? 2^n



Algoritmo Apriori

El desafío es reducir el número de asociaciones que voy a estar evaluando en mis datos. Para esto me va a interesar encontrar una forma de quedarme solamente con conjuntos de items que sean frecuentes (es decir, que superen cierto umbral de soporte).

Aquí entra en juego el llamado Algoritmo Apriori que establece que:

Si un conjunto de items no es frecuente, tampoco lo será ningún superconjunto que lo contenga.

