

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles, and others with solid blue dots. The lines are thin and grey, creating a mesh-like structure.


Reglas de Asociación

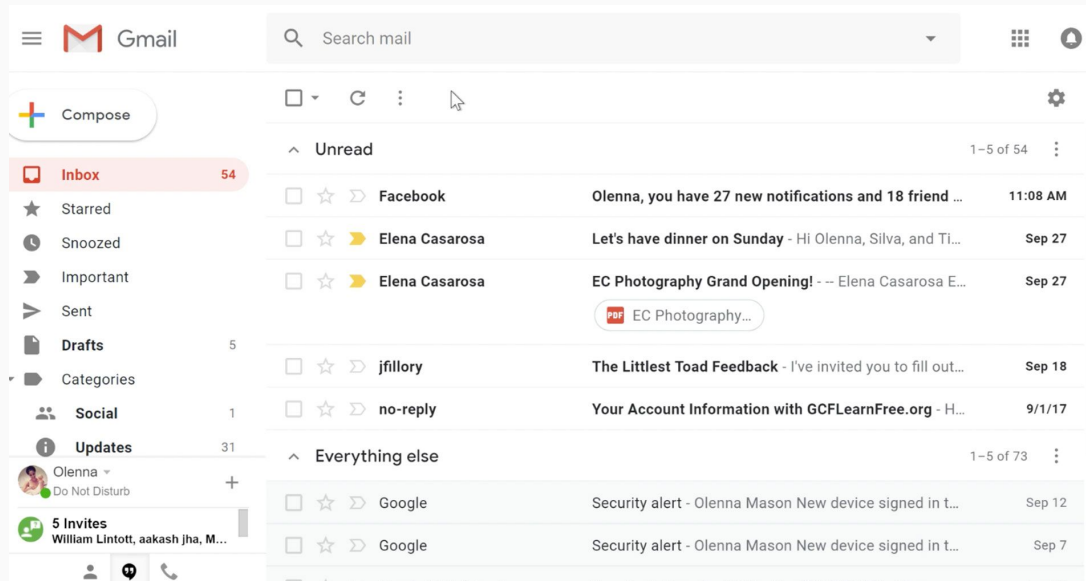
Dr. Jorge Guerra (Jorch)

jorge.guerra881215@gmail.com

A decorative network diagram in the bottom-right corner, similar to the one in the top-left, showing a web of nodes and lines with some nodes highlighted in blue.

Problema:

Supongamos que queremos ayudar a Google en su producto  Gmail creando una cola con prioridad para la bandeja de entrada. De esta manera los correos de entrada se irán ordenando no por orden de llegada sino por orden de importancia respecto a cada usuario.



Aprendizaje automático

*Datos etiquetados
con clase*

Datos no etiquetados

Aprendizaje Supervisado

Aprendizaje No Supervisado

Clase discreta

Clase continua

Dividir por similitud

*Encontrar
dependencias*

*Identificar
patrones*

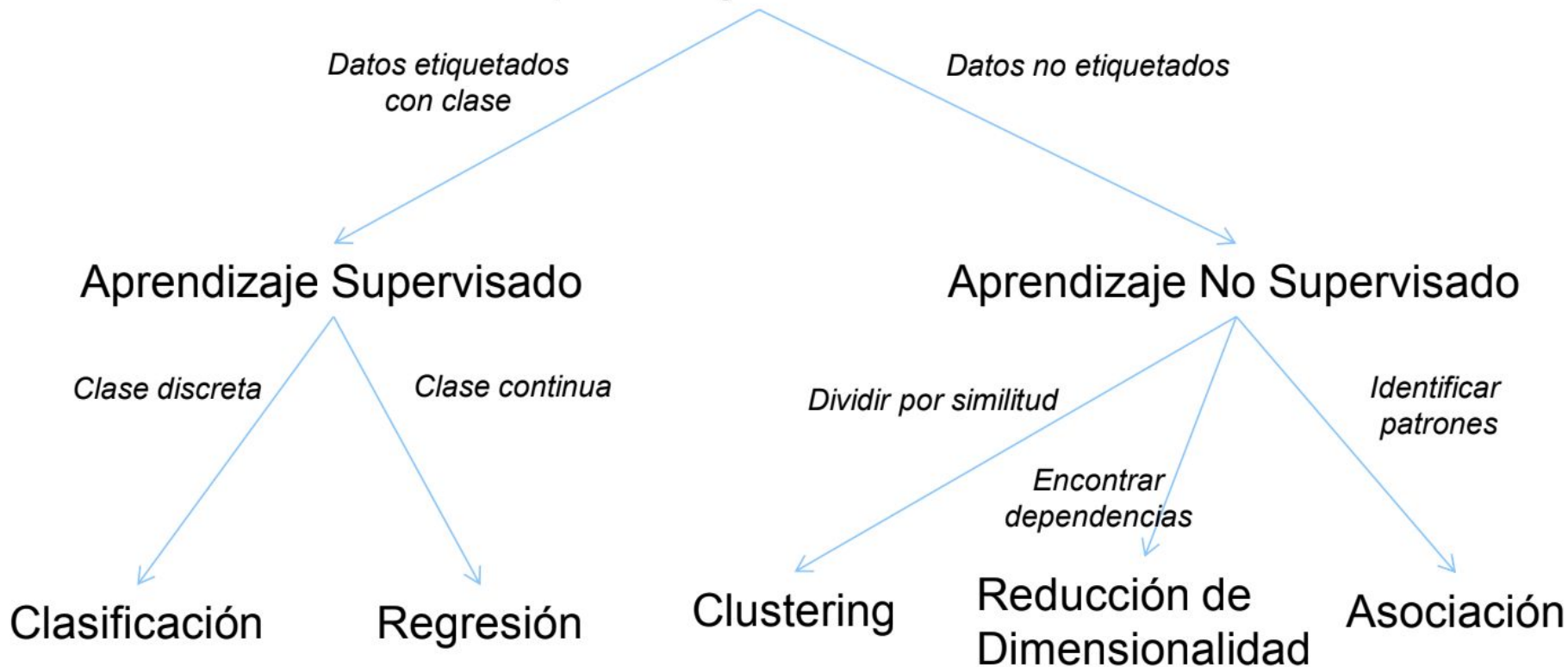
Clasificación

Regresión

Clustering

Reducción de
Dimensionalidad

Asociación



Reglas de Asociación



Regla de asociación:

Describe una relación de asociación entre los elementos de un conjunto de datos relevantes.

Ejemplos:

- Estudiantes que cursan ***Machine Learning*** tienden a cursar ***Estadística Aplicada***
- Clientes que compran productos ***lácteos*** tienden a comprar productos ***panificados***.
- Artículos que referencian a ***Srikant (1997)*** citan también a ***Agrawal. (1993)***.

Origen: Market Basket Analysis

Problema: identificar el conjunto de ítems que son adquiridos en conjunto.

{fideos, queso rallado} -> {salsa}

{viernes, persona adulta, carne} -> {fernet, coca-cola}



Definición Reglas de Asociación

De manera general: $X \rightarrow Y$ donde X e Y son conjuntos de ítems del dominio.

X se denomina el **antecedente** de la regla donde Y sería su **consecuente**.

Definición Reglas de Asociación

De manera general: $X \rightarrow Y$ donde X e Y son conjuntos de ítems del dominio.

X se denomina el **antecedente** de la regla donde Y sería su **consecuente**.

- **Soporte:** El soporte para la regla $X \rightarrow Y$ es el porcentaje de las transacciones que contienen todos los ítems de X e Y .
- **Confianza:** La confianza para la regla $X \rightarrow Y$ es el porcentaje de transacciones que contienen Y , entre las transacciones que contienen X .

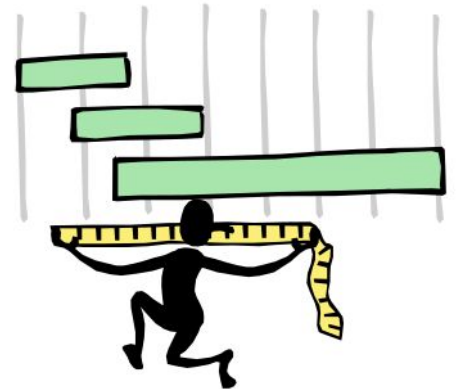
- $\text{Soporte}(X \rightarrow Y) = \text{Prob}(X \cup Y) = \text{Soporte}(X \cup Y)$

- $\text{Confianza}(X \rightarrow Y) = \text{Prob}(Y / X) = \frac{\text{Soporte}(X \cup Y)}{\text{Soporte}(X)}$

Transacciones
A B C
B C
A C
A C D

Soporte ($A \rightarrow C$): ???

Confianza ($A \rightarrow C$): ???



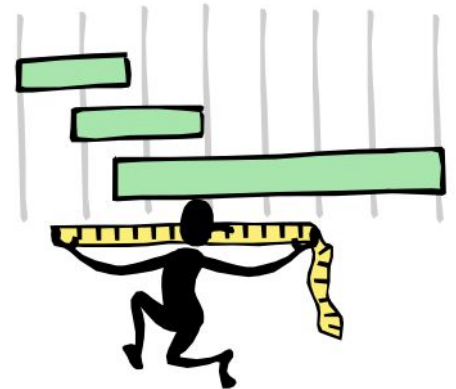
- $\text{Soporte}(X \rightarrow Y) = \text{Prob}(X \cup Y) = \text{Soporte}(X \cup Y)$

- $\text{Confianza}(X \rightarrow Y) = \text{Prob}(Y / X) = \frac{\text{Soporte}(X \cup Y)}{\text{Soporte}(X)}$

Transacciones
A B C
B C
A C
A C D

Soporte ($A \rightarrow C$): **0,75**

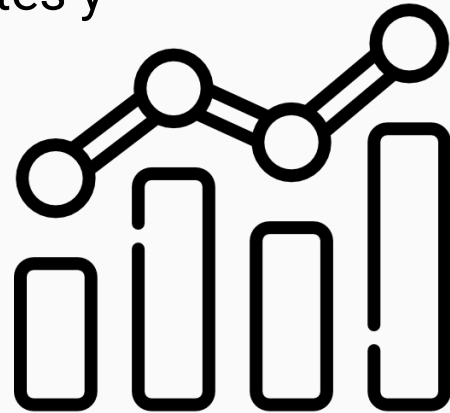
Confianza ($A \rightarrow C$): **1**



Interpretación de las Métricas

- Regla de bajo soporte
 - Puede haber aparecido por casualidad.
- Regla con baja confianza
 - Es probable que no exista relación entre antecedentes y consecuente

¿Que diferencia a $X \rightarrow Y$ de $Y \rightarrow X$?

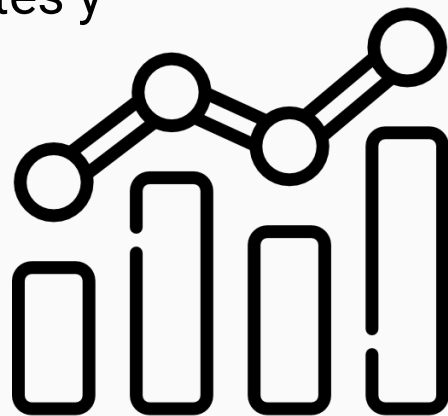
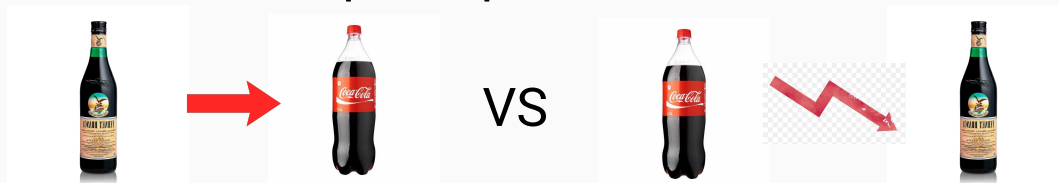


Interpretación de las Métricas

- Regla de bajo soporte
 - Puede haber aparecido por casualidad.
- Regla con baja confianza
 - Es probable que no exista relación entre antecedentes y consecuente

¿Que diferencia a $X \rightarrow Y$ de $Y \rightarrow X$?

- Tiene el mismo soporte pero distinta confianza



Algoritmo de Descubrimiento



Objetivo

Encontrar Reglas de Asociación con altos valores de soporte y confianza

- Umbrales de ***minsup*** y ***minconf*** definidos por el usuario

Importante: Encontrar dichas reglas no significa que deba existir una relación entre antecedente y consecuente. Por lo tanto, un experto en el dominio del problema debería siempre evaluar las reglas.

$I = \{ i_1, i_2, \dots, i_m \}$ es un conjunto de ítems.

D es un conjunto de transacciones T_j . Donde cada T_j es un conjunto de ítems (subconjunto de I).

TID	Transacciones
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E

- $I = \{A, B, C, D, E\}$
- $D = \{1, 2, 3, 4, 5\} = \{\{A, C, D\}, \{B, C, E\}, \{A, B, C, E\}, \{B, E\}, \{A, B, C, E\}\}$

TID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E

- Un itemset es un conjunto de ítems.
- Si X es un **itemset**, $X \subseteq I$.
- Un **itemset** que contiene k ítems es llamado **k-itemset**.
- Ej. $\{A,B\}$ es un 2-itemset

El soporte de un itemset X es el porcentaje de transacciones en D que contienen X

$$\text{Soporte}(X) = \frac{|\{T \in D / X \subset T\}|}{|D|}$$

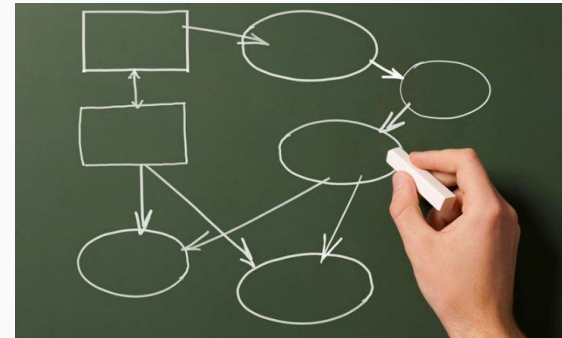
$$\text{Soporte}(\{A,B\}) = 2/5 = 0.4$$

El descubrimiento de las reglas puede ser descompuesto de dos subproblemas:

1. Encontrar todos los **itemsets** que tienen el soporte mayor que el **soporte mínimo (itemsets frecuentes)**.
2. Utilizar los **itemsets frecuentes** para generar las reglas deseadas.
 - a. Para cada **itemsets frecuentes** L ($k > 1$), encontrar todos los subconjuntos no vacíos, y para cada subconjunto $\{a\}$ generar una regla $\{a\} \rightarrow \{L - a\}$ si la confianza es mayor que el *minconf*.
 - i. Para el itemset frecuente $\{A, B, C\}$:
 $\{A\} \rightarrow \{BC\}$, $\{AB\} \rightarrow \{C\}$, $\{AC\} \rightarrow \{B\}$, $\{B\} \rightarrow \{AC\}$, $\{BC\} \rightarrow \{A\}$, $\{C\} \rightarrow \{AB\}$

Algoritmos

- Apriori y AprioriTid (Agrawal & Srikant, 1994)
- Opus (Webb, 1996)
- Direct Hasing and Pruning (DHP) (Adamo, 2001)
- Dynamic Set Counting (DIC) (Adamo, 2001)
- Charm (Zaki & Hsiao, 2002)
- FP-growth (Han, Pei & Yin, 1999)
- Closet (Pei, Han & Mao, 2000)



¿Qué los hace diferentes?

- Forma en que los datos son cargados en memoria
- Tiempo de procesamiento
- Tipos de atributos (numéricos, categóricos)
- Forma en que los itemsets son generados
- Estructura de datos utilizada

¿Qué los hace diferentes?

- Forma en que los datos son cargados en memoria
- Tiempo de procesamiento
- Tipos de atributos (numéricos, categóricos)
- Forma en que los itemsets son generados
- Estructura de datos utilizada

Los diferentes algoritmos deben siempre generar el mismo conocimiento.

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

- Soporte mínimo (*minsup*) = 0.7

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

- Soporte mínimo (*minsup*) = 0.7
- Nivel 1: Encontrar 1-itemsets frecuentes
{lapicera},{tinta},{agenda},{jabón}, {arroz}

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

- Soporte mínimo ($minsup$) = 0.7
- Nivel 1: Encontrar 1-itemsets frecuentes
~~{lapicera}, {tinta}, {agenda}, {jabón}, {arroz}~~

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

- Soporte mínimo ($minsup$) = 0.7
- Nivel 1: Encontrar 1-itemsets frecuentes
 $\{lapicera\}, \{tinta\}, \{agenda\},$ ~~$\{jabón\}, \{arroz\}$~~
- Nivel 2: Encontrar 2-itemsets frecuentes
 $\{lapicera, tinta\}, \{lapicera, agenda\},$ ~~$\{lapicera, jabón\}, \{lapicera, arroz\}, \{tinta, agenda\}, \{tinta, jabón\}, \{tinta, arroz\}, \{agenda, jabón\}, \{agenda, arroz\}$~~ ...

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

- Nivel 3:

~~{lapicera, tinta, agenda}, {lapicera, tinta, jabón},~~
~~{lapicera, tinta, arroz}, {lapicera, agenda, jabón},~~
~~{lapicera, agenda, arroz} ...~~

- Nivel 4:

~~{lapicera, tinta, agenda, jabón}, {lapicera, tinta,~~
~~agenda, arroz}, {tinta, agenda, arroz, jabón} ...~~

Ejemplo

TID	ítems
111	lapicera, tinta, agenda, jabón
112	lapicera, tinta, agenda
113	lapicera, agenda
114	lapicera, tinta, jabón, arroz

Los itemsets frecuentes son:
{lapicera}, {tinta}, {agenda}
{lapicera, tinta}, {lapicera, agenda}

¿Qué relación existe entre los n -itemsets y los $n+1$ -itemsets eliminados?

¿Qué relación existe entre los n -itemsets y los $n+1$ -itemsets eliminados?

- Nivel 1:
 - {lapicera}, {tinta}, {agenda}, {jabón}, {arroz}
- Nivel 2:
 - {lapicera, tinta}, {lapicera, agenda}, {lapicera, jabón}, {lapicera, arroz}, {tinta, agenda}, {tinta, jabón}, {tinta, arroz}, {agenda, jabón}, {agenda, arroz}
- Nivel 3:
 - {lapicera, tinta, agenda}, {lapicera, tinta, jabón}, {lapicera, tinta, arroz}, {lapicera, agenda, jabón}, {lapicera, agenda, arroz}

¿Qué relación existe entre los n -itemsets y los $n+1$ -itemsets eliminados?

- Nivel 1:
 - {lapicera}, {tinta}, {agenda}, {jabón}, {arroz}
- Nivel 2:
 - {lapicera, tinta}, {lapicera, agenda}, {lapicera, jabón}, {lapicera, arroz}, {tinta, agenda}, {tinta, jabón}, {tinta, arroz}, {agenda, jabón}, {agenda, arroz}
- Nivel 3:
 - {lapicera, tinta, agenda}, {lapicera, tinta, jabón}, {lapicera, tinta, arroz}, {lapicera, agenda, jabón}, {lapicera, agenda, arroz}

Refinamiento: extender los itemsets frecuentes de una forma que asegure que todos sus subconjuntos son itemsets frecuentes.

Propiedad Apriori

Propiedad Apriori: cada subconjunto de un itemset frecuente debe ser también un itemset frecuente.

Podemos crear itemsets frecuentes iterativamente, tomando los itemsets frecuentes de tamaño n y extendiéndolos a itemsets frecuentes de tamaño $n+1$.

Algoritmo Apriori



1. Se calcula el soporte de cada ítem individual, y se determinan los 1-itemsets frecuentes.
2. En cada paso subsecuente, los itemsets frecuentes generados en los pasos anteriores se utilizan para generar los nuevos itemsets (itemsets candidatos).
3. Se calcula el soporte de cada itemset candidato y se determinan los itemsets frecuentes.
4. El proceso continúa hasta que no pueden ser encontrados nuevos itemsets frecuentes.

D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

minsup= 0.5**L₁**

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

C₂

Itemset	Soporte
{1, 2}	1/4
{1, 3}	2/4
{1, 5}	1/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

L₂

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4

C₃

Itemset	Soporte
{2, 3, 5}	2/4

L₃

Itemset	Soporte
{2, 3, 5}	2/4

Derivación de Reglas de asociación

Para cada itemset frecuente I , se generan todos los subconjuntos no vacíos de I .

Para cada subconjunto $a \subset I$ se genera una regla de la forma $a \rightarrow (I-a)$ si la tasa entre **soporte**(I) y **soporte**(a) es al menos la confianza mínima (**minconf**).

Itemsets frecuentes

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4
{2, 3, 5}	2/4

L₁

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Reglas encontradas

Regla	Conf.	Regla	Conf
$1 \rightarrow 3$		$2,5 \rightarrow 3$	
$2 \rightarrow 3,5$		$2 \rightarrow 5$	
$3 \rightarrow 1$		$5 \rightarrow 2,3$	
$2,3 \rightarrow 5$		$5 \rightarrow 2$	
$2 \rightarrow 3$		$3 \rightarrow 5$	
$3 \rightarrow 2,5$		$5 \rightarrow 3$	
$3 \rightarrow 2$		$3,5 \rightarrow 2$	

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$



Itemsets frecuentes

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4
{2, 3, 5}	2/4

L₁

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Reglas encontradas

Regla	Conf.	Regla	Conf
$1 \rightarrow 3$	2/2	$2,5 \rightarrow 3$	
$2 \rightarrow 3,5$		$2 \rightarrow 5$	
$3 \rightarrow 1$		$5 \rightarrow 2,3$	
$2,3 \rightarrow 5$		$5 \rightarrow 2$	
$2 \rightarrow 3$		$3 \rightarrow 5$	
$3 \rightarrow 2,5$		$5 \rightarrow 3$	
$3 \rightarrow 2$		$3,5 \rightarrow 2$	

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$



Itemsets frecuentes

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4
{2, 3, 5}	2/4

L₁

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Reglas encontradas

Regla	Conf.	Regla	Conf
$1 \rightarrow 3$	2/2	$2,5 \rightarrow 3$	
$2 \rightarrow 3,5$	2/3	$2 \rightarrow 5$	
$3 \rightarrow 1$		$5 \rightarrow 2,3$	
$2,3 \rightarrow 5$		$5 \rightarrow 2$	
$2 \rightarrow 3$		$3 \rightarrow 5$	
$3 \rightarrow 2,5$		$5 \rightarrow 3$	
$3 \rightarrow 2$		$3,5 \rightarrow 2$	

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$



Itemsets frecuentes

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4
{2, 3, 5}	2/4

L₁

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Reglas encontradas

Regla	Conf.	Regla	Conf
$1 \rightarrow 3$	2/2	$2,5 \rightarrow 3$	
$2 \rightarrow 3,5$	2/3	$2 \rightarrow 5$	
$3 \rightarrow 1$	2/3	$5 \rightarrow 2,3$	
$2,3 \rightarrow 5$	2/2	$5 \rightarrow 2$	
$2 \rightarrow 3$	2/3	$3 \rightarrow 5$	
$3 \rightarrow 2,5$	2/3	$5 \rightarrow 3$	
$3 \rightarrow 2$	2/3	$3,5 \rightarrow 2$	

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$



Itemsets frecuentes

Itemset	Soporte
{1, 3}	2/4
{2, 3}	2/4
{2, 5}	3/4
{3, 5}	2/4
{2, 3, 5}	2/4

L₁

Itemset	Soporte
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

Reglas encontradas

Regla	Conf.	Regla	Conf
1 → 3	2/2	2,5 → 3	2/3
2 → 3,5	2/3	2 → 5	3/3
3 → 1	2/3	5 → 2,3	2/3
2,3 → 5	2/2	5 → 2	3/3
2 → 3	2/3	3 → 5	2/3
3 → 2,5	2/3	5 → 3	2/3
3 → 2	2/3	3,5 → 2	2/2

$$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X, Y)}{\text{Soporte}(X)}$$

minconf = 0.9



Software



- MLxtend: <http://rasbt.github.io/mlxtend/>
- Weka (<http://www.cs.waikato.ac.nz/~ml/weka/>)
- Knime: <http://www.knime.org/>
- RapidMiner: <http://rapid-i.com/content/view/181/190/>
- AIAS: Association Interestingness Analysis System
(<http://www.comp.nus.edu.sg/~dm2>)
- Gnome Data Mine (linux):
(<http://www.togaware.com/datamining/gdatamine/gdmapriori.html>)
- FIMI, Frequent Itemset Mining Implementations repository, incluye software y datasets.
- Useful links: <http://www.kdnuggets.com/software/associations.html>

Conclusiones

- Las reglas de asociación son útiles para descubrir asociaciones entre conjuntos de ítems en una base de datos de transacciones.
- Puede ser utilizado en múltiples dominios: análisis de canasta de mercado, datos de censos, recomendación, aprendizaje en sistemas multiagentes, etc..
- Varios algoritmos de descubrimiento de reglas de asociación, descubren el mismo conocimiento.
- Son necesarias diversas tareas de post-procesamiento para eliminar reglas no interesantes, podar reglas redundantes, etc.
- Extensiones: Reglas de asociación generalizadas, difusas, temporales, etc.

Bibliografía

- <https://towardsdatascience.com/association-rules-2-aa9a77241654>
- <https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84>
- R. Agrawal, R. Srikant - Fast Algorithms for Mining Association Rules - Proc. of the 20th International Conference on Very Large Databases, Santiago, Chile, Sept. 1994.
- M. Klemetinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. Proc. CIKM, 1994.
- D. Shah, L. V. S. Lakshmanan, K. Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 1999.

Reglas de Asociación

Dr. Jorge Guerra (Jorch)

jorge.guerra881215@gmail.com

