



PUCP

Curso: Python aplicado a Data Science

Sesión 4: Análisis de Asociaciones



Contenido del curso

Sesión 1: Introducción a Ciencia de Datos y Python

Sesión 2: Preprocesamiento de datos con **Numpy y Pandas**

Sesión 3: **Visualización** de datos con Seaborn, Matplotlib y Plotly

Sesión 4: Análisis de Asociaciones: Market Basket Analysis

Sesión 5: Visualización y Transformación de datos (**PCA**)

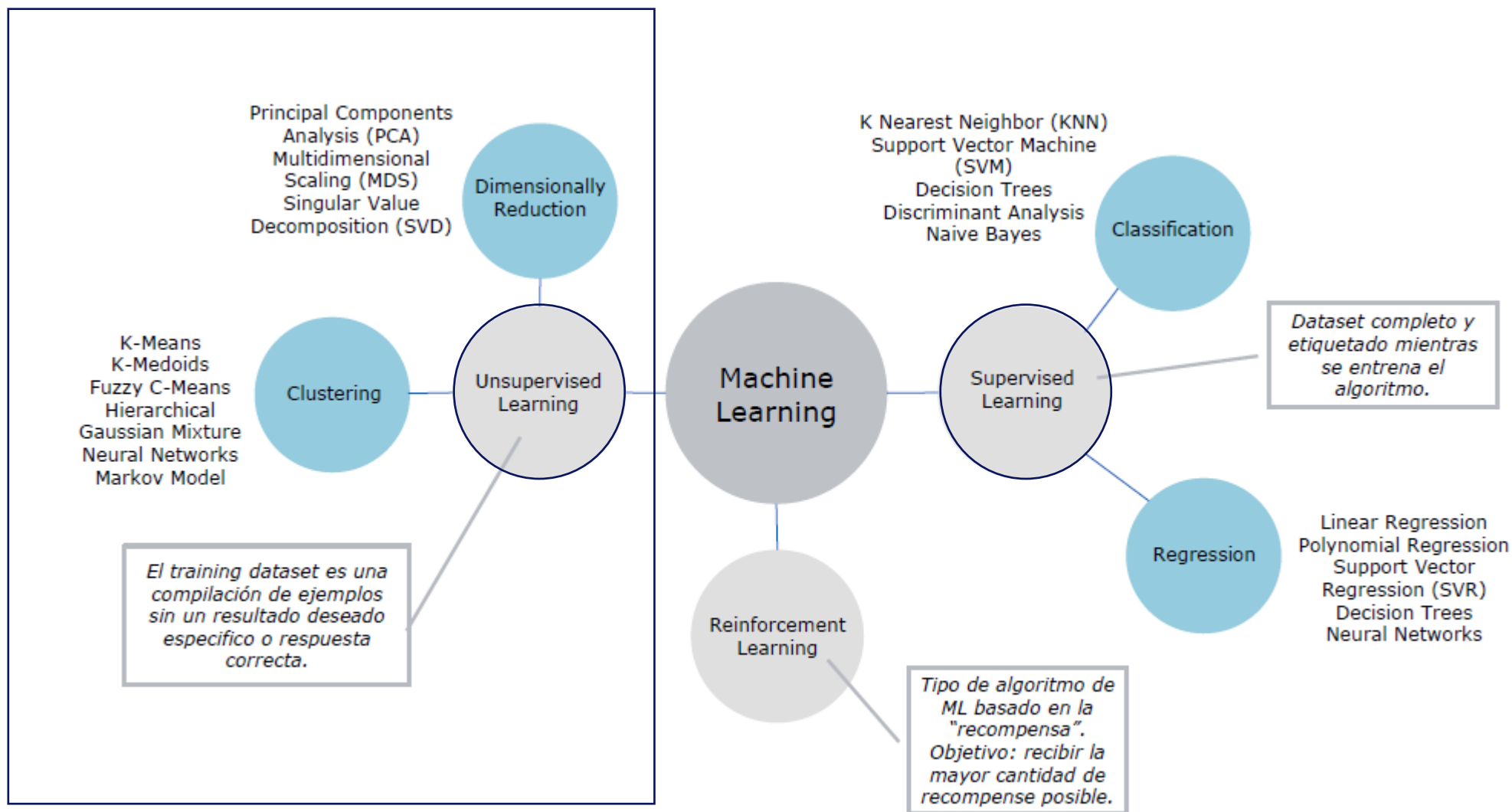
Sesión 6: Análisis de Agrupamientos: **K-Means, DBSCAN**

Sesión 7: Analítica predictiva: Modelos de **regresión**

Sesión 8: Analítica predictiva: Modelos de **clasificación**

Agenda

- Introducción: Tipos de Aprendizaje
- Reglas de Asociación
- Definiciones (soporte, confianza, lift)
- Algoritmo A-Priori
- Caso: Groceries Mall



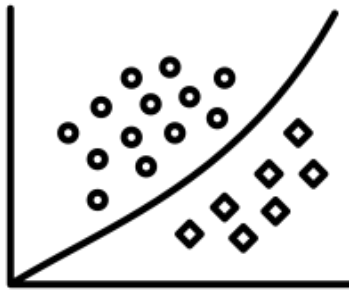
Veamos los 2 grandes tipos de aprendizaje...

Aprendizaje Supervisado

El conjunto de datos consiste **en variables y target**.

Se **entrena un modelo** para a partir de ello **predecir las etiquetas** en conjunto de datos nuevo.

Ejm. Predicción de las ventas, predicción de enfermedad.

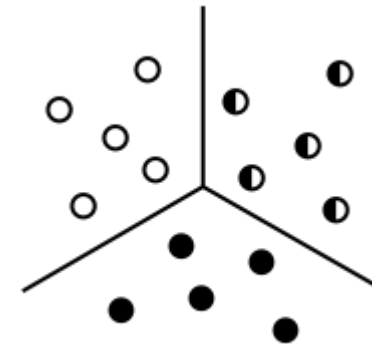


Aprendizaje No Supervisado

El conjunto de datos consiste **en variables sin etiquetar**.

Se **entrena un modelo** para encontrar patrones ocultos o **similitud entre los datos** teniendo en cuenta sus características.

Ejm. Análisis de Asociaciones, Segmentación de clientes, Reducción de la Dimensionalidad



Reglas de Asociación (Market Basket Analysis)

- Es una técnica utilizada para **descubrir relaciones** entre los productos que compran los usuarios.
- Se observan las combinaciones de **productos que son compradas conjuntamente** en las transacciones.
- Utilizando esta información, ¿**es posible que una tienda pueda tomar decisiones?**



Información sin organización



```
1 0 1 0 1 0 1 0 1 0 1 0 1
1 1 1 1 1 0 1 0 1 1 1 1 1
1 1 0 0 1 0 1 0 1 1 0 0 1
1 0 1 0 1 0 1 0 0 0 1 1 0
1 0 1 0 1 0 1 0 1 0 1 0 1
0 1 0 1 1 0 1 0 1 1 0 0 1
1 0 1 1 1 0 1 0 1 1 0 1 0
1 1 0 0 1 0 1 0 1 0 1 0 1
```

Procesamiento por el algoritmo



Asociaciones inteligentes

Casos de Uso



Colocar ambos **productos** **cerca** del otro en un supermercado (ordenamiento)



Aplicar **descuentos** para uno de los dos productos.




Ofrecer **promociones** de un producto a compradores del otro producto. (pares de productos)



Generar **nuevos productos** o bundles a partir de los productos originales.


Sistema Recomendador



See all 2 images

Read sample Audible sample

Follow the Author



Eric Carle







Follow


THE all-time classic picture book, from generation to generation, sold somewhere in the world every 30 seconds! A sturdy and beautiful book to give as a gift for new babies, baby showers, birthdays, and other new beginnings!

Featuring interactive die-cut pages, this board book edition is the perfect size for little hands and great for teaching counting and days of the week.

Read more

Report incorrect product information.

Reading age	Part of series	Print length	Language	Lexile measure	Dimensions
 1 - 3 years, from customers	 The World of Eric Carle	 26 pages	 English	 AD460L	 7.13 x 0.63 x 5 inches



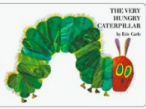
My First Book of Nursery Rhymes - Padded Board Book - Classics

★★★★★ 2,019

\$8.99 prime

Sponsored

Frequently bought together




This item: The Very Hungry Caterpillar

\$6.56

7 pts

+




Brown Bear, Brown Bear, What Do You See?

\$4.94

5 pts

+



Goodnight Moon

\$5.36

6 pts

Total price: \$16.86

Total Points: 18 pt


Add all three to Cart

Películas Series

9. Busca la Luz


43 MIN 18+

Más contenido como este



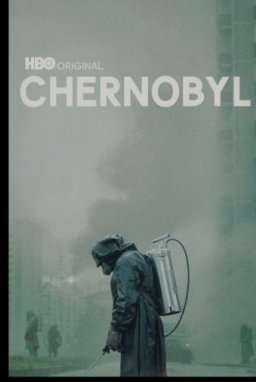
INTERESTELAR

HBO



THE WHITE LOTUS

HBO



CHERNOBYL

HBO

Reglas de Asociación

El Análisis de Asociación, o Análisis de **Reglas de Asociación**, se define como la tarea de **encontrar relaciones interesantes/relevantes** en un largo conjunto de datos.

Dicho de otro modo, se trata de descubrir cómo diferentes elementos se encuentran **asociados** entre sí.

El Algoritmo A-Priori (Agrawal, 1994) nos permite encontrar reglas de asociación en forma automática desde los datos.

$$\{X \rightarrow Y\}$$























{Antecedente → Consecuente}

Regla de asociación

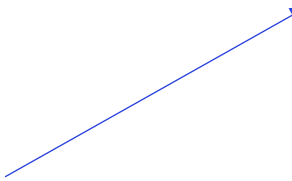
Y sucede si es que ha sucedido X
(el sentido inverso no es igual)

Definiciones - Itemset

- Colección de **1 o más ítems** dentro de un set de transacciones.
- Cada fila es una transacción (una sola compra) y todos los productos de la transacción se compraron al mismo tiempo.

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

Itemset



Soporte (support)

Esta métrica indica que “**tan popular**” es un conjunto de elementos (**itemset**).























Se mide como la proporción de las transacciones en las que el “itemset” ha aparecido en los datos.

Ejemplo:

El valor de “**support**” de {**apple**} es 4 de 8, o 50%.

$$\text{Support} \{\text{🍏}\} = \frac{4}{8}$$

Los “itemsets” pueden contener múltiples elementos.
Por ejemplo, el “support” de {**apple, beer, rice**} sería 2 de 8, o 25%.

Transaction 1	   	●
Transaction 2	  	●
Transaction 3	 	
Transaction 4	 	
Transaction 5	   	●
Transaction 6	  	●
Transaction 7	 	●
Transaction 8	 	

¿Cuál es el **soporte** de {**milk, beer**}?:

3/8

Confianza (confidence)

Esta medida nos indica que tan usual sería que un ítem Y sea comprado si es que el ítem X también fue comprado: $\{X \rightarrow Y\}$
Se calcula con la proporción de transacciones con el ítem X en las que también aparece el ítem Y:

$$\text{Confidence} \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support} \{\text{🍎, 🍺}\}}{\text{Support} \{\text{🍎}\}}$$

Puede representar erróneamente la importancia de una asociación.

¿Qué pasa si beer es un ítem muy popular y aparece en casi todas las transacciones?

Regla: Leche \rightarrow Cerveza

$$c(\text{Leche} \rightarrow \text{Cerveza}) = \frac{\sigma(\text{Leche, Cerveza})}{\sigma(\text{Leche})} = 0.7$$

Pero $\sigma(\text{Cerveza}) = 0.7$

Transaction 1	🍎 🍺 🥛 🍗
Transaction 2	🍎 🍺 🥛
Transaction 3	🍎 🍺
Transaction 4	🍎 🍏
Transaction 5	🍼 🍺 🥛 🍗
Transaction 6	🍼 🍺 🥛
Transaction 7	🍼 🍺
Transaction 8	🍼 🍏

Lift (elevación)

Esta medida indica que tan usual sería que un ítem Y sea comprado si es que el ítem X también fue comprado {X->Y} **tomando en cuenta que tan popular es el ítem Y.**

Por ejemplo, “lift” de {apple -> beer} es 1, que implica que no hay una asociación entre los ítems. Asimismo:

$$\text{Lift} \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support} \{\text{🍎}, \text{🍺}\}}{\text{Support} \{\text{🍎}\} \times \text{Support} \{\text{🍺}\}}$$

El numerador nos indica la proporción de transacciones que contienen X e Y, mientras que el denominador la proporción de X e Y como elementos independientes.

Si el valor fuera **mayor que 1**, el ítem Y sería usualmente comprado si X es comprado. Si el valor fuera **menor que 1**, el ítem Y no sería usualmente comprado si el ítem X es comprado.

Transaction 1	🍎 🍺 🍲 🍗
Transaction 2	🍎 🍺 🍲
Transaction 3	🍎 🍺
Transaction 4	🍎 🍏
Transaction 5	🍼 🍺 🍲 🍗
Transaction 6	🍼 🍺 🍲
Transaction 7	🍼 🍺
Transaction 8	🍼 🍏

Resumen de métricas

$$\text{Support}(X) = \frac{\text{Frequency}(X)}{N}$$

$$\text{Support}(X \rightarrow Y) = \frac{\text{Frequency}(X \& Y)}{N}$$

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X)}$$

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X) \text{Support}(Y)}$$

$$\text{Leverage}(X \rightarrow Y) =$$

$$\text{Support}(X \& Y) - \text{Support}(X) \text{Support}(Y)$$

$$\text{Conviction}(X \rightarrow Y) =$$

$$\frac{\text{Support}(X) \text{Support}(\bar{Y})}{\text{Support}(X \& \bar{Y})}$$

Resumen de métricas

- **Soporte**
 - **Frecuencia** relativa del itemset
- **Confianza**
 - **Probabilidad** empírica de que ocurra el consecuente dado que ocurrió el antecedente
- **Elevación**
 - Refleja el **aumento de la probabilidad** de que ocurra el consecuente cuando nos enteramos de que ocurrió el antecedente

Zhang Metric

- Introducida por Zhang (2000)
- Indica qué tan **asociado** o **disociado** es una relación
- Toma valores entre -1 y +1

$$Zhang(A \rightarrow B) =$$

$$\frac{Confidence(A \rightarrow B) - Confidence(\bar{A} \rightarrow B)}{Max[Confidence(A \rightarrow B), Confidence(\bar{A} \rightarrow B)]}$$

Ejemplos relacionados

T Items

- 1 pan, leche
- 2 pan, pañales, cerveza, huevos
- 3 **leche, pañales, cerveza**, pollo
- 4 pan, **leche, pañales, cerveza**
- 5 pan, leche, pañales, pollo

$|T| = 5$

$$Support(X) = \frac{Frequency(X)}{N}$$

$\{leche, pañales\} \rightarrow \{cerveza\}$

$$Soporte = \frac{\Omega(\{leche, pañales, cerveza\})}{|T|}$$

$$Soporte = 2/5 = \mathbf{40\%}$$

El 40% de las transacciones mostraron que leche, pañales y cerveza se compraron juntos.

Ejemplos relacionados

$\{\text{leche, pañales}\} \rightarrow \{\text{cerveza}\}$

$$\text{Confianza} = \frac{\Omega(\{\text{leche, pañales, cerveza}\})}{\Omega(\{\text{leche, pañales}\})}$$

$$\text{Confianza} = \frac{2}{3} = 0.67$$

El 67% de los consumidores que compraron leche y pañales también compraron cerveza.

T Items

- 1 pan, leche
- 2 pan, pañales, cerveza, huevos
- 3 **leche, pañales, cerveza**, pollo
- 4 pan, **leche, pañales, cerveza**
- 5 pan, **leche, pañales**, pollo

$|T| = 5$

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X)}$$

Ejemplos relacionados

$\{\text{leche, pañales}\} \rightarrow \{\text{cerveza}\}$

$$\text{Lift} = \frac{\text{Confianza}(\{\text{leche, pañales}\} \rightarrow \text{cerveza})}{\Omega(\{\text{cerveza}\})}$$

$$\text{Lift} = \frac{0.67}{0.60} = 1.117$$

La probabilidad de la cerveza **incrementa de 0.6 a 0.67** cuando sabemos que el cliente compra leche y pañales.

Si el **lift es igual a 1**, entonces el **antecedente no aporta** en nada a la probabilidad del consecuente.

Si el **lift es menor a 1** significa que el **antecedente tuvo un efecto negativo** en la ocurrencia del consecuente.

T Items

- 1 pan, leche
- 2 pan, pañales, **cerveza**, huevos
- 3 **leche**, **pañales**, **cerveza**, pollo
- 4 pan, **leche**, **pañales**, **cerveza**
- 5 pan, **leche**, **pañales**, pollo

$|T| = 5$

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X)\text{Support}(Y)}$$

Algoritmo A-Priori

Fast Algorithms for Mining Association Rules

Rakesh Agrawal Ramakrishnan Srikant*

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

Abstract

We consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Experiments with synthetic as well as real-life data show that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. We also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid. Scale-up experiments show that AprioriHybrid scales linearly with the number of transactions. AprioriHybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database.

1 Introduction

Database mining is motivated by the decision support problem faced by most large retail organizations [S⁺93]. Progress in bar-code technology has made it possible for retail organizations to collect and store massive amounts of sales data, referred to as the *basket* data. A record in such data typically consists of the transaction date and the items bought in the transaction. Successful organizations view such databases as important pieces of the marketing infrastructure [Ass92].

<https://www.macs.hw.ac.uk/~dwcorne/Teaching/agrawal94fast.pdf>

- Agrawal (1994)
- Para encontrar reglas de asociación primero debemos **encontrar los itemset frecuentes**

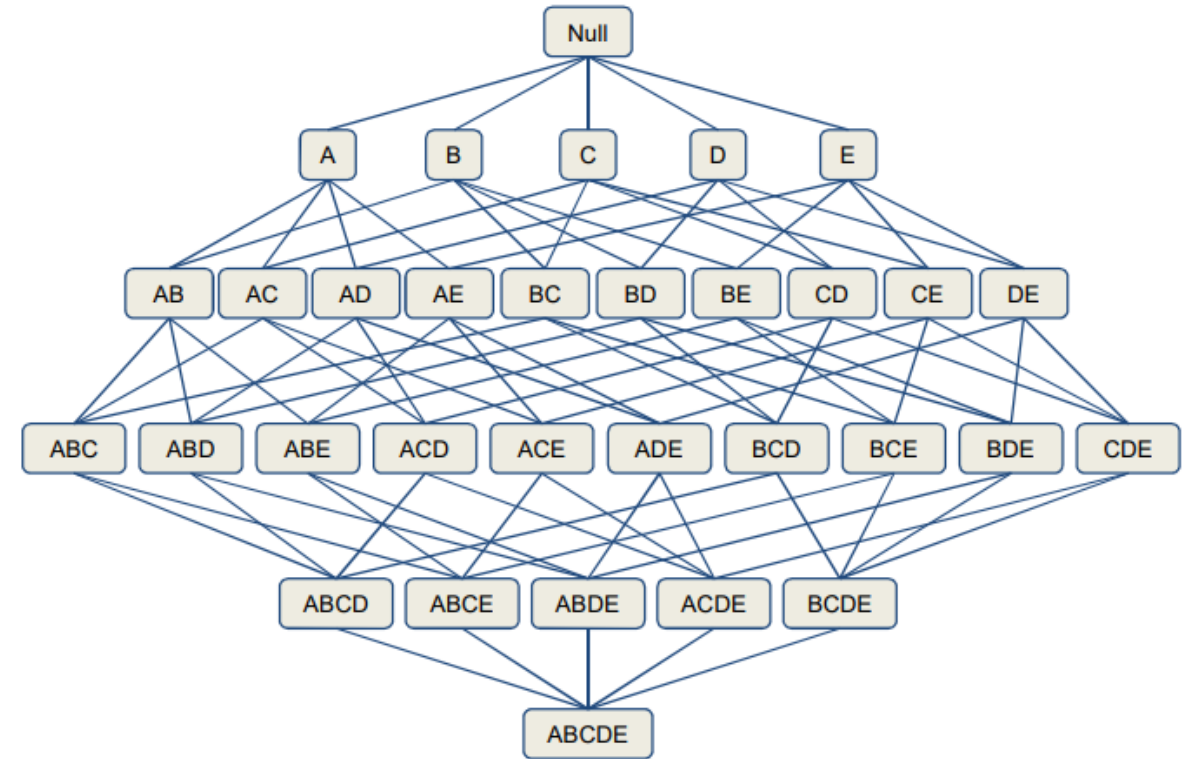
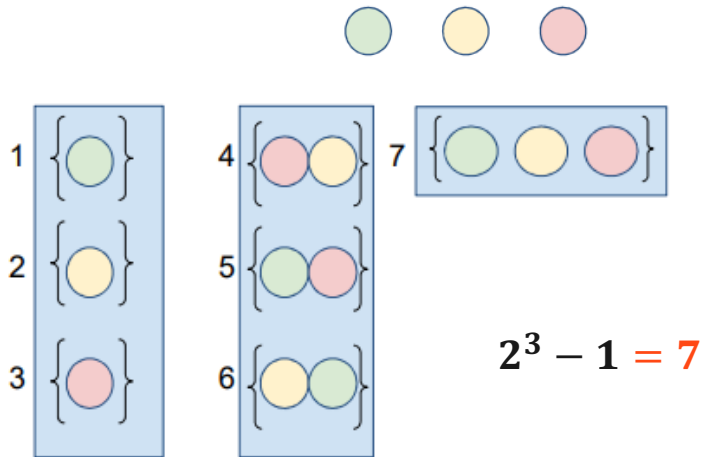
Algoritmo A-Priori

Consiste en encontrar las reglas de asociación $X \rightarrow Y$

Primera Idea:

Obtener todos los itemset posibles y para cada uno de ellos contar su frecuencia (ocurrencia) dentro de los datos.

¿Posibles Itemsets con 3 items?



$$2^5 - 1 = 31$$

La idea de generar todos los posibles itemsets **no es viable** en la práctica

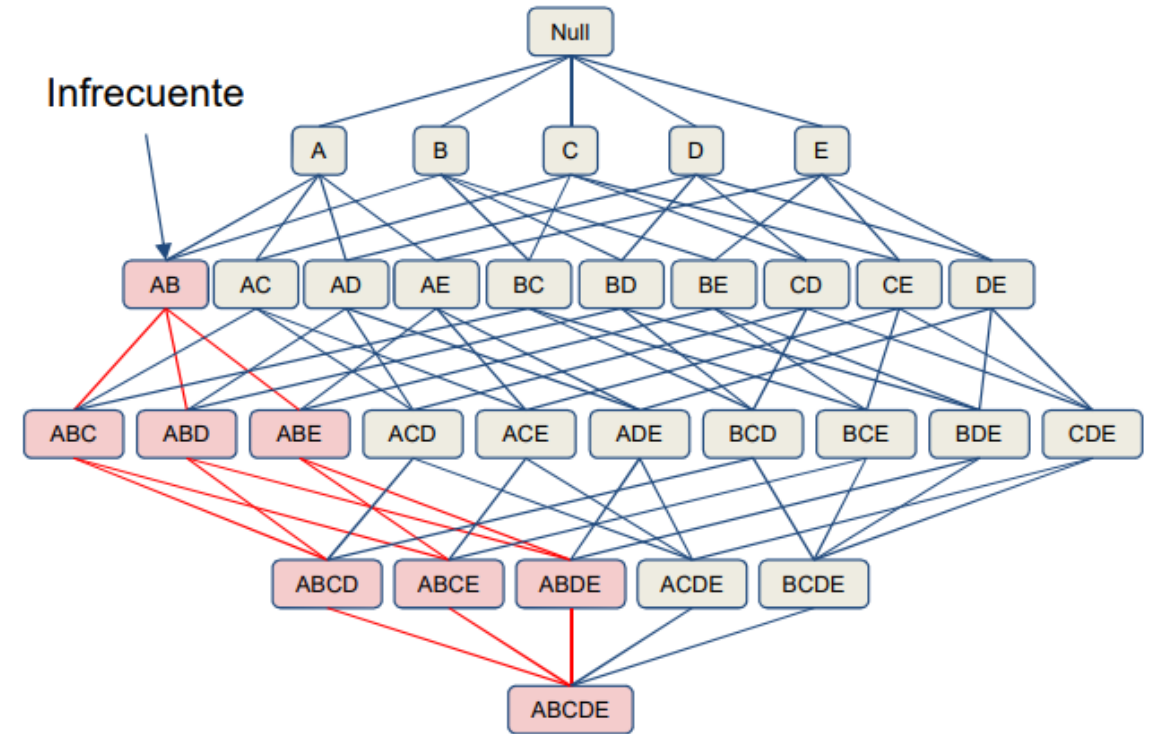
Principio de Monotonicidad

Definimos que un itemset es frecuente cuando cumple un **umbral mínimo** fijado por nosotros.

Si un itemset es **frecuente**, entonces todos los **subgrupos** de éste también son **frecuentes**.

Si un **itemset NO es frecuente**, entonces cualquier conjunto que contenga a este itemset tampoco lo será

Nos ayuda a descartar muchos itemset candidatos.



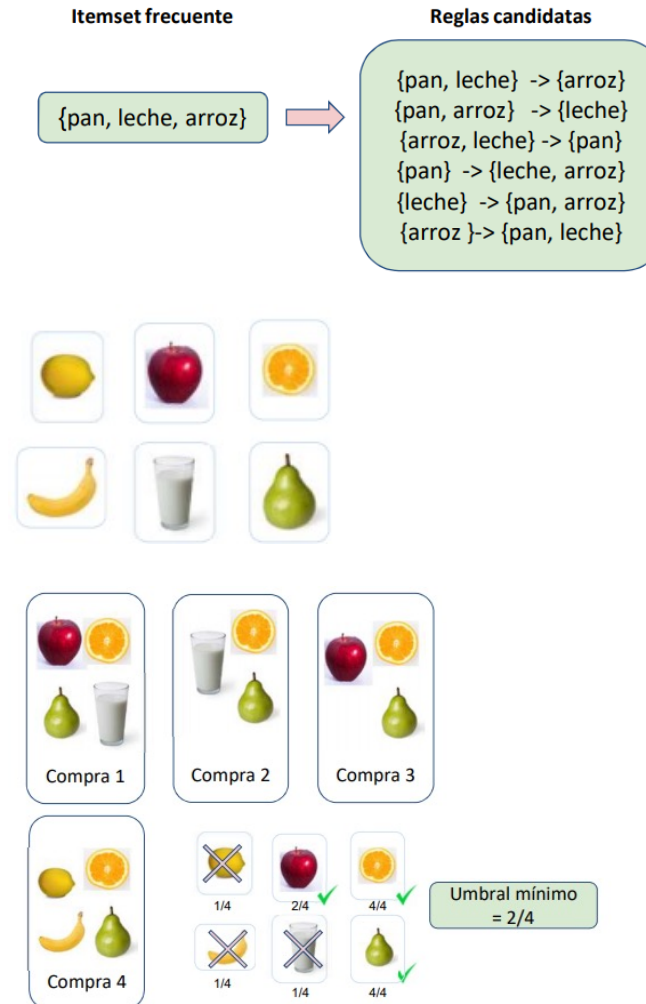
El principio de Monotonicidad nos ayuda a descartar muchos itemsets y hace posible la búsqueda

Desarrollo del Algoritmo

El algoritmo primero obtiene los **itemset frecuentes** y después calcula las reglas de asociación a partir de ellos

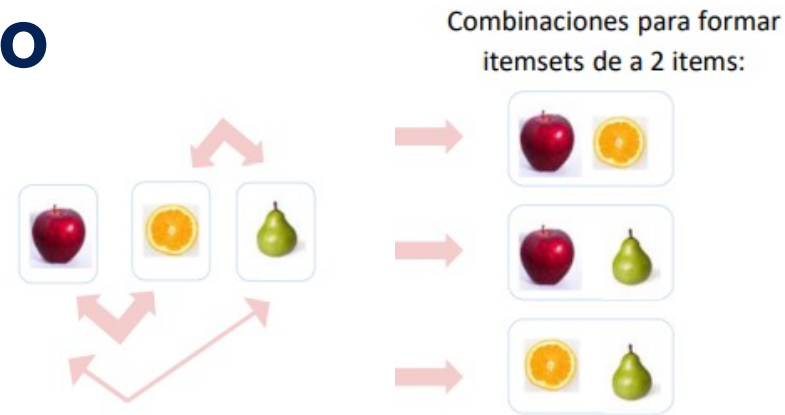
1. Al principio todos los ítems o productos son candidatos para ser un posible itemset

2. **Eliminamos** los itemset candidatos que no superan el umbral establecido.



Desarrollo del Algoritmo

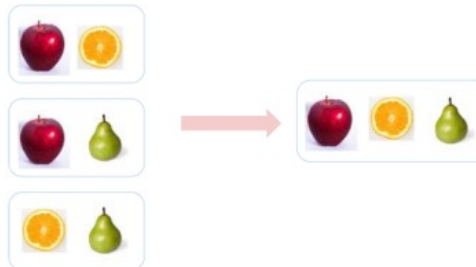
4. Una vez seleccionados los itemset de 1 ítem, formamos los itemsets candidatos que **contienen 2 ítems**.



5. Luego, **eliminamos** los itemsets candidatos que no superen el mínimo umbral



6. Luego, formamos los itemsets candidatos que **contienen 3 ítems**



7. Luego, **eliminamos** los itemsets candidatos que no superen el mínimo umbral



Ejemplo de Aplicación

Transacciones

#	Data
1	I1, I2, I4
2	I2, I4, I5
3	I1, I3
4	I1, I2, I4
5	I1, I2, I3

6	I2, I4
7	I1, I3
8	I1, I2, I4, I5
9	I1, I2, I3

Mínimo Soporte: 2/9

1. Generar el conjunto de itemsets candidatos de a un ítem

$$C_1 = \{ I1, I2, I3, I4, I5 \}$$

1. Evaluar el soporte de los candidatos y sacar los que no cumplan con el mínimo requerido

Ejemplo de Aplicación

Mínimo Soporte: 2/9

#	Data
1	I1, I2, I4
2	I2, I4, I5
3	I1, I3
4	I1, I2, I4
5	I1, I2, I3
6	I2, I4
7	I1, I3
8	I1, I2, I4, I5
9	I1, I2, I3

$$s(\{I_1\}) = \frac{7}{9} \checkmark \quad s(\{I_2\}) = \frac{7}{9} \checkmark \quad s(\{I_3\}) = \frac{4}{9} \checkmark$$

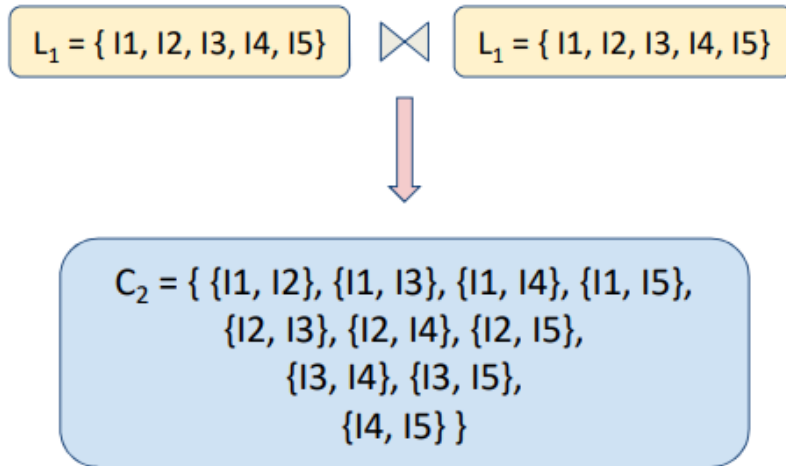
$$s(\{I_4\}) = \frac{5}{9} \checkmark \quad s(\{I_5\}) = \frac{2}{9} \checkmark$$

3. Los que cumplen el criterio pasan a ser un Itemset frecuente de la primera iteración

$$L_1 = \{I_1, I_2, I_3, I_4, I_5\}$$

3. Se generan los nuevos itemsets candidatos (tamaño 2), es decir, C_2 , a partir de L_1

Ejemplo de Aplicación



5. Los que cumplen el criterio pasan a ser un Itemset frecuente de la segunda iteración

$L_2 = \{ \{I_1, I_2\}, \{I_1, I_3\}, \{I_1, I_4\}, \{I_2, I_3\}, \{I_2, I_4\}, \{I_2, I_5\}, \{I_4, I_5\} \}$

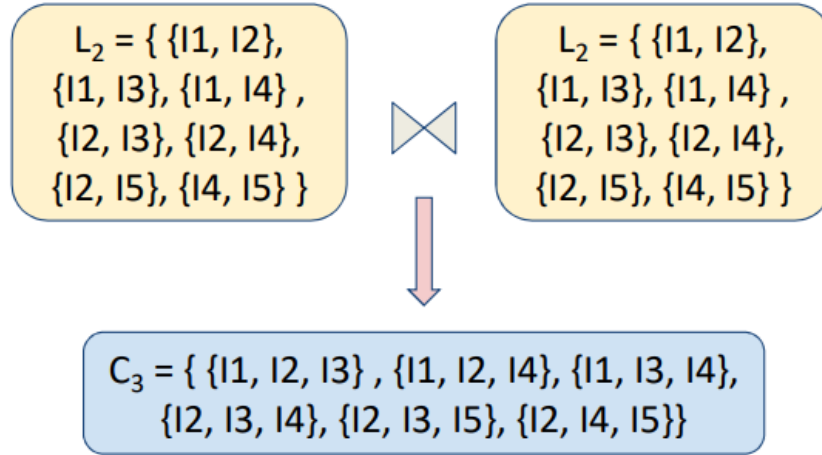
5. Se genera C_3 a partir de L_2

Mínimo Soporte: 2/9

#	Data		
		6	I2, I4
1	I1, I2, I4	7	I1, I3
2	I2, I4, I5	8	I1, I2, I4, I5
3	I1, I3	9	I1, I2, I3
4	I1, I2, I4		
5	I1, I2, I3		

$$\begin{aligned}
 s(\{I_1, I_2\}) &= \frac{5}{9} & s(\{I_1, I_3\}) &= \frac{4}{9} & s(\{I_1, I_4\}) &= \frac{3}{9} & s(\{I_1, I_5\}) &= \frac{1}{9} \\
 \checkmark & & \checkmark & & \checkmark & & \times & \\
 s(\{I_2, I_3\}) &= \frac{2}{9} & s(\{I_2, I_4\}) &= \frac{5}{9} & s(\{I_2, I_5\}) &= \frac{2}{9} \\
 \checkmark & & \checkmark & & \checkmark & & \\
 s(\{I_3, I_4\}) &= \frac{0}{9} & s(\{I_3, I_5\}) &= \frac{0}{9} \\
 \times & & \times & & \\
 s(\{I_4, I_5\}) &= \frac{2}{9} \\
 & \checkmark &
 \end{aligned}$$

Ejemplo de Aplicación



7. Los que cumplen el criterio pasan a ser un Itemset Frecuente de la segunda iteración

$L_3 = \{ \{I_1, I_2, I_3\}, \{I_1, I_2, I_4\}, \{I_2, I_4, I_5\} \}$

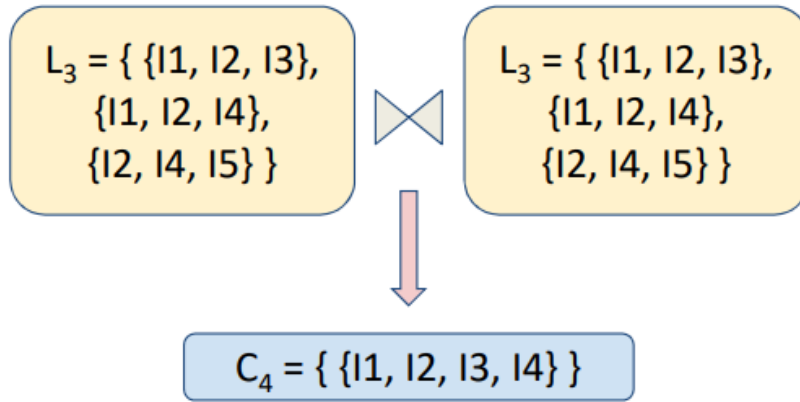
7. Se genera C_4 a partir de L_3

Mínimo Soporte: 2/9

#	Data
1	I1, I2, I4
2	I2, I4, I5
3	I1, I3
4	I1, I2, I4
5	I1, I2, I3
6	I2, I4
7	I1, I3
8	I1, I2, I4, I5
9	I1, I2, I3

$$\begin{array}{lll}
 s(\{I_1, I_2, I_3\}) = \frac{2}{9} & s(\{I_1, I_2, I_4\}) = \frac{3}{9} & s(\{I_1, I_3, I_4\}) = \frac{0}{9} \\
 \checkmark & \checkmark & \times \\
 s(\{I_2, I_3, I_4\}) = \frac{0}{9} & s(\{I_2, I_3, I_5\}) = \frac{0}{9} & s(\{I_2, I_4, I_5\}) = \frac{2}{9} \\
 \times & \times & \checkmark
 \end{array}$$

Ejemplo de Aplicación



7. Ninguno cumple el criterio de mínimo soporte por lo que L es el conjunto vacío

$L_4 = \{\}$

7. Como no hay más candidatos, dejamos de iterar

Mínimo Soporte: 2/9

#	Data		
		6	I2, I4
1	I1, I2, I4	7	I1, I3
2	I2, I4, I5	8	I1, I2, I4, I5
3	I1, I3	9	I1, I2, I3
4	I1, I2, I4		
5	I1, I2, I3		

$$s(\{I_1, I_2, I_3, I_4\}) = \frac{0}{9}$$

✗

Ejemplo de Aplicación

Los itemsets frecuentes con soporte mayor o igual 2/9

$$L_1 = \{ I1, I2, I3, I4, I5 \}$$

$$L_2 = \{ \{I1, I2\}, \{I1, I3\}, \{I1, I4\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}, \{I4, I5\} \}$$

$$L_3 = \{ \{I1, I2, I3\}, \{I1, I2, I4\}, \{I2, I4, I5\} \}$$

¡Hemos terminado!

Resumen del proceso

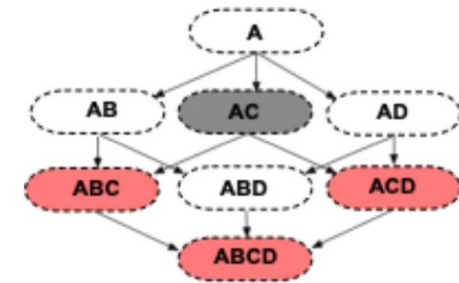
List of Lists

```
[['box'],  
 ['box'],  
 ['box'],  
 ['box'],  
 ['box'],  
 ['bag', 'box', 'sign'],  
 ['sign', 'bag', 'candle'],  
 ['bag'],  
 ['bag'],  
 ['bag'],  
 ['candle']]
```

One-Hot Encoding

	bag	box	candle	sign
0	False	True	False	False
1	False	True	False	False
2	False	True	False	False
3	False	True	False	False
4	False	True	False	False
...
14458	True	True	True	False
14459	False	True	False	True
14460	True	False	False	False
14461	True	False	False	False

Apriori Algorithm



Referencias

Curso: Introducción a la Minería de Datos – Universidad de Chile . Coursera
<https://www.coursera.org/learn/mineria-de-datos-introduccion?>

Video

Apriori Algorithm Explained - Edureka

<https://www.youtube.com/watch?v=guVvtZ7ZClw>

Video

How Netflix recommend movies?

<https://www.youtube.com/watch?v=ZspR5PZemcs&t=1601s>