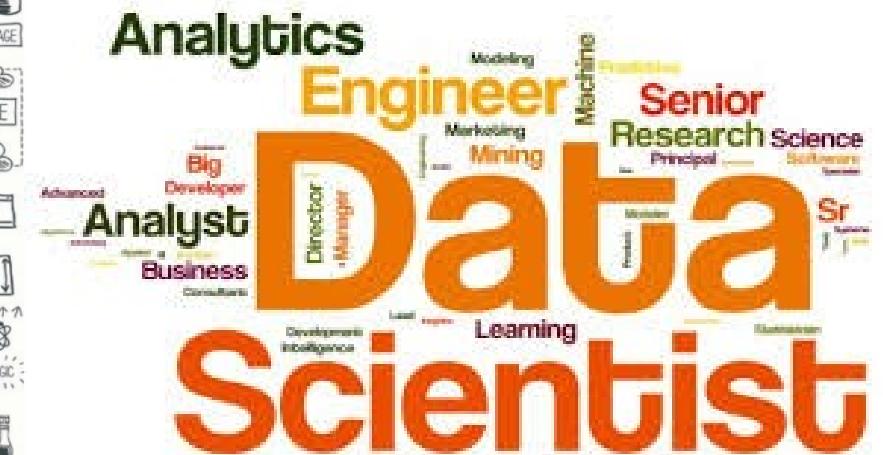
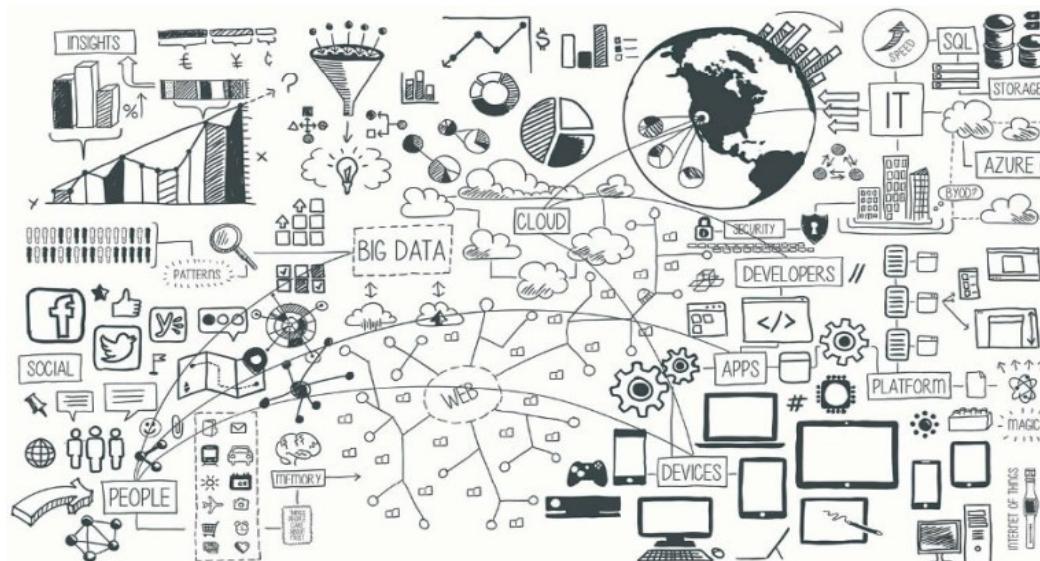


Data Mining (Minería de Datos)

Association Rules



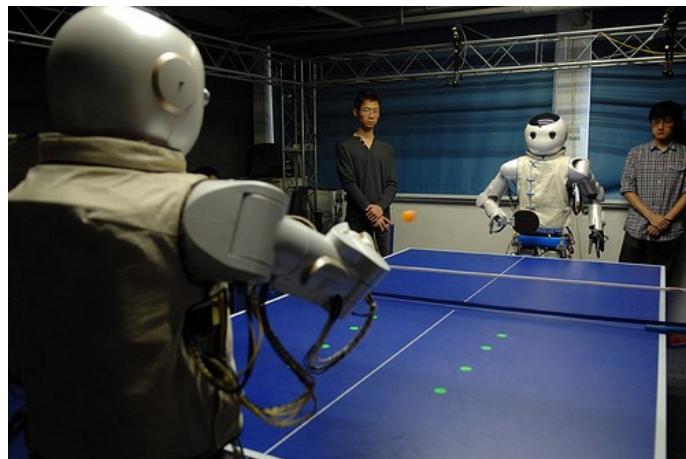
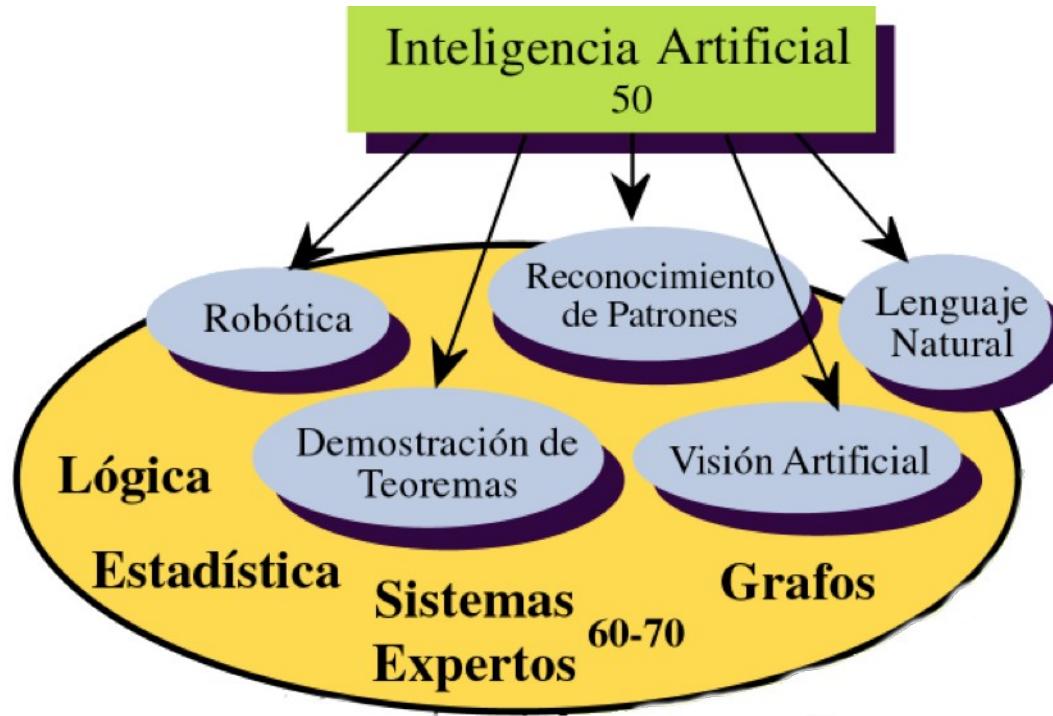
Sixto Herrera
Joaquín Bedia

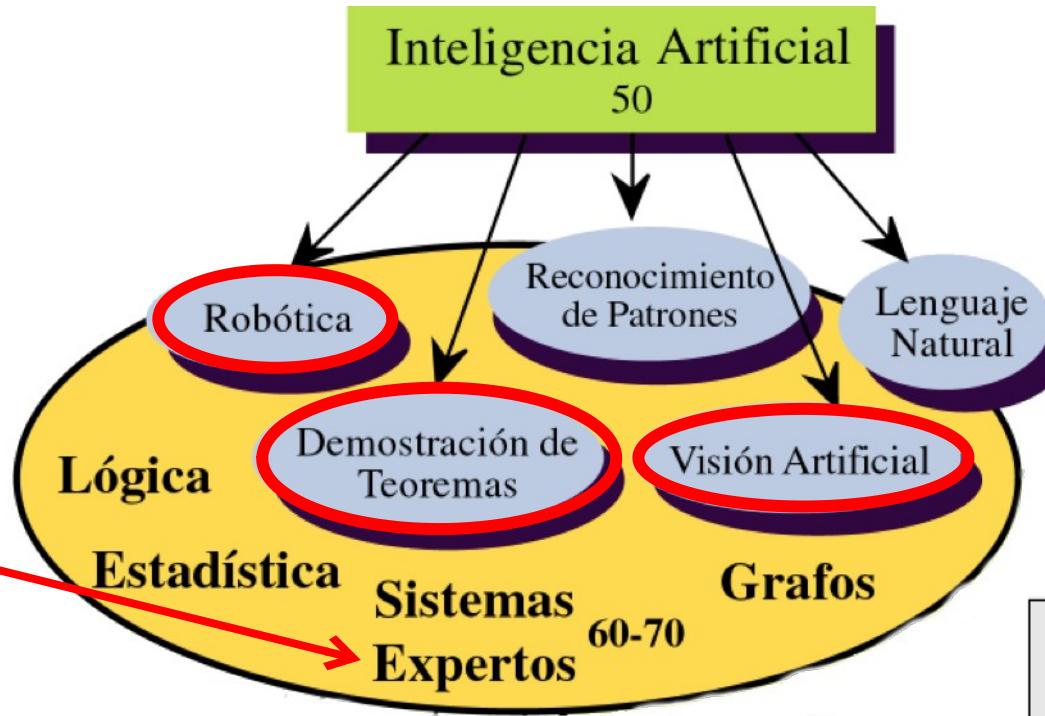
Grupo de Meteorología
Univ. de Cantabria – CSIC
MACC / IFCA



| | | |
|-----|----|---|
| Nov | 2 | Presentación, introducción y perspectiva histórica |
| | 4 | Paradigmas, problemas canonicos y data challenges |
| | 9 | Reglas de asociación |
| | 11 | Practica: Reglas de asociación |
| | 16 | Evaluación, sobrejuste y crossvalidacion |
| | 18 | Practica: Crossvalidacion |
| | 23 | Árboles de clasificacion y decision |
| | 25 | Practica: Árboles de clasificación |
| | | T01. Datos discretos |
| | 30 | Técnicas de vecinos cercano (k-NN) |
| Dic | 2 | Práctica: Vecinos cercanos |
| | 9 | Comparación de Técnicas de Clasificación. |
| | 14 | Reducción de dimensión no lineal |
| | 16 | Reducción de dimensión no lineal |
| | | T02. Clasificación |
| | 17 | Árboles de clasificación y regresion (CART) |
| | 20 | Práctica: Árboles de clasificación y regresion (CART) |
| | 21 | Practica: El paquete CARET |
| | | T03. Prediccion |
| Ene | 11 | Ensembles: Bagging and Boosting |
| | 13 | Random Forests y Gradient boosting |
| | 14 | Técnicas de agrupamiento |
| | 20 | Técnicas de agrupamiento |
| | 24 | Predicción Condicionada |
| | 26 | Sesión de refuerzo/repaso. |
| Ene | 27 | Examen |

NOTA: Las líneas de código de R en esta presentación se muestran sobre un fondo gris.





H, I cierto
L falso

G falso

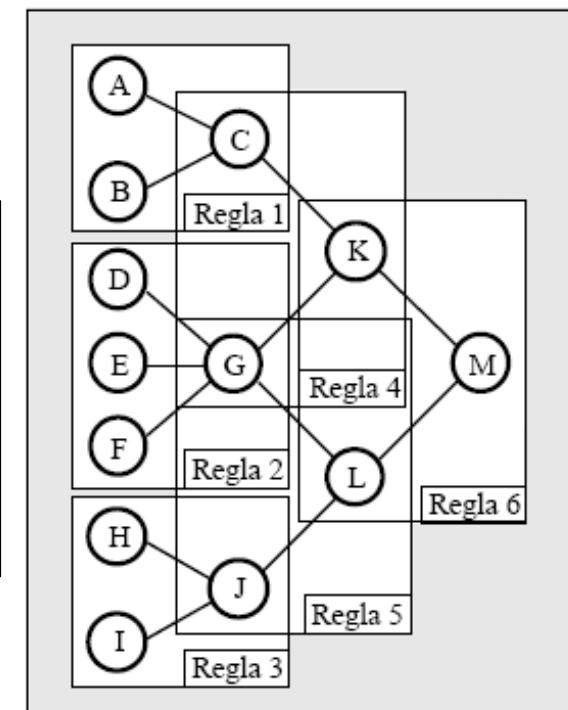
Explicit representation of knowledge
Rules, graphs, etc.

Human-like reasoning

Logical inference, look for relations on graphs, etc.

- **Modus ponens**
If P is true and $P \Rightarrow Q$ is true
then Q is true
- **Modus tolens**
if $P \Rightarrow Q$ is true and Q is false
then $\sim P$ is true

Serial processing

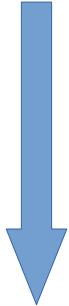




+



Si un cliente compra
pan y mantequilla ...



... entonces compra
leche el 90%



NOTA: Las líneas de código
de R en esta presentación se
muestran sobre un fondo
gris.

| | | |
|-----|----|---|
| Nov | 2 | Presentación, introducción y perspectiva histórica |
| | 4 | Paradigmas, problemas canonicos y data challenges |
| | 9 | Reglas de asociación |
| | 11 | Practica: Reglas de asociación |
| | 16 | Evaluación, sobrejuste y crossvalidacion |
| | 18 | Practica: Crossvalidacion |
| | 23 | Árboles de clasificacion y decision |
| | 25 | Practica: Árboles de clasificación |
| | | T01. Datos discretos |
| | 30 | Técnicas de vecinos cercano (k-NN) |
| Dic | 2 | Práctica: Vecinos cercanos |
| | 9 | Comparación de Técnicas de Clasificación. |
| | 14 | Reducción de dimensión no lineal |
| | 16 | Reducción de dimensión no lineal |
| | | T02. Clasificación |
| | 17 | Árboles de clasificación y regresion (CART) |
| | 20 | Práctica: Árboles de clasificación y regresion (CART) |
| | 21 | Practica: El paquete CARET |
| | | T03. Prediccion |
| Ene | 11 | Ensembles: Bagging and Boosting |
| | 13 | Random Forests y Gradient boosting |
| | 14 | Técnicas de agrupamiento |
| | 20 | Técnicas de agrupamiento |
| | 24 | Predicción Condicionada |
| | 26 | Sesión de refuerzo/repaso. |
| Ene | 27 | Examen |

¿Los registros de las ventas realizadas en el pasado son útiles para la gestión de un comercio?



1) Colocación/proposición de productos:

Si detecto productos que se compran de forma conjunta puedo planificar su colocación en el comercio.

2) Promociones y Ofertas:

Si detecto productos correlacionados puedo definir promociones que engloben a ambos (p.e. portátil + mochila, refrescos + snacks, juguetes + pilas, etc...).

3) Optimización y gestión de recursos:

Si detecto relaciones entre operaciones a realizar, puedo optimizar su planificación (p.e. en un servicio de atención al cliente si detectas problemas relacionados, puedes optimizar el servicio equipando al operario con material para resolver ambos problemas).

...

¿Los registros de las ventas realizadas en el pasado son útiles para la gestión de un comercio?



1) Colocación/proposición de productos:

Si detecto productos que se compran de forma conjunta puedo planificar su colocación en el comercio.

2) Promociones y Ofertas:

Si detecto productos correlacionados puedo definir promociones que engloben a ambos (p.e. portátil + mochila, refrescos + snacks, juguetes + pilas, etc...).

3) Optimización y gestión de recursos:

Si detecto relaciones entre operaciones a realizar, puedo optimizar su planificación (p.e. en un servicio de atención al cliente si detectas problemas relacionados, puedes optimizar el servicio equipando al operario con material para resolver ambos problemas).

...

Los algoritmos que buscan reglas de asociación analizan los datos registrados buscando relaciones entre los diferentes productos, obteniendo proposiciones del tipo: “El 90% de las compras que incluyen pan y mantequilla, compran también leche.”

¿Los registros de las ventas realizadas en el pasado son útiles para la gestión de un comercio?



1) Colocación/proposición de productos:

Si detecto productos que se compran de forma conjunta puedo planificar su colocación en el comercio.

2) Promociones y Ofertas:

Si detecto productos correlacionados puedo definir promociones que engloben a ambos (p.e. portátil + mochila, refrescos + snacks, juguetes + pilas, etc...).

**Los algoritmos que buscan reglas de asociación analizan los datos registrados buscando relaciones entre los diferentes productos, obteniendo proposiciones del tipo:
“El 90% de las compras que incluyen pan y mantequilla, compran también leche.”**

A partir de las reglas encontradas, se resuelven consultas de diferentes tipos:

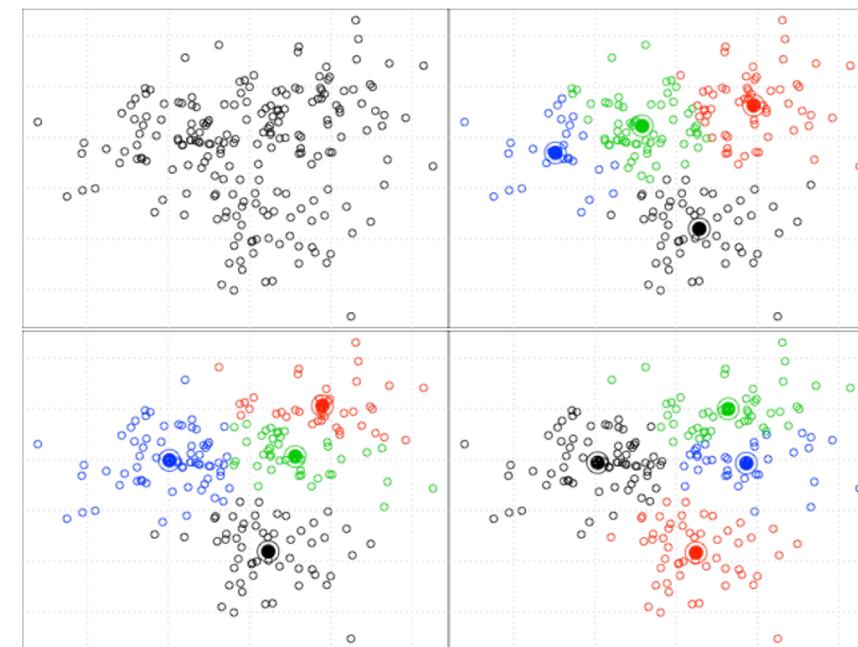
- Encontrar todas las reglas relacionando cualquier producto con uno dado (p.e. leche).
- Encontrar todas las reglas que dependen de un producto dado (p.e. pan).
- Encontrar todas las reglas que cumplen los dos criterios anteriores para ciertos productos dados.
- Encontrar el conjunto de reglas con mayor confianza cumpliendo alguno de los criterios anteriores.
- ...

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining

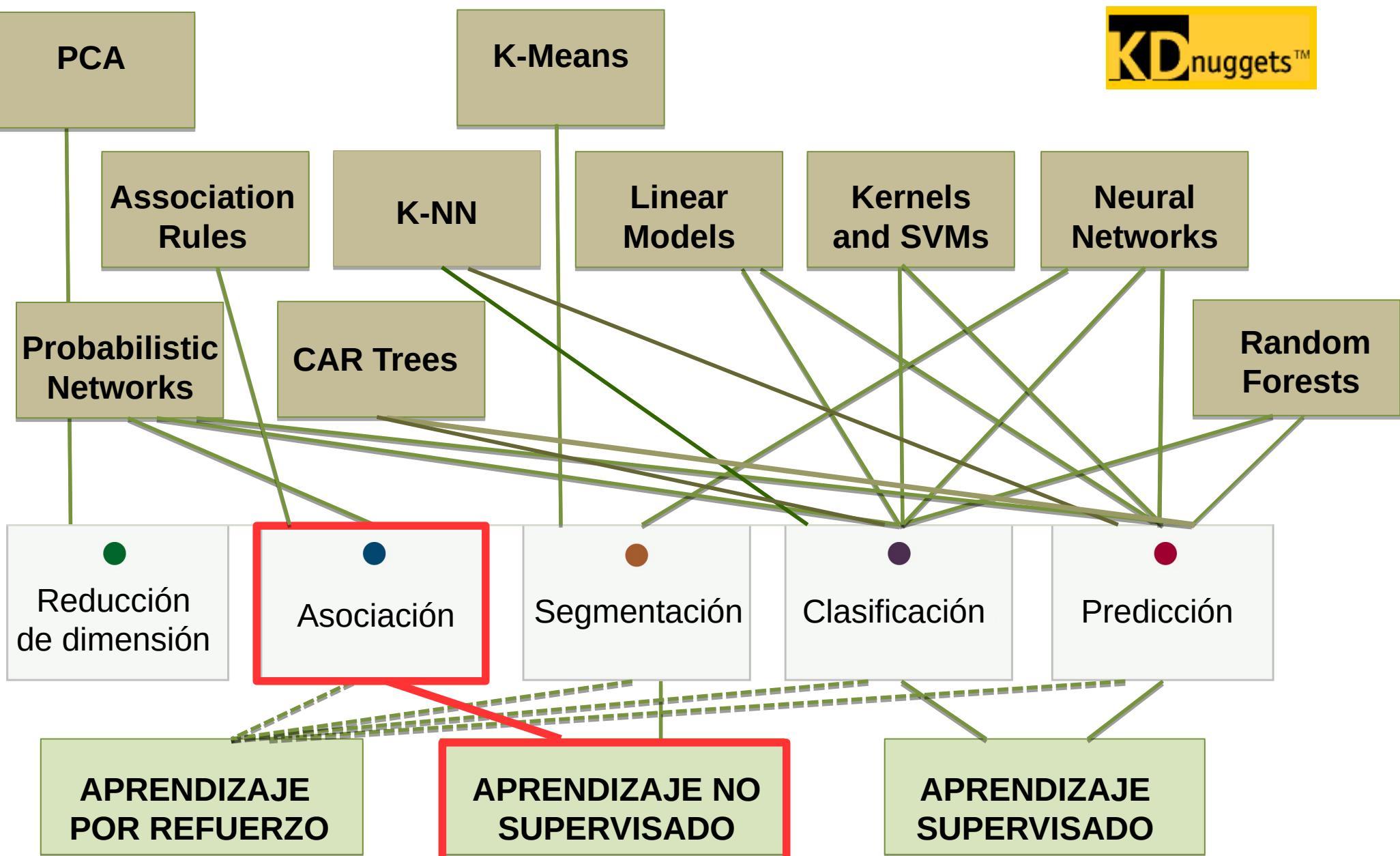
Asociación

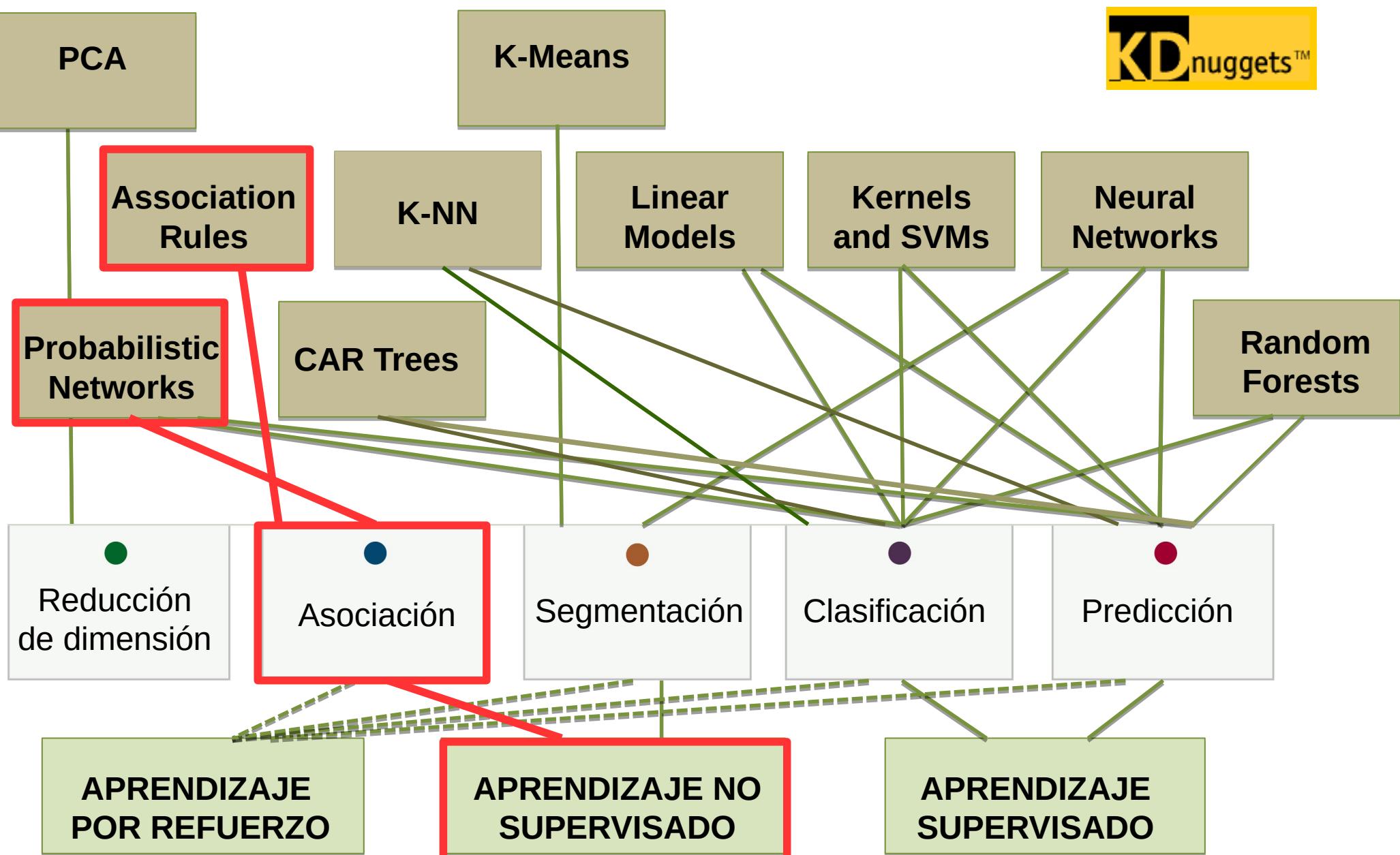
Segmentación

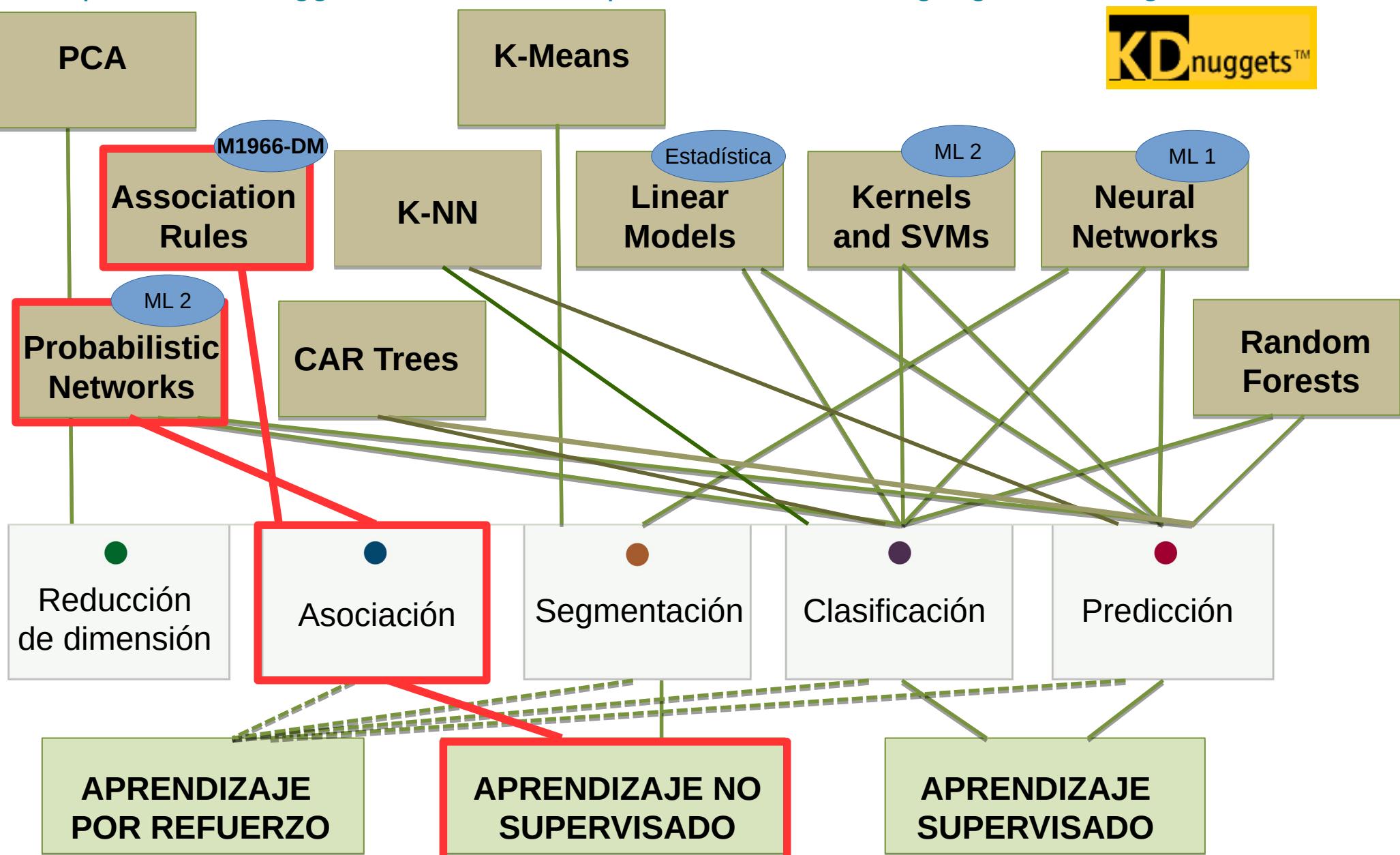
APRENDIZAJE NO
SUPERVISADO



- Target Variable: *There is no target variable (association)*
 K (*cluster*), *discrete*: #clusters (*segmentation*)
- Predictive Variables: $\{X_1, X_2, \dots, X_N\}$: *continuous or discrete*
 - “Covariates” used to make predictions.
- Predictive Model: Algorithmic, based on (X_1, X_2, \dots, X_N) .
 - Ad-hoc “learning” and “prediction” engine.







Listado de datasets utilizados en el curso

EN FUNCIÓN DE LA NATURALEZA DE LOS DATOS PODEMOS CLASIFICARLAS COMO
SÓLO CATEGÓRICAS (FACTORES)

- **Groceries.** Disponible en kaggle y en el paquete {arulesViz} de R.
- **Mushroom.** Disponible en kaggle y UCI.

MIXTOS (CONTINUOS Y FACTORES)

- **Iris.** Disponible en kaggle, UCI y el paquete {datasets} de R.
- **MNIST.** Disponible en
....
CONTINUOS
- **Meteo.** Basados en mediciones de parámetros físicos, químicos, etc...



Listado de datasets utilizados en el curso

EN FUNCIÓN DE LA NATURALEZA DE LOS DATOS PODEMOS CLASIFICARLAS COMO

SÓLO CATEGÓRICAS (FACTORES) ← Reglas de Asociación

- **Groceries.** Disponible en kaggle y en el paquete {arulesViz} de R.
- **Mushroom.** Disponible en kaggle y UCI.

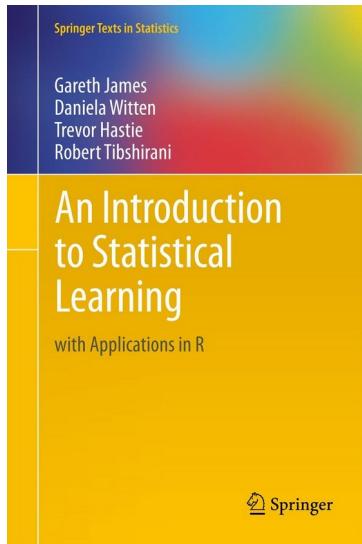
MIXTOS (CONTINUOS Y FACTORES)

- **Iris.** Disponible en kaggle, UCI y el paquete {datasets} de R.
- **MNIST.** Disponible en
....
- CONTINUOS**
- **Meteo.** Basados en mediciones de parámetros físicos, químicos, etc...



Tareas

1 30-60mins



Echa un vistazo a los datasets que hay en el paquete ISLR.

```
install.packages("ISLR")
library("ISLR")
library(help = "ISLR")
```

Analiza la estructura de los datasets: ¿de qué tipo son? ¿para qué tipo de problemas serían adecuados? e.g.

```
data("Hitters")
str(Hitters)
```

¿Hay algún dataset en este paquete que entronque en este problema? En caso de existir, ¿cuál(es) es(son)?

2 60-90mins

Lee con calma el siguiente notebook de kaggle sobre las duraciones de los trayectos de taxi en Nueva York:

<https://www.kaggle.com/headsortails/nyc-taxi-eda-update-the-fast-the-curious/notebook>

Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería Arules/ArulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat

Reglas de Asociación

- Introducción – Objetivo
- **Modelo Formal - Conceptos Básicos**
- Reglas de Asociación
- Algoritmo Apriori
- Librería Arules/ArulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat

Sean $X=(X_1, X_2, \dots, X_m)$ un conjunto de variables binarias (*items*). Sea T un conjunto de transacciones, representadas por vectores binarios t en los que $t(k)=1$ si se da la variable X_k . De este modo, decimos que la transacción t cumple X si para todas las *items* k de X se da que $t(k)=1$.

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Sean $X=(X_1, X_2, \dots, X_m)$ un conjunto de variables binarias (*items*). Sea T un conjunto de transacciones, representadas por vectores binarios t en los que $t(k)=1$ si se da la variable X_k . De este modo, decimos que la transacción t cumple X si para todas las items k de X se da que $t(k)=1$.

Definimos una **regla de asociación** como una implicación $X \rightarrow Y$, donde X es un conjunto de items e Y es un conjunto de items no incluidos en X .

{Limón, Naranja} → Galletas

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Sean $X=(X_1, X_2, \dots, X_m)$ un conjunto de variables binarias (**items**). Sea T un conjunto de transacciones, representadas por vectores binarios t en los que $t(k)=1$ si se da la variable X_k . De este modo, decimos que la transacción t cumple X si para todas las items k de X se da que $t(k)=1$.

Definimos una **regla de asociación** como una implicación $X \rightarrow Y$, donde X es un conjunto de items e Y es un conjunto de items no incluidos en X .

{Limón, Naranja} → Galletas

Dada una regla de asociación, se define el **factor de confianza** $0 < c < 1$ como el ratio entre las transacciones en T que cumplen X e Y y las transacciones que cumplen X .

¿Cuál es la confianza de la regla: **{Limón, Naranja} → Galletas?**

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Sean $X=(X_1, X_2, \dots, X_m)$ un conjunto de variables binarias (**items**). Sea T un conjunto de transacciones, representadas por vectores binarios t en los que $t(k)=1$ si se da la variable X_k . De este modo, decimos que la transacción t cumple X si para todas las items k de X se da que $t(k)=1$.

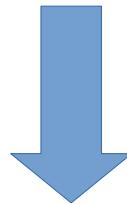
Por lo tanto, dado un conjunto de transacciones T , estamos interesados en generar todas las reglas que cumplan ciertas condiciones que podemos clasificar en dos tipos:

1. Sintácticas: se refieren a buscar reglas que involucren determinados items específicos, bien sea en el conjunto X , en el Y o en ambos.

2. Frecuencia/relevancia: se refieren a buscar reglas que se den en un número significativo de casos dentro del conjunto total. Habitualmente, estaremos interesados en reglas cuya relevancia sea superior a un umbral ya que serán aquellas con mayor impacto. De este modo, una relevancia baja implica que la regla no es muy importante y que puede, de ser necesario, ser considerada en etapas posteriores.

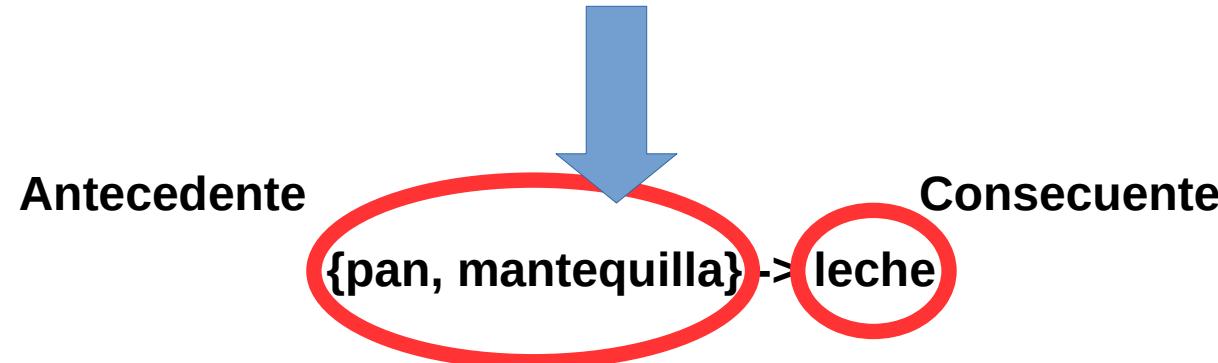
| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

“El 90% de las compras que incluyen pan y mantequilla, compran también leche.”



{pan, mantequilla} -> leche

“El 90% de las compras que incluyen pan y mantequilla, compran también leche.”



“En 90% de las compras que incluyen pan y mantequilla, compran también leche.”

Confianza (C)

Antecedente

{pan, mantequilla} ->

Consecuente

leche



$$C(\{pan, mantequilla\} \rightarrow leche) = S(\{pan, mantequilla, leche\}) / S(\{pan, mantequilla\})$$

El soporte del conjunto X, $S(X)$, es el número de observaciones en las que se da dicho evento

“En 90% de las compras que incluyen pan y mantequilla, compran también leche.”

Confianza (C)

Antecedente

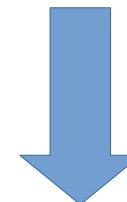
{pan, mantequilla} -> leche

Consecuente



$$C(\{\text{pan, mantequilla}\} \rightarrow \text{leche}) = S(\{\text{pan, mantequilla, leche}\}) / S(\{\text{pan, mantequilla}\})$$

El soporte del conjunto X, $S(X)$, es el número de observaciones en las que se da dicho evento. El soporte de una regla viene dada por el soporte del conjunto $\{X, Y\}$, $S(X, Y)$.



De cada 10 veces que se ha comprado pan y mantequilla en la tienda, 9 se ha comprado también leche.

“En 90% de las compras que incluyen pan y mantequilla, compran también leche.”

Confianza (C)

Antecedente

{pan, mantequilla} -> leche

Consecuente



$$C(\{\text{pan, mantequilla}\} \rightarrow \text{leche}) = S(\{\text{pan, mantequilla, leche}\}) / S(\{\text{pan, mantequilla}\})$$

El soporte del conjunto X, $S(X)$, es el número de observaciones en las que se da dicho evento. El soporte de una regla viene dada por el soporte del conjunto $\{X, Y\}$, $S(X, Y)$.



De cada 10 veces que se ha comprado pan y mantequilla en la tienda, 9 se ha comprado también leche.

Nota: El factor de confianza es una medida del peso de la regla mientras que el soporte/relevancia se corresponde con la significancia estadística.

Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- **Reglas de Asociación**
- Algoritmo Apriori
- Librería Arules/ArulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat

A partir de las definiciones de soporte y confianza de una regla surge, de forma natural, el siguiente algoritmo:

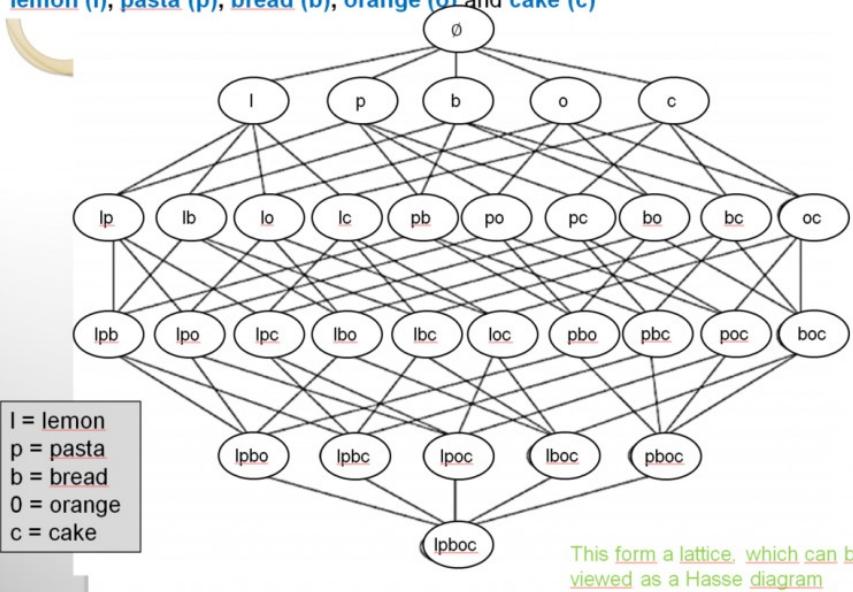
- 1) Definir umbrales ***MinSupp*** y ***MinConf*** para el soporte y la confianza de la regla.
- 2) Enumerar todas las reglas de asociación posibles.
- 3) Calcular el soporte y la confianza de cada regla.
- 4) Eliminar las reglas que no superen los umbrales establecidos. Es decir, $C(X \rightarrow Y) < MinConf$ ó $S(X, Y) < MinSupp$.

A partir de las definiciones de soporte y confianza de una regla surge, de forma natural, el siguiente algoritmo:

- 1) Definir umbrales ***MinSupp*** y ***MinConf*** para el soporte y la confianza de la regla.
- 2) Enumerar todas las reglas de asociación posibles.
- 3) Calcular el soporte y la confianza de cada regla.
- 4) Eliminar las reglas que no superen los umbrales establecidos. Es decir, $C(X \rightarrow Y) < MinConf$ ó $S(X, Y) < MinSupp$.

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



5 items → $32 = 2^5$ itemsets → $3^5 - 2^6 + 1$ reglas.



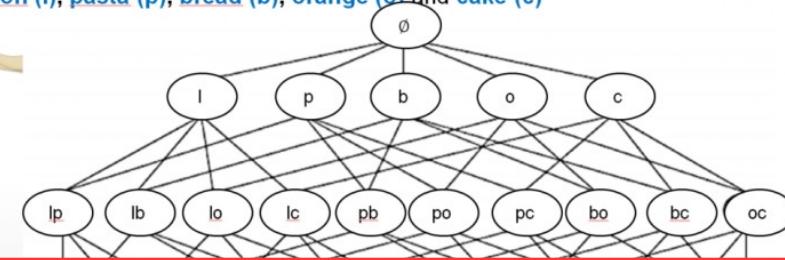
!!!!Computacionalmente Inviable!!!!

A partir de las definiciones de soporte y confianza de una regla surge, de forma natural, el siguiente algoritmo:

- 1) Definir umbrales ***MinSupp*** y ***MinConf*** para el soporte y la confianza de la regla.
- 2) Enumerar todas las reglas de asociación posibles.
- 3) Calcular el soporte y la confianza de cada regla.
- 4) Eliminar las reglas que no superen los umbrales establecidos. Es decir, $C(X \rightarrow Y) < MinConf$ ó $S(X, Y) < MinSupp$.

Search space

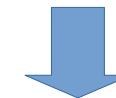
This is all the itemsets that can be formed with the items
lemon (l), pasta (p), bread (b), orange (o) and cake (c)



5 items → $32 = 2^5$ itemsets → $3^5 - 2^6 + 1$ reglas.



!!!!Computacionalmente Inviable!!!!



- 1) Reducir el **número de candidatos** → Técnicas de **poda**.
- 2) Reducir el **número de transacciones** conforme aumenta el tamaño del itemset.
- 3) Estructurar los datos de forma eficiente para almacenar los candidatos o las transacciones, de forma que se **reduzcan las comparaciones**.

Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería Arules/ArulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat

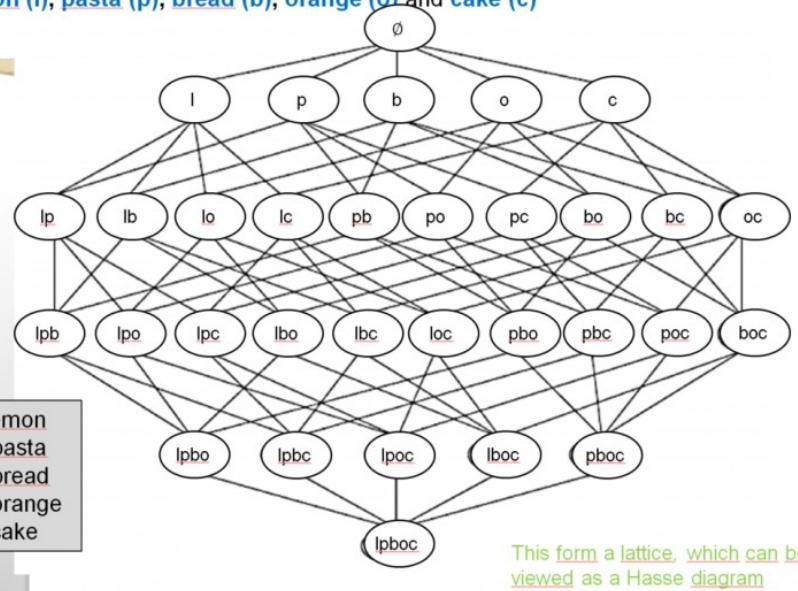
El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



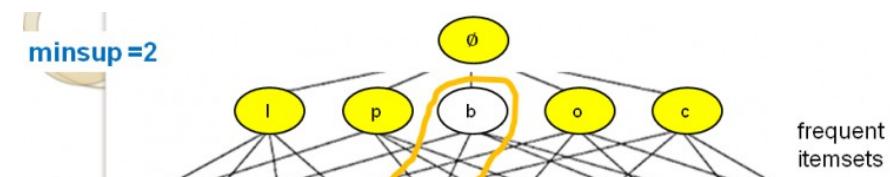
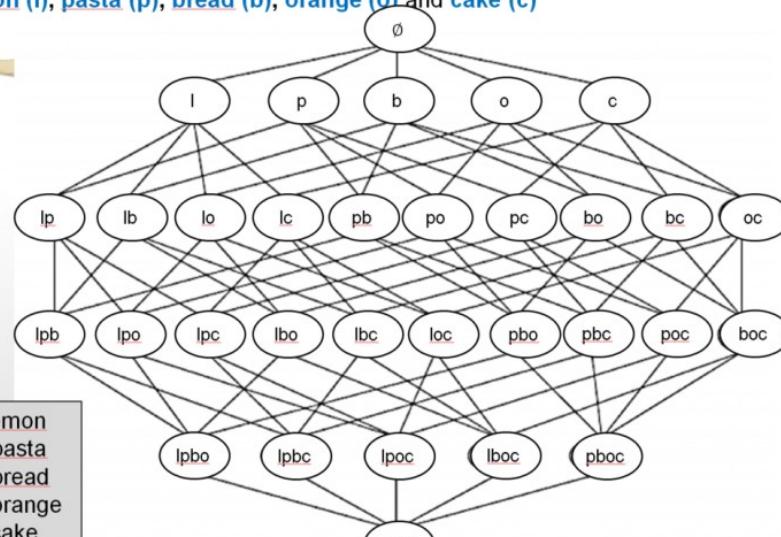
El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y: (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



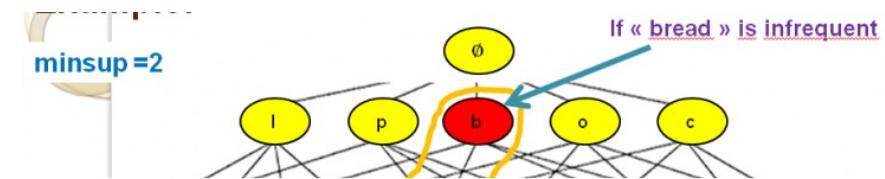
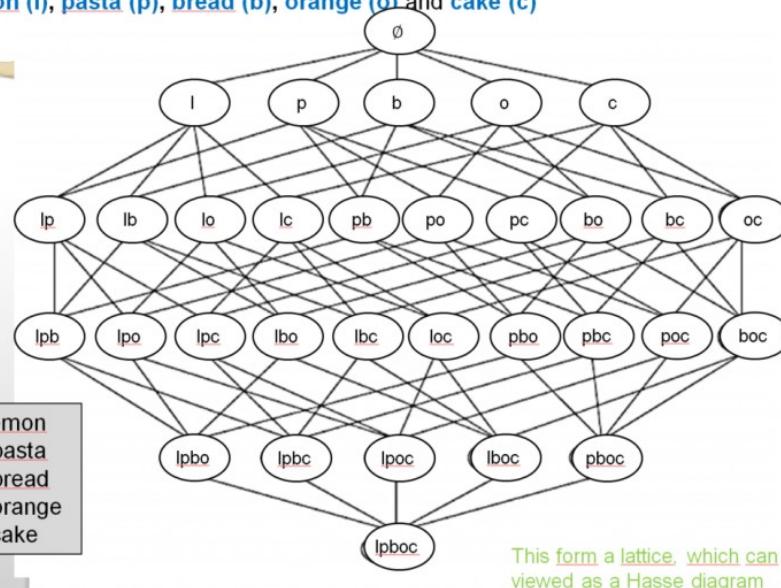
El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.

Search space

This is all the itemsets that can be formed with the items
 lemon (l), pasta (p), bread (b), orange (o) and cake (c)



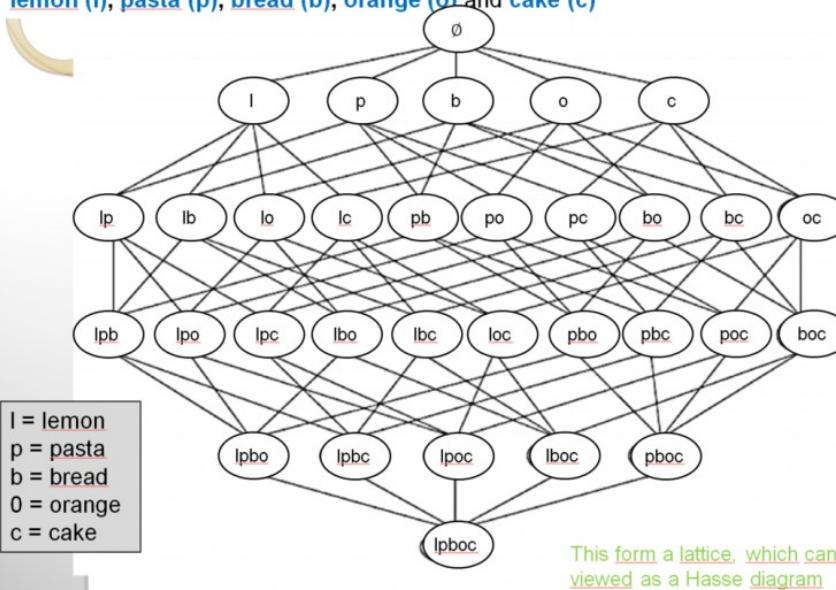
El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

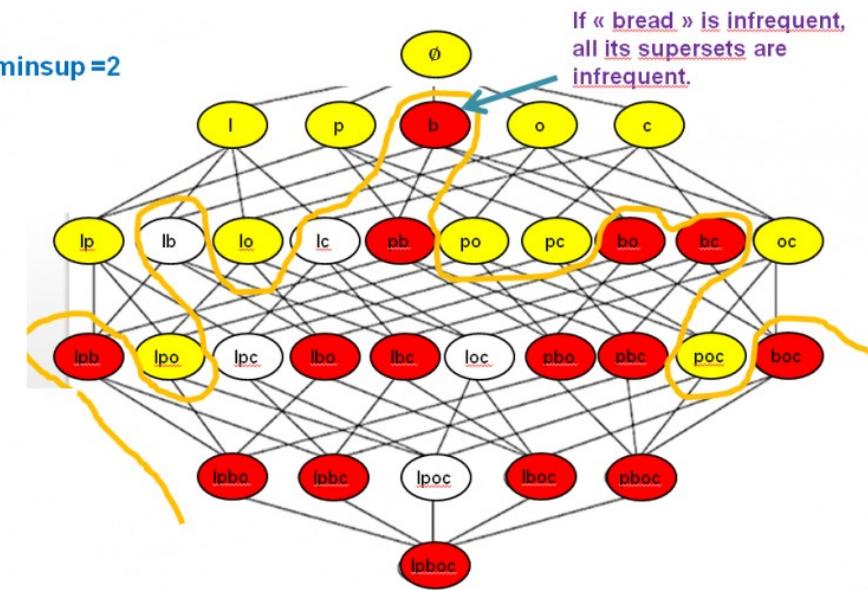
Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.
Los itemsets derivados de itemsets poco frecuentes no se incluyen en el siguiente paso.

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



minsup = 2



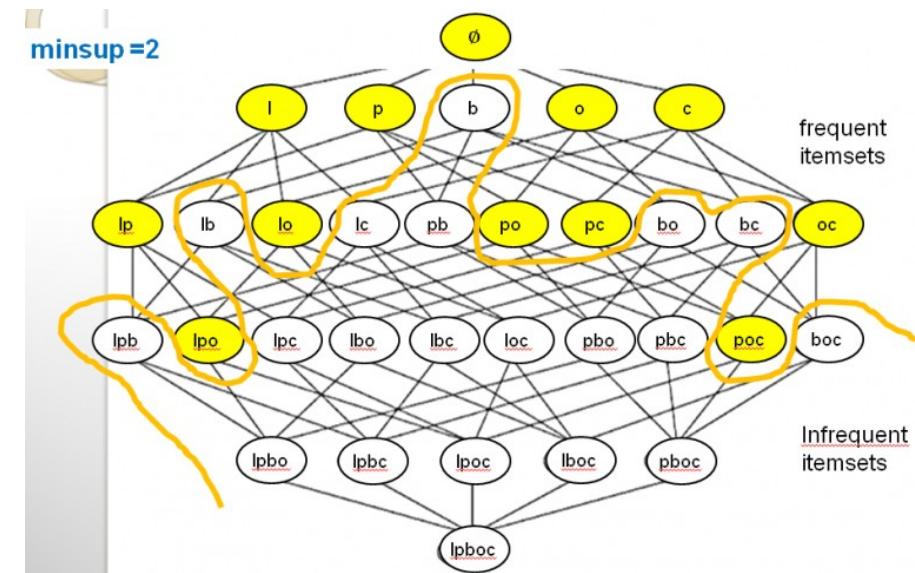
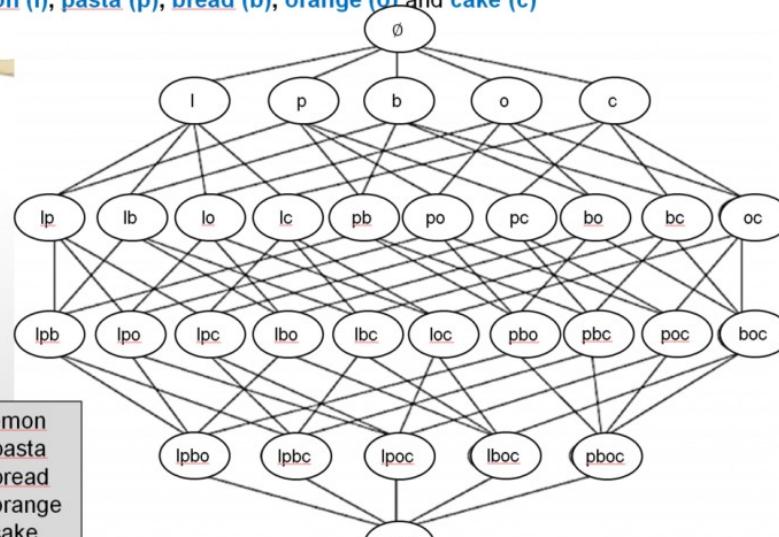
El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.
Los itemsets derivados de itemsets poco frecuentes no se incluyen en el siguiente paso.

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

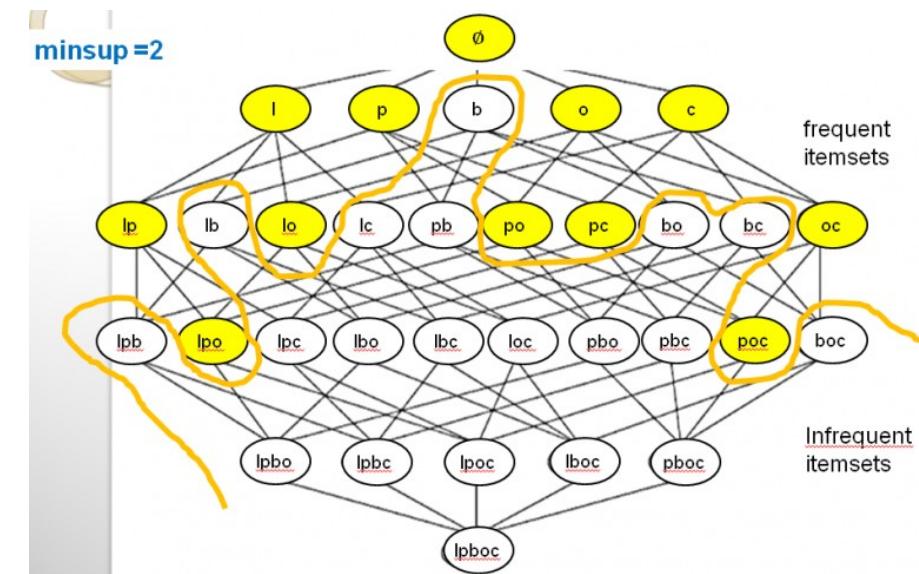
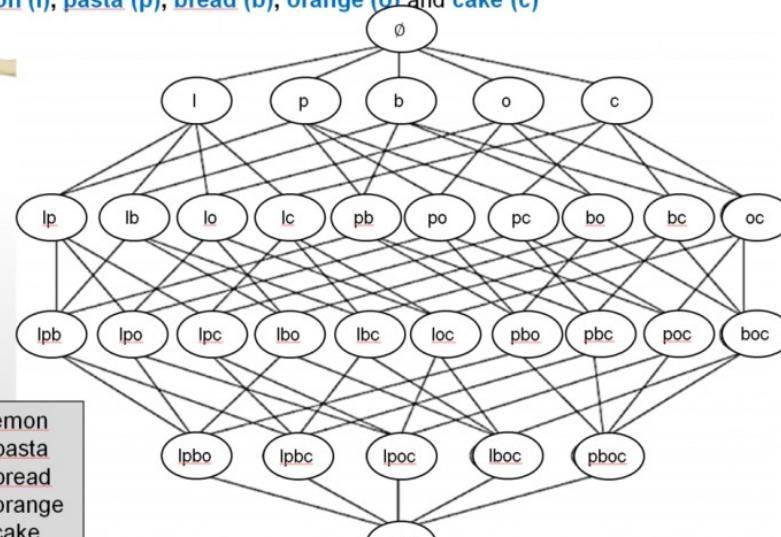
$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.
Del mismo modo,

$$\text{si } \exists X \subseteq Y : X \text{ no frecuente} \Rightarrow Y \text{ no frecuente}$$

Search space

This is all the itemsets that can be formed with the items
lemon (l), **pasta (p)**, **bread (b)**, **orange (o)** and **cake (c)**



El algoritmo **APRIORI** se enmarca dentro de las técnicas que buscar la reducción del número de candidatos a través de la “**poda**” de algunas de las ramas del “**espacio de búsqueda**”. Se fundamenta en la siguiente propiedad (**anti-monotonía del soporte**):

$$\forall X, Y : (X \subseteq Y) \Rightarrow S(X) \geq S(Y)$$

Es decir, si un itemset es frecuente, también lo son todos sus subconjuntos ya que el soporte de un itemset nunca puede ser superior al de cualquier subconjunto.

Del mismo modo,

$$si \exists X \subseteq Y : X \text{ no frecuente} \Rightarrow Y \text{ no frecuente}$$

Si bien, en cierto modo es una reformulación de la propiedad anterior, da lugar a uno de los pasos del algoritmo **APRIORI** y, por lo tanto, nos permitirá entender éste mejor. Dada una relevancia mínima **MinSupp**:

1. $i = 1$ (tamaño de los conjuntos)
2. Generar un conjunto unitario en **S1** para cada atributo.
3. Comprobar la relevancia de todos los conjuntos en **Si**, descartando aquellos tales que:
 $S(X) < MinSupp$
4. Combinar los conjuntos en **Si** creando conjuntos de tamaño $i+1$ en **Si+1**.
5. Si **Si** no es vacío entonces $i = i + 1$. Ir a 3.
6. Si no, devolver $\{S1, S2, \dots, Si\}$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Definimos la cota para el soporte/relevancia: $\text{MinSupp} = 2$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Definimos la cota para el soporte/relevancia: $\text{MinSupp} = 2$

1.- Subconjuntos de tamaño $i=1$: {I1, I2, I3, I4, I5}

2.- Soporte/relevancia: {S(I1) = 4; S(I2) = 3; S(I3) = 3; S(I4) = 1; S(I5) = 2}

¿Todos los subconjuntos cumplen la condición del soporte?

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Definimos la cota para el soporte/relevancia: $\text{MinSupp} = 2$

1.- Subconjuntos de tamaño $i=1$: {I1, I2, I3, I4, I5}

2.- Soporte/relevancia: {S(I1) = 4; S(I2) = 3; S(I3) = 3; S(I4) = 1; S(I5) = 2}

¿Todos los subconjuntos cumplen la condición del soporte? **NO**

3.- Aplicamos el MinSupp : $S1 = \{I1; I2; I3; I5\} \rightarrow$ Se excluye el I4-Pan y derivados.

4.- Construimos subconjuntos de tamaño $i=2$ a partir de $S1$: {I1I2; I1I3; I1I5; I2I3; I2I5; I3I5}

5.- Aplicamos la segunda propiedad, ¿Se excluye algún subconjunto de $S1$?

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Definimos la cota para el soporte/relevancia: $\text{MinSupp} = 2$

1.- Subconjuntos de tamaño $i=1$: {I1, I2, I3, I4, I5}

2.- Soporte/relevancia: {S(I1) = 4; S(I2) = 3; S(I3) = 3; S(I4) = 1; S(I5) = 2}

¿Todos los subconjuntos cumplen la condición del soporte? NO

3.- Aplicamos el MinSupp : $S1 = \{I1; I2; I3; I5\} \rightarrow$ Se excluye el I4-Pan y derivados.

4.- Construimos subconjuntos de tamaño $i=2$ a partir de $S1$: {I1I2; I1I3; I1I5; I2I3; I2I5; I3I5}

5.- Calculamos su soporte/relevancia: {I1I2 = 3; I1I3 = 3; I1I5 = 2; I2I3 = 2; I2I5 = 1; I3I5 = 2}

6.- Aplicamos la segunda propiedad, ¿Se excluye algún subconjunto? SI

7.- Aplicamos el MinSupp : $S2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\} \rightarrow$ Se excluye el I2I5 y derivados.

8.- Construimos subconjuntos de tamaño $i=3$ a partir de $S2$: {I1I2I3; I1I2I5; I1I3I5; I2I3I5}

- 1.- Subconjuntos de tamaño $i=1$: $\{I_1, I_2, I_3, I_4, I_5\}$
- 2.- Soporte/relevancia: $\{S(I_1) = 4; S(I_2) = 3; S(I_3) = 3; S(I_4) = 1; S(I_5) = 2\}$
- 3.- Aplicamos el ***MinSupp***: $S_1 = \{I_1; I_2; I_3; I_5\} \rightarrow$ Se excluye el I_4 -Pan y derivados.
- 4.- Construimos subconjuntos de tamaño $i=2$ a partir de S_1 : $\{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_2I_5; I_3I_5\}$
- 5.- Aplicamos la segunda propiedad. Para $i=2$ no se excluye ningún subconjunto.
- 6.- Calculamos su soporte/relevancia: $\{I_1I_2 = 3; I_1I_3 = 3; I_1I_5 = 2; I_2I_3 = 2; I_2I_5 = 1; I_3I_5 = 2\}$
- 7.- Aplicamos el ***MinSupp***: $S_2 = \{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_3I_5\} \rightarrow$ Se excluye el I_2I_5 y derivados.
- 8.- Construimos subconjuntos de tamaño $i=3$ a partir de S_2 : $\boxed{\{I_1I_2I_3; I_1I_2I_5; I_1I_3I_5; I_2I_3I_5\}}$

Se considera un orden de los items para realizar la unión sin incurrir en la repetición de conjuntos. De este modo, únicamente se combinan items que en la ordenación estén en posiciones superiores a la de los items de los itemsets considerados.

- 1.- Subconjuntos de tamaño $i=1$: $\{I_1, I_2, I_3, I_4, I_5\}$
- 2.- Soporte/relevancia: $\{S(I_1) = 4; S(I_2) = 3; S(I_3) = 3; S(I_4) = 1; S(I_5) = 2\}$
- 3.- Aplicamos el ***MinSupp***: $S_1 = \{I_1; I_2; I_3; I_5\} \rightarrow$ Se excluye el I_4 -Pan y derivados.
- 4.- Construimos subconjuntos de tamaño $i=2$ a partir de S_1 : $\{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_2I_5; I_3I_5\}$
- 5.- Aplicamos la segunda propiedad. Para $i=2$ no se excluye ningún subconjunto.
- 6.- Calculamos su soporte/relevancia: $\{I_1I_2 = 3; I_1I_3 = 3; I_1I_5 = 2; I_2I_3 = 2; I_2I_5 = 1; I_3I_5 = 2\}$
- 7.- Aplicamos el ***MinSupp***: $S_2 = \{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_3I_5\} \rightarrow$ Se excluye el I_2I_5 y derivados.
- 8.- Construimos subconjuntos de tamaño $i=3$ a partir de S_2 : $\{I_1I_2I_3; I_1I_2I_5; I_1I_3I_5; I_2I_3I_5\}$
- 9.- Aplicamos la segunda propiedad, eliminando $\{I_1I_2I_5; I_2I_3I_5\}$.
- 10.- Calculamos su soporte/relevancia: $\{I_1I_2I_3 = 2; I_1I_3I_5 = 2\}$
- 11.- Aplicamos el ***MinSupp***: $S_3 = \{I_1I_2I_3; I_1I_3I_5\}$
- 12.- Construimos subconjuntos de tamaño $i=4$ a partir de S_3 : $\{I_1I_2I_3I_5\}$
- 13.- Aplicamos la segunda propiedad, eliminando $\{I_1I_2I_3I_5\}$ al contener un subconjunto de orden 3 no frecuente ($I_1I_2I_5$ ó $I_2I_3I_5$).

- 1.- Subconjuntos de tamaño $i=1$: $\{I_1, I_2, I_3, I_4, I_5\}$
- 2.- Soporte/relevancia: $\{S(I_1) = 4; S(I_2) = 3; S(I_3) = 3; S(I_4) = 1; S(I_5) = 2\}$
- 3.- Aplicamos el ***MinSupp***: $S_1 = \{I_1; I_2; I_3; I_5\} \rightarrow$ Se excluye el I_4 -Pan y derivados.
- 4.- Construimos subconjuntos de tamaño $i=2$ a partir de S_1 : $\{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_2I_5; I_3I_5\}$
- 5.- Aplicamos la segunda propiedad. Para $i=2$ no se excluye ningún subconjunto.
- 6.- Calculamos su soporte/relevancia: $\{I_1I_2 = 3; I_1I_3 = 3; I_1I_5 = 2; I_2I_3 = 2; I_2I_5 = 1; I_3I_5 = 2\}$
- 7.- Aplicamos el ***MinSupp***: $S_2 = \{I_1I_2; I_1I_3; I_1I_5; I_2I_3; I_3I_5\} \rightarrow$ Se excluye el I_2I_5 y derivados.
- 8.- Construimos subconjuntos de tamaño $i=3$ a partir de S_2 : $\{I_1I_2I_3; I_1I_2I_5; I_1I_3I_5; I_2I_3I_5\}$
- 9.- Aplicamos la segunda propiedad, eliminando $\{I_1I_2I_5; I_2I_3I_5\}$.
- 10.- Calculamos su soporte/relevancia: $\{I_1I_2I_3 = 2; I_1I_3I_5 = 2\}$
- 11.- Aplicamos el ***MinSupp***: $S_3 = \{I_1I_2I_3; I_1I_3I_5\}$
- 12.- Construimos subconjuntos de tamaño $i=4$ a partir de S_3 : $\{I_1I_2I_3I_5\}$
- 13.- Aplicamos la segunda propiedad, eliminando $\{I_1I_2I_3I_5\}$ al contener un subconjunto de orden 3 no frecuente ($I_1I_2I_5$ ó $I_2I_3I_5$).

¡¡¡Se evalúan 13 soportes en lugar de los 32 posibles!!!

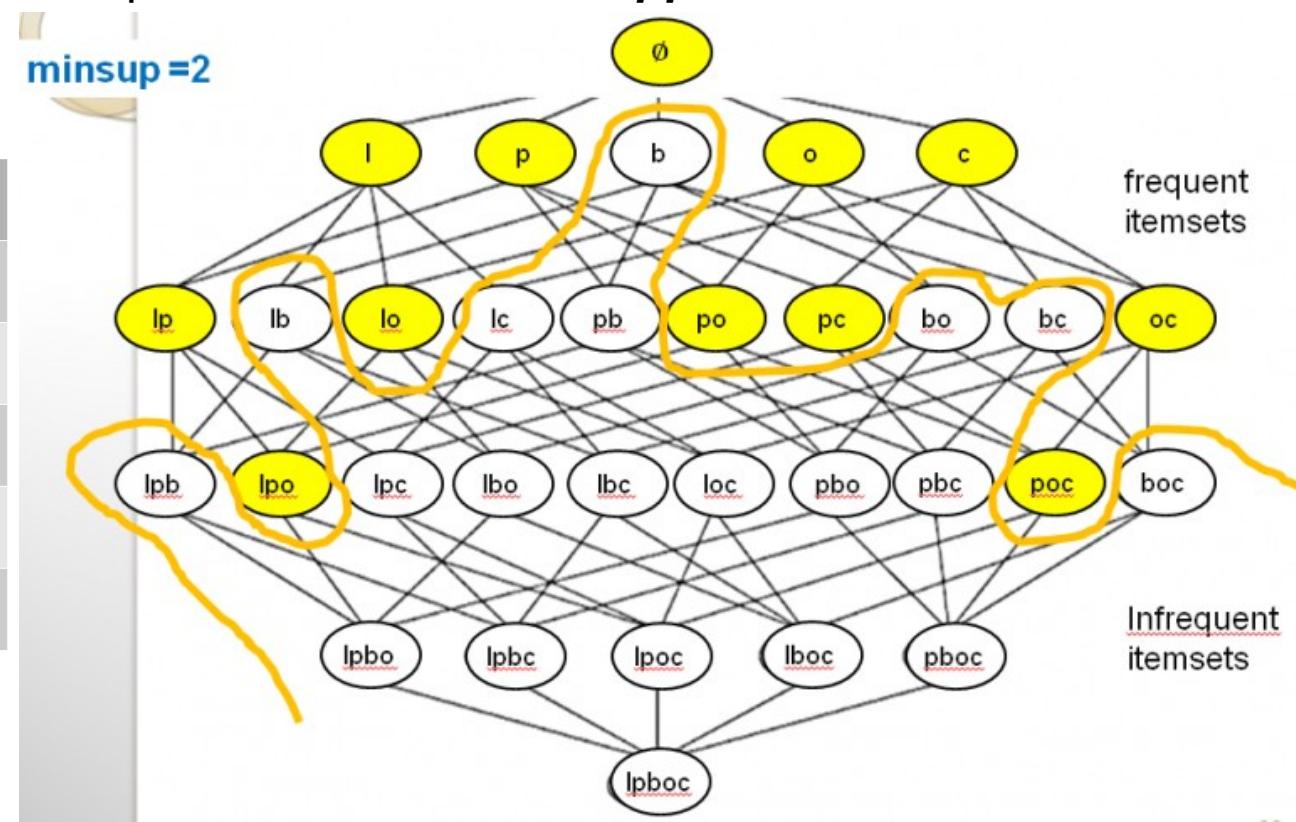
| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Definimos la cota para el soporte/relevancia: $\text{MinSupp} = 2$

| Item |
|--------------|
| Pasta = p |
| Limón = l |
| Naranja = o |
| Pan = b |
| Galletas = c |

11/32 Itemsets

13/32 Candidatos



El algoritmo **APRIORI** obtiene el conjunto de itemsets frecuentes a partir de los cuales debemos obtener las reglas de asociación. Por ejemplo, para el itemset $X=\{I1|I2|I3|I4\}$ tenemos las siguientes reglas posibles:

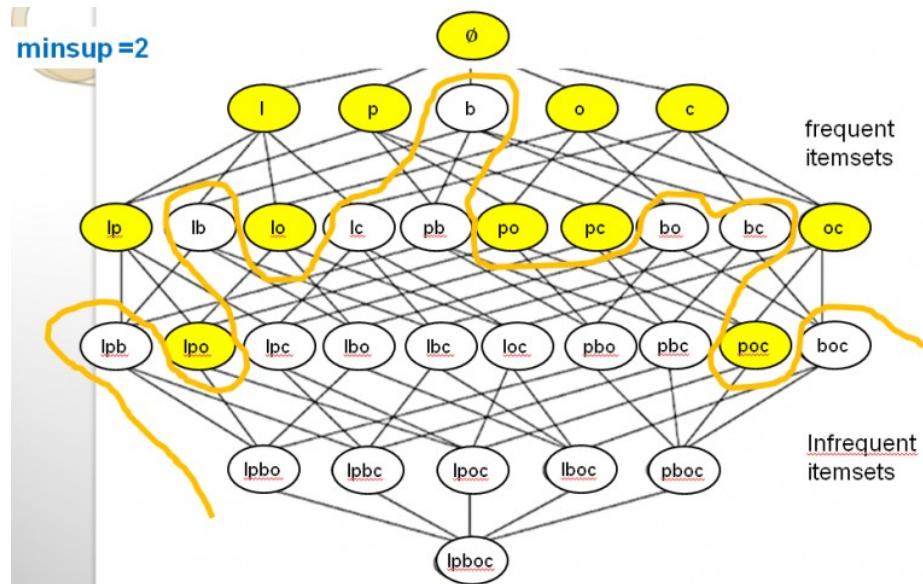
$I1|I2|I3 \rightarrow I4, I1|I2|I4 \rightarrow I3, I1|I3|I4 \rightarrow I2, I2|I3|I4 \rightarrow I1$ (4 reglas)

$I1 \rightarrow I2|I3|I4, I2 \rightarrow I1|I3|I4, I3 \rightarrow I1|I2|I4, I4 \rightarrow I1|I2|I3$ (4 reglas)

$I1|I2 \rightarrow I3|I4, I1|I3 \rightarrow I2|I4, I1|I4 \rightarrow I2|I3, I2|I3 \rightarrow I1|I4, I2|I4 \rightarrow I1|I3, I3|I4 \rightarrow I1|I2$ (6 reglas)

En general, para un itemset X con n elementos existen $2^n - 2$ reglas de asociación. Por tanto,

¿Cómo generar reglas de forma eficiente?



El algoritmo **APRIORI** obtiene el conjunto de itemsets frecuentes a partir de los cuales debemos obtener las reglas de asociación. Por ejemplo, para el itemset $X=\{I1|I2|I3|I4\}$ tenemos las siguientes reglas posibles:

$I1|I2|I3 \rightarrow I4, I1|I2|I4 \rightarrow I3, I1|I3|I4 \rightarrow I2, I2|I3|I4 \rightarrow I1$ (4 reglas)

$I1 \rightarrow I2|I3|I4, I2 \rightarrow I1|I3|I4, I3 \rightarrow I1|I2|I4, I4 \rightarrow I1|I2|I3$ (4 reglas)

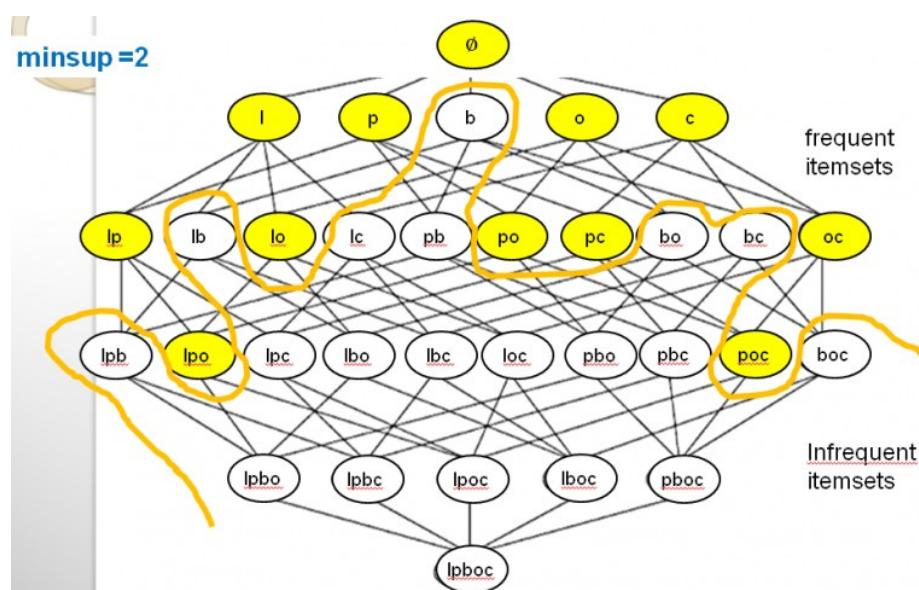
$I1|I2 \rightarrow I3|I4, I1|I3 \rightarrow I2|I4, I1|I4 \rightarrow I2|I3, I2|I3 \rightarrow I1|I4, I2|I4 \rightarrow I1|I3, I3|I4 \rightarrow I1|I2$ (6 reglas)

En general, para un itemset X con n elementos existen $2^n - 2$ reglas de asociación. Por tanto,

¿Cómo generar reglas de forma eficiente?

Uno de los criterios para filtrar las reglas es establecer una confianza mínima, **MinConf**, siendo la confianza: $C(X \rightarrow Y) = S(\{X, Y\})/S(X)$.

Sin embargo, al igual que en el caso de los itemsets y el soporte, ¿la confianza cumple alguna propiedad que permita “podar” el árbol de posibles reglas?



El algoritmo **APRIORI** obtiene el conjunto de itemsets frecuentes a partir de los cuales debemos obtener las reglas de asociación. Por ejemplo, para el itemset $X=\{I1|I2|I3|I4\}$ tenemos las siguientes reglas posibles:

$I1|I2|I3 \rightarrow I4, I1|I2|I4 \rightarrow I3, I1|I3|I4 \rightarrow I2, I2|I3|I4 \rightarrow I1$ (4 reglas)

$I1 \rightarrow I2|I3|I4, I2 \rightarrow I1|I3|I4, I3 \rightarrow I1|I2|I4, I4 \rightarrow I1|I2|I3$ (4 reglas)

$I1|I2 \rightarrow I3|I4, I1|I3 \rightarrow I2|I4, I1|I4 \rightarrow I2|I3, I2|I3 \rightarrow I1|I4, I2|I4 \rightarrow I1|I3, I3|I4 \rightarrow I1|I2$ (6 reglas)

En general, para un itemset X con n elementos existen $2^n - 2$ reglas de asociación. Por tanto,

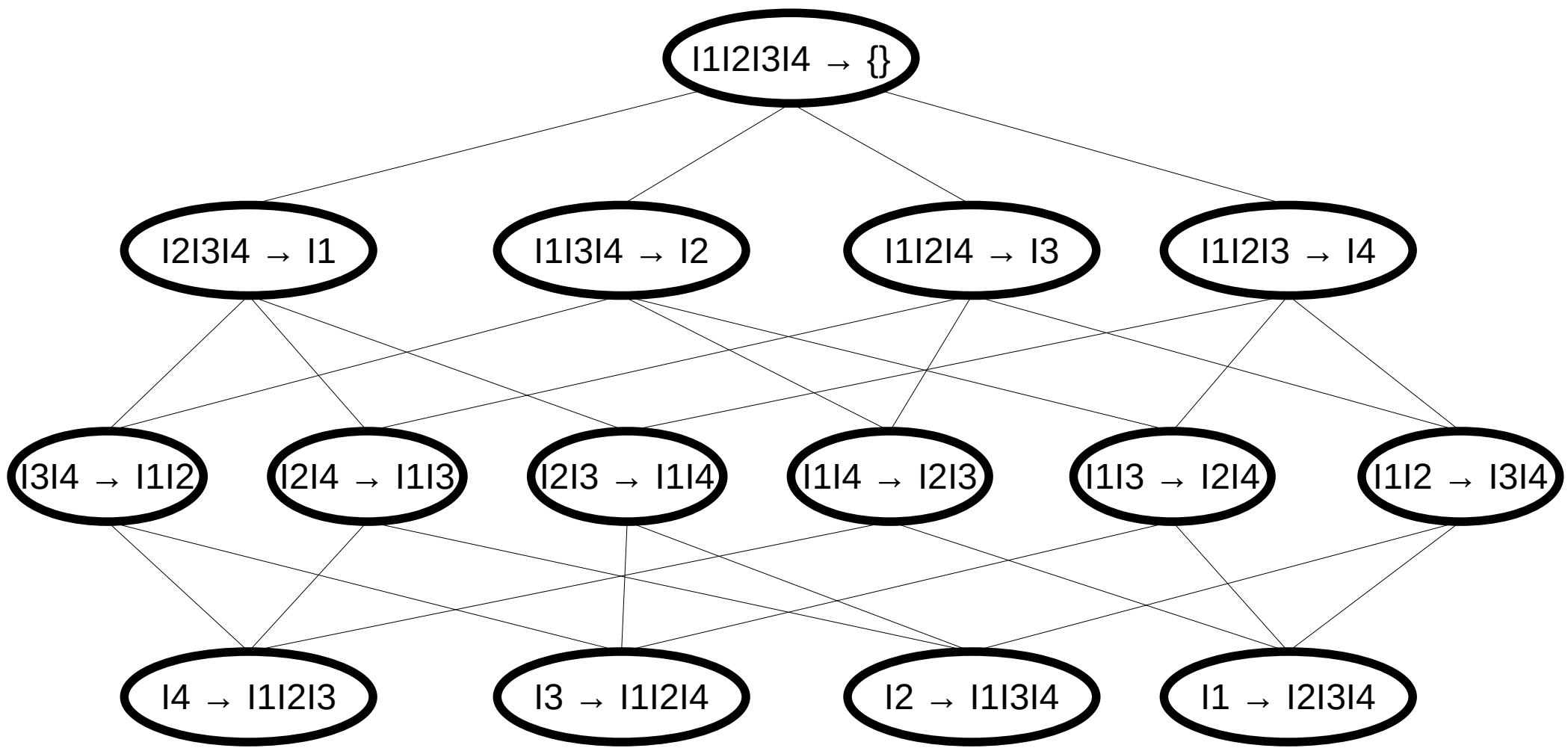
¿Cómo generar reglas de forma eficiente?

Uno de los criterios para filtrar las reglas es establecer una confianza mínima, **MinConf**, siendo la confianza: $C(X \rightarrow Y) = S(\{X, Y\})/S(X)$.

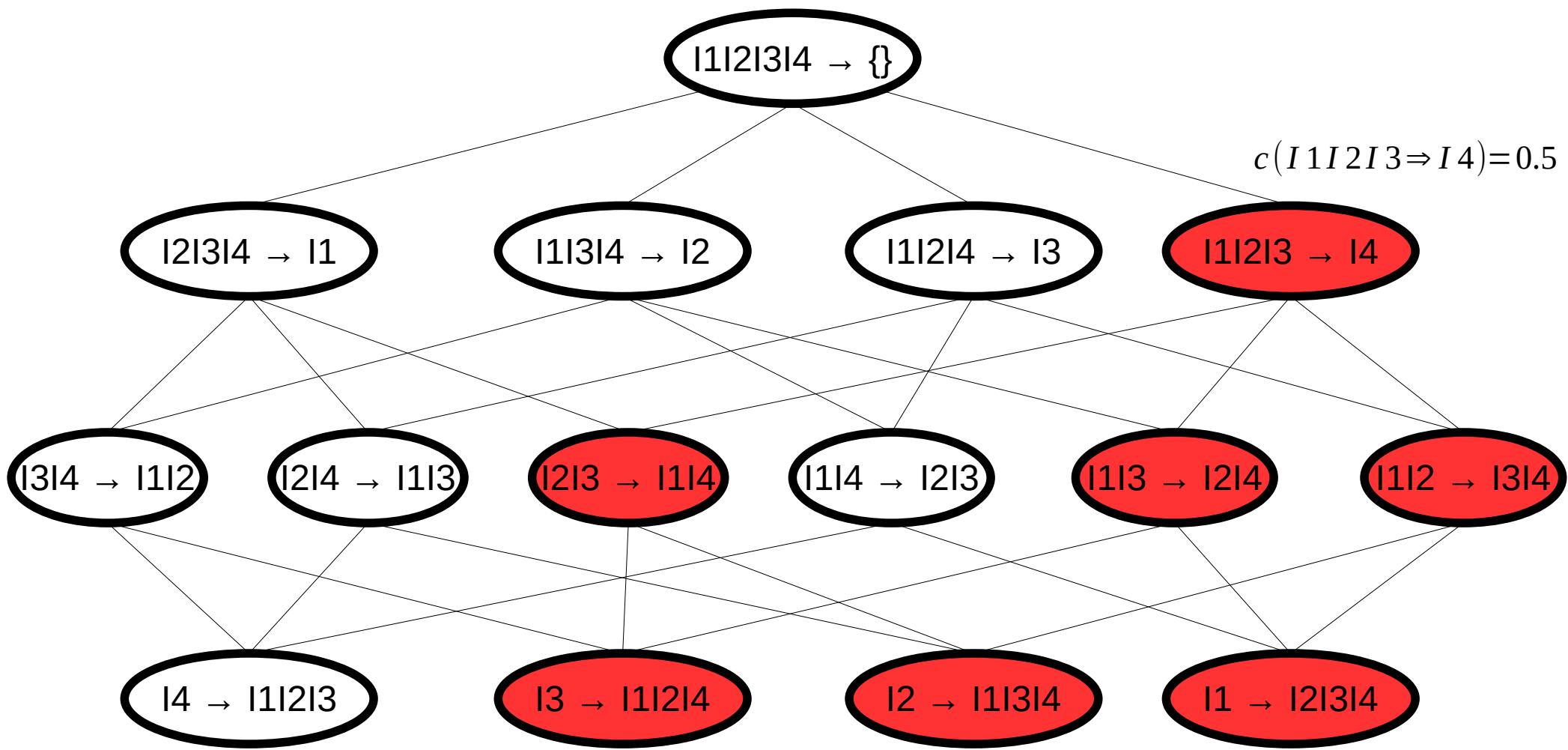
Sin embargo, al igual que en el caso de los itemsets y el soporte, **¿la confianza cumple alguna propiedad que permita “podar” el árbol de posibles reglas?**

La confianza es antimonótona con respecto al número de items en el consecuente de la regla:

$$c(I1|I2|I3 \Rightarrow I4) \geq c(I1|I2 \Rightarrow I3|I4) \geq c(I1 \Rightarrow I2|I3|I4)$$



$$c(I1I2I3 \Rightarrow I4) \geq c(I1I2 \Rightarrow I3I4) \geq c(I1 \Rightarrow I2I3I4)$$



Por la propiedad de la confianza, al no cumplir el criterio de la confianza la regla $I1I2I3 \rightarrow I4$, ninguna regla del subárbol inferior la cumplirá de modo que no es necesario evaluarlas.

$$c(I1I2I3 \Rightarrow I4) \geq c(I1I2 \Rightarrow I3I4) \geq c(I1 \Rightarrow I2I3I4)$$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: **$MinConf = 0.75$**

$$S_3 = \{I1|I2|I3; I1|I3|I5\}$$

$$C(I1|I2 \rightarrow I3) = 2/3; C(I1|I3 \rightarrow I2) = 2/3; C(I2|I3 \rightarrow I1) = 2/2;$$

$$C(I1 \rightarrow I2|I3) = 2/4; C(I2 \rightarrow I1|I3) = 2/3; C(I3 \rightarrow I1|I2) = 2/3;$$

$$C(I1|I3 \rightarrow I5) = 2/3; C(I1|I5 \rightarrow I3) = 2/2; C(I3|I5 \rightarrow I1) = 2/2;$$

$$C(I1 \rightarrow I3|I5) = 2/4; C(I3 \rightarrow I1|I5) = 2/3; C(I5 \rightarrow I1|I3) = 2/2;$$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: **$MinConf = 0.75$**

$S_3 = \{I1I2I3; I1I3I5\}$

~~$C(I1I2 \rightarrow I3) = 2/3; C(I1I3 \rightarrow I2) = 2/3; C(I2I3 \rightarrow I1) = 2/2;$~~

$C(I1 \rightarrow I2I3) = 2/4; C(I2 \rightarrow I1I3) = 2/3; C(I3 \rightarrow I1I2) = 2/3;$

No era necesario estimar esta confianza ya que, por la propiedad de la confianza, ya podía concluirse que esas reglas no iban a cumplir el criterio de la confianza para el umbral dado.

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: **MinConf = 0.75**

$$S_3 = \{I1I2I3; I1I3I5\}$$

$$C(I2I3 \rightarrow I1) = 2/2; C(I1I5 \rightarrow I3) = 2/2; C(I3I5 \rightarrow I1) = 2/2;$$

$$C(I5 \rightarrow I1I3) = 2/2;$$

$$S_2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\}$$

$$C(I1 \rightarrow I2) = 3/4; C(I2 \rightarrow I1) = 1; C(I1 \rightarrow I3) = 3/4; C(I3 \rightarrow I1) = 1; C(I5 \rightarrow I1) = 1;$$

$$C(I2 \rightarrow I3) = 2/3; C(I3 \rightarrow I2) = 2/3; C(I3 \rightarrow I5) = 2/3; C(I5 \rightarrow I3) = 1;$$

Notar que, dado que siempre se da el ítem 1, las reglas con este ítem como consecuente tienen confianza 1 aun cuando no aportan información relevante, ¿podemos filtrar reglas en base a ese criterio?

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: **$MinConf = 0.75$**

$$S_3 = \{I1I2I3; I1I3I5\}$$

$$C(I2I3 \rightarrow I1) = 2/2; C(I1I5 \rightarrow I3) = 2/2; C(I3I5 \rightarrow I1) = 2/2;$$

$$C(I5 \rightarrow I1I3) = 2/2;$$

$$S_2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\}$$

$$C(I1 \rightarrow I2) = 3/4; C(I2 \rightarrow I1) = 1; C(I1 \rightarrow I3) = 3/4; C(I3 \rightarrow I1) = 1; C(I5 \rightarrow I1) = 1;$$

$$C(I2 \rightarrow I3) = 2/3; C(I3 \rightarrow I2) = 2/3; C(I3 \rightarrow I5) = 2/3; C(I5 \rightarrow I3) = 1;$$

Notar que, dado que siempre se da el ítem 1, las reglas con este ítem como consecuente tienen confianza 1 aun cuando no aportan información relevante, ¿podemos filtrar reglas en base a ese criterio? Medidas de interés de la regla: $lift(X \rightarrow Y) = P(Y|X)/P(Y) = P(X, Y)/P(X)P(Y)$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: **MinConf = 0.75**

$$S_3 = \{I1I2I3; I1I3I5\}$$

$$C(I2I3 \rightarrow I1) = 2/2 \text{ (1)}; C(I1I5 \rightarrow I3) = 2/2 \text{ (1.33)}; C(I3I5 \rightarrow I1) = 2/2 \text{ (1)};$$

$$C(I5 \rightarrow I1I3) = 2/2 \text{ (1.33)};$$

$$S_2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\}$$

$$C(I1 \rightarrow I2) = 3/4 \text{ (1)}; C(I1 \rightarrow I3) = 3/4 \text{ (1)};$$

$$C(I2 \rightarrow I1) = 1 \text{ (1)}; C(I2 \rightarrow I3) = 2/3 \text{ (0.88)};$$

$$C(I3 \rightarrow I1) = 1 \text{ (1)}; C(I3 \rightarrow I2) = 2/3 \text{ (0.88)}; C(I3 \rightarrow I5) = 2/3 \text{ (1.33)};$$

$$C(I5 \rightarrow I1) = 1 \text{ (1)}; C(I5 \rightarrow I3) = 1 \text{ (1.33)};$$

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: $MinConf = 0.75$

$$S_3 = \{I1I2I3; I1I3I5\}$$

$$C(I2I3 \rightarrow I1) = 2/2 (1); C(I1I5 \rightarrow I3) = 2/2 (1.33); C(I3I5 \rightarrow I1) = 2/2 (1);$$

$$C(I5 \rightarrow I1I3) = 2/2 (1.33);$$

$$S_2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\}$$

$$C(I1 \rightarrow I2) = 3/4 (1); C(I1 \rightarrow I3) = 3/4 (1); \quad \text{Relación negativa}$$

$$C(I2 \rightarrow I1) = 1 (1); C(I2 \rightarrow I3) = 2/3 (0.88); \quad \text{Relación positiva}$$

$$C(I3 \rightarrow I1) = 1 (1); C(I3 \rightarrow I2) = 2/3 (0.88); C(I3 \rightarrow I5) = 2/3 (1.33)$$

$$C(I5 \rightarrow I1) = 1 (1); C(I5 \rightarrow I3) = 1 (1.33);$$

Independientes

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Otra medida es la **convicción de la regla**: $conv(X \rightarrow Y) = (1 - S(Y))/(1 - C(X \rightarrow Y))$

$S_3 = \{I1I2I3; I1I3I5\}$

$C(I2I3 \rightarrow I1) = 2/2 (1, 0); C(I1I5 \rightarrow I3) = 2/2 (1.33, Inf); C(I3I5 \rightarrow I1) = 2/2 (1, 0);$

$C(I5 \rightarrow I1I3) = 2/2 (1.33, Inf);$

$S_2 = \{I1I2; I1I3; I1I5; I2I3; I3I5\}$

$C(I1 \rightarrow I2) = 3/4 (1, 1); C(I2 \rightarrow I1) = 1 (1, 0); C(I2 \rightarrow I3) = 2/3 (0.88, 0.375);$

Notar que la convicción nula se corresponde con reglas sin interés, valores **superiores/inferiores** a 1 indica que la regla predecirá correctamente **más/menos** casos que si la relación entre ambos itemsets fuera aleatoria.

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Filtramos las reglas en base a su confianza ($C(X \rightarrow Y) = S(\{X, Y\})/S(X)$)

Definimos la cota para la confianza de la regla: ***MinConf = 0.75***

$C(I1I5 \rightarrow I3) = 2/2 (1.33)$; $C(I5 \rightarrow I1I3) = 2/2 (1.33)$;

$C(I1 \rightarrow I2) = 3/4 (1)$; $C(I1 \rightarrow I3) = 3/4 (1)$;

$C(I2 \rightarrow I3) = 2/3 (0.88)$; $C(I3 \rightarrow I2) = 2/3 (0.88)$; $C(I3 \rightarrow I5) = 2/3 (1.33)$;

$C(I5 \rightarrow I3) = 1 (1.33)$;

| Transacción | I1-Pasta | I2-Limpio | Measure | Formula |
|-------------|----------|-----------|---------------------------------|--|
| T1 | 1 | 1 | ϕ -coefficient | $\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))} \cdot \sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)} \cdot \frac{2 - \max_j P(A_j) - \max_k P(B_k)}{P(A,B)P(\bar{A},\bar{B})}$ |
| T2 | 1 | 1 | Goodman-Kruskal's (λ) | $\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$ |
| T3 | 1 | 0 | Odds ratio (α) | $\frac{P(A,B)P(\bar{A}\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(AB) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$ |
| T4 | 1 | 1 | Yule's Q | $\frac{\sqrt{P(A,B)P(AB)} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(AB)} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$ |
| | | | Yule's Y | $\frac{\sqrt{P(A,B)P(\bar{A}\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}}{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})} \cdot \frac{P(A_i, B_j)}{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}$ |
| | | | Kappa (κ) | $\min(-\sum_i P(A_i) \log P(A_i) - \sum_j P(B_j) \log P(B_j))$ |
| | | | Mutual Information (M) | $\max \left(P(A,B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}B) \log \left(\frac{P(\bar{B} A)}{P(\bar{B})} \right), P(A,B) \log \left(\frac{P(\bar{A} B)}{P(\bar{A})} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{A} B)}{P(\bar{A})} \right) \right)$ |
| | | | J-Measure (J) | $\max \left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2, P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] - P(A)^2 - P(\bar{A})^2 \right)$ |
| | | | Gini index (G) | $P(A,B)$ |
| | | | Support (s) | $\max(P(B A), P(A B))$ |
| | | | Confidence (c) | $\max \left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$ |
| | | | Laplace (L) | $\max \left(\frac{P(A)P(\bar{B})}{P(\bar{A}B)}, \frac{P(B)P(\bar{A})}{P(B\bar{A})} \right)$ |
| | | | Conviction (V) | $\frac{P(A,B)}{P(A)P(B)}$ |
| | | | Interest (I) | $\frac{P(A,B) - P(A)P(B)}{P(A,B) - P(A)P(B)}$ |
| | | | cosine (IS) | $\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$ |
| | | | Piatetsky-Shapiro's (PS) | $\frac{P(A,B) - P(A)P(B)}{P(A,B) - P(A)P(B)}$ |
| | | | Certainty factor (F) | $\max \left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$ |
| | | | Added Value (AV) | $\max(P(B A) - P(B), P(A B) - P(A))$ |
| | | | Collective strength (S) | $\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$ |
| | | | Jaccard (ζ) | $\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$ |
| | | | Klosgen (K) | $\sqrt{P(A,B)} \max(P(B A) - P(B), P(A B) - P(A))$ |

Filtramos las reglas en base a su confianza.

Definimos la cota para la confianza de la regla:

$$C(I1I5 \rightarrow I3) = 2/2 \text{ (1.33)}; C(I5 \rightarrow I1I3) = 2/2 \text{ (1.33)}$$

$$C(I1 \rightarrow I2) = 3/4 \text{ (1)}; C(I1 \rightarrow I3) = 3/4 \text{ (1)}$$

$$C(I2 \rightarrow I3) = 2/3 \text{ (0.88)}; C(I3 \rightarrow I2) = 2/3 \text{ (0.88)}$$

$$C(I5 \rightarrow I3) = 1 \text{ (1.33)};$$

Existen muchas más medidas de interés con diferentes propiedades y aplicaciones.

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Aún filtrando las reglas en base a los criterios anteriores, pueden darse reglas asociadas a una misma propiedad, o **redundantes**.

$$C(X_1 \Rightarrow Y_1) = \{ X_2 \Rightarrow Y_2 \mid X_1 \subseteq X_2 \wedge X_2 \subseteq X_1 \}$$

Estas reglas deben también eliminarse para obtener un conjunto mínimo representativo de las relaciones relevantes.

$$C(I1I5 \rightarrow I3) = 2/2 (1.33); C(I5 \rightarrow I1I3) = 2/2 (1.33);$$

$$C(I1 \rightarrow I2) = 3/4 (1); C(I1 \rightarrow I3) = 3/4 (1);$$

$$C(I2 \rightarrow I3) = 2/3 (0.88); C(I3 \rightarrow I2) = 2/3 (0.88); C(I3 \rightarrow I5) = 2/3 (1.33);$$

$$C(I5 \rightarrow I3) = 1 (1.33);$$

Aún filtrando las reglas en base a los criterios anteriores, pueden darse reglas asociadas a una misma propiedad, o **redundantes**.

$$C(X_1 \Rightarrow Y_1) = \{ X_2 \Rightarrow Y_2 \mid X_1 \subseteq X_2 \wedge X_2 Y_2 \subseteq X_1 Y_1 \}$$

Estas reglas deben también eliminarse para obtener un conjunto mínimo representativo de las relaciones relevantes.

C(I1I5 → I3) = 2/2 (1.1) is.redundant {arules}

R Documentation

C(I1 → I2) = 3/4 (1); cFind Redundant Rules

C(I2 → I3) = 2/3 (0.88) Description

C(I5 → I3) = 1 (1.33); Provides the generic functions and the S4 method is.redundant to find redundant rules.

Usage

```
is.redundant(x, ...)
## S4 method for signature 'rules'
is.redundant(x, measure = "confidence")
```

Arguments

- x a set of rules.
- measure measure used to check for redundancy.
- ... additional arguments.

Details

A rule is redundant if a more general rules with the same or a higher confidence exists. That is, a more specific rule is redundant if it is only equally or even less predictive than a more general rule. A rule is more general if it has the same RHS but one or more items removed from the LHS. Formally, a rule $X \rightarrow Y$ is redundant if

for some X' subset X , $\text{conf}(X' \rightarrow Y) \geq \text{conf}(X \rightarrow Y)$.

| Transacción | I1-Pasta | I2-Limon | I3-Naranja | I4-Pan | I5-Galletas |
|-------------|----------|----------|------------|--------|-------------|
| T1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 1 | 0 | 0 | 0 |
| T3 | 1 | 0 | 1 | 0 | 1 |
| T4 | 1 | 1 | 1 | 0 | 1 |

Aún filtrando las reglas en base a los criterios anteriores, pueden darse reglas asociadas a una misma propiedad, o **redundantes**.

$$C(X_1 \Rightarrow Y_1) = \{ X_2 \Rightarrow Y_2 \mid X_1 \subseteq X_2 \wedge X_2 \subseteq X_1 \}$$

Estas reglas deben también eliminarse para obtener un conjunto mínimo representativo de las relaciones relevantes.

~~C(I1I5 → I3) = 2/2 (1.33); C(I5 → I1I3) = 2/2 (1.33); ← Redundantes y con C(I5 → I3).~~

$C(I1 \rightarrow I2) = 3/4 (1)$; $C(I1 \rightarrow I3) = 3/4 (1)$;

$C(I5 \rightarrow I3) = 1 (1.33)$;

Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería aRules/arulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat

Los paquetes **aRules** y **arulesViz** son dos librerías de R para el aprendizaje de Reglas de Asociación a partir de un conjunto de datos.

Introduction to arules – A computational environment for mining association rules and frequent item sets

Michael Hahsler
Southern Methodist University

Bettina Grün
Johannes Kepler University Linz

Kurt Hornik
Wirtschaftsuniversität Wien

Christian Buchta
Wirtschaftsuniversität Wien

Visualizing Association Rules: Introduction to the R-extension Package arulesViz

Michael Hahsler
Southern Methodist University

Sudheer Chelluboina
Southern Methodist University

A smaller dataset “Groceries” from **arulesViz** package will be used in the course.



transactions as itemMatrix in sparse format with 9835 rows (elements/itemsets/transactions) and 169 columns (items) and a density of 0.02609146

most frequent items:

| | |
|-----------------------------|------------|
| whole milk other vegetables | rolls/buns |
| 2513 | 1903 |
| ----- | ----- |
| ----- | 1809 |

```
library("arulesViz")
data("Groceries")
```

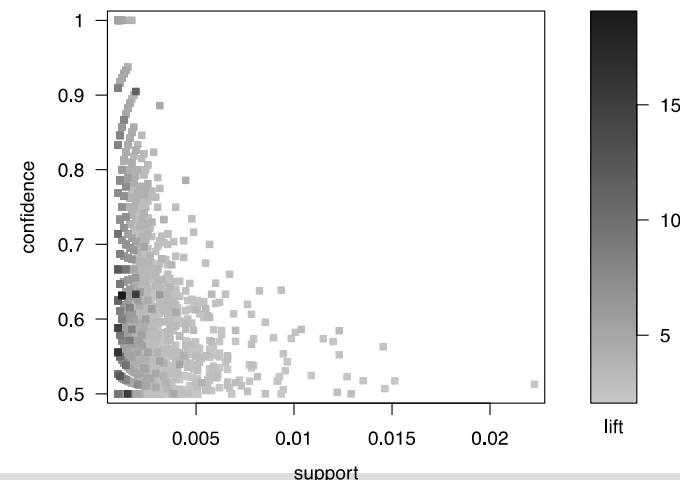
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146

most frequent items:

| | | |
|------------|------------------|------------|
| whole milk | other vegetables | rolls/buns |
| 2513 | 1903 | 1809 |
| yogurt | (Other) | |
| 1372 | 34055 | |

```
> rules <- apriori(Groceries, parameter=list(support=0.001, confidence=0.5))
> inspect(head(sort(rules, by ="lift"),3))
```

Scatter plot for 5668 rules



$$\text{lift}(X \Rightarrow Y) = \text{supp}(X \cup Y)/(\text{supp}(X)\text{supp}(Y))$$

| lhs | rhs | support | confidence | lift |
|------------------------------------|---------------------|-------------|------------|----------|
| 1 {Instant food products, soda} | => {hamburger meat} | 0.001220132 | 0.6315789 | 18.99565 |
| 2 {soda, popcorn} | => {salty snack} | 0.001220132 | 0.6315789 | 16.69779 |
| 3 {flour, baking powder} | => {sugar} | 0.001016777 | 0.5555556 | 16.40807 |

A smaller dataset “Groceries”
from **arulesViz** package will be
used in the course.



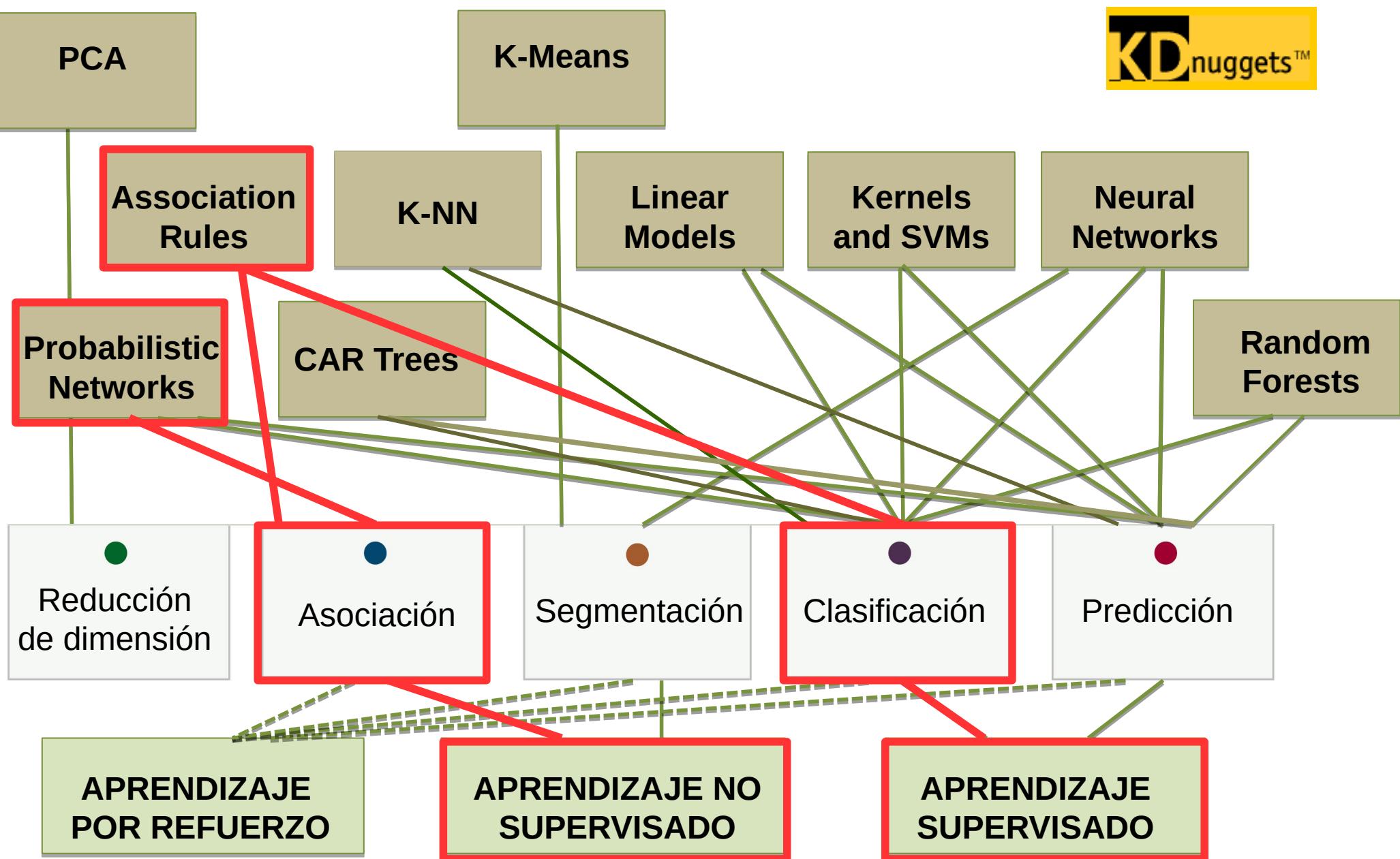
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146

most frequent items:

| | | |
|------------|------------------|------------|
| whole milk | other vegetables | rolls/buns |
| 2513 | 1903 | 1809 |

Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería aRules/arulesViz
- **Reglas de Clasificación → Árboles de Clasificación**
- Algoritmo Eclat



Machine Learning with R

Second Edition

Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R

Understanding classification rules

Classification rules represent knowledge in the form of logical if-else statements that assign a class to unlabeled examples. They are specified in terms of an **antecedent** and a **consequent**; these form a hypothesis stating that "if this happens, then that happens." A simple rule might state, "if the hard drive is making a clicking sound, then it is about to fail." The antecedent comprises certain combinations of feature values, while the consequent specifies the class value to assign when the rule's conditions are met.

Rule learners are often used in a manner similar to decision tree learners. Like decision trees, they can be used for applications that generate knowledge for future action, such as:

- Identifying conditions that lead to a hardware failure in mechanical devices
- Describing the key characteristics of groups of people for customer segmentation
- Finding conditions that precede large drops or increases in the prices of shares on the stock market

[149]

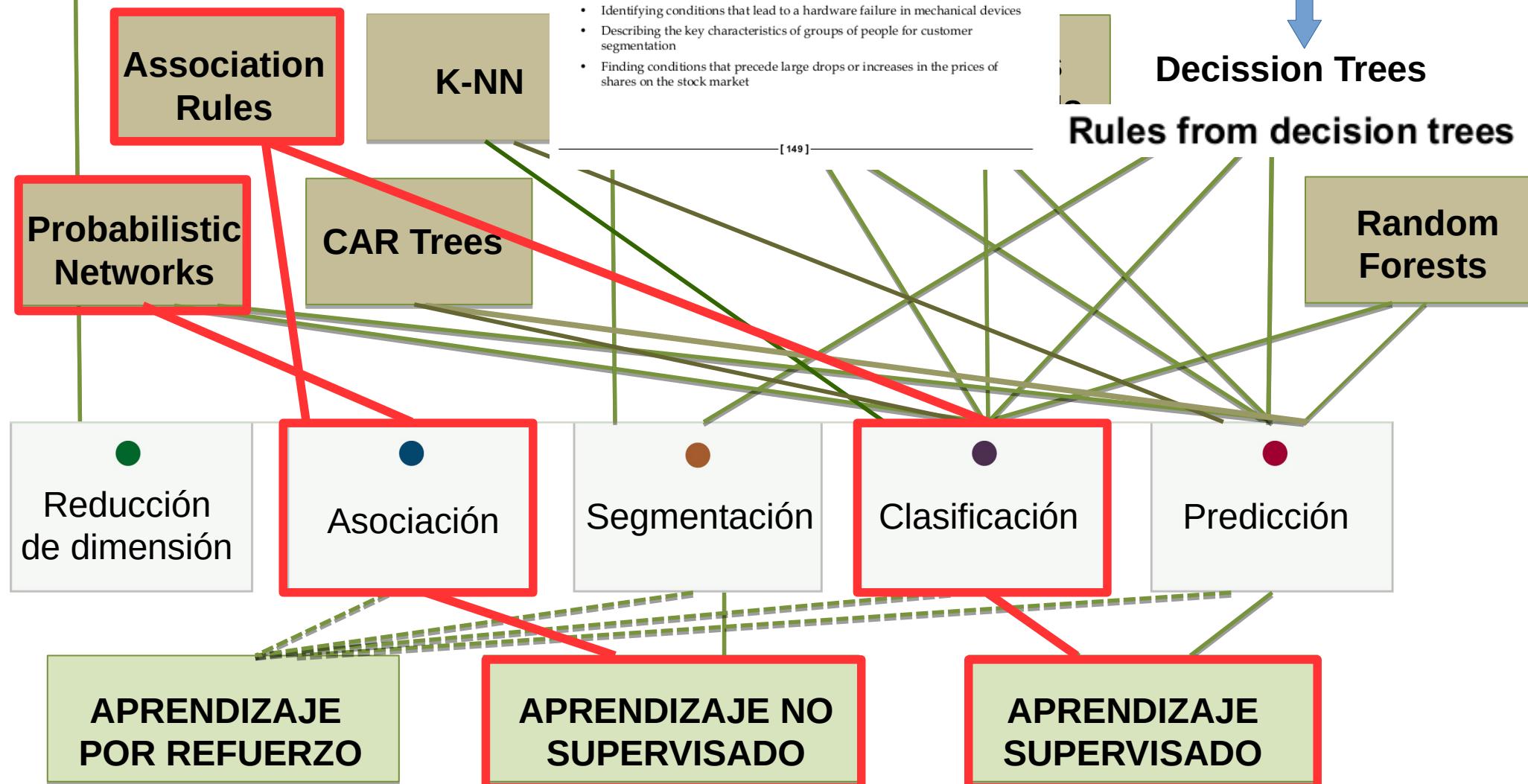
The 1R algorithm
The RIPPER algorithm



Decission Trees

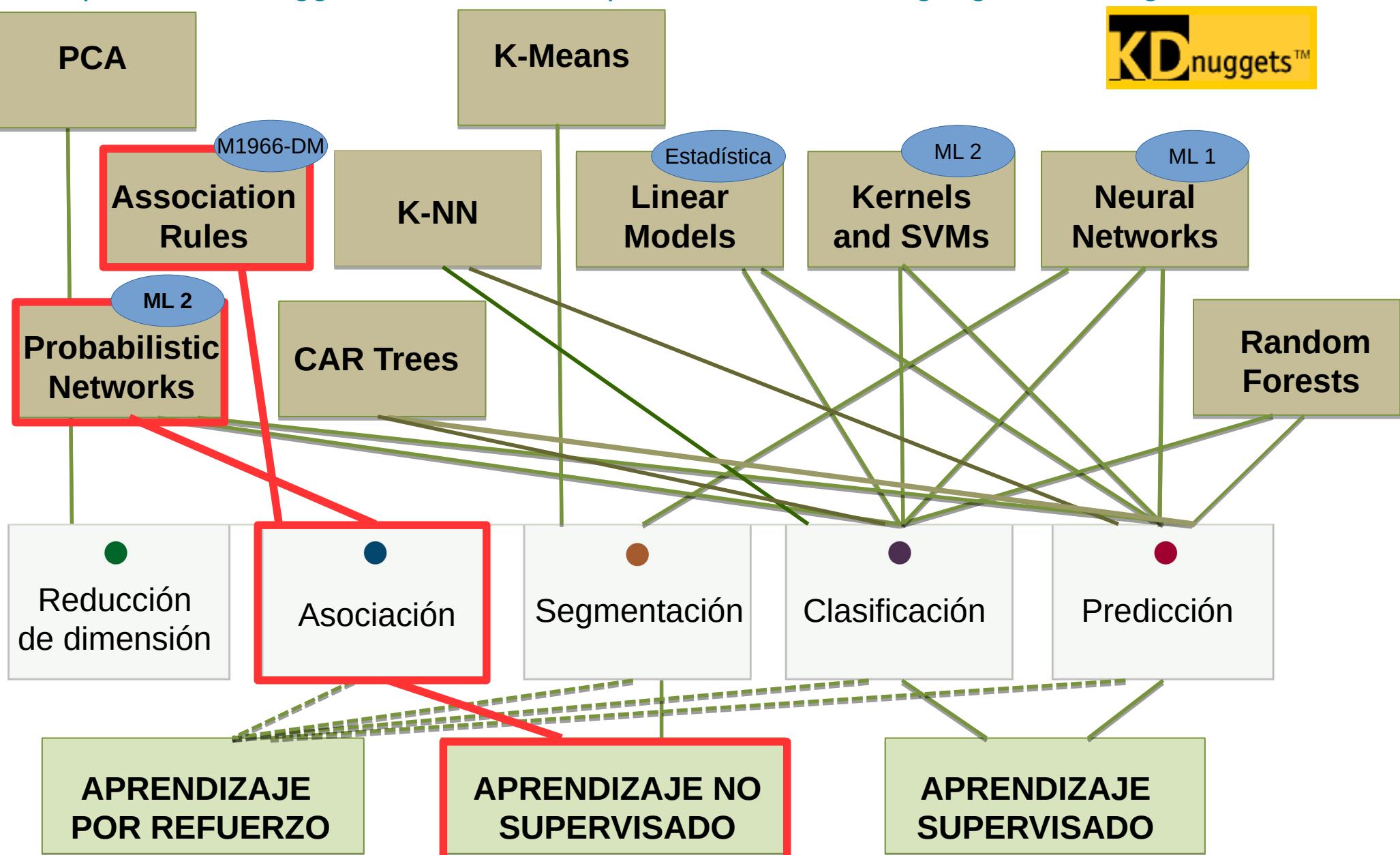
Rules from decision trees

Random
Forests

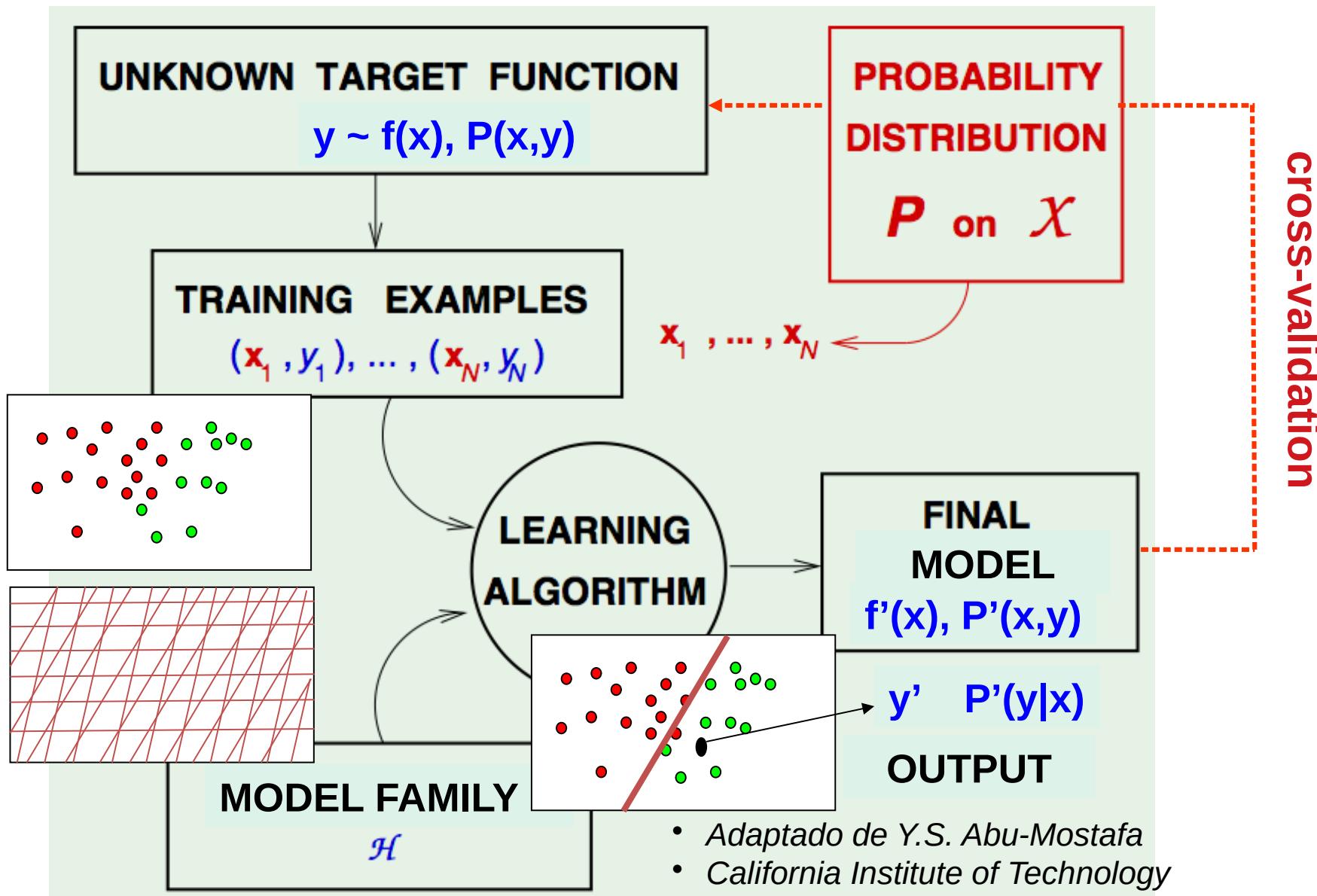


Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería aRules/arulesViz
- Redes Probabilísticas → Introducción
- Algoritmo Eclat



Generalization is the most important feature for data driven systems:
They must perform “well” when applied to new data (**cross-validation**).



| x | y | z | $p(x, y, z)$ |
|-----|-----|-----|--------------|
| 0 | 0 | 0 | 0.12 |
| 0 | 0 | 1 | 0.18 |
| 0 | 1 | 0 | 0.04 |
| 0 | 1 | 1 | 0.16 |
| 1 | 0 | 0 | 0.09 |
| 1 | 0 | 1 | 0.21 |
| 1 | 1 | 0 | 0.02 |
| 1 | 1 | 1 | 0.18 |

La especificación directa de esta función de probabilidad no es posible en casos prácticos (10^{25} parámetros para 100 variables).

Una solución es construir modelos con menos parámetros, limitando el número de posibles dependencias entre variables □ **Factorización:**

$$p(A | B) = \frac{p(A, B)}{p(B)} = \frac{p(B | A) p(A)}{p(B)}$$

¿Cómo se construye un modelo probabilístico que tenga un conjunto de independencias dado?

$$I(X_3, X_1 | X_2) \text{ and } I(X_4, \{X_1, X_3\} | X_2).$$

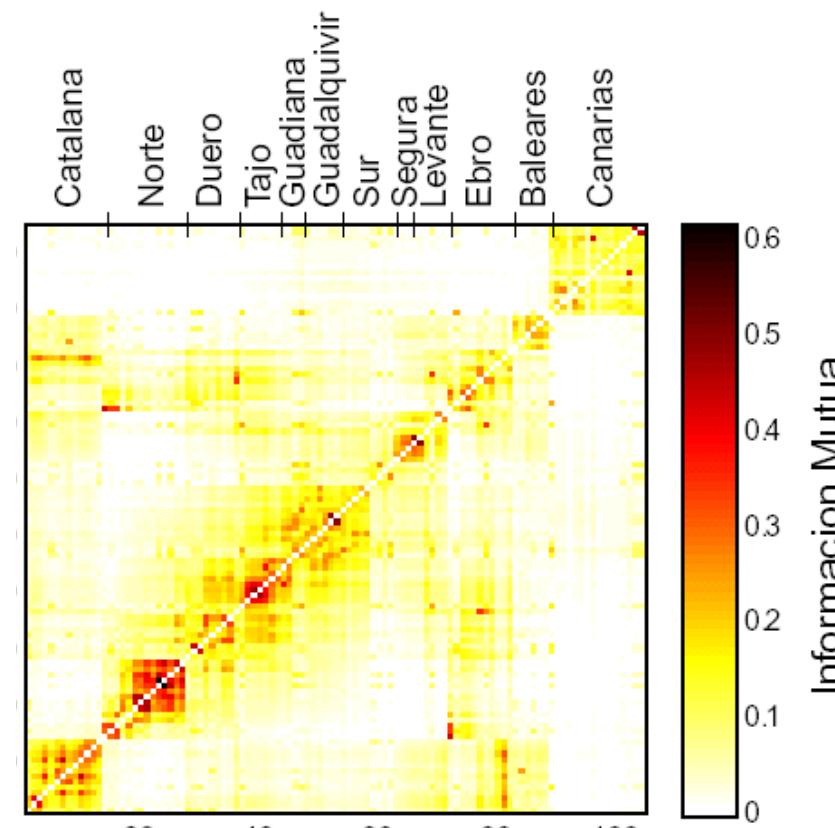
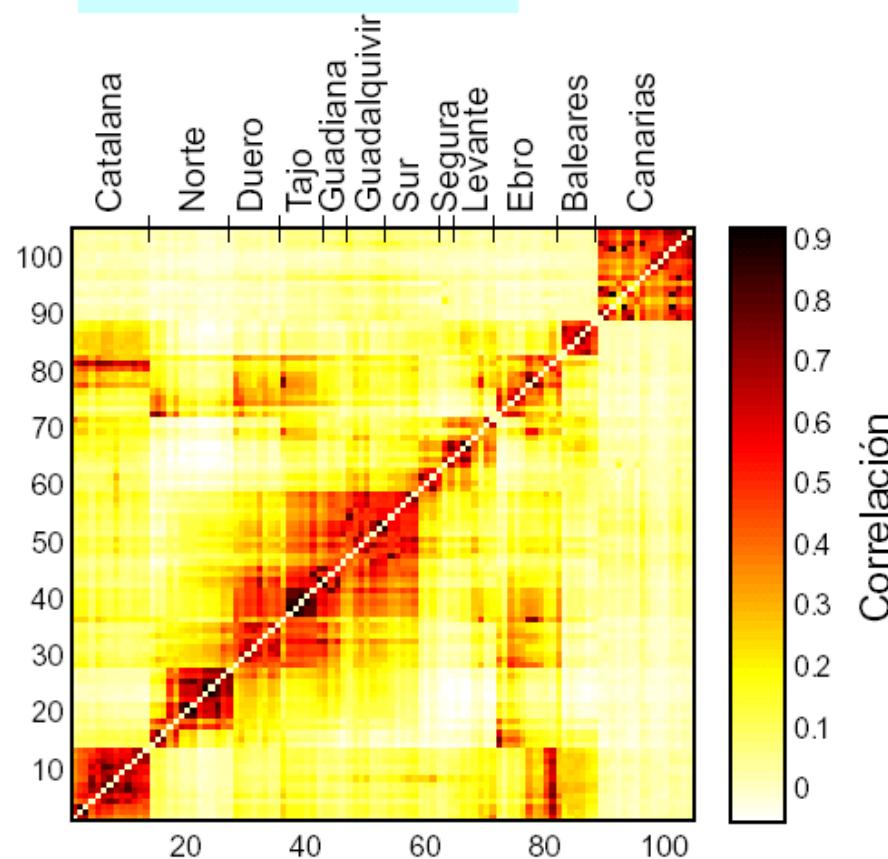
$$p(x_3 | x_1, x_2) = p(x_3 | x_2),$$

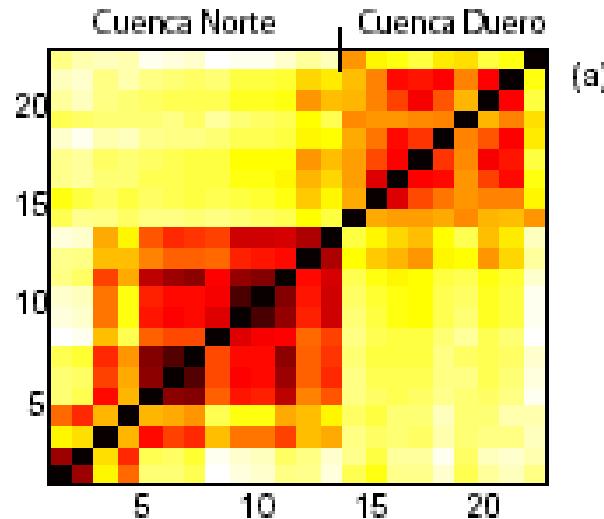
$$p(x_4 | x_1, x_2, x_3) = p(x_4 | x_2).$$

$$p(x_1, \dots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2).$$

La dependencia entre dos variables se puede calcular fácilmente con distintas medidas. Sin embargo, esta información es parcial.

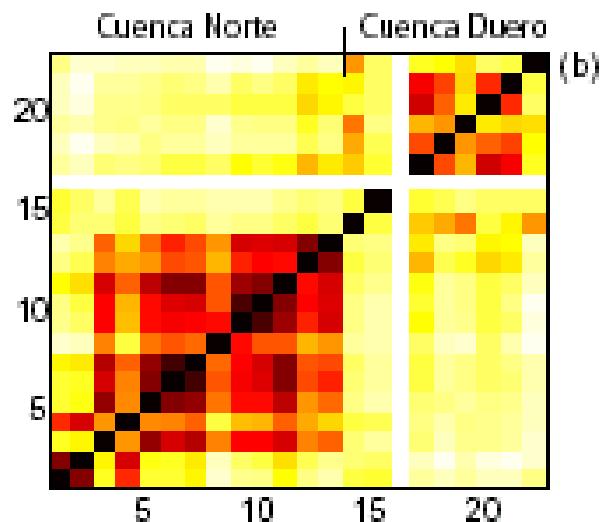
$$r = \frac{xy}{x \cdot y}$$



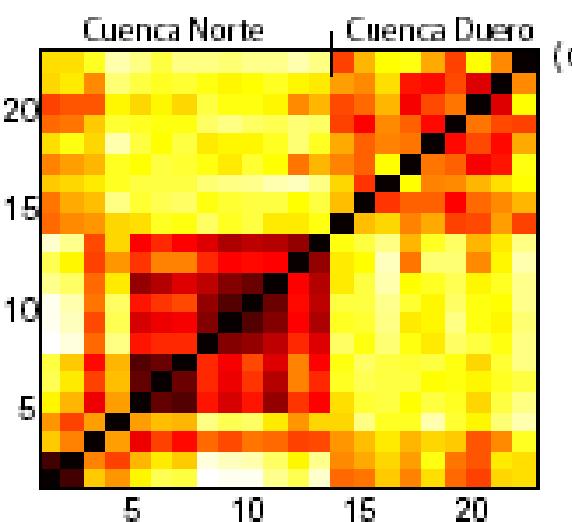


Es necesario el concepto de (in)dependencia condicional.

$$\begin{aligned} I(X, Y|Z) \text{ si } \\ P_z(Y|X) = P_z(Y) \\ P(Y|X, Z) = P(Y|Z) \end{aligned}$$



$\text{Precip(Palencia)}=0 \text{ mm}$



$\text{Precip(Palencia)}>10 \text{ mm}$

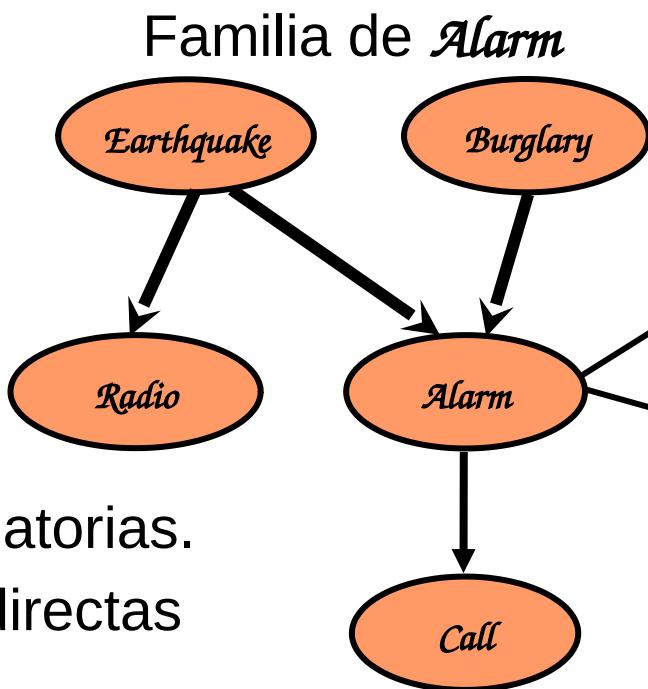
La correlación parcial tiene en cuenta la relación entre dos variables, una vez que se ha eliminado la dependencia con una tercera.

Representación compacta de una función de probabilidad conjunta mediante independencia condicional.

Parte cualitativa:

Grafo dirigido acíclico (DAG), o grafos no dirigidos (redes Markov)

Nodos – variables aleatorias.
Aristas – influencias directas



| E | B | $P(\mathcal{A} \mid E, B)$ | |
|-----------|-----------|----------------------------|------|
| e | b | 0.9 | 0.1 |
| e | \bar{b} | 0.2 | 0.8 |
| \bar{e} | b | 0.9 | 0.1 |
| \bar{e} | \bar{b} | 0.01 | 0.99 |

Junto con:

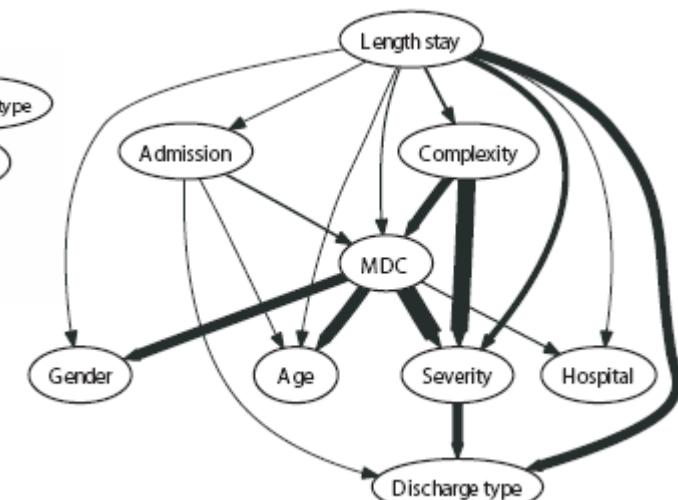
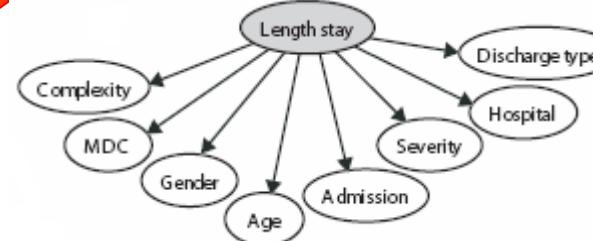
Factorización de una única función de probabilidad

Parte cuantitativa:

Conjunto de funciones/tablas de probabilidad.

$$P(B, E, \mathcal{A}, C, \mathcal{R}) = P(B)P(E)P(\mathcal{A} \mid B, E)P(\mathcal{R} \mid E)P(C \mid \mathcal{A})$$

1. Estancia Media
 2. Tasa de Mortalidad
 3. Tasa de Reingresos (a 30 días)
 4. Tasa de Infección Nosocomial
 5. Estancia Media Preoperatoria
 6. Tasa de Cesáreas
- Conjunto de indicadores**



$$\frac{1}{\#(X)\#(Z)} \sum_{x,y,z} P(y|x,z) \log_2 \left(\frac{P(y|x,z)}{\frac{1}{\#(X)} \sum_x P(y|x,z)} \right)$$

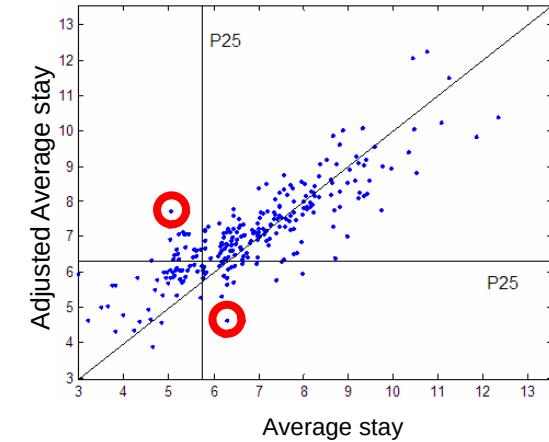
RELACIONADAS CON LA ENFERMEDAD

1. Complejidad (medido a través del peso español de GRD-AP v 18)
2. Severidad (medido a través de GRD refinados)
3. Categoría Diagnóstica Mayor de GRD-AP v 18
4. Tipo de GRD: medico, quirúrgico, indeterminado (Solo aplicado al indicador de infección nosocomial)

RELACIONADAS CON EL PACIENTE, O CON EL FUNCIONAMIENTO HOSPITALARIO

1. Edad
2. Sexo
3. Tipo de ingreso
4. Tipo de alta
5. Tipo de hospital

Nuevo sistema genérico para el ajuste de indicadores sin recurrir a regresiones, etc.



Reglas de Asociación

- Introducción – Objetivo
- Modelo Formal - Conceptos Básicos
- Reglas de Asociación
- Algoritmo Apriori
- Librería Arules/ArulesViz
- Redes Probabilísticas → Introducción
- **Algoritmo Eclat**

¿Los registros de las ventas realizadas en el pasado son útiles para la gestión de un comercio?



1) Colocación/proposición de productos:

Si detecto productos que se compran de forma conjunta puedo planificar su colocación en el comercio.

2) Promociones y Ofertas:

Si detecto productos correlacionados puedo definir promociones que engloben a ambos (p.e. portátil + mochila, refrescos + snacks, juguetes + pilas, etc...).

**Los algoritmos que buscan reglas de asociación analizan los datos registrados buscando relaciones entre los diferentes productos, obteniendo proposiciones del tipo:
“El 90% de las compras que incluyen pan y mantequilla, compran también leche.”**

A partir de las reglas encontradas, se resuelven consultas de diferentes tipos:

- Encontrar todas las reglas relacionando cualquier producto con uno dado (p.e. leche).
- Encontrar todas las reglas que dependen de un producto dado (p.e. pan).
- Encontrar todas las reglas que cumplen los dos criterios anteriores para ciertos productos dados.
- Encontrar el conjunto de reglas con mayor confianza cumpliendo alguno de los criterios anteriores.
- ...

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining

El algoritmo ECLAT (Equivalence Class Clustering and bottom up Lattice Traversal) es otro método para la generación de itemsets frecuentes. Se entiende que el algoritmo **APRIORI** parte de una ordenación **horizontal** de los datos y las transacciones, mientras que **ECLAT** parte de una representación **vertical** a partir de la cual construye las intersecciones de los itemsets:

| TID | Items |
|-----|---------------------------------|
| T1 | Pasta, Limón, Naranja, Pan |
| T2 | Pasta, Limón |
| T3 | Pasta, Naranja, Galletas |
| T4 | Pasta, Limón, Naranja, Galletas |

APRIORI ↔ **Horizontal**

| Item | TID |
|----------|------------|
| Pasta | 1, 2, 3, 4 |
| Limón | 1, 2, 4 |
| Naranja | 1, 3, 4 |
| Pan | 1 |
| Galletas | 3, 4 |

ECLAT ↔ **Vertical**

El algoritmo **ECLAT** (Equivalence Class Clustering and bottom up Lattice Traversal) es otro método para la generación de itemsets frecuentes. Se entiende que el algoritmo **APRIORI** parte de una ordenación **horizontal** de los datos y las transacciones, mientras que **ECLAT** parte de una representación **vertical**:

| Item (1) | TID |
|----------|------------|
| Pasta | 1, 2, 3, 4 |
| Limón | 1, 2, 4 |
| Naranja | 1, 3, 4 |
| Pan | 1 |
| Galletas | 3, 4 |

| Item (2) | TID |
|-------------------|---------|
| Pasta, Limón | 1, 2, 4 |
| Pasta, Naranja | 1, 3, 4 |
| Pasta, Pan | 1 |
| Pasta, Galletas | 3, 4 |
| Limón, Naranja | 1, 4 |
| Limón, Pan | 1 |
| Limón, Galletas | 4 |
| Naranja, Pan | 1 |
| Naranja, Galletas | 3, 4 |

| Item (3) | TID |
|--------------------------|------|
| Pasta, Limón, Naranja | 1, 4 |
| Pasta, Limón, Pan | 1 |
| Pasta, Limón, Galletas | 4 |
| Pasta, Naranja, Pan | 1 |
| Pasta, Naranja, Galletas | 3, 4 |
| Limón, Naranja, Pan | 1 |
| Limón, Naranja, Galletas | 4 |

| Item (4) | TID |
|---------------------------------|-----|
| Pasta, Limón, Naranja, Pan | 1 |
| Pasta, Limón, Naranja, Galletas | 4 |

El algoritmo **ECLAT** (Equivalence Class Clustering and bottom up Lattice Traversal) es otro método para la generación de itemsets frecuentes. Se entiende que el algoritmo **APRIORI** parte de una ordenación **horizontal** de los datos y las transacciones, mientras que **ECLAT** parte de una representación **vertical**:

| Item (1) | TID |
|----------|------------|
| Pasta | 1, 2, 3, 4 |
| Limón | 1, 2, 4 |
| Naranja | 1, 3, 4 |
| Pan | 1 |
| Galletas | 3, 4 |

MinSupp = 2

| Item (2) | TID |
|-------------------|---------|
| Pasta, Limón | 1, 2, 4 |
| Pasta, Naranja | 1, 3, 4 |
| Pasta, Pan | 1 |
| Pasta, Galletas | 3, 4 |
| Limón, Naranja | 1, 4 |
| Limón, Pan | 1 |
| Limón, Galletas | 4 |
| Naranja, Pan | 1 |
| Naranja, Galletas | 3, 4 |

| Item (3) | TID |
|--------------------------|------|
| Pasta, Limón, Naranja | 1, 4 |
| Pasta, Limón, Pan | 1 |
| Pasta, Limón, Galletas | 4 |
| Pasta, Naranja, Pan | 1 |
| Pasta, Naranja, Galletas | 3, 4 |
| Limón, Naranja, Pan | 1 |
| Limón, Naranja, Galletas | 4 |

| Item (4) | TID |
|---------------------------------|-----|
| Pasta, Limón, Naranja, Pan | 1 |
| Pasta, Limón, Naranja, Galletas | 4 |

1 30-60 mins

Revisar la práctica realizada en clase con el ejemplo teórico y ambos algoritmos, APRIORI y ECLAT.

Ejemplo realizado en clase:

Inicialmente consideraremos el ejemplo realizado en clase para familiarizarnos con los comandos y sus opciones para luego aplicarlo en un problema más general. Para ello, primero definiremos la tabla de transacciones del problema:

```
In [8]: table <- list(c("p","l","0","b"), c("p","l"), c("p","0","c"), c("p","l","0","c"))
transactions <- as(table, "transactions")
inspect(transactions)

  items
[1] {b,l,0,p}
[2] {l,p}
[3] {c,0,p}
[4] {c,l,0,p}
```

Podemos ver el soporte, absoluto o relativo: de cada ítem:

```
In [9]: itemFrequency(transactions, type="a")
itemFrequency(transactions, type="r")
itemFrequencyPlot(transactions)
```