IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

Cierre de la clase 6 – GMM

Modelos de mezclas:

¿En qué se basa el principio de máxima verosimilitud?

¿Qué es una probabilidad *posterior* según Bayes?

GMM:

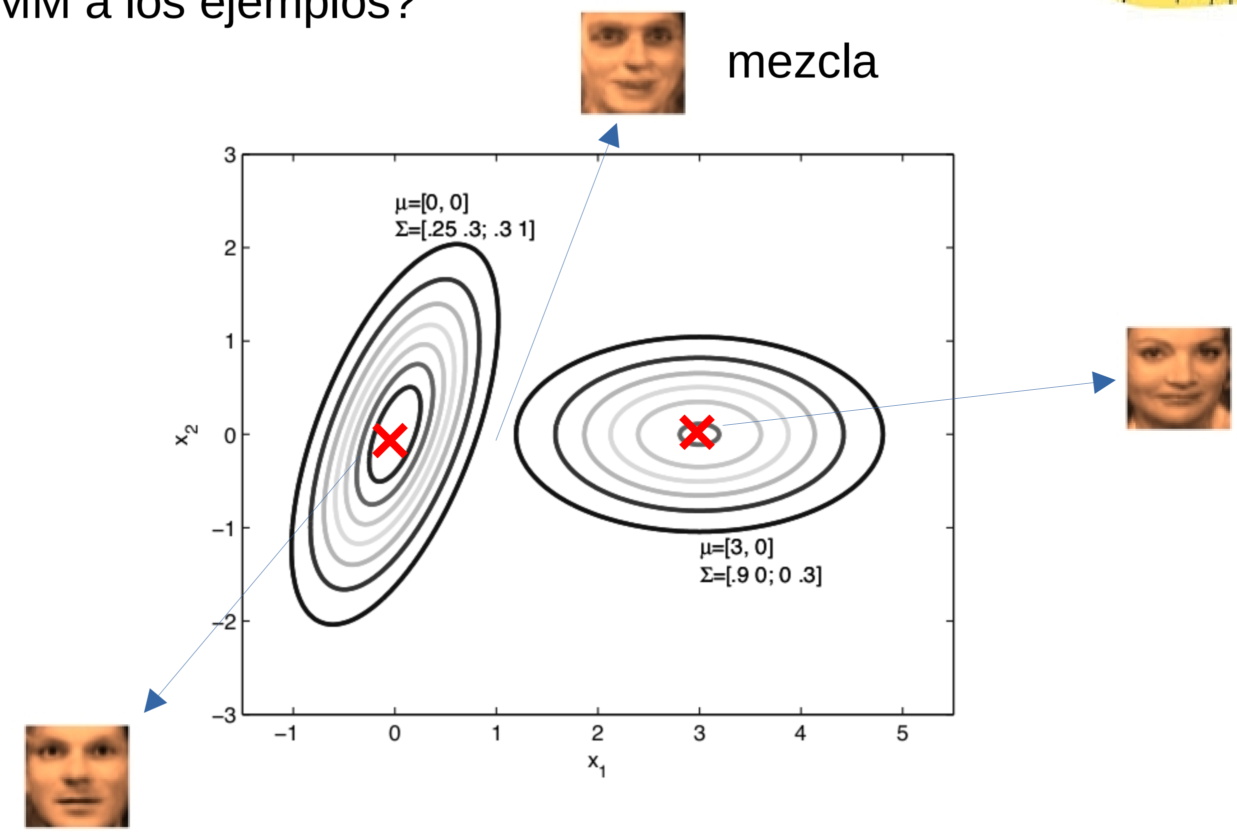
¿Qué modela la matriz de covarianza en una GMM? ¿Qué implica una entrada no diagonal?

¿Por qué el vector de medias puede aproximar un dato?

Si tengo 400 ejemplos ¿Cuántas Gaussianas pueden aproximar sin pérdida el dataset?

Cierre de la clase 6 – Actividad formativa

Si tengo dos datos ¿Cómo aproxima GMM a los ejemplos?



Dato: ✕

- Asociaciones -

Asociación

$$\{\texttt{Diapers}\} \longrightarrow \{\texttt{Beer}\}.$$

- Indica que si ocurre el antecedente, es probable que ocurra el consecuente.

- Usualmente involucran variables categóricas nominales.

- Se consolidan en reglas, las cuales usan {} para indicar asociaciones entre conjuntos denominados **itemsets**.

- En el ejemplo, ambos conjuntos son de cardinalidad 1.

- La asociación entre dos variables es un primer paso hacia la causalidad ya que define una relación de antecedente y consecuente.

- La diferencia con causalidad es que el antecedente es la causa y el consecuente es el efecto. Existe una relación lógica entre ambos que explica la observación.

- La asociación no necesariamente implica causa y efecto.

Reglas de Asociación

Datos:

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Regla: $\{\texttt{Diapers}\} \longrightarrow \{\texttt{Beer}\}$

Representación:

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Reglas de Asociación

Datos:

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Regla: $\{\texttt{Diapers}\} \longrightarrow \{\texttt{Beer}\}$

Representación:

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Items: $I = \{i_1, i_2, \ldots, i_d\}$; transacciones: $T = \{t_1, t_2, \ldots, t_N\}$

Soporte de un ítemset: $\sigma(X) = \left|\{t_i | X \subseteq t_i,\ t_i \in T\}\right|$

En cuántas transacciones ocurre X

Reglas de Asociación

Una regla de asociación es una implicancia $X \longrightarrow Y$ tal que $X \cap Y = \emptyset$.

Se definen:

$$\begin{aligned}\text{Support, } s(X \longrightarrow Y) &= \frac{\sigma(X \cup Y)}{N};\\ \text{Confidence, } c(X \longrightarrow Y) &= \frac{\sigma(X \cup Y)}{\sigma(X)}.\end{aligned}$$

\# transacciones

Reglas de Asociación

Una regla de asociación es una implicancia $X \longrightarrow Y$ tal que $X \cap Y = \emptyset$.

Se definen:

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N};$$

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

transacciones

Estrategia para generación de reglas:

1. Búsqueda de ítemsets frecuentes.
2. Generación de reglas candidatas a partir de ítemsets frecuentes.

<u>Buscar asociaciones puede ser muy ineficiente. La eficiencia de la búsqueda es un tema crítico en reglas de asociación.</u>

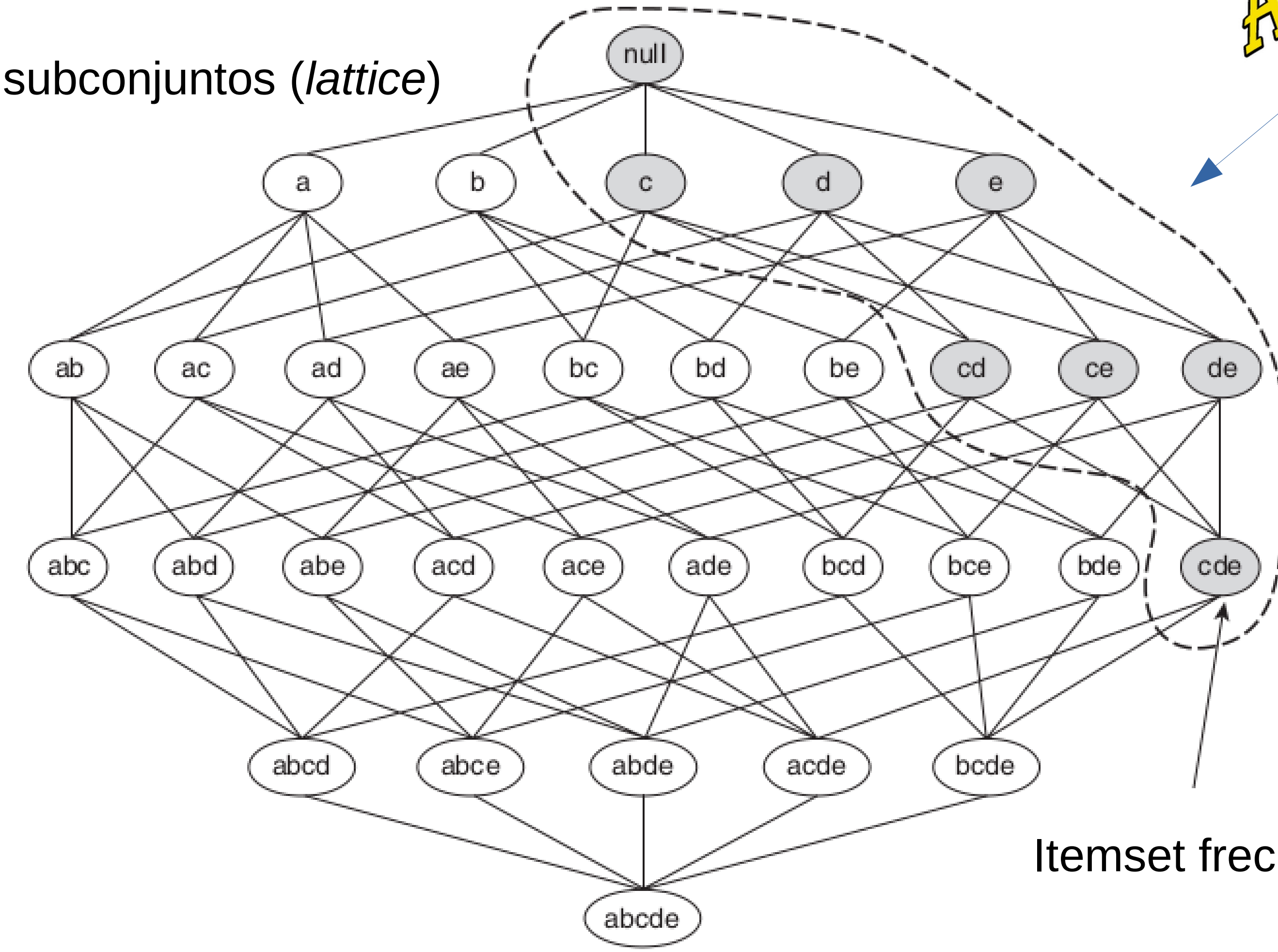
Principio Apriori

Si un itemset es frecuente, todos sus subconjuntos son frecuentes.

Reglas de Asociación

Principio Apriori:

Estructura de subconjuntos (*lattice*)



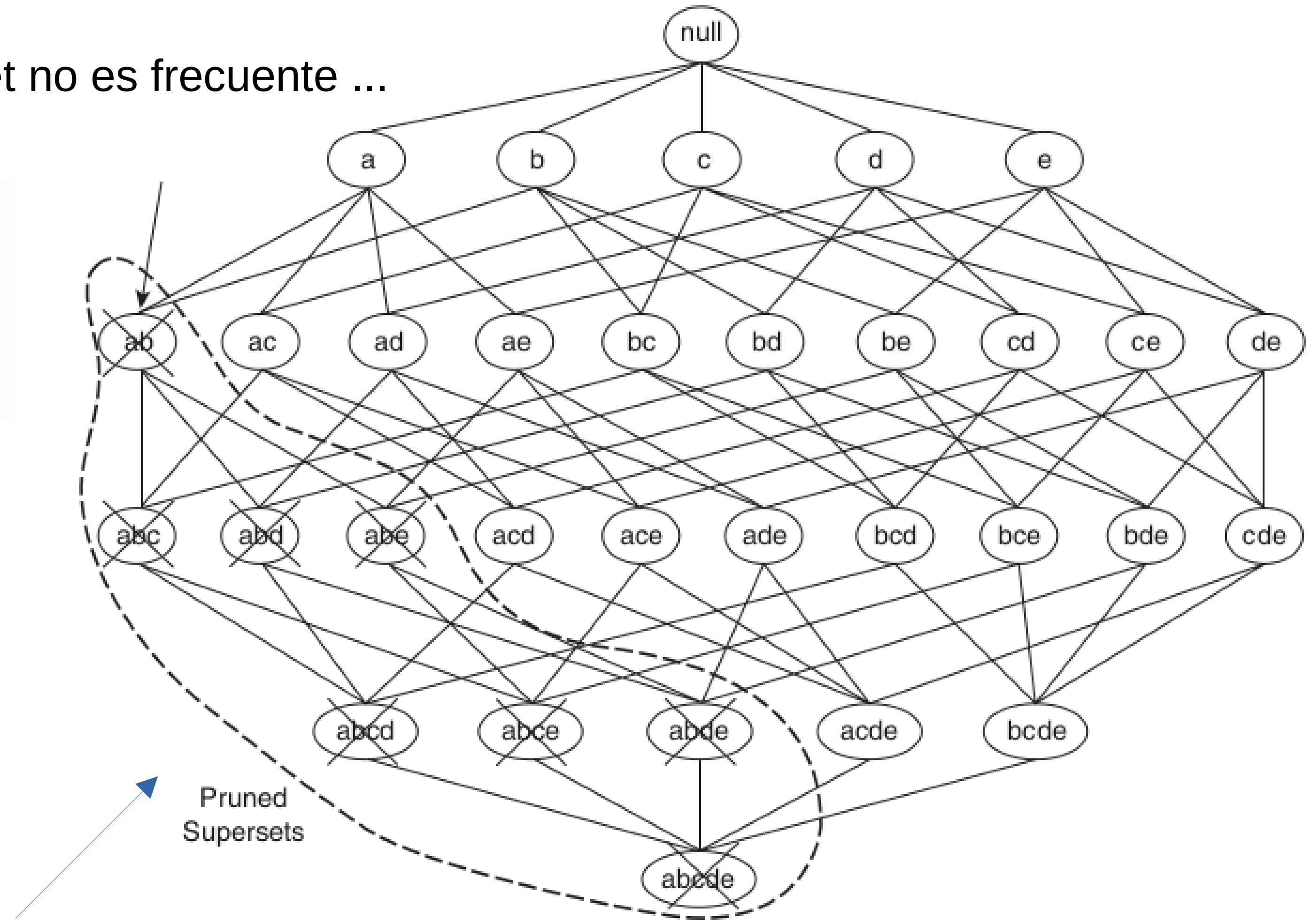
Itemset frecuente

Principio de monotonicidad: $\forall X, Y \in J: \ (X \subseteq Y) \longrightarrow f(Y) \leq f(X),$

Reglas de Asociación

Principio de monotonicidad: $\forall X, Y \in J : \ (X \subseteq Y) \longrightarrow f(Y) \leq f(X),$

Si este itemset no es frecuente ...

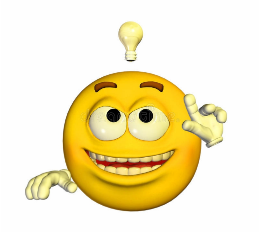


… estos tampoco son frecuentes.

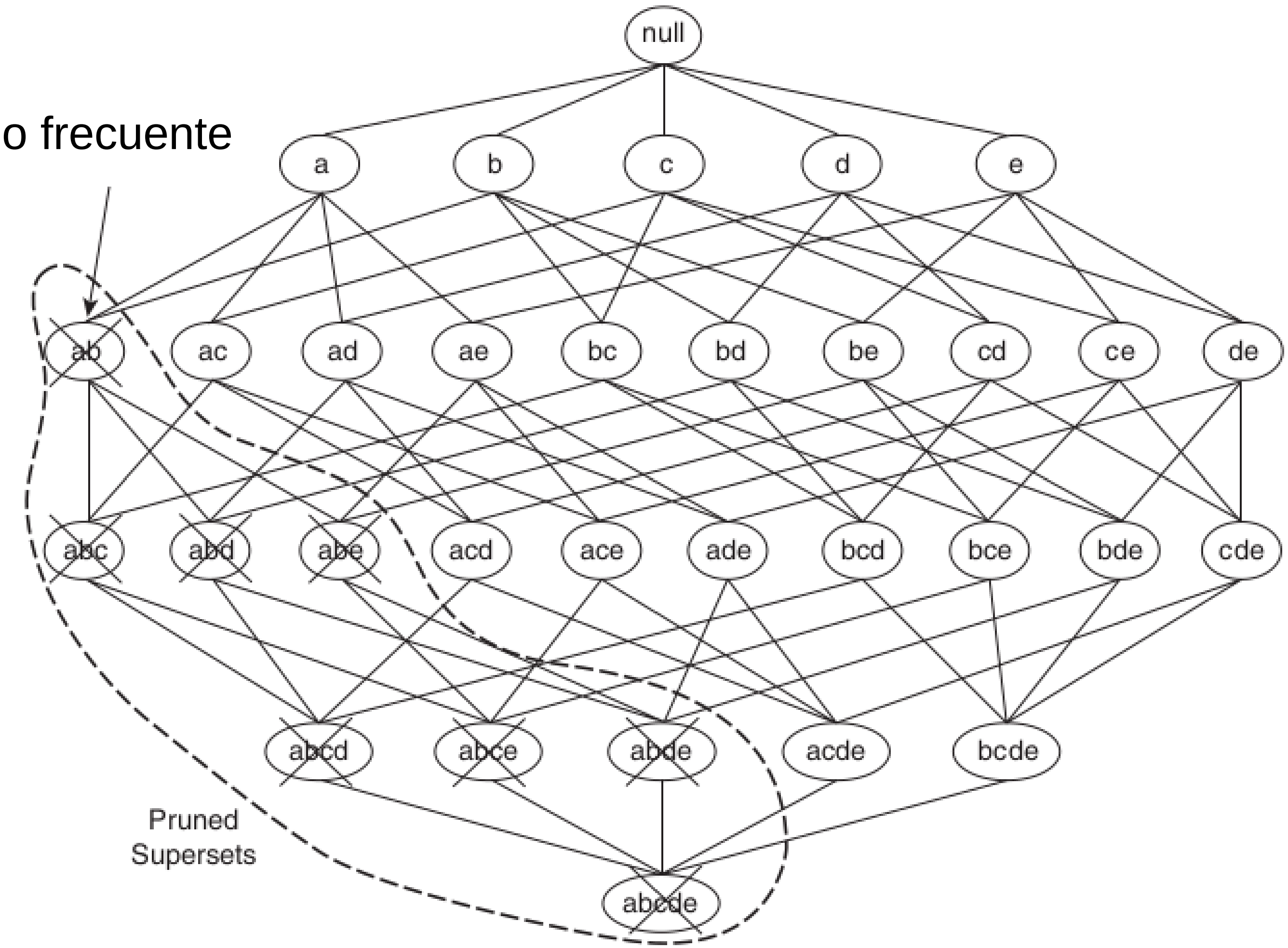
Reglas de Asociación

Principio de monotonicidad: $\forall X, Y \in J: \ (X \subseteq Y) \longrightarrow f(Y) \leq f(X),$

Itemset no frecuente



Podemos podar



Reglas de Asociación

Uso algorítmico del principio de monotonicidad: Buscamos por orden creciente de cardinalidad:

Candidate 1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Minimum support count = 3

Candidate 2-Itemsets

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

Itemsets removidos

Candidate 3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	3

Reglas de Asociación

Uso algorítmico del principio de monotonicidad: Buscamos por orden creciente de cardinalidad:

Candidate 1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Minimum support count = 3

Candidate 2-Itemsets

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

Itemsets removidos

Candidate 3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	3

Reglas de Asociación

<u>Ahora vamos a generar reglas a partir de itemset frecuentes</u>.

Recordar:

$$\text{Support}, \; s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N};$$

$$\text{Confidence}, \; c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Si una regla $X \longrightarrow Y - X$ no satisface un umbral de confianza, entonces toda regla de la forma $X' \longrightarrow Y - X'$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza.

Reglas de Asociación

<u>Ahora vamos a generar reglas a partir de itemset frecuentes</u>.

Recordar:

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N};$$

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

> Si una regla $X \longrightarrow Y - X$ no satisface un umbral de confianza, entonces toda regla de la forma $X' \longrightarrow Y - X'$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza.

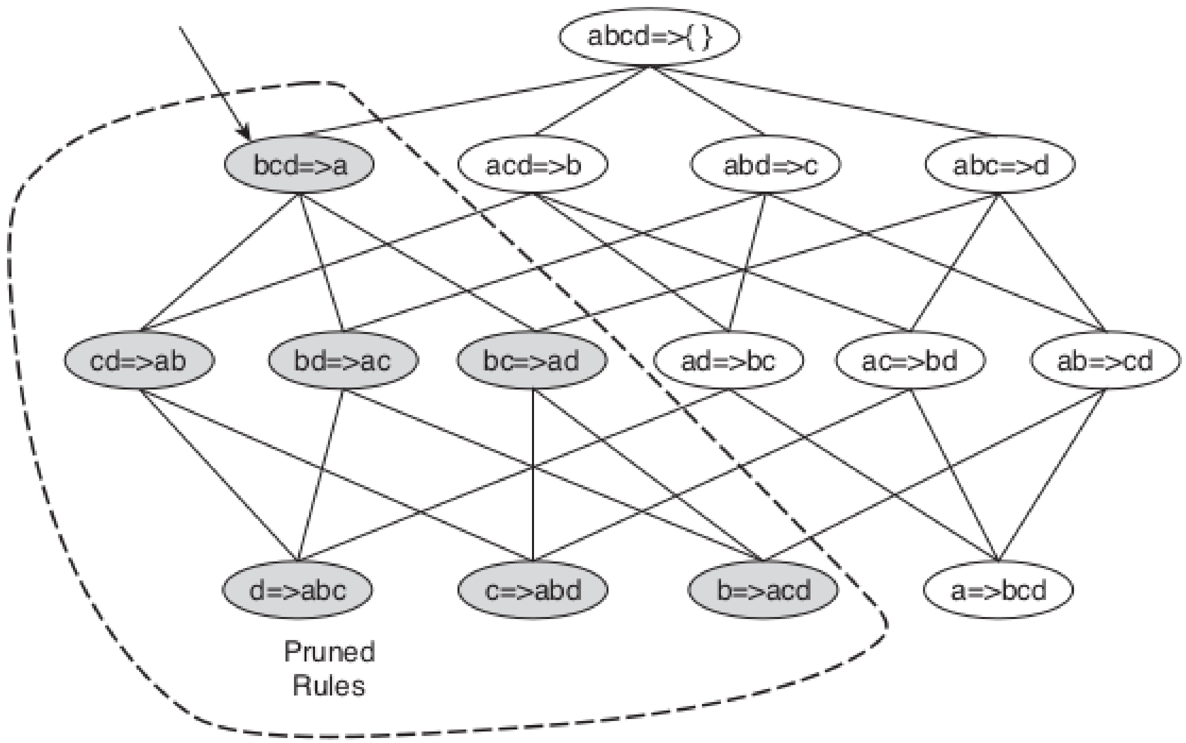
Principio de monotonicidad: $\forall X, Y \in J : (X \subseteq Y) \longrightarrow f(Y) \leq f(X),$

Dado que $\tilde{X} \subseteq X$, $\sigma(X) \leq \sigma(X')$, luego $\dfrac{\sigma(X \cup Y)}{\sigma(X)} \geq \dfrac{\sigma(X \cup Y)}{\sigma(X')}$.

Reglas de Asociación

Si una regla $X \longrightarrow Y - X$ no satisface un umbral de confianza, entonces toda regla de la forma $X' \longrightarrow Y - X',$ donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza.

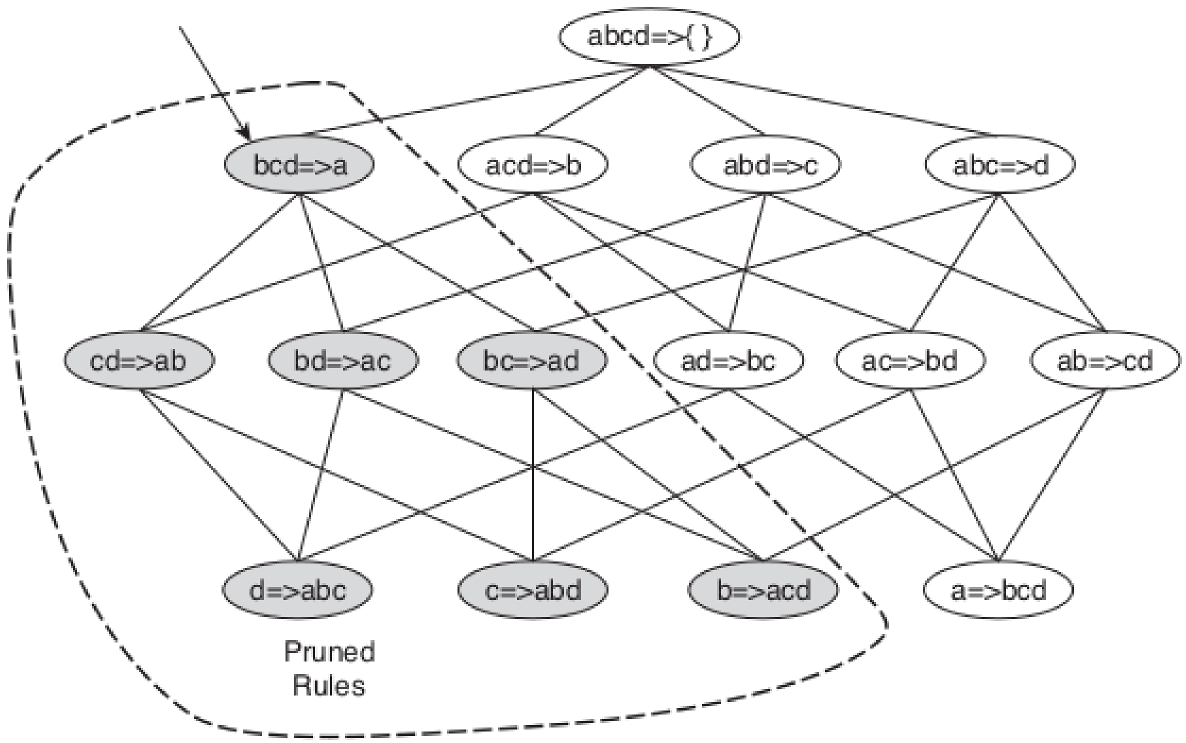
Regla de baja confianza



Reglas de Asociación

Si una regla $X \longrightarrow Y - X$ no satisface un umbral de confianza, entonces toda regla de la forma $X' \longrightarrow Y - X'$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza.

Regla de baja confianza

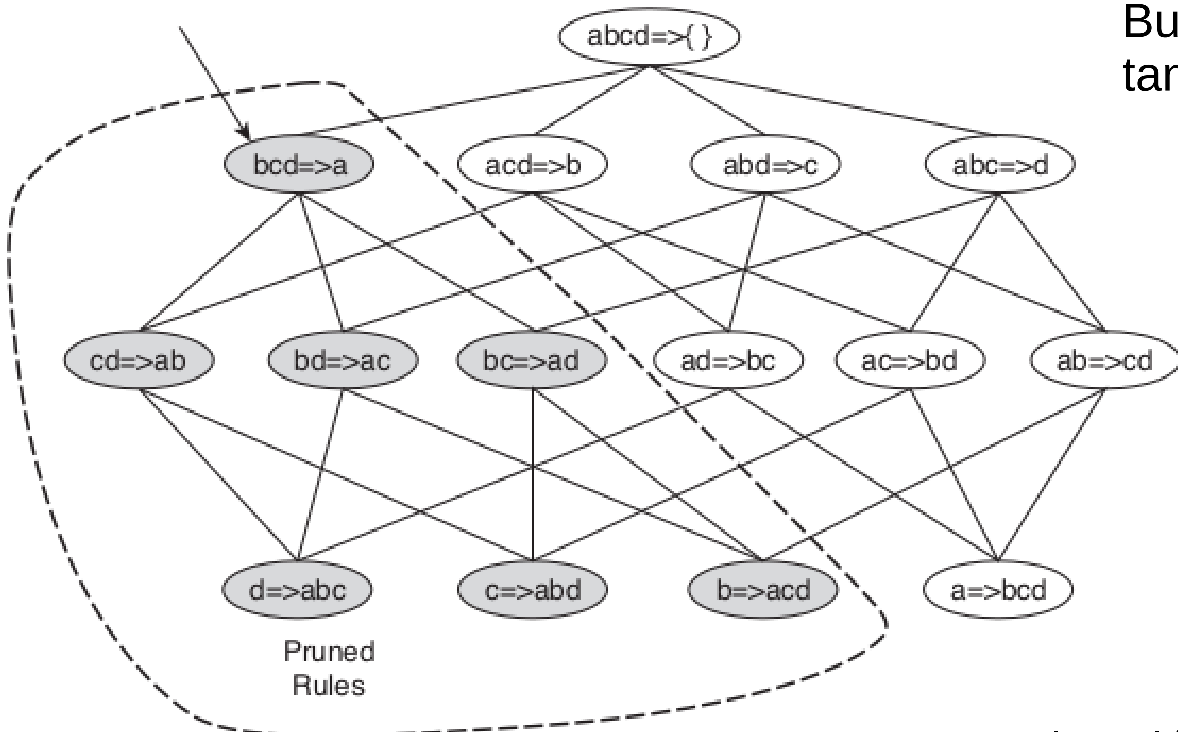


Podemos podar

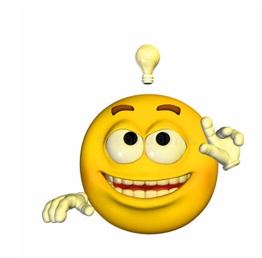
Reglas de Asociación

Si una regla $X \longrightarrow Y - X$ no satisface un umbral de confianza, entonces toda regla de la forma $X' \longrightarrow Y - X'$, donde $\tilde{X} \subseteq X$, no puede satisfacer el umbral de confianza.

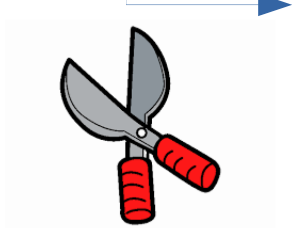
Regla de baja confianza



Buscar aumentando el tamaño del consecuente



Podemos podar



Dirección de búsqueda

- ANEXO -

Reglas de Asociación

Para k = 2 no necesitamos algoritmo

Algorithm Rule generation of the *Apriori* algorithm.

for each frequent k-itemset f_k, $k \geq 2$ **do**
  $H_1 = \{i \mid i \in f_k\}$ {1-item consequents of the rule.}
  call `ap-genrules`$(f_k, H_1.)$
end for

Algorithm Procedure `ap-genrules`$(f_k,\ H_m)$.

$k = |f_k|$ {size of frequent itemset.}
$m = |H_m|$ {size of rule consequent.}
if $k > m + 1$ **then** El antecedente tendrá al menos un ítem
  $H_{m+1} = \text{apriori-gen}(H_m)$. Aumento el consecuente en 1
  for each $h_{m+1} \in H_{m+1}$ **do**
    $conf = \sigma(f_k)/\sigma(f_k - h_{m+1})$. $\frac{\sigma(X \cup Y)}{\sigma(X)}$.
    if $conf \geq minconf$ **then**
      output the rule $(f_k - h_{m+1}) \longrightarrow h_{m+1}$.
    else
      delete h_{m+1} from H_{m+1}.
    end if
  end for
  call ap-genrules$(f_k, H_{m+1}.)$ Una vez terminamos la verificación, continuamos recursivamente.
end if

Termina cuando H_{m+1} es vacío.