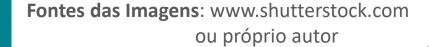
Regras de Associação

Prof. Sandro Jerônimo





### Regras de Associação

#### **Objetivo**

Encontrar padrões frequentes, associações, correlações entre conjunto de itens ou objetos de um banco de dados transacional, banco de dados relacional ou outro repositório de informação.

#### **Aplicações**

- Análise de cestas de compras, marketing, portfólio de produto/serviços, etc.
- Quais subsequentes compras após ter comprado um PC?
- Qual tipo de DNA é sensitivo a uma nova droga?
- Turistas que optam em viajar para os EUA também aceitariam viajar para a Europa

# Regras de Associação - Definições

#### Formato das Regras

A→B

- $A \in B \rightarrow C$
- $A \rightarrow B e C$

 $B \rightarrow A$ 

- . . .
- A→B <u>significa</u>: se A acontece em uma transação então B também acontece

#### **Exemplos**

Quem compra computador (A) compra software (B)

 $A \rightarrow B$ 

Quem compra computador (A) e software (B) compra webcam (W)

- $A e B \rightarrow W$
- Quem opta por ir ao cinema (C) também compra refrigerante (R) e pipoca (P)  $C \rightarrow R e P$

## Regras de Associação - Estatísticas

#### As regras possuem medidas estatísticas

Suporte: probabilidade dos elementos acontecerem

$$sup(A \to B) = \frac{n\'umero\ de\ transa\~c\~oes\ com\ A\ e\ B}{n\'umero\ total\ de\ transa\~c\~oes}$$

• Confiança: confiança em que um elemento implica a outro

$$conf(A \rightarrow B) = \frac{n \'umero\ de\ transa\~c\~oes\ que\ suportam\ (A \cup B)}{n \'umero\ de\ transa\~c\~oes\ que\ suportam\ A}$$

As medidas refletem a utilidade e o grau de certeza das regras. As regras que satisfazem aos dois valores mínimos são consideradas **fortes**.

• **EXEMPLO**: suponha a informação de clientes que compram computadores e tendem a comprar softwares de edição de texto. Na análise as medidas foram : suporte 2% e confiança 60%. O suporte significa que 2% de todas transações analisadas mostram que computador e software foram comprados juntos. Confiança de 60% indica que 60% dos clientes que compram computador também compram software.

### **Identificando Regras Fortes**



ID Transação	Itens das compras
2000	<b>A, B, C</b>
1000	A, C
4000	A, D
5000	B, E, F

# Usuário deve definir o suporte e confiança mínima. Exemplo:

Suporte: 50%

Confiança: 50%

#### Teremos as seguintes regras

•  $A \rightarrow C$ 

Suporte: 2/4 = 0.5 = 50%

Confiança: 2/3 = 0.66 = 66.6%

•  $C \rightarrow A$ 

Suporte: 2/4 = 0.5 = 50%

Confiança: 2/2 = 1 = 100 %

Produto	Núm. do Produto
Pão	1
Açúcar	2
Leite	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
logurte	9
Suco	10

# **Exemplo - Compras**

Num transação	Itens comprados
T1	{1,3,5}
T2	{2,1,3,7,5}
Т3	{4,9,2,1}
T4	{5,2,1,3,9}
T5	{1,8,6,4,3,5}
Т6	{9,2,8}



Um mercado planeja criar espaços para oferta "casada" de produtos. Quais de produtos fariam parte desses espaços assumindo um suporte de 50%?

# Alguns Itemset selecionados

Num transação	Itens comprados
T1	{1,3,5}
T2	{2,1,3,7,5}
Т3	{4,9,2,1}
T4	{5,2,1,3,9}
T5	{1,8,6,4,3,5}
Т6	{9,2,8}

Produto	Núm. do Produto	
Pão	1 (apareceu 5x)	
Açúcar	2 (apareceu 4x)	
Leite	3 (apareceu 4x)	
Manteiga	5 (apareceu 4x)	
logurte	9 (apareceu 3x)	

Itemset	Suporte (>=50%) e Confiança
{Pão, Leite} {1,3}	Suporte = 4/6 = 66,6% Confiança (Pão → Leite) = 80% Confiança (Leite → Pão) = 100 %
{Pão, Manteiga} {1,5}	Suporte = 4/6 = 66,6% Confiança (Pão → Mant.) = 80% Confiança (Mant. → Pão) = 100%
{Leite, logurte} {2,9}	Suporte = 3/6 = 50% Confiança (Leite → logurte) = 75% Confiança (logurte → Leite) = 100%
{Pão, Leite, Manteiga} {1, 3, 5}	Suporte = $3/6 = 50\%$ Confiança (Pão $\rightarrow$ Leite, Mant) = $60\%$ Confiança (Leite $\rightarrow$ Pão, Mant) = $100\%$ Confiança (Mant $\rightarrow$ Pão, Leite) = $60\%$ Confiança (Pão, Leite $\rightarrow$ Mant) = $75\%$ Confiança (Pão, Mant $\rightarrow$ Leite) = $75\%$ Confiança (Leite, Mant $\rightarrow$ Pão) = $75\%$

# Algoritmo Apriori

O algoritmo mais usado para encontrar regras de associação. Baseia-se no fato usar conhecimento já obtidos dos *itemsets* anteriores.

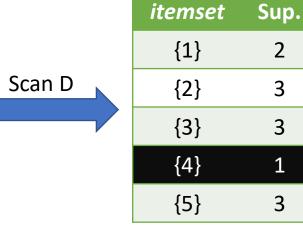
- Fase I Descobre todos os conjuntos de itens com suporte maior ou igual ao mínimo suporte especificado pelo usuário.
- Fase II A partir dos conjuntos de itens frequentes, descobre regras de associação com fator de confiança maior ou igual ao especificado pelo usuário.

### Algoritmo *Apriori – Ilustração*

*Suporte mínimo = 2 (50%)* 

Base de Dados D

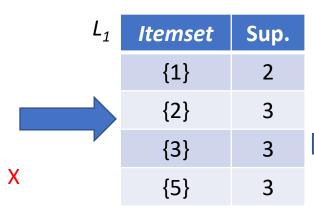
TID	Itens
100	134
200	235
300	1235
400	2 5



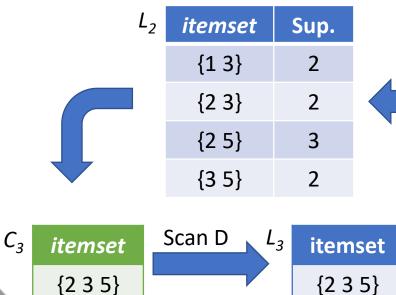
 $C_2$ 

Sup.

2



Scan D



$C_2$	itemset	Sup.
X	{1 2}	1
ı	{1 3}	2
X	{1 5}	1
	{2 3}	2
	{2 5}	3
	{3 5}	2

$C_2$	itemset
	{1 2}
)	{1 3}
	{1 5}
	{2 3}
	{2 5}
	{3 5}

PUC Minas Virtual

#### Existem outras medidas de interesse

*Lift*: indica a força de uma regra sobre a coocorrência aleatória de seus antecedentes e consequentes.

$$lift(A \to B) = \frac{sup(A \to B)}{sup(A) \times sup(B)}$$

Convicção: Assim como a confiança, é sensível à direção da regra.

$$conv(A \to B) = \frac{1 - \sup(B)}{1 - \operatorname{conf}(A \to B)}$$

**Ganho**: Ganho é calculado baseado em um valor theta ( $\theta$ ) dado. Usualmente  $\theta = 2.0$   $ganho(A \rightarrow B) = \sup(A \cup B) - \theta * \sup(A)$ 

**Laplace**: Laplace é calculado baseado em um parâmetro k. Usualmente k=1.0.

$$laplace(A \to B) = \frac{\sup(A \cup B) + 1}{\sup(A) + k}$$

Piatesky-Shaprio (P-S):

$$ps(A \rightarrow B) = \sup(A \cup B) - \sup(A) * \sup(B)$$

