



UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS  
CAMPUS BARBACENA

Bacharelado em Ciência da Computação



---

# *Mineração de dados*

## **Material de Apoio**

### *Parte VI – Regras de Associação*

Prof. José Osvano da Silva, PMP, PSM I  
joseosvano@unipac.br

*2º sem / 2023*

*Material cedido pela Profª Livia  
e Profº Osvano*

# Sumário

---

- Regras de Associação
- Aplicações mais comuns
- Exemplo
- Exemplo no Knime
- Exercício

## Regras de associação

---

- Descoberta de regras de associação é o processo de analisar os relacionamentos existentes entre atributos de uma base de dados, com o objetivo de encontrar associações ou correlações.
- A existência de associações ou a correlação entre os atributos implica que eles frequentemente aparecem juntos em uma transação.

## Regras de associação

---

- As regras de associação são comumente representadas por meio de afirmações do tipo SE ENTÃO, sendo também interpretadas como implicações do antecedente da regra (ou premissa) para o seu consequente (ou conclusão).

## Aplicações mais comuns

---

- Análise de compras (market basket analysis)
- Análise de perfis de clientes
  - SE o cliente compra determinado produto (ou subconjunto de produtos), ENTÃO o cliente também compra outro produto (ou outro subconjunto)
- Análise de mercado de ações
  - Associações a acontecimentos mundiais

## Aplicações mais comuns

---

- Análise de desempenho físico
  - Resultados de treinamento se associam a condições corporais
- Análise de comportamento eleitoral

## Regras de associação

---

- Os resultados obtidos da descoberta de regras de associação são considerados de fácil interpretação
  - Regras são expressas em “linguagem natural” e a semântica é explícita
- Grande maioria dos resultados das tarefas de mineração de dados não tem essa facilidade de avaliação, precisando realizar um processo de pós-processamento para interpretação.

## Regras de associação

---

- Uma vez mineradas as regras de associação, elas podem ser utilizadas para melhorar processos inerentes ao domínio estudado.
  - Ex: reorganização dos balcões expositores de um restaurante a partir das regras extraídas, seja colocando próximos os itens relacionadas nas regras (de forma a estimular o consumo dos mesmo); ou colocando-os longe, de forma que os clientes procurem por eles e acabem passando por diversos outros pratos (estimulando a experimentação de outros).
  - Ações de marketing



## Regras de associação

---

- Os critérios de qualidade ou avaliação do resultado estão incorporados no processo de execução, diferentemente de outros algoritmos, em que é necessário aplicar esses critérios após a execução.

# Regras de associação

---

- Um conjunto de dados transacionais é composto de vários exemplares de dados.
- Em regras de associação
  - Cada exemplar refere-se a uma transação realizada (ou evento ocorrido)
  - Cada uma das transações é composta por uma série de itens (ou elementos)

# Regras de associação

---

- Formalmente, em um domínio de aplicação, existe um conjunto de itens do domínio  $I = \{i_1, \dots, i_m\}$ .
- Uma transação  $T$  é composta pela ocorrência de um subconjunto desses itens, ou seja,  $T = \{i_1, \dots, i_l\}$ , tal que  $i_l \subset I$  e  $l \leq m$ .

# Exemplo de base de dados transacional

---

$T_1 = \{i_1, i_3\}$   
 $T_2 = \{i_1, i_2, i_3\}$   
 $T_3 = \{i_1\}$   
 $T_4 = \{i_1, i_3, i_4\}$   
 $T_5 = \{i_3\}$   
 $T_6 = \{i_1, i_2\}$   
 $T_7 = \{i_1, i_2, i_3, i_4\}$

$T_{ID}$	$i_1$	$i_2$	$i_3$	$i_4$
$T_1$	1	0	1	0
$T_2$	1	1	1	0
$T_3$	1	0	0	0
$T_4$	1	0	1	1
$T_5$	0	0	1	0
$T_6$	1	1	0	0
$T_7$	1	1	1	1

## Regras de associação

---

- Por simplicidade, em relação a representação matricial, o valor 0 na base de dados indica que o item não ocorre na transação, e o valor 1 indica que o item ocorre.
- *Itemset*: conjunto de itens, ou subconjuntos de itens do domínio. É composto por  $k$  itens ( $k$ -*itemset*).
  - Exemplo:  $\{i_1, i_3\}$  é um 2-*itemset*

## Regras de associação

---

- Uma regra de associação é do tipo  $A \rightarrow B$ , sendo que  $A$  e  $B$  são, ambos, itemsets compostos de itens pertencentes a  $I$ , e  $A \cap B = \emptyset$ .
  - Exemplo:  $\{i_2\} \rightarrow \{i_1, i_3\}$

## Regras de associação

---

- Suporte: refere-se à frequência do itemset na base transacional sob análise, ou seja, o suporte de um itemset é a frequência com que os itens que o compõem aparecem juntos em transações individuais da base de dados; geralmente é expressa em termos percentuais.
  - Exemplo: o itemset  $\{i_1, i_3\}$  tem suporte = 57% (4/7), ou seja, os itens  $i_1$  e  $i_3$  aparecem em 4 das 7 transações

## Regras de associação

---

- Quando o suporte é aplicado à uma regra diz respeito à frequência com que todos os dois itemsets envolvidos na regra (A e B) aparecem juntos em transações individuais na base de dados. Ou seja, o suporte de uma regra do tipo  $A \rightarrow B$  é o suporte do itemset  $A \cup B$ .

$$\text{suporte}_{\text{regra}}(A \cup B) = \frac{\text{cont}(A \cup B)}{\text{cont}(T)}$$

- Exemplo: o suporte da regra  $i_2 \rightarrow \{i_1, i_3\}$  é o suporte do itemset  $\{i_1, i_2, i_3\}$



# Regras de associação

---

- Equação 1:
  - O suporte de uma regra  $X \rightarrow Y$ , onde X e Y são conjuntos de itens, é dado pela seguinte fórmula:

$$\text{Suporte} = \frac{\text{Frequência de X e Y}}{\text{Total de T}}$$

- O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.

# Regras de associação

---

- Confiança: medida aplicada somente às regras, e objetiva expressar uma noção da importância e da confiabilidade de uma regra, dada a possibilidade de sua ocorrência.
- É geralmente expressa por meio de percentual e dada pela razão entre o suporte da regra e o suporte da premissa da regra (parte inicial da regra).

$$\text{confiança}_{\text{regra}}(A \cup B) = \frac{\text{suporte}_{\text{regra}}(A \cup B)}{\text{suporte}_{\text{itemset}}(A)}$$

## Regras de associação

---

- Equação 2:
  - A sua confiança é dada pela seguinte fórmula:

$$\text{Confiança} = \frac{\text{Frequência de X e Y}}{\text{Frequência de X}}$$

- O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente.
- O denominador se refere à quantidade de transações em que o item X ocorre.

# Regras de associação

---

- Exemplo: para a regra  $i_2 \rightarrow \{i_1, i_3\}$  a confiança é 65% (28/43)  
suporte da regra  $A \cup B = 28$  (itemset  $\{i_1, i_2, i_3\}$  aparece 2 das 7 transações)  
suporte da premissa  $A = 43$  (itemset  $\{i_2\}$  aparece 3 das 7 transações)
- O significado é “em 65% das vezes em que a premissa da regra ocorre, a conclusão também ocorre”, ou seja, a confiança pode ser entendida como a probabilidade de a regra  $A \rightarrow B$  ocorrer, dado que sua premissa  $A$  ocorre.

# Regras de associação

---

- Um itemset é considerado frequente quando seu suporte satisfaz a um suporte mínimo (um número entre 0 e 100, em termos percentuais) pré-estabelecido.
- Quem determina o suporte o suporte mínimo é o usuário, e essa determinação é dependente do contexto, não havendo uma regra pré-definida para seu estabelecimento (é um parâmetro a ser trabalhado).

## Regras de associação

---

- O suporte (Equação 1) pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X quanto Y;
- A confiança (Equação 2) é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X.

## Exemplo

- Dada a tabela abaixo onde cada registro corresponde a uma transação de um cliente, com itens assumindo valores binários (sim/não), indicando se o cliente comprou ou não o respectivo item, descobrir todas as regras associativas com suporte  $\geq 0,3$  (ou 30%) e grau de certeza (*confiança*)  $\geq 0,8$  (ou 80%).

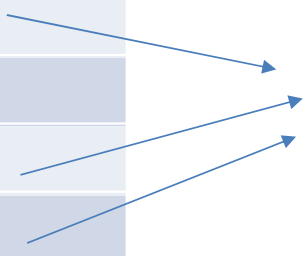
TID	leite	café	cerveja	pão	manteiga	arroz	feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

# Exemplo

---

- Calcular o suporte de conjuntos com um item.  
Determinar os itens frequentes com  $sup \geq 0,3$ .

Conjunto de itens	Suporte
{leite}	20%
{café}	30%
{cerveja}	20%
{pão}	50%
{manteiga}	50%
{arroz}	20%
{feijão}	20%



Maiores que 30%



# Exemplo

---

- Calcular o suporte de conjuntos com dois itens. Determinar conjuntos de itens frequentes com  $sup \geq 0,3$ . Obs: se um item não é frequente no primeiro passo, pode ser ignorado aqui.

Conjunto de itens	Suporte
{café, pão}	30%
{café, manteiga}	30%
{pão, manteiga}	40%

# Exemplo

---

- Calcular o suporte de conjuntos com três itens. Determinar conjuntos de itens frequentes com  $sup \geq 0,3$ .

Conjunto de itens	Suporte
{café, pão, manteiga}	30%

# Exemplo

---

- Regras candidatas com dois itens com o seu valor de certeza:

Conjunto de itens: {café, pão}

**Se** café **Então** pão *conf* = 1,0

**Se** pão **Então** café *conf* = 0,6

Conjunto de itens: {café, manteiga}

**Se** café **Então** manteiga *conf* = 1,0

**Se** manteiga **Então** café *conf* = 0,6

Conjunto de itens: {pão, manteiga}

**Se** pão **Então** manteiga *conf* = 0,8

**Se** manteiga **Então** pão *conf* = 0,8

# Exemplo

---

- Regras candidatas com três itens com o seu valor de certeza:

Conjunto de itens: {café, manteiga, pão}

<b>Se</b> café, manteiga <b>Então</b> pão	<i>conf</i> = 1,0
<b>Se</b> café, pão <b>Então</b> manteiga	<i>conf</i> = 1,0
<b>Se</b> manteiga, pão <b>Então</b> café	<i>conf</i> = 0,75
<b>Se</b> café <b>Então</b> manteiga, pão	<i>conf</i> = 1,0
<b>Se</b> manteiga <b>Então</b> café, pão	<i>conf</i> = 0,6
<b>Se</b> pão <b>Então</b> café, manteiga	<i>conf</i> = 0,6

# Regras de associação

---

- Existem outros indicadores que podem nos auxiliar quando tratamos de regras de associação, são eles:
  - Lift
  - Leverage
  - Conviction
  - Zhangs

# Regras de associação

---

- Lift

- A métrica "lift" é uma medida que é comumente usada na mineração de regras de associação para avaliar o grau de associação entre os itens em um conjunto de dados. Ela indica o quão mais provável é a ocorrência conjunta de dois itens em comparação com o que seria esperado se eles fossem independentes. O lift é usado para determinar o quão forte é uma regra de associação.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Suporte Conjunto}(X, Y)}{(\text{Suporte}(X) * \text{Suporte}(Y))}$$

# Regras de associação

---

- Lift
  - Se o lift é igual a 1, isso indica que não há associação entre os itens X e Y, ou seja, eles são independentes.
  - Se o lift é maior que 1, isso sugere uma associação positiva entre os itens X e Y. Quanto maior o valor do lift, mais forte é a associação.
  - Se o lift é menor que 1, isso sugere uma associação negativa ou uma associação fraca entre os itens X e Y. Valores menores que 1 indicam que a ocorrência conjunta é menos provável do que o esperado na ausência de associação.

# Regras de associação

---

- Leverage

- A métrica "leverage," também conhecida como "lift" bruto ou "razão de ganho," é uma medida utilizada na mineração de regras de associação para avaliar a dependência entre dois itens em um conjunto de dados. O leverage mede a diferença entre a frequência observada de coocorrência de dois itens e a frequência esperada se eles fossem independentes.

$$\text{Leverage}(X \rightarrow Y) = \text{Suporte Conjunto}(X, Y) - (\text{Suporte}(X) * \text{Suporte}(Y))$$



# Regras de associação

---

- Leverage

- Se o leverage é igual a 0, isso indica que não há dependência entre os itens X e Y, ou seja, eles são independentes.
- Se o leverage é maior que 0, isso sugere uma associação positiva entre os itens X e Y. Quanto maior o valor do leverage, mais a coocorrência de X e Y é observada do que seria esperado se fossem independentes.
- Se o leverage é menor que 0, isso sugere uma associação negativa entre os itens X e Y. Valores negativos indicam que a coocorrência de X e Y é menos frequente do que seria esperado se fossem independentes.

# Regras de associação

---

- Conviction
  - A métrica "conviction" (convicção) é uma medida utilizada na mineração de regras de associação para avaliar o grau de dependência entre dois itens em um conjunto de dados. A convicção é uma métrica que se concentra na probabilidade de que o item consequente (Y) de uma regra de associação seja comprado ou ocorra na ausência do item antecedente (X).

$$\text{Conviction (X} \rightarrow \text{Y)} = \frac{(\text{Suporte(X)} * (1 - \text{Suporte(Y)}))}{(1 - \text{Suporte Conjunto(X, Y)})}$$

# Regras de associação

---

- Conviction

- Se a convicção é igual a 1, isso indica que não há dependência entre os itens X e Y. Nesse caso, a compra de X não afeta a probabilidade de compra de Y, e vice-versa.
- Se a convicção é maior que 1, isso sugere uma dependência positiva entre os itens X e Y. Quanto maior o valor da convicção, mais forte é a associação entre eles. Isso significa que a compra de X aumenta a probabilidade de compra de Y, e vice-versa.
- Se a convicção é menor que 1, isso sugere uma dependência negativa entre os itens X e Y. Valores menores que 1 indicam que a compra de X reduz a probabilidade de compra de Y, e vice-versa.

# Regras de associação

---

- Zhang
  - A métrica "Zhang's metric," ou métrica de Zhang, é uma medida utilizada na mineração de regras de associação para avaliar a força de uma associação entre dois itens em um conjunto de dados. Essa métrica é uma alternativa ao lift e à convicção, e foi projetada para superar algumas limitações dessas métricas.

# Regras de associação

---

- Zhang
  - Se a métrica de Zhang é igual a 0, isso indica que não há associação entre os itens X e Y, ou seja, eles são independentes.
  - Se a métrica de Zhang é maior que 0, isso sugere uma associação positiva entre os itens X e Y. Quanto maior o valor da métrica de Zhang, mais forte é a associação.
  - Se a métrica de Zhang é menor que 0, isso sugere uma associação negativa entre os itens X e Y. Valores negativos indicam que a coocorrência de X e Y é menos frequente do que o esperado se fossem independentes.

# Regras de associação

---

- Algoritmos
  - Apriori: O algoritmo Apriori é um dos mais conhecidos e amplamente utilizados para mineração de regras de associação. Ele é eficiente para conjuntos de dados de tamanho moderado e identifica regras de associação frequentes com base em um limiar de suporte mínimo.

# Regras de associação

---

- Algoritmos
  - FP-Growth (Frequent Pattern Growth): O algoritmo FP-Growth é uma alternativa ao Apriori que usa uma estrutura de árvore compacta para encontrar padrões frequentes. É eficiente e geralmente mais rápido do que o Apriori em conjuntos de dados grandes.

# Regras de associação

---

- Algoritmos
  - FPMMax: O FPMMax é uma variação do FP-Growth que encontra apenas os conjuntos de itens frequentes máximos, ou seja, conjuntos de itens que não são subconjuntos de outros conjuntos frequentes.



# Exemplo

---

Usando a base de dados “associacao.csv”, os conhecimentos adquiridos crie uma mineração de dados para {pão, manteiga} usando o Python; Encontre o suporte e a confiança.

# Exercícios

---

Usando a base de dados “market.csv”, os conhecimentos adquiridos crie uma mineração de dados usando o Python; Encontre o suporte e a confiança e destaque baseando nesses critérios as 20 mais significantes.

# Referências

---

- RUSSEL, S., NORVIG, P. *Inteligência Artificial*, Editora Campus, 2ª. edição.
- OLIVEIRA, S. R. M. *Introdução à mineração de dados*, Material para aulas, 2012.
- ZARATE, L.E. *Descoberta de Conhecimento em Banco de Dados e Data Mining*, Material para aulas, 2008.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. *From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, MIT, 1996.
- SILVA, L. A. *Introdução à mineração de dados com aplicações em R*, Elsevier, 1ª ed, 2016.