UFABC

UFABC

Tecnologias e Sistemas de Informação

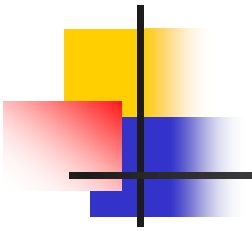
Sistemas Inteligentes

Aula 2 – Mineração de Dados e Regras de Associação

Prof. josé artur quilici-gonzalez
Email: jose.gonzalez@ufabc.edu.br

Roteiro

- Introdução
- Mineração de Dados
- Regras de Associação
- Geração de Conjuntos Frequentes
- Geração de Regras de Associação
- Referência Bibliográfica



INTRODUÇÃO

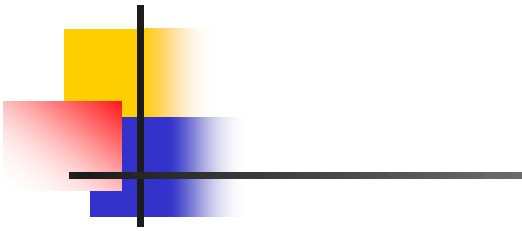
Analogia da Cesta de Compras

A **Mineração de Dados** é uma disciplina tão vasta que qualquer publicação sobre o tema obriga o autor a selecionar alguns tópicos em detrimento de outros não menos importantes

A tarefa de **Regras de Associação** foi o tópico escolhido para iniciarmos a apresentação das principais tarefas da Mineração de Dados por envolver ideias bem intuitivas

A analogia entre **Regras de Associação** e **Cesta de Compras** facilita o entendimento de como descobrir **padrões de associação entre itens** de um conjunto qualquer

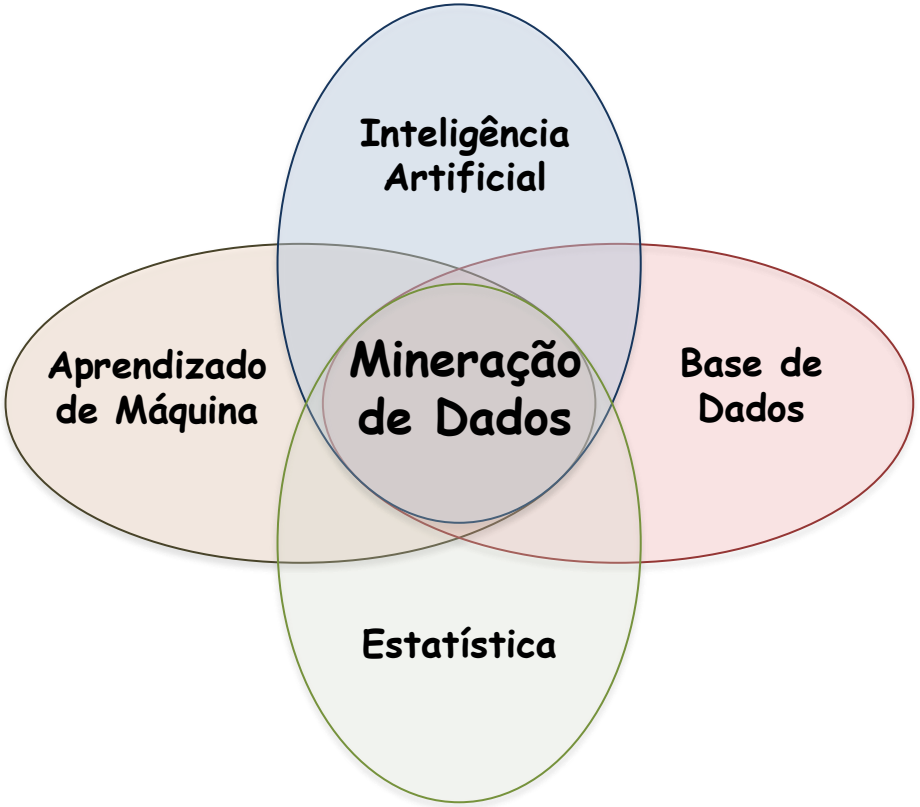
Obs.: Mineração de Dados = MD



Mineração de Dados

Mineração de Dados

A **MD** pode ser vista como a sistematização de teorias, técnicas e algoritmos desenvolvidos em outras disciplinas já consagradas, como a **Estatística**, a **Inteligência Artificial**, o **Aprendizado de Máquina**, a **Base de Dados** etc.



O propósito da **MD** é **detectar automaticamente padrões de associação** úteis e não óbvios em grandes quantidades de dados

Dados Estruturados

A MD trabalha com **dados estruturados**

TID	Itens
1	{Arroz, Feijão, Óleo}
2	{Queijo, Vinho}
3	{Arroz, Feijão, Batata, Óleo}
4	{Arroz, Água, Queijo, Vinho}
5	{Arroz, Feijão, Batata, Óleo}

Transação ou **exemplo**

A estrutura de representação de uma **Base de Dados** pode ser semelhante a uma tabela de dados, sendo cada linha dessa tabela uma **transação** ou um **exemplo**

Cada **transação** é composta por um ou mais **itens** ou, visto de outra forma, cada **exemplo** é caracterizado por seus **atributos**

Cesta de Compras

TID	Arroz	Feijão	Batata	Óleo	Água	Queijo	Vinho
1	y	y	n	y	n	n	n
2	n	n	n	n	n	y	y
3	y	y	y	y	n	n	n
4	y	n	n	n	y	y	y
5	y	y	y	y	n	n	n

Se a tabela for muito extensa, como costuma ser em casos reais, pode ser ainda mais conveniente representar cada um de seus itens na forma de um atributo associado a um valor booleano

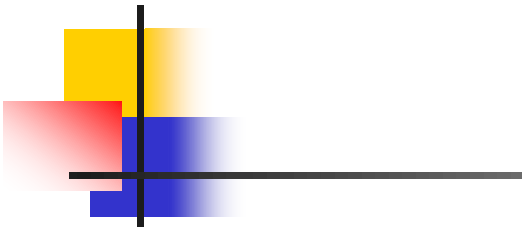
Tabela do Tempo

Exemplo clássico de uma **Base de Dados** usada em artigos sobre MD

Os dados se referem a dias de partida de um esporte não especificado

Como alguns desses atributos sempre ocorrem juntos, várias **Regras de Associação** podem ser extraídas

Dia	Temperatura	Umidade	Vento	Partida
Ensolarado	Elevada	Alta	Falso	Não
Ensolarado	Elevada	Alta	Verdadeiro	Não
Nublado	Elevada	Alta	Falso	Sim
Chuvoso	Amena	Alta	Falso	Sim
Chuvoso	Baixa	Normal	Falso	Sim
Chuvoso	Baixa	Normal	Verdadeiro	Não
Nublado	Baixa	Normal	Verdadeiro	Sim
Ensolarado	Amena	Alta	Falso	Não
Ensolarado	Baixa	Normal	Falso	Sim
Chuvoso	Amena	Normal	Falso	Sim
Ensolarado	Amena	Normal	Verdadeiro	Sim
Nublado	Amena	Alta	Verdadeiro	Sim
Nublado	Elevada	Normal	Falso	Sim
Chuvoso	Amena	Alta	Verdadeiro	Não



Regras de Associação

Regras *IF-THEN*

A representação do conhecimento através de **regras *IF-THEN*** ou **Regras de Produção** é largamente utilizada porque

- São facilmente compreendidas pelo ser humano
- Fáceis de serem alteradas, validadas e verificadas
- De baixo custo para a criação de sistemas baseados em regras

Regra de Associação

Uma **Regra de Associação** revela que a presença de um conjunto **X** de itens numa transação implica outro conjunto **Y** de itens

$\mathbf{X} = \{a, b, ...\} \Rightarrow \mathbf{Y} = \{p, ..., z\}$

O fato de um conjunto de itens **X** (antecedente) estar sempre associado a outro **Y** (consequente) não significa que um seja a causa de outro, i.e., **não há relação de causalidade** entre antecedente e consequente e sim **mera ocorrência simultânea de itens**

Estrutura de uma Regra

A **estrutura** geral de uma **Regra de Associação** assume a seguinte forma:

If (Conjunto **X** de Itens) ***then*** (Conjunto **Y** de Itens)

Exemplos de Regras

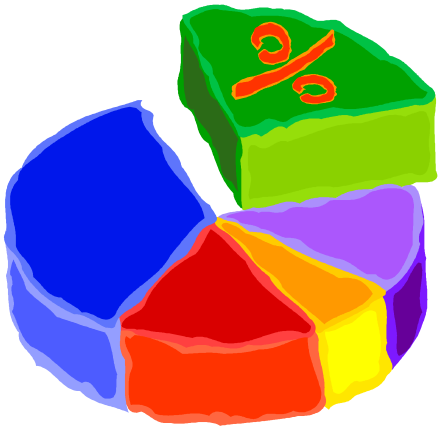
Com base na **Tabela do Tempo**, várias **Regras de Associação** podem ser formuladas

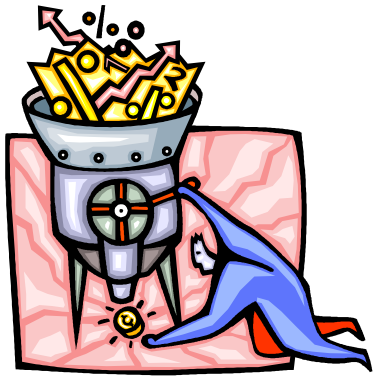
If (Temperatura = Baixa) ***then*** (Umidade = Normal) (1)

If (Umidade = Normal) ***and*** (Vento = Falso) ***then*** (Partida = Sim) (2)

If (Dia = Ensolarado) ***and*** (Partida = Não) ***then*** (Umidade = Alta) (3)

Métricas de Avaliação





- Para **selecionar as Regras de Associação** mais representativas, i.e., aquelas que se apliquem a um grande número de exemplos com alta probabilidade de acerto, precisaremos de métricas para avaliar o alcance ou a força de cada regra

- Dois indicadores conhecidos são **Suporte** e **Confiança**

Suporte

Para cada regra do tipo **X** ⇒ **Y**, o **Suporte** indica a quantos exemplos da tabela esta regra satisfaz (i.e., contém) tanto ao conjunto de itens de **X** quanto ao de **Y**, ou seja, indica sua **cobertura** com relação ao número total ***N*** de exemplos da tabela

$$Sup(X \rightarrow Y) = \frac{|X \cup Y|}{N}$$

Para a Tabela do Tempo, ***N*** = 14 exemplos

Exemplo de Suporte

Com relação à regra (1)

If (Temperatura = Baixa) **then** (Umidade = Normal)

há 4 exemplos em que {**X** ∪ **Y**} = {Temperatura=Baixa, Umidade=Normal}

Dia	Temperatura	Umidade	Vento	Partida
Ensolarado	Elevada	Alta	Falso	Não
Ensolarado	Elevada	Alta	Verdadeiro	Não
Nublado	Elevada	Alta	Falso	Sim
Chuvoso	Amena	Alta	Falso	Sim
Chuvoso	Baixa	Normal	Falso	Sim
Chuvoso	Baixa	Normal	Verdadeiro	Não
Nublado	Baixa	Normal	Verdadeiro	Sim
Ensolarado	Amena	Alta	Falso	Não
Ensolarado	Baixa	Normal	Falso	Sim
Chuvoso	Amena	Normal	Falso	Sim
Chuvoso	Amena	Alta	Verdadeiro	Não

$$Sup(regra\ 1) = \frac{|X\ U\ Y|}{N} = \frac{|\{Temperatura = Baixa, Umidade = Normal\}|}{N} = \frac{4}{14} = 0{,}29$$

As regras 2 e 3 têm Suporte 4/14 e 3/14

Confiança

A **Confiança** de uma regra reflete o número de exemplos que contêm **Y** dentre todos aqueles que contêm **X**

Em outras palavras, o parâmetro **Confiança** determina quantos são os exemplos em que **X** implica **Y**, comparado com aqueles exemplos em que **X** pode ou não implicar **Y** (podem existir regras **X** ⇒ **W** ou **X** ⇒ **Z**)

$$\boldsymbol{Conf(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} = \frac{Sup(X => Y)}{Sup(X)}}$$

Exemplo de Confiança

Com relação à regra (1)

If (Temperatura = Baixa) *then* (Umidade = Normal)

há 4 exemplos em que {**X** ∪ **Y**} = {Temperatura=Baixa, Umidade=Normal} e 4 exemplos em que {**X**} = {Temperatura=Baixa}

Dia	Temperatura	Umidade	Vento	Partida
Ensolarado	Elevada	Alta	Falso	Não
Ensolarado	Elevada	Alta	Verdadeiro	Não
Nublado	Elevada	Alta	Falso	Sim
Chuvoso	Amena	Alta	Falso	Sim
Chuvoso	Baixa	Normal	Falso	Sim
Chuvoso	Baixa	Normal	Verdadeiro	Não
Nublado	Baixa	Normal	Verdadeiro	Sim
Ensolarado	Amena	Alta	Falso	Não
Ensolarado	Baixa	Normal	Falso	Sim
Chuvoso	Amena	Normal	Falso	Sim
Nublado	Elevada	Normal	Falso	Sim
Chuvoso	Amena	Alta	Verdadeiro	Não

$$Conf(regra\ 1) = \frac{|X\ U\ Y|}{|X|} = \frac{|\{Temperatura = Baixa, Umidade = Normal\}|}{|Temperatura = Baixa|} = \frac{4}{4} = 1{,}00$$

As regras 2 e 3 têm Confiança 4/4 e 3/3

Outro Exemplo de Confiança

Para a regra (4)

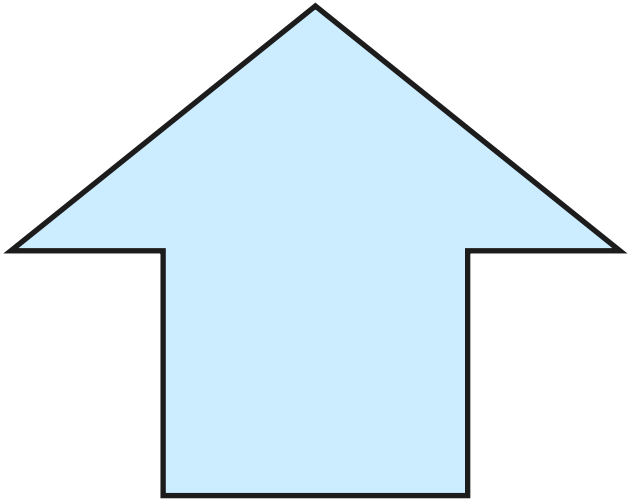
If (Vento = Falso) *and* (Partida = Não) *then* (Temperatura = Elevada) *and* (Umidade = Alta) (4)

há 1 exemplo em que {**X** ∪ **Y**} = {Vento=Falso, Partida=Não, Temperatura=Elevada, Umidade=Alta} e 2 exemplos em que {**X**} = {Vento=Falso, Partida=Não}

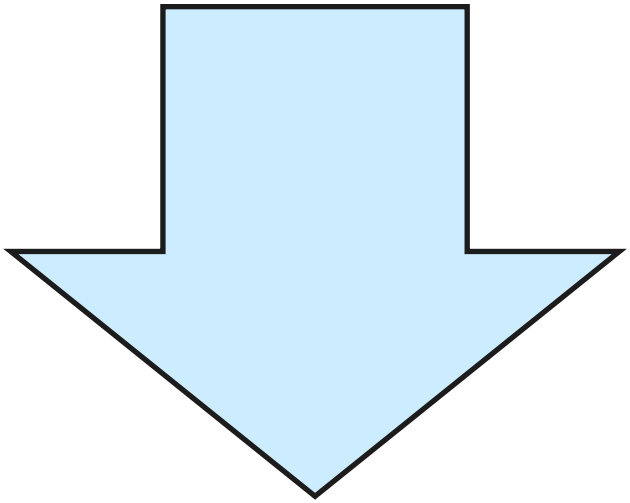
Conf (regra 4) = 1/2

Dia	Temperatura	Umidade	Vento	Partida
Ensolarado	Elevada	Alta	Falso	Não
Ensolarado	Elevada	Alta	Verdadeiro	Não
Nublado	Elevada	Alta	Falso	Sim
Chuvoso	Amena	Alta	Falso	Sim
Chuvoso	Baixa	Normal	Falso	Sim
Chuvoso	Baixa	Normal	Verdadeiro	Não
Nublado	Baixa	Normal	Verdadeiro	Sim
Ensolarado	Amena	Alta	Falso	Não
Ensolarado	Baixa	Normal	Falso	Sim
Chuvoso	Amena	Normal	Falso	Sim
Ensolarado	Amena	Normal	Verdadeiro	Sim
Nublado	Amena	Alta	Verdadeiro	Sim
Nublado	Elevada	Normal	Falso	Sim
Chuvoso	Amena	Alta	Verdadeiro	Não

Alcance ou Força de uma Regra

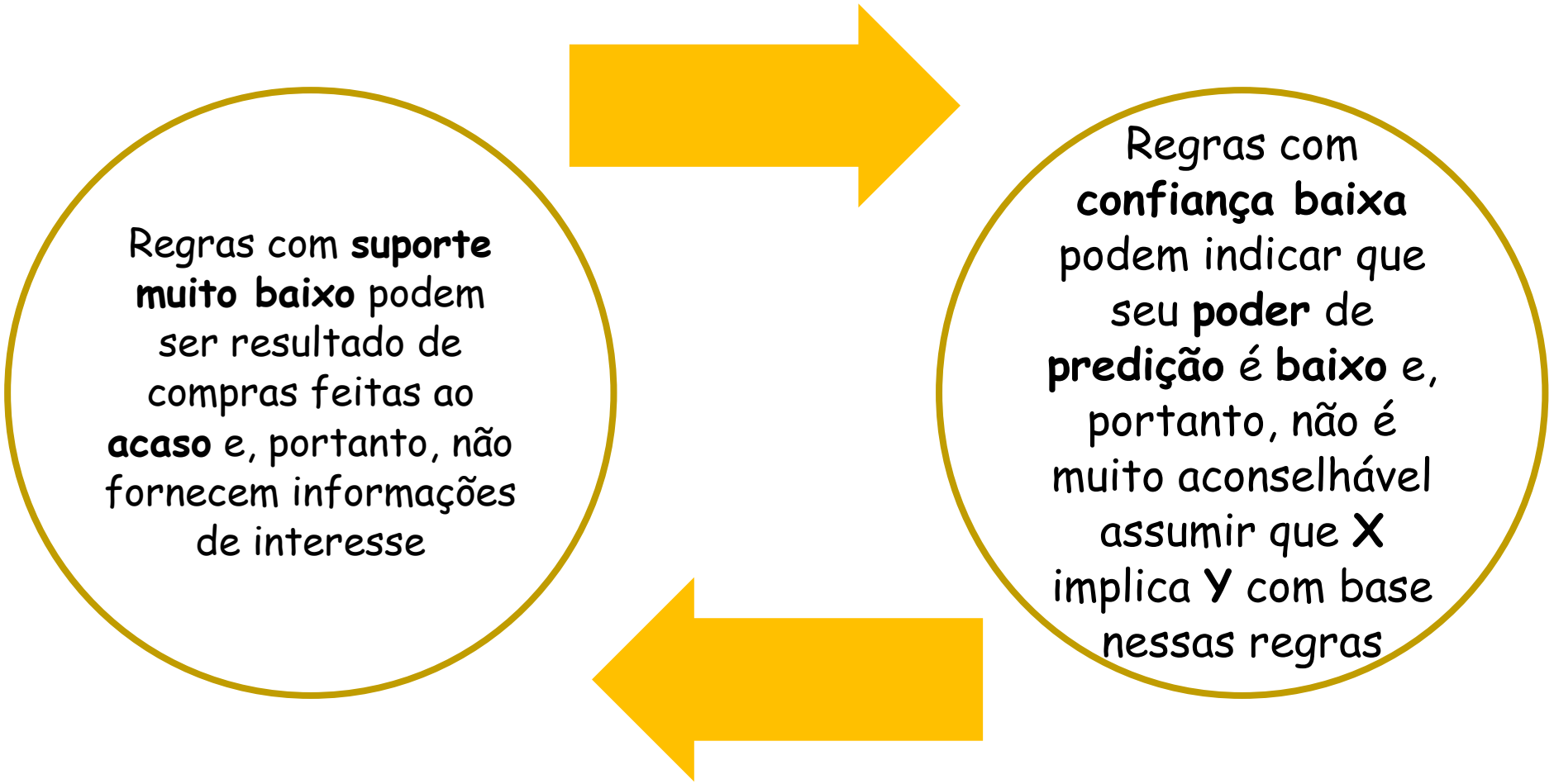


O fato de poucos itens poder gerar muitas **Regras de Associação** faz com que o número de regras geradas seja tão grande que a maioria dessas regras não tem qualquer **interesse prático**



Para contornar esta situação, antes de começar a gerar as **Regras de Associação**, é comum que sejam estabelecidos um valor de Suporte Mínimo (**SupMin**) e de Confiança Mínima (**ConfMin**)

SupMin e ConfMin

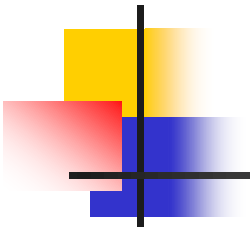


Algoritmo Apriori

Regras de Associação são geradas em duas etapas:

1. Dado um conjunto de transações **T**, primeiramente são criados conjuntos de itens frequentes, chamados de **Conjuntos Frequentes**, que devem satisfazer o limite de **SupMin**

2. A partir desses Conjuntos Frequentes são geradas **Regras de Associação** com confiança maior ou igual **ConfMin**



Geração de Conjuntos Frequentes

Conjuntos Frequentes

TID	A	B	C	D	E	F	G
1	1	1	0	1	0	0	0
2	0	0	0	0	0	1	1
3	1	1	1	1	0	0	0
4	1	0	0	0	1	1	1
5	1	1	1	1	0	0	0

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

De acordo com o algoritmo **Apriori**, para se obter Conjuntos Frequentes, inicialmente devem ser criados **Conjuntos Frequentes com 1** item apenas e que satisfaçam o **critério de Suporte Mínimo**

A seguir são criados recursivamente **Conjuntos Frequentes com 2 itens**, depois com 3 itens, e assim sucessivamente

Conjunto Frequente com 1 Item

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

5 Transações

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{E}	1/5
{F}	2/5
{G}	2/5

Possíveis CF com 1 Item

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{F}	2/5
{G}	2/5

CF com 1 Item e SupMin ≥ 2/5

Suponhamos que o **SupMin** tenha sido definido como 2/5, ou seja, 40%

Como o conjunto {E} não satisfaz **SupMin**, ele deve ser eliminado!!!!

Conjunto Frequente com 2 Itens

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{F}	2/5
{G}	2/5

CF=1 e SupMin≥2/5

Itens	Suporte
{A, B}	3/5
{A, C}	2/5
{A, D}	3/5
{A, F}	1/5
{A, G}	1/5
{B, C}	2/5
{B, D}	3/5
{B, F}	0
{B, G}	0
{C, D}	2/5
{C, F}	0
{C, G}	0
{D, F}	0
{D, G}	0
{F, G}	2/5

Possíveis CF com 2 Itens

Itens	Suporte
{A, B}	3/5
{A, C}	2/5
{A, D}	3/5
{B, C}	2/5
{B, D}	3/5
{C, D}	2/5
{F, G}	2/5

CF=2 e SupMin≥2/5

Conjunto Frequente com 3 Itens

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{F}	2/5
{G}	2/5

CF=1 e SupMin≥2/5

Itens	Suporte
{A, B}	3/5
{A, C}	2/5
{A, D}	3/5
{B, C}	2/5
{B, D}	3/5
{C, D}	2/5
{F, G}	2/5

CF=2 e SupMin≥2/5

Itens	Suporte
{A, B, C}	2/5
{A, B, D}	3/5
{A, C, D}	2/5
{A, F, G}	1/5
{B, C, D}	2/5
{B, F, G}	0
{C, D, F}	0
{C, F, G}	0

Possíveis CF com 3 Itens

Itens	Suporte
{A, B, C}	2/5
{A, B, D}	3/5
{A, C, D}	2/5
{B, C, D}	2/5

CF=3 e SupMin≥2/5

Conjunto Frequente com 4 Itens

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{F}	2/5
{G}	2/5

CF=1 e SupMin≥2/5

Itens	Suporte
{A, B}	3/5
{A, C}	2/5
{A, D}	3/5
{B, C}	2/5
{B, D}	3/5
{C, D}	2/5
{F, G}	2/5

CF=2 e SupMin≥2/5

Itens	Suporte
{A, B, C}	2/5
{A, B, D}	3/5
{A, C, D}	2/5
{B, C, D}	2/5

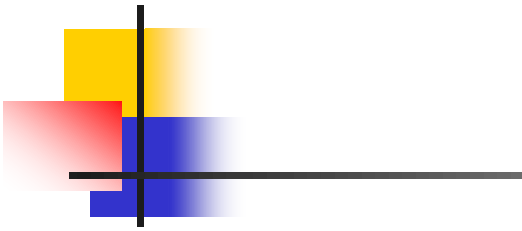
CF=3 e SupMin≥2/5

Itens	Suporte
{A, B, C, D}	2/5

CF=4 e SupMin≥2/5

Se houvesse ao menos dois CF = 4, poderíamos ainda tentar gerar CF = 5

Mas como há apenas um CF = 4, esta primeira etapa do **Apriori** termina aqui



Geração de Regras de Associação

Subconjuntos dos CFs

Na primeira etapa foram gerados:	6 x CF=1 ➔ {A}, {B}, {C}, {D}, {F}, {G}
	7 x CF=2 ➔ {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}, {F, G}
	4 x CF=3 ➔ {A, B, C}, {A, B, D}, {A, C, D}, {B, C, D}
	1 x CF=4 ➔ {A, B, C, D}

Para extrair as **Regras de Associação** de um **CF** é necessário primeiramente gerar todos os subconjuntos não-vazios desse **CF**, e para cada subconjunto **S** de CF produzir uma Regra de Associação do tipo **S ⇒ (CF - S)** que satisfaça o critério de **Confiança ≥ ConfMin**

Exemplo de Subconjunto de CF

Por exemplo, dado o CF = {A, B, C}, seus subconjuntos não-vazios possíveis são S = {{A}, {B}, {C}, {A, B}, {A, C}, {B, C}}

Portanto, é possível extrair seis **Regras de Associação** do CF = {A, B, C}:

{A} ⇒ {B, C},

{B} ⇒ {A, C},

{C} ⇒ {A, B},

{A, B} ⇒ {C},

{A, C} ⇒ {B},

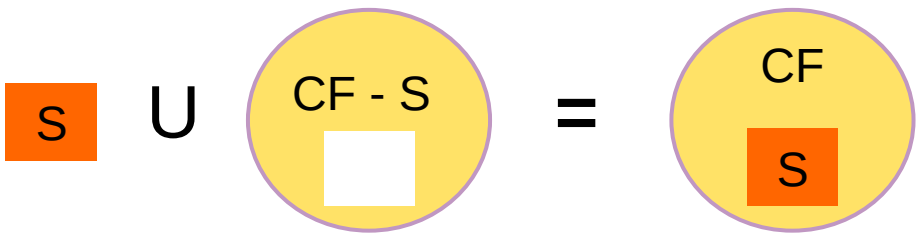
{B, C} ⇒ {A}

Suporte e Confiança

Como o **Suporte de todos os subconjuntos já foi calculado** na Etapa 1, não será necessário percorrer novamente a Base de Dados para calcular a **Confiança** de cada Regra de Associação

Basta reutilizar estes valores calculados, pois

$$Conf(S \rightarrow CF - S) = \frac{|S \cup CF - S|}{|S|} = \frac{|CF|}{|S|} = \frac{Sup(CF)}{Sup(S)}$$



Voltando ao Exemplo Inicial

TID	Itens
1	{A, B, D}
2	{F, G}
3	{A, B, C, D}
4	{A, E, F, G}
5	{A, B, C, D}

Vamos voltar ao Exemplo Inicial

O Suporte do CF = {A, B, C} é

Sup({A, B, C}) = 2/5

e de seus subconjuntos

Sup({A}) = 4/5
Sup({B}) = 3/5,
Sup({C}) = 2/5,
Sup({A, B}) = 3/5,
Sup({A, C}) = 2/5,
Sup({B, C}) = 2/5

Itens	Suporte
{A, B, C}	2/5
{A, B, D}	3/5
{A, C, D}	2/5
{B, C, D}	2/5

Itens	Suporte
{A}	4/5
{B}	3/5
{C}	2/5
{D}	3/5
{F}	2/5
{G}	2/5

Itens	Suporte
{A, B}	3/5
{A, C}	2/5
{A, D}	3/5
{B, C}	2/5
{B, D}	3/5
{C, D}	2/5
{F, G}	2/5

Confiança das Regras

Portanto, a Confiança das seis possíveis **Regras de Associação** do CF = {A, B, C} são:

$$Conf(\{A\} \Rightarrow \{B, C\}) = (2/5)/(4/5) = 0,50$$

$$Conf(\{B\} \Rightarrow \{A, C\}) = (2/5)/(3/5) = 0,66$$

$$Conf(\{C\} \Rightarrow \{A, B\}) = (2/5)/(2/5) = 1,00$$

$$Conf(\{A, B\} \Rightarrow \{C\}) = (2/5)/(3/5) = 0,66$$

$$Conf(\{A, C\} \Rightarrow \{B\}) = (2/5)/(2/5) = 1,00$$

$$Conf(\{B, C\} \Rightarrow \{A\}) = (2/5)/(2/5) = 1,00$$

Regras Aprovadas

Suponha que para o problema em questão tenha sido adotado **SupMin = 40%** e **ConfMin = 90%**, então apenas três das regras acimas seriam aproveitadas:

Conf({C} ⇒ {A, B}) = 1,00 **{Batata} ⇒ {Arroz, Feijão}**

Conf({A, C} ⇒ {B}) = 1,00 **{Arroz, Batata} ⇒ {Feijão}**

Conf({B, C} ⇒ {A}) = 1,00 **{Feijão, Batata} ⇒ {Arroz}**

TID	Itens
1	{Arroz, Feijão, Óleo}
2	{Queijo, Vinho}
3	{Arroz, Feijão, Batata, Óleo}
4	{Arroz, Água, Queijo, Vinho}
5	{Arroz, Feijão, Batata, Óleo}

Aplicando-se o procedimento explicado acima para todos os 18 CFs obtidos na Etapa 1, seriam geradas aproximadamente **30 Regras de Associação com SupMin = 40% e ConfMin = 90%**

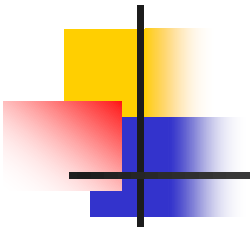
Considerações Finais

No exemplo da **Cesta de Artigos** mostramos como gerar **Regras de Associação** que indiquem **venda casada** dos artigos mais comum. Mas, frequentemente, os especialistas em vendas não estão muito interessados nestes itens porque a associação entre eles já é conhecida

Na realidade, estes especialistas buscam pares de itens dos quais um deles é um produto barato e o outro tem alta taxa de lucro. Nestes casos, lançar uma **superpromoção do produto barato faz com que as vendas do produto com alta taxa de lucro aumente**

Em nossa Cesta de Artigos está implícito o padrão de associação entre Queijo e Vinho. Talvez aí, numa campanha de inverno, cadeias de supermercados possam fazer promoções de queijos com o único propósito de vender mais vinhos

Mas como as vendas de ambos eram relativamente baixas, esta regra não satisfez os critérios estabelecidos de ***SupMin*** e ***ConfMin***. E, no entanto, é possivelmente este tipo de informação a mais procurada. O que fazer para conseguir minerar as pérolas de informação?



Referência Bibliográfica

Referência Bibliográfica

AGRAWAL, R.; IMIELINSKI, T. & SWAMI, A. **Mining Association Rules Between Sets of Items in Large Databases**. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC. New York: ACM, 1993.

PADHY, N. P. **Artificial Intelligence and Intelligent Systems**. New Delhi: Oxford University Press, 2010.

QUINLAN, J. R. **Induction of Decision Trees**. Machine Learning, Vol. 1, No. 1, pp. 81-106. Boston: Kluwer Academic Publishers, 1986.

ROCHA, M.; CORTEZ, P. & NEVES, J. M. **Análise Inteligente de Dados: Algoritmos e Implementação em Java**. Lisboa: FCA – Editora de Informática, 2008.

TAN, P.N.; STEINBACH, M. & KUMAR, V. **Introdução ao Data Mining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

WITTEN, I. H. & FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. Second Edition. Amsterdam: Morgan Kaufmann Publishers, 2005.