UNIDADE 4
Aprendizado de Máquina
Associação

Disciplina: Tópicos Especiais III (DEC7553)

Prof. Alexandre L. Gonçalves

E-mail: a.l.goncalves@ufsc.br

Regras de Associação

Visa descobrir **associações** importantes entre os itens (***k-itemsets***), tal que, a presença de um item em uma determinada transação irá implicar na presença de outro item na mesma transação.

Cada registro corresponde a uma transação, com itens assumindo valores binários (sim/não), indicando se o cliente comprou ou não o respectivo item.

Id	Leite	Café	Cerveja	Pão	Arroz	Feijão
1	S	N	N	S	N	N
2	N	S	S	S	S	N
3	S	S	S	N	N	S
4	S	S	S	S	N	N
5	S	S	N	S	N	S

Regras de Associação

Uma regra de associação é uma **implicação** na forma **X ⇒ Y.**
Por exemplo:

Se (Café) Então (Pão);
Se (Leite) Então (Café e Pão)

O número de regras que podem ser extraídas é: $R = 3^d - 2^{d+1} + 1$, onde **d** é o número de itens. No exemplo $R = 3^6 - 2^7 + 1 = 180$ regras.

Para diminuir esse número, são definidos para cada regra de associação dois parâmetros básicos: um **suporte** e uma **confiança**;

O suporte (frequência) é caracterizado pelo número mínimo de ocorrências, enquanto que a confiança (força da regra) é um percentual das transações na base de dados que satisfazem o antecedente da regra (X) e também satisfazem o consequente da regra (Y).

Regras de Associação

A função do **Suporte** é determinar a freqüência que ocorre um *itemset* dentre todas as transações da Base de Dados; representa a porcentagem de transações onde este *itemset* aparece.

Um *itemset* será considerado frequente se o seu suporte for maior ou igual a um suporte mínimo estabelecido previamente.

$$Suporte(A \Rightarrow B) = \frac{N^o de\ registros\ com\ (A \cap B)}{N^o Total\ de\ transações\ da\ BD}$$

Regras de Associação

A toda regra X => Y associa-se um grau de **confiança**. Ela é a medida da força da regra e determina a sua validade. A probabilidade condicional de se encontrar B, já tendo encontrado A é obtida pela confiança.

Assim como o suporte, também é estabelecido um nível mínimo de confiança para as regras.

$$Conf(A \Rightarrow B) = \frac{N^o\ de\ transações\ que\ \text{suportam}\ (A \cap B)}{N^o\ de\ transações\ que\ \text{suportam}\ (A)}$$

Regras de Associação

Id	Leite	Café	Cerveja	Pão	Arroz	Feijão
1	S	N	N	S	N	N
2	N	S	S	S	S	N
3	S	S	S	N	N	S
4	S	S	S	S	N	N
5	S	S	N	S	N	S

Se (Café) **Então** (Pão) S = 0.6 (3/5) C = 0.75 (3/4);
Se (Café e Pão) **Então** Leite S = 0.4 (2/5) C = 0.67 (2/3);
Se (Leite) **Então** (Café e Pão) S = 0.4 (2/5) C = 0.50 (2/4);

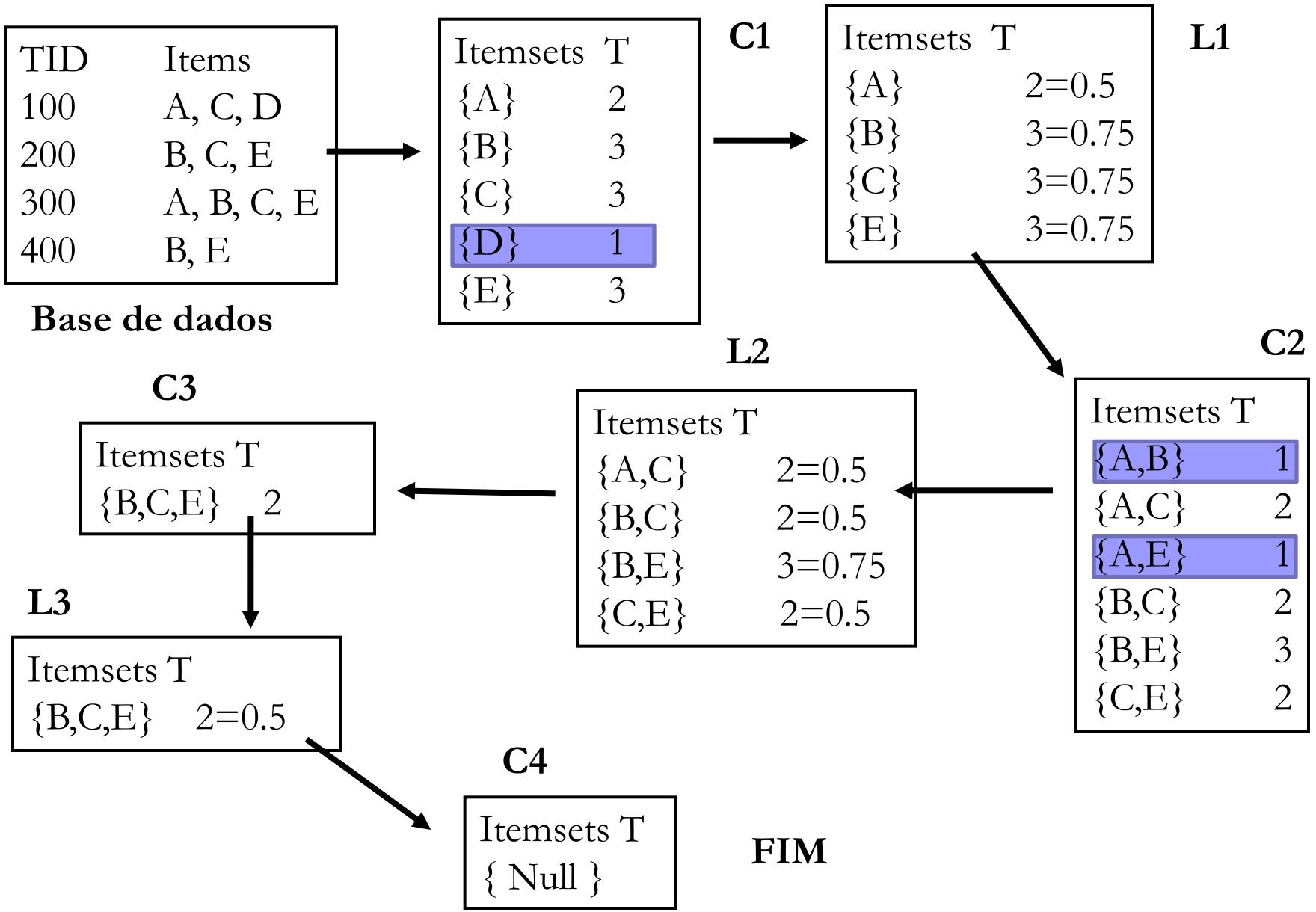
Regras de Associação

Fases do algoritmo Apriori:

1. Geração dos conjuntos candidatos, com **suporte** maior ou igual ao mínimo estabelecido;
    1. Repetir até que não haja mais possibilidades de combinações:
        1. Contagem no banco de dados
        2. Poda dos conjuntos candidatos considerando o suporte
        3. Combinação dos *itemsets*
2. Geração das regras de associação dos conjuntos candidatos gerados com **confiança** maior ou igual ao mínimo estabelecido.

Algoritmo Apriori

Suporte mínimo = 0.5 (ou 2)



Algoritmo Apriori

Regras geradas para L2 (s >= 2 (ou 50%) e c >= 60%)

Se A Então C (s = 50%, c = 100%)
Se C Então A (s = 50%, c = 66.7%)
Se B Então C (s = 50%, c = 66.7%)
Se C Então B (s = 50%, c = 66.7%)
Se B Então E (s = 75%, c = 100%)
Se E Então B (s = 75%, c = 100%)
Se C Então E (s = 50%, c = 66.7%)
Se E Então C (s = 50%, c = 66.7%)

Algoritmo Apriori

Regras geradas para L3 (s >= 2 (ou 50%) e c >= 60%)

Se B e C Então E (s = 50%, c = 100%)
Se B e E Então C (s = 50%, c = 66.7%)
Se C e E Então B (s = 50%, c = 100%)
Se B Então C e E (s = 50%, c = 66.7%)
Se C Então B e E (s = 50%, c = 66.7%)
Se E Então B e C (s = 50%, c = 66.7%)

Outras Medidas: Medidas de Interesse

A medida de interesse *lift*, também conhecida como *interest*, é utilizada para avaliar dependências.

Dada uma regra de associação $\mathbf{A \Rightarrow B}$, a medida indica o quanto mais frequente torna-se **B** quando **A** ocorre:

$$\mathbf{Lift(A \Rightarrow B) = Sup(A \Rightarrow B) \,/\, (Sup(A) \times Sup(B))}$$

Se Lift(A $\Rightarrow$ B) = 1, então A e B são independentes.
Se Lift(A $\Rightarrow$ B) > 1, então A e B tem relação positiva.
Se Lift(A $\Rightarrow$ B) < 1, então A e B tem relação negativa.

Esta medida possui interpretação bastante simples:
quanto maior o valor do *lift*, mais interessante a regra.

Outras Medidas: Medidas de Interesse

O valor do RI (regra de interesse) para A ⇒ B é computado por:

RI(A ⇒ B) = Sup(A ⇒ B) - SupEsp(A ⇒ B)

onde SupEsp, é o suporte esperado e calculado como:

SupEsp(A ⇒ B) = Sup(A) x Sup(B)

Se RI(A ⇒ B) = 0, então A e B são independentes.
Se RI(A ⇒ B) > 0, então A e B são positivamente dependentes.
Se RI(A ⇒ B) < 0, A e B são negativamente dependentes.

Nesta medida quanto maior o valor da RI, mais interessante é a regra.

Outras Medidas: Medidas de Interesse

Tanto o *lift* quanto o RI possuem como característica o fato de serem medidas simétricas, ou seja:

Lift(A ⇒ B) = Lift(B ⇒ A) e RI(A ⇒ B) = RI(B ⇒ A).

Isto ocorre porque estes índices possuem o objetivo de mensurar dependência entre os itens, ao invés de medir implicação (o sentido da seta "⇒ ").

A medida de interesse de convicção é proposta com o objetivo de avaliar uma regra de associação como uma verdadeira implicação.

$$conv\left(A \Rightarrow B\right) = \frac{1-\sup(B)}{1-conf\,(A \Rightarrow B)}$$

Categorização dos Dados

ID	Idade	Casado	Carros
ID1	23	N	1
ID2	25	S	1
ID3	29	N	0
ID4	34	S	2
ID5	38	S	2

ID	Idade	Carros	Casado
ID1	range1 [-∞ - 24]	range2 [0.500 - 1.500]	N
ID2	range2 [24 - 31.500]	range2 [0.500 - 1.500]	S
ID3	range2 [24 - 31.500]	range1 [-∞ - 0.500]	N
ID4	range3 [31.500 - ∞]	range3 [1.500 - ∞]	S
ID5	range3 [31.500 - ∞]	range3 [1.500 - ∞]	S

Categorização dos Dados

Sexo	Escolaridade	Tem Computador	Estado
M	Superior	S	RS
F	Posgraduação	S	SC
F	Posgraduação	S	SC
F	2 grau	N	PR
M	Superior	S	PR
M	2 grau	N	RS
F	1 grau	S	RS
M	1 grau	N	PR
M	2 grau	S	PR
F	Superior	S	RS

Masculino	Feminino	1 grau	2 grau	Superior	Posgraduacao	Computador
1	0	0	0	1	0	1
0	1	0	0	0	1	1
0	1	0	0	0	1	1
0	1	0	1	0	0	0
1	0	0	0	1	0	1
1	0	0	1	0	0	0
0	1	1	0	0	0	1
1	0	1	0	0	0	0
1	0	0	1	0	0	1
0	1	0	0	1	0	1

Bons Estudos!