Regras de Associação

Stanley Robson de M. Oliveira

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no Weka.

Motivação

- É possível entender o comportamento de **consumo de clientes**?
- É possível analisar uma sequência de **eventos climáticos**?
- Quais são as **pragas frequentemente encontradas** em uma determinada cultura?
- Onde devem estar localizadas os produtos **agropecuários** ?
- Que fenômenos ocorrem conjuntamente com o ***El Niño*** ?
- Qual é o percentual de produtores de **soja** e **milho** no estado de São Paulo?

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no Weka.

Associação: Definição e Exemplos

- Estuda o **relacionamento entre itens** de dados que ocorrem com uma **certa frequência**.
- É uma tarefa **descritiva**: identifica padrões em **dados históricos**.
- **Exemplos**:
  - Clientes que compram **leite** e **pão** também compram **manteiga**.
  - Em forma de regra seria: {leite, pão} ⇒ {manteiga}

Associação: Conceitos Básicos

- Alguns algoritmos para geração de **regras de associação** trabalham com **banco de dados de transações**.

Banco de Dados de Transações

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T45	Pão, Leite, Fralda, Coca

- Cada transação é composta por uma **identificação** (**TID**) e uma **lista de itens**.

Associação: Conceitos Básicos ...

- Considere o banco de dados de transações:

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

- **Itens**: I = {Pão, Leite, Fralda, Cerveja, Ovos, Coca}.
- **Banco de dados**: *D* = {T1, T2, T3, T4, T5}.
- **Exemplo de transação**:
  - T3 = {Leite, Fralda, Cerveja, Coca}.

Conjuntos Frequentes

- **Conjunto de itens (Itemset)**:
  - Uma coleção de um ou mais itens.
    - **Exemplo**: {Leite, Pão, Fralda}
  - k-itemset
    - Um conjunto que contém k itens.
- **Support count (σ)**
  - Fraquência da ocorrência de um itemset
  - **Exemplo:** σ({leite, Pão,Fralda}) = 2
- **Suporte**
  - Fração de transações que contém um itemset.
  - **Exemplo**: s({leite, Pão, Fralda}) = 2/5
- **Conjunto Frequente (Frequent Itemset)**
  - Um conjunto cujo suporte é maior ou igual a ***minsup* threshold**.

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

Associação: Conceitos Básicos ...

- Considere o banco de dados de transações:

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

- Uma **regra de associação** é uma implicação da forma (X → Y), onde X e Y são conjunto de itens e $X \cap Y = \emptyset$.
  - R1: {Cerveja} → {Fralda}.
  - R2: {Cerveja, Pão} → {Leite}.
  - R3: {Leite, Pão} → {Fralda,Coca}.

Regras de Associação

- **Regra de Associação:**
  - Uma implicação da forma **X → Y**, onde **X** e **Y** são conjuntos frequentes.
  - **Exemplo**: {Leite, Fralda} → {Cerveja}
- **Métricas para Avaliar as Regras:**
  - **Suporte** (s)
    - Fração das transações que contém ambos **X** e **Y**.
    - $Sup(X \rightarrow Y) = P(X \cup Y)$.
  - **Confiança** (c)
    - Mede a frequência de itens em **Y** que aparece nas transações que contêm **X**.
    - $Conf(X \rightarrow Y) = P(Y|X)$.
    - $Conf(X \rightarrow Y) = Sup(X \cup Y) / Sup(X)$

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

Exemplo:

{Leite, Fralda} ⇒ {Cerveja}

$$Sup = \frac{\text{Freq (Leite, Fralda, Cerveja)}}{|T|} = \frac{2}{5}$$

$$Conf = \frac{\text{Freq (Leite, Fralda, Cerveja)}}{\text{Freq (Leite, Fralda)}} = \frac{2}{3}$$

Minerando Regras de Associação

- Dado um conjunto de transações T, a tarefa de **mineração de regras de associação** é encontrar todas as regras que:
  - **suporte** ≥ fator ***minsup*** estabelecido pelo usuário.
  - **confiança** ≥ fator ***minconf*** definido pelo usuário.
- **Abordagem da Força Bruta**:
  - Listar todas as regras possíveis.
  - Calcular o **suporte** e a **confiança** para cada regra.
  - Podar regras que não atendem os fatores ***minsup*** e ***minconf***.

  ⇒ **Computacionalmente proibitivo**!

Regras de Associação ...

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Fralda, Cerveja, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Pão, Leite, Fralda, Cerveja
T5	Pão, Leite, Fralda, Coca

Exemplos de Regras:

{Leite, Fralda} → {Cerveja} (s=0.4, c=0.67)
{Leite, Cerveja} → {Fralda} (s=0.4, c=1.0)
{Fralda, Cerveja} → {Leite} (s=0.4, c=0.67)
{Cerveja} → {Leite, Fralda} (s=0.4, c=0.67)
{Fralda} → {Leite, Cerveja} (s=0.4, c=0.5)
{Leite} → {Fralda, Cerveja} (s=0.4, c=0.5)

Observações:

• Todas as regras acima são originadas do mesmo conjunto frequente: {**Leite, Fralda, Cerveja**}

• Regras originadas do mesmo conjunto frequente têm o mesmo **suporte**, mas diferentes valores para **confiança**.

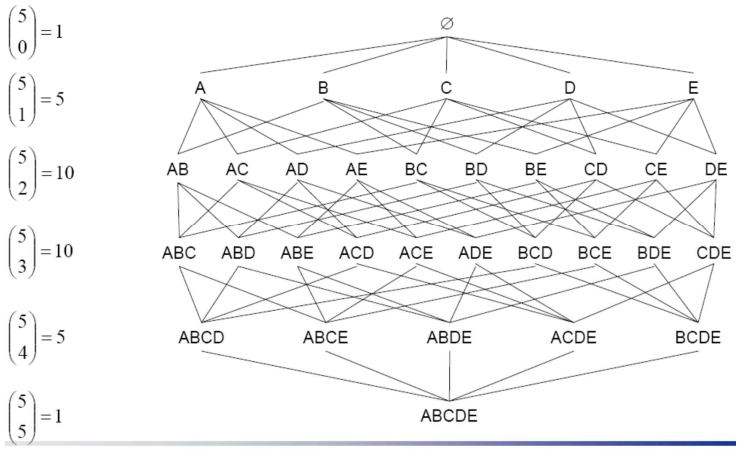
Regras de associação: Aplicações

- **Associação de produtos** em um processo de compra;
- Elaboração de **catálogos de produtos**;
- Layout de prateleiras (**produtos relacionados tendem a ser colocados perto nas prateleiras**);
- Análise de **sequências de DNA**;
- Análise de Web log (**click stream**);
- Sistemas de **recomendação**, etc.

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no Weka.

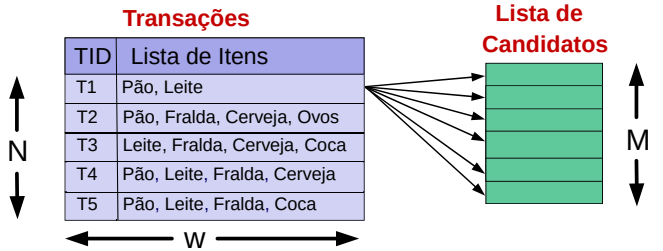
Geração de Conjuntos Frequentes



Dados d itens, existem 2^d possíveis itemsets candidatos

Geração de Conjuntos Frequentes

- Abordagem da Força Bruta:
  - Cada itemset frequente no reticulado é um candidato.
  - A contagem do suporte de cada candidato é feita "escaneando-se" todo o conjunto de dados (dataset).



  - Cada transação pode ser associado com todo candidato.
  - Complexidade ~O(NMw) => Caro já que M = 2d !!!

Estratégias para reduzir candidatos

- Reduzir o **número de candidatos** (M)
  - Busca exaustiva: $M=2^d$
  - Usar técnicas para podar e reduzir M.
- Reduzir o **número de transações** (N)
  - Reduzir o tamanho de N sempre que o tamanho dos **itemsets** crescem.
- Reduzir o **número de comparações** (NM)
  - Usar **estruturas de dados eficientes** para armazenar os **candidatos** ou **transações**.
  - Eliminar a necessidade de cada candidato ser associado com toda transação.

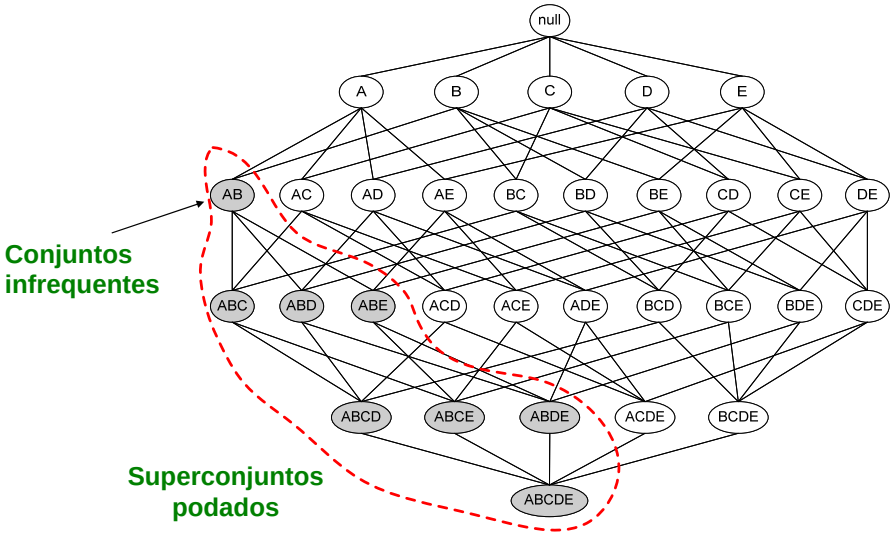
Reduzindo o Número de Candidatos

- **O princípio do Algoritmo Apriori**:
  - Se um itemset é frequente, os seus subconjuntos devem ser frequentes.
  - **Exemplo:** Se {**Leite, Fralda, Cerveja**} é frequente, todos os subconjuntos desse itemset com dois itens também o são.
- O **princípio do Apriori** assegura a seguinte propriedade para a medida **suporte**:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - O suporte de um conjunto frequente nunca excede o suporte de seus subconjuntos.
  - Essa propriedade do suporte é conhecida como **anti-monotônica**.

Ilustrando o Princípio do Apriori



Ilustrando o Princípio do Apriori ...

Id	Compras
1	1,3,5
2	1,2,3,5,7
3	1,2,4,9
4	1,2,3,5,9
5	1,3,4,5,6,8
6	2,7,8

L1 = {1}, {2}, {3},{5}

C2 = {1,2} {1,3} {1,5} {2,3} {3,5} {2,5}

L2 = {1,2} {1,3} {1,5} {3,5}

C3 = {1,2,3} {1,2,5} {1,3,5}

L3 = {1,3,5}

Suporte mínimo = 50%

Ilustrando o Princípio do Apriori ...

Item	Frequência
Pão	4
Coca	2
Leite	4
Cerveja	3
Fralda	4
Ovos	1

Itens (1-itemset)

Itemset	Freqüência
{Pão, Leite}	3
{Pão, Cerveja}	2
{Pão, Fralda}	3
{Leite, Cerveja}	2
{Leite, Fralda}	3
{Cerveja, Fralda}	3

Pares (2-itemsets)

(Não há necessidade de gerar candidatos que contém Coca ou Ovos)

Mínimo Suporte = 3

Triplas (3-itemsets)

Itemset	Frequência
{Pão, Leite, Fralda}	3

Se todo subconjunto for considerado:
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$

Usando o suporte para podar:
6 + 6 + 1 = 13

Ilustrando o Princípio do Apriori ...

Cálculo do SUPORTE

$$\frac{\text{No. Registros com X e Y}}{\text{No. Total de Registros}}$$

Nº Transação	Produtos Adquiridos
1	Café, Pão, Manteiga
2	Leite, Cerveja, Pão, Manteiga
3	Café, Pão, Manteiga
4	Leite, Café, Pão, Manteiga
5	Cerveja
6	Manteiga, Arroz
7	Pão
8	Feijão
9	Arroz, Feijão
10	Arroz

1º. Passo: Calcular suporte de conjuntos com 1 item

Leite	Sup = 0,2
Café	Sup = 0,3
Cerveja	Sup = 0,2
Pão	Sup = 0,5
Manteiga	Sup = 0,5
Arroz	Sup = 0,2
Feijão	Sup = 0,2

Sup ≥ 0,3

Ilustrando o Princípio do Apriori ...

Cálculo do SUPORTE

$$\frac{\text{No. Registros com X e Y}}{\text{No. Total de Registros}}$$

Nº Transação	Produtos Adquiridos
1	Café, Pão, Manteiga
2	Leite, Cerveja, Pão, Manteiga
3	Café, Pão, Manteiga
4	Leite, Café, Pão, Manteiga
5	Cerveja
6	Manteiga, Arroz
7	Pão
8	Feijão
9	Arroz, Feijão
10	Arroz

2º. Passo: Calcular suporte de conjuntos com 2 itens

"Se um item J não é frequente, um conjunto com 2 itens, um dos quais é o item J, não pode ser frequente. Conjuntos contendo item J são ignorados."

Café, Pão	Sup = 0,3
Café, Manteiga	Sup = 0,3
Manteiga, Pão	Sup = 0,4

Sup ≥ 0,3

Ilustrando o Princípio do Apriori ...

Cálculo do SUPORTE

$$\frac{\text{No. Registros com X e Y}}{\text{No. Total de Registros}}$$

Nº Transação	Produtos Adquiridos
1	Café, Pão, Manteiga
2	Leite, Cerveja, Pão, Manteiga
3	Café, Pão, Manteiga
4	Leite, Café, Pão, Manteiga
5	Cerveja
6	Manteiga, Arroz
7	Pão
8	Feijão
9	Arroz, Feijão
10	Arroz

3º. Passo: Calcular suporte de conjuntos com 3 itens

"Se um conjunto de itens (J,K) não é frequente, um conjunto com 3 itens que inclua o conjunto de itens (J,K) não pode ser frequente. Conjuntos contendo itens (J,K) são ignorados."

Café, Pão, Manteiga Sup = 0,3 **Sup ≥ 0,3**

Ilustrando o Princípio do Apriori ...

Cálculo da CONFIANÇA

$$\frac{\text{No. Registros com X e Y}}{\text{No. de Registros com X}}$$

Confiança mínima de 80%

Nº Transação	Produtos Adquiridos
1	Café, Pão, Manteiga
2	Leite, Cerveja, Pão, Manteiga
3	Café, Pão, Manteiga
4	Leite, Café, Pão, Manteiga
5	Cerveja
6	Manteiga, Arroz
7	Pão
8	Feijão
9	Arroz, Feijão
10	Arroz

- **(Café, Pão)** **Regra: SE (café) ENTÃO (pão)** **Conf=1,0**
  Regra: SE (pão) ENTÃO (café) **Conf=0,6**
- **(Café, Manteiga)** **Regra: SE (café) ENTÃO (manteiga)** **Conf=1,0**
  Regra: SE (manteiga) ENTÃO (café) **Conf=0,6**
- **(Pão, Manteiga)** **Regra: SE (pão) ENTÃO (manteiga)** **Conf=0,8**
  Regra: SE (manteiga) ENTÃO (café) **Conf=0,8**

Ilustrando o Princípio do Apriori ...

Cálculo da CONFIANÇA

$$\frac{\text{No. Registros com X e Y}}{\text{No. de Registros com X}}$$

Nº Transação	Produtos Adquiridos
1	Café, Pão, Manteiga
2	Leite, Cerveja, Pão, Manteiga
3	Café, Pão, Manteiga
4	Leite, Café, Pão, Manteiga
5	Cerveja
6	Manteiga, Arroz
7	Pão
8	Feijão
9	Arroz, Feijão
10	Arroz

- **(Café, Pão, Manteiga)**
  Regra: SE (café E pão) ENTÃO (manteiga) **Conf = 1,0**
  Regra: SE (café E manteiga) ENTÃO (pão) **Conf = 1,0**
  Regra: SE (manteiga E pão) ENTÃO (café) **Conf = 0,75**
  Regra: SE (café) ENTÃO (manteiga E pão) **Conf = 1,0**
  Regra: SE (pão) ENTÃO (manteiga E café) **Conf = 0,6**
  Regra: SE (manteiga) ENTÃO (pão E café) **Conf = 0,6**

O Algoritmo Apriori

Algoritmo:

- Faça k = 1;
- Gerar os conjuntos frequentes de tamanho 1;
- Repetir até que nenhum conjunto frequente seja identificado:
  - Gerar conjuntos frequentes candidatos de tamanho (k+1);
  - Podar conjuntos candidatos contendo subconjuntos de tamanho k que são infrequentes;
  - Calcular o suporte de cada conjunto candidato no DB;
  - Eliminar candidatos que são infrequentes, preservando somente aqueles que são freqUentes.

Paper Clássico: (Algoritmo Apriori, AGRAWAL et al., 1993)

Exercício

- Dado o BD de transações, abaixo, determine os conjuntos frequentes com 1, 2 e 3 itens, considerando um suporte mínimo de 30%:

TID	Lista de Produtos
1	Camisa, Algodão, Calça Jeans, Arroz, Feijão, Fralda
2	Livro, DVD, Calça Jeans, Arroz, Algodão
3	DVD
4	Calça Jeans
5	Feijão, Algodão, Arroz, Camisa
6	Camisa
7	Arroz, Feijão, Algodão
8	Livro, DVD

Fatores que afetam a complexidade

- **Escolha do suporte mínimo**:
  - Suporte baixo resulta em muitos conjuntos frequentes.
- **Dimensionalidade** (**número de itens**) **do dataset**:
  - Mais espaço será preciso para armazenar o suporte de cada item;
  - Se o número de itens frequentes aumenta, o custo computacional e o custo com I/O também aumenta.
- **Tamanho do banco de dados**:
  - Como o Apriori efetua múltiplos passos, o tempo de execução do algoritmo pode aumentar com o número de transações.
- **Número médio de itens por transação**:
  - O tamanho (# itens) da transação aumenta para datasets densos.

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no Weka.

Problemas na Geração de Regras

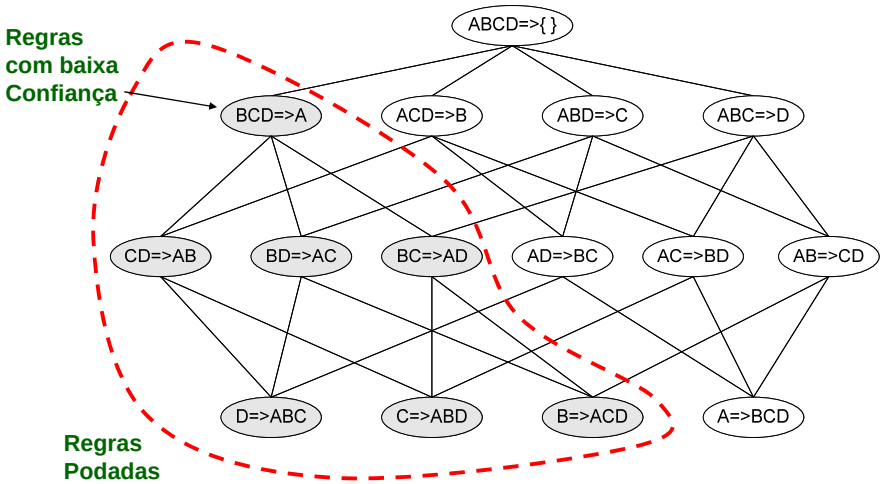
- Como gerar **regras eficientes** a partir de conjuntos frequentes?
  - Em geral, confiança não tem a propriedade anti-monotônica.

  c(ABC $\rightarrow$D) pode ter confiança maior ou menor que c(AB $\rightarrow$D)
  - Mas a **confiança** de regras geradas do **mesmo itemset** tem a propriedade anti-monotônica.
  - **Exemplo**: L = {A,B,C,D}:

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

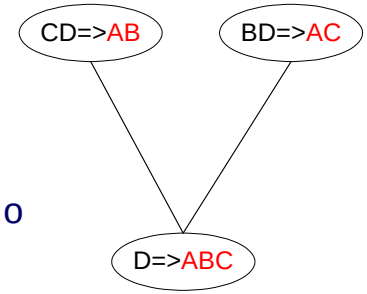
    - **Confiança** é anti-monotônica com relação ao número de itens na parte **consequente da regra**.

Reticulado para Geração de Regras



Apriori: Geração de Regras

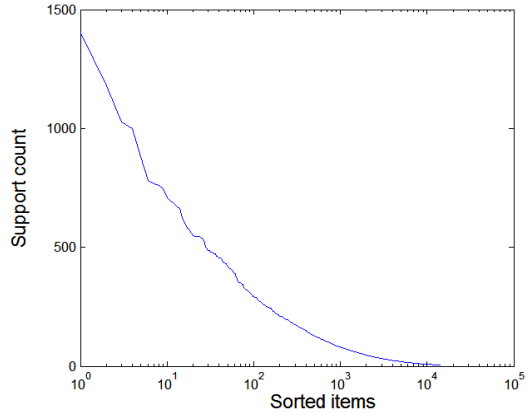
- Uma **regra candidata** é gerada pelo **merge** de duas regras que compartilham o **mesmo prefixo** no consequente da regra.
- **Junção** (CD=>AB, BD=>AC) produziria a regra candidata D => ABC
- **Podar** a regra D=>ABC se o seu subconjunto AD=>BC não tem confiança alta.



Efeito da Distribuição do Suporte

- Muitos datasets com dados reais têm distribuição do suporte distorcida (**skewed**).

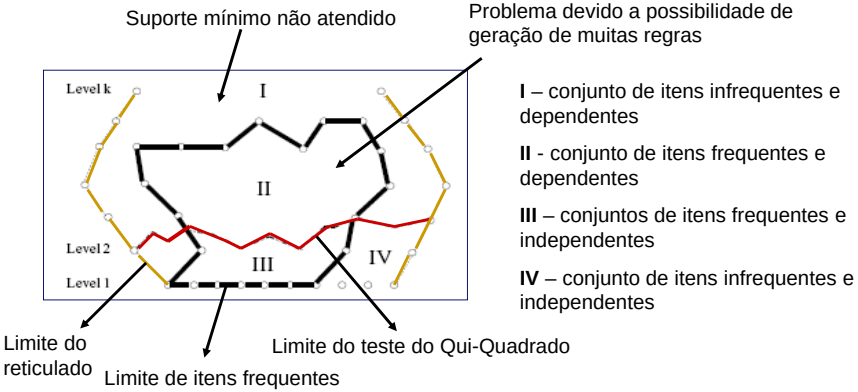
Distribuição do suporte para um dataset de vendas a varejo.



Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no Weka.

Como Selecionar Regras Relevantes



I – conjunto de itens infrequentes e dependentes

II - conjunto de itens frequentes e dependentes

III – conjuntos de itens frequentes e independentes

IV – conjunto de itens infrequentes e independentes

Reticulado de todos os conjuntos de itens

Por quê Medidas de Interesse?

- Em geral, os **algoritmos** para regras de associação **produzem muitas regras**:
  - Muitas regras são **redundantes** ou sem **utilidade**
  - Redundante se {A,B,C} → {D} e {A,B} → {D} têm o mesmo **suporte** e **confiança**.
- No arcabouço original de regras de associação, **suporte** e **confiança** são as únicas medidas.
- Outras **medidas de interesse** podem ser usadas:

P. Tan, V. Kumar, and J. Srivastava. **Selecting the right interestingness measure for association patterns**. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 2002. pages 32-41.

Existem várias medidas de interesse na literatura.

Algumas medidas são boas para certas aplicações, mas não para outras.

Que critérios deveríamos usar para determinar se uma medida é boa ou ruim?

Note que a maioria das medidas dependem do fator suporte.

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left.P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\left(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]\right.$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $\left.-P(A)^2-P(\overline{A})^2\right)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Medidas têm diferentes propriedades

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 … 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 … 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 … 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 … 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 … 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 … 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 … 1 … ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 … 1 … ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}}-1\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)$ … 0 … $\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

OBS: Detalhes sobre essa Tabela podem ser encontrados em (TAN et al., 2002).

Medidas de Interesse

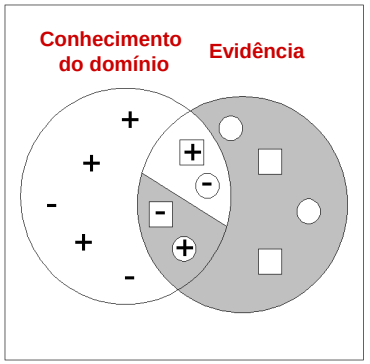
- **Medida objetiva**:
  - Ranquear os padrões com base nas estatísticas computadas nos dados em análise.
  - Existem mais de 20 medidas de interesse na literatura.
  - **Suporte** e **confiança** são as medidas de interesse tradicionais.
  - Outras medidas podem ser usadas para ajudar a selecionar padrões relevantes.
  - **Ex.** (**em inglês**): **Lift, Leverage, Conviction, Laplace, Gini, mutual information, Jaccard, etc**.

Medidas de Interesse ...

- **Medida subjetiva**:
  - Ranquear os padrões de acordo com a **interpretação do analista**.
  - Um padrão é subjetivamente interessante se ele **contradiz a expectativa do analista**.
  - Um padrão é subjetivamente interessante se ele gera informação nova (**mesmo com baixo suporte**).

Medidas de Interesse ...

- É necessário compreender as expectativas dos analistas (**conhecimento do domínio**).



- É necessário combinar as **expectativas** dos analistas com as **evidências** encontradas nos dados (**padrões descobertos**).

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no **Weka**.

Matriz de dados x Matriz de Transações

Cliente	Produto	Qtde
C1	Pão	5
C1	Leite	1
C2	Pão	3
C2	Café	1
C2	Manteiga	2
C2	Ovos	12
C3	Leite	2
C3	Fralda	2
C3	Cerveja	6
C3	Coca	3
C4	Açúcar	1
C4	Leite	3
C4	Café	1

• • • • • • • •

TID	Lista de Itens
T1	Pão, Leite
T2	Pão, Café, Manteiga, Ovos
T3	Leite, Fralda, Cerveja, Coca
T4	Açúcar, Leite, Café
T5	Pão, Leite, Fralda, Coca

Matriz de Transações

Matriz de Dados

Associação com valores somente de um atributo.

Associação: nova abordagem

- Considere o seguinte dataset sobre **qualidade de frutas**:

Casca Nominal	Cor Nominal	Tamanho Nominal	Polpa Nominal	**Risco** Nominal
aspera	marrom	grande	dura	baixo
aspera	verde	grande	dura	baixo
lisa	vermelho	grande	macia	alto
aspera	verde	grande	macia	baixo
aspera	vermelho	pequena	dura	baixo
lisa	vermelho	pequena	dura	baixo
lisa	marrom	pequena	dura	baixo
aspera	verde	pequena	macia	alto
lisa	verde	pequena	dura	alto
aspera	vermelho	grande	dura	baixo
lisa	marrom	grande	macia	baixo
lisa	verde	pequena	macia	alto
aspera	vermelho	pequena	macia	baixo
lisa	vermelho	grande	dura	alto
lisa	vermelho	pequena	dura	baixo
aspera	verde	pequena	dura	alto

Associação com vários atributos.

Associação: nova abordagem ...

- **Exemplos de regras geradas [sup. 5% e conf. 90%]**

1. Casca=aspera Tamanho=grande **4** ==> Risco=baixo **4** **conf:(1)**
2. Tamanho=pequena Risco=alto **4** ==> Cor=verde **4** **conf:(1)**
3. Cor=verde Risco=alto **4** ==> Tamanho=pequena **4** **conf:(1)**
4. Cor=verde Tamanho=pequena **4** ==> Risco=alto **4** **conf:(1)**
5. Cor=vermelho Tamanho=pequena **4** ==> Risco=baixo **4** **conf:(1)**
6. Cor=marrom **3** ==> Risco=baixo **3** **conf:(1)**
7. Casca=aspera Cor=vermelho **3** ==> Risco=baixo **3** **conf:(1)**
8. Tamanho=grande Polpa=dura Risco=baixo **3** ==> Casca=aspera **3** **conf:(1)**
9. Casca=aspera Tamanho=grande Polpa=dura **3** ==> Risco=baixo **3** **conf:(1)**

... ...

Roteiro da Aula

- Motivação e relevância.
- Regras de associação:
  - Definição e exemplos;
  - Conceitos básicos.
- Geração de regras de associação: Complexidade
- Problemas na seleção de regras.
- Medidas de interesse.
- Preparação de dados para associação.
- Exemplos de geração de regras no **Weka**.

Tipos de dados usados em Associação

- Somente **inteiro** ou **nominal**.
- Se o atributo for **numérico** (**valores reais**), ele precisa ser **discretizado**.
- **Exemplo**: Discretização dos atributos temperatura máxima (**tempMax**) e **NDVI** para cana-de-açúcar no estado de São Paulo (**Romani et al., 2008**):
- **Jaboticabal – Jaú**:
  - **REGRA**: tempMax[29-30]) ⇒ NDVI[0.56-0.63].
- **Araraquara – Luis Antônio**:
  - **REGRA**: tempMax[24-25]) ⇒ NDVI[0.56-0.66].

Exemplos no Weka

- Exemplo para associação com **atributos numéricos** (**valores reais**):

1. Selecionar o dataset **IRIS**:
2. Rodar o **Apriori** sem transformação de dados.
3. Qual foi o resultado encontrado?
4. Fazer a discretização de atributos e rodar o algoritmo **Apriori** novamente.
5. Verifique os resultados após a discretização?

Exemplos no Weka ...

- Exemplo para associação com **atributos nominais**:

1. Selecionar o dataset **WEATHER.NOMINAL**:
2. Rodar o **Apriori** sem transformação de dados.
3. Quais foram os resultados encontrados?
4. Altere o parâmetro **car** (**class association rules** para **TRUE**). Agora você pode escolher o consequente das regras geradas (**classIndex**). Depois rode o algoritmo **Apriori** novamente.
5. Verifique a diferença dos resultados gerados com relação aos valores default para **car** e **classIndex**.