Mineração de Dados

Grimaldo Lopes de Oliveira

www.aprendavirtual.com

Perfil: http://br.linkedin.com/in/grimaldo

PROGRAMA

- Motivação para utilizar Mineração de Dados
- Processo KDD
- Aplicações da Mineração de Dados
- O que é um Padrão nos Dados
- Tarefas de Mineração de Dados
- Uso da ferramenta WEKA
- Laboratórios práticos

Motivação

- A informatização dos meios produtivos permitiu a geração de grandes volumes de dados:
  - Transações eletrônicas;
  - Novos equipamentos científicos e industriais para observação e controle;
  - Dispositivos de armazenamento em massa;
- Aproveitamento da informação permite ganho de competitividade: "*conhecimento é poder (e poder = $$!)*"

Motivação

- Os recursos de análise de dados tradicionais são inviáveis para acompanhar esta evolução
- *"Morrendo de sede por conhecimento em um oceano de dados"*
- Gigantismo do problema de análise de dados para tomada de decisão:
  - BD da Wal-Mart: 20 milhões de transações por dia
  - Data Warehouse da Mobil: 100 TB
  - BD da NASA: recebe de satélites 50 GB por hora

Motivação

- Solução:
  - ferramentas de automatização das tarefas repetitivas e sistemática de análise de dados
  - ferramentas de auxílio para as tarefas cognitivas da análise
  - integração das ferramentas em sistemas apoiando o processo completo de descoberta de conhecimento para tomada de decisão

Exemplo Preliminar

- Um problema do mundo dos negócios: entender o perfil dos clientes
  - desenvolvimento de novos produtos;
  - controle de estoque em postos de distribuição;
  - propaganda mal direcionada gera maiores gastos e desestimula o possível interessado a procurar as ofertas adequadas;
- Quais são meus clientes típicos?

Exemplo

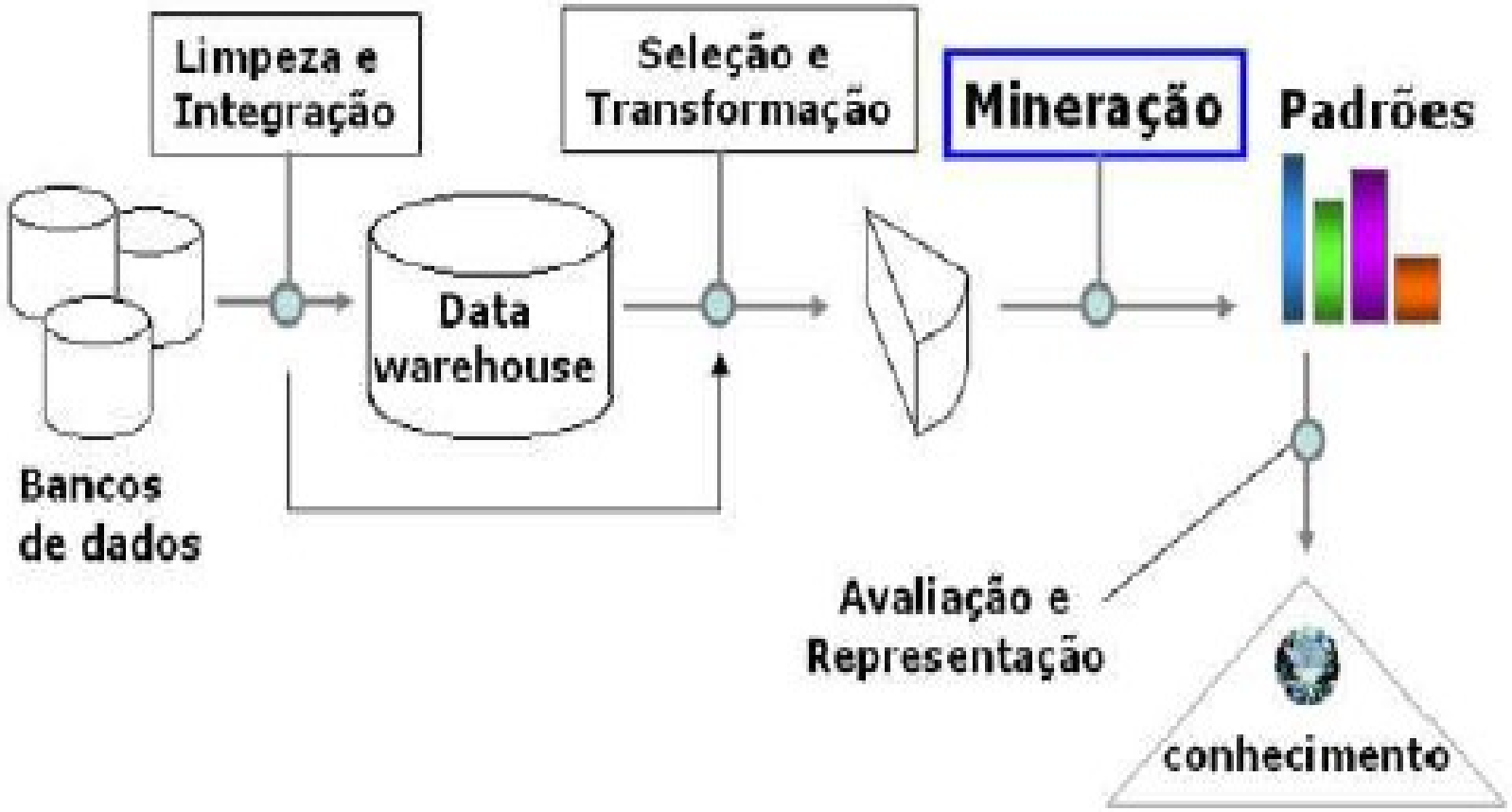
Como descubro estes DADOS ????

Descoberta de Conhecimento em Bancos de Dados = KDD

- “O processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis de uma fonte de dados”;
- “Torture os dados até eles confessarem”;
- O que é um padrão interessante ? (válido, novo, útil e interpretável)

Conceitos

Etapas do KDD



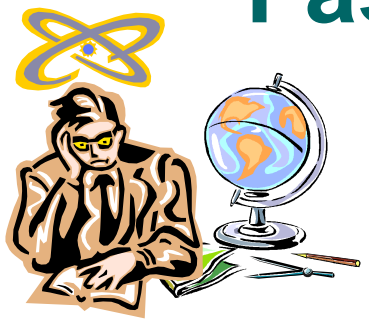
KDD x Data Mining

- Mineração de dados é o passo do processo de KDD que produz um conjunto de padrões sob um custo computacional aceitável;
- KDD utiliza algoritmos de *data mining* para extrair padrões classificados como "conhecimento". Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados;

Mineração de Dados

A mineração de dados também conhecida como "**garimpagem os dados"** é o processo de extração de informações, sem conhecimento prévio em um grande banco de dados, das características dos dados e seu uso são para tomada de decisões

Fases da Mineração de Dados



Definição do problema

Datamining

Análise das relações descobertas

Aplicação das relações descobertas

Análise dos Resultados

Aplicações da Mineração de dados

- Comércio
  - Real
  - Virtual
- Medicina
- Detecção de Fraudes
- Inteligência Competitiva
  - Concorrentes
  - Tendências do Mercado

Exemplos

- Áreas de aplicações potenciais:
  - Vendas e Marketing
    - *Identificar padrões de comportamento de consumidores*
    - *Associar comportamentos à características demográficas de consumidores*
    - *Campanhas de marketing direto (mailing campaigns)*
    - *Identificar consumidores "leais"*

Exemplos

- Áreas de aplicações potenciais:
  - Bancos
    - *Identificar padrões de fraudes (cartões de crédito)*
    - *Identificar características de correntistas*
    - *Mercado Financeiro ($$$)*

Exemplos

- Áreas de aplicações potenciais
  - Médica
    - *Comportamento de pacientes*
    - *Identificar terapias de sucessos para diferentes tratamentos*
    - *Fraudes em planos de saúdes*
    - *Comportamento de usuários de planos de saúde*

Localizando Padrões

Veja se você consegue localizar um padrão na estrutura abaixo:

ABCVYABCVKABDKCABCVUABEWLABCVO

Localizando Padrões

Passos:

1-A primeira etapa é perceber que existe uma seqüência de letras que se repete bastante;

2-Após as sequências determinadas, verificamos que elas segmentam o padrão original em diversas unidades independentes

Localizando Padrões

Veja se você encontrou estas possíveis **respostas:**

"ABCXY"
"ABCVK“
"ABDKC“
"ABCVU“
"ABEWL"
"ABCVO"

Generalize os padrões

"ABC??"
"ABD??"
"ABE??"
"AB???",

Substitua

'A‘ -> "aquisição de pão”

B' -> "aquisição de leite“

Localizando Padrões

Regras Sugeridas

- Quer dizer, na prática, que toda vez que alguém comprou pão(A), também comprou leite(B). Esses dois atributos estão associados e isto foi revelado pelo processo de descoberta de padrões.

Exemplo Padrões

- Hora de ver um exemplo

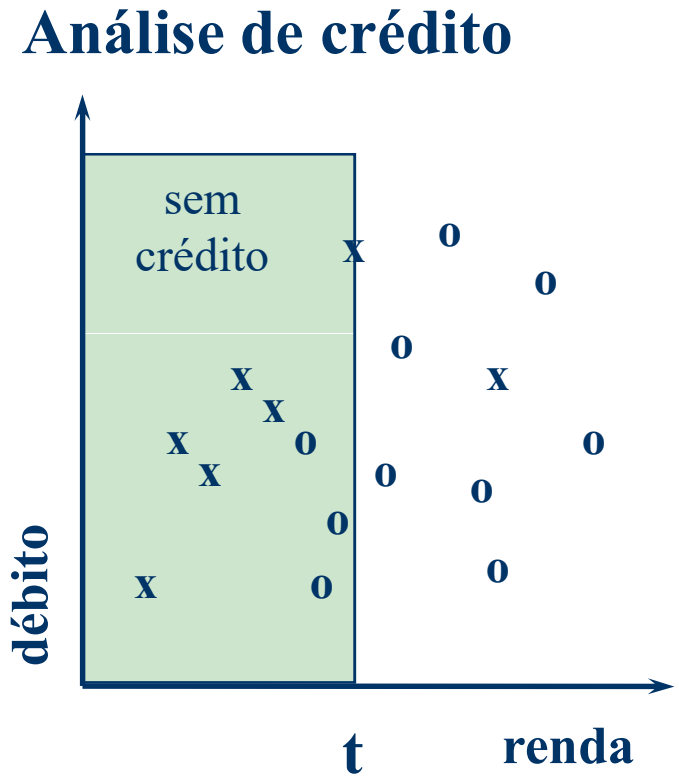
- Abra o arquivo:Exercício inicial mineração de dados - localizar padrão.doc

Quais Tarefas de Mineração são utilizadas?

Tarefas de Mineração de Dados

- Análise de Regras de Associação
- Análise de Padrões Sequenciais
- Classificação
- Análise de Clusters (agrupamentos) – Segmentação
- Análise de Outliers (exceções)
- Estimativa (ou regressão)
- Sumarização

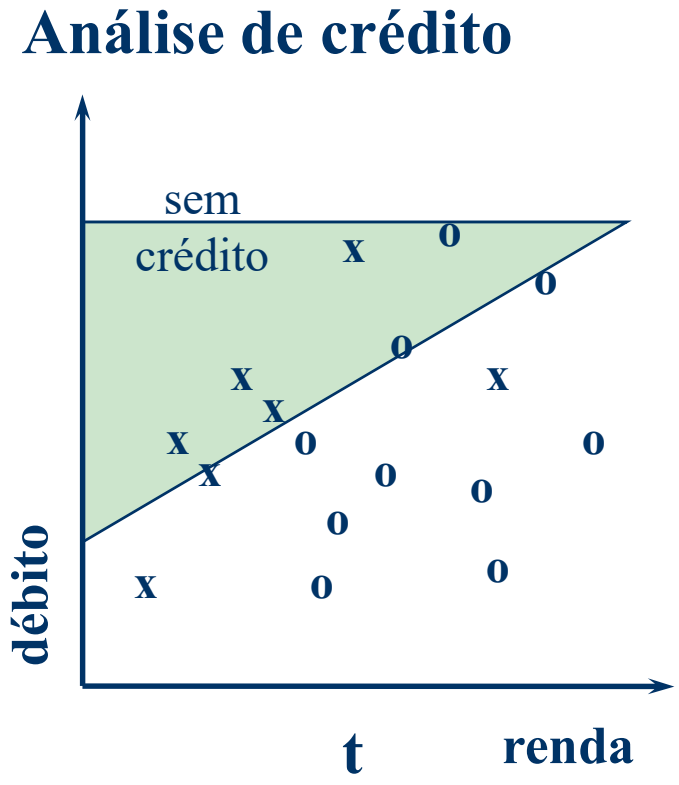
Exemplo de previsão (I)



x: exemplo recusado
o: exemplo aceito

- Um hiperplano paralelo de separação: pode ser interpretado diretamente como uma regra:
  - se a renda é menor que *t*, então o crédito não deve ser liberado
- Exemplo:
  - árvores de decisão;
  - indução de regras

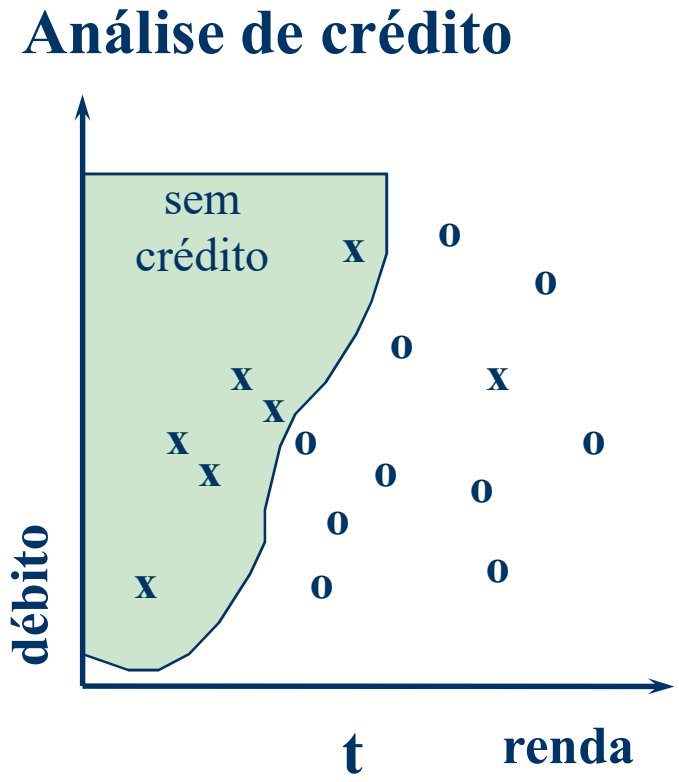
Exemplo de previsão (II)



- Hiperplano oblíquo: melhor separação:
- Exemplos:
  - regressão linear;
  - perceptron;

x: exemplo recusado
o: exemplo aceito

Exemplo de previsão (III)

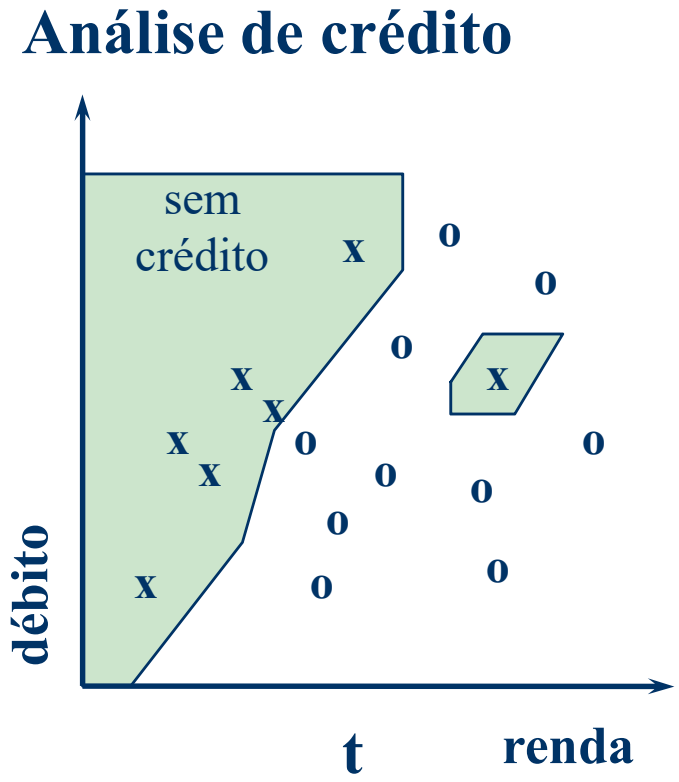


x: exemplo recusado
o: exemplo aceito

- Superfície não linear: melhor poder de classificação, pior interpretação;
- Exemplos:
  - perceptrons multicamadas;
  - regressão não-linear;

Métodos

Exemplo de previsão (IV)



x: exemplo recusado
o: exemplo aceito

- Métodos baseado em exemplos;
- Exemplos:
  - k-vizinhos mais próximos;
  - raciocínio baseado em casos;

Métodos

Análise de Clusters (agrupamentos) – Segmentação

- Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos

Análise de Outliers (exceções)

- Identificação de dados que não apresentam o comportamento geral

Estimativa (ou regressão)

- Usada para definir um valor para alguma variável contínua desconhecida

Sumarização

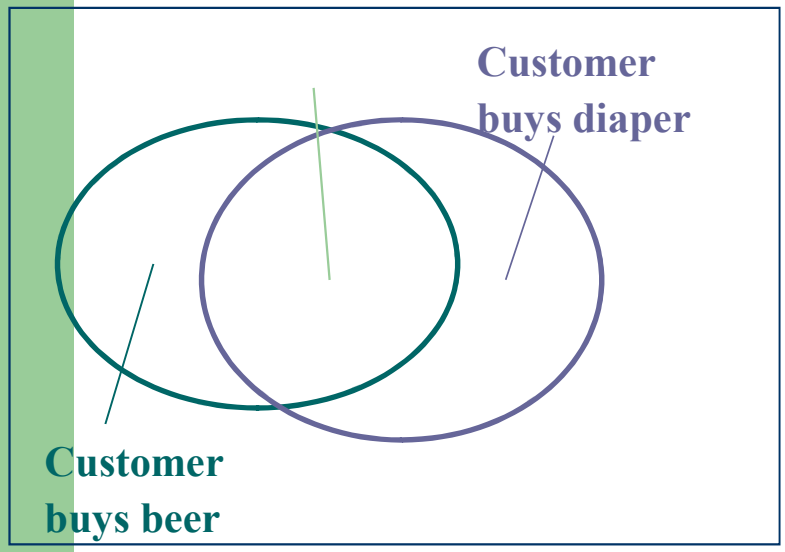
- Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados

Regras de Associação

Regras de Associação

- Mineração de associações ou de regras de associação:
  - Encontrara padrões frequentes, associações, correlações, ou estruturas causais a partir de conjuntos de itens ou objetos em DB de transações, relacionais, ou em outros repositórios de informações.
- Aplicações:
  - Análise de cestas de dados (basket data), marketing cruzado, projeto de catálogos, agrupamento, etc.

Regras de Associação



Encontrar regras $X \& Y \Rightarrow Z$ **com** suporte e confiança mínimos

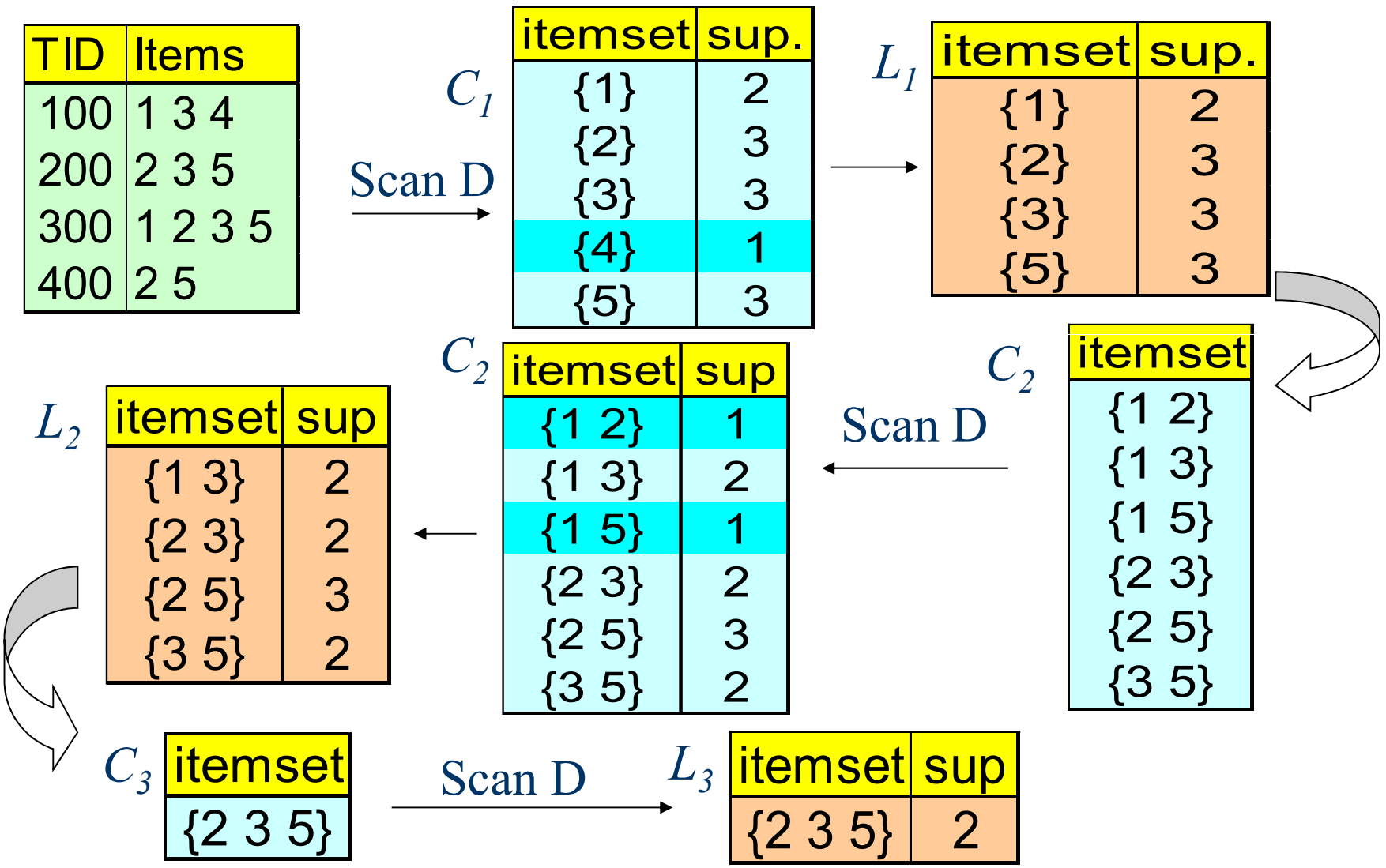
- **Suporte, s**, é a probabilidade de uma transação conter $\{X \cap Y \cap Z\}$
- **Confiança, c**, é a probabilidade condicional da transação tendo $\{X \cap Y\}$ também conter Z

Transação	Itens
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Para um suporte mínimo de 50%, e confiança mínima de 50%, tem-se:

- $A \Rightarrow C$ **(50%, 66.6%)**
- $C \Rightarrow A$ **(50%, 100%)**

Regras de Associação



Análise de Regras de Associação

ID	Compras
1	Pão, **Leite**, **Manteiga**
2	**Leite**,Açucar
3	**Leite**, **Manteiga**
4	Manteiga, Açúcar

Leite ⟶ Manteiga

$$\text{Suporte} = \frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{Total de clientes}} = 50\%$$

$$\text{Confiança} = \frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{número de clientes que compraram Leite}} = 66{,}6\%$$

Análise de Padrões Sequenciais

Itens = { TV, Vídeo , DVD, FitaDVD, ... }

ITEMSET >> ITEMSET >> ITEMSET >> ... >>ITEMSET

Análise de Padrões Sequenciais

1	{TV , Rádio} >>{DVD}
2	{Computador}
3	{TV} >> {Rádio, DVD}
4	{Rádio} >>{Comp}
5	{Comp} >> {Impressora}

< {TV} , {DVD} >

$$\text{Suporte} = \frac{\text{número de clientes que compraram TV, DVD em seqüência}}{\text{Total de clientes}} = 40\%$$

Exercício
Vamos lá!!!!

Classificação

Classificação

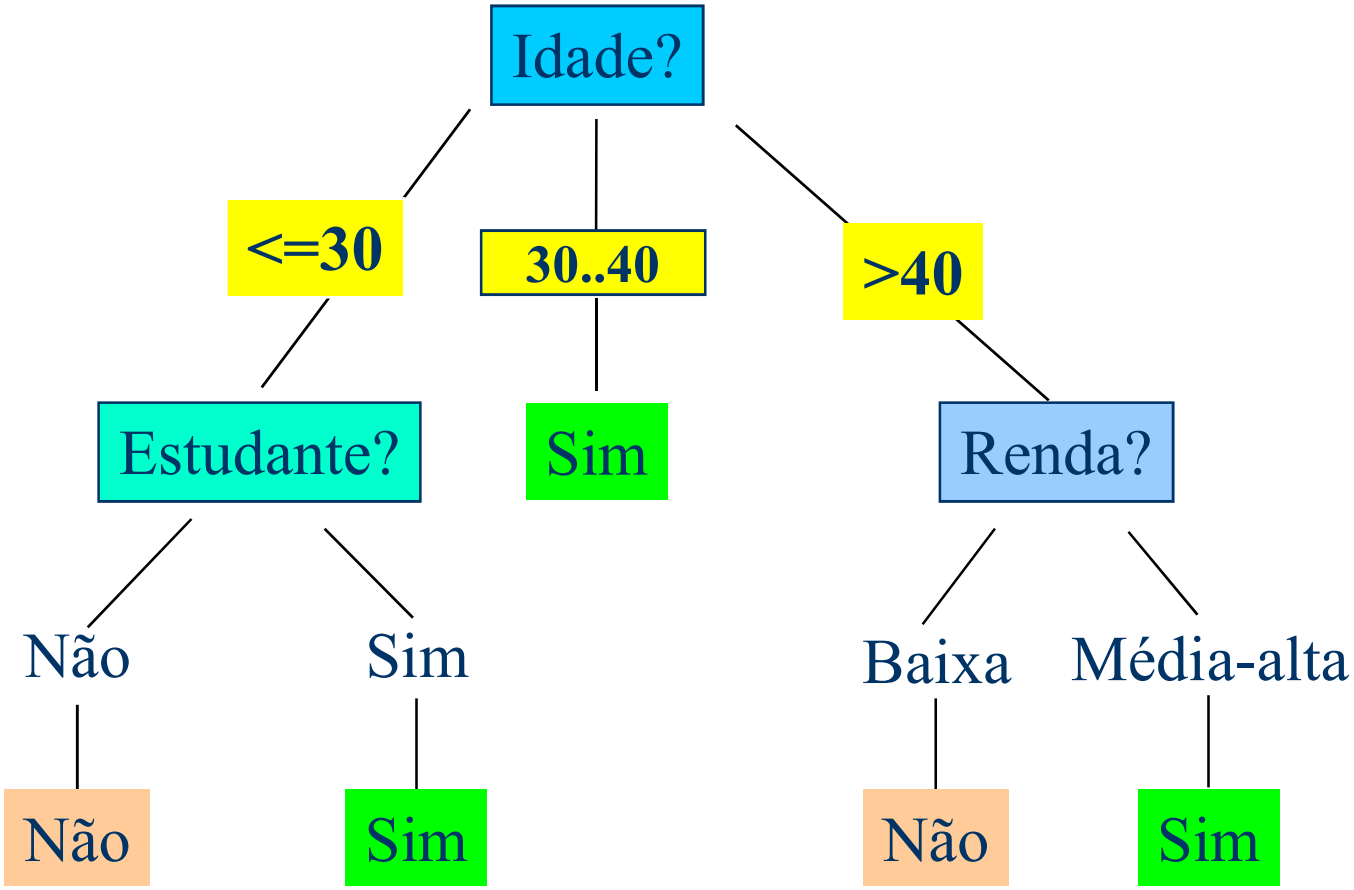
- Classificação
  - Predição dos nomes (rótulos) das classes;
  - Classifica os dados (constrói um modelo) com base no conjunto de treinamento e nos valores (rótulos) do atributo classificador, de forma a determinar a classe dos novos dados;

- Aplicações típicas
  - Aprovação de crédito, marketing dirigido, diagnóstico médico ...

Classificação

Nome	Idade	Renda	Profissão	Bom Pagador
Daniel	≤ 30	Média	Estudante	Sim
João	31..50	Média-Alta	Professor	Sim
Carlos	31..50	Média-Alta	Engenheiro	Sim
Maria	41..50	Baixa	Vendedora	Não
Paulo	≤ 30	Baixa	Porteiro	Não
Otavio	> 60	Baixa	Aposentado	Não

Classificação : Árvore de Decisão

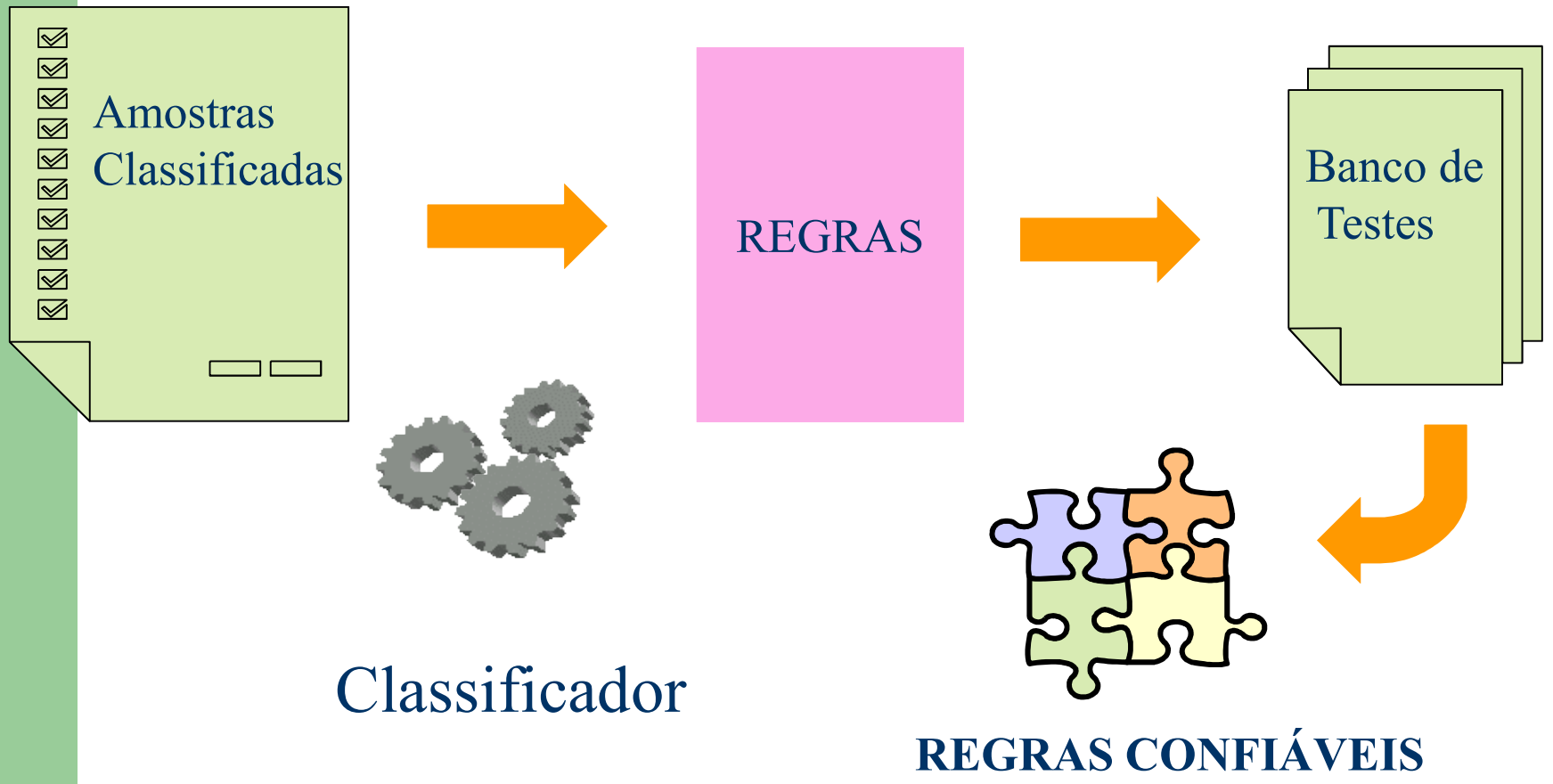


É Bom Pagador?

Exemplo : Árvore de Decisão

- Representação por regras IF-THEN:
  - Cada par (atributo, valor) forma uma conjunção;
- Regras são de mais fácil compreensão aos usuários:
  - IF Idade = “<=30” AND Estudante = “Não”
    THEN Bom_Pagador = “Não”
  - IF Idade = “>40” AND Renda = “Média-Alta”
    THEN Bom_Pagador = “Sim”

Classificação



Classificador Bayesiano

Classificador Bayesiano

- Aprendizagem probabilista: cálculo da probabilidade explícita da hipótese, de ampla aplicação em vários domínios;

- Incremental:
  - cada exemplo de treinamento pode aumentar / diminuir a probabilidade da hipótese;
  - Conhecimento a priori pode ser combinado com os dados observados;

- Previsão probabilista:
  - Várias hipótese podem ser previstas, ponderadas por suas probabilidades;
  - Fornece uma referência a ser comparada a outros métodos.

Fundamento: Teorema de Bayes

- Dado um conjunto de dados D, a probabilidade a posteriori de uma hipótese h, P(h|D) é dada por:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- A probabilidade máxima a posteriori MAP é:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Dificuldade prática: requer conhecimento inicial de muitas probabilidades, custo computacional elevado;

Exemplo: Jogar ou não Tênis

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 2/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Exemplo: Jogar ou não Tênis

- Um novo exemplo: X = <rain, hot, high, false>

- $P(X|p)\cdot P(p) =$
  $P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) = 3/9\cdot 2/9\cdot 3/9\cdot 6/9\cdot 9/14 =$
  0.010582
- $P(X|n)\cdot P(n) =$
  $P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) = 2/5\cdot 2/5\cdot 4/5\cdot 2/5\cdot 5/14 =$
  0.018286

- O exemplo X é classificado como da classe **n (não jogar).**

Redes Neurais

Redes Neurais

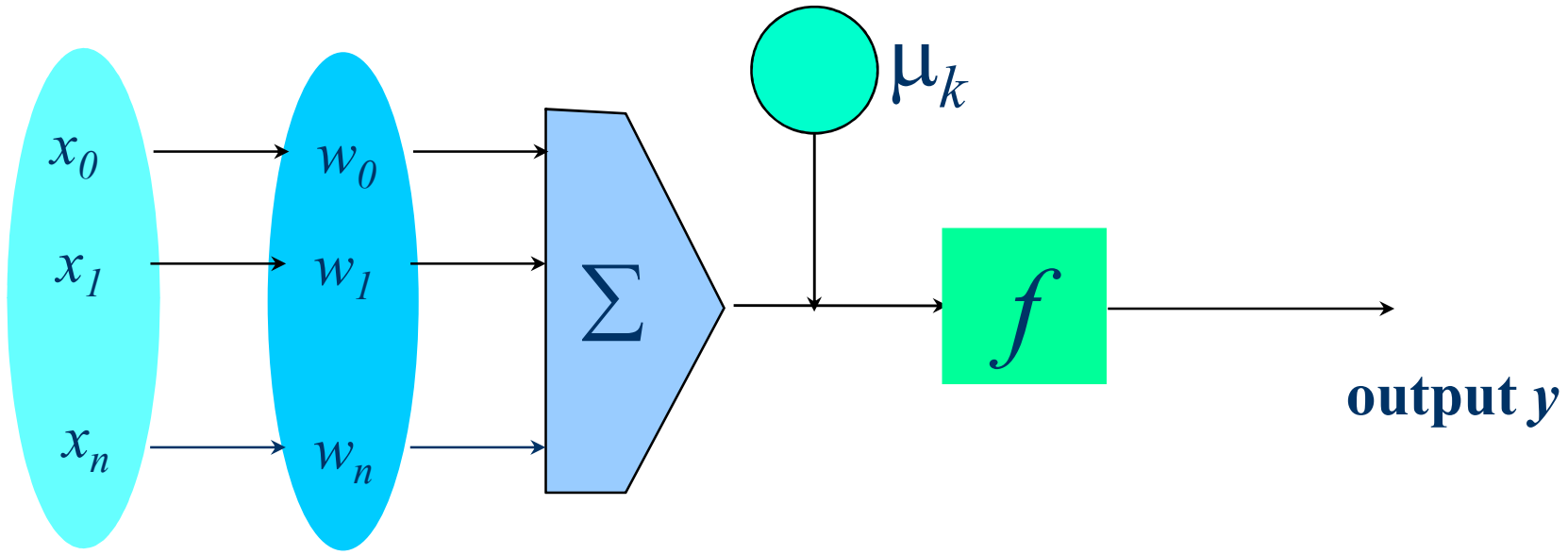
Vantagens:

- Correção de predição em geral elevada;
- Robustez, bom funcionamento na presença de ruídos;
- Saídas discretas, reais, ou mistas;
- Avaliação rápida da função de aprendizagem.

Desvantagens / crítica:

- Tempo de treinamento lento;
- Dificuldade no entendimento da função de aprendizagem (pesos);
- Difícil incorporação de conhecimento de domínio.

Um neurônio



Input vector x **weight vector w** **weighted sum** **Activation function**

- Um vetor n-dimensional x de entrada é mapeado em uma variável y por meio de um produto escalar e de um mapeamento não-linear.

Agrupamento(Clustering)

Agrupamento

Cluster: uma coleção de objetos de dados;

- Similares entre si no mesmo cluster;
- Não similares aos objetos fora do respectivo cluster;

Análise de clusters:

- Agrupamento de dados em clusters;

Agrupamento (*clustering*) é uma classificação não-supervisionada: não há classes pré-definidas.

Aplicações típicas:

- Como ferramenta para análise da distribuição dos dados;
- Como pré-processamento para outros métodos.

Aplicações gerais do agrupamento

- Reconhecimento de padrões;
- Análise de dados espaciais:
  - Criação de mapas temáticos em GIS por agrupamento de espaços de características;
  - Detecção de clusters espaciais e sua explanação em data mining;
- Processamento de imagens;
- Pesquisas de mercado;
- WWW:
  - Classificação de documentos;
  - Agrupamento de dados de weblogs para descobrir padrões similares de acesso;

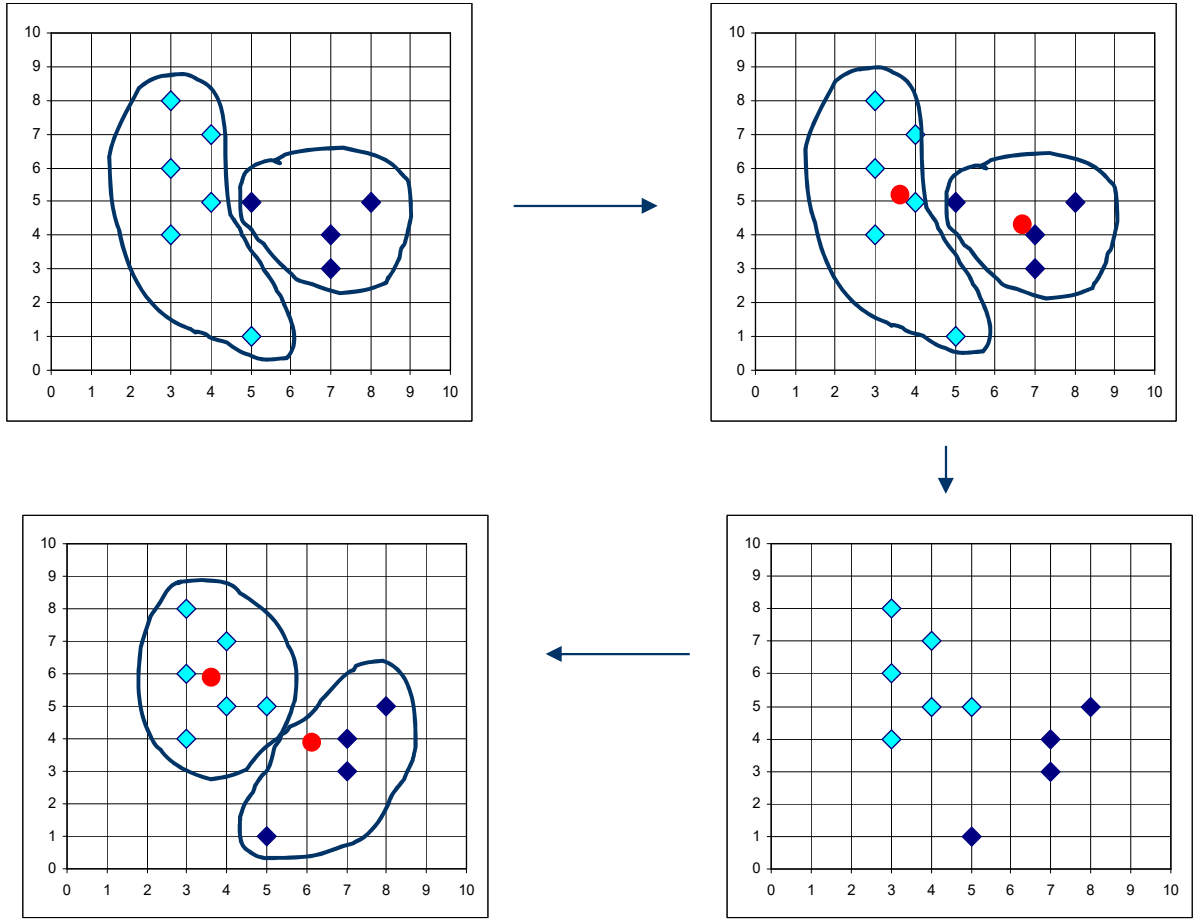
O método k-means (k-médias)

- Dado k, o algoritmo k-means é implementado em quatro passos:

1. Partição dos objetos em k conjuntos não vazios;
2. Cálculo de pontos "semente" como os centróides (médias) dos clusters das partições correntes;
3. Assinalação de cada objeto ao cluster (centróide) mais próximo de acordo com a função de distância;
4. Retorno ao passo 2 até que não haja mais alterações de assinalação.

O método k-means (k-médias)

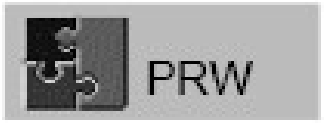
- Exemplo

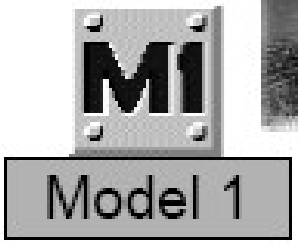


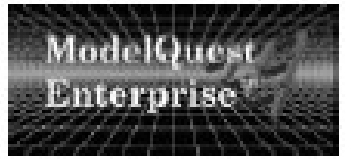
Técnicas de Mineração de Dados

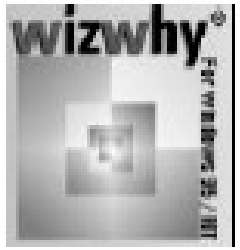
Técnica	Tarefas	Exemplos
Descoberta de Regras de Associação	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen *et al.*., 1996).
Árvores de Decisão	Classificação Regressão	CART, CHAID, C5.0, Quest (Two Crows, 1999); ID-3 (Chen *et al.*., 1996); SLIQ (Metha *et al.*., 1996); SPRINT (Shafer *et al.*., 1996).
Raciocínio Baseado em Casos ou MBR	Classificação Segmentação	BIRCH (Zhang *et al.*., 1996); CLARANS (Chen *et al.*., 1996); CLIQUE (Agrawal *et al.*., 1998).
Algoritmos Genéticos	Classificação Segmentação	Algoritmo Genético Simples (Goldberg, 1989); Genitor, CHC (Whitley, 1993); Algoritmo de Hillis (Hillis, 1997); GA-Nuggets (Freitas, 1999); GA-PVMINER (Araújo *et al.*., 1999).
Redes Neurais Artificiais	Classificação Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (Azevedo, 2000), (Braga *et al.*, 2000), (Haykin, 2001)

Data Mining Products









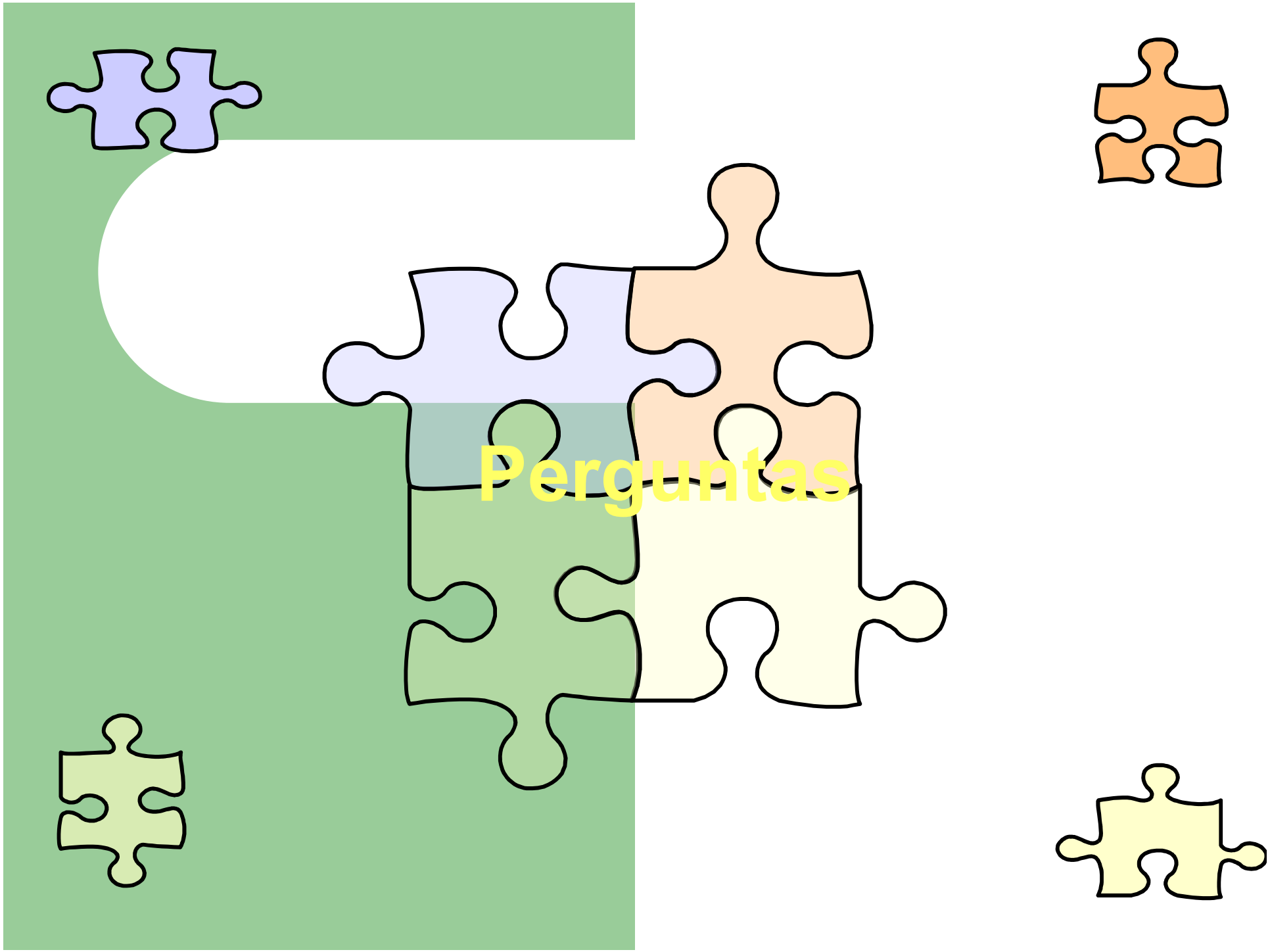
Exemplos

- Empresas de software para Data mining:
  - SAS http://www.sas.com
  - Information Havesting http://www.convex.com
  - Red Brick http://www.redbrick.com
  - Oracle http://www.oracle.com
  - Sybase http://www.sybase.com
  - Informix http://www.informix.com
  - IBM http://www.ibm.com

Algorithms	Decision Trees	Linear/Statistical	Multi-layer Perceptrons	Nearest Neighbor	Radial Basis Functions	Bayes	Rule Induction	Polynomial Networks	Generalized Linear Models	Time Series	Sequential Discovery	K Means	Association Rules	Kohonen
Clementine	√	√	√				√					√	√	√
Darwin	√		√	√										
Datamind							√							
Enterprise Miner	√	√	√		√				√	√		√	√	
GainSmarts	√	√+												
Intelligent Miner	√	√–	√		√–					√	√	√+	√	
MineSet	√					√						√	√	
Model 1	√+	√	√									√		
ModelQuest	√	√		√				√		√–				
PRW		√+	√	√	√	√						√		
CART	√													
Cognos	√													
NeuroShell			√+		√					√–				
OLPARS		√	√	√	√	√						√		√
See5	√						√							
SPlus	√	√+							√	√		√		
WizWhy							√							

Conclusões

- Data mining é um processo que permite compreender o comportamento dos dados.
- Data mining analisa os dados usando técnicas de aprendizagem para encontrar padrões e regulariedades nestes conjuntos de dados.
- É um problema pluridisciplinar, envolve Inteligência Artificial, Estatística, Computação Gráfica, Banco de Dados.
- Pode ser bem aplicado em diversas áreas de negócios


Perguntas

Referências Bibliográficas

- Técnicas de Mineração de Dados -JAI - SBC2004
  - http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf (Acesso 02/06/2005)
  - http://www.deamo.prof.ufu.br/arquivos/JAI-slides.ppt (Acesso 02/06/2005)
- Gimenes, Eduardo. “Data Mining – Data Warehouse” – Importância da Mineração de Dados em tomadas de decisão. Taquaritinga, 2000. Monografia sobre Mineração de Dados
  - http://geocities.yahoo.com.br/dugimenes/arquivos/data_mining.zip (Acesso 8/07/2005)
- Neto, Manoel Gomes de Mendonça. “Mineração de Dados”.
  - http://www.nuperc.unifacs.br/publicacoes.htm(Acesso 10/07/2005)
- Parâmetros na escolha de técnicas e ferramentas de mineração de dados
  - http://www.ppg.uem.br/Docs/ctf/Tecnologia/2002/18_279_02_Maria%20Dias_Parametros%20na%20escolha.pdf (Acesso 9/7/2005)

Referências Bibliográficas

- A Comparison of Leading Data Mining Tools (PDF format). A presentation by John F. Elder IV and Dean W.
  - http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf (Acesso 9/7/2005)
- Oliveira, Aracele G.; Garcia, Denise F. Mineração da Base de Dados de um Processo Seletivo Universitário. p.38-43.
  - http://www.dcc.ufla.br/infocomp/artigos/v3.2/art07.pdf (Acesso 31/05/2005)

Referências

- Fayyad et al. (1996). Advances in knowledge discovery and data mining, AAAI Press/MIT Press.
- Holsheimer, M. & Siebes, A.P.J.M. Data Mining: The Search for Knowledge in Databases, 1994.
- http://www-pcc.qub.ac.uk/tec/courses/datamining
- http://www.rio.com.br/~extended
- http://www.datamining.com
- http://www.santafe.edu/~kurt
- http://www.datamation.com
- http://www-dse.doc.ic.ac.uk/~kd
- http://www.cs.bham.ac.uk/~anp
- http://www.dbms.com
- http://www.infolink.com.br/~mpolito/mining/mining.html
- http://www.lci.ufrj.br/~labbd/semins/grupo1

Referências