CEFET-MG

PÓS-GRADUAÇÃO *LATO SENSU*

Curso: Banco de Dados

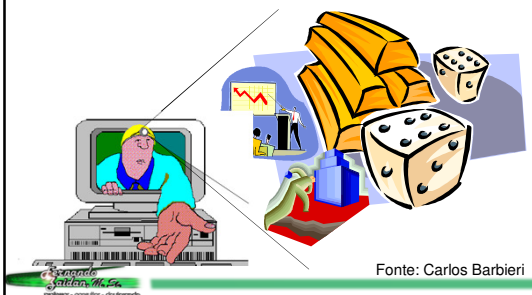
Disciplina: *Data Warehouse e Business Intelligence*

Professor: Fernando Zaidan

Unidade 5 – Data Mining

2012

DATA MINING



Fonte: Carlos Barbieri

INTELIGÊNCIA NOS NEGÓCIOS

- TOMADA DE DECISÃO TRANFORMANDO DADO EM INFORMAÇÃO
- DUAS VERTENTES:
  - ANALÍTICAS=>INTERPRETAÇÃO
  - GARIMPAGEM=>DESCOBRIMENTO DE TENDÊNCIAS-PADRÕES-CRIANDO MODELOS

TOMADA DE DECISÃO

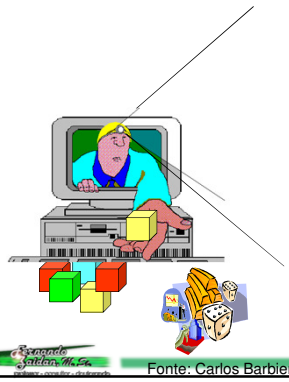
BUSINESS INTELLIGENCE

INFERÊNCIA-PADRÕES

Fonte: Carlos Barbieri

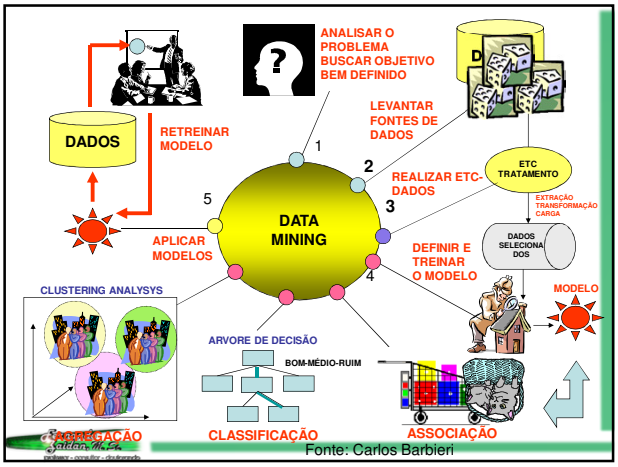
INFORMAÇÕES ANALÍTICAS

DATA MINING



- DESCOBERTA DE PADRÕES E TENDÊNCIAS EM GRANDES VOLUMES DE DADOS
- VIABILIZADO POR:
  - < CUSTO ARMAZENAMENTO
  - < CUSTO DE PROCESSADORES
  - > COMPETITIVIDADE NEGOCIAL
- DIFERE DE OLAP-ANÁLISE ONDE CERTAS INFLUÊNCIAS SÃO COLOCADAS NA PESQUISA OLAP
- MINING-ANÁLISE DE TODAS AS INTERRELAÇÕES DE INFLUÊNCIA
- CHAMADA DE KD(KNOWLEDGE DISCOVERY)
- USA MODELOS-CRIADO A PARTIR DOS PADRÕES DOS DADOS
- PADRÕES: COMPORTAMENTOS REPETITIVOS QUE PERMITAM PREVISÃO:
  - EX: 13451345134?
- ALGORÍTMOS ESTATÍSTICOS:
  - MECANISMOS PARA CRIAÇÃO DE MODELOS
  - ÁRVORE DE DECISÃO
  - ASSOCIAÇÃO
  - ANÁLISE DE REGRESSÃO
  - ALGORITMOS GENÉTICOS
  - REDES NEURAIS ARTIFICIAIS

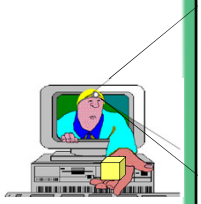
Fonte: Carlos Barbieri



Fonte: Carlos Barbieri

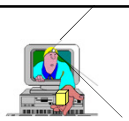
FATORES CRÍTICOS DE PROJETOS DE DATA MINING

- SABE COM DETALHES QUAL É O SEU PROBLEMA?
- VOCÊ SABE EXATAMENTE O QUE NECESSITA?
  - ANÁLISES COMPLEXAS, TENDÊNCIAS ESCONDIDAS, INFERÊNCIAS?
- VOCÊ TEM OS DADOS NECESSÁRIOS, NA GRANULARIDADE NECESSÁRIA, NO DETALHE NECESSÁRIO?
  - PREPARAÇÃO É FUNDAMENTAL
  - DADOS INTERNOS E EXTERNOS
- DATA MINING É SOLUÇÃO?
  - ACERTO DE EXPECTATIVA
- TEM USUÁRIO PATROCINADOR?
- VOCÊ SABE QUE DM É UM PROJETO DE ASPECTO NEGÓCIOS(NÃO SÓ UM PROJETO ISOLADO)
- VOCÊ TEM, OU ESTÁ DISPOSTO A TER:
- BOA ARQUITETURA TECNOLÓGICA?
  - CUSTO, COERENTE COM AS NECESSIDADES
  - GRANDES VOLUMES DE DADOS E PROCESSAMENTO
- EQUIPE COM TÉCNICAS E TÉCNICOS COM SKILL EM DM E ANALISTAS EM BUSINESS INTELLIGENCE?
  - CONHECIMENTOS EM PACOTES CLÁSSICOS COMO SAS, SPSS, ORACLE



Fonte: Carlos Barbieri

DATA MINING ESTRATÉGIAS



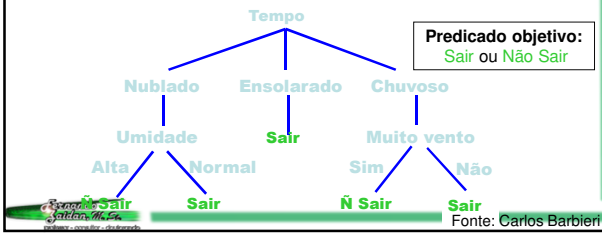
- **MINING DIRETO**
- **OS ATRIBUTOS CONHECIDOS(INPUTS) SÃO FORNECIDOS AO ENGINE DE DM PARA PRODUZIR(BASEADO NO MODELO E PADRÕES DEFINIDOS), OS VALORES PREVISTOS(PREDICTED/ TARGETS)**

- **MINING INDIRETO**
- **OS ATRIBUTOS CONHECIDOS SÃO FORNECIDOS AO ENGINE DE DM PARA PRODUZIR CONJUNTOS DE COMPORTAMENTO SEMELHANTES. NÃO PRODUZEM VALORES. SERVEM PARA FORMAR NUVENS OU COLONIAS DE OBJETOS COM CARACTERÍSTICAS SEMELHANTES**
- **NORMALMENTE SERÃO USADAS COMO PREPARAÇÃO PARA OUTRAS TÉCNICAS. POR EXEMPLO AGRUPA-SE DADOS PARA TRATAMENTO VIA ARVORE DE DECISÃO**

Fonte: Carlos Barbieri

Técnicas

- Árvores de decisão- 1º Exemplo
  - Representações simples do conhecimento
  - Utilização de regras condicionais
  - A partir de um conjunto de valores decide SIM ou NÃO
  - Mais rápida e mais compreensível que redes neurais
  - Exemplo: Sair ou não de acordo com o tempo

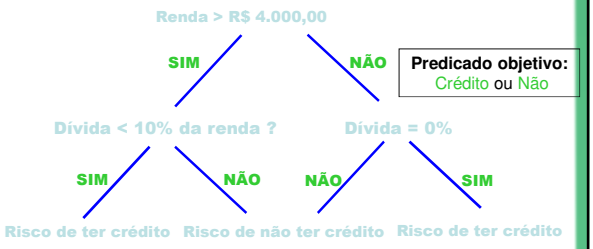


Fonte: Carlos Barbieri

Técnicas

- Árvores de decisão: 2º Exemplo

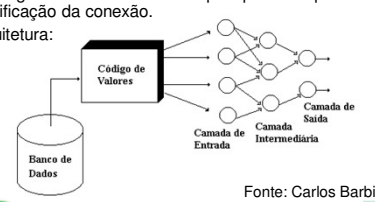
Classificação de um indivíduo com risco de ter ou não crédito



Fonte: Carlos Barbieri

Técnicas

- Redes Neurais:
  - É uma abordagem computacional que envolve desenvolvimento de estruturas matemáticas com a habilidade de aprender. (modelo do sistema nervoso para aprender)
  - Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios/nós), que possuem entrada, saída e processamento.
  - São organizados em camadas que aprendem pela modificação da conexão.
  - Arquitetura:



Fonte: Carlos Barbieri

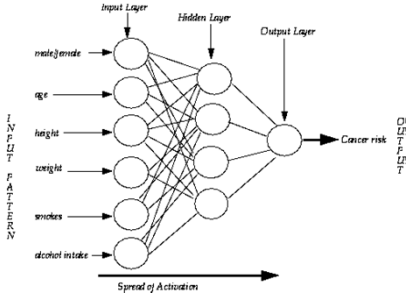
Técnicas

- Redes Neurais:
  - Para construir um modelo neural, nós primeiramente "adestramos" a rede em um dataset de treinamento e então usamos a rede já treinada para fazer predições.
  - Problemas:
    - Não retorna informação a priori
    - Complicação para treinar em uma grande base de dados
    - Entrada não pode ser dados alfa-numéricos (mapear para numérico)
    - Nenhuma explanação dos dados é fornecida (caixa preta)

Fonte: Carlos Barbieri

Técnicas

- Redes Neurais:
  - Exemplo prático: risco de câncer

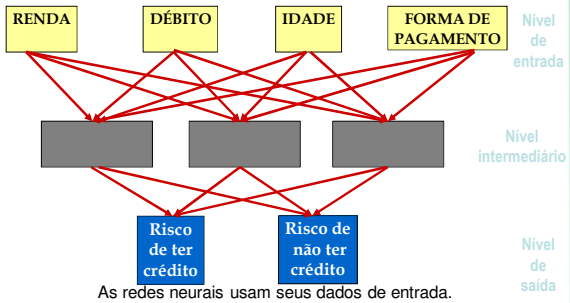


Data mining - Clementine User Guide

Fonte: Carlos Barbieri

- Redes Neurais: - 2º Exemplo

Técnicas



As redes neurais usam seus dados de entrada.
Atribui pesos nas conecções entre os atributos (neurônios).
E obtém um resultado (risco de ter ou não crédito) no nível de saída.

Fonte: Carlos Barbieri

DATA MINING E ESTATÍSTICA CONCEITOS COMUNS

- **POPULAÇÃO : GRUPO DE CASOS INCLUIDOS NO MODELO. AS CARACTERÍSTICAS DEVEM SER BEM DEFINIDAS.**
- **AMOSTRA: SE O CONJUNTO DE CASOS INCLUÍDOS NÃO É GERENCIÁVEL, CRIA-SE UMA AMOSTRA REPRESENTATIVA DAQUELA POPULAÇÃO. HÁ QUE SE REPRESENTAR O TODO.**
- **VALORES ESTATÍSTICOS:**
  - **MÉDIA-SOMA DE VALORES / N**
  - **MODA-VARIÁVEL DE > FREQUÊNCIA**
  - **MEDIANA-VALOR CENTRAL DE UMA DISTRIBUIÇÃO**
  - **VARIÂNCIA: SOMA DAS DIFERENÇAS QUADRADAS ENTRE OS VALORES E A MÉDIA**
  - **DESVIO PADRÃO: RAIZ QUADRADA DA VARIÂNCIA**

Fonte: Carlos Barbieri

DATA MINING E ESTATÍSTICA CONCEITOS COMUNS

- **CORRELAÇÃO: QUANDO UMA VARIÁVEL MUDA EM FUNÇÃO DA MUDANÇA DE OUTRA**
- **REGRESSÃO: MEDE NUMERICAMENTE O VALOR DESTA INFLUÊNCIA/RELAÇÃO. O QUANTO UMA MUDA, MUDANDO X NA OUTRA**

Fonte: Carlos Barbieri

DATA MINING TREINAMENTO DO MODELO

- **TREINAR O MODELO SIGNIFICA ANALISAR AS VARIÁVEIS DE INPUT E OBTER OS PADRÕES**
- **OS PADRÕES SÃO COLOCADOS EM MODELOS CONSTRUÍDOS**
- **O MODELO TREINADO FAZ PARTE DO "LEARNING MACHINE", ONDE OS ALGORÍTMOS SÃO AUTOMATICAMENTE ADAPTADOS PARA MELHORAR BASEADO NA EXPERIÊNCIA DE ANÁLISE DE DADOS HISTÓRICOS,CRIAÇÃO DE PADRÕES E SUA VALIDAÇÃO CONTRA NOVOS DADOS**

Fonte: Carlos Barbieri

DATA MINING CUIDADOS NA OBTENÇÃO SIGNIFICÂNCIA DOS MODELOS

- **MINING OBJETIVA DESCOBRIR PADRÕES ESCONDIDOS**
- **POIS BEM: ALGUMAS ANÁLISES PODEM LEVAR A INFERÊNCIAS INCONSISTENTES. POR EX:**
- **UM BANCO FRANCÊS DEFINIU QUE DONOS DE CARROS VERMELHOS ERAM MELHORES PAGADORES DE EMPRÉSTIMOS.**
- **ISSO É CHAMADO "OVERFITTING"**
- **TÉCNICAS ESTATÍSTICAS EXISTEM PARA ABALISAR A RELEVÂNCIA DE CERTAS VARIÁVEIS NUM MODELO ESTATÍSTICO:**
  - **ANÁLISE DE CHI-QUADRADO**
  - **PRUNNING(PODA-CASO DE ÁRVORES DE DECISÃO)**
  - **CROSS VALIDATION**

Fonte: Carlos Barbieri

- Exemplo (1) - Fraldas X cervejas
  - O que as cervejas tem a ver com as fraldas ?
  - homens casados, entre 25 e 30 anos;
  - compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa;
  - Wal-Mart otimizou às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas;
  - Resultado: o consumo cresceu 30%

Fonte: Carlos Barbieri

- Exemplo (2) - Lojas de Departamento
  - Aplicou 1 milhão de dólares em técnicas de data mining
  - Reduziu de 51000 produtos para 14000 produtos oferecidos em suas lojas.
  - Exemplo de anomalias detectadas:
    - Roupas de inverno e guarda chuvas encalhadas no nordeste
    - Batedeiras 110v a venda em SC onde a corrente é 220v

Fonte: Carlos Barbieri

Exemplos

- Exemplo (3) - Banco
  - Selecionou entre seus 36 milhões de clientes
    - Aqueles com menor risco de dar calotes
    - Tinham filhos com idades entre 18 e 21 anos
    - Resultado em três anos o banco lucrou 30 milhões de dólares com a carteira de empréstimos.

Fonte: Carlos Barbieri

Obrigado e Bons Estudos!

Zaidan

A persistência é o caminho do êxito.

Charles Chaplin