Tarefas de Mineração de Dados

Prof. Me. Ricardo Ávila

ricardo.avila@outlook.com.br

Tarefas de DCBD (KDD)

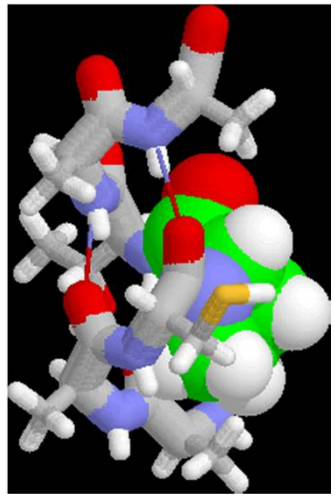
- Classificação [Preditivo]
- Clustering [Descritivo]
- Regras de associação [Descritivo]
- Padrões seqüenciais [Descritivo]
- Outliers [Preditivo]

Exemplos de Tarefas de Classificação

- Predizer se um tumor é **benigno** ou **maligno**

- Classificar transações de cartões de crédito como **legítimas** ou **fraudulentas**

- Classificar estruturas secundárias de proteínas como alpha-helix, beta-sheet, or random coil



- Categorizar textos como da área de finanças, previsão de tempo, esportes, cultura, etc.

Classificação: Aplicação 1

- Marketing direto
  - Objetivo: Reduzir o custo de postagem na oferta para um *conjunto alvo* de consumidores mais prováveis de comprar um novo produto.
  - Abordagem:
    - Usar os dados de um produto similar oferecido anteriormente.
    - Sabemos quais consumidores compraram e quais não compraram. Esta decisão *{compra, não compra}* forma o *atributo classe*.
    - Coletar várias informações demográficas, de estilo de vida e de interações com a empresa de todos estes clientes.
      - Tipo de atividade, local da moradia, rendimentos, estado civil, etc.
    - Usar esta informação como atributos de entrada para gerar um modelo de classificação.

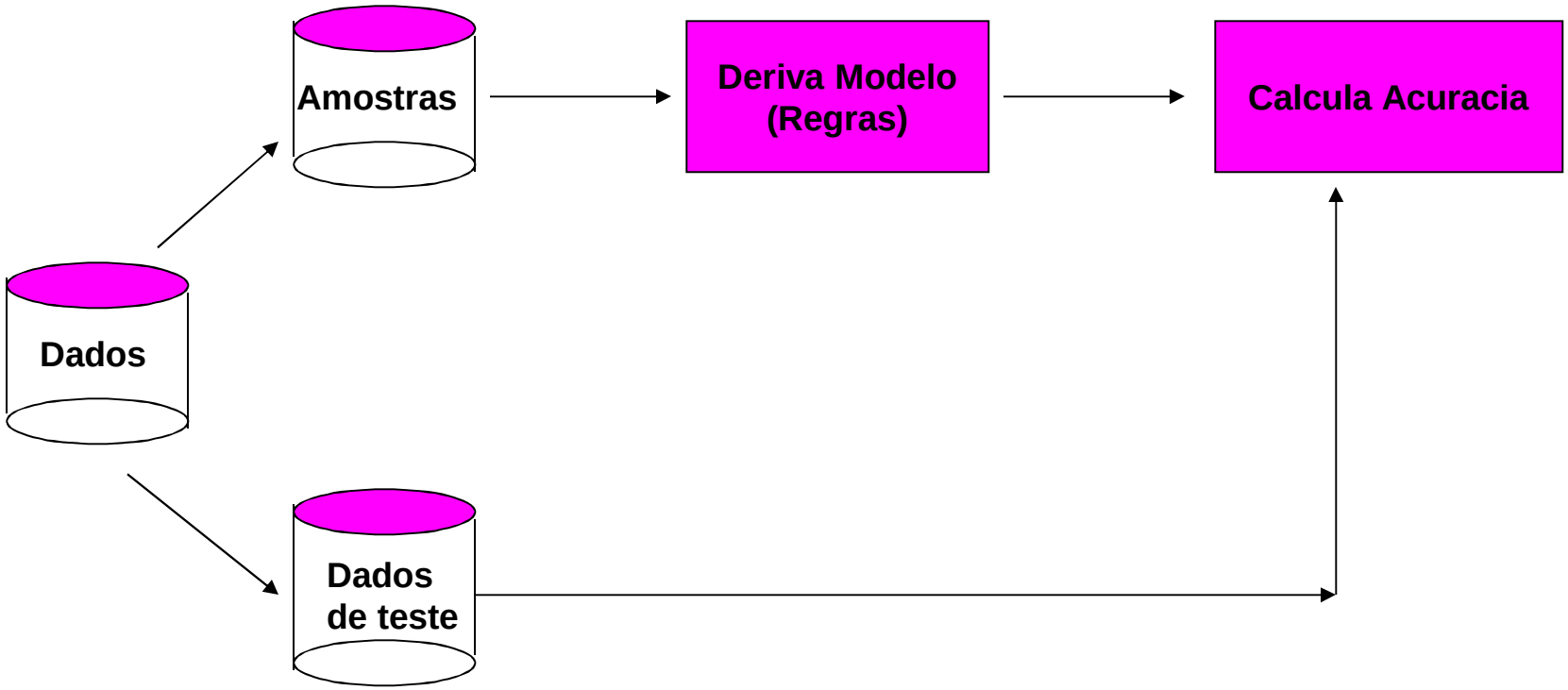
Classificação: Aplicação 2

- Detecção de fraudes
  - Objetivo: identificar casos de fraude em transações com cartão de crédito.
  - Abordagem:
    - Usar as transações do cartão de crédito e as informações do proprietário como atributos.
      - Quando um consumidor compra, o que ele compra, onde ele compra, compra a vista ou a prazo, etc
    - Rotular as transações passadas como fraude ou não. Isto forma o atributo classe.
    - Gerar um modelo de classificação para as transações.
    - Usar este modelo para detectar fraudes observando as novas transações .

Classificação: Aplicação 3

- Conservação de clientes:
  - Objetivo: prever se é provável que um cliente de uma empresa de telefone celular passe para um concorrente.
  - Abordagem:
    - Usar um registro detalhado das transações de cada cliente antigo e atual para obter os atributos.
      - Com que frequência o cliente faz ligações, para quem ele liga, a que horas ele liga mais freqüentemente, sua renda, estado civil, etc.
    - Rotular os clientes como fiéis ou infiéis a empresa.
    - Gerar um modelo.

Processo de Classificação



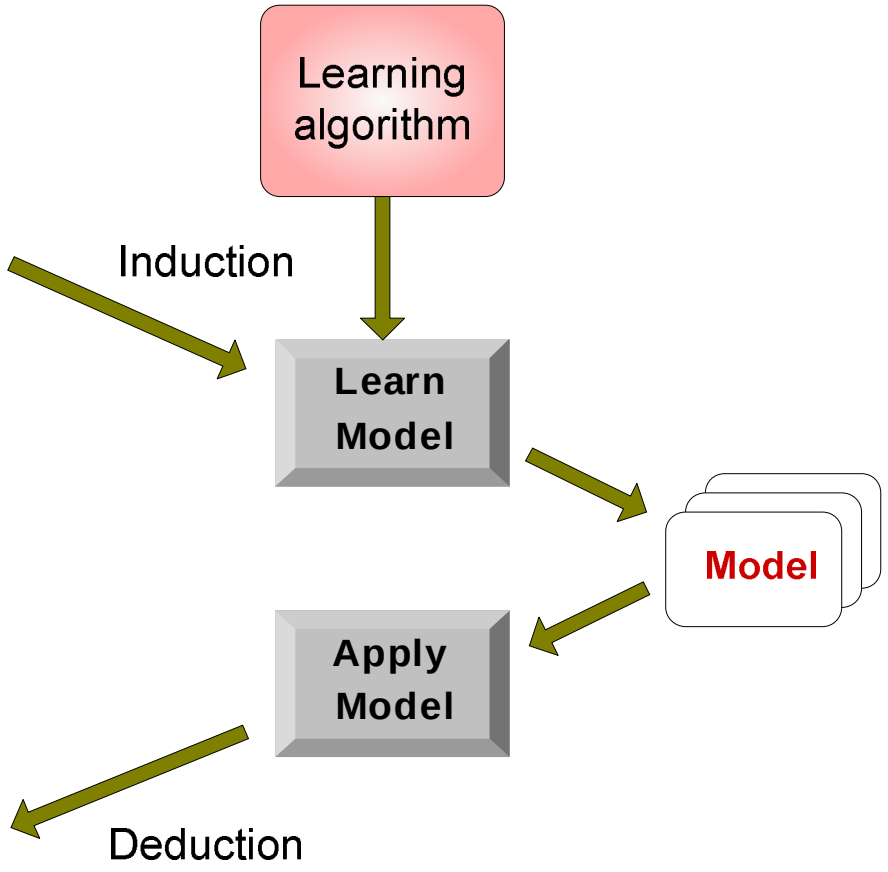
Ilustrando a Tarefa de Classificação

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classificação: definição

- Dada uma coleção de registros (*conjunto de treinamento*)
  - Cada registro contém um conjunto de *atributos*, e um dos atributos é a *classe*.
- Encontre um *modelo* para o atributo classe como uma função dos valores dos outros atributos
- Objetivo: definir a classe para novos registros tão acuradamente quanto possível.
  - Um *conjunto de teste* é usado para determinar a acurácia do modelo. Normalmente, o conjunto de dados é dividido em conjunto de treinamento e conjunto de teste, com o conjunto de treinamento usado para a construção do modelo e o conjunto de teste para validação.

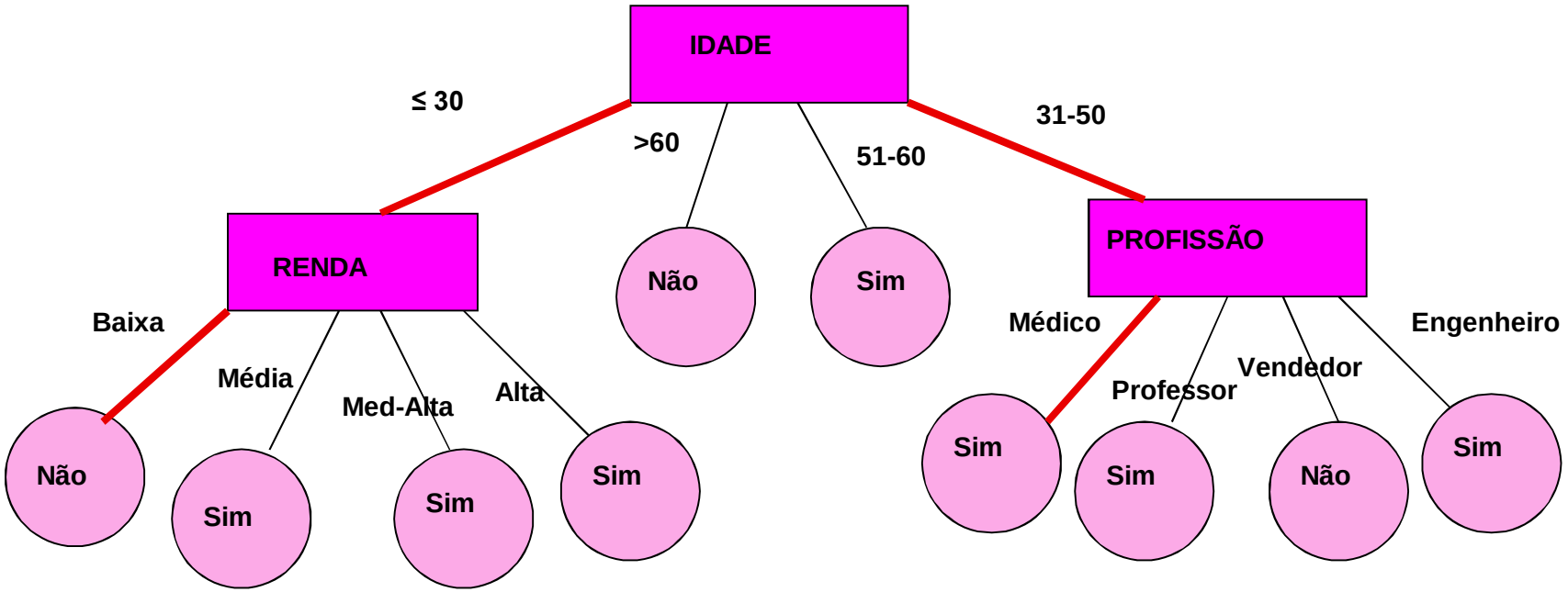
Exemplo...

Classe: compra produto Eletrônico

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

SE. **Idade ≤ 30** **E** **Renda = Média** **ENTÃO** **Compra-Produto-Eletrônico = SIM.**

Exemplo de Árvore de Decisão



Se Idade ≤ 30 e Renda= Baixa então Não compra Eletrônico

Se Idade = 31-50 e Profissão=Médico então compra Eletrônico

Clustering

Clustering (formação de agrupamentos)

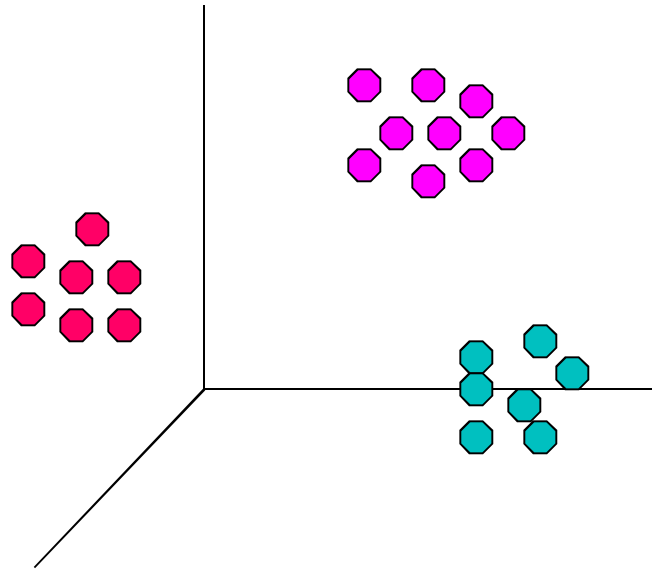
- Dado um conjunto de dados, cada um com um conjunto de atributos, e uma medida de similaridade entre eles, encontre clusters (grupos) tais que:
  - Dados de um grupo são mais similares entre si que com dados de outros grupos
  - Dados de grupos diferentes são menos similares entre si.
- Medidas de similaridade:
  - Distância Euclidiana, para atributos contínuos
  - Outras medidas específicas do problema.

Clustering: exemplo

- Clustering em espaço 3-D baseado em distância euclidiana.

Distâncias intracluster são minimizadas

Distâncias intercluster são maximizadas



Clustering: Aplicação 1

- Segmentação de mercado:
  - Objetivo: subdividir um mercado em diferentes subconjuntos de clientes onde cada subconjunto possa ser selecionado como objetivo específico de marketing a ser alcançado.
  - Abordagem:
    - Obter diferentes atributos de clientes baseado em informações geográficas e de estilo de vida dos clientes
    - Encontrar grupos (clusters) de clientes similares.
    - Medir a qualidade dos clusters observando padrões de compra entre clientes do mesmo cluster versus entre clientes de outros clusters

Clustering: Aplicação 2

- Clustering de documentos:
  - **Objetivo**: encontrar grupos de documentos que são similares entre si baseado em termos importantes que aparecem nos documentos.
  - **Abordagem**: identificar termos que ocorrem freqüentemente em cada documento. Criar uma medida de similaridade baseada na freqüência dos diferentes termos. Usar esta medida para a formação dos grupos.
  - **Ganho**: os clusters podem ser usados em Recuperação de Informações para relacionar um novo documento ou termo de pesquisa a clusters de documentos.

Exemplo de clustering de documentos

- Dados utilizados: 3204 artigos do jornal Los Angeles Times.
- Medida de similaridade: quantas palavras são comuns nestes documentos (após a filtragem de algumas palavras).

Category	*Total Articles*	*Correctly Placed*
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering de ações da bolsa

- Observe os movimentos das ações a cada dia.
- Dados: ação-{UP/DOWN}
- Medida de similaridade: Duas ações são similares se os eventos descritos por elas freqüentemente acontecem juntos no mesmo dia.
  - Foram usadas regras de associação para quantificar a medida de similaridade.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Regras de Associação

Regras de associação: Definição

- Dado um conjunto de registros, cada um com um conjunto de itens de uma certa coleção;
  - Produza regras de dependência que vão predizer a ocorrência de um item baseado na ocorrência de outros.

TID	*Items*
1	guaraná, leite, pão
2	cerveja, pão
3	cerveja, fralda, guaraná, leite
4	cerveja, fralda, leite, pão
5	fralda, guaraná, leite

Regras descobertas:

{leite} --> {guaraná}
{fralda, leite} --> {cerveja}

Regras de associação: Aplicação 1

- Marketing e promoção de vendas:
  - Considere a seguinte regra descoberta

    {Paçoquinha, … } → {Batata Frita}
  - Batata Frita como consequente: Pode ser usada para determinar o que deve ser feito para incrementar a sua venda.
  - Paçoquinha no antecedente: Pode ser usado para ver que produtos podem ser afetados se a loja deixar de vender Paçoquinha.
  - Paçoquinha no antecedente *e* Batata Frita no consequente: Pode ser usado para ver que produtos poderiam ser vendidos com Paçoquinha para promover a venda de Batata Frita!

Regras de associação: Aplicação 2

- Gerenciamento de prateleiras de supermercado.
  - Objetivo: identificar itens que são comprados juntos por um grande número de clientes.
  - Abordagem: processar os dados das transações de compra obtidos com os códigos de barras para encontrar dependências entre itens.
  - Uma regra clássica--
    - Se um cliente compra fralda e leite ele tem uma boa probabilidade de comprar também cerveja.
    - Portanto, não fique surpreso de encontrar pacotes de cerveja próximo das fraldas!

Regras de associação: Aplicação 3

- Gerência de inventário:
  - Objetivo: uma empresa de consertos de eletrodomésticos quer antecipar a natureza dos consertos nos aparelhos dos seus clientes de forma a ter em seus veículos de serviço peças de reposição, de modo a poder realizar o conserto na hora, sem precisar voltar à casa dos clientes
  - Abordagem: Analisar os dados de consertos anteriores em termos de ferramentas e peças necessárias para descobrir padrões de co-ocorrência.

Padrões sequenciais: Definição

- Dado um conjunto de *objetos*, com cada objeto associado com a sua *linha de eventos*, encontre regras com forte dependência sequencial entre diferentes eventos.

(A B) (C) ⟶ (D E)

Padrões sequenciais: exemplos

- Em transações de vendas
  - Livraria de informática:

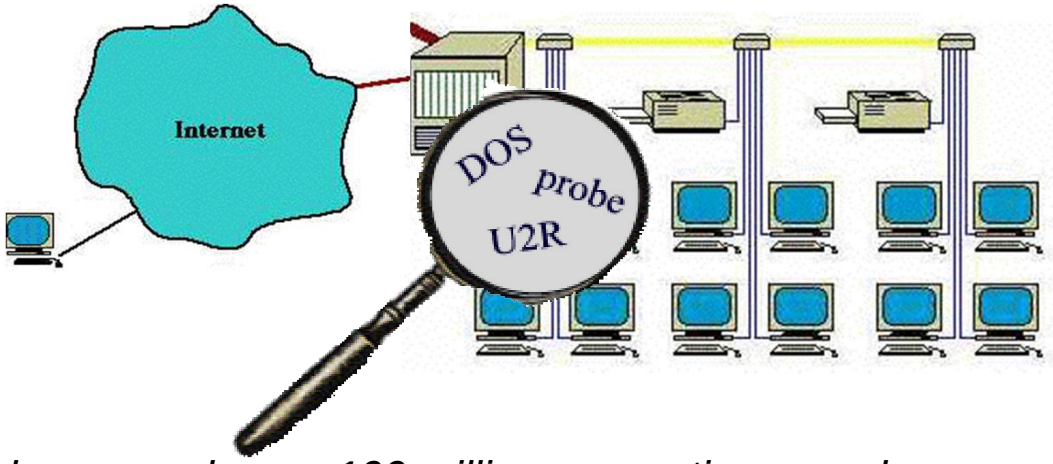
    (Intro_To_Visual_C) (C++_Primer) → (Perl_for_dummies,Tcl_Tk)
  - Loja de artigos esportivos:

    (tenis) (raquete, bolas) → (moleton)

Detecção de desvios

- Determinar desvios significativos do comportamento normal
- Aplicações:
  - Detecção de fraudes em cartões de crédito

  - Detecção de invasão em redes de computadores



Typical network traffic at University level may reach over 100 million connections per day

Desafios para Data Mining

- Escalabilidade
- Dimensionalidade
- Dados complexos e heterogêneos
- Qualidade dos dados
- Propriedade e distribuição dos dados
- Preservação da privacidade
- Dados em fluxo contínuo