Preparação de Dados para Mineração de Dados (**AP-532A** / **MT803B**)

Stanley Robson de Medeiros Oliveira
Professor Colaborador

Agosto de 2017

Agenda

- Apresentação dos **Alunos** + **Docente**
  - Formação + Situação Atual
  - Expectativas
- Apresentação da Disciplina
  - Critérios de Avaliação
  - Programa
- Dicas sobre revisão de artigos
- Breve Introdução à **Mineração de Dados**

Informações Gerais

- **Professor Responsável**: Stanley Robson de Medeiros Oliveira.
- **AP532A**: 45 horas (3 créditos)
- **MT803B**: 60 horas (4 créditos)

Metodologia

- Aulas teóricas expositivas.
- Discussões em sala de aula.
- Atividades em sala de aula.
- Exercícios (**revisões** e **listas de exercícios**).
- Aulas práticas em laboratório.
- Apresentação de estudos de caso.

Sobre o professor

- Formação
  - **Graduação** – Ciência da Computação (UFCG – 1990).
  - **Mestrado** – Ciência da Computação (UFCG – 1996).
  - **Doutorado** – Ciência da Computação **Mineração de Dados** (Universidade de Alberta, Canadá – 2004).
- Atuação Profissional
  - **Desenvolvedor** (Embrapa).
  - **Analista de Sistemas** (Embrapa).
  - **Pesquisador** (Embrapa).
  - **Docência** (ITI/EEP, IPEP, IBTA, Unicamp/Feagri, IMECC).

Critérios de Avaliação

- **Frequência**: Indispensável.
- **Horário das aulas**: Sextas das 9:00 às 12:00h.
- **Exercícios para sedimentar o conhecimento:**
  - Revisões de artigos (**pelo menos 2 artigos p/ aluno**).
  - Listas de Exercícios (**individual + grupo**).
  - Proposta do projeto final (**entregar até a 8ª aula**).
- **Avaliação**:
  - Prova em Sala de Aula (35%) + Exercícios (20%).
  - Trabalho de Conclusão (45%).

Planejamento das Aulas

- **Aula 1 – 11/08/2017**
  - Informações gerais sobre o curso:
    - **Ementa, metodologia, avaliação dos alunos**.
  - Dicas sobre como revisar artigos científicos.
  - **Exercício**: Definição de um artigo para revisão.
  - Breve revisão sobre tarefas de mineração de dados.

Planejamento das Aulas ...

- **Aula 2 – 18/08/2017**
  - **Aspectos relevantes** na fase de preparação de dados.
  - Problemas relacionados à **qualidade dos dados**:
    - Valores faltantes, ruídos e redundância.
  - Procedimentos para **limpeza dos dados**:
    - Técnicas para substituição de valores faltantes.
    - Técnicas para reduzir o ruído nos dados.
    - Técnicas para eliminar a redundância nos dados.
  - Integração de dados de **múltiplas fontes**.

Planejamento das Aulas ...

- **Aula 3 – 25/08/2017**
  - Preparação de dados para **Regras de Associação**:
    - Conceitos básicos.
    - Geração de regras de associação.
    - O Algoritmo Apriori.
    - Efeitos da distribuição do suporte nos datasets.
    - Problemas na seleção de regras.
    - Medidas de avaliação de regras de associação.
    - Tipos de dados usados em associação.
    - Exemplos de geração de regras no Weka.

Planejamento das Aulas ...

- **Aula 4 – 01/09/2017**
  - **Clusterização ou Agrupamentos de Dados:**
    - Conceitos e aplicações
    - Tipos de dados em clusterização
    - Transformação (**normalização**) de dados.
    - Medidas de similaridade
    - Análise da qualidade de clusters gerados
    - Métodos de Clusterização:
      - Particionamento
      - Métodos hierárquicos
      - Métodos baseados em densidade
      - Outros métodos.

Planejamento das Aulas ...

- **Aula 5 – 15/09/2017**
  - **Aula Prática – Laboratório – WEKA**
  - Aula em laboratório cobrindo os tópicos sobre **transformação de dados**, **associação** e **agrupamento de dados**, apresentados em sala de aula, usando os softwares R e WEKA.

Planejamento das Aulas ...

- **Aula 6 – 22/09/2017**
  - Análise Multivariada (**Regressão Linear**):
    - Simples.
    - Múltipla.
    - Penalizada (**LASSO**).

Planejamento das Aulas ...

- **Aula 7 – 29/09/2017**
  - Introdução ao **Aprendizado de Máquina**:
    - O processo de classificação de dados
    - Principais métodos de classificação
    - Árvores de decisão
    - Entropia e ganho de informação
    - Principais algoritmos existentes (ID3, C4.5, CART)
    - Escolha do atributo “split”
    - Mecanismos de poda .

Planejamento das Aulas ...

- **Aula 8 – 06/10/2017**
  - Introdução ao **Aprendizado de Máquina**:
    - Métodos Boosting e Bagging.
    - O algoritmo Naïve Bayes.
  - Medidas de **avaliação de modelos**:
    - Hold-out, cross validation e percentage split.
    - Ajustes de hiperparâmetros.
    - Medidas clássicas.

Planejamento das Aulas ...

- **Aula 9 – 20/10/2017**
  - Aprendizado com **classes desbalanceadas**:
    - Classes desbalanceadas: problema e desafios;
    - O algoritmo k-vizinhos mais próximos;
    - Precisão, taxa de erro e classes desbalanceadas;
    - Técnicas para medir desempenho de classificadores;
    - Tratamento para classes desbalanceadas;
    - Qual proporção de classes é melhor para aprender;
    - Como descartar ou duplicar exemplos;
    - Resultados da avaliação dos tratamentos em diversos conjuntos de dados.

Planejamento das Aulas ...

- **Aula 10 – 27/10/2017**
  - Redução de dimensionalidade:
    - Aspectos relevantes:
      - Necessidades, motivação e aplicações.
    - Principais abordagens:
      - Extração de atributos (**não-Supervisionada**);
      - Seleção de atributos (**Supervisionada**).
    - Métodos para extração de atributos:

      Análise de Componentes Principais (**PCA**).

Planejamento das Aulas ...

- **Aula 10 – 27/10/2017 (continuação) ...**
  - Métodos Supervisionados para Seleção de Atributos:
    - Filtros;
    - Força Bruta (**Brute-Force**);
    - Métodos ***Embedded***;
    - Métodos ***Wrappers***;
    - Método baseado no teste do **Qui-quadrado**;
    - Método baseado na correlação de atributos (**CFS**).
  - Estudo de caso – comparação dos métodos acima.

Planejamento das Aulas ...

- **Aula 11 – 10/11/2017**
  - **Aula Prática** – **Laboratório** – **WEKA e R**
  - Aula em laboratório cobrindo os tópicos sobre métodos para **redução de dimensionalidade** (**PCA**), métodos para **seleção de atributos** e **aprendizado com classes desbalanceadas**.

Planejamento das Aulas ...

- **Aula 12 – 17/11/2017** (Dr. Thiago T. Santos)
  - Processamento de imagens com **SciPy** e **Jupyter**:
    - A linguagem Python e o ambiente de computação interativa.
    - Arrays multimensionais em Python com NumPy.
    - Imagens com arrays.
    - Processamento de imagens com NumPy.
    - Aprendizado de máquina com scikit-learn.
    - Processamento de imagem com scikit-image.

Planejamento das Aulas ...

- **Aula 13 – 24/11/2017** (Dr. Thiago T. Santos)
  - Processamento de imagens com **SciPy** e **Jupyter**:
    - Naïve Bayes.
    - SVM.
    - Gradient Boosting.
    - Problemas de visão computacional como problemas de Aprendizado de Máquina.
    - **Exemplo**: classificação de dígitos (MINIST dataset).
    - **Exemplo**: classificação de frutos do café.

Planejamento das Aulas ...

- **Aula 14 – 01/12/2017**
  - Prova em sala de aula
  - **Assunto**: todo o conteúdo sobre preparação de dados apresentado.
  - A prova é Individual e **SEM** Consulta.
  - **Não** é permitido o empréstimo de qualquer material.
  - Usar caneta **azul** ou **preta**.

Planejamento das Aulas ...

- **Aula 15 – 15/12/2017**
  - Apresentação oral dos trabalhos finais;
  - Entrega de relatórios com resultados (**artigo**).

Disponibilização do Material Didático

https://www.ggte.unicamp.br/ea/

Bibliografia

Referências básicas:

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**, 3rd edition, Morgan Kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Springer, 2nd Edition, 2009.

PYLE, D., **Data Preparation for Data Mining**, Morgan Kaufmann, 1999.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining – Mineração de Dados**. Rio de Janeiro: Ed. Ciência Moderna Ltda, 2009. 900p.

WITTEN, I.H.; FRANK, E.; HALL, M.A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd ed. Morgan Kaufmann, Burlington, MA, 2011.

Referências Complementares:

HAND, D.J.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**, The MIT Press, 2001.

NISBET, R.; ELDER, J.; MINER, G. **Handbook of Statistical Analysis and Data Mining Applications**. Elsevier, 2009.

Dicas sobre Revisão de Artigos

Critérios para avaliar uma revisão

- **Entendimento**:
  - Compreensão sobre o problema de pesquisa e sua solução;
  - Identificação dos **pontos fracos** e **fortes** do artigo;
  - Apresentação de críticas pessoais sobre os pontos analisados.
- **Relevância**:
  - Tente identificar algum relacionamento entre o artigo revisado com outros artigos relacionados (**citados ou não no artigo revisado**).

Critérios para avaliar uma revisão

- **Respostas**:
  - Procure mencionar o problema e sua solução com as **suas próprias palavras**.
  - Descreva os principais pontos do artigo **sem detalhes desnecessários**.
  - Suas respostas devem ser **claras** e **precisas**.

Detalhes importantes

- Comece lendo o **resumo**; depois vá para as **conclusões**.
- Em seguida, leia a **introdução** e demais **seções**.
- A **introdução** deveria conter:
  - O problema de pesquisa e sua solução;
  - O objetivo – em geral é o último parágrafo;
  - Um mapa (**roteiro**) das seções do artigo.

Detalhes importantes

- Em geral, as **conclusões** ressaltam:
  - A solução do problema de pesquisa (**de forma sucinta**);
  - As contribuições do artigo;
  - Trabalhos futuros.
- Em geral, as **conclusões não** são repetições dos **resultados alcançados**.

Questões a serem respondidas

1. Qual é o **problema** que está sendo abordado?
2. Como os autores **solucionaram** o problema?
3. Quais são as **diferenças** entre o artigo que está sendo revisado com os trabalhos relacionados?
4. Quais são as **principais contribuições** do artigo?
5. Quais são os **pontos fracos** e **fortes** do artigo?
6. Como os autores poderiam **melhorar** o conteúdo e a **apresentação** do artigo?
7. Qual seria o **público alvo** para o artigo revisado?

Próximos Passos

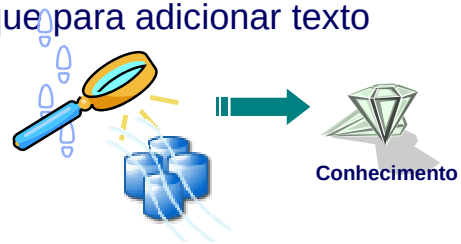
- **Revisar o artigo abaixo**:

  AMO, Sandra de. **Técnicas de Mineração de Dados**. XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informática, 31 de Julho a 6 de Agosto 2004, Salvador, Brasil.

- **Importante**:
  - Entregar a revisão até a próxima aula;
  - Será dado um "**feedback**" a todos os alunos;
  - Essa revisão **não** terá nota. Servirá como exercício.

Introdução à Mineração de dados

Clique para adicionar texto



O Problema da Explosão de Dados

- Informatização generalizada
  - negócios, governos, pesquisas
- Armazenamento
  - maior capacidade, menor custo
- Evolução da coleta de dados
  - Leitores de código de barras, Sensores, ...

O Problema da Explosão de Dados

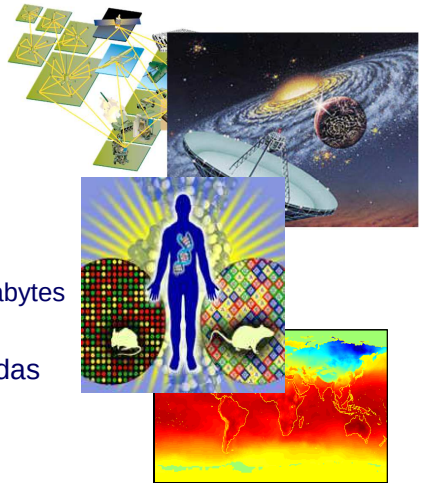
Ponto de Vista Comercial

- Quantidades gigantescas de dados são **coletados** e **armazenados** em empresas, corporações, etc:
  - Dados de **comércio eletrônico**;
  - Dados de **navegação na internet**;
  - Dados de **compras de clientes** em grandes lojas de departamentos, supermercados, etc;
  - Dados de **transações bancárias**, ou de cartão de crédito.

O Problema da Explosão de Dados

Ponto de Vista Científico: Medicina, Biologia, Engenharia

- Dados coletados e armazenados a velocidades enormes (GB/hora)
  - Sensores remotos em satélites;
  - Telescópios;
  - Microarrays gerando dados de expressões de genes;
  - Simulações científicas gerando terabytes de dados.
- Técnicas tradicionais não apropriadas para analisar tais dados:
  - Ruídos e grande dimensionalidade.

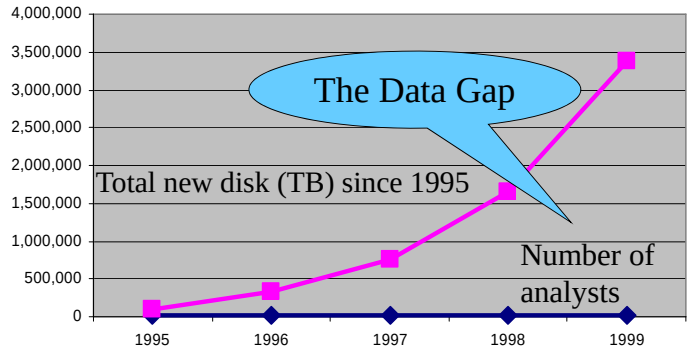


Efeitos da Explosão de Dados

- A quantidade de informação é **duplicada** a cada ano.
- Muitas empresas usam **apenas 7%** dos dados coletados.
- Somos **ricos em dados** e **pobres em informação** e **conhecimento**.
- **Necessidade**: técnicas para explorar grande volume de dados e transformar esses dados em **conhecimento**.

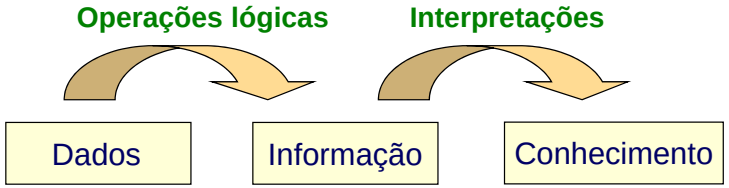
Efeitos da Explosão de Dados ...

- Existem **informações escondidas** nos dados que **não** são **evidentes** – muitos conjuntos de dados nunca são analisados.
- Analistas levariam semanas para analisar **parte** desses **dados**.



Fonte: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Dado, Informação e Conhecimento



- **Dado** é algo bruto; é a matéria-prima da qual podemos extrair informação.
- **Informação** é o dado processado, com significado e contexto bem definido.
- **Conhecimento** é o uso inteligente da informação; é a informação contextualizada e utilizada na prática.

Dado, Informação e Conhecimento

Considere o conjunto de dados abaixo sobre **eficiência de cana-de-açúcar em São Paulo** (em 2002). Dê exemplo de um tipo de **informação** e de **conhecimento** que pode ser extraído desses dados.

Município	Chuva	Tmax	Tmin	Def_Hidrica	Usina	Eficiencia
Adamantina	1286,6	30,2	17,8	353,0	Sim	media
Aguaí	1342,3	28,0	16,0	384,4	Nao	alta
Águas da Prata	1357,8	27,8	15,9	387,6	Nao	muito_baixa
Águas de Lindóia	1368,8	27,5	15,8	367,9	?	muito_baixa
Agudos	1313,2	28,7	16,8	349,7	Nao	alta
Alambari	1377,7	28,2	16,2	334,2	Nao	muito_baixa
Altair	1291,8	30,2	17,8	384,3	Nao	muito_baixa
Altinopolis	1392,3	28,7	16,5	393,3	Nao	media
Alto Alegre	1260,8	30,1	17,6	364,6	Nao	media
Álvares Machado	1338,8	30,0	17,7	337,0	Nao	muito_baixa
Americana	1323,7	28,2	16,3	358,4	Nao	alta

O que é mineração de dados?

- O processo da **descoberta de novas informações** a partir de grandes massas de dados.
- Descoberta de **padrões não triviais**, **implícitos** e **desconhecidos**) em grandes bases de dados.
- Processo de extração de **modelos úteis** de um conjunto de dados.

Diferentes pontos de vista

- **Databases**: Concentra-se em conjuntos de dados gigantes (**non-main-memory**).
- **AI** (**machine-learning**): concentra-se em modelos complexos, mesmo que os conjuntos de dados sejam pequenos.
- **Statistics**: concentra-se em modelos.

Modelos x Processamento Analítico

- Para um **especialista em Banco de Dados**, data mining é uma forma de **processamento analítico** – consultas que examinam grandes quantidades de dados.
  - O resultado é um conjunto de respostas (padrões, sequências, regras, agrupamento de objetos, etc).
- Para um **estatístico**, data mining é um processo de **inferência de modelos**.
  - O resultado é uma lista de **parâmetros do modelo**.

Mineração de Dados: Por que ?

- Técnicas de Mineração podem ajudar **analistas**:
  - Entender e prever as necessidades de clientes;
  - Descobrir fraudes;
  - Descobrir perfis de comportamento de clientes.
- Técnicas de Mineração podem ajudar **cientistas** a:
  - Classificar e segmentar dados;
  - Formular hipóteses.

Consulta versus Mineração

O que é Mineração de Dados?

Cliente	Produto	Classe	Tempo
1	Vinho	A	T1
2	Açúcar	B	T2
1	Queijo	A	T1
3	Pão	C	T3
3	Leite	C	T3
1	Vinho	A	T4
1	Queijo	A	T4
4	Leite	C	T5

- Que produtos são comprados por clientes **Classe A** ?

Consulta e Resultado

SELECT Clientes.Produto
FROM Clientes
WHERE Clientes.Classe = 'A'

RESPOSTA: Vinho, Queijo

Consulta versus Mineração

Cliente	Produto	Classe Social	Tempo
1	Vinho	A	T1
2	Açúcar	B	T2
1	Queijo	A	T1
3	Pão	C	T3
3	Leite	C	T3
1	Vinho	A	T4
1	Queijo	A	T4
4	Leite	C	T5

- Existe ligação entre a **classe social** e **produtos** comprados numa mesma transação ?

Resultado da Mineração

PADRÕES:

Classe A -> vinho, queijo

Classe C -> pão, leite

...

Mineração de Dados: O que é ?

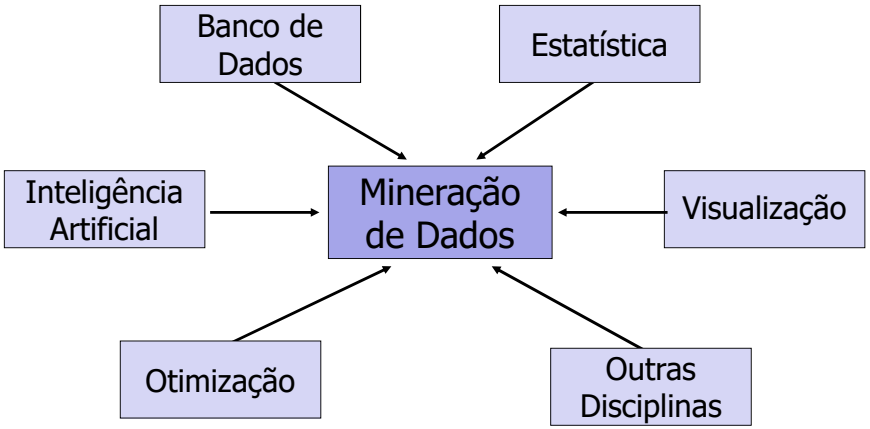
- **Não**
  1. Fazer uma consulta no Google sobre "Data Mining ".
  2. Procurar um nome numa lista telefônica.
  3. Fazer uma consulta SQL a um banco de dados.

- **Sim**
  1. Agrupar documentos similares retornados pelo Google de acordo com seu contexto.
  2. Descobrir se certos nomes aparecem com mais frequência em determinadas regiões da cidade (periferia, centro, bairros abastados,…)

Exemplos

- Qual o **perfil do cliente** que consome mais ?
- Que **produtos** são **comprados conjuntamente** ? E em sequência ?
- Meu **site web** tem uma boa estrutura ?
- É possível agrupar **produtores agrícolas** de acordo com seus **perfis**?
- Qual o **tipo de seguro** um banco poderia oferecer a cada um de seus **clientes** ?
- Que **produtos** podem ser **recomendados** para os **clientes** de uma **loja de eletrônicos** ?

Mineração: Área Multidisciplinar



Influência das Disciplinas

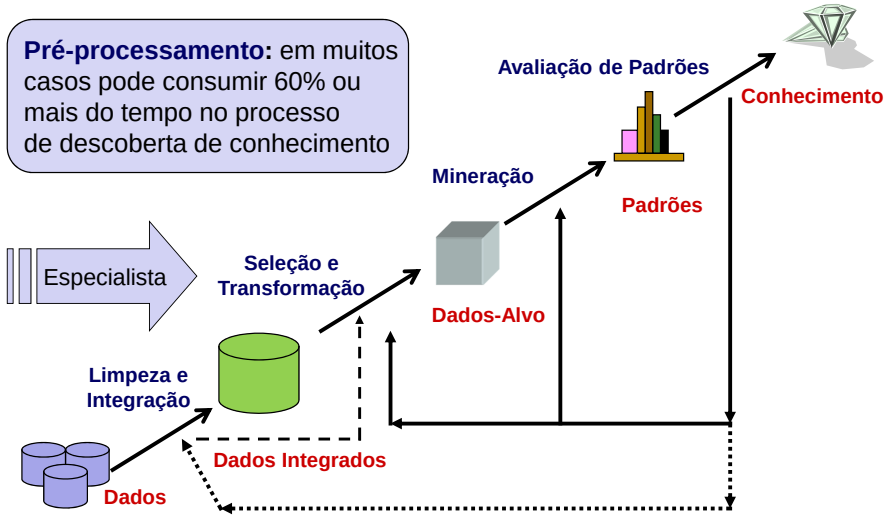
Aprendizado de Máquina (**Inteligência Artificial**)
- Algoritmos de aprendizado indutivo

Estatística
- Métodos de análise de dados
- Seleção e amostragem de dados
- Técnicas de avaliação do conhecimento

Banco de Dados
- *Data Warehousing*
- Estruturas de dados e mecanismos de busca

Visualização
- Técnicas que estimulam a percepção e a inteligência

A descoberta do conhecimento



A descoberta do conhecimento ...

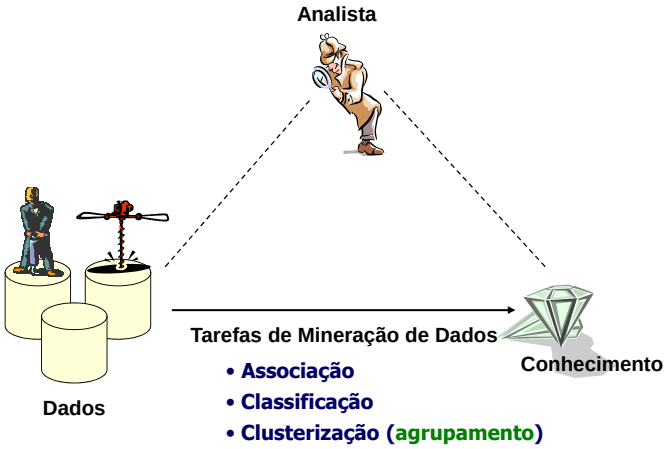
- **Característica de padrões interessantes:**
  - **Novos:** os padrões descobertos devem possuir um certo grau de **novidade.**
  - **Úteis:** os padrões descobertos devem ter potencial de conduzir a ações com **utilidade.**
  - **Compreensíveis:** padrões compreensíveis pelos humanos é um objetivo (**simplicidade**).

Padrões interessantes representam **CONHECIMENTO**

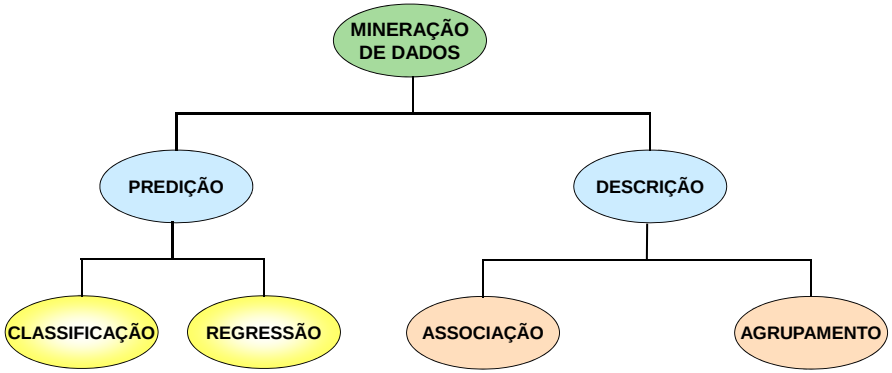
Estatística versus Mineração

Estatística	Mineração de Dados
Uma amostra dos dados é suficiente.	Quanto maior a quantidade de dados, melhor.
É baseada em hipótese (**rejeita** ou **aceita**).	É uma atividade **exploratória**.
Análise de dados primários ou **top-down** (**confirmação**)	Análise de dados secundários ou **bottom-up** (**descoberta de padrões**).
Sumariza uma amostra ou dataset (**ex.**: média, variância, distribuição).	Busca informações úteis e desconhecidas, **implícitas** nos dados (**padrões**).

Principais Tarefas de Mineração



Principais Tarefas de Mineração



Tarefas de Mineração de Dados

- **Tarefas Preditivas**
  - Predizer o valor de um determinado atributo baseado nos valores de outros atributos:

    Classificação – Predição.
- **Tarefas Descritivas**
  - Derivar « **padrões** » : correlações, tendências, anomalias, agrupamentos dentro de uma grande massa de dados:

    Regras de Associação – Padrões Sequenciais – Agrupamentos – Anomalias.

Tarefas de Mineração

- Tarefa ➡ ato de descobrir um certo **tipo de padrão**.
  - Regras de Associação;
  - Análise de Sequências;
  - Classificação;
  - Agrupamento;
  - Outliers, etc.

Classificação de Dados

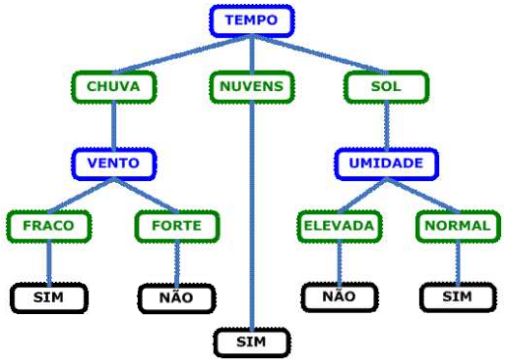


Fig.: Árvore de decisão para jogar vôlei

Tarefas versus Técnicas

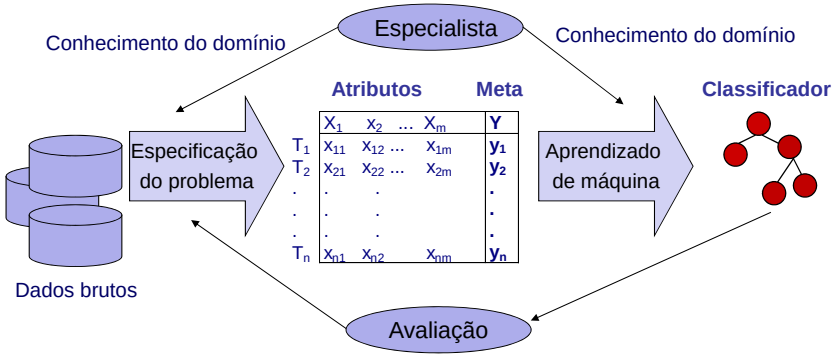
- A **TAREFA** consiste no ato de descobrir um certo **tipo de padrão**, ou seja, que tipo de regularidades temos interesse em encontrar.
  - **Exemplo**: um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos.
- A **TÉCNICA** consiste na especificação de métodos que garantam a descoberta de padrões que nos interessam.
- **Principais técnicas** utilizadas:
  - Estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda-validação.

Classificação de Dados

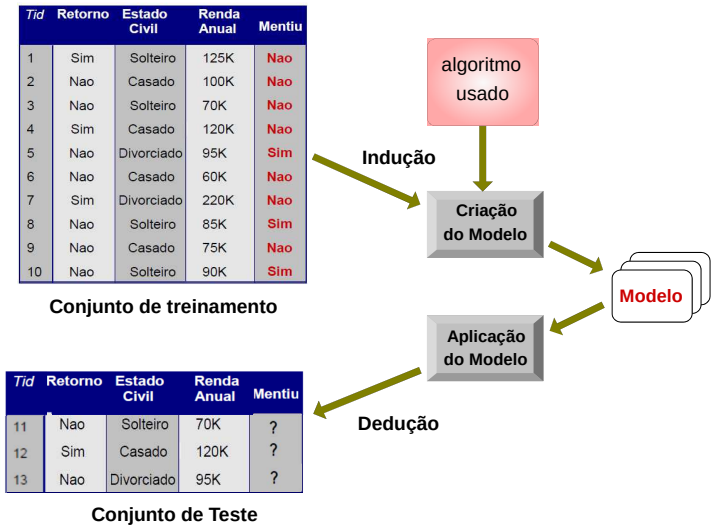
- É uma tarefa **preditiva**: define o valor de uma variável desconhecida a partir de variáveis conhecidas.
- **Passo 1**: encontrar um **modelo** para o **atributo alvo** como uma função dos valores dos outros atributos.
- **Passo 2**: registros não conhecidos devem ser associados à classe com a maior precisão possível.

Classificação de Dados ...

- **Processo de Classificação**



O que é classificação? ...



Tid	Retorno	Estado Civil	Renda Anual	Mentiu
1	Sim	Solteiro	125K	Nao
2	Nao	Casado	100K	Nao
3	Nao	Solteiro	70K	Nao
4	Sim	Casado	120K	Nao
5	Nao	Divorciado	95K	Sim
6	Nao	Casado	60K	Nao
7	Sim	Divorciado	220K	Nao
8	Nao	Solteiro	85K	Sim
9	Nao	Casado	75K	Nao
10	Nao	Solteiro	90K	Sim

Conjunto de treinamento

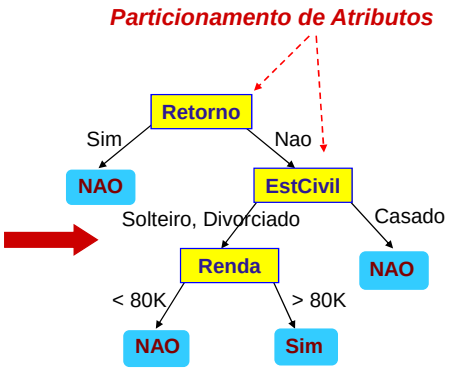
Tid	Retorno	Estado Civil	Renda Anual	Mentiu
11	Nao	Solteiro	70K	?
12	Sim	Casado	120K	?
13	Nao	Divorciado	95K	?

Conjunto de Teste

Exemplo de árvore de decisão

Tid	Retorno	Estado Civil	Renda Anual	Mentiu
1	Sim	Solteiro	125K	Nao
2	Nao	Casado	100K	Nao
3	Nao	Solteiro	70K	Nao
4	Sim	Casado	120K	Nao
5	Nao	Divorciado	95K	Sim
6	Nao	Casado	60K	Nao
7	Sim	Divorciado	220K	Nao
8	Nao	Solteiro	85K	Sim
9	Nao	Casado	75K	Nao
10	Nao	Solteiro	90K	Sim

Conjunto de treinamento



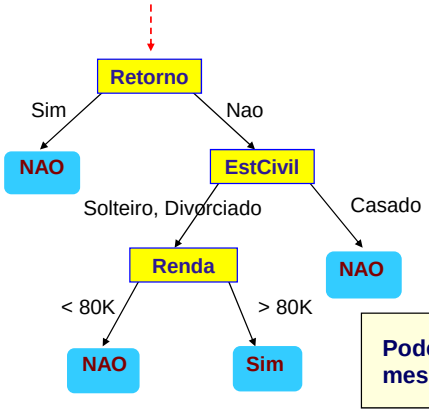
Modelo: árvore de decisão

Aplicando o conjunto de teste ao modelo

Início na raiz da árvore

Conjunto de teste

Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?

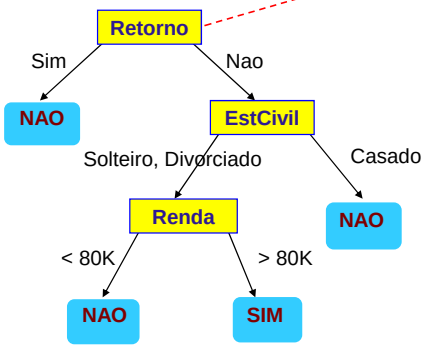


Pode haver mais de uma árvore para o mesmo conjunto de dados!

Aplicando o conjunto de teste ao modelo

Conjunto de teste

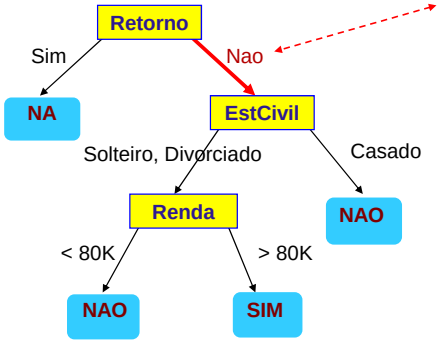
Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?



Aplicando o conjunto de teste ao modelo

Conjunto de teste

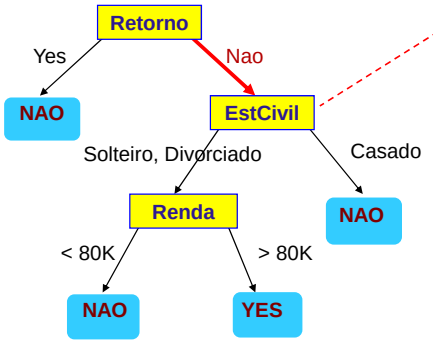
Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?



Aplicando o conjunto de teste ao modelo

Conjunto de teste

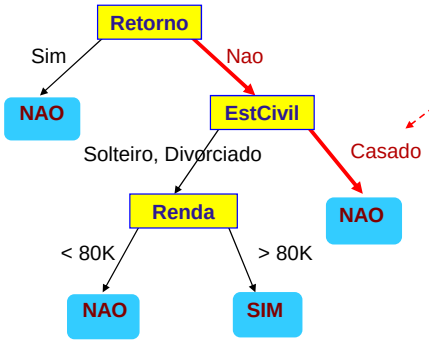
Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?



Aplicando o conjunto de teste ao modelo

Conjunto de teste

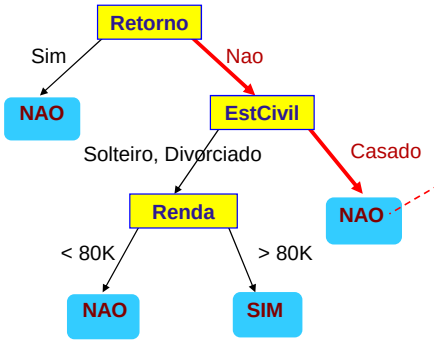
Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?



Aplicando o conjunto de teste ao modelo

Conjunto de teste

Retorno	Estado Civil	Renda Anual	Mentiu
Nao	Casado	80K	?



Associa "**Nao**" para o atributo **Mentiu**

Classificação: Aplicações

- Avaliação de concessão de créditos;
- Detecção de fraudes;
- Diagnósticos médicos;
- Sistema de alerta de geada;
- Previsão de mortalidade de frangos;
- Agrupamento de família de proteínas;
- etc.

Perguntas sobre Classificação

1. Qual é a principal diferença entre **Técnicas** e **Tarefas** de **Mineração de Dados**?

2. Considerando que o **conjunto de teste** é formado, em geral, por 33% das amostras no banco de dados, o que aconteceria se esse conjunto fosse reduzido para 5%?

Regras de Associação

- Estuda o **relacionamento** entre itens de dados que ocorrem com uma certa **frequência**.
- É uma tarefa **descritiva**: identifica padrões em dados históricos.

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)

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**), etc.

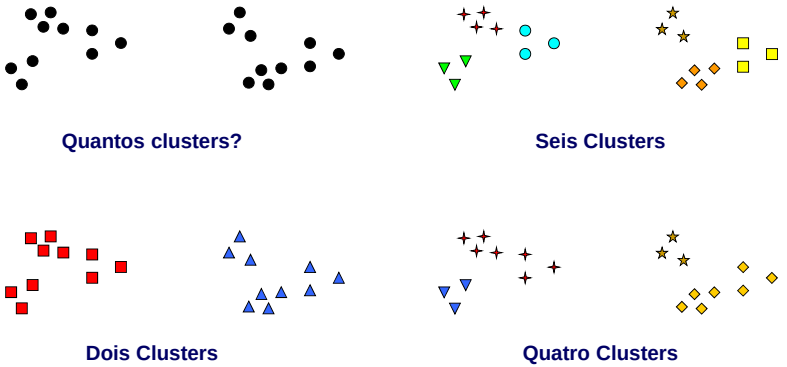
Exercícios sobre Associação

- Considere o conjunto de transações, abaixo, para uma **locadora de filmes**:

TID	Lista de Filmes
T1	A, B, C, D, F, H, I, J, R
T2	A, C, D, P, Q
T3	C, P, Q
T4	B, C, H, I, J
T5	A, D, R, J

1. Existem **padrões de consumo**? Quais?
2. Com base em um padrão de consumo, dê um exemplo de uma **recomendação**, isto é, uma oferta que a locadora poderia sugerir para os seus clientes.

Clusterização ou Agrupamento de Dados



Clusterização ou Agrupamento

- É também considerada uma tarefa **descritiva**.
- **Clusterização**
  - Agrupamento de conjuntos de dados em clusters.
- **Cluster**: uma coleção de objetos
  - **Similares** aos objetos do mesmo cluster.
  - **Dissimilares** aos objetos de outros clusters.
- **Clusterização** é uma classificação **não supervisionada** - sem classes predefinidas.

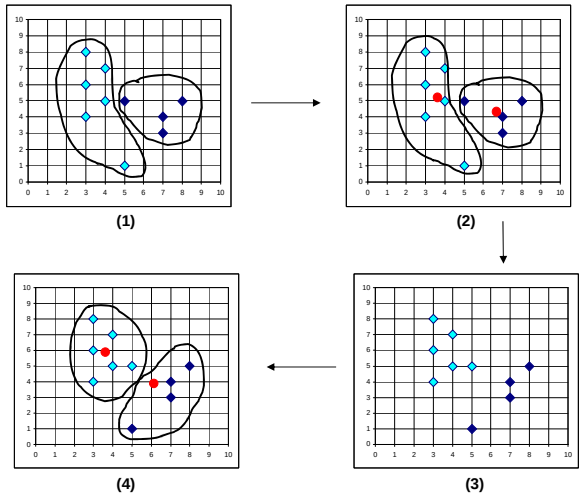
Classificação x Clusterização

- **Aprendizado supervisionado** (**classificação**)
  - **Supervisão**: As observações no conjunto de treinamento são acompanhadas por "**labels**" indicando a classe a que elas pertencem.
  - Novas ocorrências são classificadas com base no conjunto de treinamento.
- **Aprendizado não-supervisionado** (**clusterização**)
  - Não existe **classe** pré-definida para nenhum dos atributos.
  - Um conjunto de observações é dado com o propósito de se estabelecer a existência das classes ou clusters.

Clusterização: Aplicações

- **Marketing**: Segmentação de mercado;
- **Medicina**: Agrupamento de pacientes com sintomas semelhantes;
- **Bioinformática**: Agrupamento de famílias de proteínas.
- **Seguro**: Identifica grupos de clientes que fazem comunicação de sinistro com alta freqüência.
- **Web**: Agrupamento de documentos.

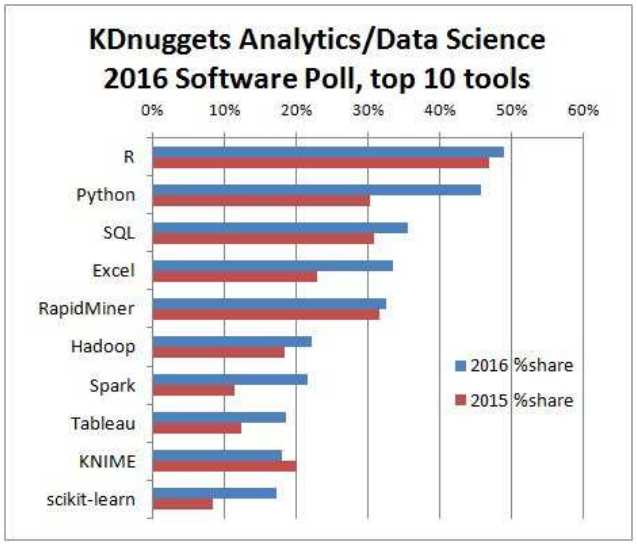
Exemplo do algoritmo K-means



Exercícios sobre Clusterização

1. Como um analista de dados poderia transformar um **problema de séries temporais curtas** em um **problema de clusterização**?
2. Por que **clusterização** é considerada uma tarefa **não-supervisionada**?

Principais softwares para Mineração



Exercícios

1. Qual é a diferença básica entre **classificação** e **clusterização**?
2. Complete a tabela a seguir:

Aplicação	Tarefa de Mineração
Detecção de fraudes	
Segmentação de mercado	
Análise de sequências de DNA	
Recomendação de produtos para clientes	
Layout de prateleiras	
Diagnósticos médicos	
Perfil de compra de clientes	
Avaliação de riscos de empréstimos	