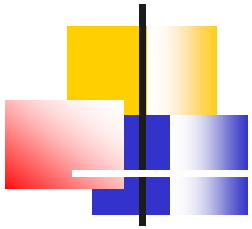
Ciência de Dados

Revisado: Roseli Romero

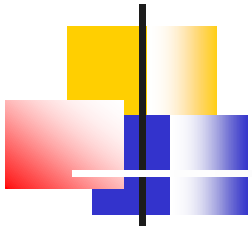
Pré-processamento de dados

Prof. Dr. André C. P. L. F. de Carvalho
Dr. Isvani Frias-Blanco
ICMC-USP



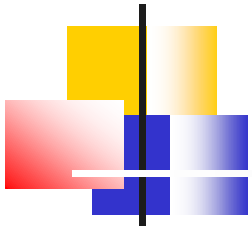
Tópicos

- Introdução
- Qualidade de dados
  - Fontes de problemas
- Limpeza de dados
- Desbalanceamento
- Transformação de dados



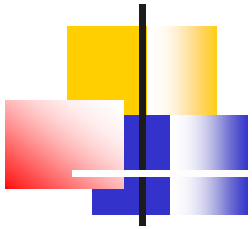
Pré-processamento

- Prepara os dados para seu uso por algoritmos de AM
- Procura melhorar desempenho do algoritmo
  - Custo
    - Tempo
    - Memória
  - Qualidade do modelo gerado
    - Acurácia preditiva



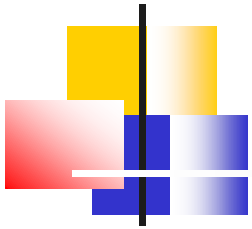
Qualidade de dados

- Em geral, dados não foram gerados para AM
  - Produzidos para outros propósitos
  - Frequentemente apresentam problemas
    - Pelo menos 5% dos exemplos de um conjunto têm problemas
- Algoritmos de AM aprendem melhor com dados "limpos"
  - Entra lixo, sai lixo
  - Problemas nos dados precisam ser detectados e corrigidos
    - Limpeza de dados (data cleansing)



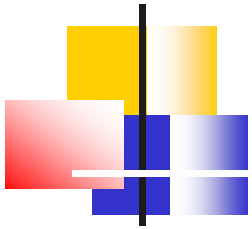
Qualidade de dados

- Problemas nos dados podem ter causa:
  - Sistemática (determinística)
    - Mais fácil de detectar e corrigir
  - "Aleatória"



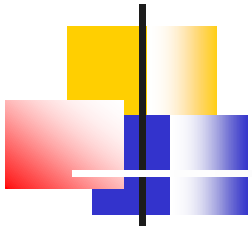
Possíveis causas de problemas

- Falha humana
- Má fé
- Falha no processo ou dispositivo de coleta ou de medição de dados
- Limitações do dispositivo de coleta ou de medição
- Mudanças de conceito



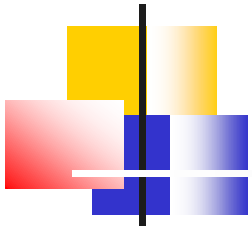
Possíveis consequências

- Valores de atributos ou de exemplos inteiros podem ser perdidos
- Obtenção de exemplos que sejam:
  - Espúrios ou duplicados
    - Ex.: diferentes registros para mesma pessoa que morou em endereços diferentes
  - Inconsistentes
    - Ex.: engenheiro de 3 anos de idade



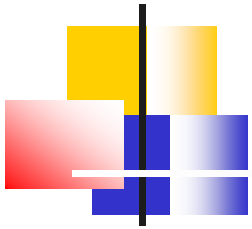
Limpeza

- Correção de problemas detectados nos dados deve lidar com:
  - Atributos com valores ausentes
  - Atributos e objetos redundantes
  - Atributos e objetos com valores inconsistentes
  - Atributos com ruídos
  - *Outliers*



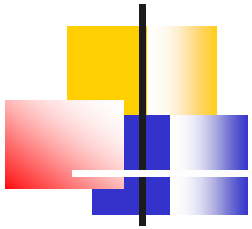
Valores ausentes

- Dados faltosos, faltantes, incompletos
- Várias técnicas de AM não foram projetadas para lidar com valores ausentes
  - Têm dificuldades ou não conseguem induzir um modelo



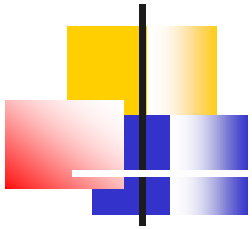
Valores ausentes

- Não é raro um objeto não ter valores para um ou mais atributos
- Possíveis causas:
  - Atributo não foi considerado quando os primeiros dados foram coletados
  - Desconhecimento do valor do atributo por ocasião do preenchimento
  - Distração, mal entendido ou declinação na hora do preenchimento
  - Problema com dispositivo / processo de coleta de valores para o atributo



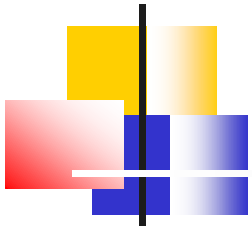
Exemplo de valores ausentes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
	não	não	baixo	não	1100	saudável
Maria	sim	sim		não	600	saudável
José	sim	não	baixo	sim		doente
Sérgio	não	não	baixo	não	1100	saudável
Ana	sim	não	alto	sim	1800	saudável
Leila		não	alto		900	doente
Marta	sim	não	baixo	sim	2000	doente



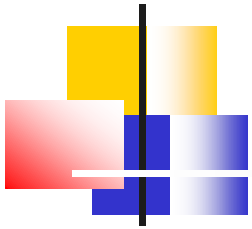
Lidar com valores ausentes

- Agir como se não houvessem valores ausentes
  - Utilizar apenas os valores que estão presentes
    - Ex.: Menos atributos no cálculo da distância entre objetos
  - Modificar algoritmo de AM para lidar com valores ausentes
- Descartar objetos com atributos sem valores
- Preencher valores ausentes



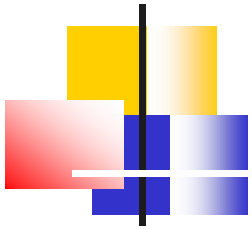
Descarte de objetos

- Geralmente empregado quando:
  - Um dos atributos ausentes é o atributo classe
  - Objeto tem muitos valores ausentes
- Não é indicado quando:
  - Ocorre com poucos atributos do objeto
  - Há risco de perder dados importantes



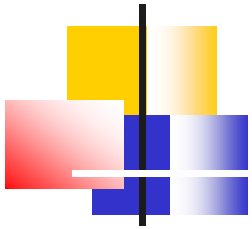
Preenchimento de valor

- Criação de um novo valor que significa ausência
  - Para valores nominais (sem ordem)
- Criação de um novo atributo preditivo
  - Marcando objetos em que um dado atributo tinha valor ausente
- Estimativa de um valor para suprir a ausência



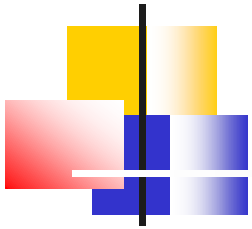
Estimativa do valor

- Usar medida de localidade
  - Média (mediana, moda) dos valores do atributo
    - Todos os valores
    - Dos objetos mais próximos e/ou da mesma classe
  - Para série temporais, medida de localidade entre valores anterior e posterior



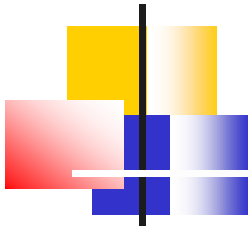
Estimativa do valor

- Induzir valor induzido por algum estimador
  - Valor presente em objetos semelhantes
  - Utilizar algoritmo de AM
  - Alternativa mais eficiente



Valores ausentes

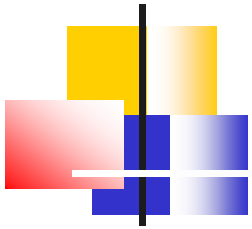
- Observações
  - Em alguns casos, a ausência de valor é uma informação importante sobre o objeto
  - Existem situações em que o valor pode ou precisa estar ausente
    - Ex.: Atributo número do apartamento para quem mora em uma casa
    - Ao invés de ausente, é um valor inexistente
    - Difícil tratar de forma automática
      - Criação de um novo atributo



Exercício

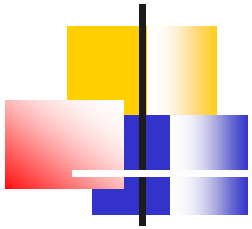
- Tratar dos valores ausentes da tabela abaixo

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	adimplente
Lia		Superior	200	174	7000	inadimplente
Maria	Advogado	Médio		180	600	adimplente
José	Médico	Superior	100		2000	inadimplente
Sérgio	Bancário		82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	36	2000	inadimplente
José	Médico	Médio	340		800	



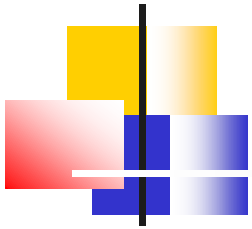
Valores inconsistentes

- Dados podem conter valores inconsistentes
  - Atributos preditivos
    - Ex. Código postal invalido para uma cidade
      - Erro / engano
      - Proposital (fraude)
  - Atributo alvo
    - Podem levar a objetos conflitantes (ambiguidade)
      - Ex.: valores iguais para atributos preditivos e diferentes para atributo alvo
    - Podem ser causados por erro na rotulação do objeto



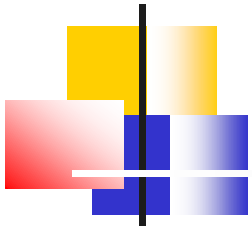
Valores inconsistentes

- Algumas inconsistências são de fácil detecção
  - Violação de relações conhecidas entre atributos
    - Ex.: Valor de atributo A é sempre menor que valor de atributo B
  - Valor inválido para o atributo
    - Ex.: altura com valor negativo
  - Em outros casos, informações adicionais precisam ser consideradas
- Podem indicar presença de ruído



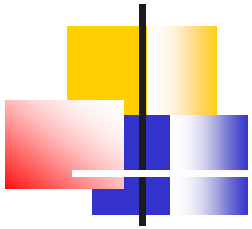
Exemplo de objetos inconsistentes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
Pedro	não	não	baixo	não	1100	saudável
Maria	sim	sim	alto	não	600	saudável
José	sim	não	baixo	sim	2000	doente
Sérgio	não	não	baixo	não	1100	doente
Ana	sim	não	alto	sim	1800	saudável
Leila	não	não	alto	sim	900	doente
Marta	sim	não	alto	sim	3000	doente



Exemplo de atributos inconsistentes

Nome	Idade	Enjoo	Batimentos	Estudo	Diagnóstico
João	30	sim	baixo	2	doente
Pedro	42	não	baixo	4	saudável
Maria	27	sim	alto	3	saudável
José	4	não	baixo	10	doente
Sérgio	38	não	baixo	3	doente
Ana	63	não	alto	2	saudável
Leila	22	não	alto	21	doente
Marta	53	não	alto	30	doente



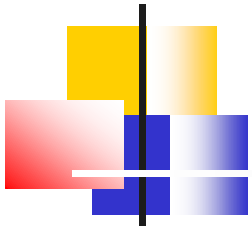
Objetos redundantes

- Objetos ou atributos preditivos (quase) duplicados
  - Não trazem informação nova
  - Ex.: Pessoas em diferentes BDs com mesmo nome, mas endereço com pequenas diferenças
    - Diferença real ou erro no preenchimento
- Deduplicação
  - Detectar e eliminar (ou combinar) duplicações
  - Cuidado para não eliminar ou combinar objetos ou atributos que representam dados diferentes

Exemplo

- objetos redundantes

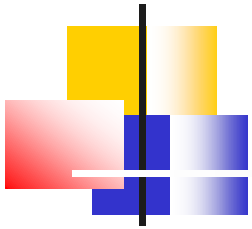
Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
Segio	não	não	baixo	não	1100	saudável
Maria	sim	sim	alto	não	600	saudável
Marta	sim	não	baixo	sim	2000	doente
Sérgio	não	não	baixo	não	1100	saudável
Ana	sim	não	alto	sim	1800	saudável
Leila	não	não	alto	sim	900	doente
Marta	sim	não	baixo	sim	2000	doente



Exercício

- Definir problemas existentes na tabela abaixo:

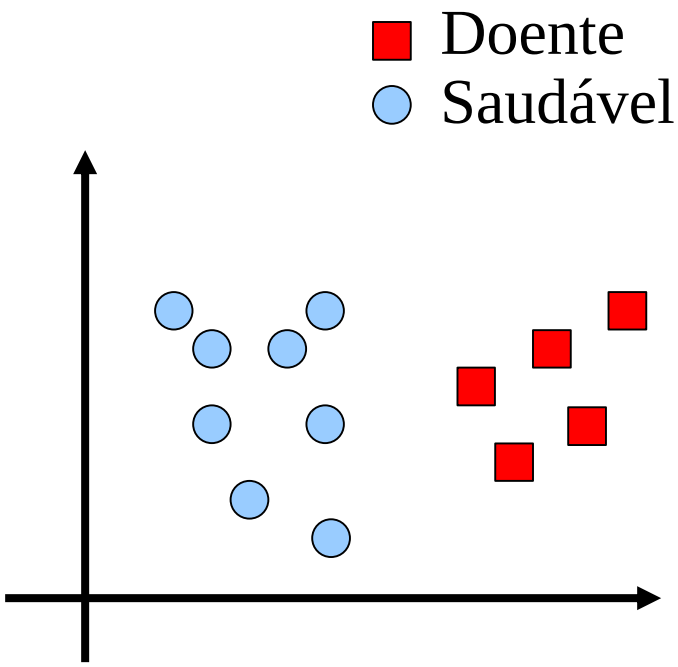
Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador		70	180	3000	adimplente
Lia	Médico	Superior	200	174	7000	inadimplente
Maria	Advogado	Médio	90	180	600	adimplente
José	Médico	Superior	200	174	7000	inadimplente
Sérgio	Bancário	Superior	82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	-6	2000	inadimplente



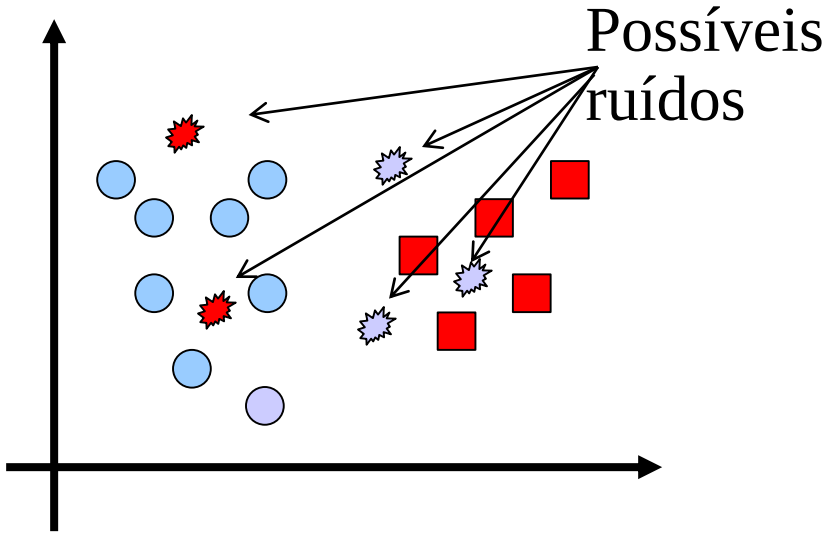
Ruídos

- Podem levar a um superajuste do modelo induzido por um algoritmo de AM
- Difícil ter certeza que um valor é ruído
  - Tem-se apenas um indício
    - A menos que valor seja inconsistente
  - Se identificados, podem ser tratados como valores ausentes
- Nos atributos preditivos ou no atributo alvo
  - Consequências diferentes

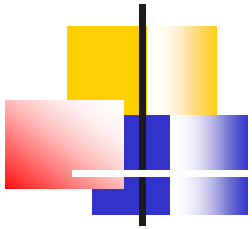
Exemplo



Dados sem ruído



Dados com possíveis ruídos



Outliers

- Objetos ou valores anômalos
  - Objetos que têm características diferentes da grande maioria dos demais objetos
    - Valor(es) de um ou mais atributos que destoa(m) dos valores típicos
- *Outliers* podem sugerir a presença de ruído ou ser valores legítimos
  - Em várias aplicações, objetivo é encontrar *outliers*

Outliers