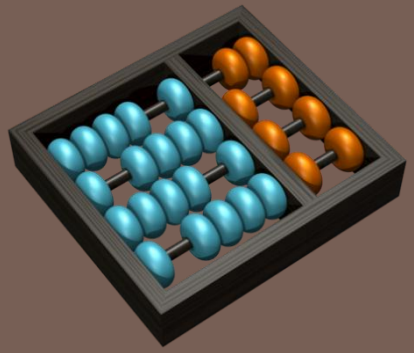


UNICAMP

TRATAMENTO DE DADOS

Prof. Julio Cesar dos Reis

jreis@ic.unicamp.br

www.ic.unicamp.br/~jreis

CT-0611

Vídeo

Objetivos da aula

- Aprender tipos de dados e dos conjunto de dados
- Estudar o problema da qualidade dos dados
- Introduzir aspectos relacionados ao **pré-processamento e transformação de dados**

Tipos de dados e tipos de conjunto de dados

Atributos, instâncias e classes

Instâncias

Atributos

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

Valores de atributos

- Valores de atributos são números ou símbolos atribuídos a um atributo

- Distinção entre atributos e valores de atributos
  - Algum atributo pode ser mapeado a diferentes valores de atributos
    - Exemplo: altura pode ser medida em pés e metros

Valores de atributos

- Valores de atributos são números ou símbolos atribuídos a um atributo
- Distinção entre atributos e valores de atributos
  - Algum atributo pode ser mapeado a diferentes valores de atributos
    - Exemplo: altura pode ser medida em pés e metros
  - Diferentes atributos podem ser mapeados para o mesmo conjunto de valores
    - Exemplo: Valores de atributos para ID e idade são inteiros
    - Propriedades de atributos podem ser diferentes
      - ID não tem limite mas idade tem um valor máximo e um valor mínimo

Tipos de atributos

- A maior parte dos algoritmos diferenciam
  - Atributo nominal e ordinal

- Atributos nominais são também chamados "*categorical*", "*enumerated*" ou "*discrete*" [eg: cor dos olhos]
  - Porém "*enumerated*" e "*discrete*" implicam uma ordem
  - Caso especial: dicotomia ("boolean")

Tipos de atributos

- A maior parte dos algoritmos diferenciam
  - Atributo nominal e ordinal

- Atributos nominais são também chamados "*categorical*", "*enumerated*" ou "*discrete*" [eg: cor dos olhos]
  - "*enumerated*" e "*discrete*" implicam uma ordem
  - Caso especial: dicotomia ("boolean")

- Atributos ordinais são chamados de "*numeric*", ou "*continuous*"
  - "*continuous*" implica continuidade matemática
  - Eg: altura

Atributos nominais

- Os valores são símbolos diferentes
- Exemplo: atributo "*outlook*" da base condições tempo
  - Valores: "sunny", "overcast", e "rainy"
- Não existe relação entre os valores nominais (sem ordem ou medida de distância)
- Somente testes de igualdade podem ser realizados

Atributos ordinais

- Impõem uma ordem nos valores
  - Exemplo
    - Atributo “temperature” nos dados de condição do tempo
    - Valores: “hot” > “mild” > “cool”
- Adição e subtração não tem sentido
  - Exemplo de regra:

    temperature < hot => play = yes
- A diferença entre atributos nominais e ordinais nem sempre é clara

Atributos intervalares

- Os intervalos são ordenados e medidos em unidades fixas e iguais (numéricos)
  - Exemplo 1: atributo "*temperature*" expresso em graus Fahrenheit
  - Exemplo 2: atributo "*year*"
- A diferença entre 2 valores faz sentido
- A soma ou produto podem não fazer sentido

Propriedades de valores de atributos

- O tipo de um atributo depende das propriedades que possui:
  - **Distinção:** $= \neq$
  - **Ordem:** $< >$
  - **Adição/Sub:** $+ -$
  - **Multiplicação:** $* /$

  - Atributo nominal: dinstinção
  - Atributo ordinal: distinção & ordem
  - Atributo Intervalo: distinção, ordem & adição
  - Atributo racional: todas as 4 propriedades

Tipo de Atributo	Descrição	Exemplos	Operações
Nominal	Os valores de atributos nominais são apenas nomes diferentes, i.e., atributos nominais têm informação suficiente para **distinguir** um objeto de outro.	zip codes, número de ID, cor dos olhos, sexo: {*homem, mulher*}	moda, entropia, contingência correlação, teste χ^2
Ordinal	Os valores de um atributo ordinal têm informação suficiente para **ordenar** objetos. (<, >)	dureza de minerais, {*good, better, best*}, notas números de ruas	mediana, porcentagens, correlação de rank, testes de corrida, testes do sinal
Intervalo	Para atributos intervalares, as diferenças entre valores são significativas, *i.e.*, uma unidade de medida existe. (+, -)	datas, temperatura em Celsius ou Fahrenheit	média, desvio padrão, correlação de Pearson, teste *t* e teste *F*
Racional	Para variáveis racionais, tanto diferenças como proporções são importantes (*, /)	Quantidades monetárias, contagens, idade, massa, comprimento, corrente elétrica	média geométrica, média harmônica, variação percentual

Tipos de conjuntos de dados

- **Registro**
  - Matriz de dado
  - Documentos
  - Transações
- **Grafos**
  - WWW
  - Estruturas moleculares
- **Ordenados**
  - Dados espaciais
  - Dados temporais
  - Dados sequenciais
  - Dados de sequenciamento genético

Tabelas

- Dados como coleção de registros definidos em termos de um conjunto fixo de atributos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Matrizes

- Se os objetos têm o mesmo conjunto de **atributos numéricos**, então os objetos podem ser pensados como pontos em um espaço multidimensional
  - Cada dimensão representa um atributo distinto

- Conjunto de dados representado por uma matriz m x n
  - m colunas, uma para cada atributo, e n linhas, uma para cada objeto

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Documentos

- Cada documento é representado como um vetor de ‘termos’
  - Cada termo é um componente (atributo) do vetor
  - O valor de cada componente é o número de vezes que o termo ocorre no documento

Exemplo de dados em documentos

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

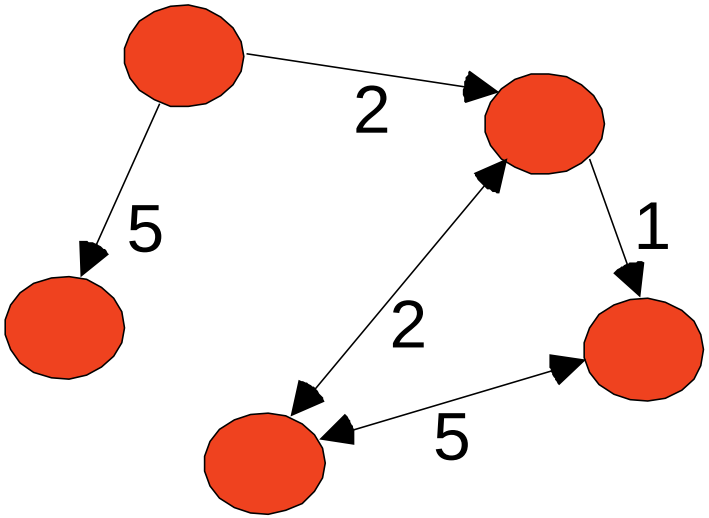
Transações

- Tipo especial de registro
  - Cada registro (transação) inclui um conjunto de itens
  - Por exemplo, considere um supermercado
    - O conjunto de produtos comprados por um cliente em um determinado dia constitui uma **transação**, enquanto os produtos comprados são **itens**

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Grafos

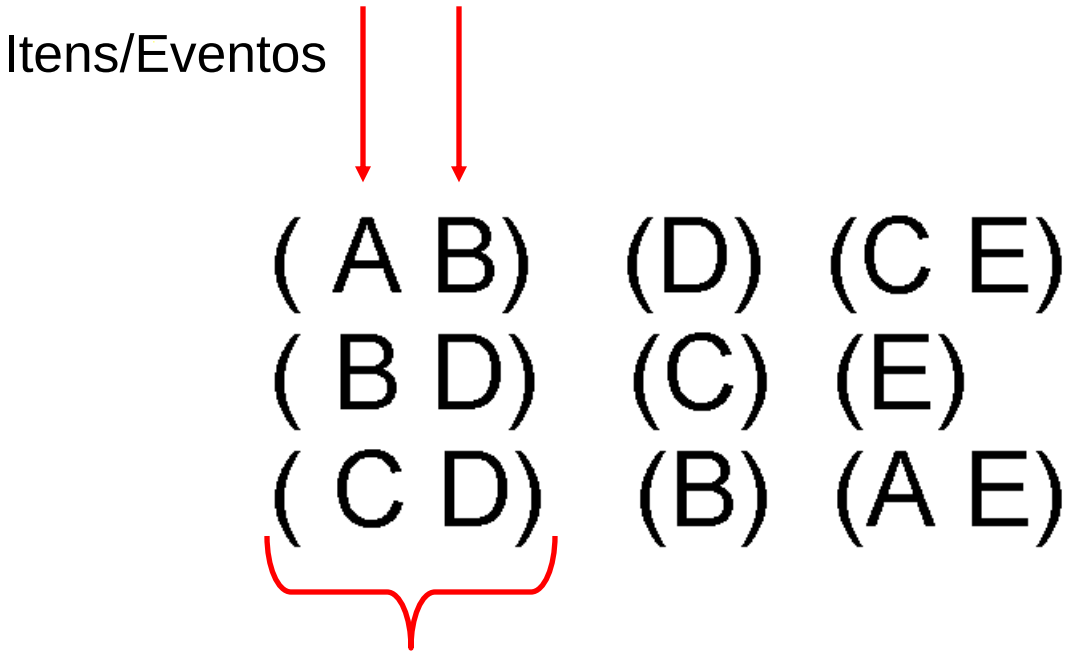
- Exemplos: Grafo genérico e ligações HTML




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Dados ordenados

- Sequência de transações



Exemplo de dados ordenados

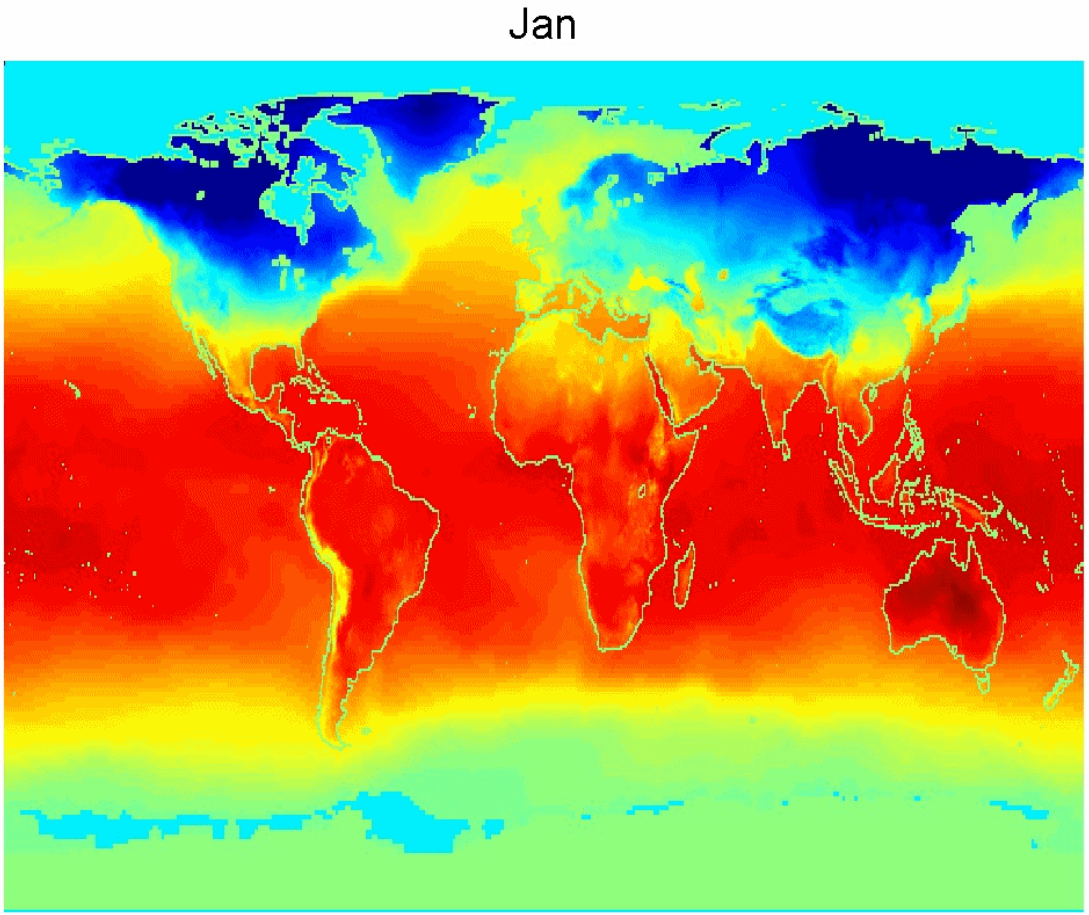
- Dados sequenciais genômicos

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Exemplo de dados ordenados

- Dados espaço-temporais

Temperatura média mensal dos continentes e oceanos



Cuidados no tratamento de dados

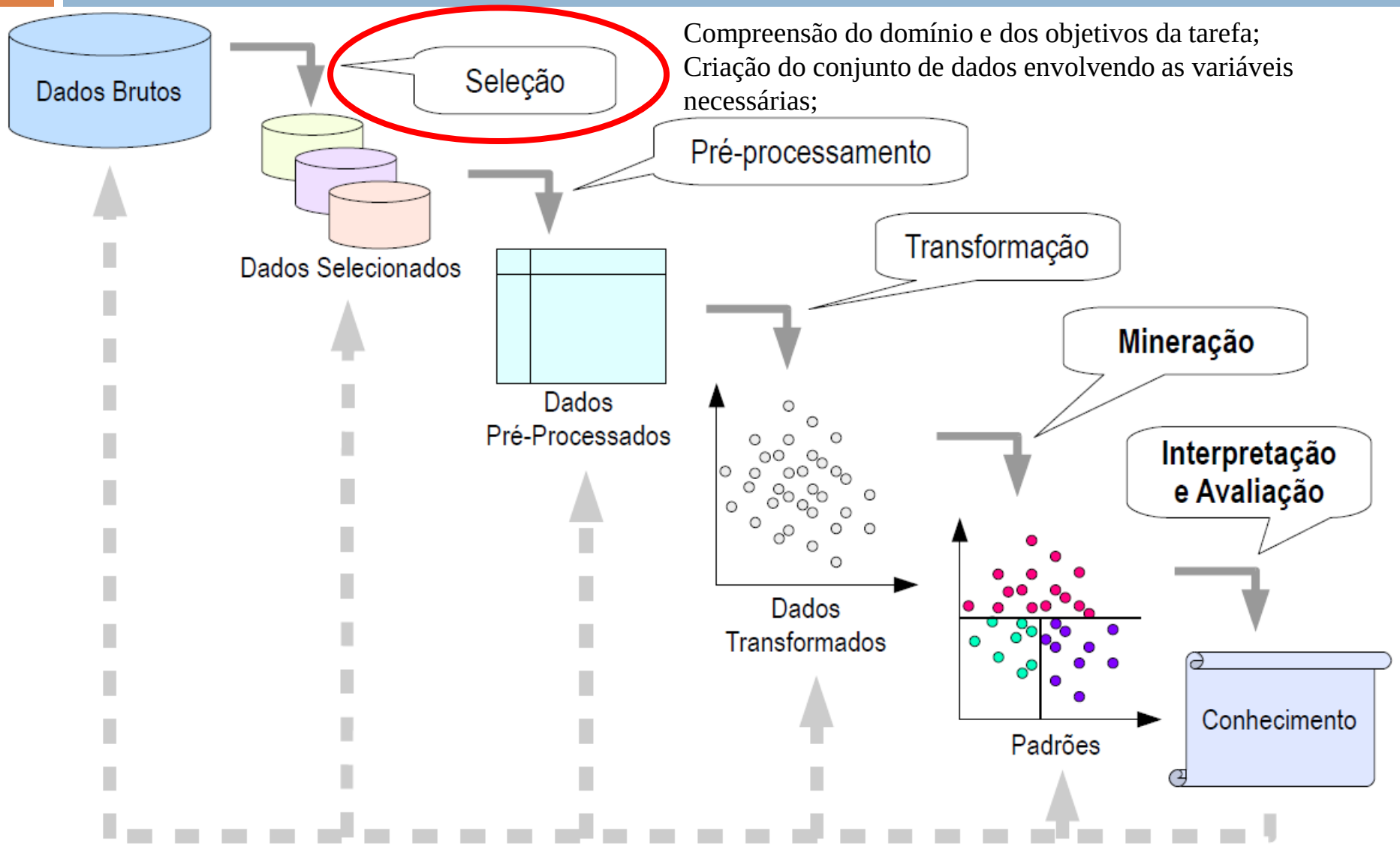
- Atributos com representação inadequada para tarefa e algoritmo
- Atributos cujos valores não tenham informações adequadas
- Excesso de atributos (podem ser redundantes ou desnecessários)

Cuidados no tratamento de dados

- Atributos com representação inadequada para tarefa e algoritmo
- Atributos cujos valores não tenham informações adequadas
- Excesso de atributos (podem ser redundantes ou desnecessários)
- Atributos insuficientes
- Excesso de instâncias (afetam tempo de processamento)
- Instâncias insuficientes
- Instâncias incompletas (sem valores para alguns atributos)

Seleção de dados

Processo de descoberta de conhecimento



Compreensão do domínio e dos objetivos da tarefa;
Criação do conjunto de dados envolvendo as variáveis necessárias;

Amostragem

- Técnica principal empregada para seleção de dados
  - Usada frequentemente tanto para estudos preliminares quanto para análise final de dados
- Estatísticos realizam amostragens porque obter o conjunto inteiro de dados de interesse é caro ou demanda muito tempo
- Amostragem é usada em mineração de dados porque processar o conjunto inteiro de dados de interesse é muito caro ou demanda muito tempo

Amostragem

- Princípio chave para amostragens bem sucedidas
  - Usar uma amostra é tão bom quanto usar o conjunto inteiro de dados
  - Amostra precisa ser representativa
  - Uma amostra é representativa se tem aproximadamente a mesma propriedade (de interesse) do conjunto original de dados
- Há 4 tipos

Tipos de amostragem

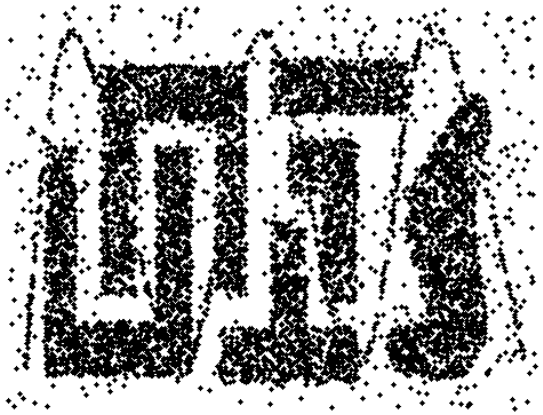
- **Amostragem aleatória**
  - Há uma probabilidade igual de selecionar um item particular
- **Amostragem sem reposição**
  - Na medida em que cada item é selecionado, é removido da população

Tipos de amostragem

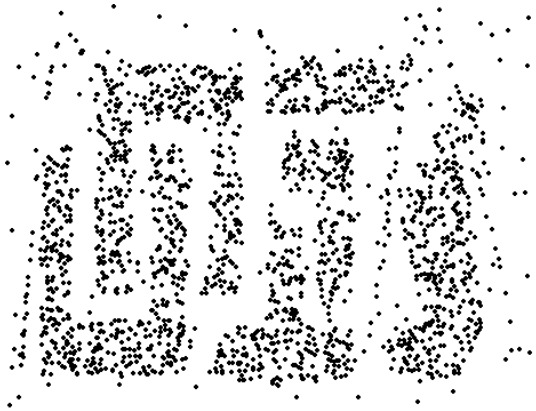
- **Amostragem com reposição**
  - Objetos não são removidos da população quando são selecionados para a amostra
    - Na amostragem com reposição, o mesmo objeto pode ser selecionado mais de uma vez

- **Amostragem estratificada**
  - Particiona o dado em várias partições
  - Seleciona-se amostras aleatórias de cada partição

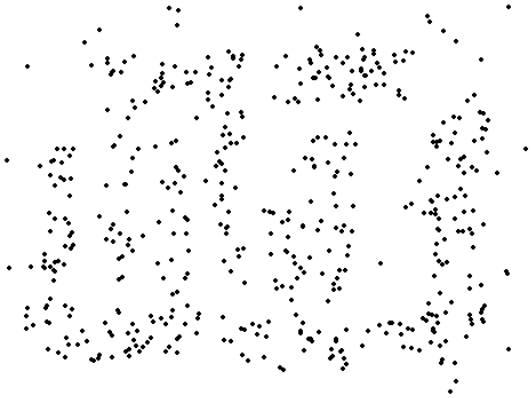
Tamanho da amostra



8000 Pontos



2000 Pontos



500 Pontos

Pré-processamento de dados

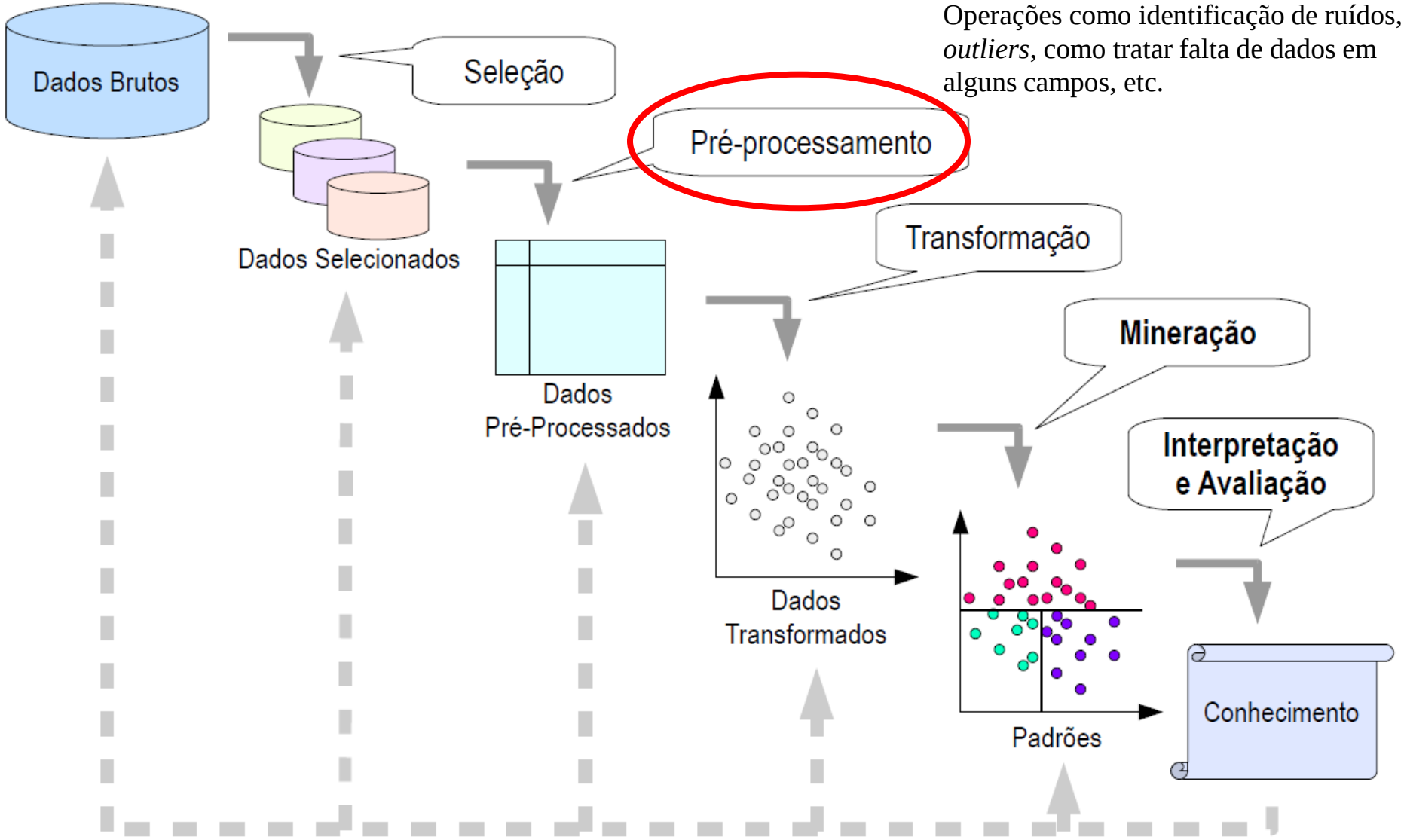
Qualidade dos dados

- Quais os tipos de problemas de qualidades de dados?
- Como podemos detectar problemas com os dados?
- Como podemos proceder para tratar desses problemas?

Problemas de qualidade dos dados

- Ruído
- *Outliers*
- Valores faltantes
- Dados duplicados

Processo de descoberta de conhecimento



Operações como identificação de ruídos, *outliers*, como tratar falta de dados em alguns campos, etc.

Pré-processamento de dados

- Rotinas de limpeza de dados visam
  - Suprir valores ausentes
  - Reduzir discrepâncias de valores ruidosos
  - Corrigir inconsistências

Valores faltantes

- Motivos para valores faltantes
  - Informação não é coletada
    - Exemplo: Pessoas não informam idade ou peso
  - Atributos podem não ser aplicados a todos os casos
    - Exemplo: rendimento anual não é aplicável às crianças

Tratamento de valores faltantes

- Eliminar objetos
- Estimar valores faltantes
- Ignorar valores faltantes durante análise
- Trocar por média de todos os valores possíveis (ponderados por suas probabilidades)

Técnicas sobre valores ausentes

1. Ignorar a tupla
2. Suprir valores ausentes

   a) Manualmente;

   b) Através de uma constante global;

   c) Utilizando a média do atributo;

   d) Utilizando a média do atributo para todas as instâncias da mesma classe;

   e) Com o valor mais provável (regressão, inferência etc.)

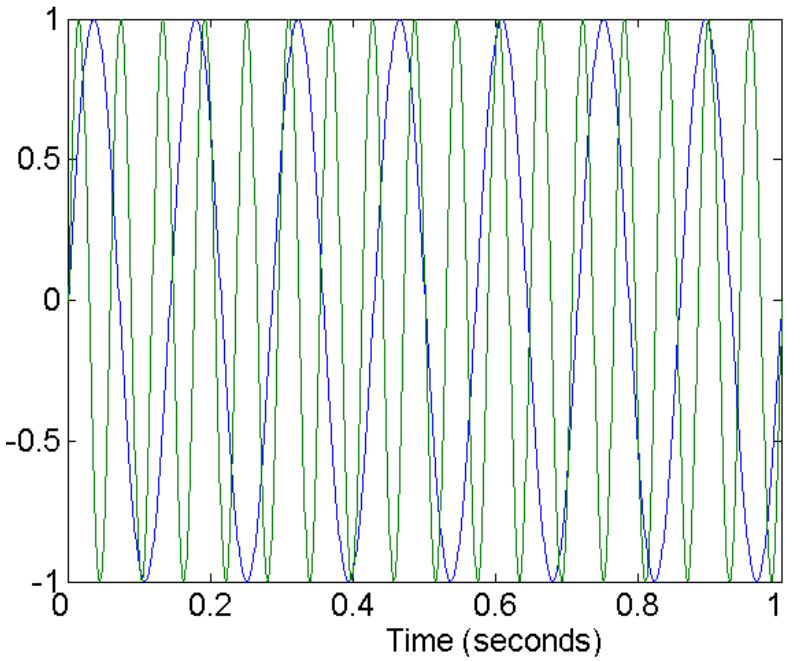
Técnicas sobre valores ausentes

- As técnicas 2b, 2c, 2d e 2e podem "viciar" os dados

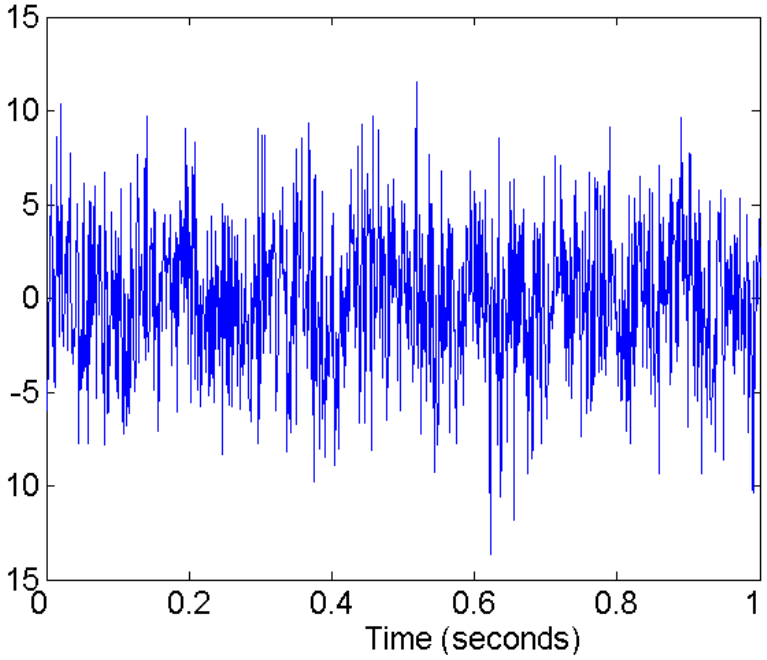
- A técnica 2e é uma estratégia interessante, pois em comparação com outros métodos utiliza um maior número de informações dos dados disponíveis

Ruído nos dados

- Ruído refere-se as modificações dos valores originais
  - Exemplos: distorções da voz de uma pessoa quando falando em um telefone ruim



Duas ondas senoidais



Duas ondas senoidais + Ruído

Ruídos nos dados

- São erros aleatórios ou variâncias numa variável mensurada

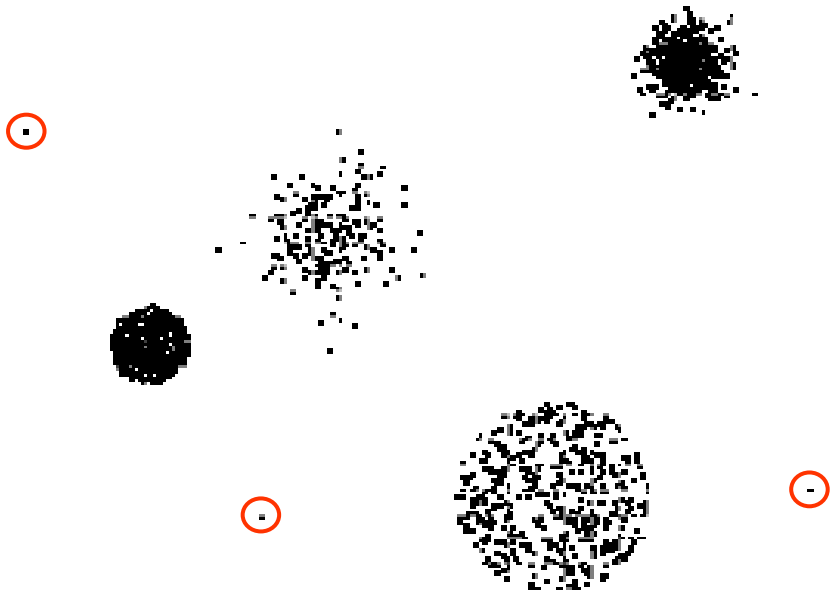
- Eliminação de ruídos pode ser realizada através de:
  1 – Interpolação
  2 – Agrupamento
  3 – Inspeção humana e computacional combinadas
  4 – Regressão

Inconsistências

- Corrigidas manualmente através de referências externas
- Rotinas de consistência evitam a inserção de dados incorretos
- Discrepâncias podem ser combatidas através de dependências funcionais

Outliers

- São objetos com características que são consideravelmente diferentes da maioria dos outros objetos do conjunto de dados

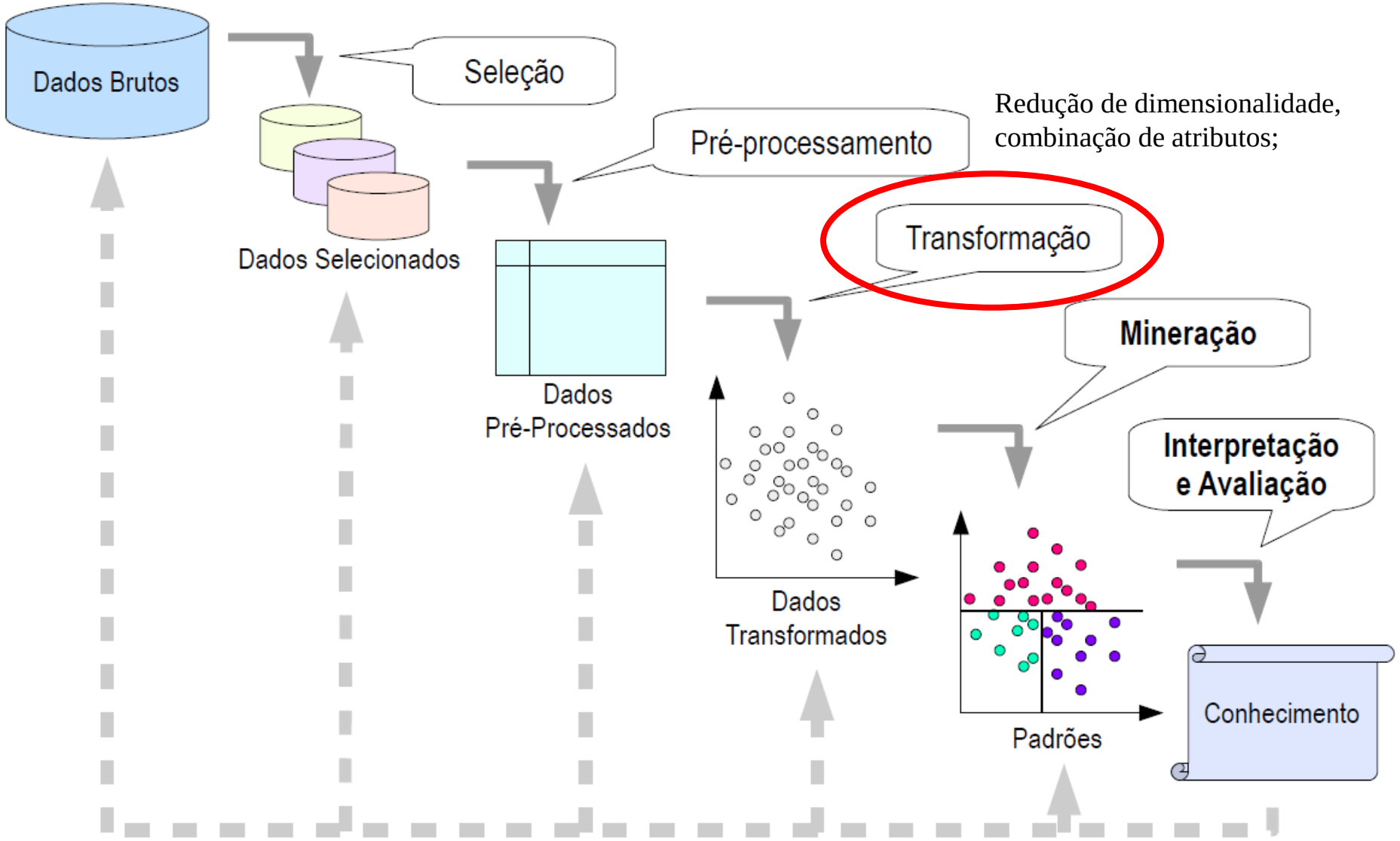


Dados duplicados

- Conjuntos de dados podem conter objetos que são duplicatas, ou quase duplicatas de algum outro
  - Questão relevante quando se integra dados de fontes heterogêneas
- Exemplos:
  - Mesma pessoa com múltiplos endereços de e-mail
- Necessário um procedimento para tratar questões de dados duplicados
  - Ex. removê-los

Transformação de dados

Processo de descoberta de conhecimento



Transformação de dados

- **Seleção de atributos**
  - Redução de dimensionalidade (grande quantidade de atributos)
    - Seleção de características
    - Agregação
- **Criação de atributos**
  - Discretização e "binarização"
  - Transformação de atributos

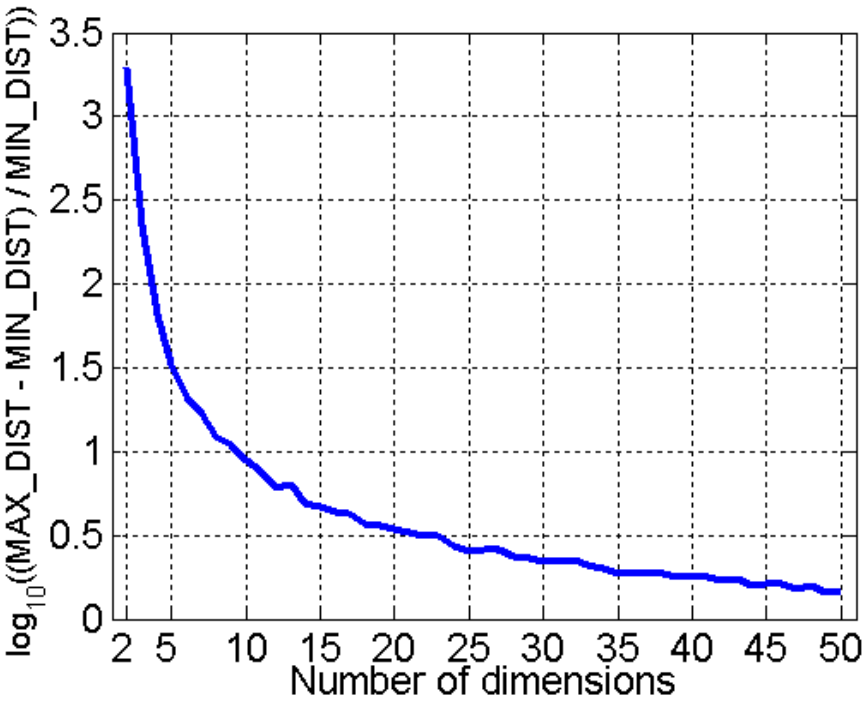
Transformação de dados

- **Seleção de atributos**
  - Redução de dimensionalidade (grande quantidade de atributos)
    - Seleção de características
    - Agregação
- Criação de atributos
  - Discretização e "binarização"
  - Transformação de atributos

Maldição da dimensionalidade

- Quando a dimensionalidade cresce, os dados tornam-se esparsos no espaço que ocupam

- Definições de densidade e distância entre pontos, que são críticos para técnicas de agrupamento e detecção *outliers,* tornam-se menos significativos



- **Gere 500 pontos aleatoriamente**
- **Compute a diferença entre a distância máxima e a mínima entre pares de pontos**

Redução de dimensionalidade

- Objetivo
  - Evitar a maldição de dimensionalidade
  - Reduzir a quantidade de tempo e memória necessários considerando algoritmos de mineração
  - Permitir que dados sejam mais facilmente visualizados
  - Poder ajudar a eliminar características irrelevantes ou reduzir ruído

Seleção de características

- Características redundantes
  - Duplicar muito ou toda informação contida em um ou mais atributos
  - Exemplo: preço de compra de um produto e quantidade de impostos pagos na venda
- Características irrelevantes
  - Contém nenhuma informação que seja útil para tarefa de mineração
  - Exemplo: ID de estudante não é relevante para cálculo de média de notas

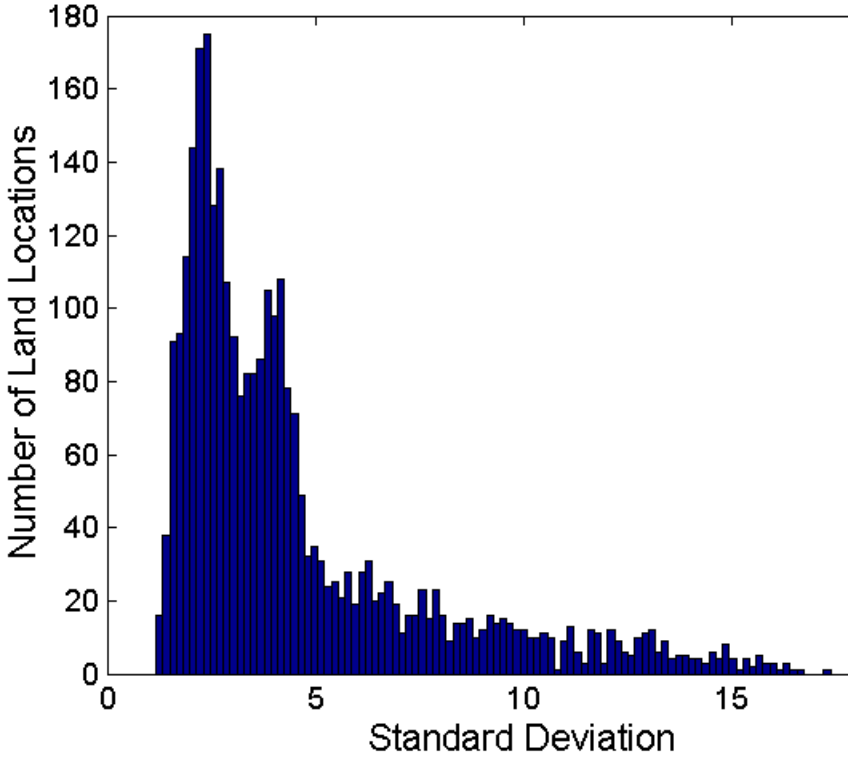
Agregação

- Combinar dois ou mais atributos (ou objetos) em um atributo único (ou objeto)

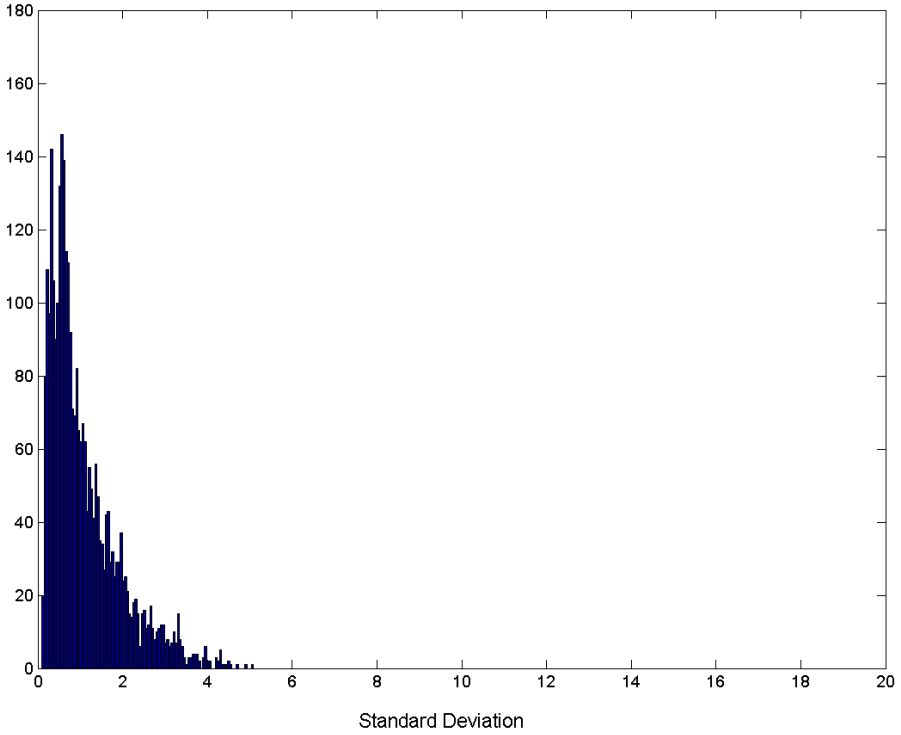
- Objetivo
  - Redução de dados
    - Reduzir o número de atributos ou objetos
  - Mudança de escala
    - Cidades agregadas em regiões, estados, países, *etc.*
  - Dados mais "estáveis"
    - Dados agregados tendem a ter menos variabilidade

Agregação

Variação de precipitação na Austrália



Desvio padrão da precipitação média mensal



Desvio padrão da precipitação média anual

Transformação de dados

- Seleção de atributos
  - Redução de dimensionalidade (grande quantidade de atributos)
    - Seleção de características
    - Agregação

- **Criação de atributos**
  - Discretização e "binarização"
  - Transformação de atributos

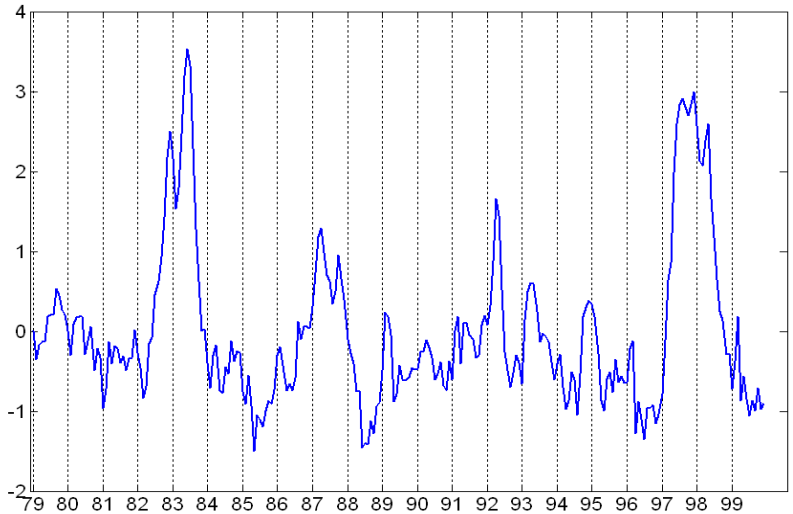
Criação de características

- Novos atributos podem capturar informação importante no conjunto de dados muito mais eficientemente do que nos atributos originais

- Três metodologias gerais
  - Extração de característica
    - Específico de domínio
  - Transformação de característica
    - Mapear dados para novo espaço
  - Construção de característica
    - Combinação de característica

Mapeamento de dados para novo espaço

- Uma função que mapeia o conjunto de valores de um atributo em um novo conjunto de tal forma que o valor antigo possa ser identificado com um dos novos valores
  - Funções simples: x^k, log(x), e^x, |x|
  - Padronização e normalização



Discretização de variáveis contínuas

- Transforma atributos contínuos em atributos categóricos
  - Categóricos: Somente conjunto de valores finito e contável
  - Contínuos: Tem números reais como valores de atributos
- Absolutamente essencial se a tarefa de mineração apenas manuseia atributos categóricos
- Em alguns casos, mesmo métodos que manuseiam atributos contínuos têm melhor desempenho com atributos categóricos

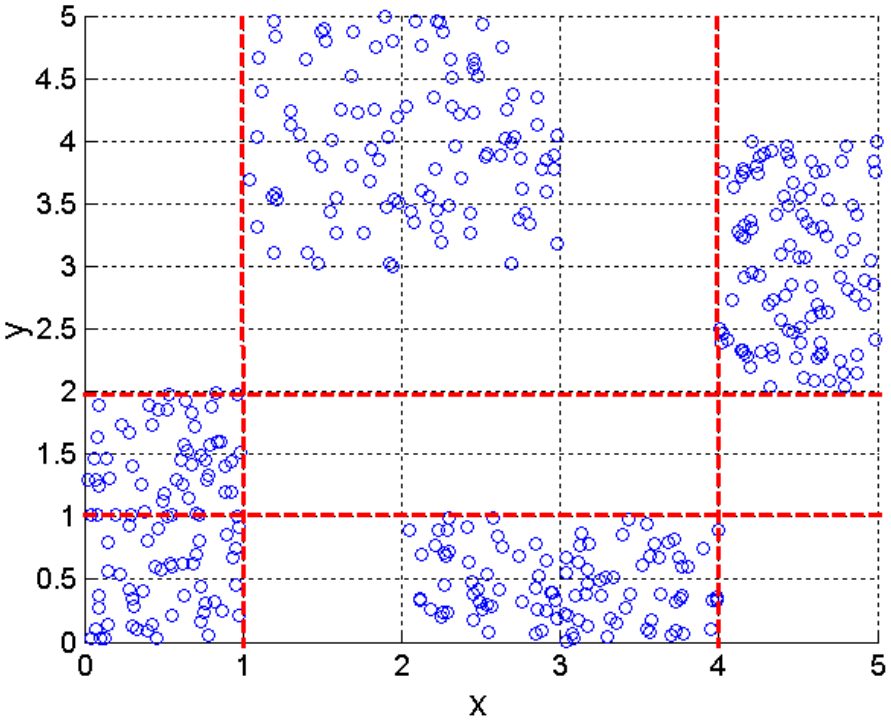
Discretização de variáveis contínuas

- Métodos de discretização
  - **Discretização supervisionada**
    - Leva em consideração o atributo classe
  - **Discretização não-supervisionada**
    - Atributo contínuo é discretizado ignorando-se o atributo classe
    - Não leva em consideração o atributo classe

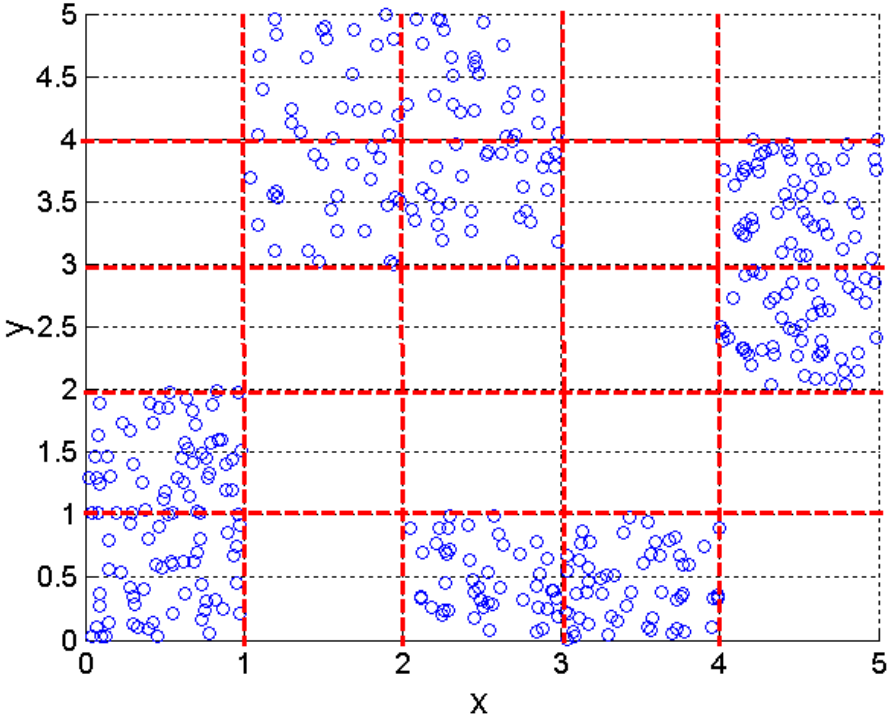
Discretização supervisionada

- Discretização pelo Método 1R (1-rule)
  - Sub-produto de uma técnica de extração automática de regras
  - Utiliza as classes de saída para discretizar cada atributo de entrada separadamente
  - Ex:
    - Base de dados hipotética: Meteorologia
    - Decisão: Realizar ou não um certo jogo

Discretização usando rótulos de classe



3 categorias tanto para x quanto para y



5 categorias tanto para x quanto para y

Discretização não-supervisionada

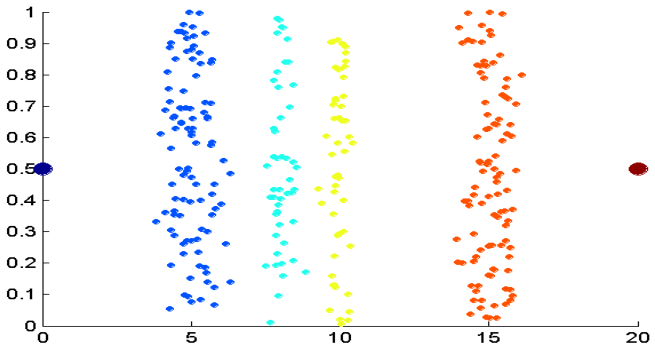
- O método 1R é supervisionado. Considera a variável de saída (classe) na discretização

- Métodos **Não-Supervisionados** consideram somente o atributo a ser discretizado
  - São a única opção no caso de problemas de agrupamento (*clustering*), em que não se conhecem as classes de saída

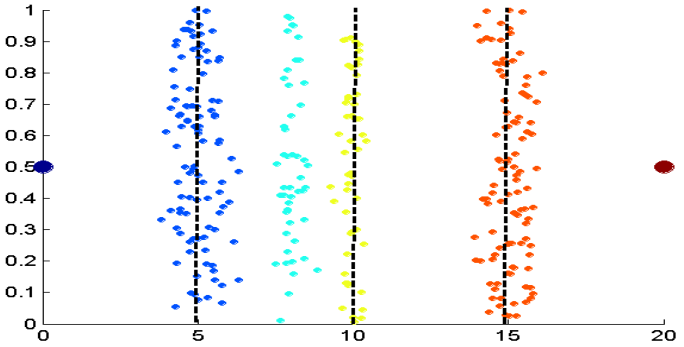
Discretização não-supervisionada

- Três abordagens básicas:
  - Número pré-determinado de intervalos
    - uniformes (*equal-interval binning*)
  - Número uniforme de amostras por intervalo
    - (*equal-frequency binning*)
  - Agrupamento (*clustering*): intervalos arbitrários

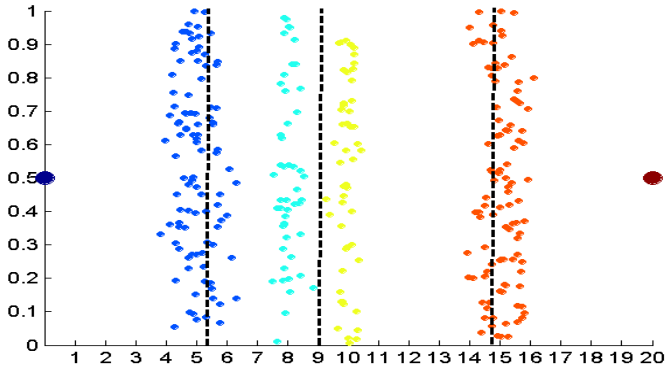
Discretização sem usar rótulos de classes



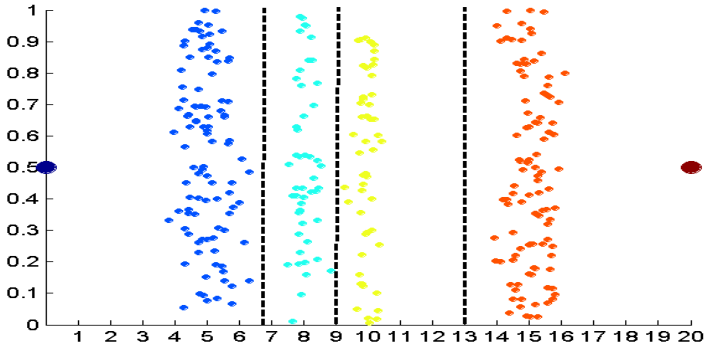
Dados



Mesmo tamanho de intervado



Frequência igual



K-médias (agrupamento)

Método com mesmo tamanho de intervalo

- Número pré-determinado de intervalos uniformes
  - (*equal-interval binning*)

- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85

- Bins com largura 6:

  $x \leq 60$
  $60 < x \leq 66$
  $66 < x \leq 72$
  $72 < x \leq 78$
  $78 < x \leq 84$
  $84 < x \leq 90$

Método com mesmo tamanho de intervalo

- Número pré-determinado de intervalos uniformes
  - (equal-interval binning)

- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85

- Bins com largura 6:

  $x \leq 60$: n.a.

  $60 < x \leq 66$: 64, 65

  $66 < x \leq 72$: 68, 69, 70, 71, 72, 72

  $72 < x \leq 78$: 75, 75

  $78 < x \leq 84$: 80, 81, 83

  $84 < x \leq 90$: 85

Problemas com mesmo tamanho intervalo

- Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido às divisões muito grandes ou fronteiras inadequadas

- Distribuição das amostras muito irregular
  - alguns bins com muitas amostras
  - outros com poucas amostras

Método de intervalo por frequência

- Número uniforme de amostras por intervalo
  - (*equal-frequency binning*)
- Chamado de equalização do histograma
- Cada *bin* tem o mesmo número aproximado de amostras
- Histograma é plano
- Heurística para o número de bins: $\sqrt{N}$
  - N = número de amostras

Método de intervalo por frequência

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- No exemplo (temperatura):
  - 64 65 68 69 |70 71 72 72 |75 75 80| 81 83 85

- 14 amostras: 4 Bins
  - $x \leq 69{,}5$: 64, 65, 68, 69
  - $69{,}5 < x \leq 73{,}5$: 70, 71, 72, 72
  - $73{,}5 < x \leq 80{,}5$: 75, 75, 80
  - $x > 80{,}5$: 81, 83, 85

Método de agrupamento

- Agrupamento (*Clustering*)
- Pode-se aplicar um algoritmo de agrupamento
- No caso unidimensional
- Para cada grupo (*cluster*) se atribui um valor discreto

Síntese da aula

- Atributos podem ser de diversos tipos: nominal, ordinal, intervalado e racional
- Dados podem ser organizados de diferentes maneiras: registros, matriz, grafos, etc.
- Necessário tratar problemas de qualidade dos dados como ruído, outliers, valores faltantes e duplicados
- Tratamento de dados envolve agregar dados e efetuar seleção e criação de características

Fontes*

- Slides sobre mineração de dados. Prof. Júlio Cesar Nievola Data Mining (PUCPR)

Este material foi parcialmente construído usando conteúdo dessas fontes