CIÊNCIA DE DADOS - 03

Prof. Júlio Cesar Nievola

PPGIa – PUCPR

11/maio/2019

O que são Dados?

- Coleção de objetos de dados e seus atributos

- Um atributo é uma propriedade ou característica de um objeto
  - Exemplos: cor dos olhos de uma pessoa, temperatura, etc.
  - Atributo também é conhecido como variável, campo ou característica
- Uma coleção de atributos descreve um objeto
  - Objeto também é conhecido como registro, ponto, caso, amostra, entidade, ou instância

Atributos

Objetos

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

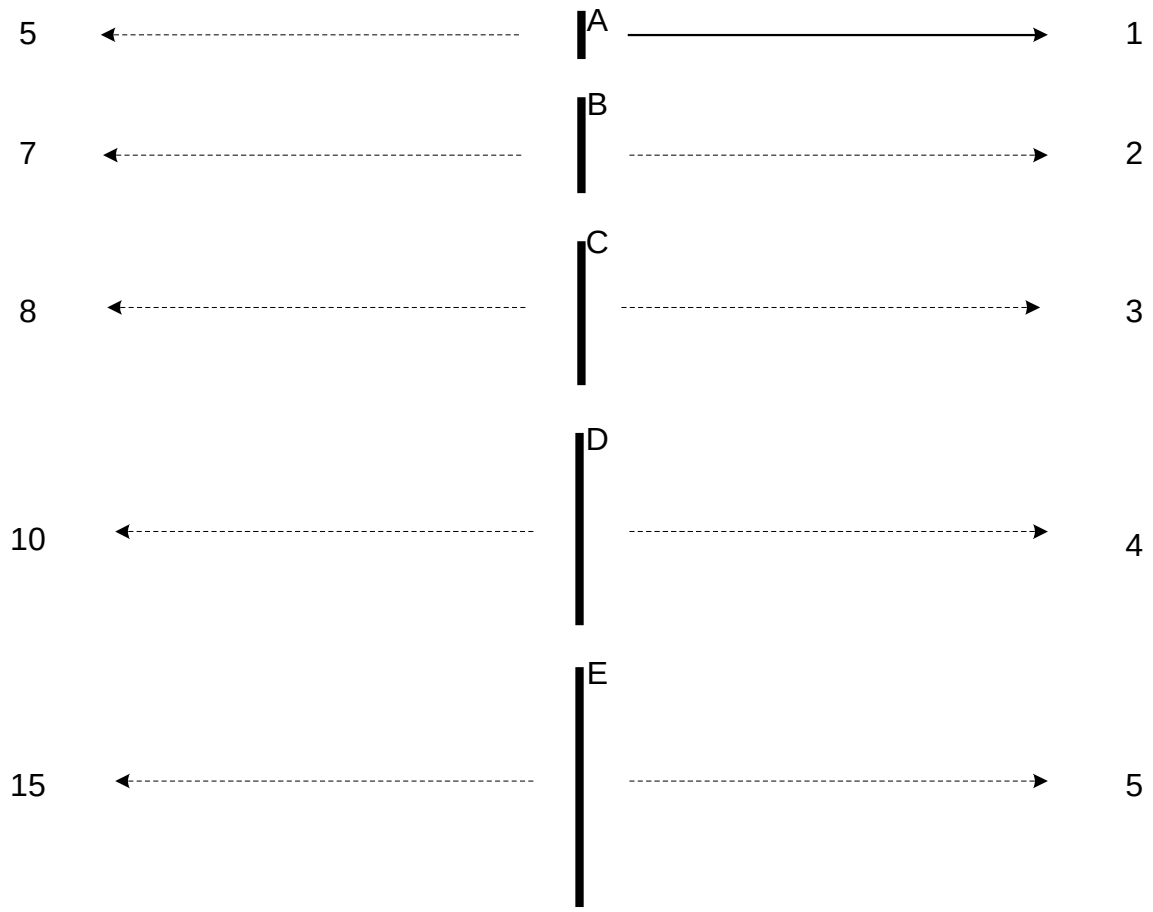
Valores de Atributos

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

- Distinção entre atributos e valores de atributos
  - Um mesmo atributo pode ser mapeado em diferentes valores de atributo
    - Exemplo: altura pode ser medida em pés ou metros
  - Diferentes atributos podem ser mapeados no mesmo conjunto de valores
    - Exemplo: Valores de atributo para ID e idade são inteiros
    - Mas propriedades dos valores dos atributos podem ser diferentes
      - ID não tem limite mas idade tem um valor máximo e um mínimo

Medidas de Comprimento

A forma com que se mede um atributo pode, às vezes, não estar de acordo com as propriedades dos atributos.



Tipos de Atributos

- Há diferentes tipos de atributos
  - Nominal
    - Exemplos: números de ID, cor dos olhos, códigos de CEP
  - Ordinal
    - Exemplos: ordem (e.g., gosto de batata frita em uma escala entre 1-10), graus, altura em {alto, médio, baixo}
  - Intervalar
    - Exemplos: datas de calendário, temperaturas em Celsius ou Fahrenheit.
  - Razão
    - Exemplos: temperatura em Kelvin, comprimento, tempo, contagem

Propriedades dos Valores dos Atributos

- O tipo de um atributo depende de quais das seguintes propriedades ele possui:
  - Distinção: = ≠
  - Ordem: < >
  - Adição: + -
  - Multiplicação: * /

  - Atributo Nominal: distinção
  - Atributo Ordinal: distinção & ordem
  - Atributo Intervalar: distinção, ordem & adição
  - Atributo Razão: todas as quatro propriedades

Tipo de atributo	Descrição	Exemplos	Operações
Nominal	Os valores de um atributo nominal são apenas nomes diferentes, i.e., atributos nominais fornecem só informação suficiente para distinguir um objeto de outro. (=, ≠)	Códigos CEP, números de ID de empregados, cor dos olhos, sexo: {*masculino, feminino*}	Moda, entropia, correlação de contingência, teste χ^2
Ordinal	Os valores de um atributo ordinal fornecem informação suficiente para ordenar objetos. (<, >)	Dureza de minerais, {*bom, melhor, o melhor*}, graus, número de ruas	Mediana, correlação de ordem, percentis, testes de execução, testes de sinal
Intervalar	Para atributos intervalares, as diferenças entre valores tem sentido, i.e., existe uma unidade de medida. (+, -)	Datas de calendário, temperatura em Celsius ou Fahrenheit	Média, desvio padrão, correlação de Pearson, testes *t* e *F*
Razão	Para variáveis do tipo razão, tanto diferenças quanto razão (divisão) tem sentido. (*, /)	Temperatura em Kelvin, quantidades monetárias, contagem, idade, massa, comprimento, corrente elétrica	Média geométrica, média harmônica, variação percentual

Nível do atributo	**Transformação**	**Comentários**
Nominal	Qualquer permutação de valores.	Se todos os números de ID dos empregados fosse re-assinalada, isto faria alguma diferença?
Ordinal	Uma alteração de valores que preserve a ordem, i.e., *novo_valor = f(valor_antigo)* em que *f* é uma função monotônica.	Um atributo abrangendo a noção de *bom*, *melhor*, *o melhor* pode ser igualmente representado pelos valores {1, 2, 3} ou { 0.5, 1, 10}.
Intervalar	*novo_valor =a * valor_antigo + b* em que *a* e *b* são constantes	Escalas de temperaturas em Fahrenheit e Celsius diferem em termos de onde o valor zero está e do tamanho da unidade (grau).
Razão	*novo_valor = a * valor_antigo*	Comprimento pode ser medido em metros ou pés.

Atributos Discretos e Contínuos

- Atributo Discreto
  - Tem um conjunto de valores finito ou contavelmente infinito
  - Exemplos: código CEP, contagens, ou o conjunto de palavras em uma coleção de documentos
  - Freqüentemente representados como variáveis inteiras.
  - Nota: atributos binários são um caso especial de atributos discretos.

- Atributos Contínuos
  - Tem números reais como atributos de valores
  - Exemplos: temperatura, altura, ou peso.
  - Na prática, valores reais somente podem ser medidos e representados usando um número finito de dígitos.
  - Atributos Contínuos são representados tipicamente como variáveis de ponto flutuante.

Tipos de conjuntos de dados

- **Registro**
  - **Matriz de dados**
  - **Dados de documentos**
  - **Dados de transações**
- **Grafo**
  - **World Wide Web**
  - **Estruturas Moleculares**
- **Ordenados**
  - **Dados espaciais**
  - **Dados temporais**
  - **Dados seqüenciais**
  - **Dados de seqüências genéticas**

Características Importantes de Dados Estruturados

- **Dimensionalidade**
  - **Maldição da Dimensionalidade**
- **Esparsidade**
  - **Somente a presença importa**
- **Resolução**
  - **Padrões dependem da escala**

Dados de Registros

- Dados que consistem de um coleção de registros, cada um dos quais consiste de um conjunto fixo de atributos

ID	Restituição?	Estado Civil	Receita anual	Declaração correta?
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Dados Matriciais

- Se os objetos de dados tem o mesmo conjunto fixo de atributos numéricos, então os objetos de dados podem ser vistos como pontos em um espaço multidimensional, em que cada dimensão representa um atributo distinto

- Tal conjunto de dados pode ser representado por uma matriz *m* por *n*, em que há *m* linhas, uma para cada objeto, e *n* colunas, uma para cada atributo

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

Dados de Documentos

- Cada documento torna-se 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 correspondente ocorre no documento.

	Time	Treinador	Jogos	Bola	Placar	Jogada	Win	Lost	Timeout	Temporadas
Documento 1	3	0	5	0	2	6	0	2	0	2
Documento 2	0	7	0	2	1	0	0	3	0	0
Documento 3	0	1	0	0	1	2	2	0	3	0

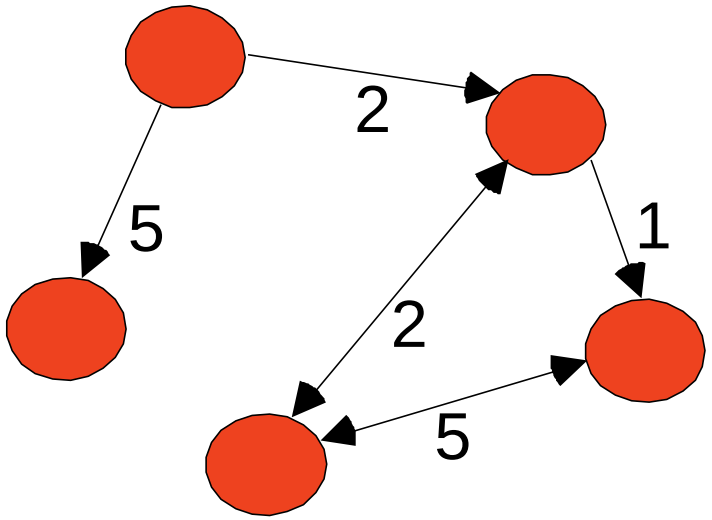
Dados de Transações

- São dados de registro de um tipo especial, em que
  - cada registro (transação) envolve um conjunto de itens.
  - Por exemplo, considere um supermercado. O conjunto de produtos comprados por um cliente durante constitui uma transação, enquanto os produtos individuais comprados são os itens.

ID	*Itens*
1	Pão, Refri, Leite
2	Cerveja, Pão
3	Cerveja, Refri, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Refri, Fralda, Leite

Dados de Grafos

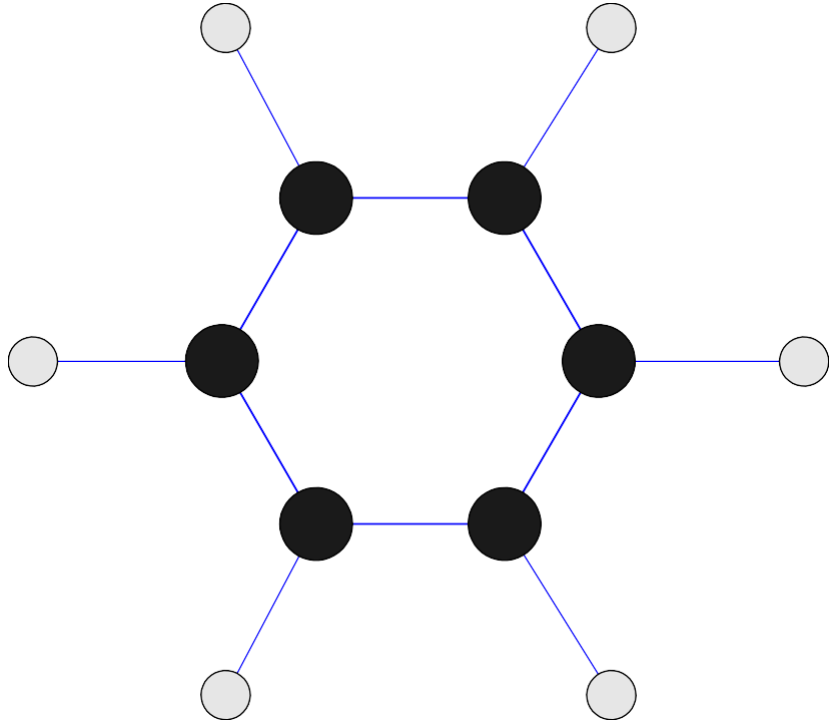
- Exemplos: Grafos genéricos e links HTML




```
<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
```

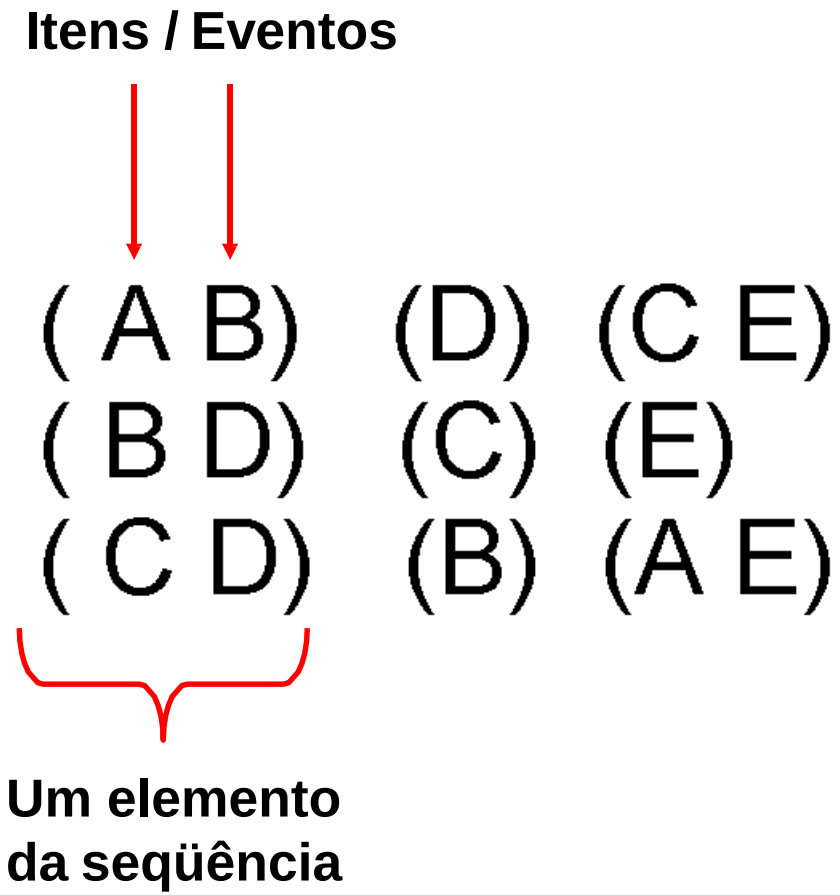
Dados Químicos

- Molécula de Benzeno: C_6H_6



Dados Ordenados – 1

- Seqüências de transações



Dados Ordenados – 2

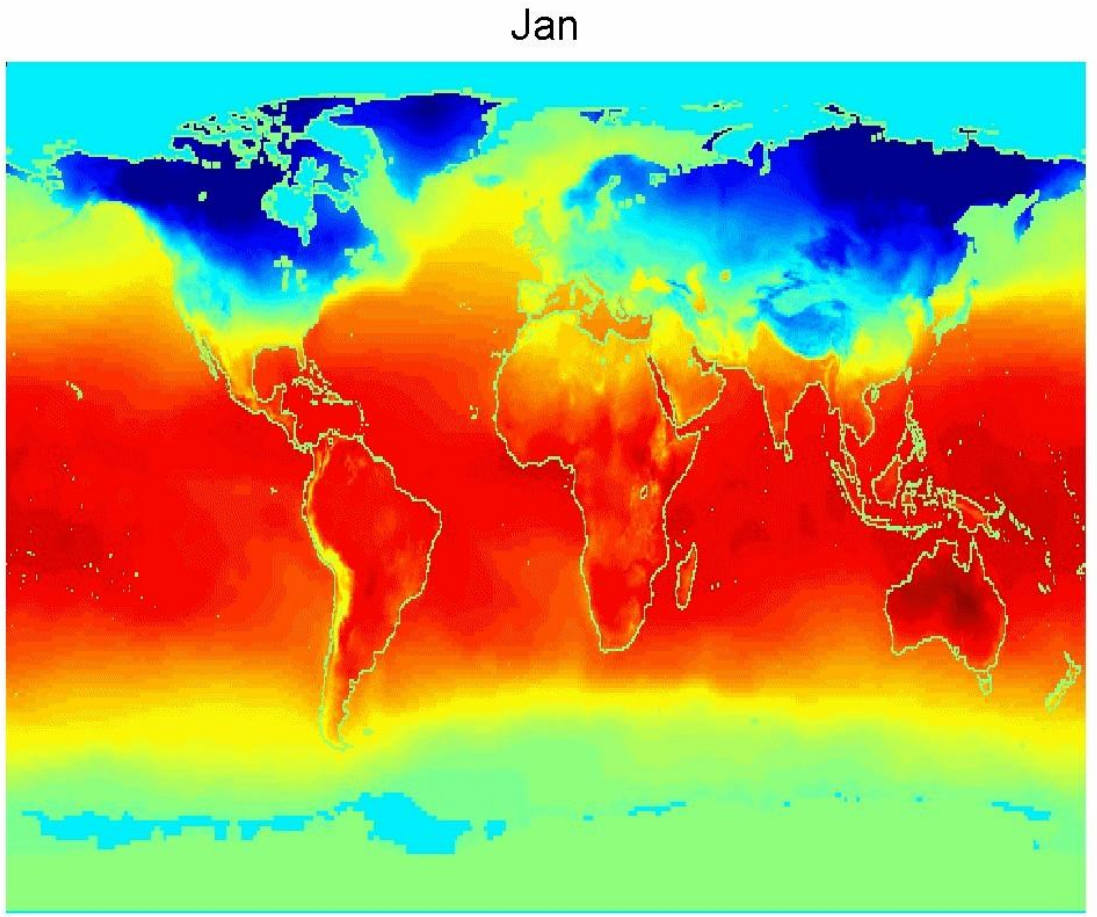
- Dados de seqüência genômica

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

Dados Ordenados – 3

- Dados Espaço-Temporais

Temperatura Média Mensal das terras e oceanos



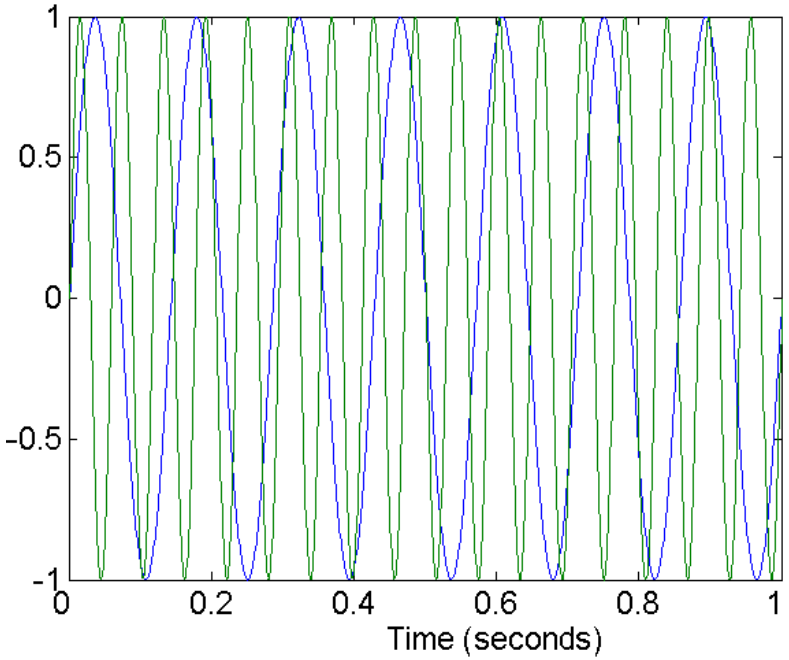
Qualidade dos Dados

- Que tipo de problemas de qualidade de dados?
- Como se pode detectar problemas nos dados?
- O que se pode fazer a respeito destes problemas?

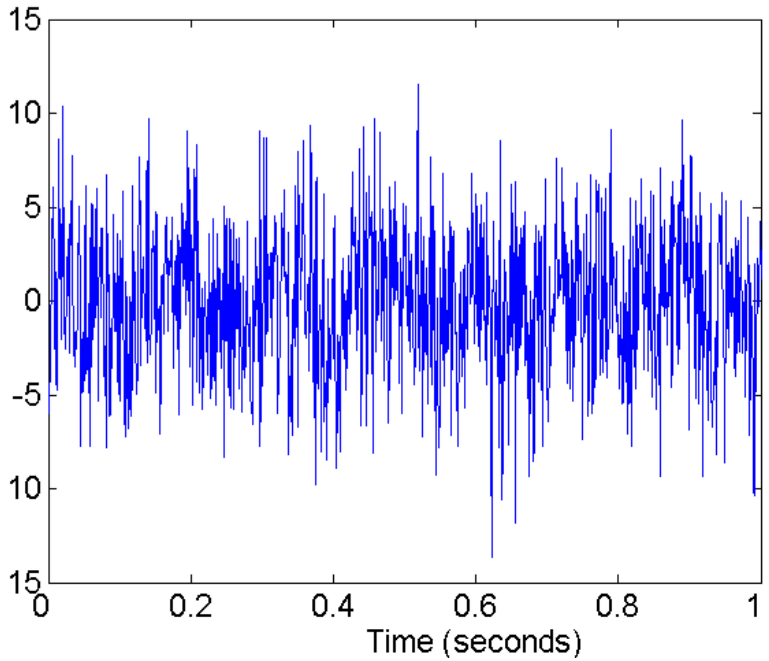
- Exemplos de problemas de qualidade nos dados:
  - Ruídos e outliers
  - Dados faltantes
  - Dados duplicados

Ruído

- Ruído refere-se à modificação de valores originais
  - Exemplos: distorção da voz de uma pessoa falando



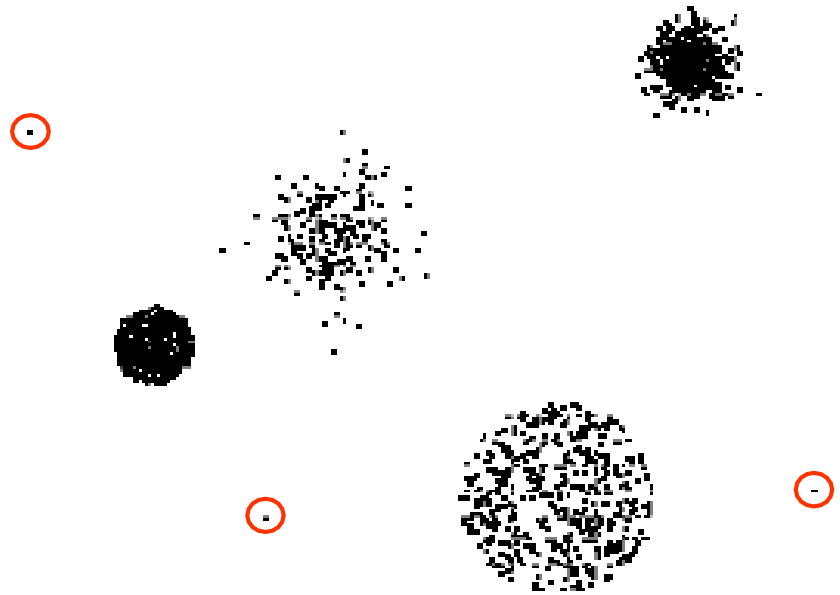
Duas ondas senoidais



Duas ondas senoidais + Ruído

Outliers

- Outliers são objetos de dados com características que são consideravelmente diferentes da maioria dos outros objetos de dados no conjunto de dados



Valores Faltantes

- Razões para valores faltantes
  - Informação não foi coletada
    (e.g., pessoas não fornecem sua idade e peso)
  - Atributos podem não ser aplicáveis a todos os casos (e.g., salário anual não é aplicável a crianças)

- Manipulando valores faltantes
  - Eliminar objetos de dados
  - Estimar valores faltantes
  - Ignorar valores faltantes durante análise
  - Substituir por todos os valores possíveis (ponderados por suas probabilidades)

Dados Duplicados

- Conjunto de dados pode incluir objetos de dados que são duplicatas, ou quase duplicadas de outros
  - Grande problema quando unindo dados de fontes heterogêneas
- Exemplos:
  - Mesma pessoa com múltiplos endereços de email
- Limpeza dos dados
  - Processo de trabalho com dados duplicados

Pré-processamento de Dados

- Agregação
- Amostragem
- Redução de Dimensionalidade
- Seleção de Subconjuntos de Características
- Criação de Características
- Discretização e Binarização
- Transformação de Atributos

Agregação

- Combinar dois ou mais atributos (ou objetos) em um único atributo (ou objeto)
- Finalidade
  - Redução de dados
    - Reduzir o número de atributos ou objetos
  - Alteração de escala
    - Cidades agregadas em regiões, estados, países, etc
  - Dados mais “estáveis”
    - Dados agregados tendem a ter menor variabilidade

Amostragem

- **Amostragem é a principal técnica empregada na seleção de dados**
  - **Usada freqüentemente tanto para investigação preliminar dos dados quanto para análise final dos dados.**

- **Estatísticos amostram porque obter o conjunto completo dos dados de interesse é muito caro ou consome tempo demais.**

- **Amostragem é usada em mineração de dados porque o processamento do conjunto inteiro dos dados de interesse é muito caro ou consome tempo demasiado.**

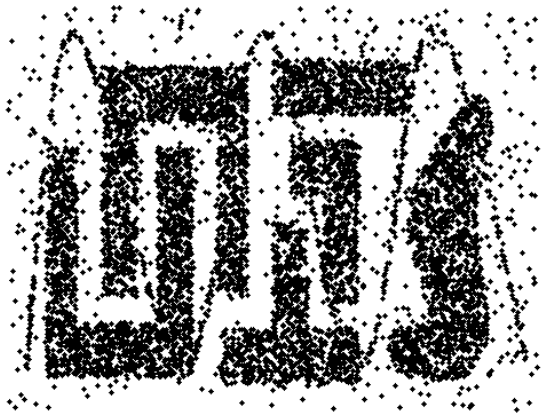
Amostragem ...

- O princípio básico para amostragem efetiva é o seguinte:
  - usando uma amostra funcionará tão bem quanto usando o conjunto completo de dados se a amostra é representativa
  - uma amostra é representativa se ela tem aproximadamente as mesmas propriedades (de interesse) quanto o conjunto original de dados

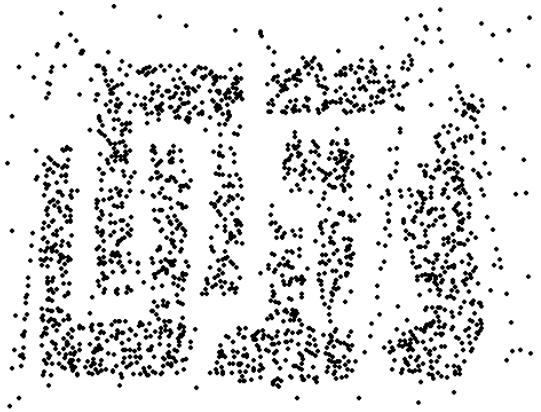
Tipos de amostragem

- Amostragem simples aleatória
  - Há uma probabilidade igual de selecionar qualquer item particular
- Amostragem sem reposição
  - À medida que cada item é selecionado, ele é removido da população
- Amostragem com reposição
  - Objetos não são removidos da população quando são selecionados para compor a amostra.
    - Na amostragem com reposição, o mesmo objeto pode ser escolhido mais de uma vez
- Amostragem estratificada
  - Divide os dados em várias partições; retira então amostras aleatórias de cada uma das partições

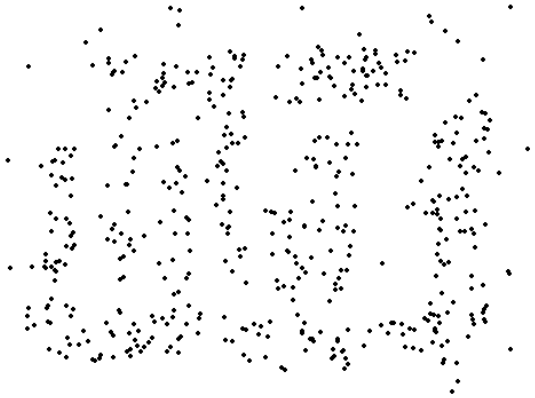
Tamanho da amostra



8000 pontos



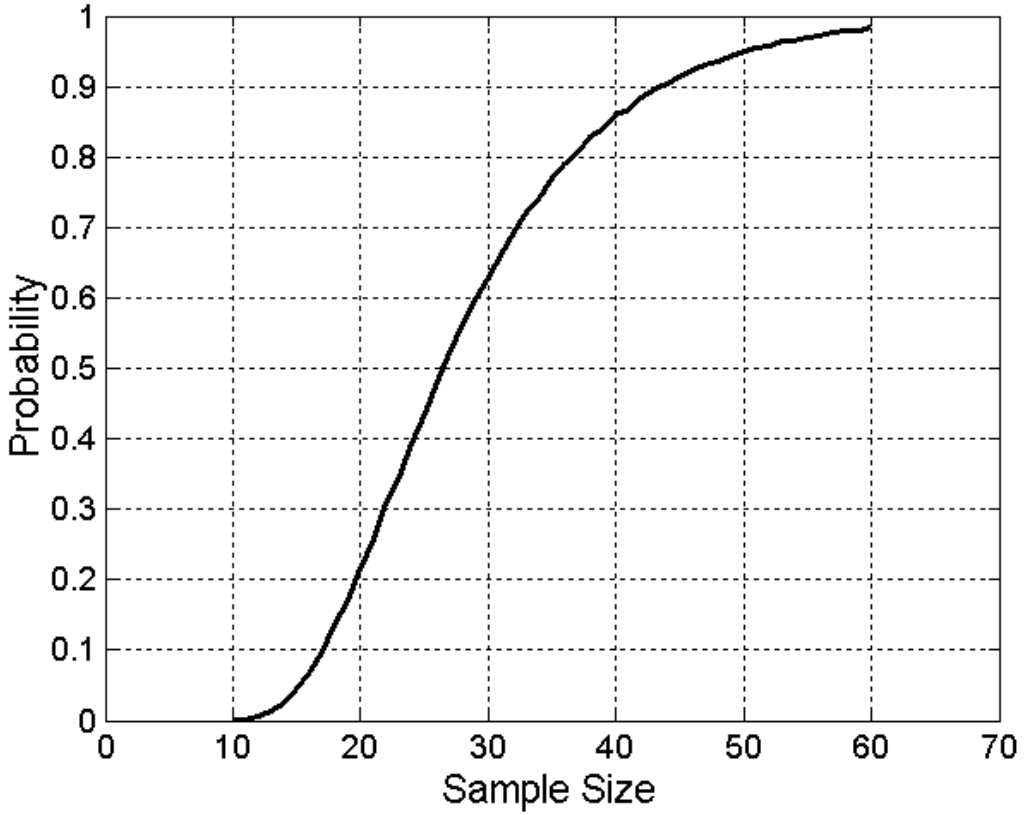
2000 pontos



500 pontos

Tamanho da amostra

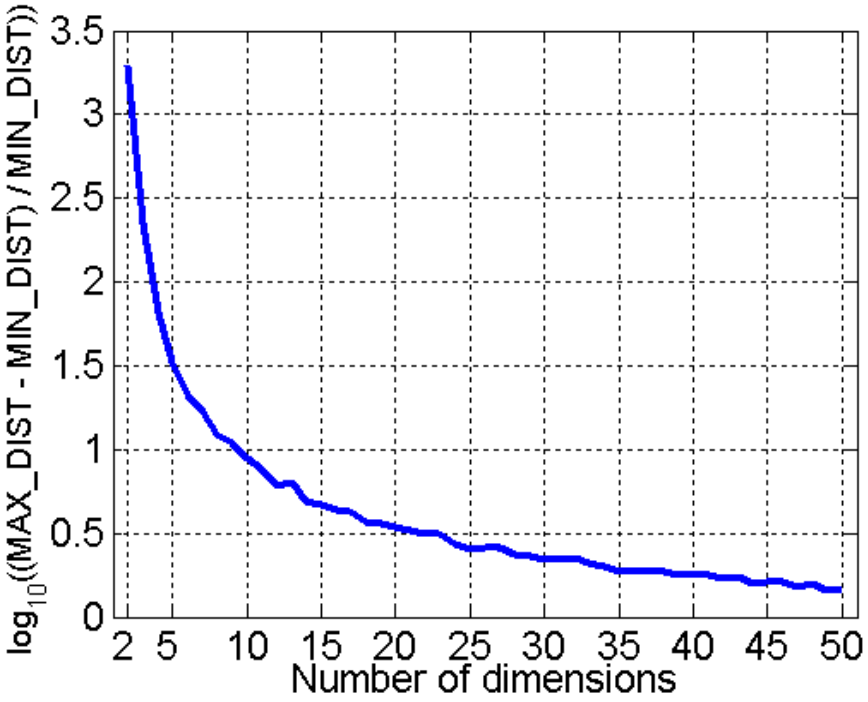
Que tamanho de amostra é necessário para obter pelo menos um objeto de cada um de 10 grupos?



Maldição da Dimensionalidade

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

- Definições de densidade e distância entre pontos, que são críticas para agrupamento e detecção de outliers, passam a ter menos significado



- Gerar aleatoriamente 500 pontos
- Calcular a diferença entre a distância máxima e mínima entre quaisquer pares de pontos

Redução de Dimensionalidade

- Finalidade:
  - Reduzir a maldição da dimensionalidade
  - Reduzir a quantidade de tempo e memória necessárias pelos algoritmos de mineração de dados
  - Permitir que os dados sejam mais facilmente visualizados
  - Ajudar a eliminar características irrelevantes ou a reduzir o ruído

- Técnicas
  - Análise de Componentes Principais – PCA
  - Singular Value Decomposition – SVD
  - Outros: técnicas supervisionadas e não-lineares

Seleção de Subconjuntos de Características

- Outra forma de reduzir a dimensionalidade dos dados
- Características redundantes
  - Duplicam muita ou toda a informação contida em um ou mais atributos
  - Exemplo: preço de venda de um produto e a quantidade de taxas de venda pagas
- Características irrelevantes
  - Não contém informação que seja útil para a tarefa de mineração de dados sendo executada
  - Exemplo: ID do estudante é freqüentemente irrelevante na tarefa de prever o seu desempenho

Seleção de Subconjuntos de Características

- Técnicas:
  - Abordagem de força bruta:
    - Tenta todos os subconjuntos possíveis de características como entrada para o algoritmo de mineração de dados
  - Abordagem embutidas:
    - Seleção de características ocorre naturalmente como parte do algoritmo de mineração de dados
  - Abordagem filtro:
    - Características são selecionadas antes que o algoritmo de mineração de dados seja executado
  - Abordagem wrapper:
    - Uso o algoritmo de mineração de dados como uma caixa preta para encontrar o melhor subconjunto de atributos

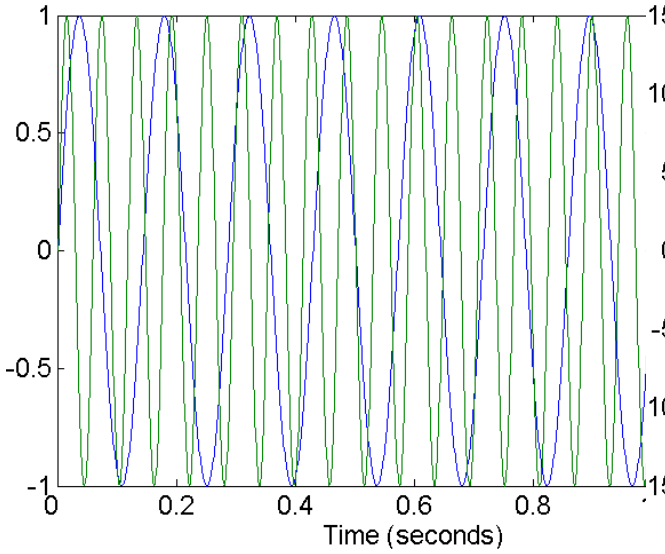
Criação de Características

- Cria novos atributos que podem capturar informação importante em um conjunto de dados muito mais eficientemente que os atributos originais

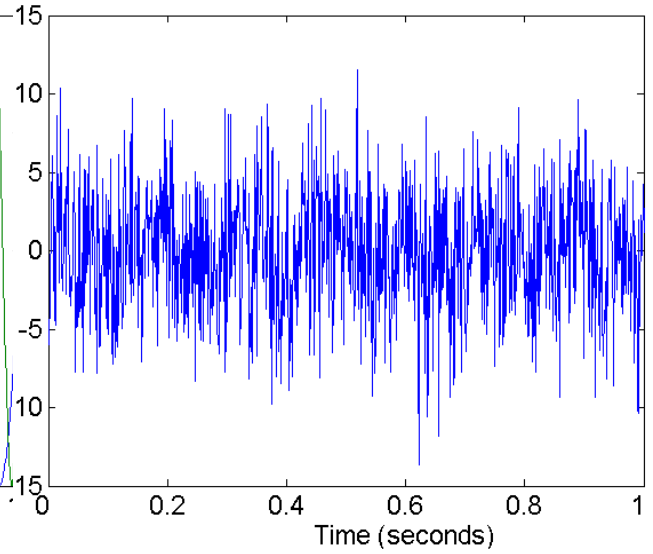
- Três metodologias gerais:
  - Extração de características
    - específicas do domínio
  - Mapeamento de dados para novo espaço
  - Construção de características
    - combinando características

Mapeando Dados para um Novo Espaço

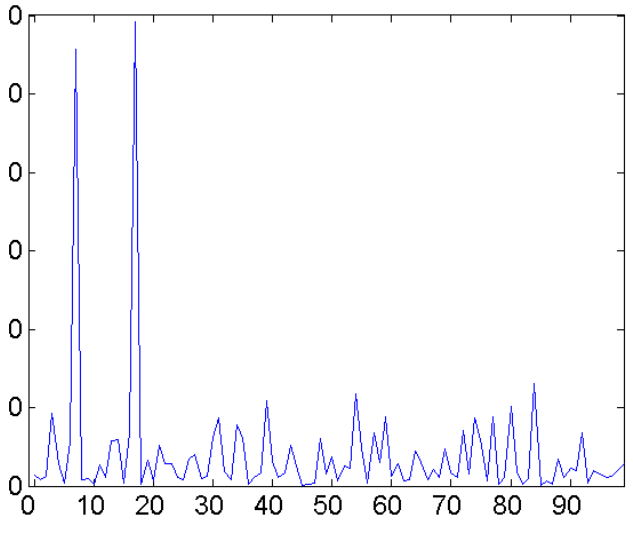
- Transformada de Fourier
- Transformada Wavelet



Duas ondas senoidais



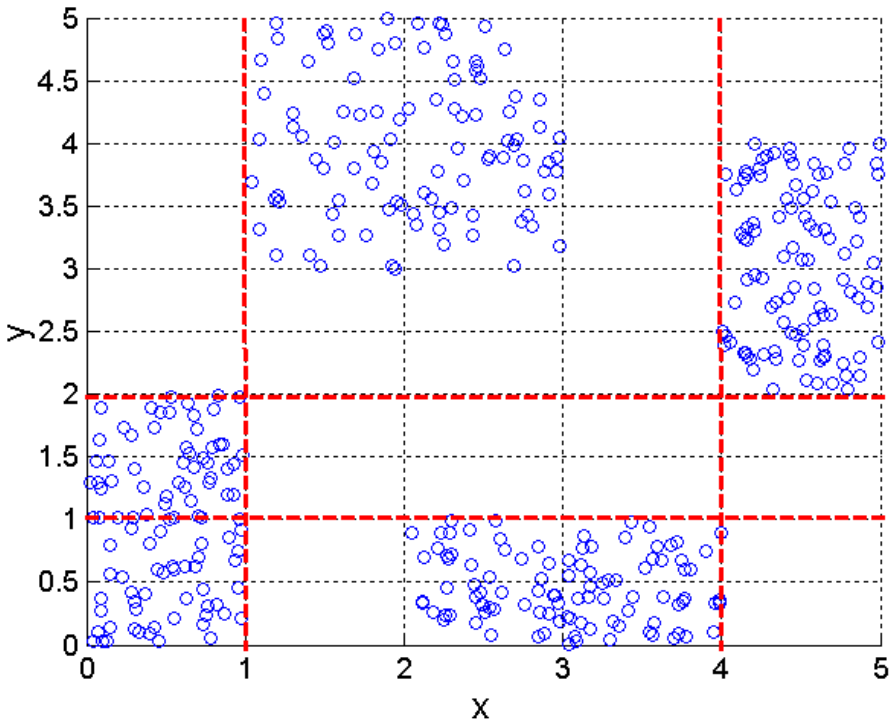
Duas ondas senoidais + Ruído



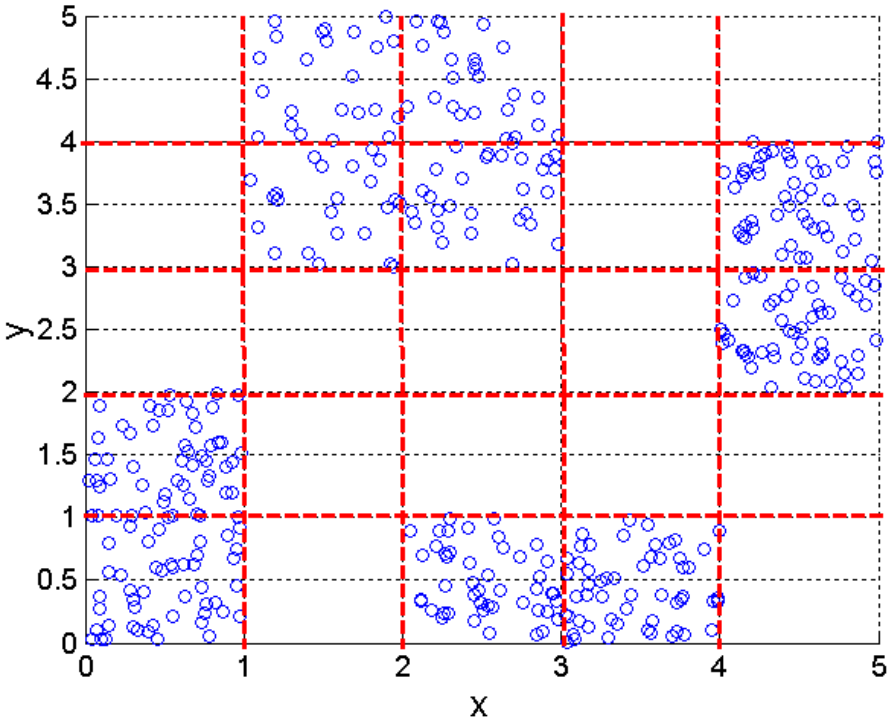
Freqüência

Discretização Usando Rótulos das Classes

Abordagem baseada em Entropia

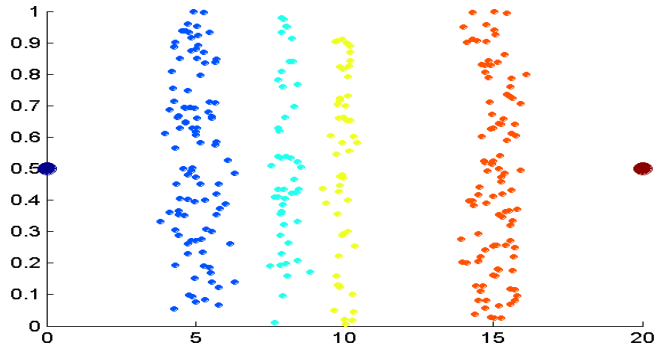


3 categorias tanto para *x* quanto *y*

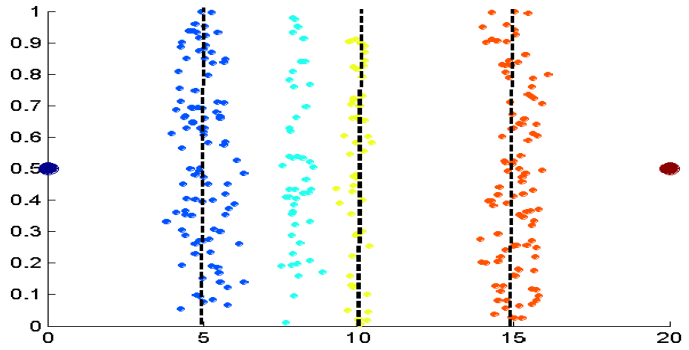


5 categorias tanto para *x* quanto *y*

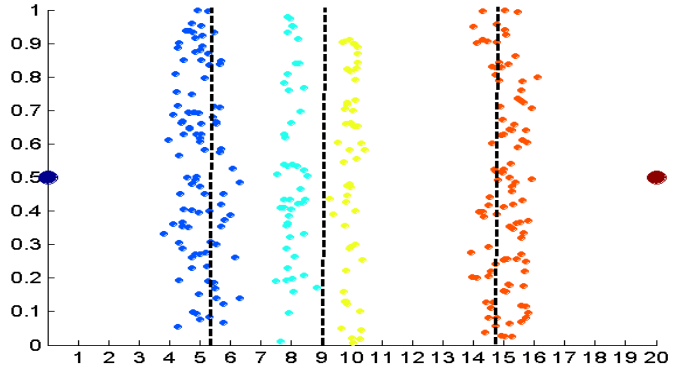
Discretização sem Usar Rótulos das Classes



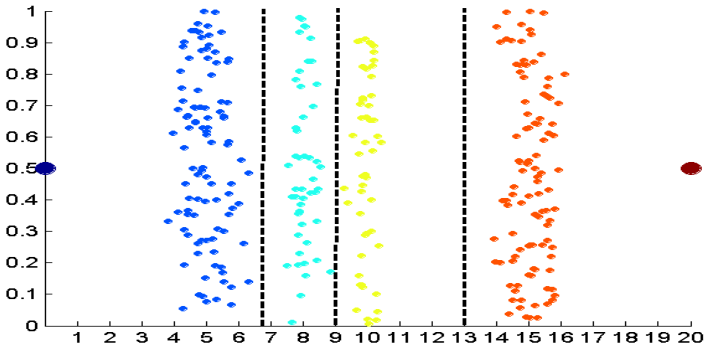
Dados



Intervalos de largura igual



Freqüência igual



K-médio

Transformação de Atributos

- Uma função que mapeia o conjunto inteiro de valores de um dado atributo para um novo conjunto de valores de substituição tal que cada valor antigo pode ser identificado com um dos novos valores
  - Funções simples: x^k, log(x), e^x, |x|
  - Padronização e Normalização