INTELIGÊNCIA COMPUTACIONAL

PREPARANDO OS DADOS

FELIPE TORRES

CONHECER OS DADOS DO SEU DATASET

- O primeiro passo para a análise de um conjunto de dados ou dataset é conhecer os seus dados. Esta etapa auxilia:
  - A definir as técnicas de pré-processamento.
  - Ajuda a projetar e definir os algoritmos que serão utilizados no data mining.
  - Entender os resultados das suas análises

CONHECER OS DADOS DO SEU DATASET

- Dados podem ser de diversos tipos:
  - Registros (em um BD)
  - Matriz de dados (numéricos)
  - Palavras de um documento
  - Dados de um grafo
  - Ordem temporal ou sequencial
  - Dados espaciais
  - Imagens, multimídia

CONJUNTO DE DADOS

- Conjuntos de dados são formados por instâncias
  - Amostras, exemplos, objetos, tuplas, pontos, entidades, casos, vetores
  - Instâncias são descritas por atributos
  - Linhas em um BD são instâncias
  - Colunas em um BD são atributos

TIPOS DE DADOS

- Atributo (ou dimensões, features, variáveis):
  - um campo representando uma característica da instância. ex.: nome, endereço, ID

- Tipos:
  - Nominal: qualitativo
  - Binário: qualitativo
  - Ordinal: qualitativo
  - Numérico: quantitativo

TIPOS DE DADOS

- Nominal
  - valor do atributo é um nome para algo
  - uma categoria, um código, um estado
  - exemplos: estado civil, cor do cabelo, ocupação
  - podem ser representados por números arbitrários
    - mas não tem sentido efetuar operações entres estes valores, não são quantitativos, nem tem ordenação
  - não há média nem mediana

TIPOS DE DADOS

- Binário
  - atributos com dois valores: 0 e 1
  - ausente ou presente, sim ou não
  - exemplo: fumante? possui carro?
    - simétrico: ambos valores são relevantes
    - assimétrico: um valor é mais relevante
    - (normalmente, o valor 1 é utilizado)

TIPOS DE DADOS

- Ordinal
  - valores possuem uma ordem (ranking)
  - o valor em si não tem significado
  - exemplo: notas, tamanho P exemplo: notas, tamanho P-M-G, escala de G, escala de satisfação
  - podem vir da discretização de quantidades numéricas

TIPOS DE DADOS

- Numérico
  - quantitativo, quantidade mensurável
  - escala por intervalo (interval-scaled)
    - escala de unidades de mesmo tamanho
    - ordem, há diferença entre valores
    - não há zero verdadeiro, indicando ausência
    - exemplo: temperatura em celsius, dias de calendário
- escala por razão (ratio-scaled):
  - há zero verdadeiro
  - exemplo: temperatura em K, valor monetário

TIPOS DE DADOS

- Atributo Discreto
  - Somente um conjunto finito ou contável de valores
  - ex.: cep, profissão, palavras em coleção de documentos
  - As vezes, representado por variáveis inteiras
- Atributo Contínuo
  - Número reais
  - ex.: temperatura, altura, peso
  - Na prática, medimos e representamos usando um número finito de dígitos ⇨ ponto-flutuante

TIPOS DE DADOS

- Características de conjuntos de dados
  - Dimensionalidade
    - Número de atributos
    - Maldição da dimensionalidade (dimensionality curse): redução?
  - Esparcidade (Sparcity)
    - Atributos ausentes ou 0
  - Resolução
    - Padrões dependem da escala

ESTATÍSTICA BÁSICA DOS DADOS

- Medidas para compreender melhor os seus dados.
- A utilização de medidas de tendência central para entender o comportamento dos dados.
- Medidas de dispersão e visualização dos dados, auxiliam também no entendimento do dataset.

ESTATÍSTICA BÁSICA DOS DADOS – TENDÊNCIA CENTRAL

- Média Aritmética
  - n é tamanho da amostra, N é tamanho da população

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Média Ponderada
  - Média ponderada é a média aritmética com um quociente x_i ponderando os valores

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

ESTATÍSTICA BÁSICA DOS DADOS – TENDÊNCIA CENTRAL

- Mediana
  - valor do meio se quantidade impar, ou média dos valores do meio se quantidade par
  - atributos numéricos e ordinais

2, 2, 3, **7**, 8, 9, 9
Mediana = 7

1, 4, 4, **5**, **6**, 7, 7, 7
Mediana = (5+6) ÷ 2
= 5.5

ESTATÍSTICA BÁSICA DOS DADOS – TENDÊNCIA CENTRAL

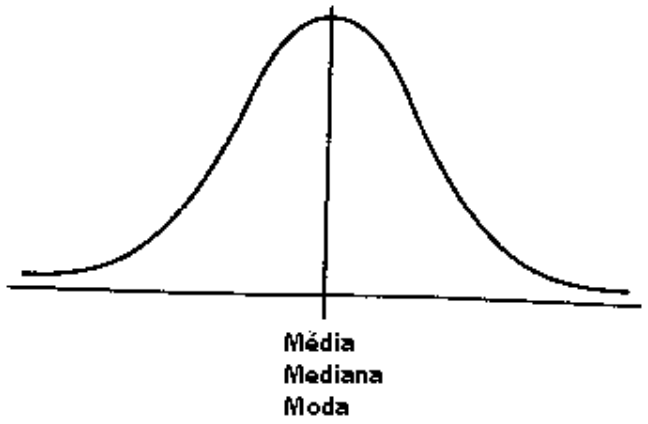
- Moda
  - valor mais frequente
  - uma moda: unimodal
  - bimodal, trimodal, multimodal
  - atributos numéricos, ordinais e nominais

$S1 = \{1,1,1,1,1,1,1,1,1,1\}$, $\boldsymbol{a\ moda}$ é $\mathbf{1}$

$S2 = \{1,1,1,2,2,3,4\}$, $\boldsymbol{a\ moda}$ é $\mathbf{1\ e\ 2}$

$S3 = \{1,1,2,2,3,3,4,4\}$, $\boldsymbol{a\ moda}$ é $\mathbf{1, 2, 3\ e\ 4}$

ESTATÍSTICA BÁSICA DOS DADOS – DISTRIBUIÇÃO SIMÉTRICA



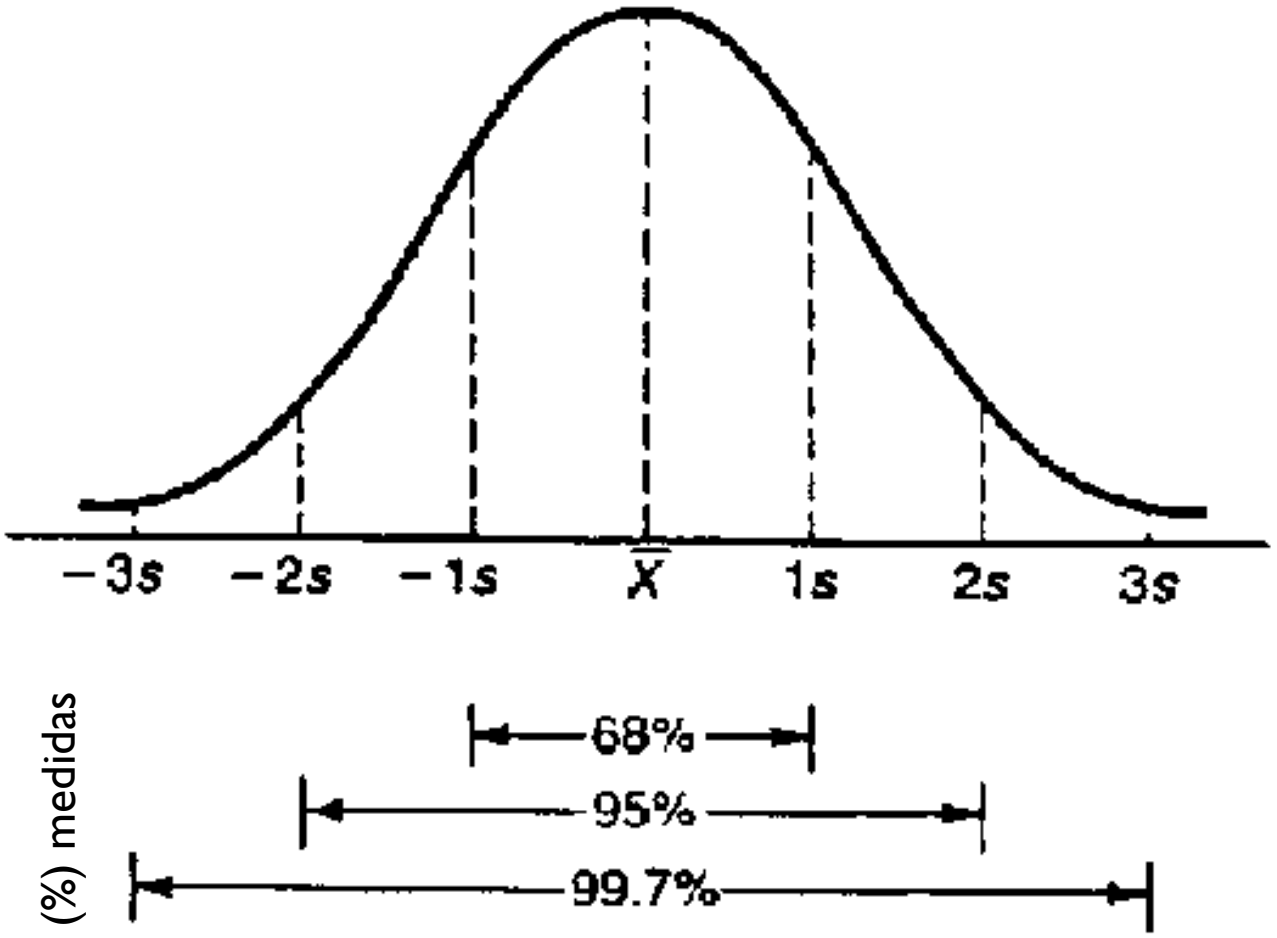
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Variância e desvio padrão
  - amostra: s, população: σ
  - variância

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

  - Desvio padrão é a raiz quadrada da variância

ESTATÍSTICA BÁSICA DOS DADOS – DESVIO PADRÃO



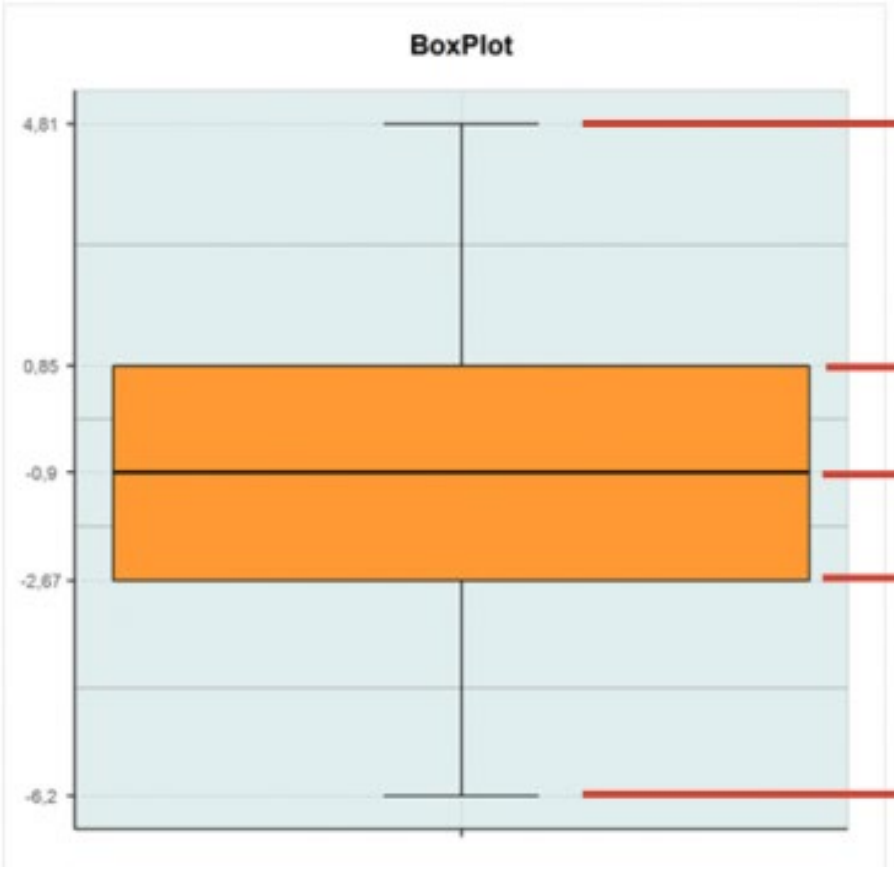
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Quantil
  - Pontos que dividem dados ordenados em q subconjuntos de tamanho igual
  - Cada subconjunto é um q-quantil, teremos (q-1) q-quantis
  - q=100: os 100-quatil são percentis
  - q=4: os 4-quantil são quartis

ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Quartil, outliers e boxplots
  - Quartil: Q1 (percentil de 25%), Q3 (percentil de 75%)
  - Inter-quartil range: IQR = Q3 quartil range: IQR = Q3 – Q1
  - Resumo de 5 valores: min, Q1, mediana, Q3, max
  - Boxplot: final da caixa são os quartis, mediana é marcada, além de bigodes e outliers
  - Outlier: usualmente, um valor maior/menor que 1.5*IQR

ESTATÍSTICA BÁSICA DOS DADOS – BOX PLOT



Limite Superior

Terceiro Quartil Q_3

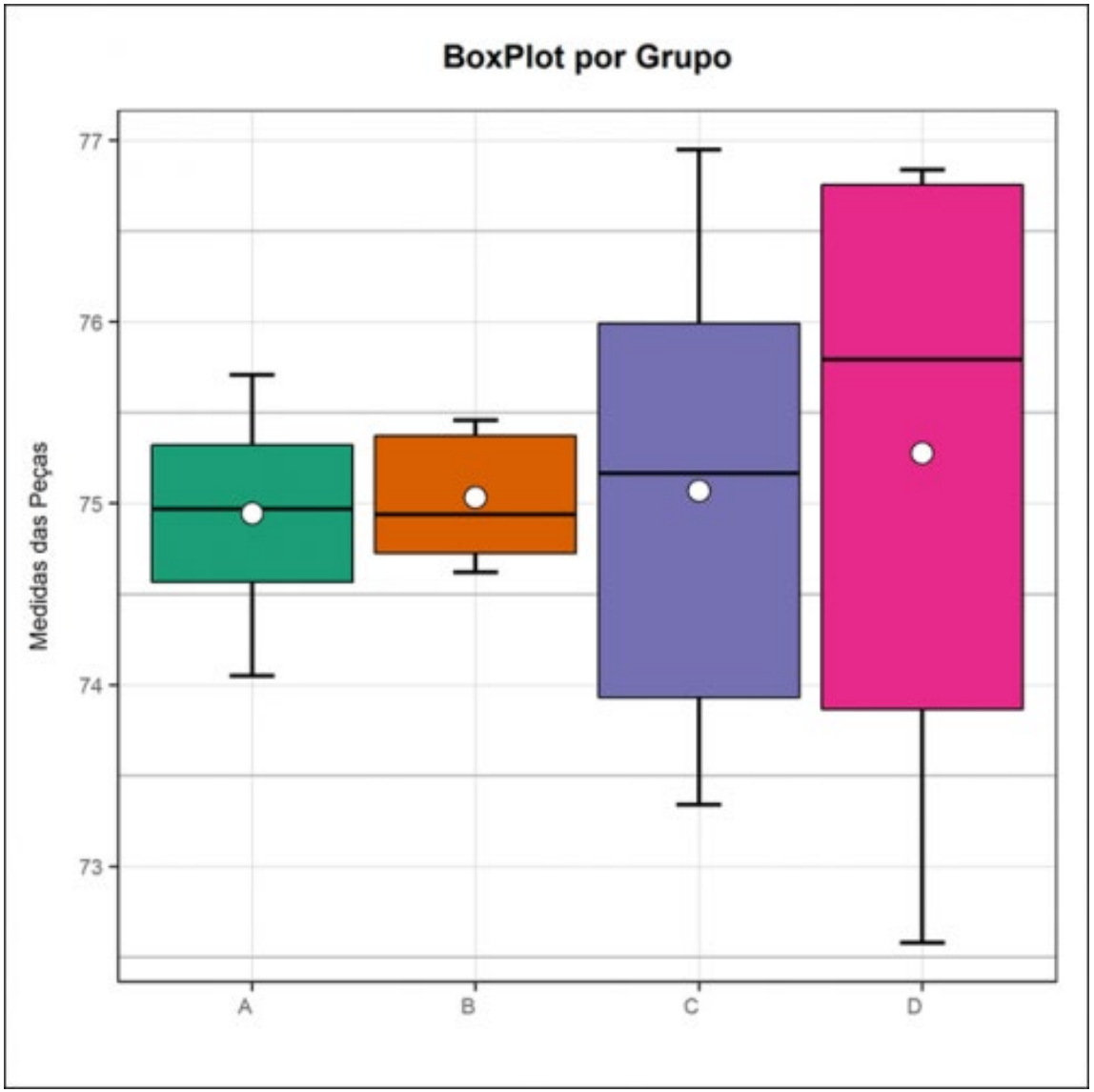
Segundo Quartil ou Mediana Q_2

Primeiro Quartil Q_1

Limite Inferior

ESTATÍSTICA BÁSICA DOS DADOS – BOX PLOT

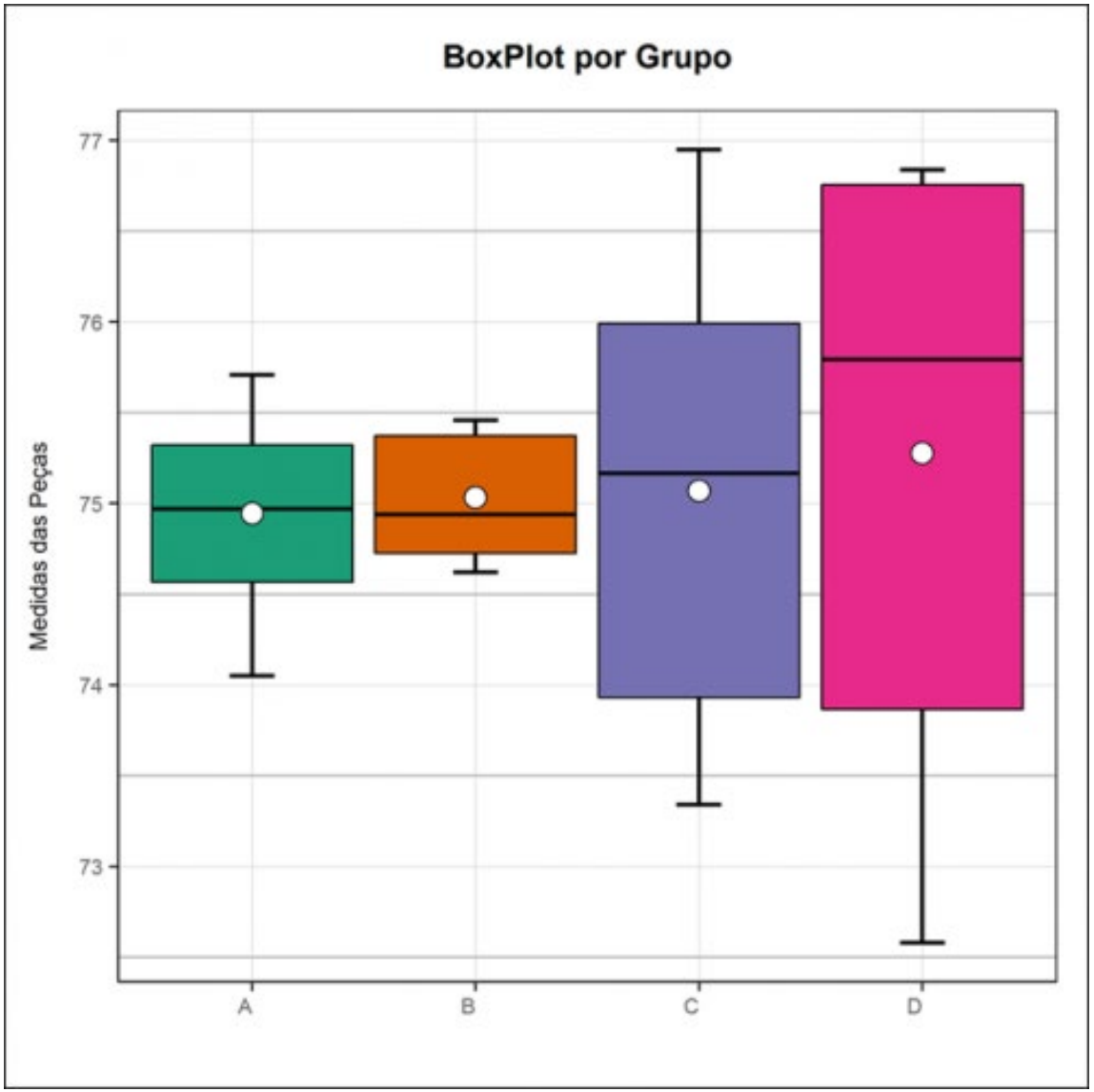
Uma indústria produz uma peça automotiva cujo valor de referência é 75cm. Após verificar lotes com peças fora de especificação, enviaram duas equipes de trabalhadores (A e B) para um treinamento. Para verificar a eficiência do treinamento, foram selecionadas 10 peças produzidas pelas equipes A e B e 10 peças produzidas pelas equipes C e D que não participaram do treinamento.



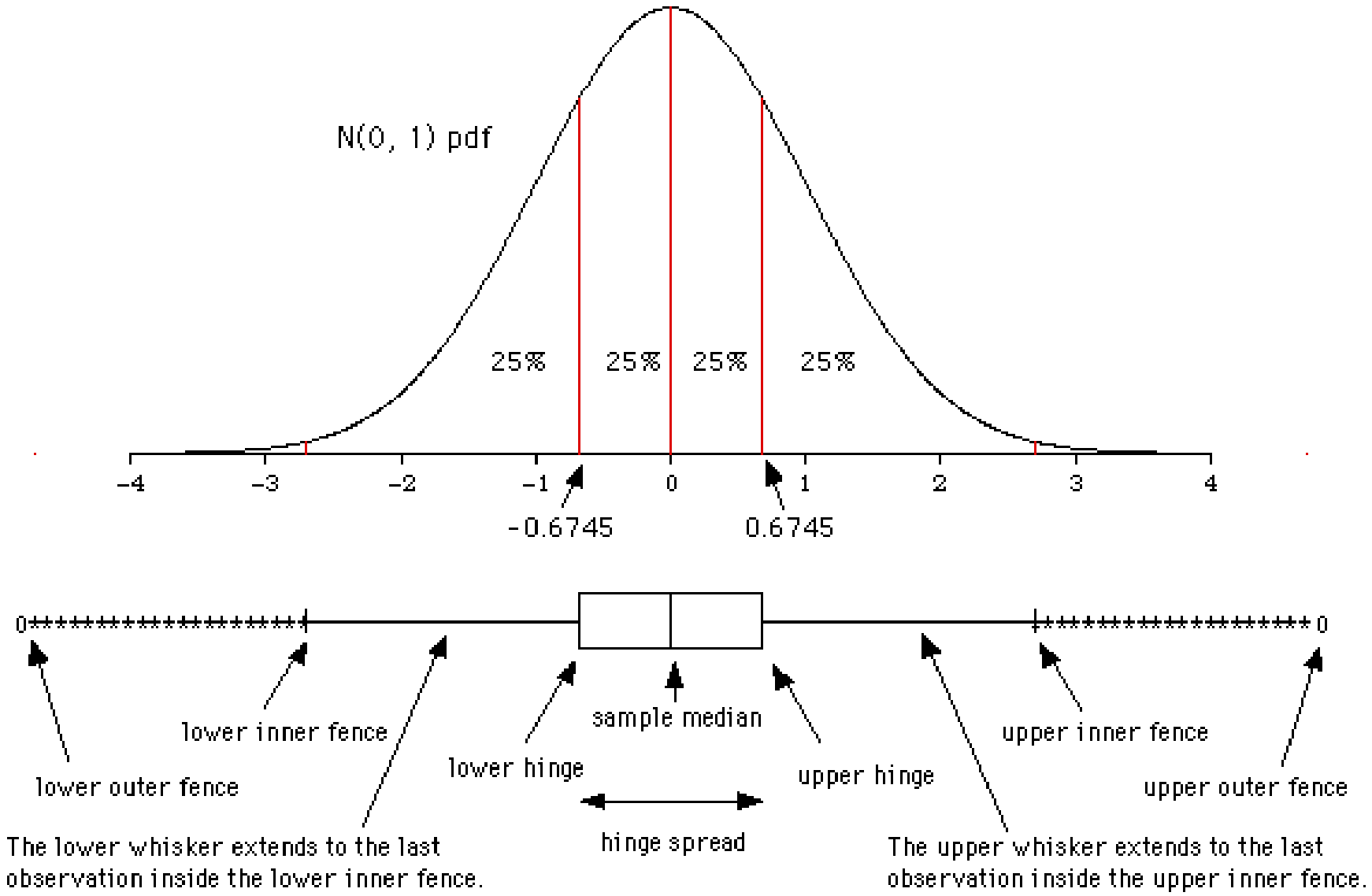
ESTATÍSTICA BÁSICA DOS DADOS – BOX PLOT

1. As equipes A e B produzem peças com menor variabilidade, indicando que o treinamento teve o efeito desejado;
2. A equipe D é a que produz peças com maior variabilidade;
3. A equipe B é a que produz peças com menor variabilidade.

Como as peças das equipes A e B tem menor variabilidade e com valor médio próximo do valor de referência, vale a pena enviar as demais equipes para o treinamento.

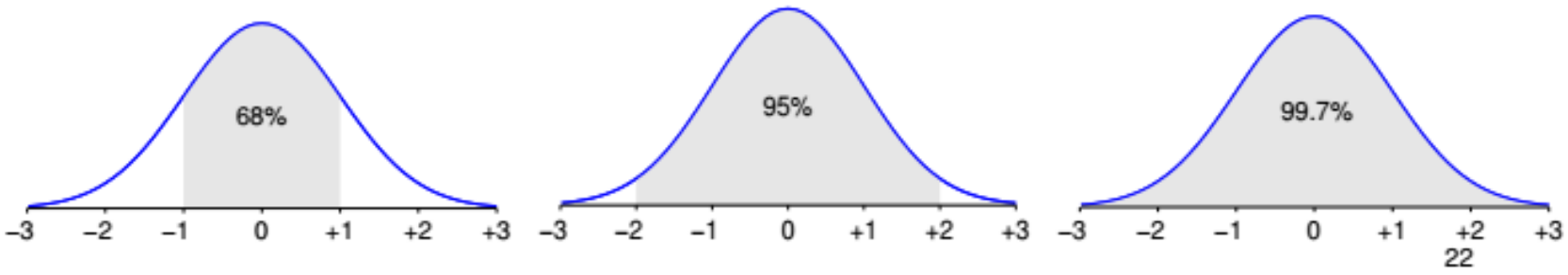


ESTATÍSTICA BÁSICA DOS DADOS – BOX PLOT



ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Curva de distribuição normal
  - De μ–σ até μ+σ: contém até 68% das medidas (μ: média, σ: desvio padrão)
  - De μ–2σ até μ+2σ: contém cerca de 95%
  - De μ–2σ até μ+2σ: contém cerca de 95%
  - De μ–3σ até μ+3σ: contém cerca de 99.7%



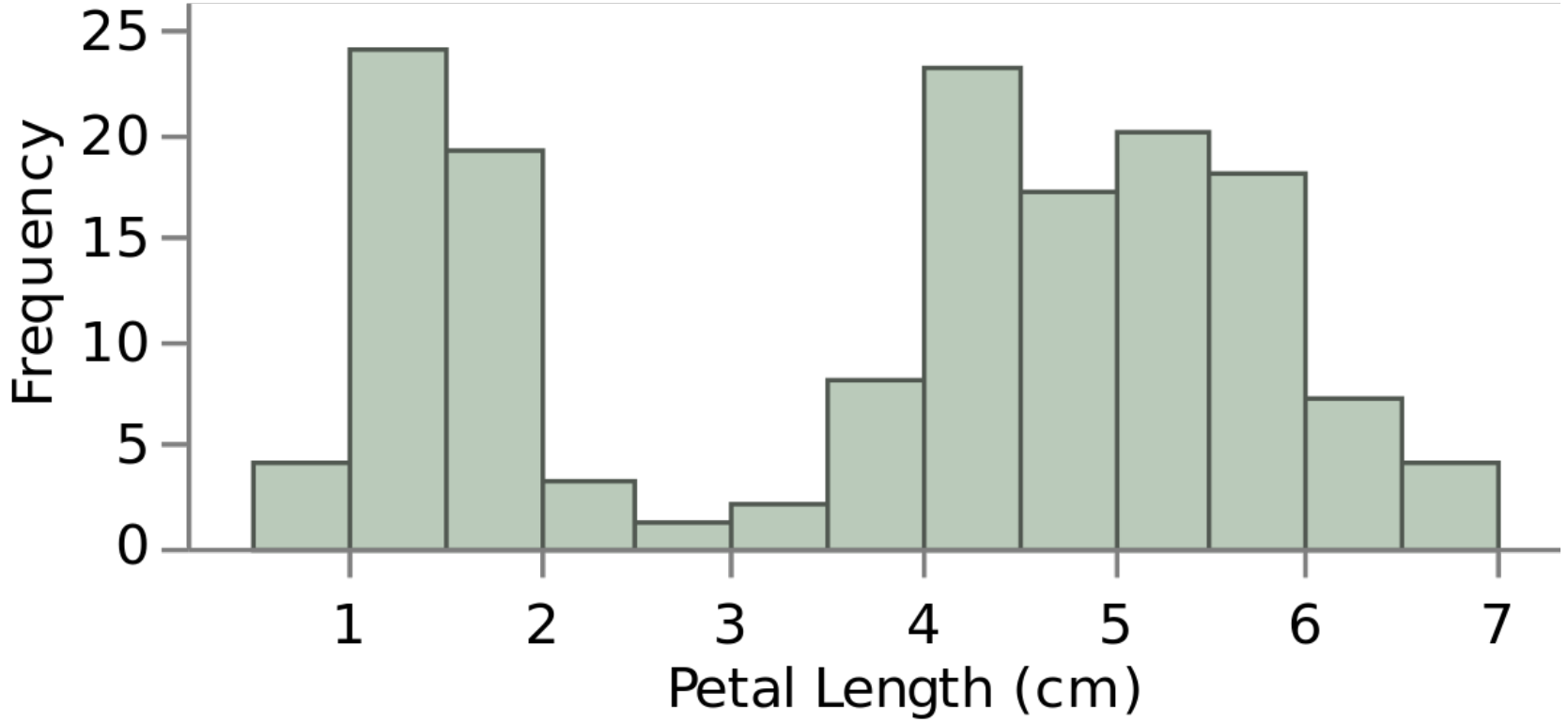
Nem sempre os dados tem distribuição normal!

ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Histograma: gráfico exibe frequência tabulada por meio de barras
  - Mostra proporção de casos que caem em várias categorias ou intervalos
  - Difere de um gráfico de barras pois a área da barra denota o valor e não a altura como em gráfico de barra, distinção importante quando as categorias não tem largura uniforme

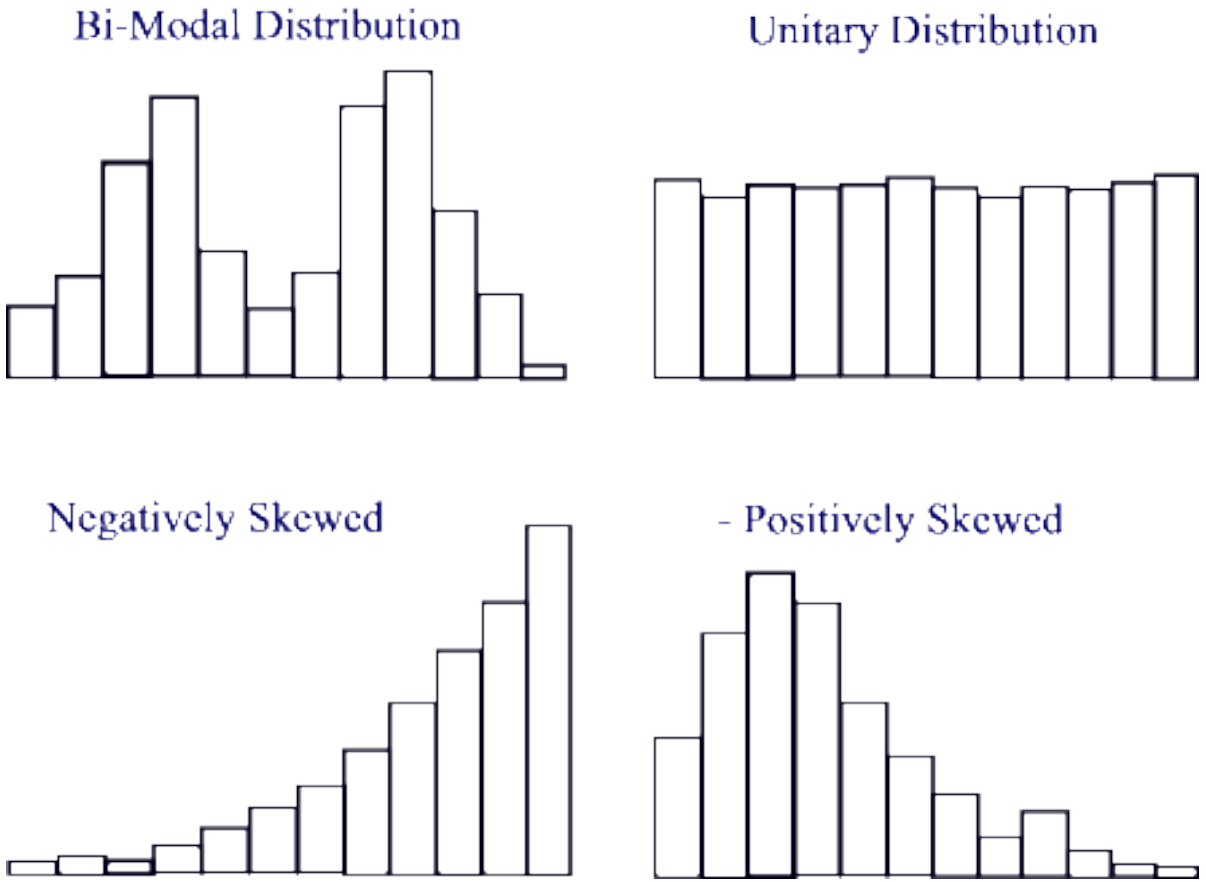
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Histograma



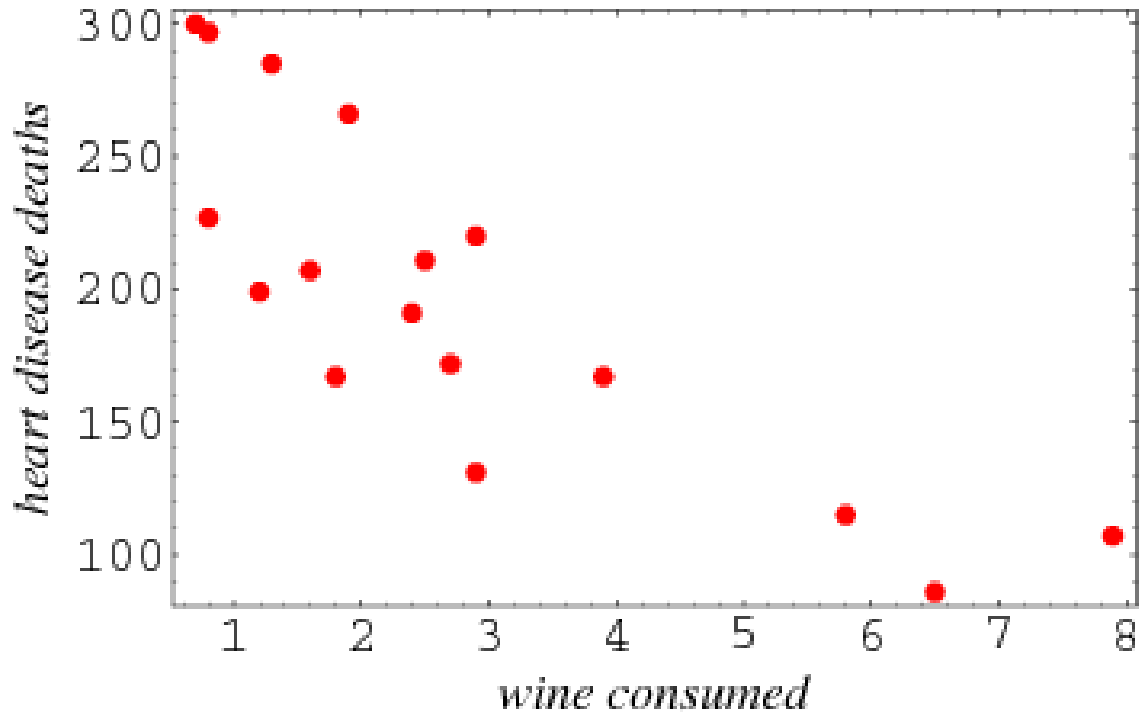
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Histograma:
- Histogramas de distribuições diferentes



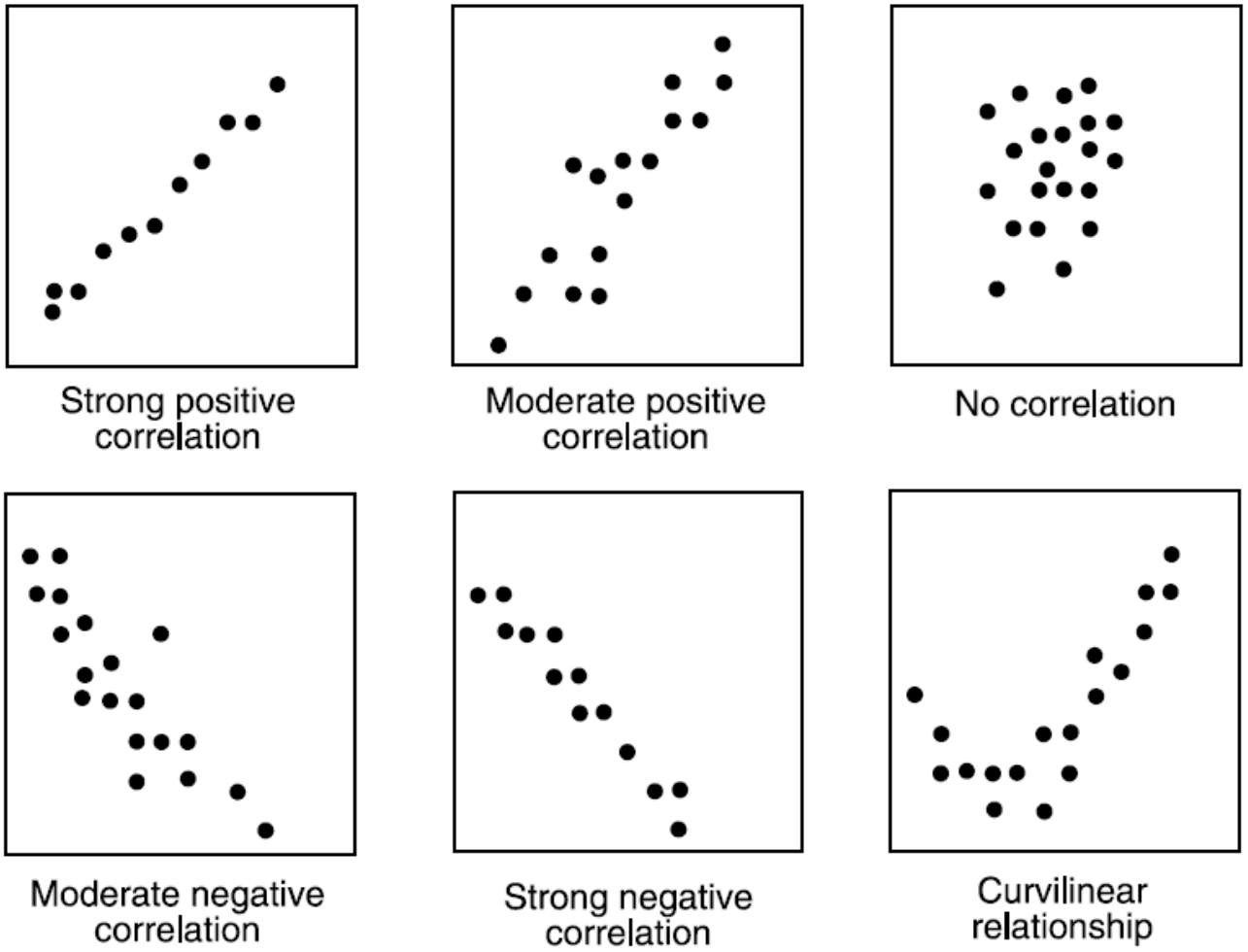
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Scatter plot
  - Permite visualização de correlação entre duas variáveis
  - Cada par de valores é indicado como um ponto em um plano



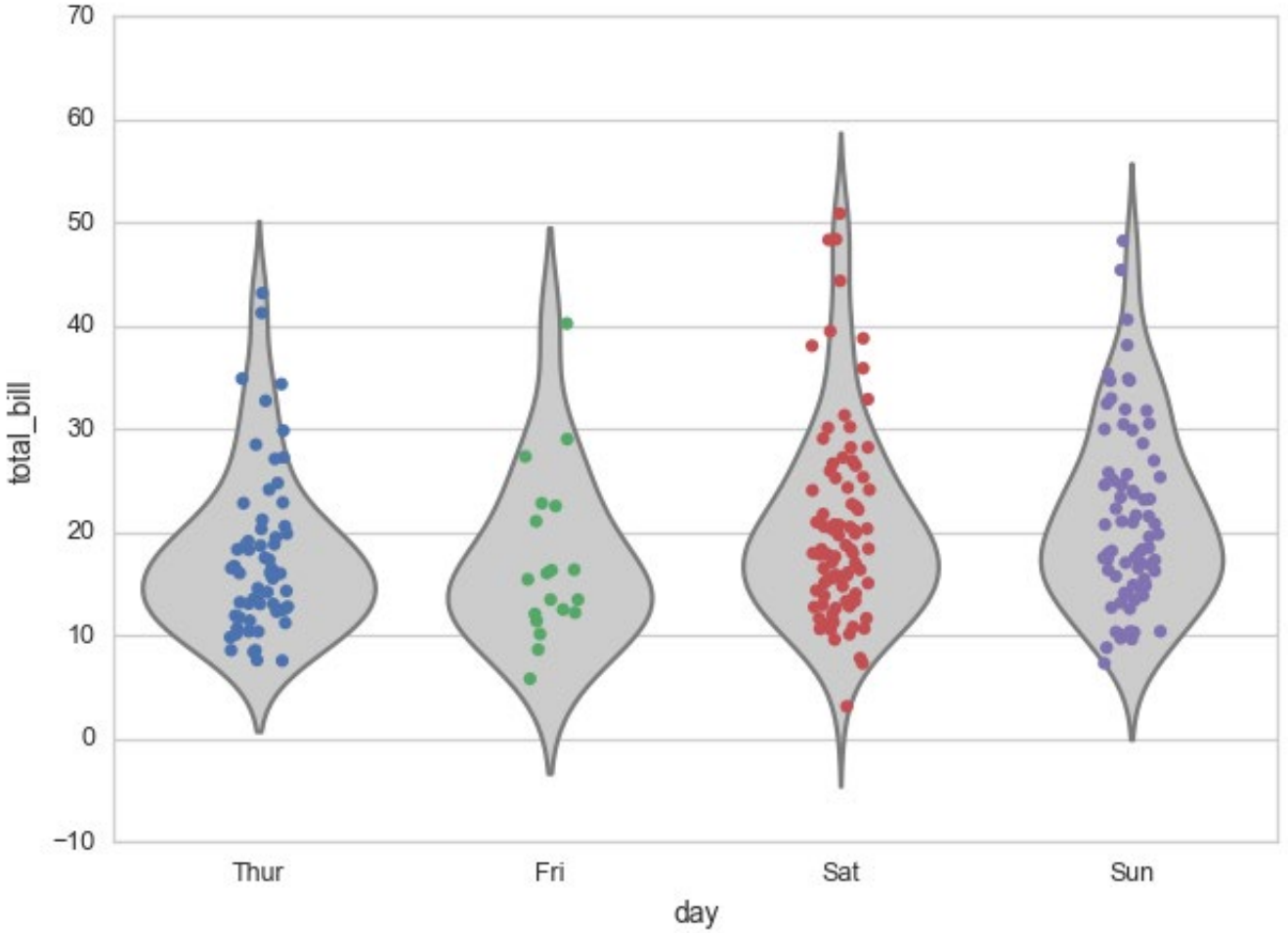
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Scatter plot



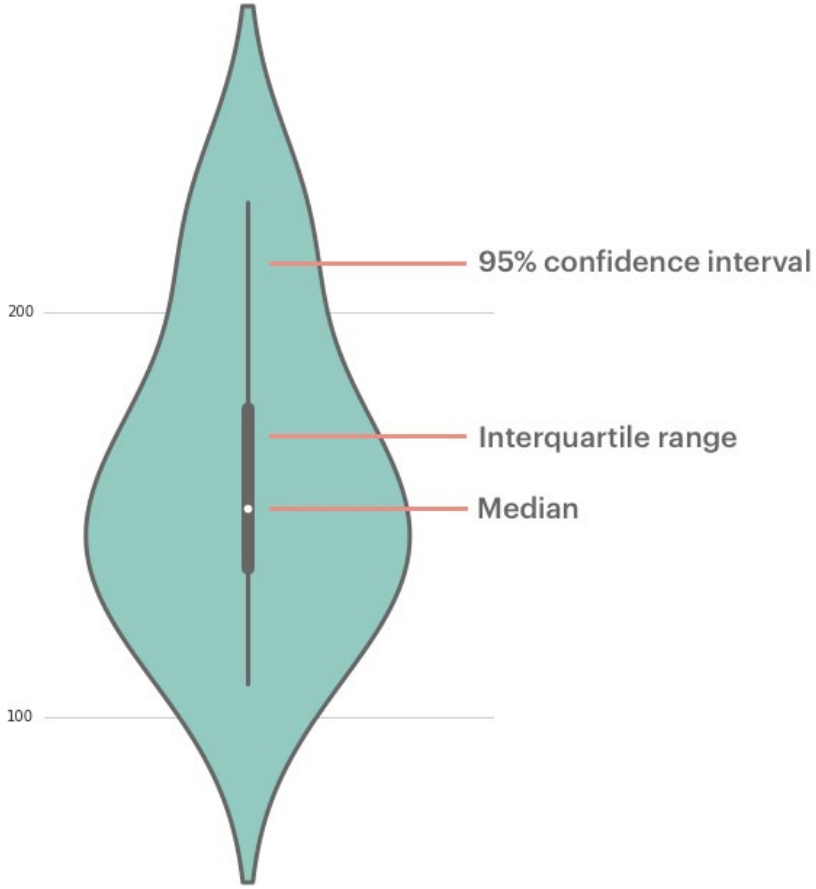
ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Violin plot

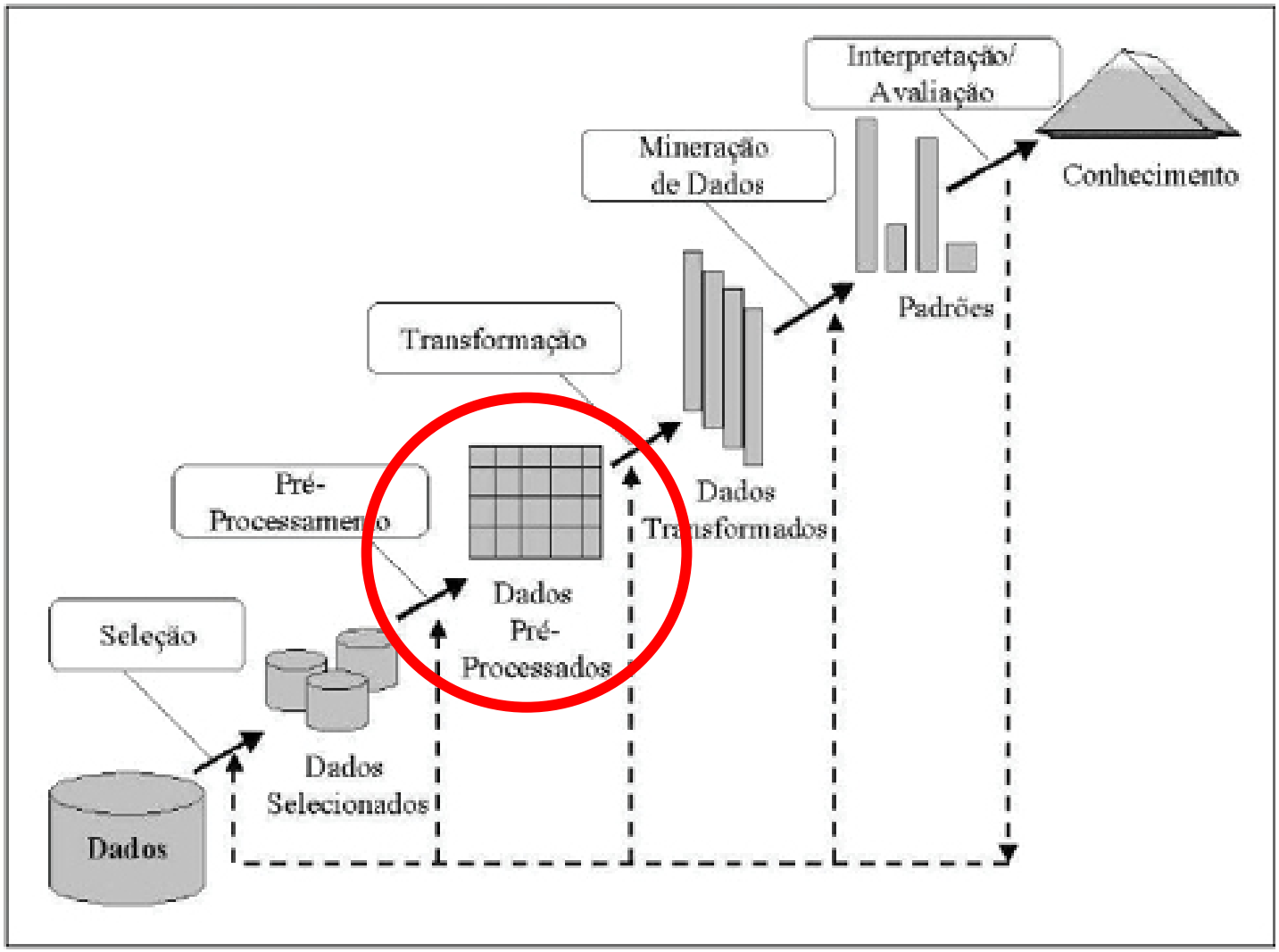


ESTATÍSTICA BÁSICA DOS DADOS – DISPERSÃO

- Violin plot



PRÉ-PROCESSAMENTO



PRÉ-PROCESSAMENTO

- Aplicação de várias técnicas para captação, organização, tratamento e a preparação dos dados.

- É uma etapa que possui fundamental relevância no processo de KDD.

- Compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados.

PRÉ-PROCESSAMENTO

- Qualidade dos dados
  - Limpeza dos dados
- Preparação dos dados
  - Integrar dados e atributos
  - Reduzir dados
  - Transformar dados e atributos

PRÉ-PROCESSAMENTO

- Que tipos de problemas podemos ter com os dados?
- Como detectar estes problemas?
- Como tratar estes problemas?

QUALIDADE DE DADOS

- Medidas de Qualidade
  - Acurácia: corretos ou errados
  - Completude: não registrado, não disponível
  - Consistência Consistência: alguns modificados modificados, outros não, sem referência, padrões diferentes, ...
  - Temporalidade: atualizados no tempo correto?
  - Credibilidade: quão confiáveis são os dados?
  - Interpretabilidade: quão fácil é o entendimento dos dados?

QUALIDADE DE DADOS

- Problemas com dados
  - Ruído
  - Outlier
  - Duplicações
  - Inconsistências
  - Valores ausentes
  - Limpeza dos Dados (Data Cleaning)!

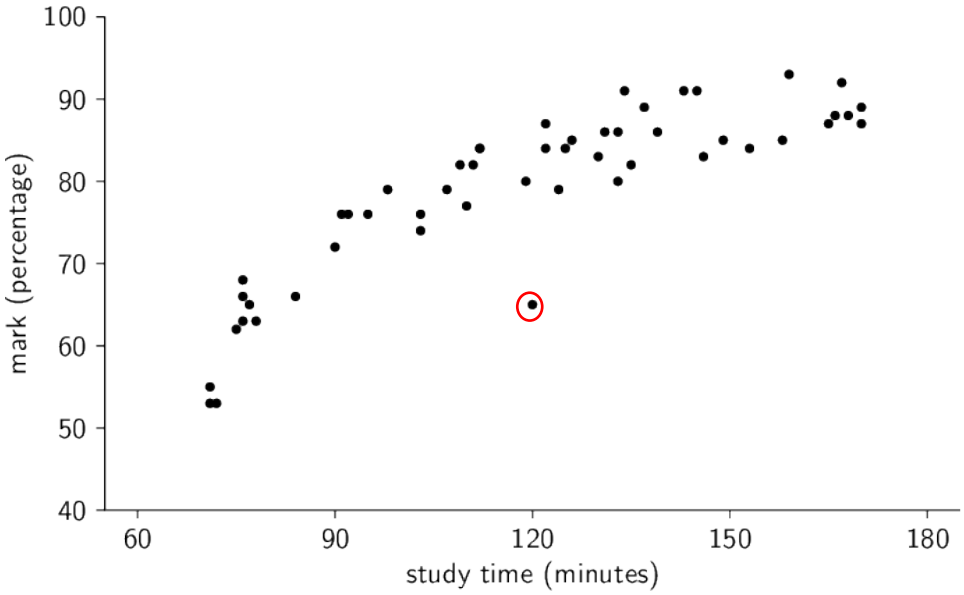
QUALIDADE DE DADOS

- Ruído: mudança nos valores originais
  - gerado na captura, armazenamento, transmissão, processamento, conversão
  - ex.: barulho barulho de fundo captado captado em conjunto conjunto com o áudio de uma voz, sensibilidade do sensor de luz na captura de uma imagem fotográfica
  - pode ser feita filtragem espectral ou suavização dos dados

QUALIDADE DE DADOS

- Outliers: instâncias de dados com características consideravelmente diferentes da maioria das outras instâncias

- Identificar ou remover outliers



QUALIDADE DE DADOS

- Valores ausentes
  - informação não coletada (ex.: pessoa se recusou a oferecer)
  - atributos não aplicáveis a todos casos (ex.: renda não é aplicável a crianças)

- Lidando com valores ausentes
  - Eliminar instâncias de dados
  - Corrigir manualmente
  - Utilizar um valor de tendência central
  - Estimar valores ausentes com base nos demais
  - Ignorar valores ausentes na análise

QUALIDADE DE DADOS

- Base de dados pode possuir instâncias duplicadas, ou quase duplicadas
  - Pode ocorrer na mesclagem de dados de fontes heterogêneas
  - Ex.: mesma pessoa com vários endereços de email
- Dados incompletos ou inconsistentes
  - Identificar, preencher, estimar, corrigir

PRÉ-PROCESSAMENTO

- Agregação
- Amostragem
- Redução de dimensionalidade
- Redução de numerosidade
- Discretização e Binarização
- Transformações de atributos

PRÉ-PROCESSAMENTO

- Agregação
- Combinar dois ou mais atributos (ou instâncias) em um único atributo (ou instância)
- Motivação
  - Redução de dados
    - Reduzir o número de atributos ou instâncias
    - ex.: vendas mensais agregadas em vendas anuais
  - Mudança de escala
    - Cidade agregadas em regiões, estados, países, etc
  - Dados mais 'estáveis'
    - Dados agregados podem ter menor variabilidade
    - ex.: temperatura do dia vs temperatura do mês

PRÉ-PROCESSAMENTO

- Amostragem
  - Obter um conjunto de amostra menor que o conjunto de dados original
  - Escolha um subconjunto representativo
  - Cuidado com o tamanho da amostra
  - Pode reduzir tempo de processamento e espaço de memória
  - Poder ser usado para testes iniciais antes de execução na base completa

PRÉ-PROCESSAMENTO

- Amostragem
  - Aleatório simples
    - Cada instância tem mesma probabilidade
  - Sem reposição
    - Instância selecionada não pode mais ser escolhida
  - Com reposição
    - Instância selecionada poder ser escolhida de novo
  - Amostragem estratificada
    - Dados são particionados, instâncias são sorteadas de cada partição (proporcionalmente)

PRÉ-PROCESSAMENTO

- Amostragem

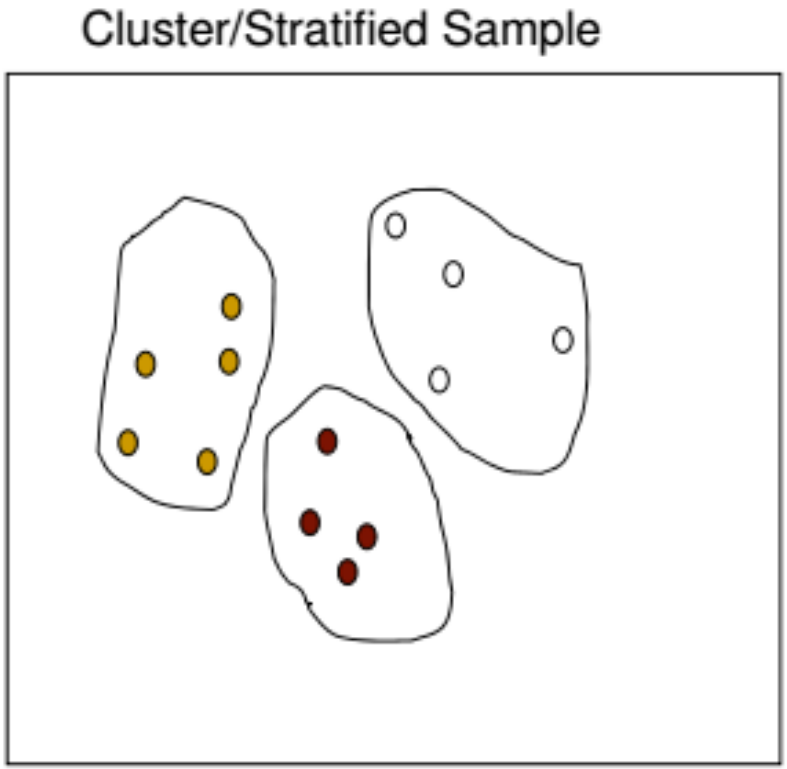
8000 points

2000 Points

500 Points

PRÉ-PROCESSAMENTO

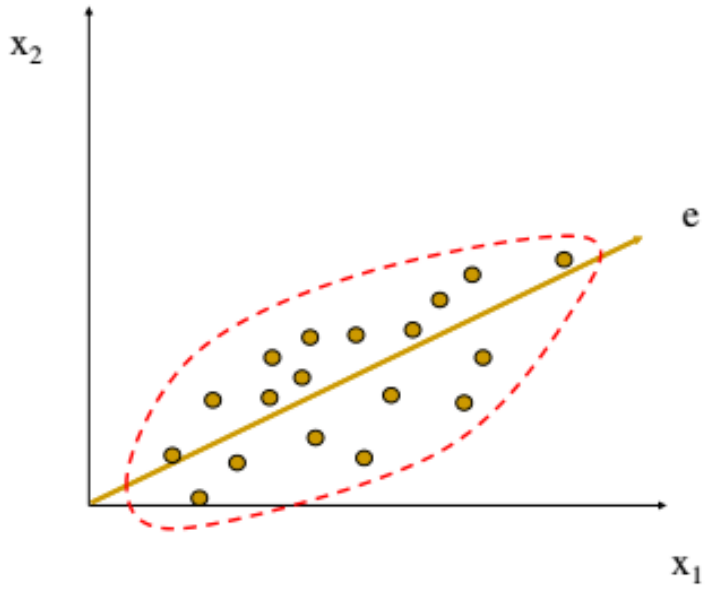
- Amostragem



PRÉ-PROCESSAMENTO

- Maldição da dimensionalidade: aumento do número de dimensões deixa dados mais esparsos
- Redução de dimensionalidade
  - tenta eliminar eliminar atributos atributos irrelevantes irrelevantes ou reduzir reduzir ruído
  - reduz tempo e memória para mineração de dados
  - facilita visualização
- Técnicas
  - Principal Components Analysis
  - Singular Vector Decomposition
  - Feature subset selection (subconjunto de atributos)

- Principal Component Analysis (PCA)
  - Ache uma projeção que captura a maior parte da variação dos dados
  - Dados originais são projetados em um espaço menor, composto pelos autovetores da matriz de co-variância



PRÉ-PROCESSAMENTO

- Seleção de subconjunto de atributos
  - Podem haver atributos redundantes ou irrelevantes
  - Para n atributos, existem 2n Para n atributos, existem 2 subconjuntos n subconjuntos
  - Podemos tentar busca exaustiva pelo melhor subconjunto
    - Pode-se avaliar por teste de significância estatística, por ganho de informação, executando a tarefa de mineração,...
  - Mas busca exaustiva pode ser proibitiva

PRÉ-PROCESSAMENTO

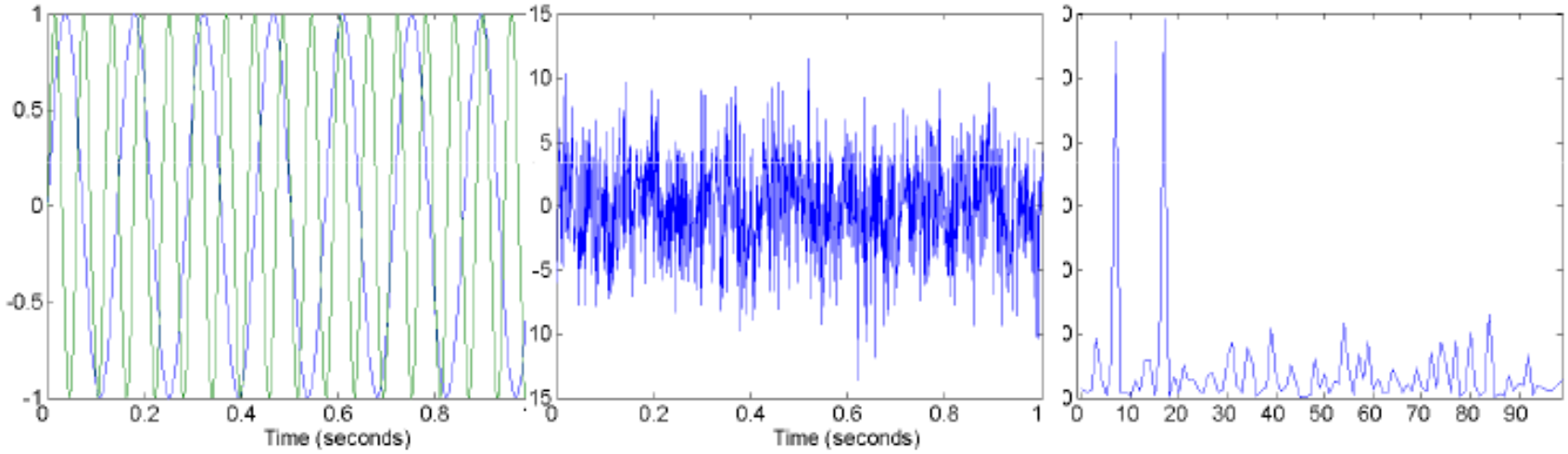
- Seleção de subconjunto de atributos
  - Uso de heurísticas
    - Tipicamente uso de algoritmo guloso: busca localmente a melhor solução
    - Passo-a-passo para frente: começa com subconjunto vazio e adiciona o melhor atributo a cada passo
    - Passo-a-passo para trás: começa com todos atributos e elimina o pior a cada passo
    - Combinação de adicionar e eliminar
    - Árvore de decisão: atributos mais relevantes para particionamento são nós da árvore

PRÉ-PROCESSAMENTO

- Criação de atributos
  - Criar novos atributos que capturam a informação importante nos dados de forma mais eficiente que os originais
  - Feature extraction dependente de domínio
    - ex.: extração de descritores de imagens
  - Mapeamento de dados em outro espaço
    - transformada de Fourier
  - Feature construction
    - combinar dados existentes, dependente de domínio

PRÉ-PROCESSAMENTO

- Criação de atributos



Two Sine Waves

Two Sine Waves + Noise

Frequency

PRÉ-PROCESSAMENTO

- Redução de numerosidade
  - Reduza volume de dados escolhendo uma representação de dados alternativa, menor
  - Métodos paramétricos (ex. regressão)
    - Ajuste os dados a um modelo e guarde os parâmetros do modelo, descartando os dados
  - Métodos não paramétricos
    - Não assuma modelos
    - histogramas, agrupamento, amostragem,…

PRÉ-PROCESSAMENTO

- Redução de numerosidade
  - Histograma: divida os dados em intervalos e guarde somente a média de cada
  - Agrupamento: descoberta de grupos com similaridade interna e dissimilaridade externa
    - Guarde os centróides (protótipos) dos grupos

PRÉ-PROCESSAMENTO

- Discretização
  - Transformação de atributos numéricos em categóricos (ordinais ou nominais)
  - Não supervisionado (sem uso da classe): intervalos iguais, frequências iguais, agrupamento
  - Supervisionado: maximizar pureza de classe nos intervalos, ex: minimizar entropia
- Binarização
  - Transformação de atributos em um ou mais atributos binários

PRÉ-PROCESSAMENTO

- Transformações de atributos
  - função que mapeia um conjunto de valores de um atributo em um novo conjunto de valores
    - funções simples: xk, log(x), ex, |x|
    - ex: bytes transmitidos de 1 a 109, pode aplicar log(X)
    - normalização ou padronização

INTELIGÊNCIA COMPUTACIONAL

PREPARANDO OS DADOS

FELIPE TORRES