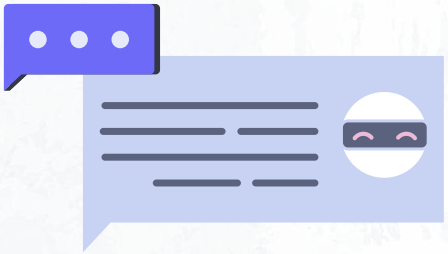
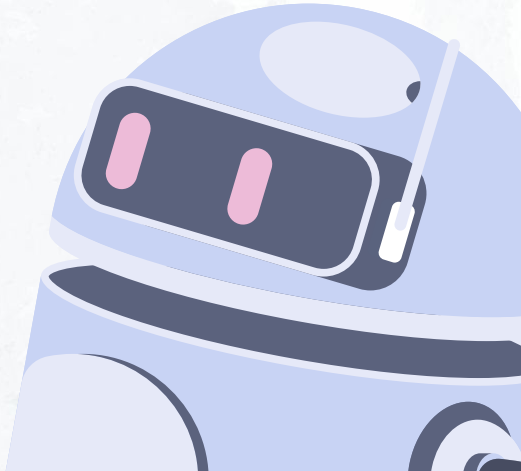
Introdução à Inteligência Artificial

João Paulo Aires

(IA)





Índice

01 →

Caracterização de Dados

(IA)

Dados

Conjuntos de Dados / Matriz de Dados (Dataset)

- **Linhas (*N*)**
  - Instâncias (instances)
  - Objetos (objects)
  - Exemplos (examples)
  - Tuplas (tuples)
  - Amostras (samples)
- **Colunas (*m*)**
  - Atributos (attributes)
  - Características (features)
  - Campos (fields)
  - Variáveis (variables)
  - Dimensões (dimensions)

	$\mathbf{x}_1$	$\mathbf{x}_2$		$\mathbf{x}_m$
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	…	$x_m^{(1)}$
$\mathbf{x}^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	…	$x_m^{(2)}$
			…	
$\mathbf{x}^{(N)}$	$x_1^{(N)}$	$x_2^{(N)}$	…	$x_m^{(N)}$

Dados

Conjuntos de Dados (Dataset)

Categórico (qualitativo)

Numérico (quantitativo)

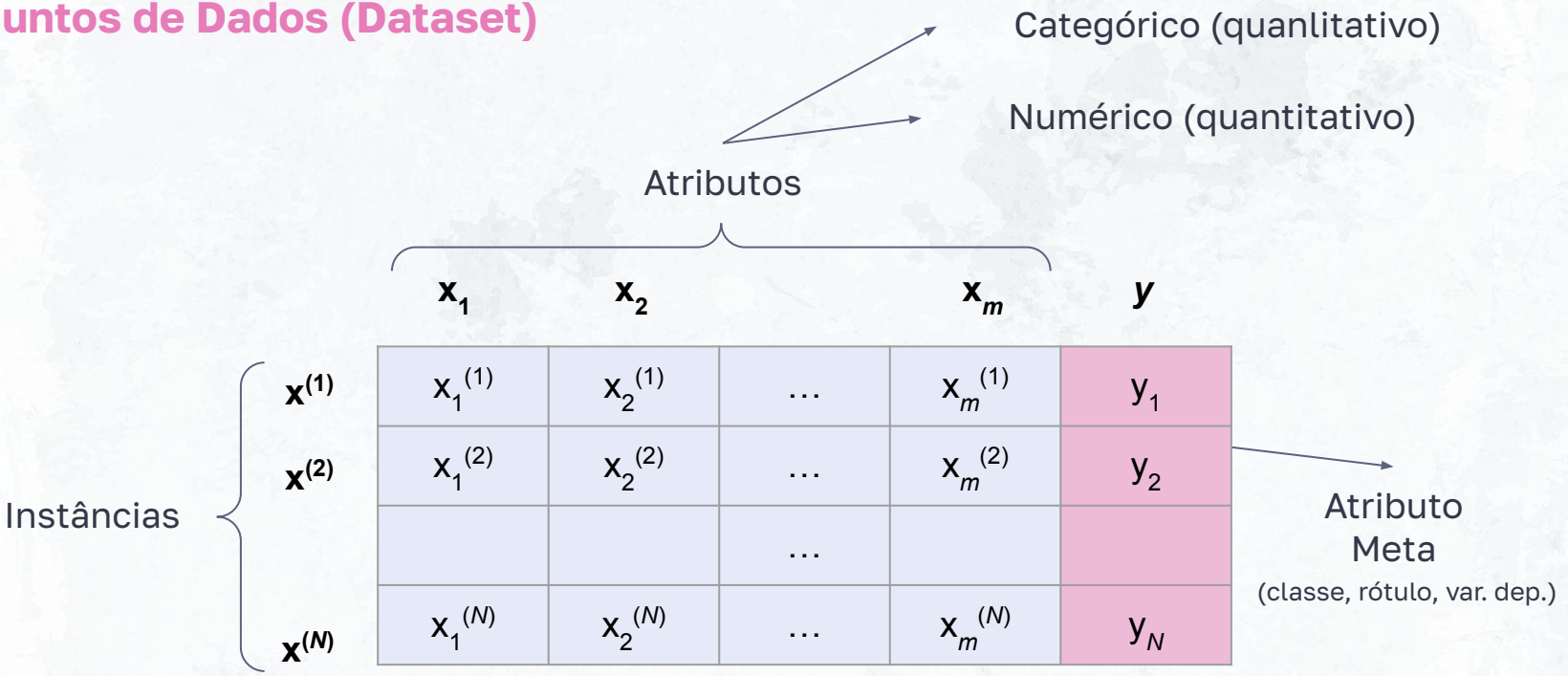
Atributos

Instâncias

	x_1	x_2	...	x_m
$x^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	...	$x_m^{(1)}$
$x^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	...	$x_m^{(2)}$
			...	
$x^{(N)}$	$x_1^{(N)}$	$x_2^{(N)}$	...	$x_m^{(N)}$

Dados

Conjuntos de Dados (Dataset)



Dados

Exemplo: Diagnóstico de uma doença

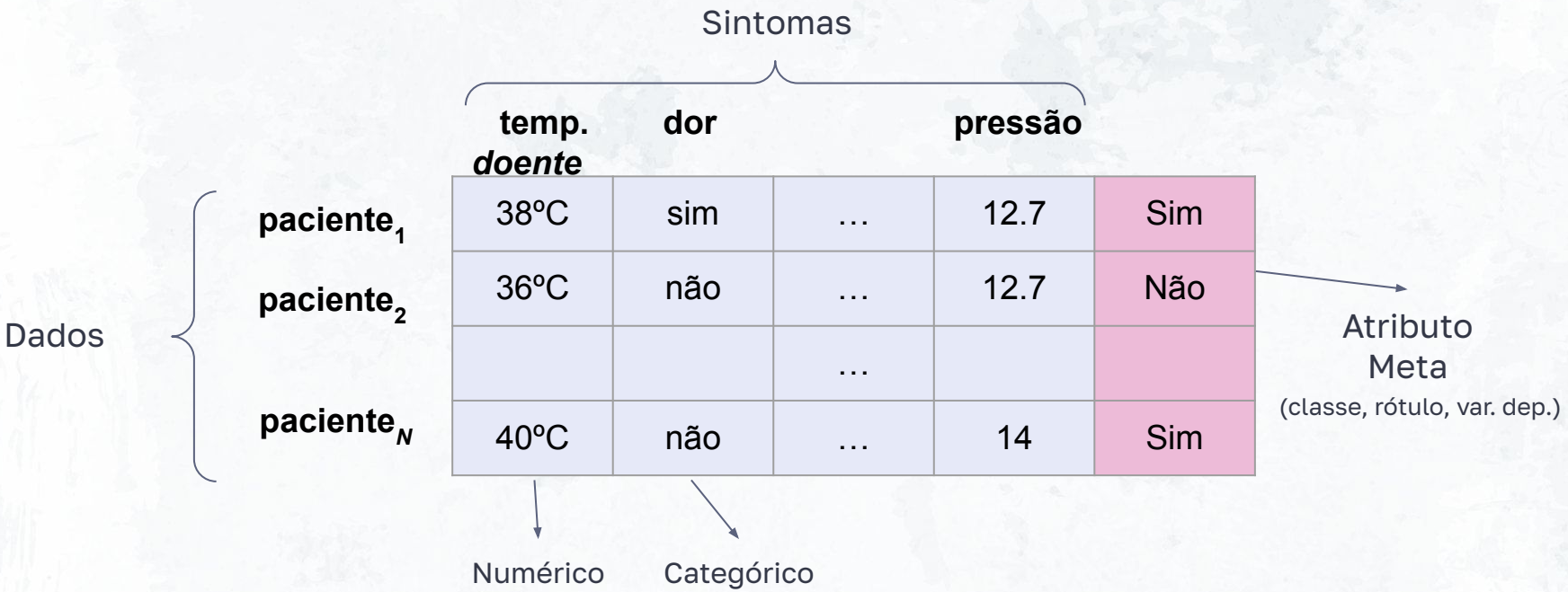
Sintomas

Dados

	temp.	dor	…	pressão
paciente$_1$	38ºC	sim	…	12.7
paciente$_2$	36ºC	não	…	12.7
			…	
paciente$_N$	40ºC	não	…	14

Dados

Exemplo: Diagnóstico de uma doença



Dados

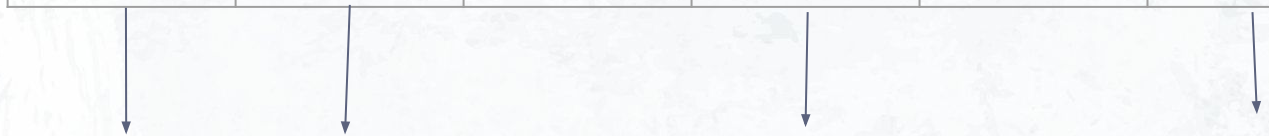
Tipos de Atributos

- **Nominal (qualitativo, categórico)**
  - Ex.: cor, profissão, tipo sanguíneo
- **Ordinal (qualitativo, categórico)**
  - Ex: qualidade (ruim, médio, bom), dias da semana
- **Intervalar (quantitativo, numérico)**
  - Ex: data, temperatura em Célcius
- **Racional (quantitativo, numérico)**
  - Ex: peso, tamanho, idade, temperatura em Kelvin

Dados

Exemplo

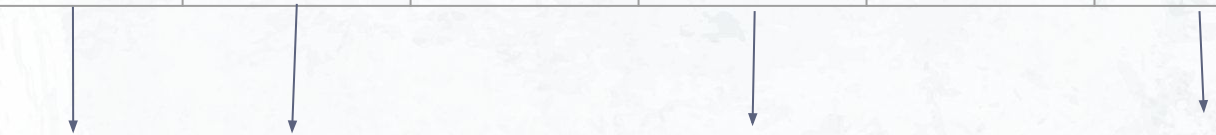
Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável



Dados

Exemplo

Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável

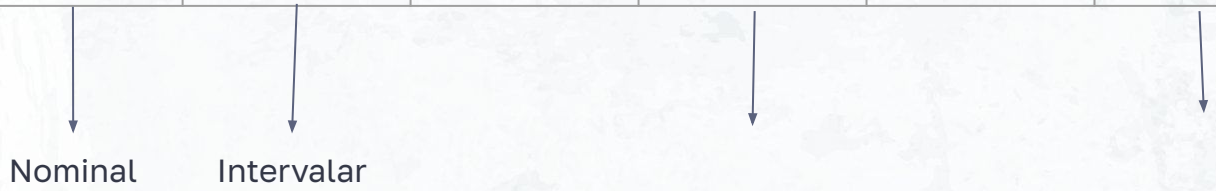


Nominal

Dados

Exemplo

Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável



Dados

Exemplo

Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável

Dados

Exemplo

Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável

Nominal

Intervalar

Ordinal

Racional

Dados

Exercícios

Definir o tipo dos seguintes atributos como nominal, ordinal, intervalar ou racional:

- Tempo (em termos de AM ou PM)
- Brilho (medido por medidor de luz)
- Brilho (medido pelo julgamento humano)
- Bronze, Prata e Ouro (medalhas)
- Número de pacientes em hospital
- Rank militar

Dados

Tipos de Atributos

Uma taxonomia alternativa para atributos pode ser estabelecida pelo número de valores

- Atributo Contínuo
  - Assume uma quantidade incontável de valores
- Atributo Discreto
  - Assume um número contável de valores
    - Finito ou infinito

02 →

Análise Exploratória de Dados

(IA)

Motivação

Exploração preliminar dos dados

- Facilita entendimento de suas características
- Ajuda a selecionar melhor técnica de pré-processamento ou aprendizado
- Faz uso principalmente de:
  - Estatística descritiva
  - Visualização

Estatística Descritiva

Permite capturar

- **Frequência** dos dados
- **Localização** ou tendência central
- **Dispersão** ou espalhamento
- **Distribuição** ou formato

Frequência

Proporção de vezes que um atributo assume um dado valor

- Frequentemente utilizada para análise de atributos categóricos

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Frequência

Proporção de vezes que um atributo assume um dado valor

- Frequentemente utilizada para análise de atributos categóricos

66% das manchas são grandes

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Frequência

Proporção de vezes que um atributo assume um dado valor

- Frequentemente utilizada para análise de atributos categóricos

50% dos pacientes são doentes

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Medidas de Localidade

- Dados Categóricos
  - Moda
- Dados numéricos
  - Média
  - Mediana
  - Percentil

Exemplo de Moda

- Moda para o atributo mancha: **grande**

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Média

- **Pode ser calculada facilmente:**

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Sensível a *outliers*

Mediana

- Menos sensível a *outliers* que a média
- Necessário ordenar valores

$$mediana(x_j) = \begin{cases} x_j^{r+1} & n\%2! = 0 \\ \frac{1}{2}(x_j^{(r)} + x_j^{(r+1)}) & n\%2 = 0 \end{cases}$$

Média Podada

- *Trimmed mean*
- Minimiza problema da média descartando valores extremos
  - Dados são ordenados
  - Porcentagem *p* de valores são eliminados de cada extremidade

```
import numpy as np
from scipy.stats import trim_mean

x = np.array([1, 2, 3, 4, 5])

# Eliminar 20% (0.2) dos valores em cada extremidade
media_podada = trim_mean(x, 0.2)

print(f'Média podada: {media_podada}')
~
```

Média podada: 3

Exercício

- Dado um atributo $\mathbf{x}_j$ = [1, 2, 3, 4, 5, 80], calcule:
  - Média
  - Mediana
  - Média podada com p = 16.5%

03 →

Pré-Processamento de Dados

(IA)

Pré-Processamento de Dados

Transformação de Dados

- Conversão de valores **simbólicos** para **numéricos;**
- Conversão de valores **numéricos** para **simbólicos;**
- **Normalização** de valores numéricos

Pré-Processamento de Dados

Conversão Categórico → Numérico

- **Muitos algoritmos** de AM trabalham **apenas com variáveis numéricas**
  - Redes Neurais, SVMs, etc.
  - Variáveis categóricas precisam ser **convertidas**
- Conversão depende da **existência de ordem**
  - Variáveis são nominais ou ordinais?

Pré-Processamento de Dados

Conversão de valores Ordinais

- Para variáveis ordinais, a **ordem** dos valores deve ser mantida de alguma maneira
  - Estratégia comum: associar valores **inteiros crescentes**
    - Ex: {frio, morno, quente} = {1, 2, 3}
  - Tal estratégia pode inserir distorções relativas entre os conceitos (qualquer política de pesos também insere!)
    - Diferenças entre símbolos são subjetivas.

Pré-Processamento de Dados

Conversão de Valores Nominais

- Atributos **nominais**
  - Conversão é feita por **binarização**
  - Codificação mais usual:
    - Codificação **1-de-n (canônica, one-hot)**

Pré-Processamento de Dados

Conversão de Valores Nominais

- Codificação **1-de-n** (*one-hot encoding*)
  - Um atributo binário associado a cada valor nominal
  - Exemplo:
    - Codificar {amarelo, vermelho, verde, azul laranja, branco}
      - 100000 - amarelo
      - 010000 - vermelho
      - 001000 - verde
      - 000100 - azul
      - 000010 - laranja
      - 000001 - branco

Pré-Processamento de Dados

Atributos Nominais com Muitos Valores

- Codificação **1-de-n** pode levar a dados muito esparsos quando **n** é grande
- Solução pode estar no uso de **conhecimento de domínio** do problema em questão
- Ex: atributo = nome de país
  - Existem 193 países membros da ONU
    - Codificação 1-de-n demandaria 192 atributos adicionais
    - Dados esparsos / maldição da dimensionalidade!

Pré-Processamento de Dados

Atributos Nominais com Muitos Valores

- Ex: atributo = nome de país
  - Possível solução:
    - Utilizar 1 atributo nominal com apenas 7 valores (continentes)
    - Tentar discriminar entre os países com um conjunto menor de **pseudo-atributos** numéricos
      - PIB, população, IDH, temperatura média, ...
    - Funcionamento satisfatório depende da aplicação
    - Não existe abordagem de pré-processamento sempre melhor ou pior
      - **No free-lunch**

Pré-Processamento de Dados

Discretização

- Alguns algoritmos de AM aceitam apenas valores categóricos
  - Valor numérico precisa ser **discretizado em intervalos**
- Melhor discretização depende de:
  - Algoritmo que utilizará os valores discretizados
  - Demais atributos, ...
- Em geral, é realizada a priori, como pré-processamento

Pré-Processamento de Dados

Discretização

- Transformar valores contínuos em intervalos
  - Atributo se transforma em **categórico ordinal**
- Passos necessários
  - Definição do número de intervalos (categorias)
    - Geralmente ad-hoc (feito pelo usuário)
  - Definição de como mapear os valores contínuos para as novas categorias
    - Definir limites/tamanho dos intervalos
    - Geralmente feito pelo algoritmo

Pré-Processamento de Dados

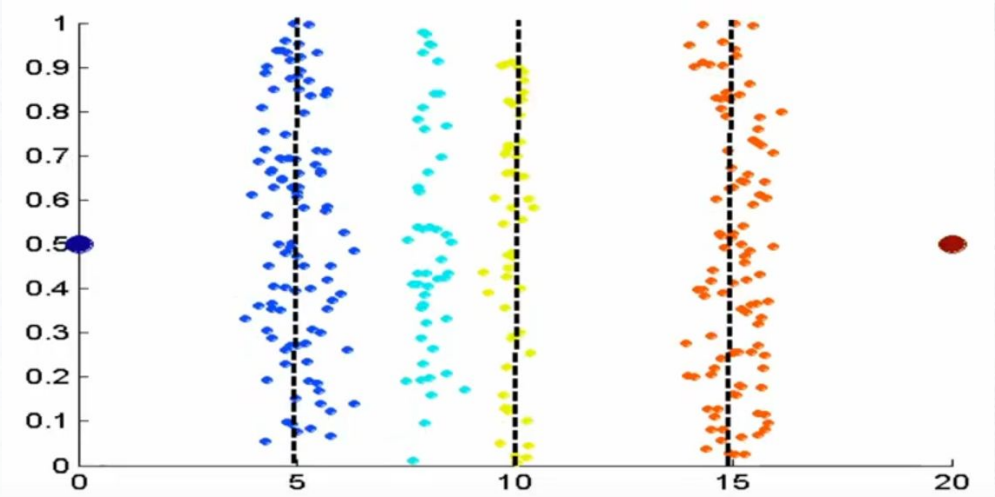
Discretização Não-Supervisionada

- Algoritmo Simples
  - **Larguras Iguais**
    - Divide intervalo original de valores em *n* sub-intervalos com mesma largura
    - Simples de implementar, porém:
      - Assume que valores possuem distribuição uniforme
      - Muito ineficaz em distribuições não uniformes
      - Muito sensível à presença de *outliers*

Pré-Processamento de Dados

Discretização Não-Supervisionada

- Larguras iguais (n = 4)



Pré-Processamento de Dados

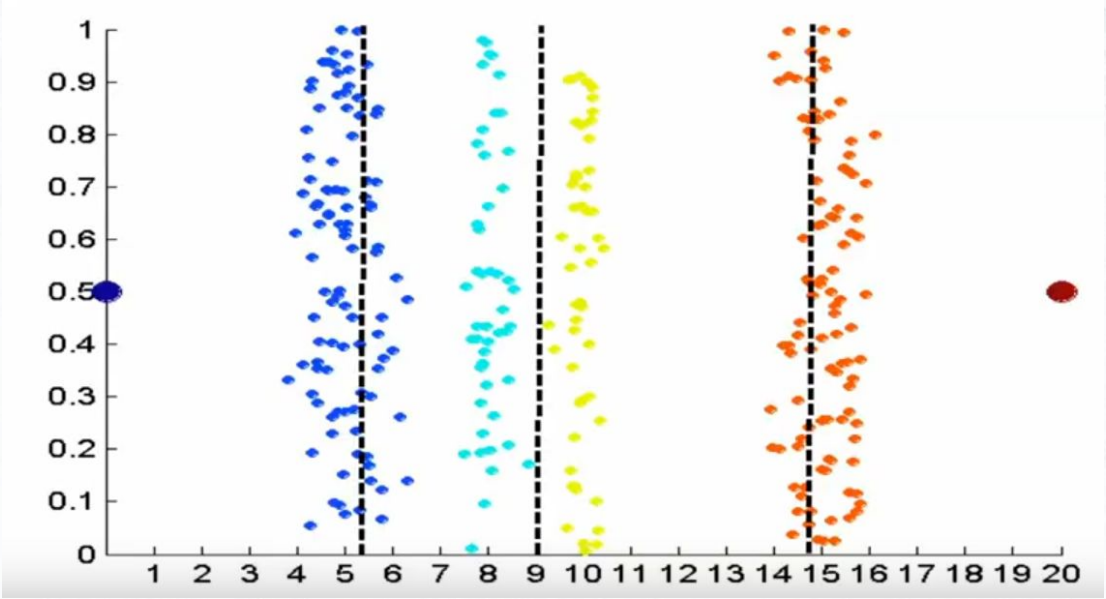
Discretização Não-Supervisionada

- Algoritmo Simples
  - **Frequências Iguais**
    - Atribui o mesmo número de instâncias por intervalo
    - Simples de implementar, porém:
      - Assume que valores estão em grupos balanceados
      - Muito ineficaz em distribuições desbalanceadas

Pré-Processamento de Dados

Discretização Não-Supervisionada

- Frequências iguais (n = 4)



Pré-Processamento de Dados

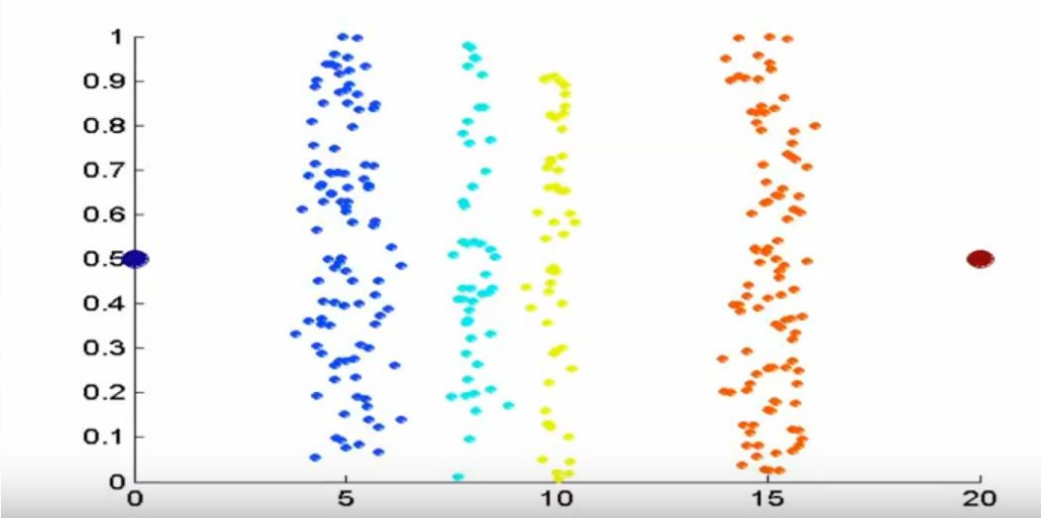
Discretização Não-Supervisionada

- Algoritmo Simples
  - **Inspeção Visual**
    - Observar gráfico com os valores do atributo e determinar visualmente os intervalos de acordo com a distribuição natural dos dados
    - Simples, eficaz e permite determinação eficiente de n
    - Porém:
      - Pré-processamento manual (*time consuming*)

Pré-Processamento de Dados

Discretização Não-Supervisionada

- Inspeção Visual



Exercício

Discretização atributo numérico

- Em 3 intervalos

$$x_j = [0, 1, 2, 6, 6, 9, 10, 10, 10, 13, 18, 20, 21, 21, 25]$$

- Usar:
  - Larguras iguais
  - Frequências iguais
  - Inspeção visual

Normalização

- **Transformação** aplicada aos dados de forma com que estes exibam **propriedades em comum**;
- Normalizações mais usuais em Ciência de Dados são as **lineares**
  - Re-escalar
  - Padronizar

Normalização

Re-Escalar

- Re-escalar os valores de um atributo:
  - Adicionar ou subtrair uma constante
  - Multiplicar ou dividir por uma constante
- utilizada para mudar unidade de medida dos dados;
- Uso mais comum é para converter valores de atributos para os intervalos [0, 1] ou [-1, +1]
  - Ex: converter valores para intervalo [novoMin, novoMax]

Normalização

Re-Escalar

$$x' = \frac{(x - min(x)) \times (R))}{(max(x) - min(x))} + novoMin$$

$$R = (novoMax - novoMin)$$

Normalização

Re-Escalar

- Muito utilizada em algoritmos baseados em **otimização**
  - Ex: redes neurais e SVM
  - Principalmente para evitar problemas numéricos oriundos da estratégia de otimização
- Problema: extremamente influenciada por ***outliers***
  - Deve-se evitar re-escalar em aplicações sujeitas a outliers e/ou ruídos

Normalização

Padronizar

- Padronizar os valores de um atributo
  - Adicionar/subtrair uma medida de localização;
  - Multiplicar/dividir por uma média de escala
- Para atributos com distribuição Gaussiana
  - Subtrair cada valor da média (μ)
  - Dividir pelo desvio padrão (σ)
  - Resultado: distribuição normal padrão: $N(0, 1)$
    - Chamada de normalização **z-score**

Normalização

Padronizar

- Normalização **z-score**:

$$x' = \frac{(x - \mu_x)}{(\sigma_x)}$$

- Muito utilizado para pré-processar dados de algoritmos de **agrupamento** (e demais tarefas que exijam **cálculos de distância**)

Normalização

Exemplo

- Considere um dataset com atributos **idade** e **salário**
  - Diferenças em salário são bem maiores que diferenças em idade
    - Influencia algoritmos de AM que se utilizam de informações sobre diferenças (ex: distância Euclidiana)
    - Tal influência serve como “peso”
      - Salário está tendo mais importância que idade
      - Se isso não é desejável, padronizar!

Exercício

Converter os dados abaixo para valores numéricos no intervalo [0, 1]

Febre	Enjoo	Mancha	Dor	Diagnóstico
baixa	sim	pequena	A	doente
média	não	média	C	saudável
alta	sim	grande	B	saudável
alta	não	pequena	A	doente
baixa	não	grande	D	saudável
média	não	sem	C	doente

Obs: não existe relação de ordem entre os tipos de dor