



Pré-Processamento de Dados

Luciano Barbosa



Motivação

- “Todo mundo faz mas ninguém fala”
- Boa parte do tempo gasto nesta etapa
- Dados incorretos podem distorcer os resultados das análises e modelos
- “Better data beats fancier algorithms”





Roteiro

- Ajuste de tipos
- Dados ausentes
- Discretização
- Normalização



Dados Ausentes

- Viés na ausência de dados
 - Ex: pessoas mais ricas tendem a não dizer renda em pesquisas



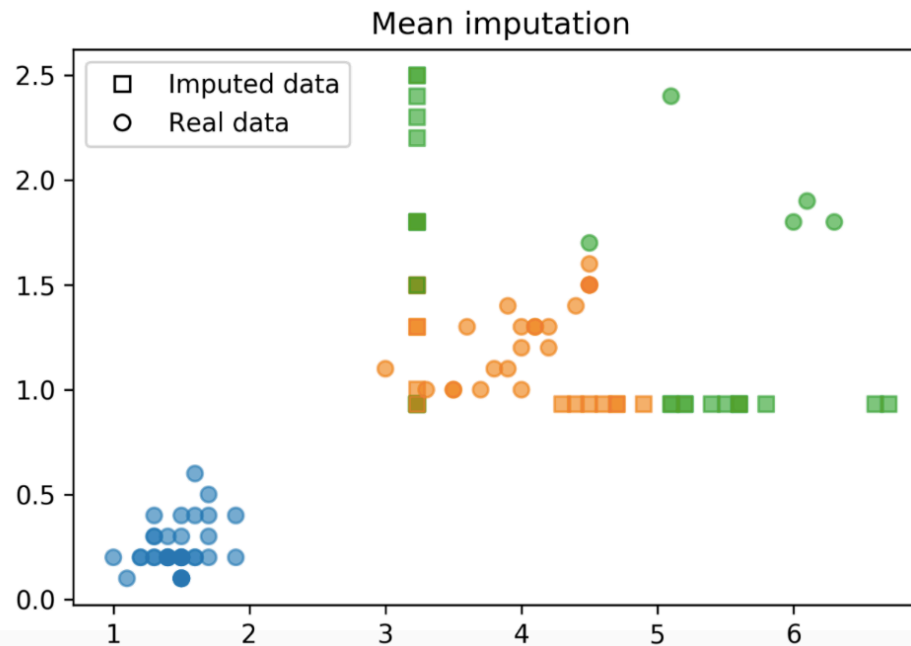
Remoção de Dados

- Remoção das instâncias: instâncias nas quais ao menos um dos atributos está faltando
 - Pro: simples
 - Con:
 - Possível viés se as instâncias removidas forem diferentes das não removidas
 - Pode remover boa parte dos dados
- Remoção de instâncias baseado na análise
- Remoção de variáveis com muitos valores ausentes



Imputação de Dados

- Imputar dados ausentes
- Média, mediana ou moda da variável
 - Problema se existem muitos valores ausentes
 - Pode atrapalhar no cálculo de correlações entre variáveis





Formas de Imputação

- Imputar um valor e criar outra variável indicando que o dado foi imputado
- Para dados categóricos: adicionar uma categoria de dados ausentes
- Amostra aleatória dos valores da variável ou baseado em uma distribuição (ex.: normal)



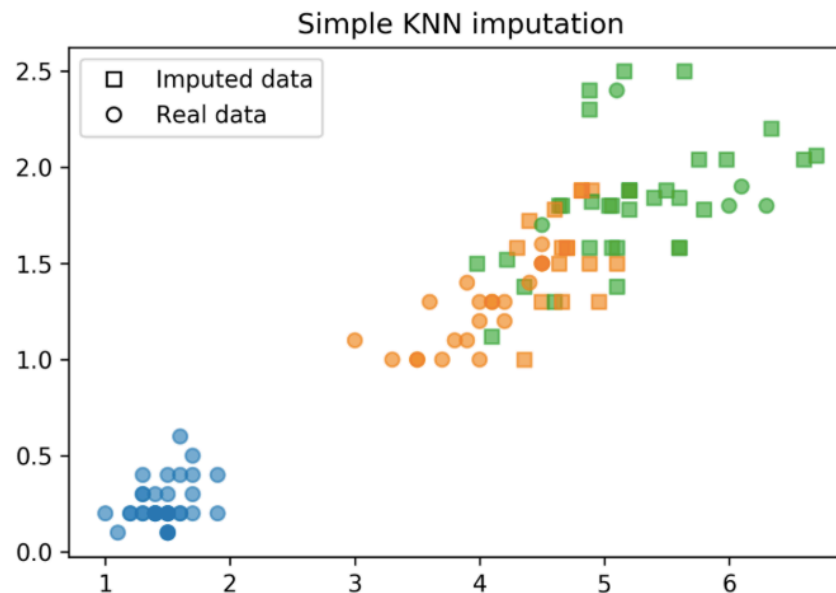
Formas de Imputação

- Para atributos categóricos:
 - Usar classificação para prever os valores ausentes de uma variável a partir das outras
- Para atributos numéricos:
 - Usar regressão para prever os valores ausentes de uma variável a partir das outras



Formas de Imputação

- KNN (<https://github.com/iskandr/fancyimpute>):
 - Valor ausente definido pelos valores das instâncias mais próximas
 - Valor: média dos valores dos vizinhos





Discretização

- Transformar atributos contínuos em categóricos
- Decidir número de intervalos e seu tamanho
- Dois tipos
 - Supervisionada
 - Baseada em entropia: intervalos com mesma proporção de um rótulo
 - Não-supervisionada
 - Intervalos de mesma largura
 - Intervalos de mesma frequência



Supervisionada

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

- Passo 1: Calcular a entropia da classe

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

O-Ring Failure	
Y	N
7	17

$$E(\text{Failure}) = E(7, 17) = E(0.29, .71) = -0.29 \times \log_2(0.29) - 0.71 \times \log_2(0.71) = \mathbf{0.871}$$



Supervisionada

O-Ring Failure Temperature

Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

- Passo 2: Calcular entropia para a classe dada uma discretização

$$E(S|A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

		O-Ring Failure	
		Y	N
Temperature	<= 60	3	0
	> 60	4	17

$$E(\text{Failure}|\text{Temperature}) = P(<=60) \times E(3,0) + P(>60) \times E(4,17) = 3/24 \times 0 + 21/24 \times 0.7 = \mathbf{0.615}$$



Supervisionada

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

- Passo 3: Calcular o Information Gain

$$\text{Information Gain} = E(S) - E(S|A)$$

Information Gain (Failure, Temperature) = **0.256**

Gain = **0.256**

		O-Ring Failure	
		Y	N
Temperature	≤ 60	3	0
	> 60	4	17

Gain = **0.101**

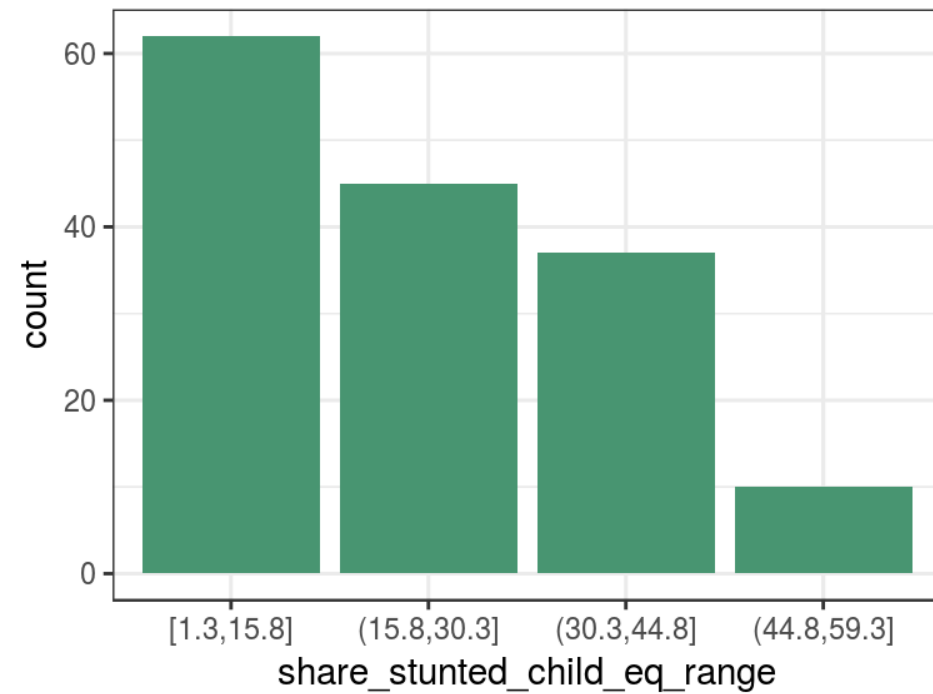
		O-Ring Failure	
		Y	N
Temperature	≤ 70	6	8
	> 70	1	9

Gain = **0.148**

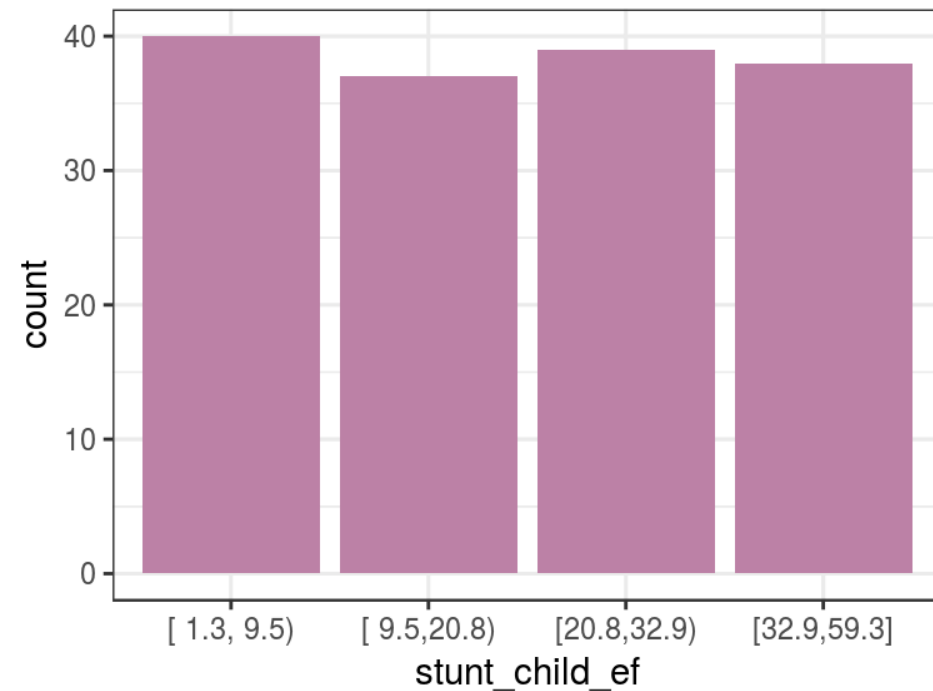
		O-Ring Failure	
		Y	N
Temperature	≤ 75	7	11
	> 75	0	6



Não-Supervisionada



Mesmo Intervalo



Mesmo frequência



Normalização

- Útil para cálculo de distância entre instâncias com atributos com diferentes intervalos
- Intervalos: $[0,1]$ ou $[-1,1]$
- Atributos com valores maiores podem dominar os com valores menores



Normalização por Min-Max

- Valores do atributo são ajustados para o intervalo $[a,b]$ baseado no valor máximo e mínimo:

$$v[i] = \frac{v[i] - \min(v)}{\max(v) - \min(v)}(b - a) + a$$

For:

$$(X_1, Y_1) = (1, 40)$$

$$(X_2, Y_2) = (2, 100)$$

Distance Equation Solution:

$$d = \sqrt{(2 - 1)^2 + (100 - 40)^2}$$

$$d = \sqrt{(1)^2 + (60)^2}$$

$$d = \sqrt{1 + 3600}$$

$$d = \sqrt{3601}$$

$$d = 60.008333$$

For:

$$(X_1, Y_1) = (0, 0)$$

$$(X_2, Y_2) = (1, 1)$$

Distance Equation Solution:

$$d = \sqrt{(1 - 0)^2 + (1 - 0)^2}$$

$$d = \sqrt{(1)^2 + (1)^2}$$

$$d = \sqrt{1 + 1}$$

$$d = \sqrt{2}$$

$$d = 1.414214$$



Normalização por Média e Desvio Padrão

- Média = 0
- Desvio padrão = 1

$$v[i] = \frac{v[i] - \bar{v}}{\sigma_v}$$

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix} .$$