

Visualização de Dados

Preparação de dados

Tópico 1

O pré-processamento

Pequeno contexto

- Os dados podem ser **recolhidos** ou ser **produzidos** por sensores, também podem ser criados por **simulações** computacionais, ou até podem ser **artificialmente gerados**.
- Os dados podem vir em “bruto” (**raw**), isto é, não processados, ou podem já estar **transformados/processados** através de remoção de ruído, outliers, escalados ou interpolados.

Álvaro Figueira • VD • 2023 • 1ª ed.

3

Tipos de Dados

- 1. Nominais / nominals:** Este é um tipo de dado que é usado para criar labels para variáveis sem valor quantitativo. Exemplos: género, cor do cabelo, ou nacionalidade.
- 2. Ordinais / ranked:** Este tipo de dados tem uma ordem (i.e., pode ser ordenado). Um exemplo pode ser uma questão em que se pede aos respondentes para classifiquem a sua felicidade numa escala de 1 a 7.

Além de nominal e ordinal, existem outros dois tipos principais de dados:

- 3. Intervalares:** São **dados numéricos em que a diferença entre dois valores é significativa**. Exemplos incluem temperatura em Celsius ou Fahrenheit, Intervalos horários, longitude, etc.
- 4. Rácio:** São como dados intervalares, mas têm **uma clara definição de zero**. Exemplos são idade, rendimento, ou altura. É nesta categoria que o conceito de "discreto" (contável, por exemplo, número de animais de estimação que uma pessoa possui) e de "contínuo" (mensurável, por exemplo, altura, peso) se podem diluir.

Nota: Dados "binários" podem ocorrer em diferentes categorias, dependendo do que representam. Se representam presença/ausência, sucesso/falha, etc., podem ser considerados nominais. Se representam maior/menor, concordar/discordar, etc., podem ser considerados ordinais.

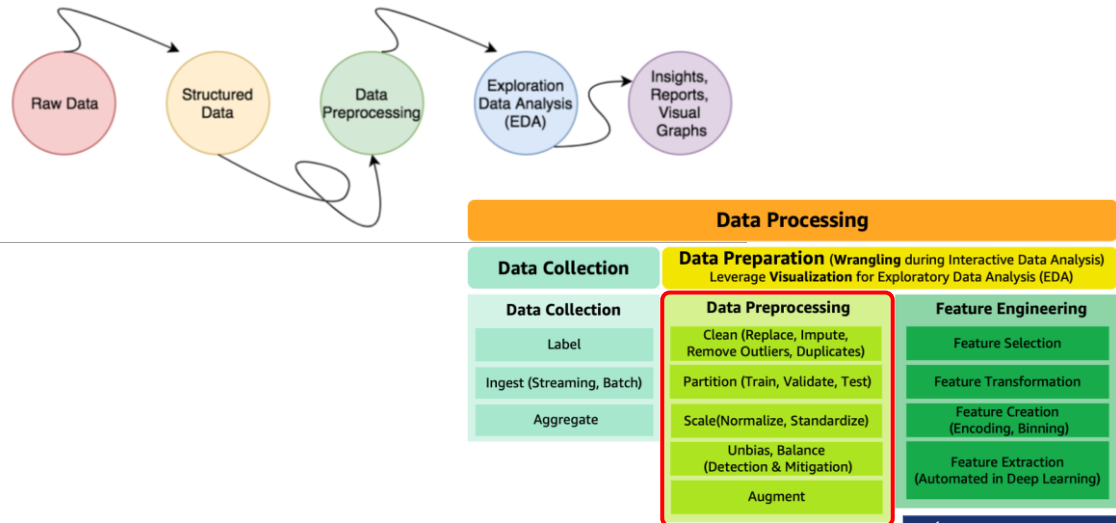
Portanto, as classes principais de dados são: **Nominal**, **Ordinal**, **Intervalar** e **Rácio**.

Álvaro Figueira • VD • 2023 • 1ª ed.

4

Pré-processamento de dados (Data Pre-processing)

Muitas vezes é preferível visualizar os dados em bruto – “raw” (tipicamente em na área médica). Mas, para algumas aplicações é necessário proceder a algum **pré-processamento**.



Álvaro Figueira • VD • 2023 • 1ª ed.

5

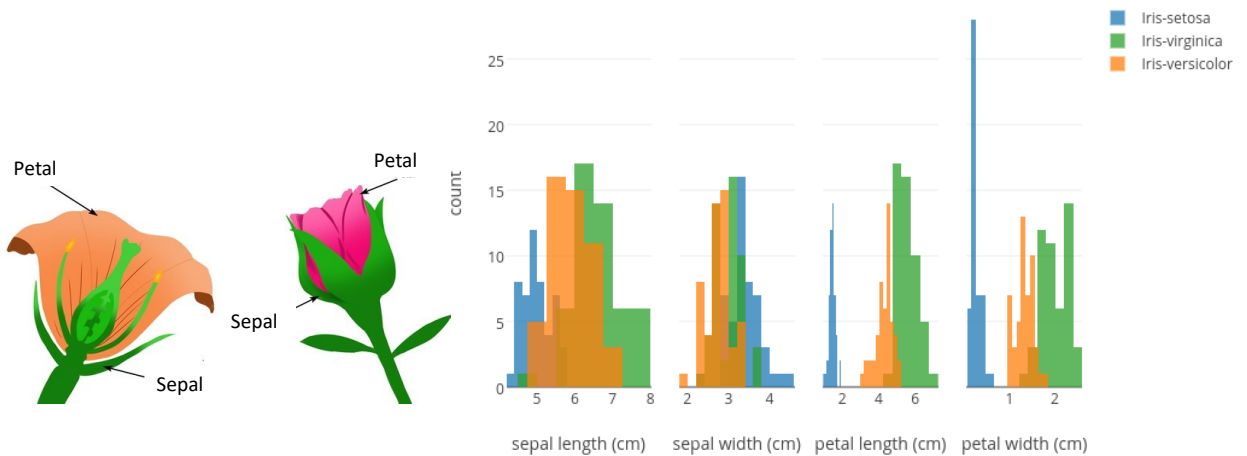
Estatísticas preliminares

- Uma **análise estatística simples** pode fornecer informação útil tal como:
 - Estatísticas de **sumarização**
 - Análise de **variância**
 - Detecção de **outliers** (valores fora do normal)
 - Identificação de grupos (**clusters**) semelhantes
 - Perceber a **distribuição** para identificar possibilidade de aumento de dados
 - Identificação de variáveis redundantes usando a **correlação**
- **Histogramas** e **Violin plots** podem ser usados para analisar a **distribuição dos dados**
- **Matriz de correlação** ajuda na identificação de correlações

Álvaro Figueira • VD • 2023 • 1ª ed.

6

Histogramas



Álvaro Figueira • VD • 2023 • 1ª ed.

7

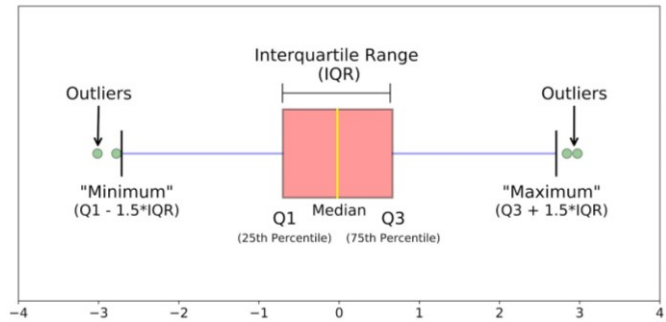
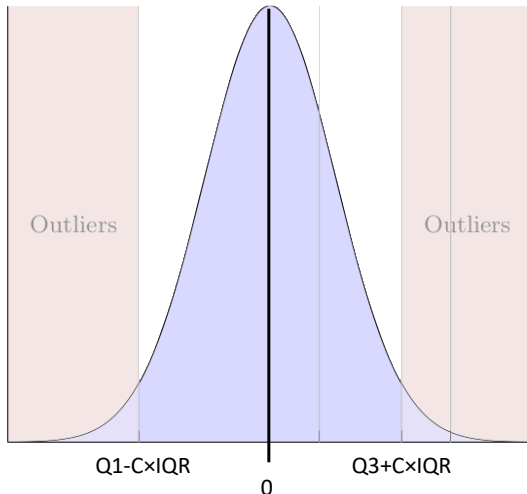
Método simples para deteção de outliers

- Supondo que a variável j tem uma distribuição Gaussiana $N(\mu_j, \sigma_j)$. Então, primeiro deve-se transformar os dados para que se verifique $\mu_j = 0$ e $\sigma_j = 1$
- Se x_{ij} é o valor da variável j na instância i , e c é uma constante, então a probabilidade de $|x_{ij}| \geq c$ decresce rapidamente à medida que c cresce
- **Método simples:**
 x é um outlier se x está **fora do intervalo** $[Q1 - 2.5 \times IQR, Q3 + 2.5 \times IQR]$
- Geralmente usa-se $c = 1.5, 2, 2.5, \dots$
- IQR ié o "interquartile range"

Álvaro Figueira • VD • 2023 • 1ª ed.

8

Método simples para detecção de outliers



Álvaro Figueira • VD • 2023 • 1ª ed.

9

Método simples para identificar variáveis redundantes

- Sejam x_i e x_j duas variáveis. Primeiro, calcular a sua correlação

$$\text{cor}(x_i, x_j) = \frac{\text{cov}(x_i, x_j)}{\sqrt{\text{var}(x_i) \text{var}(x_j)}}$$

- Em que $\text{cov}(x_i, x_j)$ é dada por

$$\text{cov}(x_i, x_j) = \frac{1}{m-1} \sum_{k=1}^m (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

- E $\text{var}(x_i)$ é dada por

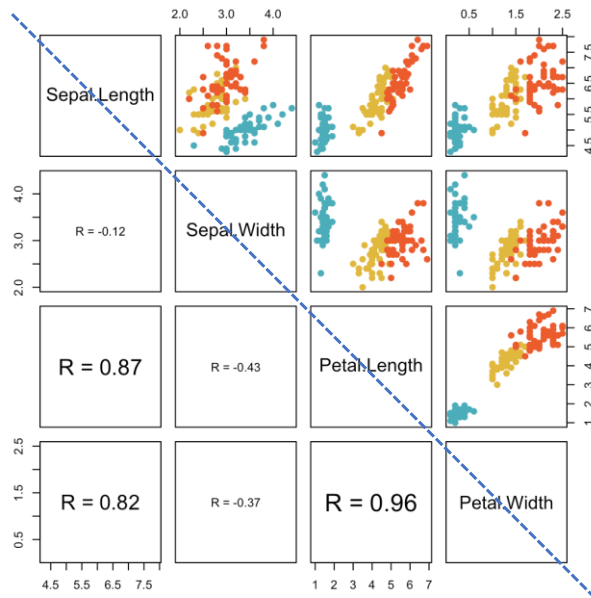
$$\text{var}(x_i) = \sigma^2$$

- Então, valores de $|\text{cor}(x_i, x_j)|$ próximos de 1 indicam correlação alta. Nesse caso, uma das duas variáveis, x_i ou x_j pode ser descartada

Álvaro Figueira • VD • 2023 • 1ª ed.

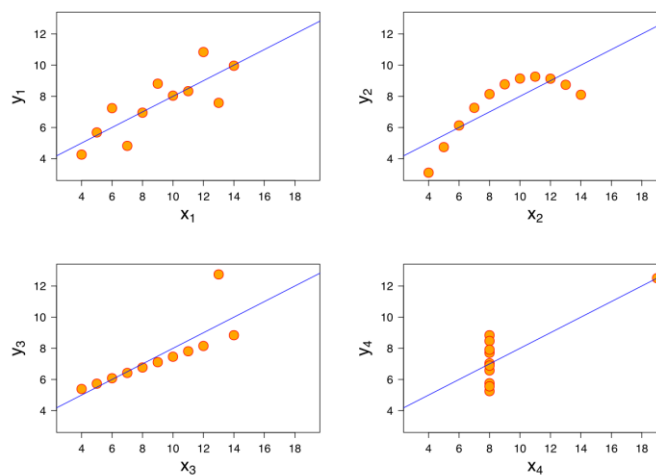
10

Método simples para identificar variáveis redundantes

Álvaro Figueira • VD • 2023 • 1ª ed.¹¹

11

Atenção: a correlação descreve muito pouco!



Todos têm
correlação igual
(de 0.816)

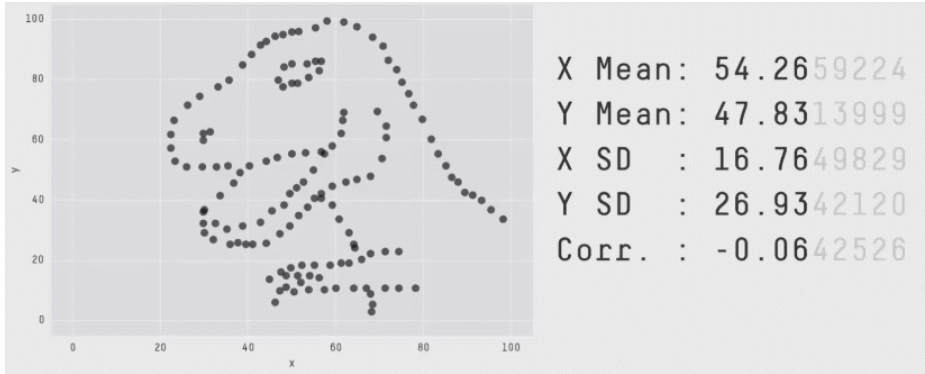
By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=9838454>

Álvaro Figueira • VD • 2023 • 1ª ed.

12

As propriedades estatísticas das distribuições tornaram-se piadas!

animação, muda para estrela vagina, e demais formas ...



Álvaro Figueira • VD • 2023 • 1ª ed.¹³

13

Topic 2

Data cleaning (cleansing)

Álvaro Figueira • VD • 2023 • 1ª ed.

14

Missing Values and Data Cleaning

- On "real" data it is normal that some **data** be **missing** or be **wrong**
- Common strategies to address such issue
 - **Discard** the data instance with defect.
 - Note: it can represent an important data loss
 - Assign a **sentinel** value. Example: NA
 - However, the sentinel value cannot be used on the calculations
 - Calculate a **replacement** value.
 - However, *data imputation* might be risky

Álvaro Figueira • VD • 2023 • 1ª ed.¹⁵

15

Data Imputation

- Simple data imputation methods
 - Assign the **average / median / mode** value.

Note: it can hide outlier values
 - Assign the **most common** value
 - Assign a value based on the **k nearest neighbors** (using a mean or median).

Note: on the neighborhood calculation it is difficult to know if there are more relevant attributes
 - Assign a value computed by **linear interpolation** or **regression interpolation**.
 - In a time series, use the **last observation carried forward** (LOCF) or the reverse (NOCB).


Note: not useful when there are many sequential missing values.

Álvaro Figueira • VD • 2023 • 1ª ed.

16


Data Imputation

It is probably one of the most difficult steps in data mining.
Must be **done with care!**



originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
Macbeth	0	1908-01-01	<NA>	9	Drama,Short
Making Moving Pictures	0	1908-01-01	<NA>	<NA>	Documentary,Short
Mallorca, isla dorada	0	1908-01-01	<NA>	14	Documentary,Short
The Man and the Woman	0	1908-01-01	<NA>	13	Drama,Short
The Man in the Box	0	1908-01-01	<NA>	<NA>	Crime,Drama,Short
Maria Marten	0	1908-01-01	<NA>	<NA>	Drama,Short
María Rosa	0	1908-01-01	<NA>	18	Short
The Merchant of Venice	0	1908-01-01	<NA>	9	Drama,Short

Activity duration



16/06/17, 10:36	Person X	-	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)
16/06/17, 09:55	Person X	-	Página: Notas dos testes
16/06/17, 09:54	Person X	-	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)

Álvaro Figueira • VD • 2023 • 1ª ed.¹

17

Several problems happen in real-world data

color	director_name	duration	gross	movie_title	language	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall Street	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful Eight	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	Margaret	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of Smaug	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of Justice	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected Journey	English	USA	180000000	2012	7.9
Color	Edward Hall	180		Restless	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud Atlas	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon Tattoo	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of God	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five Armies	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6

The data cleaning process addresses these issues

Álvaro Figueira • VD • 2023 • 1ª ed.¹

18

Topic 3

Normalization

Álvaro Figueira • VD • 2023 • 1ª ed.

19

Normalization

- On applications that involve instances' comparison, one scenario can **distort** the result and introduce **bias**:
 - Example: When the Euclidean **norm** of the vectors (line or column) they represent is **too different**
- A possible solution is **normalization**:
 - Transform the data so that they present a **desired statistical property in a more regular ("normal") data distribution**

The goal is to create a single common scale

Álvaro Figueira • VD • 2023 • 1ª ed.

20

Normalization

- To address the first scenario, it is possible to transform the vector so that the data instances have **unit norms**

$$x_{ij} = \frac{x_{ij}}{\|x_i\|} \text{ for } 1 \leq j \leq m$$

- The vectors can be the lines or columns of the data matrix

Álvaro Figueira • VD • 2023 • 1ª ed.

21

Min-Max Normalization

- Then, the process consists of transforming the data so that the values range in $[0,1]$ ("*min-max scaling*")
- If the maximum x_j^{max} and minimum x_j^{min} values are known, we can do

$$x_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}}$$

- On specific cases it could be interesting to use the **known maximum and minimum** values, such as on **percentages**

Álvaro Figueira • VD • 2023 • 1ª ed.

22

Z-Score Normalization

- Another known normalization is the **standardization**. It transforms the data so that the **average is 0** and the **standard deviation is 1**

$$x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

- However, **normalization can distort** the data, for instance, in the presence of outliers, the data can be flattened

Álvaro Figueira • VD • 2023 • 1ª ed.

23

Max Scaling and Feature Scaling Normalization

- **Max Scaling:**
 - In this method, data is divided by the maximum value of each feature to bring it in the range [0, 1]
- **Unit Vector Normalization (Feature Scaling)**
 - This method rescales a data point to the length of 1 (like a unit vector). It's used when direction of the data matters, but not the length of the feature vector.

Note: Normalization is not always desirable! Some algorithms are immune to the scale.

Álvaro Figueira • VD • 2023 • 1ª ed.

24

Topic 4

Interpolation

Álvaro Figueira • VD • 2023 • 1ª ed.

25

Interpolation

- Sometimes it is necessary to fill the "space" between samples, this is done through interpolation
- Given x_j and x_k , the linear interpolation between them can be computed using

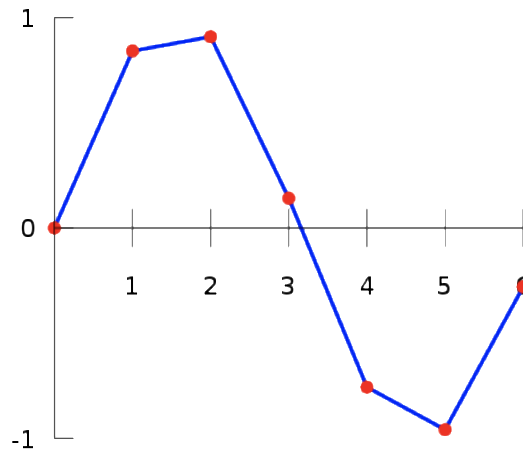
$$x = (1 - \alpha) \times x_j + \alpha \times x_k$$

- With α ranging in $[0,1]$

Álvaro Figueira • VD • 2023 • 1ª ed.

26

Linear Interpolation



```
# in R
# Create two vectors
x <- c(1, 2, 3, 4, 5)
y <- c(2, 3, 5, 8, 14)

# Interpolate at x = 2.5
result <- approx(x, y, xout = 2.5)
result
```

```
# Load the package
library(forecast)

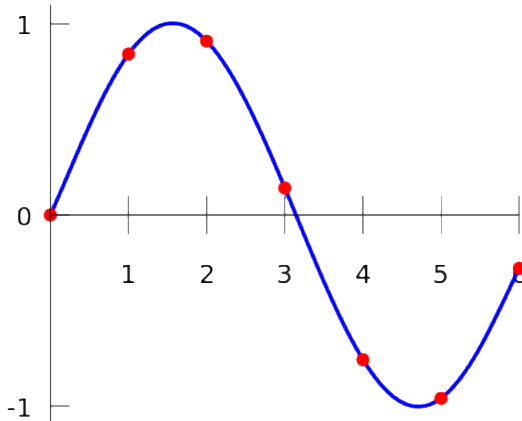
# Create a vector with NA values
data <- c(1, 2, NA, 4, 5, NA, 7)

# Perform interpolation with NAs
result <- na.interp(data)
result
```

Álvaro Figueira • VD • 2023 • 1ª ed.

27

Polynomial Interpolation



```
library(pracma)

# Create two vectors
x <- c(1, 2, 3, 4, 5)
y <- c(2, 3, 5, 8, 14)

# Perform polynomial interpolation (degree 2)
p <- polyfit(x, y, 2)

# Print the polynomial coefficients
p
```

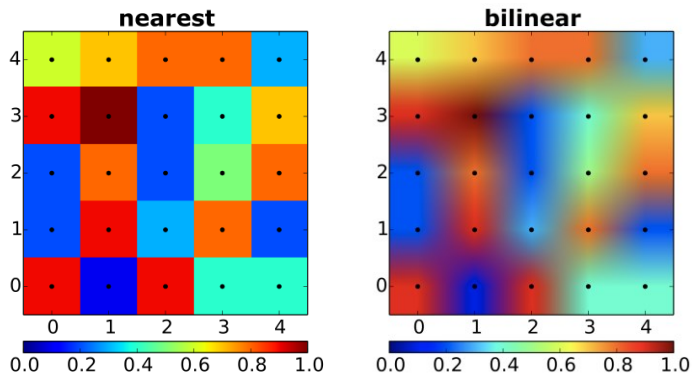
```
# Predict y at x = 2.5
y_pred <- polyval(p, 2.5)
y_pred
```

$$f(x) = -0.0001521x^6 - 0.003130x^5 + 0.07321x^4 - 0.3577x^3 + 0.2255x^2 + 0.9038x.$$

Álvaro Figueira • VD • 2023 • 1ª ed.

28

Interpolation in Higher Dimensions



```
library(akima)

# Create data
x <- c(1, 2, 3, 3, 1, 2)
y <- c(1, 1, 1, 2, 2, 2)
z <- c(1, 2, 3, 4, 5, 6)

# Define grid for interpolation
xo = seq(min(x), max(x), length = 100)
yo = seq(min(y), max(y), length = 100)

# Perform interpolation
ir <- akima::interp(
  x = x, y = y, z = z,
  xo=xo, yo=yo, linear = TRUE)

# Print result
ir
```