

*Jeidsan A. da C. Pereira*

---

# ***Estatística e Ciência de Dados***

***Notas e solução dos exercícios***



---

## Conteúdo

---

<b>Prefácio</b>	<b>ix</b>
<b>Prefácio</b>	<b>ix</b>
Pendências . . . . .	ix
<b>1 Estatística, Ciência de Dados e Megadados</b>	<b>1</b>
1.1 Introdução . . . . .	1
1.2 Aprendizado com estatística . . . . .	1
1.3 Aprendizado automático . . . . .	1
1.4 Uma cronologia do desenvolvimento da estatística . . . . .	1
1.5 Notação e tipos de dados . . . . .	1
1.6 Paradigmas para o aprendizado com estatística . . . . .	1
1.7 Este livro . . . . .	1
1.8 Conjuntos de dados . . . . .	1
1.9 Notas do capítulo . . . . .	1
<b>I Análise Exploratória de Dados</b>	<b>3</b>
<b>2 Preparação dos dados</b>	<b>5</b>
2.1 Considerações preliminares . . . . .	5
2.2 Planilhas de dados . . . . .	5
2.3 Construção de tabelas . . . . .	5
2.4 Construção de gráficos . . . . .	5
2.5 Notas de capítulo . . . . .	5
2.6 Exercícios . . . . .	5

<b>3</b>	<b>Análise de dados de uma variável</b>	<b>7</b>
3.1	Introdução . . . . .	9
3.2	Distribuição de frequências . . . . .	9
3.3	Medidas resumo . . . . .	9
3.4	<i>Boxplots</i> . . . . .	9
3.5	Modelos probabilísticos . . . . .	9
3.6	Dados amostrais . . . . .	9
3.7	Gráficos QQ . . . . .	9
3.8	Desvio padrão e erro padrão . . . . .	9
3.9	Intervalo de confiança e tamanho da amostra . . . . .	9
3.10	Transformação de variáveis . . . . .	9
3.11	Notas de capítulo . . . . .	9
3.12	Exercícios . . . . .	9
<b>4</b>	<b>Análise de dados de duas variáveis</b>	<b>11</b>
4.1	Introdução . . . . .	11
4.2	Duas variáveis qualitativas . . . . .	11
4.3	Duas variáveis quantitativas . . . . .	11
4.4	Uma variável qualitativa e outra quantitativa . . . . .	11
4.5	Notas de capítulo . . . . .	11
4.6	Exercícios . . . . .	11
<b>5</b>	<b>Análise de dados de várias variáveis</b>	<b>13</b>
5.1	Introdução . . . . .	13
5.2	Gráficos para três variáveis . . . . .	13
5.3	Gráficos para quatro ou mais variáveis . . . . .	13
5.4	Medidas resumo multivariadas . . . . .	13
5.5	Tabelas de contingência de múltiplas entradas . . . . .	13
5.6	Notas de capítulo . . . . .	13
5.7	Exercícios . . . . .	13

<b>6</b>	<b>Análise de Regressão</b>	<b>15</b>
6.1	Introdução . . . . .	15
6.2	Regressão linear simples . . . . .	15
6.3	Regressão linear múltipla . . . . .	15
6.4	Regressão para dados longitudinais . . . . .	15
6.5	Regressão logística . . . . .	15
6.6	Notas de capítulo . . . . .	15
6.7	Exercícios . . . . .	15
<b>7</b>	<b>Análise de Sobrevivência</b>	<b>17</b>
7.1	Introdução . . . . .	17
7.2	Estimação da função de sobrevivência . . . . .	17
7.3	Comparação de curvas de sobrevivência . . . . .	17
7.4	Regressão para dados de sobrevivência . . . . .	17
7.5	Notas de capítulo . . . . .	17
7.6	Exercícios . . . . .	17
<b>II</b>	<b>Aprendizado Supervisionado</b>	<b>19</b>
<b>8</b>	<b>Regularização e Modelos Aditivos Generalizados</b>	<b>21</b>
8.1	Introdução . . . . .	21
8.2	Regularização . . . . .	21
8.3	Modelos aditivos generalizados (GAM) . . . . .	21
8.4	Notas de capítulo . . . . .	21
8.5	Exercícios . . . . .	21
<b>9</b>	<b>Classificação por meio de técnicas clássicas</b>	<b>23</b>
9.1	Introdução . . . . .	23
9.2	Classificação por regressão logística . . . . .	23
9.3	Análise discriminante linear . . . . .	23
9.4	Classificador do vizinho mais próximo . . . . .	23
9.5	Algumas extensões . . . . .	23
9.6	Notas de capítulo . . . . .	23
9.7	Exercícios . . . . .	23

<b>10 Algoritmos de Suporte Vetorial</b>	<b>25</b>
10.1 Introdução . . . . .	25
10.2 Fundamentação dos algoritmos de suporte vetorial . . . . .	25
10.3 Classificador de margem máxima . . . . .	25
10.4 Classificador de margem flexível . . . . .	25
10.5 Classificador de margem não linear . . . . .	25
10.6 Regressão por algoritmos de suporte vetorial . . . . .	25
10.7 Notas de capítulo . . . . .	25
10.8 Exercícios . . . . .	25
<b>11 Árvores e Florestas</b>	<b>27</b>
11.1 Introdução . . . . .	27
11.2 Classificação por árvores . . . . .	27
11.3 <i>Bagging, boosting</i> e florestas . . . . .	27
11.4 Árvores para regressão . . . . .	27
11.5 Notas de capítulo . . . . .	27
11.6 Exercícios . . . . .	27
<b>12 Redes neurais</b>	<b>29</b>
12.1 Introdução . . . . .	29
12.2 <i>Perceptron</i> . . . . .	29
12.3 Redes com camadas ocultas . . . . .	29
12.4 O algoritmo de retropropagação ( <i>backpropagation</i> ) . . . . .	29
12.5 Aprendizado profundo ( <i>Deep learning</i> ) . . . . .	29
12.6 Notas de capítulo . . . . .	29
12.7 Exercícios . . . . .	29
<b>III Aprendizado não Supervisionado</b>	<b>31</b>

<i>Contents</i>	vii
<b>13 Análise de Agrupamentos</b>	<b>33</b>
13.1 Introdução . . . . .	33
13.2 Estratégias de agrupamento . . . . .	33
13.3 Algoritmos hierárquicos . . . . .	33
13.4 Algoritmos de partição: K-médias . . . . .	33
13.5 Notas de capítulo . . . . .	33
13.6 Exercícios . . . . .	33
<b>14 Redução de dimensionalidade</b>	<b>35</b>
14.1 Introdução . . . . .	35
14.2 Análise de Componentes Principais . . . . .	35
14.3 Análise fatorial . . . . .	35
14.4 Análise de componentes independentes . . . . .	35
14.5 Notas de capítulo . . . . .	35
14.6 Exercícios . . . . .	35
<b>Apêndice</b>	<b>35</b>
<b>A Otimização numérica</b>	<b>37</b>
A.1 Introdução . . . . .	37
A.2 O método de Newton-Raphson . . . . .	37
A.3 O método scoring . . . . .	37
A.4 O método de Gauss-Newton . . . . .	37
A.5 Métodos Quase-Newton . . . . .	37
A.6 Aspectos computacionais . . . . .	37
A.7 Notas de capítulo . . . . .	37
A.8 Exercícios . . . . .	37
<b>B Noções de simulação</b>	<b>39</b>
B.1 Introdução . . . . .	39
B.2 Método Monte Carlo . . . . .	39
B.3 Simulação de variáveis discretas . . . . .	39

B.4	Simulação de variáveis contínuas . . . . .	39
B.5	Simulação de vetores aleatórios . . . . .	39
B.6	Métodos de reamostragem . . . . .	39
B.7	Notas de capítulo . . . . .	39
B.8	Exercícios . . . . .	39
<b>C</b>	<b>Algoritmos para dados aumentados</b>	<b>41</b>
C.1	Introdução . . . . .	41
C.2	O algoritmo EM . . . . .	41
C.3	O algoritmo EM Monte Carlo . . . . .	41
C.4	Cálculo de erros padrões . . . . .	41
C.5	O algoritmo para dados aumentados . . . . .	41
C.6	Exercícios . . . . .	41



---

## ***Prefácio***

---

Esta página contém notas e solução para os exercícios propostos no livro **Estatística e Ciência de Dados**, de autoria de Pedro Alberto Morettin e Júlio da Motta Singer, publicado pela LTC em 2022 [Morettin and Singer, 2022].

É importante destacar que trata-se de um produto não oficial, as anotações e soluções de exercícios aqui apresentadas são de cunho pessoal e não possuem qualquer revisão ou análise por parte dos autores da obra ou da editora. Dessa forma e por se tratar de um produto construído durante o processo de aprendizagem, o conteúdo pode conter erros, tanto no texto em si, como na lógica utilizada para solução dos exercícios.

Dúvidas ou sugestões de melhoria podem ser encaminhadas para o e-mail *jeidsan.pereira@gmail.com*<sup>1</sup>.

---

## **Pendências**

•

---

<sup>1</sup>mailto:jeidsan.pereira@gmail.com



# 1

---

## *Estatística, Ciência de Dados e Megadados*

---

### 1.1 Introdução

---

### 1.2 Aprendizado com estatística

---

### 1.3 Aprendizado automático

---

### 1.4 Uma cronologia do desenvolvimento da estatística

---

### 1.5 Notação e tipos de dados

---

### 1.6 Paradigmas para o aprendizado com estatística

---

### 1.7 Este livro

---

### 1.8 Conjuntos de dados

---

### 1.9 Notas do capítulo

---



## **Parte I**

# **Análise Exploratória de Dados**



## 2

---

### *Preparação dos dados*

---

#### 2.1 Considerações preliminares

---

#### 2.2 Planilhas de dados

---

#### 2.3 Contrução de tabelas

---

#### 2.4 Construção de gráficos

---

#### 2.5 Notas dde capítulo

---

#### 2.6 Exercícios





# 3

## *Análise de dados de uma variável*



---

### 3.1 Introdução

---

### 3.2 Distribuição de frequências

---

### 3.3 Medidas resumo

---

### 3.4 *Boxplots*

---

### 3.5 Modelos probabilísticos

---

### 3.6 Dados amostrais

---

### 3.7 Gráficos QQ

---

### 3.8 Desvio padrão e erro padrão

---

### 3.9 Intervalo de confiança e tamanho da amostra

---

### 3.10 Transformação de variáveis

---

### 3.11 Notas de capítulo

---

### 3.12 Exercícios



# 4

---

## *Análise de dados de duas variáveis*

---

### 4.1 Introdução

---

### 4.2 Duas variáveis qualitativas

---

### 4.3 Duas variáveis quantitativas

---

### 4.4 Uma variável qualitativa e outra quantitativa

---

### 4.5 Notas de capítulo

---

### 4.6 Exercícios



# 5

---

## *Análise de dados de várias variáveis*

---

---

### 5.1 Introdução

---

### 5.2 Gráficos para três variáveis

---

### 5.3 Gráficos para quatro ou mais variáveis

---

### 5.4 Medidas resumo multivariadas

---

### 5.5 Tabelas de contingência de múltiplas entradas

---

### 5.6 Notas de capítulo

---

### 5.7 Exercícios

---





# 6

---

## *Análise de Regressão*

---

### 6.1 Introdução

---

### 6.2 Regressão linear simples

---

### 6.3 Regressão linear múltipla

---

### 6.4 Regressão para dados longitudinais

---

### 6.5 Regressão logística

---

### 6.6 Notas de capítulo

---

### 6.7 Exercícios

---



# 7

---

## *Análise de Sobrevivência*

---

---

### 7.1 Introdução

---

### 7.2 Estimação da função de sobrevivência

---

### 7.3 Comparação de curvas de sobrevivência

---

### 7.4 Regressão para dados de sobrevivência

---

### 7.5 Notas de capítulo

---

### 7.6 Exercícios



## **Parte II**

# **Aprendizado Supervisionado**



# 8

---

## *Regularização e Modelos Aditivos Generalizados*

---

### 8.1 Introdução

---

### 8.2 Regularização

---

### 8.3 Modelos aditivos generalizados (GAM)

---

### 8.4 Notas de capítulo

---

### 8.5 Exercícios





# 9

---

## *Classificação por meio de técnicas clássicas*

---

### 9.1 Introdução

---

### 9.2 Classificação por regressão logística

---

### 9.3 Análise discriminante linear

---

### 9.4 Classificador do vizinho mais próximo

---

### 9.5 Algumas extensões

---

### 9.6 Notas de capítulo

---

### 9.7 Exercícios



# 10

---

## *Algoritmos de Suporte Vetorial*

---

---

### 10.1 Introdução

---

### 10.2 Fundamentação dos algoritmos de suporte vetorial

---

### 10.3 Classificador de margem máxima

---

### 10.4 Classificador de margem flexível

---

### 10.5 Classificador de margem não linear

---

### 10.6 Regressão por algoritmos de suporte vetorial

---

### 10.7 Notas de capítulo

---

### 10.8 Exercícios

---



# 11

---

## *Árvores e Florestas*

---

### 11.1 Introdução

---

### 11.2 Classificação por árvores

---

### 11.3 *Bagging, boosting* e florestas

---

### 11.4 Árvores para regressão

---

### 11.5 Notas de capítulo

---

### 11.6 Exercícios



# 12

---

## *Redes neurais*

---

---

### 12.1 Introdução

---

### 12.2 *Perceptron*

---

### 12.3 Redes com camadas ocultas

---

### 12.4 O algoritmo de retropropagação (*backpropagation*)

---

### 12.5 Aprendizado profundo (*Deep learning*)

---

### 12.6 Notas de capítulo

---

### 12.7 Exercícios





## **Parte III**

# **Aprendizado não Supervisionado**



# 13

---

## *Análise de Agrupamentos*

---

---

### 13.1 Introdução

---

### 13.2 Estratégias de agrupamento

---

### 13.3 Algoritmos hierárquicos

---

### 13.4 Algoritmos de partição: K-médias

---

### 13.5 Notas de capítulo

---

### 13.6 Exercícios



# 14

---

## *Redução de dimensionalidade*

---

### 14.1 Introdução

---

### 14.2 Análise de Componentes Principais

---

### 14.3 Análise fatorial

---

### 14.4 Análise de componentes independentes

---

### 14.5 Notas de capítulo

---

### 14.6 Exercícios



# A

---

## *Otimização numérica*

---

### A.1 Introdução

---

### A.2 O método de Newton-Raphson

---

### A.3 O método scoring

---

### A.4 O método de Gauss-Newton

---

### A.5 Métodos Quase-Newton

---

### A.6 Aspectos computacionais

---

### A.7 Notas de capítulo

---

### A.8 Exercícios





# B

---

## *Noções de simulação*

---

---

### B.1 Introdução

---

### B.2 Método Monte Carlo

---

### B.3 Simulação de variáveis discretas

---

### B.4 Simulação de variáveis contínuas

---

### B.5 Simulação de vetores aleatórios

---

### B.6 Métodos de reamostragem

---

### B.7 Notas de capítulo

---

### B.8 Exercícios



# C

---

## *Algoritmos para dados aumentados*

---

### C.1 Introdução

---

### C.2 O algoritmo EM

---

### C.3 O algoritmo EM Monte Carlo

---

### C.4 Cálculo de erros padrões

---

### C.5 O algoritmo para dados aumentados

---

### C.6 Exercícios



---

## ***Bibliografia***

---

Pedro Alberto Morettin and Julio da Motta Singer. *Estatística e Ciência de Dados*. LTC, Rio de Janeiro, 2022.