

Julia

数据科学应用

[美] 扎卡赖亚斯·弗格里斯 (Zacharias Voulgaris) 著 陈光欣 译



中国工信出版集团

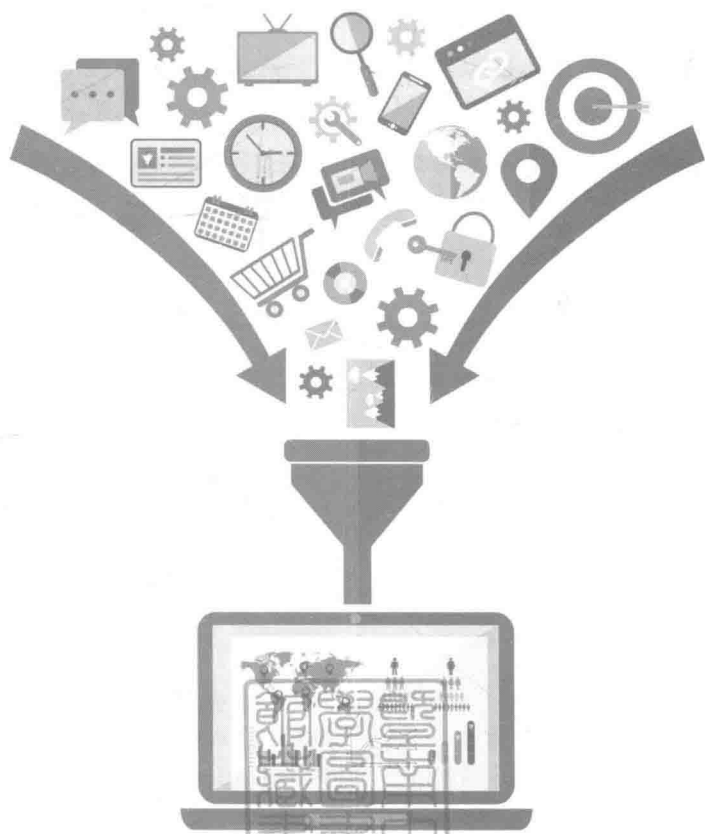


人民邮电出版社
POSTS & TELECOM PRESS

Julia

数据科学应用

[美] 扎卡赖亚斯·弗格里斯 (Zacharias Voulgaris) 著 陈光欣 译



人民邮电出版社
北京

图书在版编目(CIP)数据

Julia数据科学应用 / (美) 弗格里斯
(Zacharias Voulgaris) 著; 陈光欣译. — 北京: 人
民邮电出版社, 2018.2
ISBN 978-7-115-47328-8

I. ①J… II. ①弗… ②陈… III. ①程序语言—研究
IV. ①TP312

中国版本图书馆CIP数据核字(2017)第309720号

版权声明

Simplified Chinese translation copyright ©2017 by Posts and Telecommunications Press

ALL RIGHTS RESERVED

Julia for Data Scientist, by Zacharias Voulgaris, ISBN 9781634621304

Copyright © 2016 by Technics Publications, LLC

本书中文简体版由 **Technics Publications** 授权人民邮电出版社出版。未经出版者书面许可,
对本书的任何部分不得以任何方式或任何手段复制和传播。
版权所有, 侵权必究。

-
- ◆ 著 [美] 扎卡赖亚斯·弗格里斯 (Zacharias Voulgaris)
 - 译 陈光欣
 - 责任编辑 陈冀康
 - 责任印制 焦志炜
 - ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
 - 邮编 100164 电子邮件 315@ptpress.com.cn
 - 网址 <http://www.ptpress.com.cn>
 - 大厂聚鑫印刷有限责任公司印刷
 - ◆ 开本: 720×960 1/16
 - 印张: 19.25
 - 字数: 240千字 2018年2月第1版
 - 印数: 1—2400册 2018年2月河北第1次印刷

著作权合同登记号 图字: 01-2016-8081号

定价: 69.00元

读者服务热线: (010)81055410 印装质量热线: (010)81055316

反盗版热线: (010)81055315

广告经营许可证: 京东工商广登字 20170147号



内容提要

数据科学通过各种统计学和机器学习的技术与方法，将数据转换为有用的信息或知识。Julia 是一种在数据科学领域逐渐流行起来的语言。

本书会提出一系列在数据科学流程中常见的、有代表性的实际问题，并指导读者使用 Julia 去解决这些问题。全书共 13 章，涵盖了 Julia 基础知识、工作环境搭建、语言基础和高级内容、数据科学应用、数据可视化、机器学习方法（包括无监督式学习方法和监督式学习方法）、图分析方法等重要话题。附录部分给出了学习和使用 Julia 的一些有用的参考资料，还给出了各章的思考题的答案。

本书适合对数据科学的知识和应用方法感兴趣的读者阅读，特别适合有志于学习 Julia 并从事数据科学相关工作的人员学习参考。



前言

我是在几年前发现 Julia 的，从此就被它的强大能力与巨大潜力所吸引。Julia 具有用户友好的集成开发环境（Integrated Development Environment, IDE），这使它很容易上手；它还具有高级的逻辑表达能力（非常类似 Matlab 和其他高级语言）和极高的性能，这使它的功能非常强大。但是，当时我正致力于研究其他更成熟的平台，比如 R 和 Java，未能给予 Julia 太多的关注。

因此，我只了解了 Julia 的一些基础知识，以及当时的教程中提供的一些具体应用，并没有进行更多的研究。除了 Julia 之外，我还知道不断有一些新的有趣的语言被开发出来，但大多数是昙花一现。

那么，为什么我现在又对 Julia 感兴趣了呢？一个原因就是，这些年它一直保持着良好的发展势头，Julia 会议的参与人数每年都有显著的增长。尽管我曾经很熟悉它的基本知识，但当我重拾 Julia 时，发现有很多新的知识需要学习。从我初识 Julia 之后，它已经有了很大的发展。

更重要的原因是，Julia 已经跨过了大西洋，引起了欧洲从业者的极大兴趣，其中一位已经为这种相当年轻的语言创建了一系列视频和练习资料。

在试用了 Julia 0.2 版之后，我开始琢磨，除了快速分解质因数和计算第 n 个斐波那契数之外，是否能使用 Julia 来做些真正有用的事情。虽然 0.2 版仅有几个软件包，文档也做得很差，我只能找到零星几个介绍这门语言的视频，多数还是来自某个 Python 会议上的发言。但是，我还是在计算机上保留着 Julia，并时不时地用它写个脚本，来解决 Project Euler、Programming Praxis 或类似站点上的编程问题。当时我是个项目经理，所以没有很大的积极性去掌握一门新的编程语言。我在 Julia 上所做的一切都是出于兴趣。

但是，几个月之后，我重新开始从事数据科学工作，并更加正式地使用 Julia 编程。我很快就发现，使用 Julia 编写代码比使用 Python 更容易，例如，即使使

用 Python 完成一个基本的数据加工任务，也需要一大堆扩展包。

在使用 Julia 解决小问题之后，我决定使用 Julia 独立完成一个完整的数据科学项目。在经历了不可避免的学习曲线和成长阵痛之后，我终于达到了目标。这并不是我最得意的成果，但它证明了在进行一些训练、尝试和纠错之后，Julia 可以高效地完成正式的数据科学任务。

在本书中，我会分享在这个项目以及随后的项目中获得的经验，阐述如何在数据科学的各个环节使用 Julia。尽管现在已经有了一些介绍 Julia 的书籍，但还没有一本全面介绍如何在数据科学领域内应用 Julia 的专著。我曾非常期待有这样一本书，但有了多年使用 Julia 的经验之后，我决定亲自上阵，撰写这样的一本书。

我完全清楚，撰写一本介绍正处于发展时期的语言的书风险有多大，但是 Julia 这门语言不会停止发展，如果我等待它完全成熟，这本书就永远不会完成。

我并不期待你能够全面掌握 Julia，或成为一个成熟的数据科学家。如果你渴望扩展技能，学习解决老问题的新方法，并严格按照本书的进度进行学习，那么 Julia 就会成为你进行数据分析的一个有效工具。



目录

第 1 章 Julia 简介	1
1.1 Julia 如何提高数据科学水平	2
1.1.1 数据科学工作流程	3
1.1.2 Julia 被数据科学社区接受的过程	5
1.2 Julia 扩展	6
1.2.1 包的质量	6
1.2.2 找到新的包	6
1.3 关于本书	7
第 2 章 建立数据科学工作环境	9
2.1 Julia IDE	9
2.1.1 Juno	10
2.1.2 IJulia	11
2.1.3 其他 IDE	12
2.2 Julia 扩展包	13
2.2.1 找到并选择扩展包	13
2.2.2 安装扩展包	14
2.2.3 使用扩展包	15
2.2.4 破解扩展包	16
2.3 IJulia 基础	16
2.3.1 文件处理	16
2.3.2 在 .jl 文件中组织代码	19
2.3.3 引用代码	20
2.3.4 工作目录	20

2.4 要使用的数据集	21
2.4.1 数据集描述	21
2.4.2 下载数据集	23
2.4.3 加载数据集	24
2.5 在 Julia 中实现一个简单的机器学习算法	25
2.5.1 算法描述	26
2.5.2 算法实现	27
2.5.3 算法测试	30
2.6 将工作区保存到数据文件	32
2.6.1 将数据保存为分隔值文件	32
2.6.2 将数据保存为 Julia 数据文件	33
2.6.3 将数据保存为文本文件	35
2.7 帮助	36
2.8 小结	36
2.9 思考题	37
第3章 Julia 入门	39
3.1 数据类型	39
3.2 数组	42
3.2.1 数组基础	42
3.2.2 在数组中引用多个元素	43
3.2.3 多维数组	44
3.3 字典	44
3.4 基本命令与函数	45
3.4.1 print()和 println()	46
3.4.2 typemax()和 typemin()	46
3.4.3 collect()	47
3.4.4 show()	47
3.4.5 linspace()	48

3.5 数学函数	48
3.5.1 round()	48
3.5.2 rand()和 randn()	49
3.5.3 sum()	52
3.5.4 mean()	53
3.6 数组与字典函数	53
3.6.1 in	53
3.6.2 append!()	54
3.6.3 pop!()	54
3.6.4 push!()	55
3.6.5 splice!()	55
3.6.6 insert!()	56
3.6.7 sort()和 sort!()	57
3.6.8 get()	57
3.6.9 keys()和 values()	58
3.6.10 length()和 size()	58
3.7 其他函数	59
3.7.1 time()	59
3.7.2 条件语句	59
3.7.3 string()	61
3.7.4 map()	62
3.7.5 versioin()	62
3.8 运算符、循环语句与条件语句	62
3.8.1 运算符	63
3.8.2 循环语句	64
3.8.3 break 命令	66
3.9 小结	66
3.10 思考题	67

第4章 Julia 进阶	68
4.1 字符串处理	68
4.1.1 split()	69
4.1.2 join()	70
4.1.3 正则表达式函数	70
4.2 定制函数	74
4.2.1 函数结构	74
4.2.2 匿名函数	75
4.2.3 多分派	75
4.2.4 函数示例	76
4.3 实现简单算法	77
4.4 创建完整解决方案	79
4.5 小结	83
4.6 思考题	84
第5章 Julia 数据科学应用概述	85
5.1 数据科学工作流程	85
5.2 数据工程	88
5.2.1 数据准备	88
5.2.2 数据探索	90
5.2.3 数据表示	92
5.3 数据建模	93
5.3.1 数据发现	93
5.3.2 数据学习	94
5.4 信息萃取	96
5.4.1 数据产品创建	96
5.4.2 知识、交付物与可视化产品	97
5.5 保持开放型思维	99
5.6 在实际问题中应用数据科学流程	99

5.6.1	数据准备	99
5.6.2	数据探索	100
5.6.3	数据表示	101
5.6.4	数据发现	101
5.6.5	数据学习	102
5.6.6	数据产品创建	102
5.6.7	知识、交付物和可视化产品	103
5.7	小结	103
5.8	思考题	105
第 6 章	Julia 数据工程	106
6.1	数据框	106
6.1.1	创建并填充数据框	107
6.1.2	数据框基础	108
6.1.3	引用数据框中的特定变量	109
6.1.4	探索数据框	109
6.1.5	筛选数据框	110
6.1.6	在数据框变量上应用函数	111
6.1.7	使用数据框进行工作	111
6.1.8	修改数据框	113
6.1.9	对数据框的内容进行排序	113
6.1.10	数据框的一些补充建议	114
6.2	导入与导出数据	115
6.2.1	使用.json 数据文件	115
6.2.2	保存数据到.json 文件	115
6.2.3	将数据文件加载到数据框	116
6.2.4	保存数据框到数据文件	116
6.3	数据清洗	117
6.3.1	数值型数据的清洗	117

6.3.2 文本型数据的清洗	118
6.4 数据格式化与转换	119
6.4.1 数值型数据的格式化	119
6.4.2 文本数据的格式化	119
6.4.3 数据类型的重要性	120
6.5 对数值型数据进行转换	120
6.5.1 标准化	121
6.5.2 离散化(分箱)与二值化	122
6.5.3 二值变量转换为连续型变量(仅对于二值分类问题)	123
6.5.4 文本数据转换	124
6.5.5 大小写标准化	124
6.5.6 向量化	124
6.6 初步的特征评价	126
6.6.1 回归	126
6.6.2 分类	126
6.6.3 特征评价补充说明	127
6.7 小结	128
6.8 思考题	129
第7章 探索数据集	130
7.1 倾听数据	130
本章要使用的扩展包	131
7.2 计算基本统计量和相关性	131
7.2.1 变量概要	133
7.2.2 变量之间的相关性	134
7.2.3 两个变量之间的可比性	136
7.3 绘制统计图	136
7.3.1 图形语法	137
7.3.2 为可视化准备数据	137

7.3.3 箱线图	138
7.3.4 条形图	138
7.3.5 折线图	139
7.3.6 散点图	140
7.3.7 直方图	143
7.3.8 导出统计图到文件	144
7.4 假设检验	145
7.4.1 检验的基础知识	145
7.4.2 错误类型	146
7.4.3 灵敏度与特异度	146
7.4.4 显著性水平与检验力	146
7.4.5 KRUSKAL-WALLIS 检验	147
7.4.6 T-检验	147
7.4.7 卡方检验	149
7.5 其他检验	151
7.6 统计检验附加说明	151
7.7 案例研究：探索 OnlineNewsPopularity 数据集	151
7.7.1 变量统计	152
7.7.2 可视化	153
7.7.3 假设	154
7.7.4 奇妙的 T-SNE 方法	155
7.7.5 结论	156
7.8 小结	156
7.9 思考题	159
第 8 章 构建数据空间	160
8.1 主成分分析	161
8.1.1 在 Julia 中使用 PCA	162
8.1.2 独立成分分析：主成分分析的最常用替代方法	164

8.2 特征评价与选择	165
8.2.1 方法论概述	165
8.2.2 在 Julia 中使用余弦相似度进行特征评价与选择	166
8.2.3 在 Julia 中使用 DID 进行特征评价与选择	168
8.2.4 特征评价与选择方法的优缺点	170
8.3 其他数据降维技术	170
8.3.1 其他降维方法概述	171
8.3.2 何时使用高级降维方法	172
8.4 小结	172
8.5 思考题	173
第 9 章 数据抽样与结果评价	175
9.1 抽样技术	175
9.1.1 基本抽样	176
9.1.2 分层抽样	176
9.2 分类问题的性能指标	177
9.2.1 混淆矩阵	177
9.2.2 准确度	178
9.2.3 精确度与召回度	180
9.2.4 F1 指标	181
9.2.5 误判成本	181
9.2.6 受试者工作特征 (ROC) 曲线及相关指标	182
9.3 回归问题的性能指标	185
9.3.1 MSE 及其变种 RMSE	186
9.3.2 SSE	187
9.3.3 其他指标	187
9.4 K 折交叉验证 (KFCV)	188
9.4.1 在 Julia 中应用 KFCV	189
9.4.2 KFCV 小提示	189

9.5 小结	190
9.6 思考题	192
第 10 章 无监督式机器学习	193
10.1 无监督式学习基础知识	193
10.1.1 聚类的类型	194
10.1.2 距离的度量	195
10.2 使用 K-均值算法分组数据	196
10.2.1 使用 Julia 实现 K-均值聚类	197
10.2.2 对 K-均值算法的使用建议	198
10.3 密度和 DBSCAN 算法	199
10.3.1 DBSCAN 算法	199
10.3.2 在 Julia 中应用 DBSCAN	200
10.4 层次聚类	201
10.4.1 在 Julia 中使用层次聚类	201
10.4.2 何时使用层次聚类	203
10.5 聚类的验证方式	203
10.5.1 Silhouettes	203
10.5.2 关于聚类验证的一些建议	204
10.6 关于有效进行聚类的一些建议	204
10.6.1 处理高维数据	205
10.6.2 标准化	205
10.6.3 可视化建议	205
10.7 小结	206
10.8 思考题	207
第 11 章 监督式机器学习	209
11.1 决策树	210
11.1.1 在 Julia 中使用决策树	211
11.1.2 关于决策树的一些建议	214

11.2 回归树	214
11.2.1 在 Julia 中实现回归树	215
11.2.2 关于回归树的一些建议	216
11.3 随机森林	216
11.3.1 在 Julia 中使用随机森林进行分类	216
11.3.2 在 Julia 中使用随机森林进行回归	218
11.3.3 关于随机森林的一些建议	219
11.4 基本神经网络	220
11.4.1 在 Julia 中使用神经网络	221
11.4.2 关于神经网络的一些建议	223
11.5 极限学习机	224
11.5.1 在 Julia 中使用 ELM	224
11.5.2 关于 ELM 的一些建议	226
11.6 用于回归分析的统计模型	227
11.6.1 在 Julia 中使用统计回归	227
11.6.2 关于统计回归的一些建议	230
11.7 其他监督式学习系统	230
11.7.1 提升树	230
11.7.2 支持向量机	230
11.7.3 直推式系统	231
11.7.4 深度学习系统	232
11.7.5 贝叶斯网络	232
11.8 小结	233
11.9 本章思考题	235
第 12 章 图分析	236
12.1 图的重要性	237
12.2 定制数据集	239
12.3 图的统计量	240

12.4 环的检测	242
用 Julia 检测环	243
12.5 连通子图	244
12.6 团	245
12.7 图的最短路径	246
12.8 最小生成树	248
12.8.1 在 Julia 中实现 MST	249
12.8.2 用文件保存和加载图	250
12.9 Julia 在图分析中的作用	251
12.10 小结	252
12.11 思考题	254
第 13 章 更上一层楼	255
13.1 Julia 社区	255
13.1.1 与其他 Julia 用户进行交流	255
13.1.2 代码库	256
13.1.3 视频文件	256
13.1.4 新闻	257
13.2 学以致用	257
13.2.1 从这些特征开始	258
13.2.2 关于这个项目的一些思考	259
13.3 在数据科学中使用 Julia 的最后思考	260
13.3.1 不断提高 Julia 编程水平	260
13.3.2 贡献 Julia 项目	261
13.3.3 Julia 在数据科学中的未来	262
附录 A 下载安装 Julia 与 IJulia	264
附录 B 与 Julia 相关的一些常用站点	266
附录 C 本书所用的扩展包	268
附录 D Julia 与其他平台的集成	269