

《数据科学》

图书基本信息

书名：《数据科学》

13位ISBN编号：978711152926X

出版时间：2016-4-1

作者：尼娜·朱梅尔 (Nina Zumel), 约翰·芒特 (John Mount)

页数：321

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《数据科学》

内容概要

编辑推荐

《数据科学:理论、方法与R语言实践》从实用的角度较为全面地展现了数据科学的主要内容。并结合大量的实际项目案例，利用R语言详细地讲解了数据项目的开发过程和关键技术。《数据科学:理论、方法与R语言实践》适合作为高等院校高年级本科生和研究生及从事数据管理与分析的工程技术人员的主要参考书。

名人推荐

本书是所有数据科学家都应该拥有的一部独特、举足轻重的书籍。

——引自Jim Porzak的序言，Bay Area R Users Group联合创始人

覆盖了端到端的全部过程，从数据探索到建模再到交付结果。

——Nezih Yigitbasi，Intel公司

对志向高远的年轻数据科学家和经验丰富的数据科学家而言，本书充满了有用的宝石。

——Fred Rahmanian，西门子医疗

使用真实的示例进行数据分析，强烈推荐。

——Kostas Passadis博士，IPTO

作者简介

作者：（美国）尼娜·朱梅尔（Nina Zumel） 约翰·芒特（John Mount） 译者：于戈 鲍玉斌 王大玲
尼娜·朱梅尔（Nina Zumel），现在是Win—Vector LLC的首席顾问。她曾是SRI International（SRI International是一个独立的非盈利研究机构）的科学家，及一家定价优化公司的首席科学家，并创办了一家合同研究公司。
约翰·芒特（John Mount），现在是Win—Vector LLC的首席顾问。他曾是生物技术领域的计算科学家和股票交易算法的设计者，并且在Shopping.com领导一个研究团队。

书籍目录

译者序

序言

前言

第一部分数据科学引论

第1章数据科学处理过程2

1.1数据科学项目中的角色2

1.2数据科学项目的阶段4

1.2.1制定目标5

1.2.2收集和管理数据5

1.2.3建立模型7

1.2.4模型评价和批判8

1.2.5展现和编制文档9

1.2.6模型部署和维护10

1.3设定预期11

1.4小结12

第2章向R加载数据14

2.1运用文件中的数据14

2.1.1在源自文件或URL的良结构数据上使用R15

2.1.2在欠结构数据上使用R17

2.2在关系数据库上使用R19

2.2.1一个生产规模的示例20

2.2.2从数据库向R系统加载数据23

2.2.3处理PUMS数据25

2.3小结28

第3章探索数据29

3.1使用概要统计方法发现问题30

3.2用图形和可视化方法发现问题34

3.2.1可视化检测单变量的分布35

3.2.2可视化检测两个变量间的关系42

3.3小结51

第4章管理数据52

4.1清洗数据52

4.1.1处理缺失值52

4.1.2数据转换56

4.2为建模和验证采样61

4.2.1测试集和训练集的划分61

4.2.2创建一个样本组列62

4.2.3记录分组63

4.2.4数据溯源63

4.3小结63

第二部分建模方法

第5章选择和评价模型66

5.1将业务问题映射到机器学习任务67

5.1.1解决分类问题67

5.1.2解决打分问题68

5.1.3目标未知情况下的处理69

5.1.4问题到方法的映射71

- 5.2模型评价71
 - 5.2.1分类模型的评价72
 - 5.2.2打分模型的评价76
 - 5.2.3概率模型的评价78
 - 5.2.4排名模型的评价82
 - 5.2.5聚类模型的评价82
- 5.3模型验证84
 - 5.3.1常见的模型问题的识别84
 - 5.3.2模型可靠性的量化85
 - 5.3.3模型质量的保证86
- 5.4小结88
- 第6章记忆化方法89
 - 6.1KDD和KDD Cup 200989
 - 6.2构建单变量模型91
 - 6.2.1使用类别型特征92
 - 6.2.2使用数值型特征94
 - 6.2.3使用交叉验证估计过拟合的影响96
 - 6.3构建多变量模型97
 - 6.3.1变量选择97
 - 6.3.2使用决策树99
 - 6.3.3使用最近邻方法102
 - 6.3.4使用朴素贝叶斯105
 - 6.4小结108
- 第7章线性回归与逻辑斯谛回归110
 - 7.1使用线性回归110
 - 7.1.1理解线性回归110
 - 7.1.2构建线性回归模型113
 - 7.1.3预测114
 - 7.1.4发现关系并抽取建议117
 - 7.1.5解读模型概要并刻画系数质量118
 - 7.1.6线性回归要点122
 - 7.2使用逻辑斯谛回归123
 - 7.2.1理解逻辑斯谛回归123
 - 7.2.2构建逻辑斯谛回归模型124
 - 7.2.3预测125
 - 7.2.4从逻辑斯谛回归模型中发现关系并抽取建议129
 - 7.2.5解读模型概要并刻画系数130
 - 7.2.6逻辑斯谛回归要点136
 - 7.3小结137
- 第8章无监督方法138
 - 8.1聚类分析138
 - 8.1.1距离139
 - 8.1.2准备数据140
 - 8.1.3使用hclust () 进行层次聚类142
 - 8.1.4k—均值算法150
 - 8.1.5分派新的点到簇154
 - 8.1.6聚类要点156
 - 8.2关联规则156
 - 8.2.1关联规则概述156

- 8.2.2问题举例157
- 8.2.3使用arules程序包挖掘关联规则158
- 8.2.4关联规则要点165
- 8.3小结165
- 第9章高级方法探索166
 - 9.1使用bagging和随机森林方法减少训练方差167
 - 9.1.1使用bagging方法改进预测167
 - 9.1.2使用随机森林方法进一步改进预测170
 - 9.1.3bagging和随机森林方法要点173
 - 9.2使用广义加性模型学习非单调关系173
 - 9.2.1理解GAM174
 - 9.2.2一维回归示例174
 - 9.2.3提取非线性关系178
 - 9.2.4在真实数据上使用GAM179
 - 9.2.5使用GAM实现逻辑斯谛回归182
 - 9.2.6GAM要点183
 - 9.3使用核方法提高数据可分性183
 - 9.3.1理解核函数184
 - 9.3.2在问题中使用显式核函数187
 - 9.3.3核方法要点190
 - 9.4使用SVM对复杂的决策边界建模190
 - 9.4.1理解支持向量机190
 - 9.4.2在人工示例数据中使用SVM192
 - 9.4.3在真实数据中使用SVM195
 - 9.4.4支持向量机要点197
 - 9.5小结197
- 第三部分结果交付
- 第10章文档编制和部署200
 - 10.1buzz数据集200
 - 10.2使用knitr产生里程碑文档202
 - 10.2.1knitr是什么202
 - 10.2.2knitr技术详解204
 - 10.2.3使用knitr编写buzz数据文档205
 - 10.3在运行时文档编制中使用注释和版本控制208
 - 10.3.1编写有效注释208
 - 10.3.2使用版本控制记录历史209
 - 10.3.3使用版本控制探索项目213
 - 10.3.4使用版本控制分享工作217
 - 10.4模型部署220
 - 10.4.1将模型部署为RHTTP服务220
 - 10.4.2按照输出部署模型222
 - 10.4.3要点223
 - 10.5小结224
- 第11章有效的结果展现226
 - 11.1将结果展现给项目出资方227
 - 11.1.1概述项目目标228
 - 11.1.2陈述项目结果229
 - 11.1.3补充细节230
 - 11.1.4提出建议并讨论未来工作231

11.1.5	向项目出资方展现的要点	232
11.2	向最终用户展现模型	232
11.2.1	概述项目目标	232
11.2.2	展现模型如何融入用户的工作流程	233
11.2.3	展现如何使用模型	235
11.2.4	向最终用户展现的要点	236
11.3	向其他数据科学家展现你的工作	236
11.3.1	介绍问题	236
11.3.2	讨论相关工作	237
11.3.3	讨论你的方法	238
11.3.4	讨论结果和未来工作	239
11.3.5	向其他数据科学家展现的要点	240
11.4	小结	240
附录A	使用R和其他工具	241
附录B	重要的统计学概念	263
附录C	更多的工具和值得探索的思路	292
	参考文献	297
	索引	299

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu000.com