

数据科学实战



[数据科学实战_下载链接1](#)

著者:[美] Rachel Schutt

出版者:人民邮电出版社

出版时间:2015-3

装帧:平装

isbn:9787115383495

- 统计推断、探索性数据分析（EDA）及数据科学工作流程
- 算法

- 垃圾邮件过滤、朴素贝叶斯和数据清理
- 逻辑回归
- 金融建模
- 推荐引擎和因果关系
- 数据可视化
- 社交网络与数据新闻
- 数据工程、MapReduce、Pregel和Hadoop

作者介绍:

作者简介:

Rachel Schutt

美国新闻集团旗下数据科学部门高级副总裁、哥伦比亚大学统计系兼职教授、约翰逊实验室高级研究科学家，同时也是哥伦比亚大学数据科学及工程研究所教育委员会的发起人之一。她曾在谷歌研究院工作数年，负责设计算法原型并通过建模理解用户行为。

Cathy O'Neil

约翰逊实验室高级数据科学家、哈佛大学数学博士、麻省理工学院数学系博士后、巴纳德学院教授，曾发表过大量算术代数几何方面的论文。他曾在著名的全球投资管理公司D.E.

Shaw担任对冲基金金融师，后加入专门评估银行和对冲基金风险的软件公司RiskMetrics，个人博客：mathbabe.org。

译者简介:

冯凌秉

澳大利亚国立大学统计学博士，本科和研究生分别毕业于中南财经政法大学和中国人民大学。现在，他任职于江西财经大学金融管理国际研究院，任讲师、硕士生导师，研究方向为应用统计与金融计量。

王群锋

毕业于西安电子科技大学，现任职于IBM西安研发中心，从事下一代统计预测软件的开发运维工作。

目录: 作者介绍 XII

关于封面图 XIII

前言 XIV

第1章 简介：什么是数据科学 1

1.1 大数据和数据科学的喧嚣 1

1.2 冲出迷雾 2

1.3 为什么是现在 3

1.4 数据科学的现状和历史	5
1.5 数据科学的知识结构	8
1.6 思维实验：元定义	10
1.7 什么是数据科学家	11
1.7.1 学术界对数据科学家的定义	12
1.7.2 工业界对数据科学家的定义	12
第2章 统计推断、探索性数据分析和数据科学工作流程	14
2.1 大数据时代的统计学思考	14
2.1.1 统计推断	15
2.1.2 总体和样本	16
2.1.3 大数据的总体和样本	17
2.1.4 大数据意味着大胆的假设	19
2.1.5 建模	21
2.2 探索性数据分析	26
2.2.1 探索性数据分析的哲学	27
2.2.2 练习：探索性数据分析	29
2.3 数据科学的工作流程	31
2.4 思维实验：如何模拟混沌	34
2.5 案例学习：RealDirect	35
2.5.1 RealDirect是如何赚钱的	36
2.5.2 练一练：RealDirect公司的数据策略	36
第3章 算法	39
3.1 机器学习算法	40
3.2 三大基本算法	41
3.2.1 线性回归模型	42
3.2.2 k 近邻模型 (k-NN)	55
3.2.3 k 均值算法	64
3.3 练习：机器学习算法基础	68
3.4 总结	72
3.5 思维实验：关于统计学家的自动化	73
第4章 垃圾邮件过滤器、朴素贝叶斯与数据清理	74
4.1 思维实验：从实例中学习	74
4.1.1 线性回归为何不适用	75
4.1.2 k 近邻效果如何	77
4.2 朴素贝叶斯模型	78
4.2.1 贝叶斯法则	79
4.2.2 个别单词的过滤器	80
4.2.3 直通朴素贝叶斯	82
4.3 拉普拉斯平滑法	83
4.4 对比朴素贝叶斯和k 近邻	85
4.5 Bash代码示例	85
4.6 网页抓取：API和其他工具	87
4.7 Jake的练习题：文章分类问题中的朴素贝叶斯模型	88
第5章 逻辑回归	92
5.1 思维实验	93
5.2 分类器	94
5.2.1 运行时间	95
5.2.2 你自己	95
5.2.3 模型的可解释性	95
5.2.4 可扩展性	96
5.3 逻辑回归：一个来自M6D 的真实案例研究	96
5.3.1 点击模型	96
5.3.2 模型背后	97
5.3.3 α 和 β 的参数估计	99

5.3.4 牛顿法	101
5.3.5 随机梯度下降法	101
5.3.6 操练	101
5.3.7 模型评价	102
5.4 练习题	105
第6章 时间戳数据与金融建模	110
6.1 Kyle Teague与GetGlue公司	110
6.2 时间戳	112
6.2.1 探索性数据分析 (EDA)	113
6.2.2 指标和新变量	117
6.2.3 下一步怎么做	117
6.3 轮到Cathy O'Neill了	118
6.4 思维实验	118
6.5 金融建模	119
6.5.1 样本期内外以及因果关系	120
6.5.2 金融数据处理	121
6.5.3 对数收益率	123
6.5.4 实例：标准普尔指数	124
6.5.5 如何衡量波动率	126
6.5.6 指数平滑法	128
6.5.7 金融模型的反馈	128
6.5.8 聊聊回归模型	130
6.5.9 先验信息量	130
6.5.10 一个小例子	131
6.6 练习：GetGlue提供的时间戳数据	134
第7章 从数据到结论	136
7.1 William Cukierski	136
7.1.1 背景介绍：数据科学竞赛	136
7.1.2 背景介绍：众包模式	137
7.2 Kaggle模式	139
7.2.1 Kaggle的参赛者	140
7.2.2 Kaggle的客户	141
7.3 思维实验：关于作业自动评分系统	143
7.4 特征选择	145
7.4.1 例子：留住用户	146
7.4.2 过滤型	149
7.4.3 包装型	149
7.4.4 决策树与嵌入型变量选择	151
7.4.5 熵	153
7.4.6 决策树算法	155
7.4.7 如何在决策树模型中处理连续性变量	156
7.4.8 随机森林	157
7.4.9 用户黏性：模型的预测能力与可解释性	159
7.5 David Huffaker：谷歌社会学研究的新方法	160
7.5.1 从描述性统计到预测模型	161
7.5.2 谷歌的社交研究	163
7.5.3 隐私保护	163
7.5.4 思维实验：如何消除用户的顾虑	164
第8章 构建面向大量用户的推荐引擎	165
8.1 一个真实的推荐引擎	166
8.1.1 最近邻算法回顾	167
8.1.2 最近邻模型的已知问题	168
8.1.3 超越近邻模型：基于机器学习的分类模型	169
8.1.4 高维度问题	171

8.1.5 奇异值分解 (SVD)	172
8.1.6 关于SVD的重要特性	172
8.1.7 主成分分析 (PCA)	173
8.1.8 交替最小二乘法	174
8.1.9 固定矩阵V, 更新矩阵U	175
8.1.10 关于这些算法的一点思考	176
8.2 思维实验: 如何过滤模型中的泡沫	176
8.3 练习: 搭建自己的推荐系统	176
第9章 数据可视化与欺诈侦测	179
9.1 数据可视化的历史	179
9.1.1 Gabriel Tarde	180
9.1.2 Mark 的思维实验	181
9.2 到底什么是数据科学	181
9.2.1 Processing	182
9.2.2 Franco Moretti	182
9.3 一个数据可视化的方案实例	183
9.4 Mark 的数据可视化项目	186
9.4.1 《纽约时报》大厅里的可视化: Moveable Type	186
9.4.2 屏幕上的生命: Cascade可视化项目	188
9.4.3 Cronkite广场项目	189
9.4.4 eBay与图书网购	190
9.4.5 公共剧场里的“莎士比亚机”	192
9.4.6 这些展览的目的是什么	193
9.5 数据科学和风险	193
9.5.1 关于Square公司	194
9.5.2 支付风险	194
9.5.3 模型效果的评估问题	197
9.5.4 建模小贴士	200
9.6 数据可视化在Square	203
9.7 Ian的思维实验	204
9.8 关于数据可视化	204
第10章 社交网络与数据新闻学	207
10.1 Morning Analytics与社交网络	207
10.2 社交网络分析	209
10.3 关于社交网络分析的相关术语	209
10.3.1 如何衡量向心性	210
10.3.2 使用哪种向心性测度	211
10.4 思维实验	212
10.5 Morningside Analytics	212
10.6 从统计学的角度看社交网络分析	215
10.6.1 网络的表示方法与特征值向心度	215
10.6.2 随机网络的第一个例子: Erdos-Renyi模型	217
10.6.3 随机网络的第二个例子: 指数随机网络图模型	217
10.7 数据新闻学	220
10.7.1 关于数据新闻学的历史回顾	220
10.7.2 数据新闻报告的写作: 来自专家的建议	220
第11章 因果关系研究	222
11.1 相关性并不代表因果关系	223
11.1.1 对因果关系提问	223
11.1.2 干扰因子: 一个关于在线约会网站的例子	224
11.2 OK Cupid的发现	225
11.3 黄金准则: 随机化临床实验	226
11.4 A/B测试	228
11.5 退一步求其次: 关于观察性研究	229

- 11.5.1 辛普森悖论 230
- 11.5.2 鲁宾因果关系模型 231
- 11.5.3 因果关系的可视化 232
- 11.5.4 定义：因果关系 233
- 11.6 三个小建议 235
- 第12章 流行病学 236
 - 12.1 Madigan的学术背景 236
 - 12.2 思维实验 237
 - 12.3 统计学在现代 238
 - 12.4 医学文献与观察性研究 238
 - 12.5 分层法不解决干扰因子的问题 239
 - 12.6 就没有更好的办法吗 241
 - 12.7 研究性实验 (OMOP) 242
 - 12.8 最后的思维实验 246
- 第13章 从竞赛中学到的：数据泄漏和模型评价 247
 - 13.1 Claudia作为数据科学家的知识结构 247
 - 13.1.1 首席数据科学家的生活 248
 - 13.1.2 作为一名女数据科学家 248
 - 13.2 数据挖掘竞赛 249
 - 13.3 如何成为出色的建模者 250
 - 13.4 数据泄漏 250
 - 13.4.1 市场预测 251
 - 13.4.2 亚马逊案例学习：出手阔绰的顾客 251
 - 13.4.3 珠宝抽样问题 251
 - 13.4.4 IBM 客户锁定 252
 - 13.4.5 乳腺癌检测 253
 - 13.4.6 预测肺炎 253
 - 13.5 如何避免数据泄漏 254
 - 13.6 模型评价 255
 - 13.6.1 准确度重要吗 256
 - 13.6.2 概率的重要性，不是非0 即1 256
 - 13.7 如何选择算法 259
 - 13.8 最后一个例子 259
 - 13.9 临别感言 260
- 第14章 数据工程：MapReduce、Pregel、Hadoop 261
 - 14.1 关于David Crawshaw 262
 - 14.2 思维实验 262
 - 14.3 MapReduce 263
 - 14.4 单词频率问题 264
 - 14.5 其他MapReduce案例 267
 - 14.6 Pregel 268
 - 14.7 关于Josh Wills 269
 - 14.8 思维实验 269
 - 14.9 给数据科学家的话 269
 - 14.9.1 数据丰富和数据匮乏 270
 - 14.9.2 设计模型 270
 - 14.10 算算Hadoop的经济账 270
 - 14.10.1 Hadoop简介 271
 - 14.10.2 Cloudera 271
 - 14.11 Josh 的工作流程 272
 - 14.12 如何开始使用Hadoop 272
- 第15章 听听学生们怎么说 273
 - 15.1 重在过程 273
 - 15.2 不再简单 274

- 15.3 援助之手 275
- 15.4 殊途同归 277
- 15.5 逢山开路，遇水架桥 279
- 15.6 作品展示 279
- 第16章 下一代数据科学家、自大狂和职业道德 281
- 16.1 前面都讲了些什么 281
- 16.2 什么是数据科学（再问一次） 282
- 16.3 谁是下一代的数据科学家 283
 - 16.3.1 成为解决问题的人 284
 - 16.3.2 培养软技能 284
 - 16.3.3 成为提问者 285
- 16.4 做一个有道德感的数据科学家 286
- 16.5 对于职业生涯的建议 289
- • • • • [\(收起\)](#)

[数据科学实战_下载链接1](#)

标签

数据科学

数据分析

数据挖掘

机器学习

大数据

统计

计算机

数据

评论

很不错的一本湿货，翻译好的没话说，连“无厘头”都被翻出来了，很想知道原文是不是nonsense。。。以后还会翻看里面的R程序

翻译还算通顺，但是错别字太多了。里面有一些东西还是能给人启示的。

對於ML 來說，這本書講得太淺。興許對於DS 還尚可。不過最後對於data scientist 所應具備的技能，還是蠻認可的。scientist 比engineer 做好了還是技高一籌，對於研究結果，論文輸出或報告輸出擴大影響還是蠻有好處。只是本書翻譯欠佳，讀來不禁搖頭嘆息。要是要理解各算法，還是MMDS 更值得推薦。

译者也太能放飞自我了。

意外地不错，更多是思维方式。干货不少，比如欺诈侦测那一节，虽然只是大概描述了一下square的风险管理模型的架构，但是因为我自己做这一块，所以就知道所言非虚。

涵盖面很广，个别章节还是有一定难度，给出的代码可操作性不高。整本书更像是一系列课程教案的汇总，含金量还是不错的，学到了一些知识。

这本书涉及的面包含数据科学的各个层面，是相关专家客座课堂讲义的分析和总结。相对比较烧脑的是算法的原理解释与过程，从通用的分类、聚类算法的，k近邻、k均值，朴素贝叶斯，决策树，到应用在金融、推荐系统等业务领域的案例与分析，深浅适宜，不过依然对数学一般的我产生不小压力。

特别好，给了结实的框架，接下来想要探究什么就可以按图索骥了。书里给出的详细的参考书目几乎每一本都是在别的地方交叉推荐过的，含金量很高。（处在入门阶段，非数学专业）

结合一线数据科学家日常工作的一手资料，在山顶眺望全局，非工具书，也有不明觉厉之处，但对打开视野大有裨益，技能和方法岂是朝夕之间，实战嘛，纸上得来终觉浅

强调数据科学总体视角，区别于算法教材，不难但适合有一定基础后阅读，有很多大牛的经验

优秀!
基于哥大的《数据科学导论》讲义，作者想挑战的是"数据科学是不可能在大学里或者书本里学会的"这一论断。作者并未对"数据科学"下精确定义，而是通过"数据科学家在做的事"来定义"数据科学"。"师者，所以传道授业解惑也"。授人以鱼不如授人以渔。身为新兴和交叉学科，全书通过从业者本身在做的事，给出了"数据科学"的各个面向。不同学科背景的人，可以结合自身优势和兴趣点，按图索骥，找到适合自身的发展路径。"我希望他们能从中得到激励，或者学到一些有用的工具，来让这个世界变得更好，而不是更坏"。
发展初期，也即规则制定期，学的开心，玩的尽兴。"不必现在就规划好未来，走点弯路也没什么，谁知道这一路上你会发现什么呢?"

师者，所以传道授业解惑也。这本书做到了

看后依旧手无寸铁，却隐约有了去挖矿打铁器的动力。随便摘抄一段：下一代数据科学家会怎么做？1. 对一切保持怀疑态度：怀疑模型本身，模型在什么情况下会失败，如何使用，因何会被无用；2. 认识到「模型的反馈循环与潜在模型之间的博弈」

我的数据科学开蒙。

数据驱动的政治分析

什么是data science

内容比较简单。但是一些平时不会注意到的细节的讨论还是不错的。

对具体算法细节未做深入探讨，但属于名副其实的实战，值得一读，尤其推荐六九两章的部分内容。

主要指明大数据的应用领域, 偏应用范畴, 对Data Scientist来说很赞...

一翻就不明觉厉看不下去的类型，拜服

数据科学实战 下载链接1

书评

这本书蛮不错的，就是看的时候碰到一些小错误，记录如下，如果本书的编者看到了，也方便勘误。P43 第11行 “事” 改为 “是” P45 第9行 “歌” 改为 “个” P52 图3-6说明文字第2行 “直” 改为 “致” P96 正文第6行 “Emprical” 改为 “Empirical” P103 倒数第4行 “...

很喜欢此书，但首先要说这本书不是用来入门算法看的。data science的方法是各种统计学计算机方法的综合，所以所有对统计学有较好的数理基础，对各种统计推断方法或数据挖掘算法有较好理解的童鞋可以通过翻阅此书，从各个角度打开对data science的认知。如果没有很好的相关知...

Now that answering complex and compelling questions with data can make the difference in an election or a business model, data science is an attractive discipline. But how can you learn this wide-ranging, interdisciplinary field? With this book, you'll get...

我看看过了我看看过了我看看过了我看看过了我看看过了我看看过了我看看过了我看看过了我看看过了我看看过了

我看过了 我看...

[数据科学实战_下载链接1](#)