

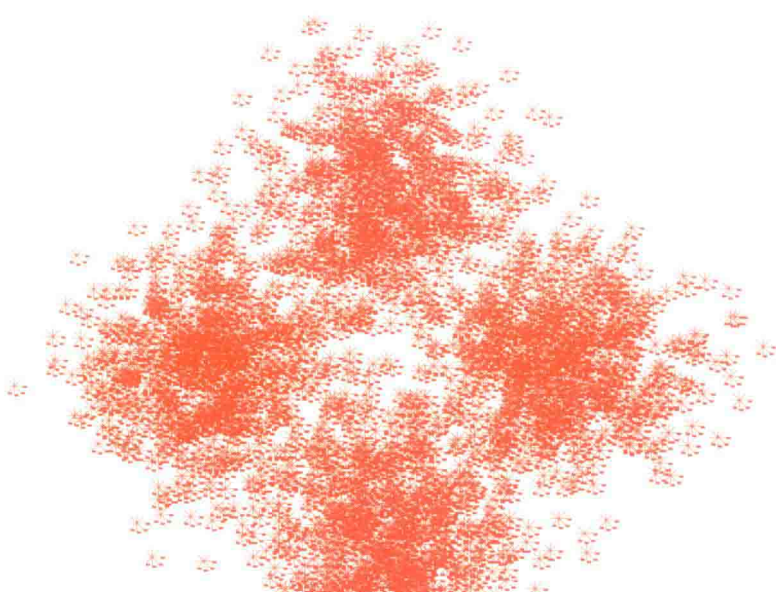
MathWorks Fellow推荐用书!

系统论述MATLAB数据挖掘的享誉全球的著作!

英文原书网站提供实例代码、工具箱、数据集、彩色图片等资源!

清华

开发者书库



MATLAB

Exploratory Data Analysis with MATLAB
Second Edition

MATLAB数据探索性分析

(原书第2版)

温迪·L. 马丁内兹 (Wendy L. Martinez)
[美] 安吉尔·R. 马丁内兹 (Angel R. Martinez) 著
杰弗瑞·L. 索卡 (Jeffrey L. Solka)

迟冬祥 黎明 赵莹 ◎ 译

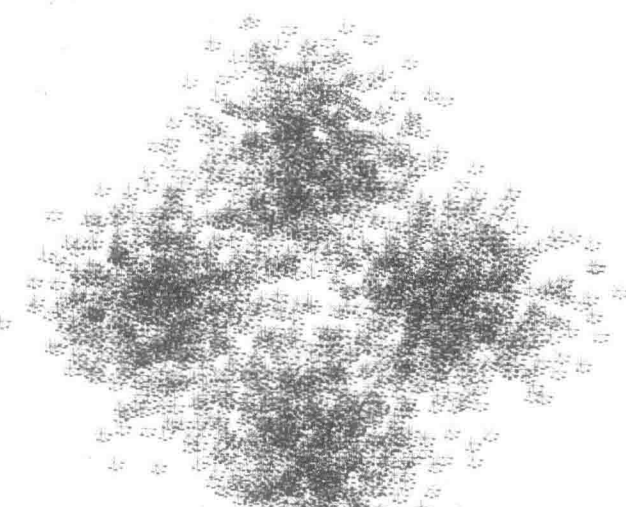
实例代码
工具箱

清华大学出版社



清华

开发者书库



MATLAB

Exploratory Data Analysis with MATLAB
Second Edition



MATLAB数据探索性分析

(原书第2版)

温迪·L. 马丁内兹 (Wendy L. Martinez)
[美] 安吉尔·R. 马丁内兹 (Angel R. Martinez) 著
杰弗瑞·L. 索卡 (Jeffrey L. Solka)

迟冬祥 黎明 赵莹 ◎ 译

清华大学出版社
北京

内 容 简 介

本书系统介绍了基于 MATLAB 语言的探索性数据分析与实现方法。本书共分 10 章,从实际数据集与探索性数据分析的基本概念讲起,内容涉及数据模式的发现、线性与非线性降维方法、数据巡查方法、聚类分析,以及用于探索性数据分析的数据可视化方法。除了基本分析与实现方法,书中也给出了丰富的应用实例,并提供了大量免费的相关资源,全部实例代码都可以直接用于探索性数据分析。

Exploratory Data Analysis with MATLAB, Second Edition/by Wendy L. Martinez, Angel R. Martinez, Jeffrey L. Solka/
ISBN: 9781439812204

Copyright© 2011 by CRC Press.

Authorized translation from English language edition published by CRC Press, part of Taylor & Francis Group LLC; All rights reserved; 本书原版由 Taylor & Francis 出版集团旗下 CRC 出版公司出版,并经其授权翻译出版。版权所有,侵权必究。

Tsinghua University Press is authorized to publish and distribute exclusively the Chinese (Simplified Characters) language edition. This edition is authorized for sale and distribution in the People's Republic of China exclusively (except Taiwan, Hong Kong SAR and Macao SAR). No part of the publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

本书中文简体翻译版授权由清华大学出版社独家出版并仅限在中国大陆地区销售。未经出版者书面许可,不得以任何方式复制或发行本书的任何部分。

Copies of this book sold without a Taylor & Francis sticker on the cover are unauthorized and illegal.

北京市版权局著作权合同登记号 图字: 01-2016-9793

本书封面贴有 Taylor & Francis 公司防伪标签,无标签者不得销售。
版权所有,侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

MATLAB 数据探索性分析: 原书第 2 版: 英文/(美)温迪·L. 马丁内兹(Wendy L. Martinez),(美)安吉尔·R. 马丁内兹(Angel R. Martinez),(美)杰弗瑞·L. 索卡(Jeffrey L. Solka)著;迟冬祥等译. —北京:清华大学出版社,2018

书名原文: Exploratory Data Analysis with MATLAB, Second Edition

(清华开发者书库)

ISBN 978-7-302-47499-9

I. ①M… II. ①温… ②安… ③杰… ④迟… III. ①Matlab 软件—高等学校—教材—英文
IV. ①TP317

中国版本图书馆 CIP 数据核字(2017)第 142134 号

责任编辑: 盛东亮

封面设计: 李召霞

责任校对: 时翠兰

责任印制: 丛怀宇

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座

社 总 机: 010-62770175

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

课件下载: <http://www.tup.com.cn>, 010-62795954

印 装 者: 三河市金元印装有限公司

经 销: 全国新华书店

开 本: 186mm×240mm 印 张: 23

版 次: 2018 年 9 月第 1 版

定 价: 79.00 元

邮 编: 100084

邮 购: 010-62786544

字 数: 511 千字

印 次: 2018 年 9 月第 1 次印刷

产品编号: 070444-01

译者序

FOREWORD

简单来说,探索性数据分析是连接经典的统计分析与流行的数据挖掘和机器学习的一座桥梁,它使得我们能够深刻理解数据,是数据挖掘和机器学习成功的基石,本书介绍的探索性数据分析,就是告诉我们,可以用什么方法,怎样从不同的角度分析并了解数据。

如果说有件事情能够让人不计收获而愿意付出,那一定是兴趣使然。我发现这本书,是因为我自己正在从事以数据可视化方法做探索性数据分析的工作,并且,我刚好有十几年MATLAB编程经验,这本书正是上述几个领域的交集。

本书译稿能够顺利完成,要感谢我的朋友贺瑞君和盛东亮。瑞君先生是CRC Press and Taylor & Francis Asia Pacific的执行主编(Managing Editor),他介绍我认识清华大学出版社盛东亮编辑,没有他们两人的热心帮助和专业指导,就没有这本书的中文版。

感谢我的同事兼好友黎明博士和赵莹博士。黎明博士翻译了第4章、第5章、第7章和附录A、附录B、附录C,赵莹博士翻译了第3章、第6章和附录D、附录E。他们在繁忙的工作中抽出时间,为本书译稿的完成做出重要贡献。中国科学院上海生命科学研究院的博士生岑凯对本书的1.4.2节关于“基因表达数据”中的很多内容,提出了非常专业的翻译建议,在此表示诚挚感谢。

此外,也要感谢我的妻子和女儿,我的妻子让我“逃脱”了很多家庭责任,我的女儿允许我“不理睬”她的学业。感谢她们!

译者

2018年3月

第2版前言

PREFACE

在过去几年里,EDA 领域的进步颇多,这本书也到了需要更新的时候了,尤其是在降维、聚类 and 可视化等方面,出现了很多新方法。

以下,我们列出了第 2 版中一些主要改进和增加的内容。

- 在线性降维这章增加了很多内容。新方法是非负矩阵分解和线性判别分析,也扩充了数据集本征维数的估计方法。
- 在第 3 章中,也描述了曲元分析这种非线性降维方法。曲元分析是作为自组织映射的一种改进方法提出的。
- 在数据巡查中加入了独立成分分析的内容。
- 几种新的聚类方法,包括非负矩阵分解、概率潜语义分析和谱聚类。
- 增加了关于平滑样条以及用于均匀间隔数据的快速样条方法。
- 在文中增加了几个可视化方法,包括用于二元数据的测距仪箱线图、带有边际直方图的散点图、双标图和一种叫作安德鲁图像的新方法。
- 文中很多方法可以通过图形用户界面(GUI)访问。这个免费的 EDA GUI 工具箱的详细信息列在附录 E 中。

与第 1 版类似,本书内容并未聚焦于方法的理论。确切地说,本书主要集中在 EDA 方法的运用。因此,本书并未纠缠于方法的实现和算法细节上。相反地,通过提供实例和应用,为学生和实践者提供了用于 EDA 的操作方法。

MATLAB 的示例代码、工具箱、数据集和大部分图像的彩色版本均可下载。下载站点来自 Carnegie Mellon StatLib,如下:

<http://lib.stat.cmu.edu>

或者是本书网站,如下:

<http://pi-sigma.info>

关于安装和变更信息,请参见 readme 文件。

关于 MATLAB 产品的信息,请联系:

The MathWorks, Inc.

3 Apple Hill Drive

Natick, MA, 01760-2098 USA

Tel: 508-647-7000

Fax: 508-647-7001

E-mail: info@mathworks.com

Web: www.mathworks.com

很多研究者写出了本书中所述方法的 MATLAB 代码并免费提供,对于他们的宝贵帮助,表示感谢。尤其是,作者感谢 Michael Berry 在非负矩阵分解方面的有益讨论,感谢 Ata Kaban 允许使用她的 PLSI 代码,也对 Mia Hubert 与 Sabine Verboven 授权使用他们的 bagplot 函数和非常耐心地与我们通信表示感谢。

感谢这套计算机科学和数据分析丛书的编辑们收录本书,感谢 CRC 出版社的 David Grubbs、Bob Stern 和 Michele Dimont 的帮助和耐心。一如既往地,感谢 MathWorks 公司的 Naomi Fernandes 和 Tom Lane 在 MATLAB 方面的特殊援助。

免责声明

(1) 随 EDA 工具箱提供的一些 MATLAB 函数由其他研究者编写,他们保留其著作权。在附录 B 和各自函数的帮助部分给出了参考文献。特别指出,EDA 工具箱在 GNU 协议许可下提供:

<http://www.gnu.org/copyleft/gpl.html>

(2) 本书中表达的观点来自作者,并不代表美国国防部或者其分支机构观点。

Wendy L. Martinez, Angel R. Martinez, Jeffrey L. Solka

第1版前言

PREFACE

我们的第一本书——《基于 MATLAB 的计算统计学手册》(*Computational Statistics Handbook with MATLAB*)[2002],其目的之一就是展示计算统计学的一些基本概念和方法,以及如何用 MATLAB 实现^①。计算统计学的一个核心部分就是探索性数据分析(exploratory data analysis),或称 EDA。因此,这本书可以看作是第一本书的补充,并有类似的目标——使得 EDA 技术为广大读者所用。

EDA 属于统计学和数据分析,其思路是先探索数据,常采用描述性统计学、科学可视化、数据巡查、降维等方法。这种探索没有任何预设观点或者假设。相反,这种方法使用探索的结果来引导和展开后续的假设检验和建模等。它与数据挖掘领域紧密关联,本书讨论的很多 EDA 工具是知识发现和数据挖掘工具箱的一部分。

本书旨在服务于进行原始数据分析的广大读者,包括科学家、统计学家、数据挖掘者、工程师、计算机科学家、生物统计学家、社会科学家以及其他学科工作者,也希望本书可以用于大学高年级学生或者研究生课堂教学中。每章包含的练习题目使其适合作为 EDA 课程、数据挖掘、计算统计学、机器学习等方面的课本或者补充材料。我们鼓励读者仔细看一下练习,因为有时练习中会介绍一些新的概念。练习本质上是计算性的或者探索性的,所以往往没有唯一的答案。

至于本书所需的背景,假设读者有线性代数基础。比如,应该熟悉线性代数的名词、数组乘法、矩阵逆、行列式和数组转置等,也假设读者学习过概率与统计学课程。读者应该在这门课程里了解随机变量、概率分布和密度函数、基本的描述性度量和回归等。

与第一本书类似,本书并未纠结于方法的理论。确切地说,本书的重点在于 EDA 方法的运用。方法的实现是第二位的,但只要适宜,本书为学生们和实践者展示了方法实现的算法、过程和 MATLAB 代码。很多方法是复杂的, MATLAB 的实现细节并不重要。在这些例子中,展示了如何使用函数和技巧。感兴趣的读者(或者程序员)可以查看 M 文件,获得更多信息。这样,喜欢使用其他编程语言的读者应该可以自行实现算法。

虽然本书不探究理论,但希望重申书中描述的方法都有其理论基础。因此,在各章最后,提供参考文献等资源,供那些想进一步了解理论信息的读者查阅。

^① MATLAB® 和 Handle Graphics® 是 MathWorks 公司的注册商标。

MATLAB 代码以 EDA 工具箱的形式随书提供。这包括函数、图形用户界面和书中使用的数据集。上述内容可以在以下网站下载:

<http://lib.stat.cmu.edu>

关于安装和变更信息,请参见 readme 文件。练习中包含 MATLAB 命令的 M 文件也可以下载。

本书也作了免责声明,说明本书中的 MATLAB 代码并不是最有效的方案。在很多情况下,为了(代码)清晰易懂而牺牲了效率。请参看示例的 M 文件,感谢 MathWorks 公司的 Tom Lane。

附录 B 对 EDA 工具箱做了非常详细的描述,也提供了可供下载的(免费的)其他站点信息。这里的一些工具箱和函数在本书中使用,另外的那些提供了参考信息。只要可能和适合,本书都会使用 EDA 工具箱的免费函数,使作者很容易学习示例和练习。

假设读者有 MathWorks 公司的统计工具箱(版本 4 或者更高版本)。在恰当的时候,本书会指出函数是来自 MATLAB 主程序包、统计工具箱或者 EDA 工具箱。EDA 工具箱的开发主要是基于 MATLAB 6.5 版(版本 4 的统计工具箱),所以如果你有这些条件,代码就可以正常运行。然而,在本书写作时,有新版的 MATLAB 及其统计工具箱发布,所以也整合了这个版本提供的新功能。

感谢以下校稿人的宝贵帮助: Chris Fraley、David Johannsen、Catherine Loader、Tom Lane、David Marchette 和 Jeffrey Solka。他们的很多意见和建议让本书变得更好,本书的任何不足之处由作者承担责任。特别感谢 Jeffrey Solka 在有限混合方法编程方面的协助,感谢 Richard Johnson 允许使用他的数据可视化工具箱并更新函数。也感谢所有本书所涉及方法的研究者,他们编写了 MATLAB 代码并免费提供使用。感谢计算机科学和数据分析丛书的编辑们收录本书。非常感谢 CRC 出版社的 Bob Stern、Rob Calver、Jessica Vakili 和 Andrea Demby 的帮助和耐心。最后,感谢 MathWorks 公司的 Naomi Fernandes 和 Tom Lane 在 MATLAB 方面的特殊援助。

免责声明

(1) 随 EDA 工具箱提供的一些 MATLAB 函数由其他研究者编写,他们保留其著作权。在附录 B 和各自函数的帮助部分给出了参考文献。除非特别指出,EDA 工具箱在 GNU 协议许可下提供:

<http://www.gnu.org/copyleft/gpl.html>

(2) 本书中表达的观点来自作者,并不代表美国国防部或者其分支机构观点。

Wendy L. Martinez, Angel R. Martinez

目 录

CONTENTS

译者序	I
第 2 版前言	III
第 1 版前言	V

第一部分 探索性数据分析综述

第 1 章 绪论	3
1.1 何为探索性数据分析	3
1.2 全文概述	5
1.3 关于符号表示法	6
1.4 本书使用的数据集	7
1.4.1 非结构化文本文档	7
1.4.2 基因表达数据	10
1.4.3 Oronsay 数据集	14
1.4.4 软件检测	15
1.5 数据变换	15
1.5.1 幂变换	16
1.5.2 标准化	17
1.5.3 数据球面化	17
1.6 深入阅读	19
练习	20

第二部分 模式发现的 EDA 方法

第 2 章 降维——线性方法	23
2.1 简介	23
2.2 主成分分析——PCA	25
2.2.1 基于样本协方差矩阵的 PCA	25
2.2.2 基于样本相关矩阵的 PCA	27

2.2.3	应该保留多少个维度	27
2.3	奇异值分解——SVD	31
2.4	非负矩阵分解	34
2.5	因子分析	37
2.6	Fisher 线性判别	40
2.7	本征维数	44
2.7.1	最近邻法	44
2.7.2	关联维数	47
2.7.3	最大似然法	48
2.7.4	包数估计	49
2.8	总结与深入阅读	50
练习		51
第3章	降维——非线性方法	55
3.1	多维尺度分析——MDS	55
3.1.1	度量 MDS	56
3.1.2	非度量 MDS	62
3.2	流形学习	68
3.2.1	局部线性嵌入	68
3.2.2	等距特征映射——ISOMAP	69
3.2.3	海赛特征映射	70
3.3	人工神经网络方法	73
3.3.1	自组织映射	74
3.3.2	生成式拓扑映射	76
3.3.3	曲元分析	79
3.4	总结与深入阅读	82
练习		83
第4章	数据巡查	85
4.1	总体巡查法	85
4.1.1	Torus Winding 法	86
4.1.2	伪总体巡查法	89
4.2	插值巡查法	91
4.3	投影追踪法	92
4.4	投影追踪索引	98
4.4.1	Posse 卡方索引	98
4.4.2	矩索引	99

4.5 独立成分分析	100
4.6 总结与深入阅读	103
练习	104
第5章 发现类	105
5.1 简介	105
5.2 层次聚类法	107
5.3 优化方法—— k 均值聚类	111
5.4 谱聚类	115
5.5 文本聚类	117
5.5.1 非负矩阵分解——回顾	117
5.5.2 概率潜在语义分析	121
5.6 聚类评估	124
5.6.1 Rand 索引	125
5.6.2 同型相关	127
5.6.3 上尾法	127
5.6.4 轮廓图	129
5.6.5 间隙统计	131
5.7 总结与深入阅读	135
练习	137
第6章 基于模型的聚类	140
6.1 基于模型的聚类方法概述	140
6.2 有限混合模型	142
6.2.1 多元有限混合模型	143
6.2.2 分量模型——协方差矩阵约束	144
6.3 最大期望算法	149
6.4 基于模型的层次聚合聚类	152
6.5 基于模型的聚类	154
6.6 基于模型聚类的密度估计和判决分析	158
6.6.1 模式识别介绍	158
6.6.2 贝叶斯决策理论	158
6.6.3 基于模型聚类的概率密度估计	160
6.7 由混合模型生成随机数据	163
6.8 总结与深入阅读	165
练习	167

第7章 平滑散点图	169
7.1 简介	169
7.2 loess	170
7.3 鲁棒 loess 拟合	177
7.4 loess 残差分析与诊断	179
7.4.1 残差图	180
7.4.2 散布平滑	183
7.4.3 loess 包络——向上和向下平滑	184
7.5 平滑样条及应用	185
7.5.1 样条回归	185
7.5.2 平滑样条	187
7.5.3 均匀间隔数据的平滑样条	191
7.6 选择平滑参数	193
7.7 二元分布平滑	195
7.7.1 中间平滑对	196
7.7.2 极平滑	198
7.8 曲线拟合工具箱	201
7.9 总结与深入阅读	202
练习	203

第三部分 EDA 的图形方法

第8章 聚类可视化	209
8.1 树状图	209
8.2 树图	211
8.3 矩形图	213
8.4 ReClus 图	216
8.5 数据图像	220
8.6 总结与深入阅读	223
练习	224
第9章 分布图形	226
9.1 直方图	226
9.1.1 一元直方图	226
9.1.2 二元直方图	231
9.2 箱线图	232

9.2.1	基本箱线图	233
9.2.2	基本箱线图的变形	237
9.3	分位数图	240
9.3.1	概率图	240
9.3.2	q-q 图	241
9.3.3	分位数图	245
9.4	袋状图	246
9.5	测距仪箱线图	247
9.6	总结与深入阅读	250
	练习	251
第 10 章	多元可视化	253
10.1	象形图	253
10.2	散点图	255
10.2.1	2D 和 3D 散点图	255
10.2.2	散点图矩阵	257
10.2.3	六边形分组散点图	259
10.3	动态图	260
10.3.1	识别数据	261
10.3.2	关联	263
10.3.3	笔刷	265
10.4	协同图	267
10.5	点阵图	269
10.5.1	基本点阵图	269
10.5.2	多路点阵图	270
10.6	绘点为线	274
10.6.1	平行坐标图	274
10.6.2	安德鲁曲线	276
10.6.3	安德鲁图像	278
10.6.4	其他绘图矩阵	280
10.7	再看数据巡查	282
10.7.1	总体巡查	282
10.7.2	组合巡查	283
10.8	双标图	285
10.9	总结与深入阅读	287
	练习	289

附录 A 近似度量	292
A.1 定义	292
A.1.1 相异性	293
A.1.2 相似性度量	294
A.1.3 二值数据的相似性度量	294
A.1.4 概率密度函数的相异性	295
A.2 变换	295
A.3 进阶阅读	296
附录 B EDA 相关软件资源	297
B.1 MATLAB 程序	297
B.2 其他 EDA 程序	300
B.3 EDA 工具箱	301
附录 C 数据集的描述	302
附录 D MATLAB 工具使用要点	306
D.1 MATLAB 简介	306
D.2 在 MATLAB 中获得帮助	307
D.3 文件和工作空间管理	307
D.4 MATLAB 的标点符号	308
D.5 算术运算符	309
D.6 MATLAB 的数据结构	309
D.6.1 基本数据结构	309
D.6.2 构建数组	309
D.6.3 元胞数组	310
D.6.4 结构体	311
D.7 脚本文件与函数	311
D.8 控制流	311
D.8.1 for 循环	312
D.8.2 while 循环	312
D.8.3 条件分支语句	312
D.8.4 开关语句	313
D.9 基本绘图	313
D.10 如何获取 MATLAB 信息	315

附录 E MATLAB 函数	316
E.1 MATLAB	316
E.2 统计工具箱	318
E.3 EDA 工具箱	320
E.4 EDA 图形界面工具箱	322
参考文献	332

第一部分

探索性数据分析综述



