HZ BOOKS
华章科技

机器学习领域经典著作，智能计算专家多年经验结晶，以全新的角度诠释机器学习的算法理论，通过案例系统阐述机器学习的实践方法和应用技巧，指导读者轻松步入工程应用阶段

大数据技术丛书

The Practice of Machine Learning

机器学习实践指南

案例应用解析

麦好◎著

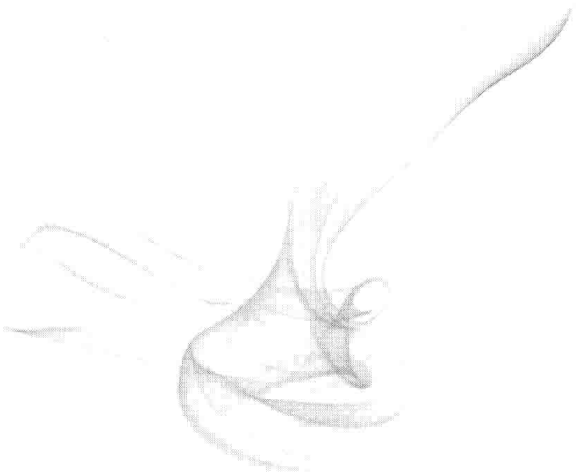
机械工业出版社
China Machine Press

大数据技术丛书

The Practice of Machine Learning

机器学习实践指南
案例应用解析

麦好◎著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习实践指南：案例应用解析/麦好著. —北京：机械工业出版社，2014.4
（大数据技术丛书）

ISBN 978-7-111-46207-1

I. 机… II. 麦… III. 机器学习—指南 IV. TP181-62

中国版本图书馆CIP数据核字（2014）第054810号

本书是机器学习及数据分析领域不可多得的一本著作，也是为数不多的既有大量实践应用案例又包含算法理论剖析的著作，作者针对机器学习算法既抽象复杂又涉及多门数学学科的特点，力求理论联系实际，始终以算法应用为主线，由浅入深以全新的角度诠释机器学习。

全书分为准备篇、基础篇、统计分析实战篇和机器学习实战篇。准备篇介绍了机器学习的发展及应用前景以及常用科学计算平台，主要包括统计分析语言 R、机器学习模块 mlpy 和 Neurolab、科学计算平台 Numpy、图像识别软件包 OpenCV、网页分析 BeautifulSoup 等软件的安装与配置。基础篇先对数学基础及其在机器学习领域的应用进行讲述，同时推荐配套学习的数学书籍，然后运用实例说明计算平台的使用，以 Python 和 R 为实现语言，重点讲解了图像算法、信息隐藏、最小二乘法拟合、因子频率分析、欧氏距离等，告诉读者如何使用计算平台完成工程应用。最后，通过大量统计分析和机器学习案例提供实践指南，首先讲解回归分析、区间分布、数据图形化、分布趋势、正态分布、分布拟合等数据分析基础，然后讲解神经网络、统计算法、欧氏距离、余弦相似度、线性与非线性回归、数据拟合、线性滤波、图像识别、人脸辨识、网页分类等机器学习算法。此书可供算法工程师、IT 专业人员以及机器学习爱好者参考使用。

机器学习实践指南：案例应用解析

麦好 著

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：陈佳媛 杨绣国
印 刷：藁城市京瑞印刷有限公司
开 本：186mm×240mm 1/16
书 号：ISBN 978-7-111-46207-1
版 次：2014年4月第1版第1次印刷
印 张：21.25（含0.25印张彩插）
定 价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066
购书热线：（010）68326294 88379649 68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

华章计算机

HZBOOKS | Computer Science and Technology

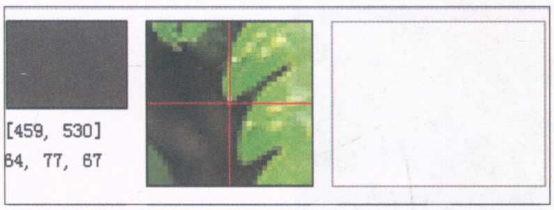


图 4-6　树叶放大的颗粒效果

图 4-9 随机产生若干像素点

图 4-10　图像变暗

图 4-11　图像变亮

图 4-12　图像日落效果

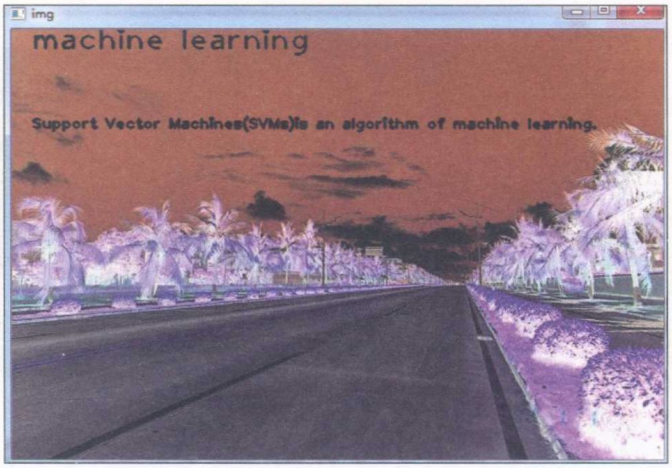


图 4-13　负片和水印效果

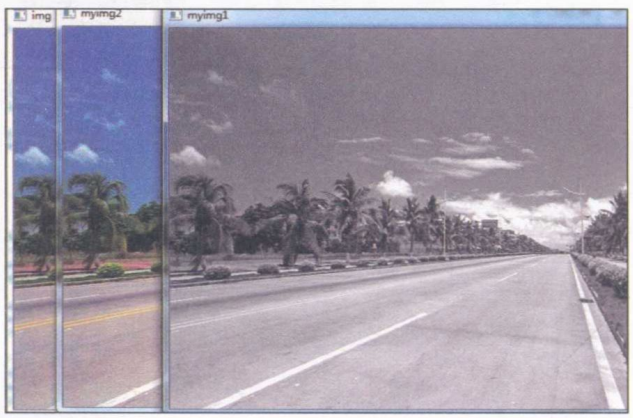


图 4-18　图像灰度化

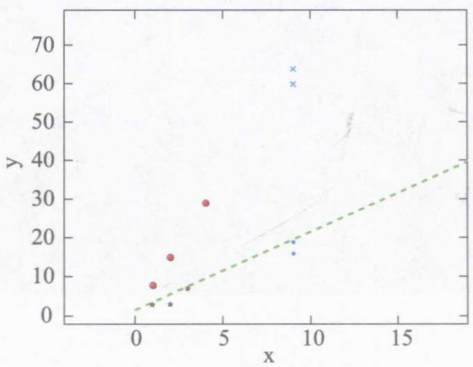


图 7-4　神经网络分类

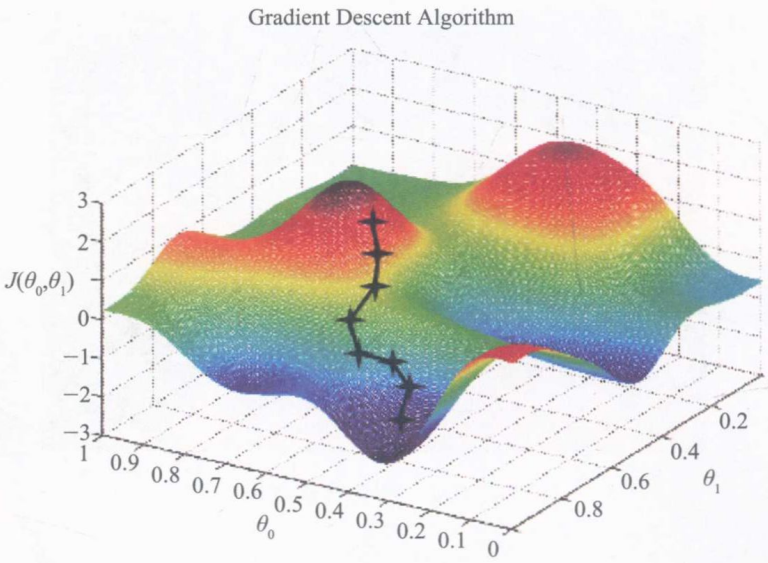


图 7-6　误差曲面及梯度下降

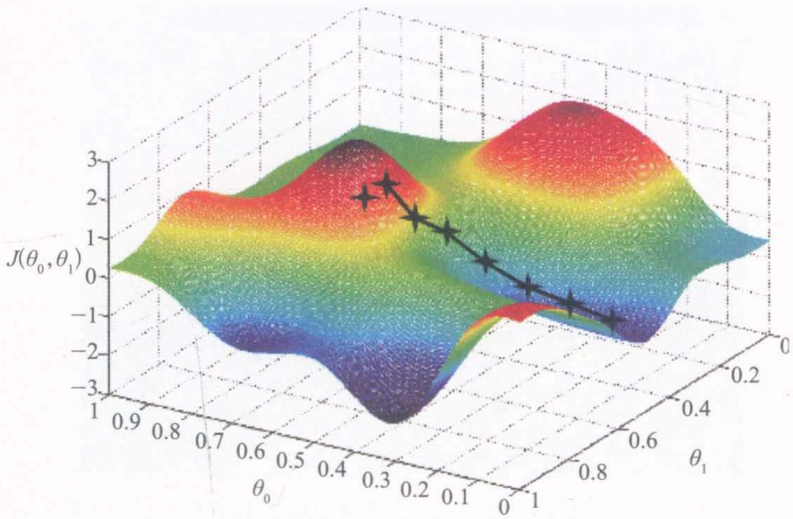


图 7-7　落到局部最小点的梯度下降

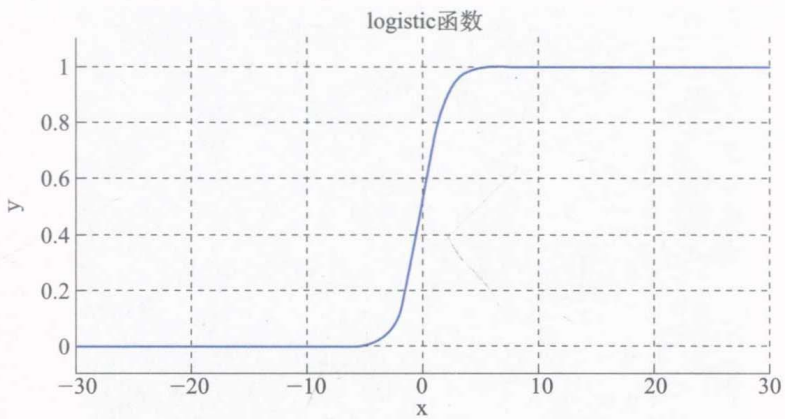


图 7-12　sigmoid 曲线

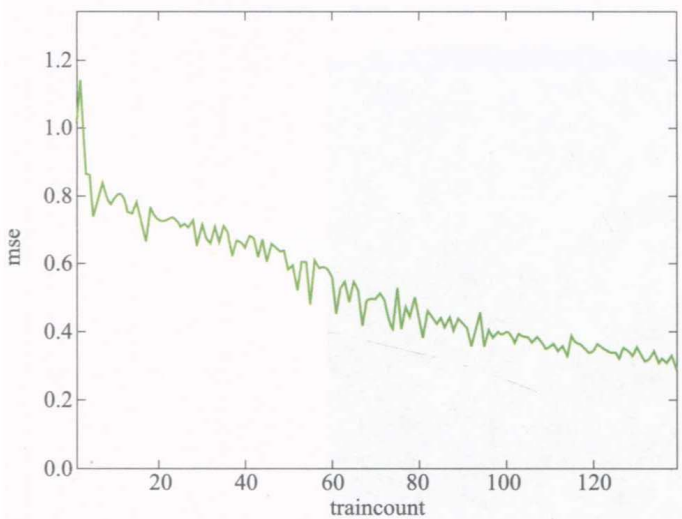


图 7-18　误差曲线

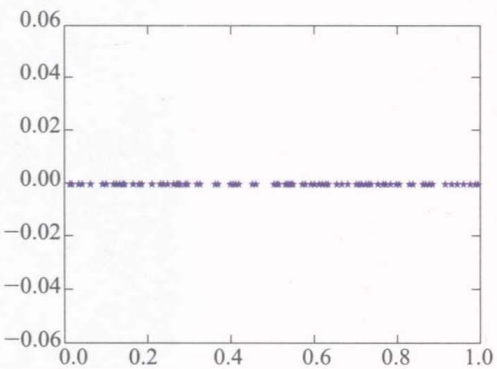


图 7-23　数据点分布

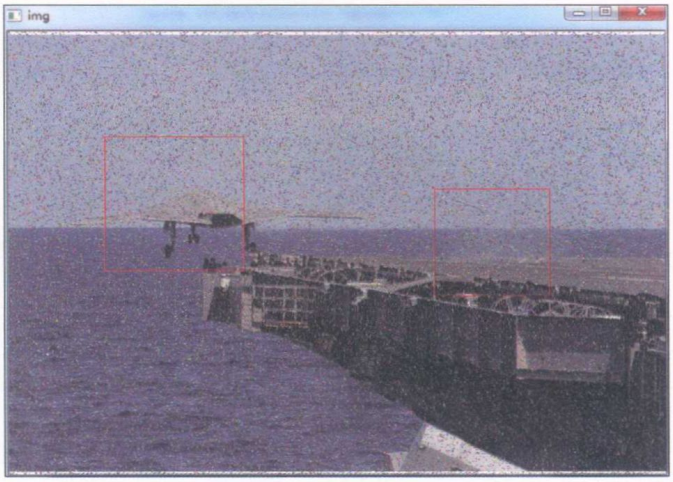


图 9-5　弱噪声切片识别效果图

图 9-6　强噪声图像

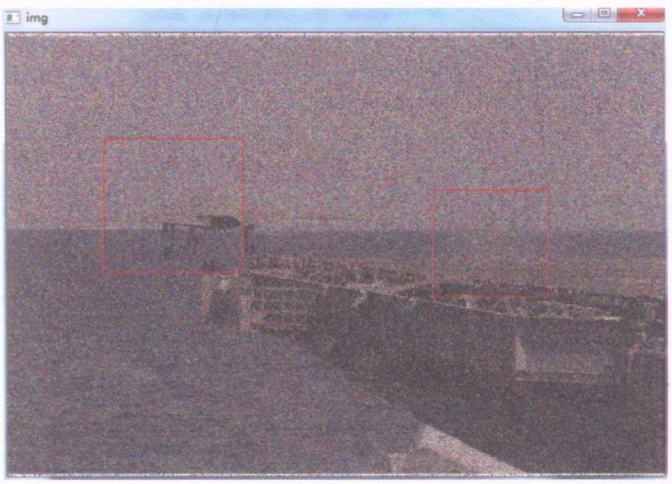


图 9-7　强噪声切片识别效果图

前　言

为什么要写这本书

自从计算机问世以来，人们就想知道，机器是否能像人类一样具有学习能力。1959年，美国的塞缪尔设计了一个下棋程序，这个程序具有学习能力，它可以在不断的对弈中提高自己的棋艺。4年后，这个程序战胜了设计者本人。又过了3年，这个程序战胜了美国一个保持常胜不败战绩8年之久的冠军。不难看出，这个程序向人们展示了机器学习的能力。如果我们理解了计算机学习的内在机制，即怎样使它们根据经验来自动提高，那么影响将是空前的。

机器学习作为一门多领域交叉学科，在近20年里异军突起。机器学习涉及概率论、统计学、代数学、微积分、算法复杂度理论等多门学科，通过设计和分析一些让计算机可以自动“学习”的算法，人类对机器学习的不断研究开辟出许多全新的应用领域，使智能机器的计算能力和可定制性上升到新的高度。

在国外，机器学习技术大量应用于军事领域，X-47B 验证机已经完成首飞，这款由诺斯罗普·格鲁曼公司为美国海军研制、外形极似B-2 战略轰炸机的飞机，是世界上第一架完全由计算机控制的“无尾翼、喷气式无人驾驶飞机”，它意味着在未来战场，将会出现无人机先出动， 打击对方的防空阵地、雷达、机场等重要目标，而有人机编队则在战场外，负责拦截对方空中支援的战斗机，这将彻底改变人类战争的方式。X-47B代表了人类在机器学习研究方面的巨大进步，是智能机器全面参与人类战争的标志，代表了人类在模仿自己智能水平方面进入了一个新的阶段，同时也给机器学习带来了全新的发展机会。

在国内，机器学习正展现出巨大的潜力，在计算机领域中扮演着日益重要的角色。机器学习的应用领域包括数据挖掘、语音识别、图像识别、机器人、生物信息学、信息安全、车辆自动驾驶、遥感信息处理、计算金融学、工业过程控制、智能家居等。在不久的将来，机器的学习能力更接近人类智能：计算机能通过学习医疗记录，获取治疗新疾病最有效的方法；住宅管理系统可分析住户的用电模式，以降低能源消耗；个人助理软件则可跟踪用户的兴趣，为其选择最感兴趣的在线信息。

随着机器学习技术在国内外的大量应用，机器学习工程师成了备受关注的人才。Google、Microsoft等公司早已经尝到了机器学习商业化带来的甜头，所以对机器学习人才提出了大量的需

求。国内很多知名的公司如阿里巴巴、淘宝等为迎接大数据时代带来的挑战，已经大量引进机器学习方面的人才。百度、搜狗等由于拥有能与Google竞争的搜索引擎，早已经开始了机器学习人才的猎取。奇虎作为中国领先的互联网软件与技术公司，其重头产品360安全卫士成为网络安全领域的领先品牌，也对引进机器学习研发工程师表现出了强烈的渴求。

现在中国已经悄然兴起了机器学习的学习热潮，掌握机器学习的工程师成为了各大IT巨头手中疯抢的“香饽饽”。机器学习成为了进入国内知名IT公司和跨国IT巨头比如Microsoft、Google的敲门砖，良好的发展势头和较高的职业薪水，吸引着越来越多的软件工程师和数据分析师涌入机器学习领域。

但是，机器学习的入门门槛较高，尤其对研究者的数学理解能力有较高要求，相对于数据结构、计算机算法以及系统架构知识来说，机器学习是一个全新的领域，也是一个全新的高度。希望本书能帮助读者进入机器学习的精彩世界。

理解机器学习算法往往要从理解它涉及的数学公式和数学知识开始，本书作者也是通过攀登数学这座大山一步步走入机器学习领域的，对此深有体会。打好数学基础是非常必要的，一旦你掌握了数学分析、线性代数、概率与统计、统计学、离散数学、抽象代数、数学建模等数学理论后，理解机器学习算法就容易多了，就不会畏惧那些让人生厌和烦琐的数学符号和数学公式，反而会喜欢上这些数学公式，并尝试亲自推导一番。

读者对象

- 开发人员。在理解机器学习算法的基础上，调用机器学习的中间库进行开发，将机器学习应用于各种场景，如数据分析、图像识别、文本分类、搜索引擎、中文智能输入法等。
- 架构师。在理解机器学习算法的基础上，适应现代云计算平台的发展，将机器学习算法应用在大规模并行计算上。同时，机器学习算法是大数据分析的基础，如神经网络、SVM、相似度分析、统计分析等技术。
- 机器学习的初、中级读者。人类对机器学习的研究只是一个开始，还远远没有结束。近年来，机器学习一直保持着强劲的发展势头，并拥有广阔的发展前景，而不同于某些软件开发领域中的程序语言或架构知识。掌握机器学习有一定的难度，属于“金领”行业，对读者来说，掌握机器学习知识就意味着更高的薪水、更具前景的职业。

如何阅读本书

全书分为准备篇、基础篇、 统计分析实战篇和机器学习实战篇。机器学习算法建立在复杂的计算理论基础之上，并涉及多门数学学科。抽象的理论加上成堆的数学公式，对部分读者来说，带来了极大的挑战，也许会将渴求学习的人们挡在门外。针对这种情况，本书力求理论联系实际，在介绍理论基础的同时，注重机器学习算法的实际运用，让读者明白其中的原理。

准备篇中首先介绍机器学习的发展及应用前景，使读者对其产生深厚的兴趣，同时也介绍目前

常用的科学计算平台和本书将用到的工程计算平台，使读者消除对机器学习的畏难心理。这些平台的使用，也降低了机器学习软件实现的难度。

基础篇将对数学知识基础、计算平台应用实例进行介绍，推荐配置学习的数学教科文档，介绍计算平台开发的基本知识，应用这些平台实现计算应用。

最后，本书将针对统计分析实战和机器学习实战两个部分帮助读者建立机器学习实战指南。还将大量应用计算平台对统计分析以及机器学习算法，并进行软件的实现和应用。本书附有效果图，使读者对机器学习的应用和理论基础有形象的理解。

勘误和支持

由于作者的水平有限，编写的时间也很仓促，书中难免会出现一些错误或者不准确的地方，有不妥之处恳请读者批评指正。您如果遇到任何问题，或有更多的宝贵意见，欢迎发送邮件至我的邮箱myhaspl@myhaspl.com，很期待能够收到您的真挚反馈。此外，本书的代码及相关资源请在华章网站（http://www.hzbook.com/）本书页面上下载。

致谢

我首先要感谢伟大的电影《机械公敌》及其主角威尔·史密斯，这位美国演员主演了《当幸福来敲门》、《拳王阿里》、《绝地战警》、《全民超人汉考克》、《黑衣人》、《机械公敌》等影片，他曾获奥斯卡奖和金球奖提名。他主演的《当幸福来敲门》让很多人理解到了幸福是什么，而《机械公敌》让我看到了人工智能的未来，我相信《机械公敌》描述的以下场景一定能在将来实现：

公元2035年，智能型机器人已被人类广泛利用，作为最好的生产工具和人类伙伴，机器人在各个领域扮演着日益重要的角色。而由于有众所周知的机器人“三大安全法则”的限制，人类对这些能够胜任各种工作且毫无怨言的伙伴充满信任，它们中的很多甚至已经成为了一个家庭的组成成员。

但是我不希望看到电影中的NS-5型机器人追杀和控制人类的场景在将来某一天上演，这将是人类的悲剧，我想这并不是人工智能学者希望看到的。也许将来有一天，人工智能技术很成熟了，机器人与人之间的关系可以作为一个社会伦理和哲学问题被大家热议，机器人也能和人类一起参与讨论自己在人类社会中的角色和定位。

我衷心感谢机械工业出版社华章公司的编辑们，由于他们的努力和远见，让我顺利地完成了全部书稿。最后我还要感谢家人的大力支持和无私奉献，正因为有他们的关心和照顾，我才有足够的时间和精力来完成本书的撰写工作。

谨以此书献给热爱机器学习技术的朋友以及喜欢威尔·史密斯的影迷。

麦好（Myhaspl）

中国，广东，2013年12月

目　录

前　言

第一部分　准备篇

第1章　机器学习发展及应用前景 …… 2

1.1　机器学习概述 …… 2

1.1.1　什么是机器学习 …… 3

1.1.2　机器学习的发展 …… 3

1.1.3　机器学习的未来 …… 4

1.2　机器学习应用前景 …… 5

1.2.1　数据分析与挖掘 …… 5

1.2.2　模式识别 …… 5

1.2.3　更广阔的领域 …… 6

1.3　小结 …… 7

第2章　科学计算平台 …… 8

2.1　科学计算软件平台概述 …… 8

2.1.1　常用的科学计算软件 …… 9

2.1.2　本书使用的工程计算平台 …… 10

2.2　计算平台的配置 …… 11

2.2.1　Numpy等Python科学计算包的安装与配置 …… 11

2.2.2　OpenCV 安装与配置 …… 13

2.2.3 mlpy 安装与配置 …… 14
2.2.4 BeautifulSoup安装与配置 …… 15
2.2.5 Neurolab安装与配置 …… 15
2.2.6 R安装与配置 …… 15
2.3 小结 …… 16

第二部分 基础篇

第3章 机器学习数学基础 …… 18
3.1 数学对我们有用吗 …… 18
3.2 机器学习需要哪些数学知识 …… 20
3.3 小结 …… 25

第4章 计算平台应用实例 …… 26
4.1 Python计算平台简介及应用实例 …… 26
4.1.1 Python语言基础 …… 26
4.1.2 Numpy库 …… 37
4.1.3 pylab、matplotlib绘图 …… 44
4.1.4 图像基础 …… 46
4.1.5 图像融合与图像镜像 …… 55
4.1.6 图像灰度化与图像加噪 …… 57
4.1.7 声音基础 …… 60
4.1.8 声音音量调节 …… 63
4.1.9 图像信息隐藏 …… 68
4.1.10 声音信息隐藏 …… 72
4.2 R语言基础 …… 78
4.2.1 基本操作 …… 78
4.2.2 向量 …… 81
4.2.3 对象集属性 …… 87
4.2.4 因子和有序因子 …… 88
4.2.5 循环语句 …… 89

4.2.6 条件语句 …… 89
4.3 R语言科学计算 …… 90
4.3.1 分类（组）统计 …… 90
4.3.2 数组与矩阵基础 …… 91
4.3.3 数组运算 …… 94
4.3.4 矩阵运算 …… 95
4.4 R语言计算实例 …… 103
4.4.1 学生数据集读写 …… 103
4.4.2 最小二乘法拟合 …… 105
4.4.3 交叉因子频率分析 …… 106
4.4.4 向量模长计算 …… 107
4.4.5 欧氏距离计算 …… 108
4.5 小结 …… 109
思考题 …… 109

第三部分 统计分析实战篇

第5章 统计分析基础 …… 112
5.1 数据分析概述 …… 112
5.2 数学基础 …… 113
5.3 回归分析 …… 118
5.3.1 单变量线性回归 …… 118
5.3.2 多元线性回归 …… 121
5.3.3 非线性回归 …… 121
5.4 数据分析基础 …… 124
5.4.1 区间频率分布 …… 124
5.4.2 数据直方图 …… 126
5.4.3 数据散点图 …… 127
5.4.4 五分位数 …… 129
5.4.5 累积分布函数 …… 130
5.4.6 核密度估计 …… 130

5.5 数据分布分析 …… 132
5.6 小结 …… 134
思考题 …… 135

第6章 统计分析案例 …… 136

6.1 数据图形化案例解析 …… 136
6.1.1 点图 …… 136
6.1.2 饼图和条形图 …… 137
6.1.3 茎叶图和箱线图 …… 138
6.2 数据分布趋势案例解析 …… 140
6.2.1 平均值 …… 140
6.2.2 加权平均值 …… 140
6.2.3 数据排序 …… 141
6.2.4 中位数 …… 142
6.2.5 极差、半极差 …… 142
6.2.6 方差 …… 143
6.2.7 标准差 …… 143
6.2.8 变异系数、样本平方和 …… 143
6.2.9 偏度系数、峰度系数 …… 144
6.3 正态分布案例解析 …… 145
6.3.1 正态分布函数 …… 145
6.3.2 峰度系数分析 …… 146
6.3.3 累积分布概率 …… 146
6.3.4 概率密度函数 …… 147
6.3.5 分位点 …… 148
6.3.6 频率直方图 …… 151
6.3.7 核概率密度与正态概率分布图 …… 151
6.3.8 正太检验与分布拟合 …… 152
6.3.9 其他分布及其拟合 …… 154
6.4 小结 …… 155
思考题 …… 155

第四部分　机器学习实战篇

第7章　机器学习算法 …… 158

7.1　神经网络 …… 158

7.1.1　Rosenblatt感知器 …… 159

7.1.2　梯度下降 …… 173

7.1.3　反向传播与多层感知器 …… 180

7.1.4　Python神经网络库 …… 199

7.2　统计算法 …… 201

7.2.1　平均值 …… 201

7.2.2　方差与标准差 …… 203

7.2.3　贝叶斯算法 …… 205

7.3　欧氏距离 …… 208

7.4　余弦相似度 …… 209

7.5　SVM …… 210

7.5.1　数学原理 …… 210

7.5.2　SMO算法 …… 212

7.5.3　算法应用 …… 212

7.6　回归算法 …… 217

7.6.1　线性代数基础 …… 217

7.6.2　最小二乘法原理 …… 218

7.6.3　线性回归 …… 219

7.6.4　多元非线性回归 …… 221

7.6.5　岭回归方法 …… 223

7.6.6　伪逆方法 …… 224

7.7　PCA降维 …… 225

7.8　小结 …… 227

思考题 …… 227

第8章　数据拟合案例 …… 228

8.1　数据拟合 …… 228

8.1.1　图像分析法 …… 228

8.1.2 神经网络拟合法 …… 240
8.2 线性滤波 …… 256
8.2.1 WAV声音文件 …… 256
8.2.2 线性滤波算法过程 …… 256
8.2.3 滤波Python实现 …… 257
8.3 小结 …… 262
思考题 …… 262

第9章 图像识别案例 …… 264

9.1 图像边缘算法 …… 264
9.1.1 数字图像基础 …… 264
9.1.2 算法描述 …… 265
9.2 图像匹配 …… 266
9.2.1 差分矩阵求和 …… 267
9.2.2 差分矩阵均值 …… 269
9.2.3 欧氏距离匹配 …… 271
9.3 图像分类 …… 277
9.3.1 余弦相似度 …… 277
9.3.2 PCA图像特征提取算法 …… 283
9.3.3 基于神经网络的图像分类 …… 284
9.3.4 基于SVM的图像分类 …… 289
9.4 人脸辨识 …… 291
9.4.1 人脸定位 …… 291
9.4.2 人脸辨识 …… 293
9.5 手写数字识别 …… 300
9.5.1 手写数字识别算法 …… 300
9.5.2 算法的Python实现 …… 301
9.6 小结 …… 303
思考题 …… 304

第10章 文本分类案例 …… 305

10.1 文本分类概述 …… 305