HZ BOOKS
华章科技

机器学习领域经典著作全新升级版，增补内容达40%以上。以算法应用为主线，以全新的角度诠释机器学习的算法理论及实际运用，透过案例系统阐述机器学习的实践方法和应用技巧

大数据技术丛书

The Practice of Machine Learning, Second Edition

机器学习实践指南

案例应用解析

第2版

麦好◎著

机械工业出版社
China Machine Press

大数据技术丛书

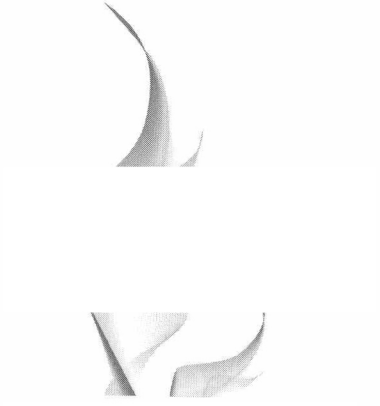
The Practice of Machine Learning, Second Edition

机器学习实践指南

案例应用解析

第2版

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 . —2 版 . —北京：机械工业出版社，2016.7
（大数据技术丛书）

ISBN 978-7-111-54021-2

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（2016）第 130361 号

机器学习实践指南：案例应用解析（第 2 版）

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：佘 洁 责任校对：殷 虹
印 刷：北京诚信伟业印刷有限公司 版 次：2016 年 7 月第 2 版第 1 次印刷
开 本：186mm×240mm 1/16 印 张：34（含 0.25 印张彩插）
书 号：ISBN 978-7-111-54021-2 定 价：89.00 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）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

图 3-6　树叶放大的颗粒效果

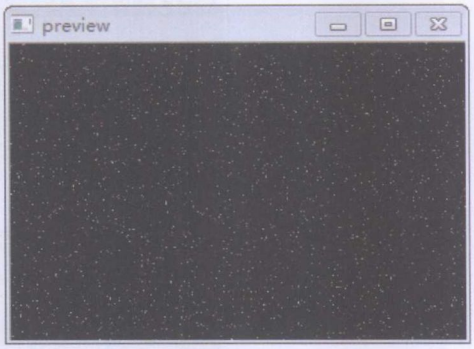


图 3-9　随机产生若干像素点

图 3-10　图像变暗

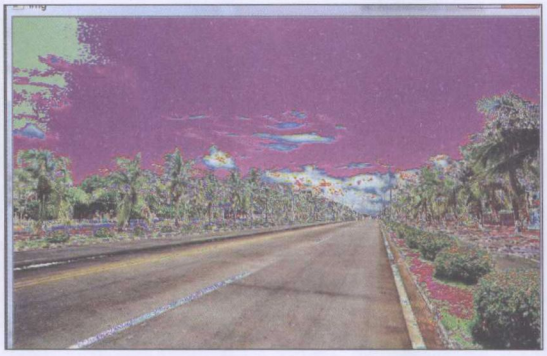


图 3-11　图像变亮

图 3-12　图像日落效果

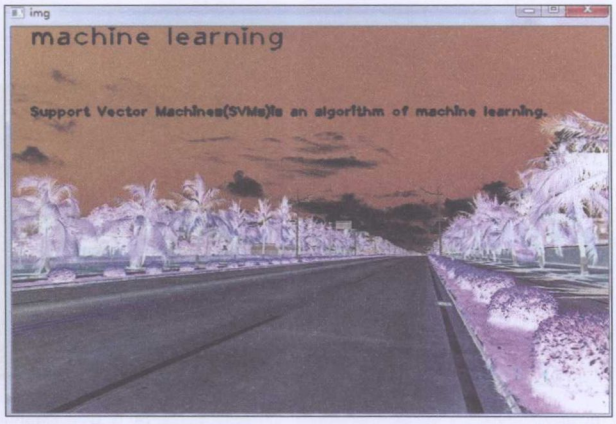


图 3-13　负片和水印效果

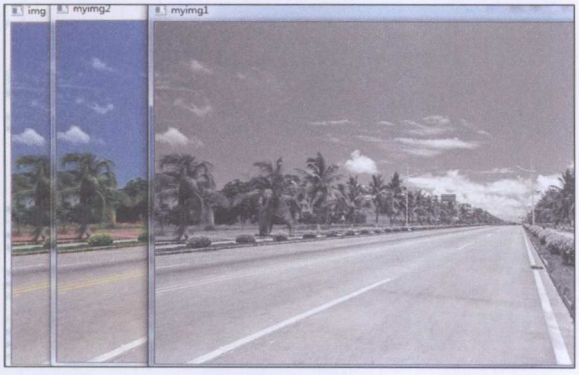


图 3-18　图像灰度化

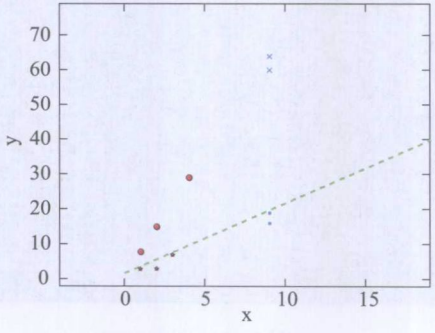


图 8-4　神经网络分类

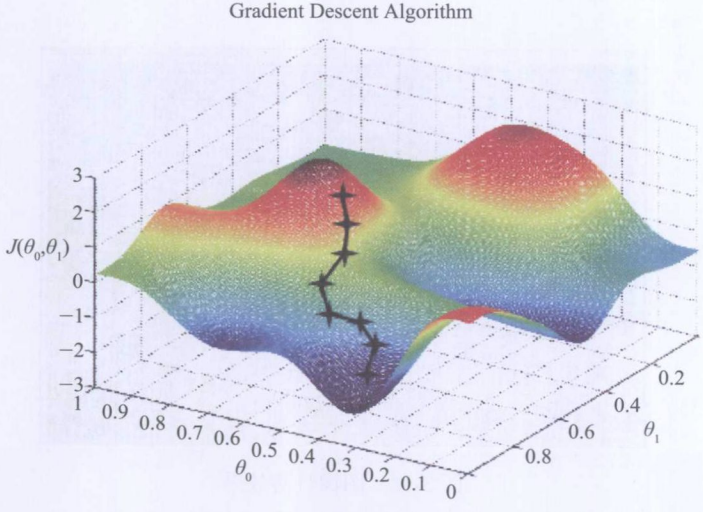


图 8-6　误差曲面及梯度下降

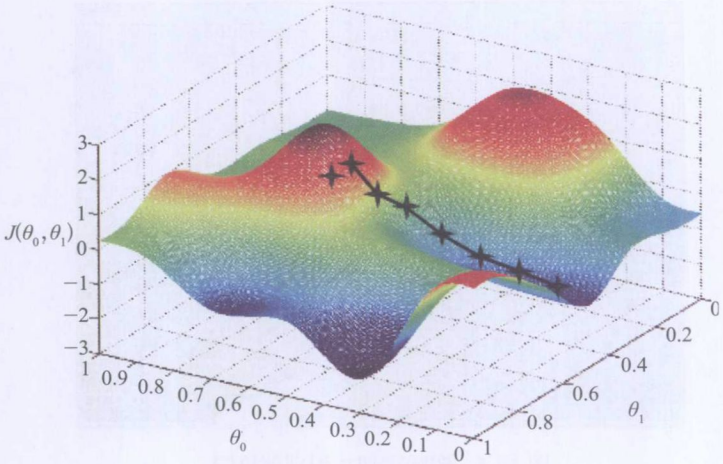


图 8-7　落到局部最小点的梯度下降

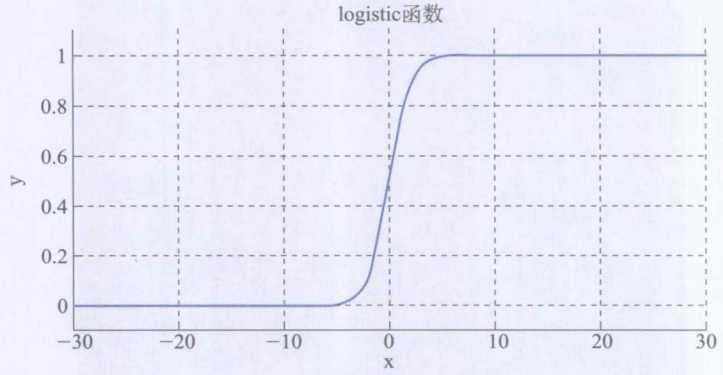


图 8-12　sigmoid 曲线

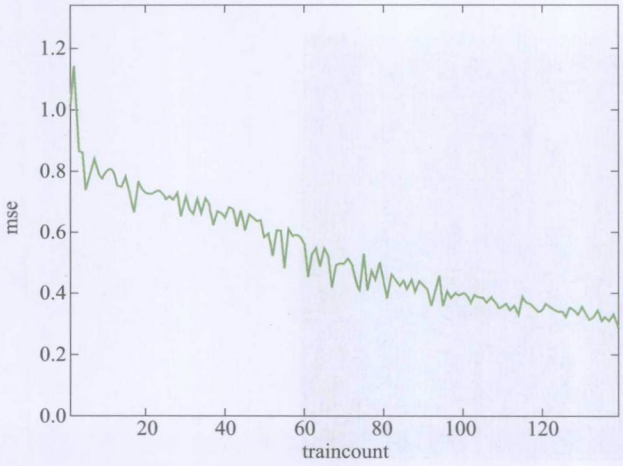


图 8-18　误差曲线

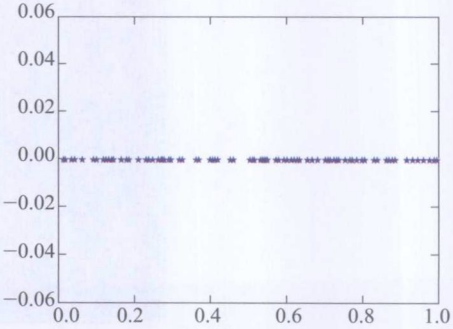


图 8-23　数据点分布

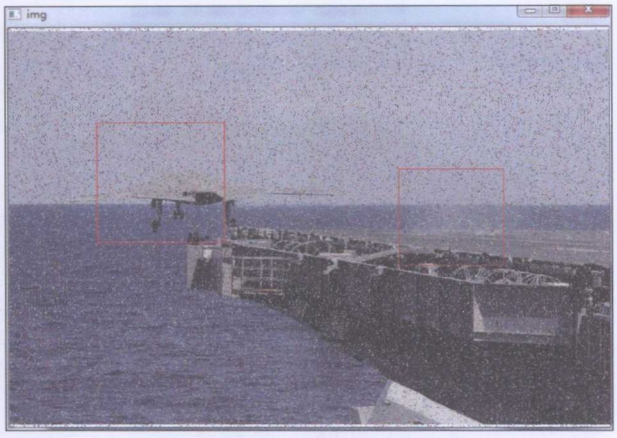


图 10-5　弱噪声切片识别效果图

图 10-6　强噪声图像

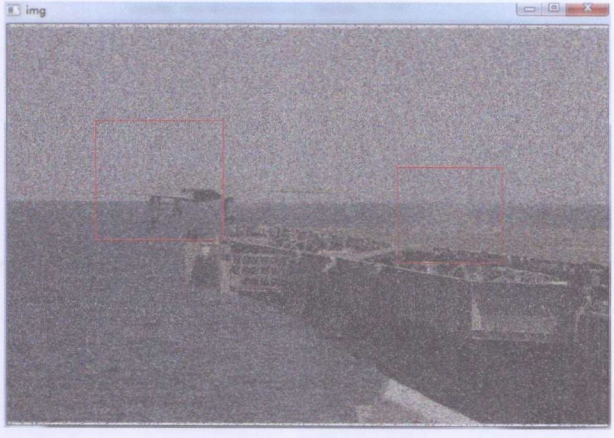


图 10-7　强噪声切片识别效果图

Forword 推 荐 序

在2013年的中国大数据创新峰会上，我偶然结识了作者，期间聊到了人工智能革命和机器学习的话题，被作者的渊博知识所折服，慢慢地结下了深厚的友谊。时间一晃而过，机器学习现在已经得到了井喷式发展，按照麻省理工学院罗德尼·布鲁克斯的预测：到2100年以前，我们的日常生活中将充满智能机器人，而且人类无法将自己同它们区分开来，我们也将是机器人，同机器人互相联系。

追忆2011年，当时我在吉林大学读研三，幸运地拿到了百度研发工程师的offer，进入百度商务搜索架构部，一直做着与凤巢广告相关的工作。现代广告业的奠基人大卫·奥格威曾经说过，除非你的广告建立在伟大的创意之上，否则它就像夜航的船，不为人所注意。广告的创意是广告的灵魂，我也一直沿着广告内容技术的方向，优化创意，提升用户的体验，提升广告主的转化。在这个方向上，我采用了机器学习的相关技术，取得了毕昇（获得2014年度百度最高奖）、图片凤巢、知识凤巢、地域识别等项目的成功，深刻地体会到了机器学习的强大，正是有了机器学习的闪闪发光，才推动了很多令人惊艳的产品的诞生。对于互联网、IT从业人员，机器学习已经成为必备利器，掌握了它，就等于站在了巨人的肩膀上工作，可帮助自己提高个人的核心竞争力。

我和作者认识近3年，同时也是《机器学习实践指南》第1版的读者，并在工作之余与作者一起管理《机器学习实践指南》的读者QQ群（群号：192029861），在群里认识了更多专注机器学习的朋友和学者。《机器学习实践指南》第1版主要针对初、中级读者，作者出书的目标就是：以机器学习算法的实践应用为主，将更多的“门外汉”带入机器学习殿堂，让更多拥有机器学习理论却无法下手的朋友掌握机器学习实践思维，轻松步入机器学习实战领域。实践思维对IT行业非常重要，一旦形成了适当的思维方式，很多工作中遇到的技术难题将迎刃而解，学习新知识的速度也更快，因为只有实践与理论相结合才能更精准地理解知识。也希望对机器学习有兴趣的读者能从中受益。

《机器学习实践指南》第 2 版出版在即，我高兴地接受了作者的邀请——为本书写推荐序。第 2 版比第 1 版增加了更多的案例和算法解析，全书详细介绍了机器学习发展及应用前景、科学计算平台、Python 计算平台应用、R 语言计算平台应用、生产环境基础、统计分析基础、描述性分析案例、假设检验与回归模型案例、神经网络、统计算法、欧氏距离与余弦相似度、SVM、回归算法、PCA 降维、关联规则、聚类与分类算法、数据拟合案例、图像算法案例、机器视觉案例、文本分类案例等机器学习实践与应用。

第 2 版致力推动机器学习理论在国内的普及和应用，为公司创建更多的商业价值；同时，力争让更多的学生、IT 工程师等进入人工智能相关领域，适应智能时代工作的需要。

最后，希望大家喜欢这本书，进而从中受益。

徐培治

百度在线网络技术（北京）有限公司

2016 年 3 月于北京

Preface 前　言

为什么要写这本书

随着全球第三次工业革命的迅猛发展，机器学习技术异军突起，人类对机器学习技术的研究也开辟出了许多全新的应用领域，这使智能机器的计算能力和可定制性上升到了一个新的层次。到了 2015 年，人类在机器学习领域取得了一系列重大的突破，这项技术已悄无声息地潜入我们的日常生活，而在未来，机器学习也将拥抱变化，持续发力。如今，它已经在各行各业的技术革新中扮演着日益重要的角色，从各方面影响和改变着我们的生活。

近年来，机器学习技术在国外得到了海量应用和深入发展。2015 年 11 月，谷歌开源了全新的 TensorFlow 机器学习系统，该系统更快、更智能，也更具有弹性。2015 年 1 月，机器学习平台 GraphLab 改名为 Dato，并获得了 1850 万美元的新融资（投资方为 Vulcan Capital 、Opus Capital 、New Enterprise Associates、Madrona Venture Group），此前他们曾获得 680 万美元的融资。2015 年 8 月，Facebook 推出了“ M”，Facebook 认为人类不仅会回答人工智能所不能回答的问题，而且从长远来看，人类也会帮助改善人工智能技术，“ M”除了能做到回答问题、查阅信息等基本功能外，还可以帮助用户完成如购买商品、餐厅定位、安排旅行计划等操作。在 2015 年 12 月召开的“2015 年神经信息处理系统”（NIPS）会议上，微软研究人员和工程师公开了 20 多篇机器学习最新研究成果的论文。此外，微软还宣布，机器学习正在成为 Windows 10 的一部分：Skype 翻译可以将口语几乎实时地翻译成其他语言，就像《星际迷航》中的通用翻译器那样，可以做到面对面的交流。Cortana 个人数字助理在与用户的互动中不断学习与改进，从而帮助用户管理日历、跟踪快递，甚至能与用户聊天和讲笑话，实现真正的个性化互动体验。Clutter 是微软 Office 2016 的成员，通过学习它可以识别出哪些电子邮件对用户来说最重要，并自动将不重要的邮件重定向到一个单独的文件夹中，从而保持用户收件箱的整洁。2015 年 9 月，美军军队医疗中心指挥官少将 Steve Jones 在美军陆军的一次

会议上发言表示，未来可以让智能机器人代替人类上战场运送伤员，美国军方甚至高调宣布：未来战场上机器人救起的可能不是人，而是机器人，因为智能机器人军团将代替人类出征。

在国内，机器学习掀起了技术革新的热潮，智能技术得到了广泛的普及和应用。隶属于中国科学院的新松机器人自动化公司生产了智能复合型机器人，这个安装了眼睛和感知器件的智能机器人，可以在车间里自由地行走并十分精确地完成任务，当其他工位人手不足时，接到指令的他还会主动上前帮忙，马上进入角色并开始工作。百度创造和完善了大规模机器学习的技术，搭建了一个能容纳万亿特征数据的、分钟级别模型更新的、高效训练的点击率预估系统；为进一步深入地发展机器学习技术，百度开始研究如何从“机器学习”到“复制人类大脑”；此外，百度甚至在 2016 年提出，百度的产品和服务都靠机器学习等技术来驱动。

随着机器学习技术在国内外的大量应用，机器学习工程师成为炙手可热的职位。现在中国已经悄然兴起了机器学习的学习热潮，掌握了机器学习技术的工程师将成为各大 IT 巨头疯抢的“香馍馍”，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的领域。国内知名的公司百度、阿里巴巴、腾讯（俗称 BAT）为迎接大数据时代带来的挑战，早已全面引进机器学习方面的人才，并有组织地对机器学习技术展开大规模的、更深入的研究。其他各大公司（包括非 IT 行业的公司）也提出了引进机器学习研发工程师的渴求。

但是，机器学习的入门门槛较高，尤其是对研究者的数学理解能力有较高的要求，相对于数据结构、算法导论中讲述的计算机算法及系统架构知识来说，机器学习是一个全新的领域，理解机器学习算法往往要从理解它所涉及的数学公式和数学知识开始，打好数学基础是非常有必要的，一旦掌握了数学分析、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会容易很多，不再畏惧那些让人生厌的、麻烦的数学符号和数学公式，说不定还会喜欢上这些数学公式，并亲自推导一番。希望本书能帮助朋友们进入机器学习的精彩世界。

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的并行计算上。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，并拥有美好的发展前景，这点不同于某些软件开发领域中的程序语言或架构知识。掌握机器学习技术有一定的难度，但也意味着，掌握机器学习的技术就能获得更高的薪水和更具前景的职业。

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，给部分读者带来了极大的挑战，将渴求学习的人们挡在了门外。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，让读者更好地明白其中的原理。

准备篇中首先将介绍机器学习的发展及应用前景，使读者产生浓厚的兴趣，同时也将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情绪，这些平台的使用也降低了机器学习软件实现的难度。

基础篇将介绍数学知识基础和计算平台应用实例，介绍计算平台的开发基本知识，并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，应用计算平台对统计分析及机器学习算法进行实现和应用，同时还会附上效果图，让读者对机器学习的基本应用和理论基础有一个形象的理解。

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不准确的地方，不妥之处恳请读者批评指正。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，很期待能够听到您的真挚反馈。此外，本书的代码及相关资源（包括思考题中涉及的数据等）的下载地址为：https://yunpan.cn/cYjhBYGLKkKTb（提取码：65ad）。

致谢

首先我要感谢伟大的电影《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》，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》描述的以下场景在将来一定能

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各个家庭的组成成员。

在此，我衷心地感谢机械工业出版社华章公司的编辑杨福川老师和策划编辑杨绣国老师，由于他们的魄力和远见，让我顺利地完成了全部书稿。最后我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，献给热爱机器学习的朋友，以及喜欢威尔 • 史密斯的影迷。

麦好（Myhaspl）

2016 年 3 月于中国广东

目　录 Contents