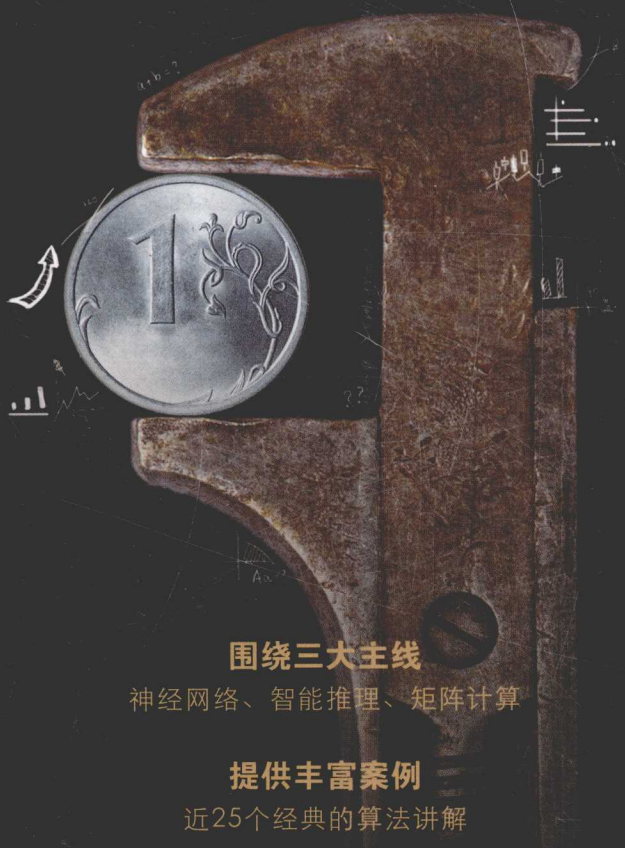
深度学习，让产品和生活具备人工智能
数据时代，让数字产生最大价值

机器学习

算法原理与编程实践

郑捷◎著



围绕三大主线

神经网络、智能推理、矩阵计算

提供丰富案例

近25个经典的算法讲解

解剖有代表性的算法库

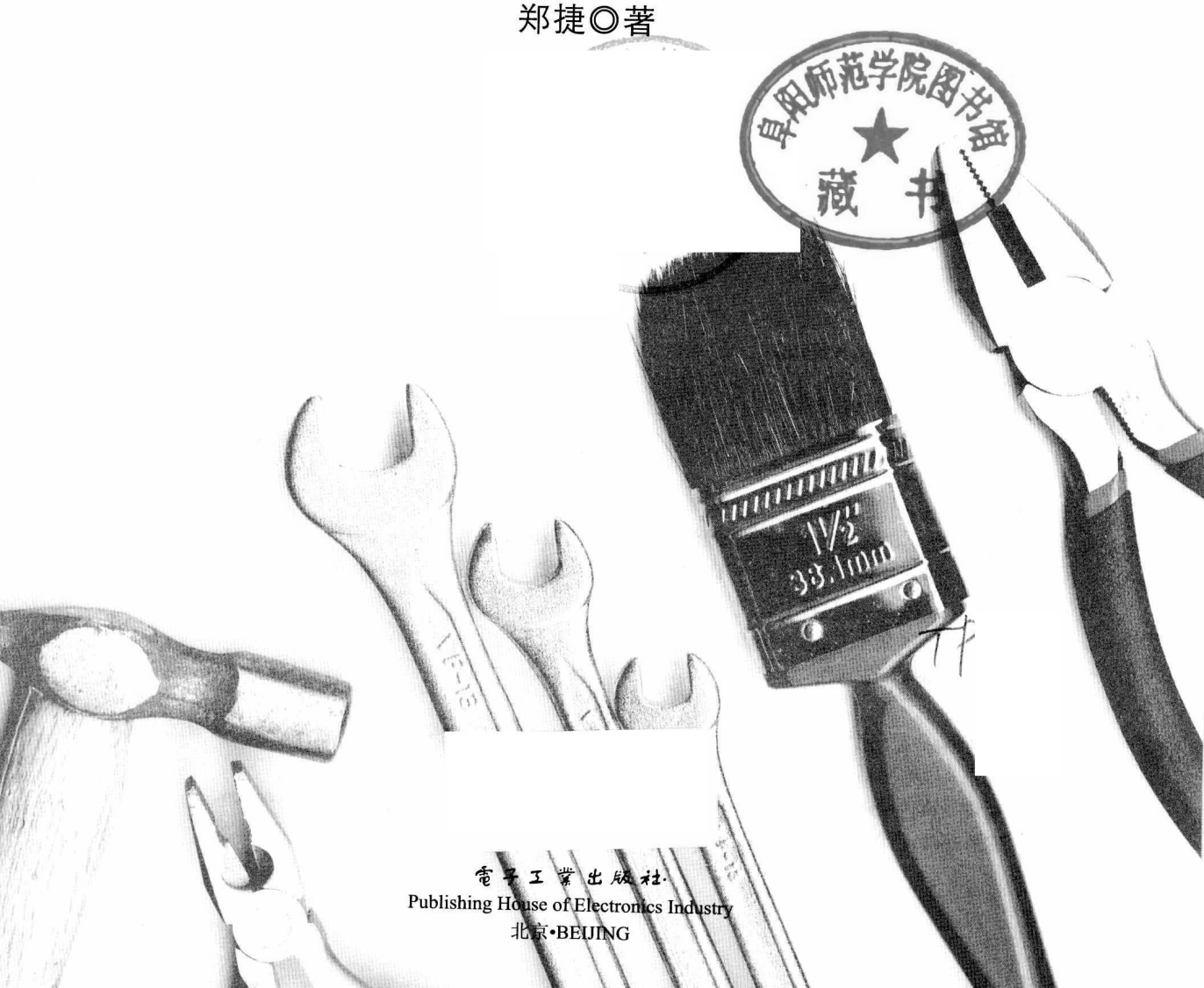
Scikit-Learn算法库、OpenCV机器视觉、
Theano深度学习库

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

机器学习

算法原理与编程实践

郑捷◎著



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书从结构上阐明了研究机器学习理论和算法的方法：最重要的不是数学，也不是这些算法本身，而是思想的发展过程，这与之前所有的书籍有所不同。全书分为三条主线：第一条主线是从第一代神经网络（线性分类器）、第二代神经网络（非线性）及其在预测领域的应用，到支持向量机，最后是深度学习；第二条主线是贝叶斯理论，从朴素贝叶斯算法到贝叶斯网，最后是隐马尔科夫模型，这部分属于智能推理的范畴；第三条主线是矩阵降维、奇异值分解（SVD）和 PCA 算法，因为算法简单，本书都使用真实案例进行讲解。

目前机器学习主要由这三条主线贯穿始终，本书着力于讲解三条主线的理论发展、思想变迁、数学原理，而具体算法就是在其上的一颗颗明珠。希望读者在学习完本书之后，能够将机器学习的各种理论融会贯通。

图书在版编目（CIP）数据

机器学习算法原理与编程实践 / 郑捷著. —北京：电子工业出版社，2015.11

ISBN 978-7-121-27367-4

Ⅰ. ①机… Ⅱ. ①郑… Ⅲ. ①机器学习－算法 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字（2015）第 239481 号

策划编辑：李　冰
责任编辑：李　冰
特约编辑：赵海红　罗树利
印　　刷：北京京科印刷有限公司
装　　订：三河市皇庄路通装订厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编 100036
开　　本：787×1092　1/16　　印张：27　　字数：704 千字　　彩插：4
版　　次：2015 年 11 月第 1 版
印　　次：2015 年 11 月第 1 次印刷
定　　价：88.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

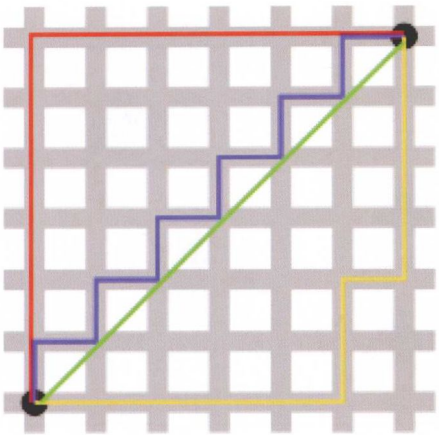


图1.10　A、B两点间的曼哈顿距离为红色、蓝色、黄色线条

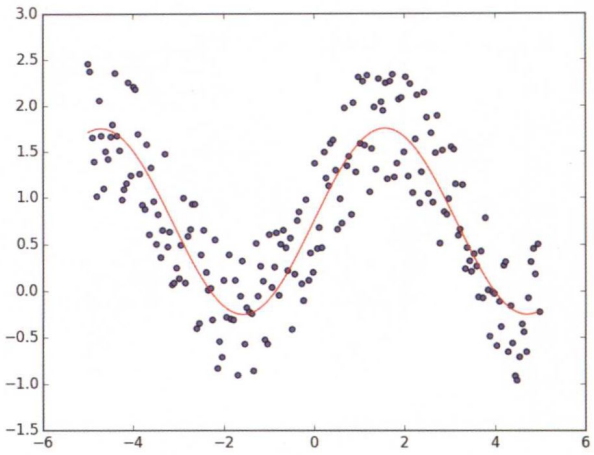


图1.17　数据可视化1

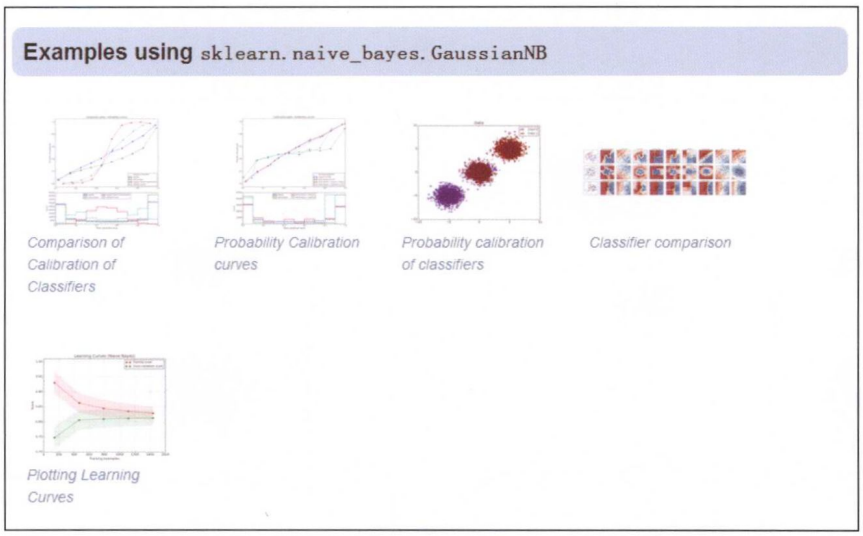


图2.7　高斯朴素贝叶斯实例

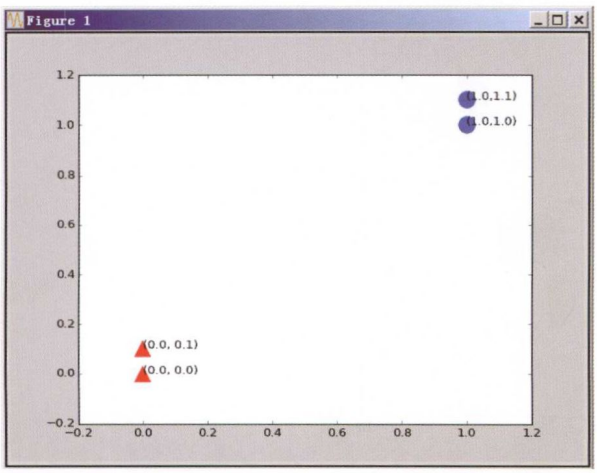


图2.9　4个点构成的训练集

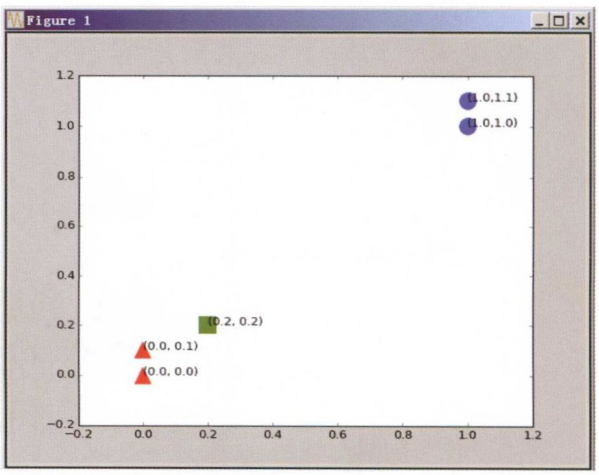


图2.10　kNN算法的基本原理

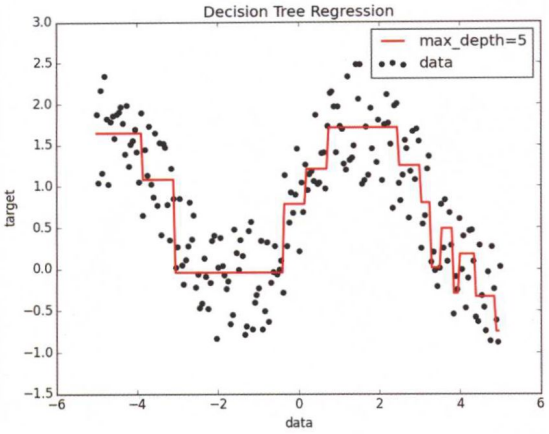


图3.7　Scikit-Learn生成的决策树

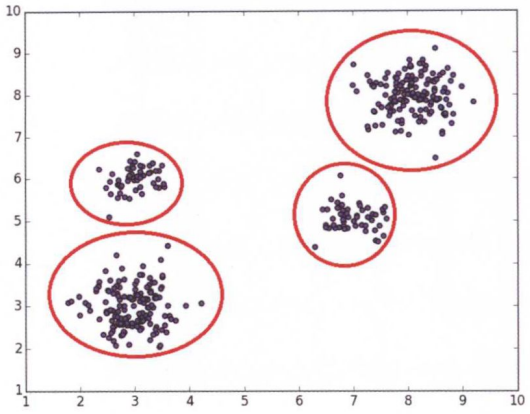


图4.8　散点示意图

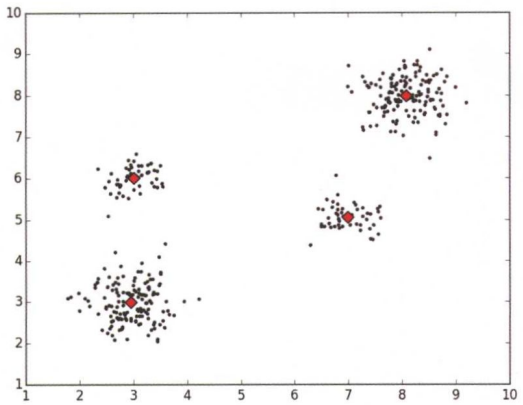


图4.9　KMeans聚类结果示意图

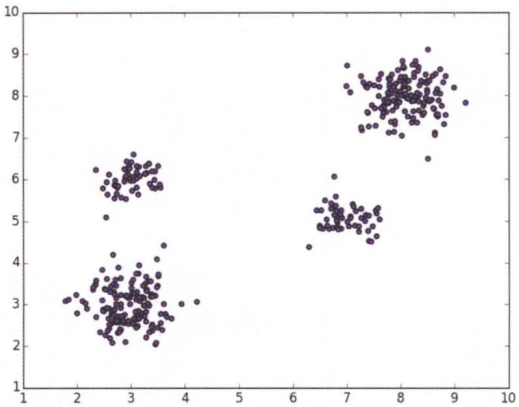


图4.10　随机散点图

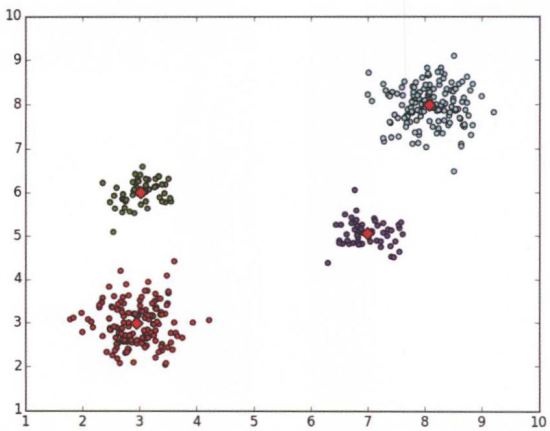


图4.12　KMeans聚类结果

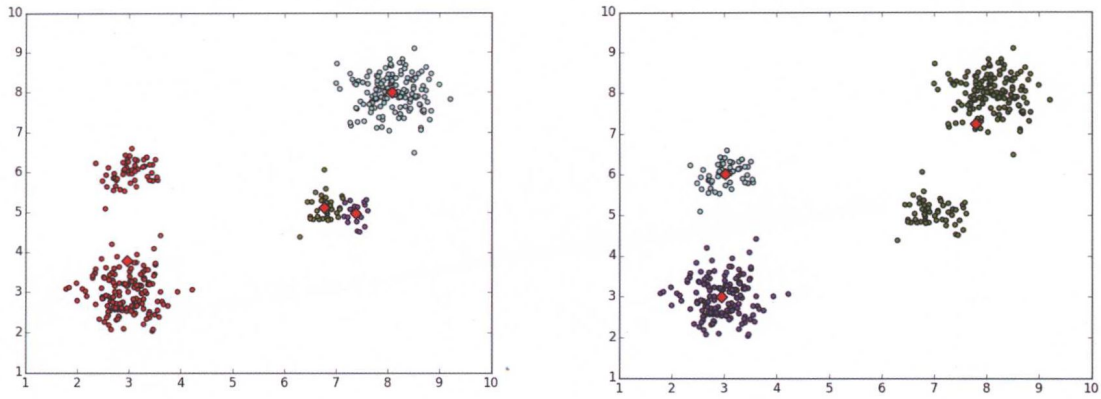


图4.13　KMeans聚类的错误结果

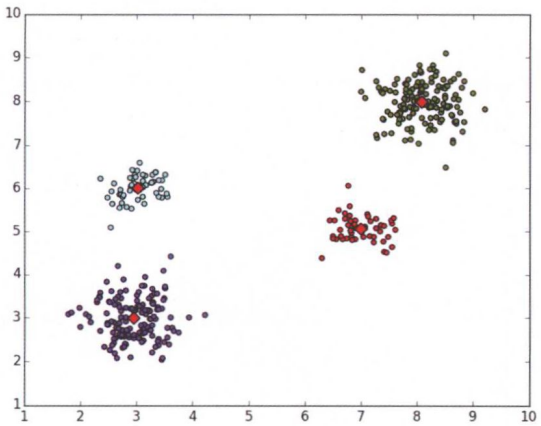


图4.14　二分KMeans算法的聚类结果

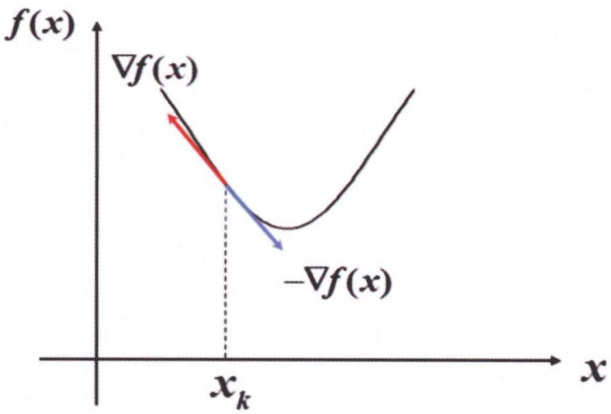


图5.7　目标函数的梯度和负梯度方向

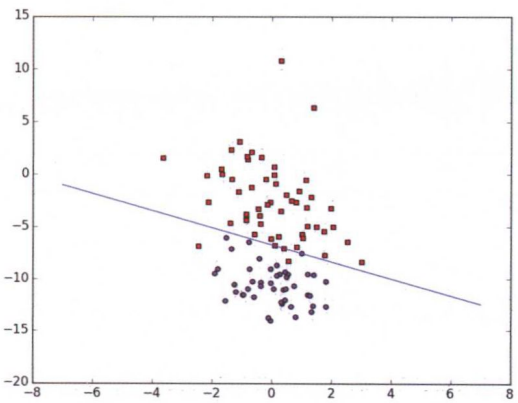


图5.14　权重向量构成的分类超平面

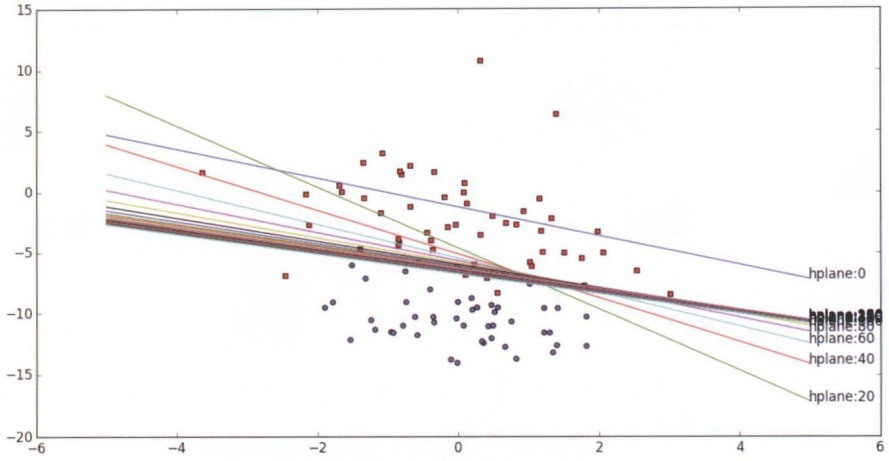


图5.15　分类超平面（权重向量）的变化

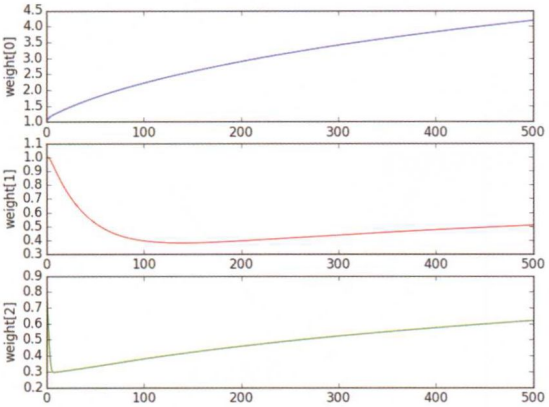


图5.18　权重向量的变化趋势

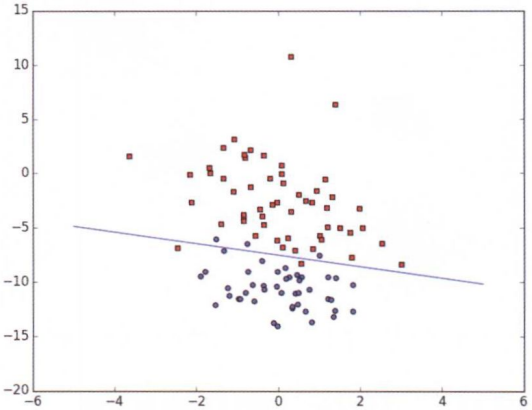


图5.19　随机梯度下降法输出

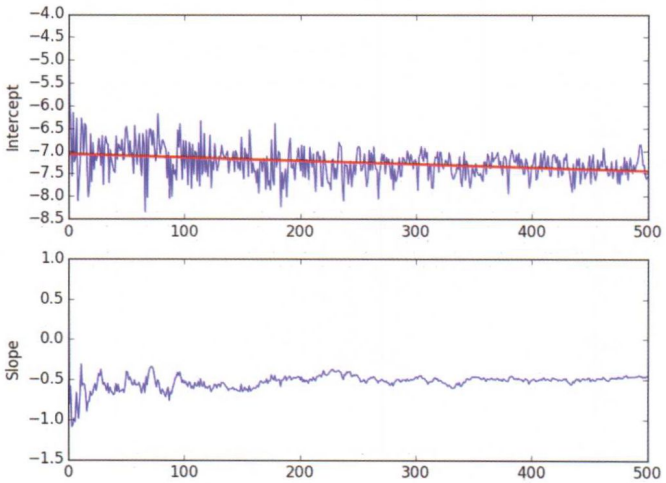


图5.21　斜率和截距变化

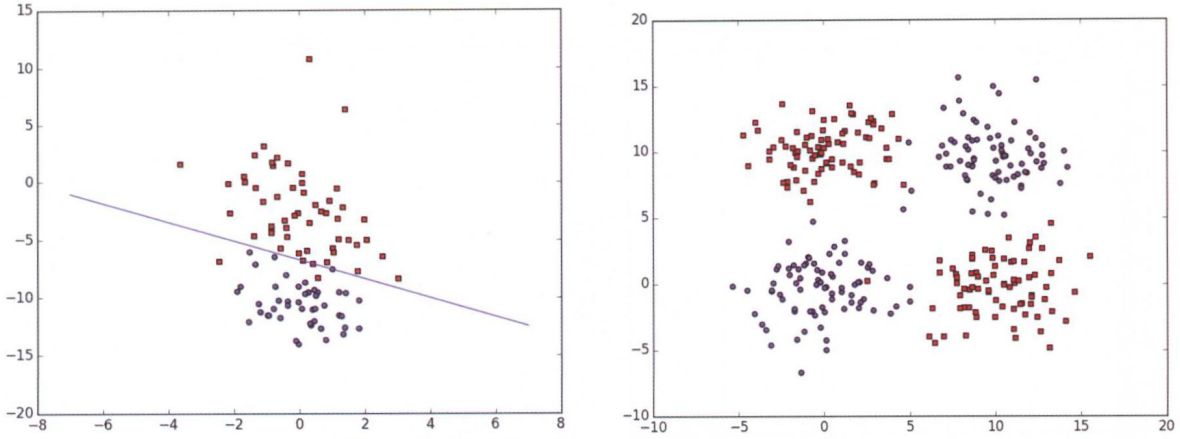


图6.1　线性可分与线性不可分

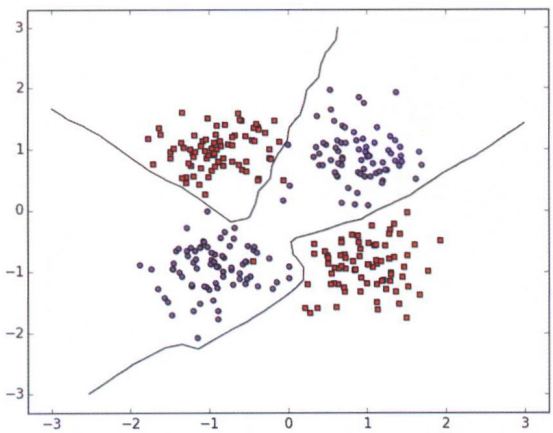


图6.4　数据集分类结果

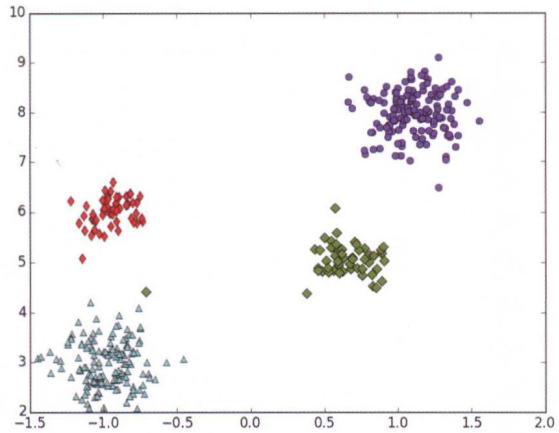


图6.10　聚类结果

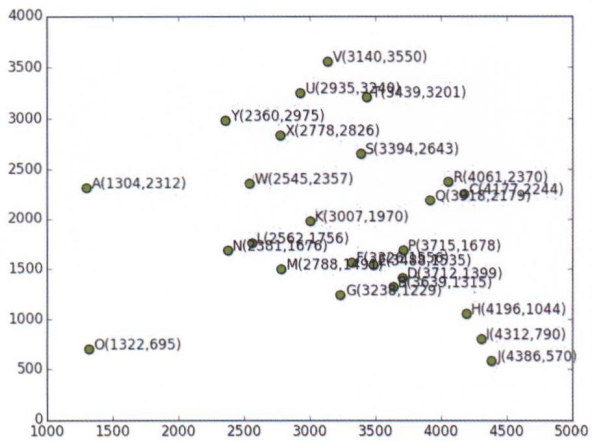


图6.11　TSP示意图

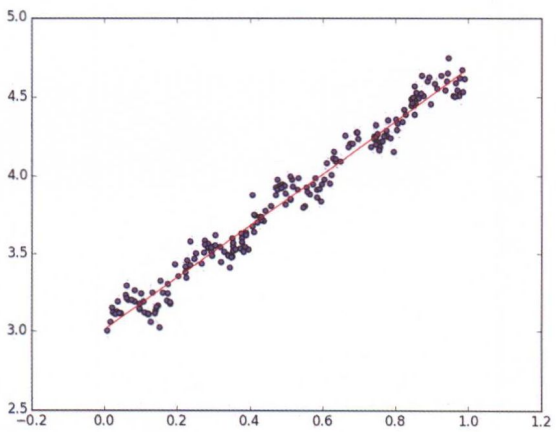


图7.2　最小二乘法的回归线

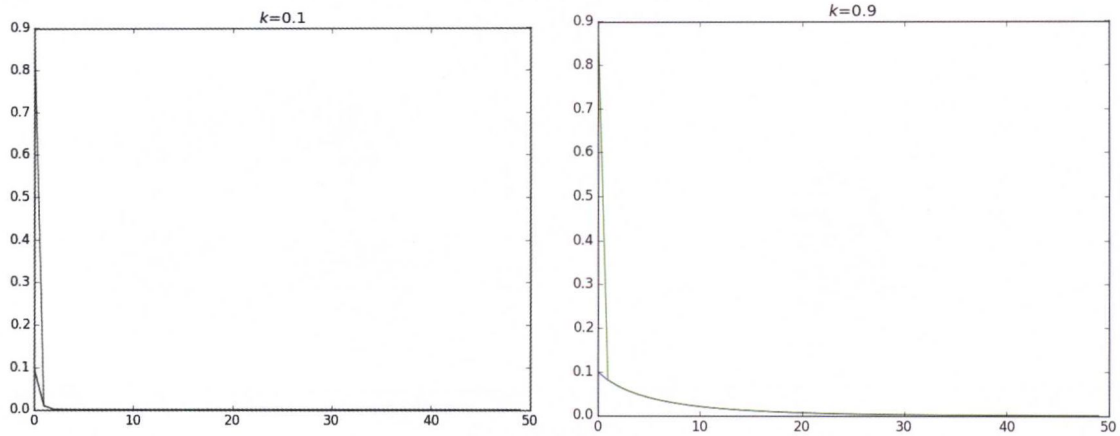


图7.15　Logistic映射的稳定点

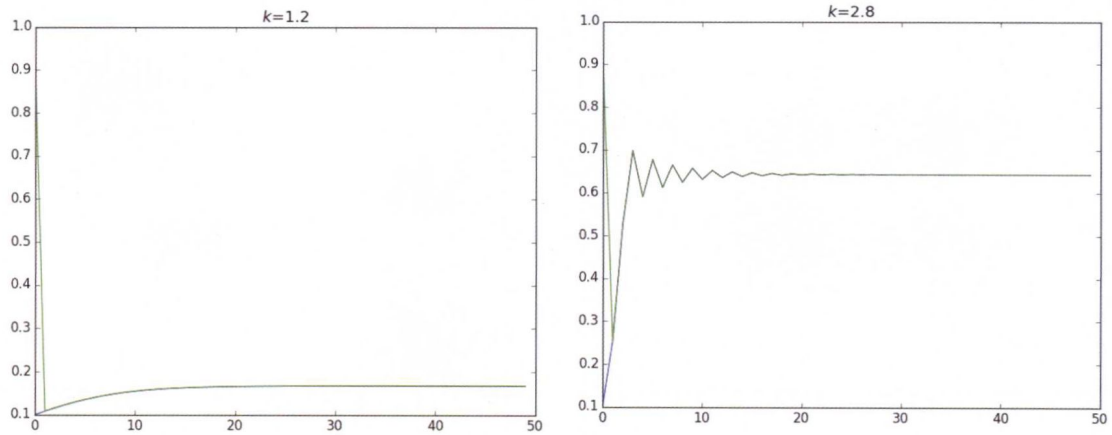


图7.16　Logistic映射的随k变动的稳定点

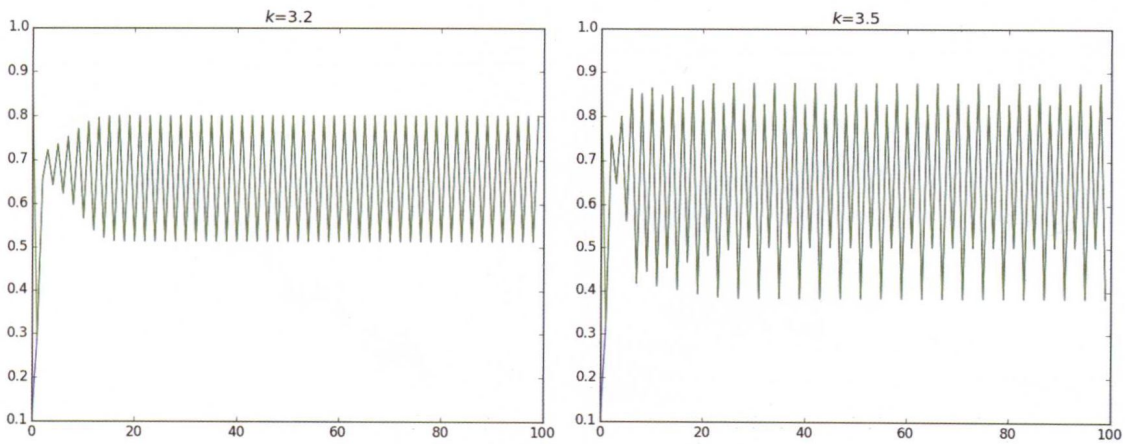


图7.17　Logistic映射的周期吸引子

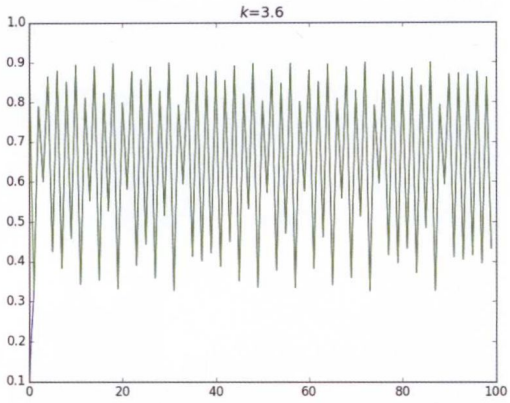


图7.18　Logistic映射的混沌吸引子

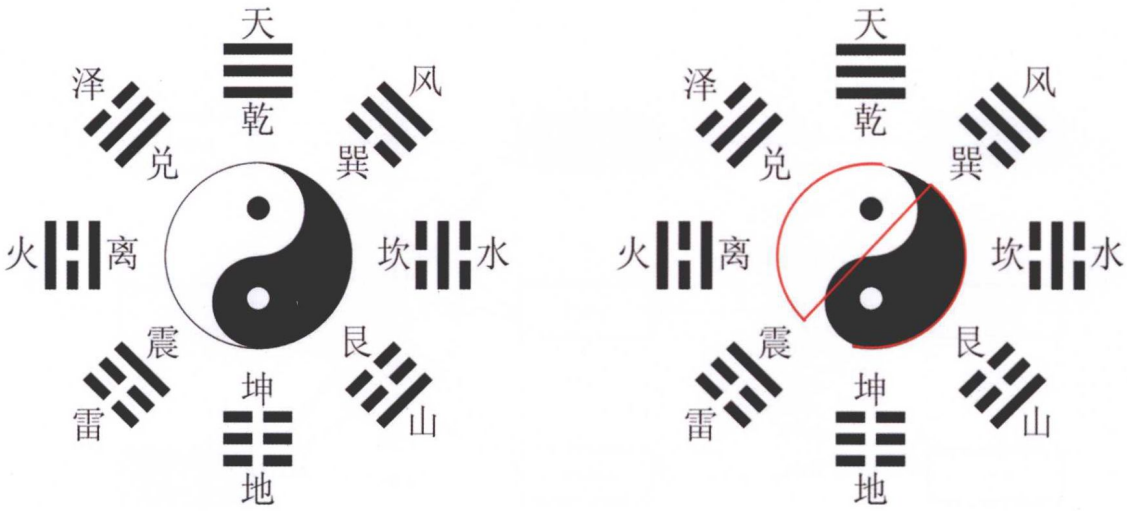


图7.21　八卦图

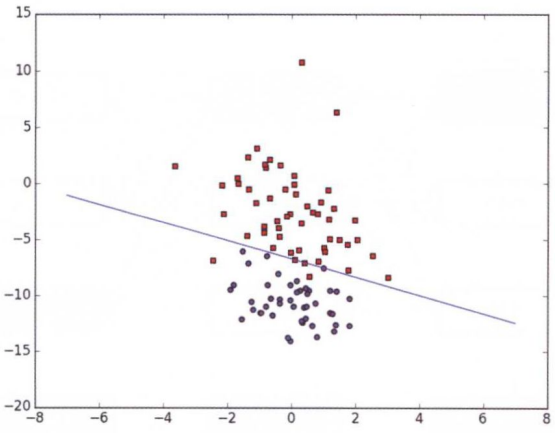


图8.8　离群点

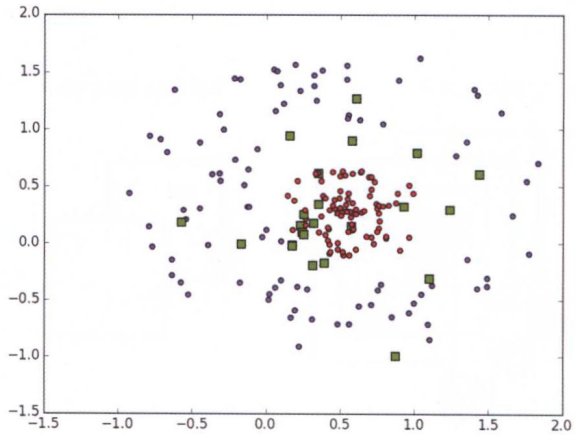


图8.11　数据集分类结果

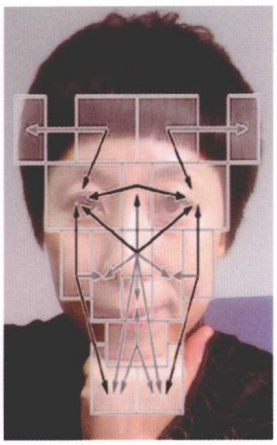


图9.1　人脸检测特征模板

图9.8　人脸的特征化

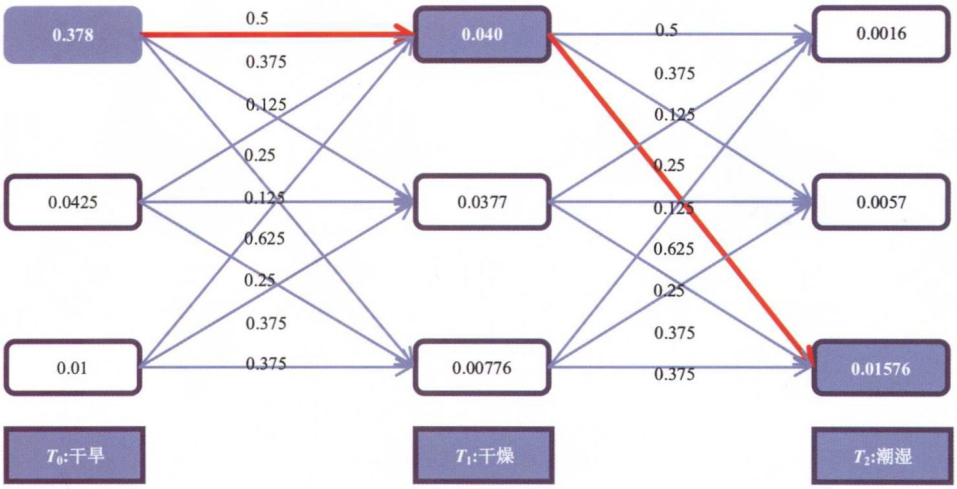


图11.11　计算完成后的HMM

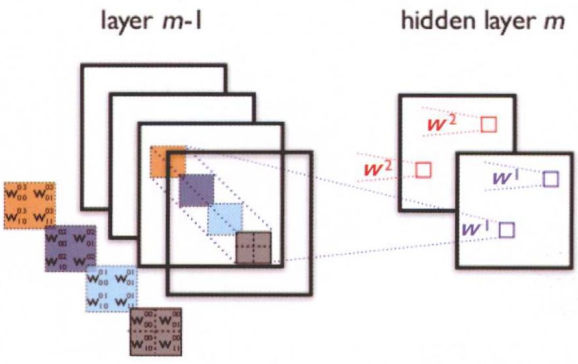


图10.8　卷积层

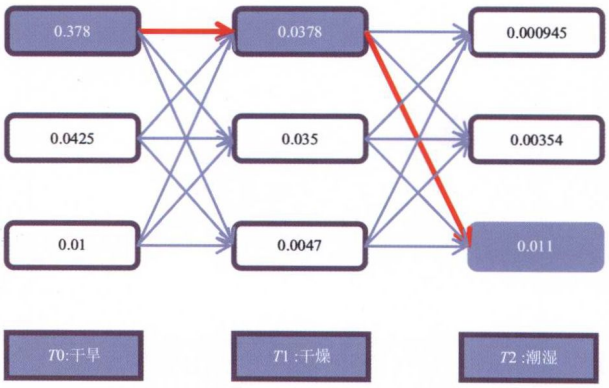


图11.12　Vertibi计算结果

前　言

动机

2011 年 1 月 14 日，史上最强的人机对抗在美国纽约约克镇高地拉开序幕。*Jeopardy!* 是美国具有 25 年历史的众所皆知的电视问答节目秀。每次三名参赛者相互角逐，在竞赛中需要迅速理解屏幕上提出的各类智力问题，并作出回答。问题涉及的领域十分广泛，就像一套世界知识的百科全书，超过个人所能掌握的知识容量的极限。而这次，一名特殊的参赛者名列其中，它就是 IBM 公司的计算机参赛者 Watson，挑战两位人类选手 Ken 和 Brad。经过激烈的角逐，Watson 同时击败了两位人类选手，赢得 100 万美元奖金而一举成名。这一具有历史意义的比赛被 *Jeopardy!* 的哥伦比亚广播公司连续在 2011 年 2 月 14—16 日三天晚上进行了重播，也成为计算机发展史上一个重要的时刻。IBM 评论为：

“在 *Jeopardy!* 比赛中，计算机打败人类选手是开放领域问答系统的一个里程碑！”

事实上，这次比赛有力地证明了，在广泛的知识和智能领域，机器有能力全面超越人类。开放领域问答软件的一个重要核心就是机器学习。从很多方面来看，这才仅仅是一个开始。近年来，计算机行业取得的最重要成就或多或少地都与机器学习领域的技术突破密切相关。2010 年前后，多伦多大学的 Geoffrey Hinton 提出的深度学习（Deep Learning）算法，突破了产生抽象概念的技术瓶颈，被评价为：

“借助于 DeepLearning 算法，人类终于找到了如何处理‘抽象概念’这个亘古难题的方法。”

该算法与衍生的卷积神经网络（CNN——有监督）和深度置信网络（DNN——无监督）在计算机视觉、语音识别和部分自然语言处理领域获得巨大的成功，其与另一个并行处理架构 Map Reduce 并称“大数据”技术的基石。

2012 年 11 月，微软在天津的一次活动上公开演示了一个全自动的同声传译系统，讲演者用英文演讲，后台的计算机一气呵成自动完成语音识别、英中机器翻译和中文语音合成，效果非常流畅。据报道，后面支撑的关键技术就是 DNN，或者深度学习（DeepLearning，DL）。

人与动物最本质的区别之一就是人类具有高度发达的智能。千百年来，人类从未停

止过对智慧本身的研究与探索。20 世纪 50 年代，图灵就在论文《机器能思考吗》中提出了一个著名测试，后世称为图灵测试：

“假如一台机器通过特殊的方式与人沟通，若有一定比例的人（超过 30%）无法在特定时间内（5 分钟）分辨出与自己交谈的是人还是机器，则可认为该机器具有‘思考’的能力。”

这里的思考能力就是指智能。而对于计算机领域而言，它是一个多么奢侈而艰难的字眼。以 IBM Watson 为例，它由 90 台 IBM 服务器、360 个计算机芯片驱动组成，是一个有着 10 台冰箱那么大的计算机系统。它拥有 15TB 内存，2880 台处理器，每秒可进行 80 万亿次运算。系统配置的处理器是 Power 7 系列处理器，这是当前 RISC（精简指令集计算机）架构中最强的处理器。它采用 45nm 工艺打造，拥有 8 个核心、32 个线程，主频最高可达 4.1GHz，仅其二级缓存就达到 32MB。

在大数据领域，据 Google 称，其知识图谱的信息有许多来源，包括 CIA 的世界概况、Freebase 和维基百科，其功能与 Ask.com 和 Wolfram Alpha 等问题问答系统相似。截至 2012 年，其语义网络包含超过 570 亿个对象，超过 18 亿个介绍，用来理解搜索关键词含义的、不同对象之间的链接关系更是不可计数。2012 年 11 月 4 日，知识图谱新增了 7 种语言：西班牙语、法语、德语、葡萄牙语、日语、俄罗斯语及意大利语。

历经半个多世纪，在各个领域的商业机构和科研机构的共同努力下，几经沉浮，人们逐渐意识到，高度并行的计算（硬件）能力和大规模数据的学习（算法）能力是“思考”的基础。距离让机器像人类一样思考的目标已经不远了！

本书特色

本书的最大特色就是理论讲解深入浅出、通俗易懂，入门门槛不高，理论与实践并重。降低学习门槛是我们主要的努力方向。对于中国读者，特别是广大的工程技术人员，无论是在职还是学生，进入机器学习领域不外乎两条路。

第一条路是从开源代码学习，辅助一些书籍资料。大多数软件设计人员都做过几年源码解析工作，源码解析这条路是比较辛苦的，但一旦掌握，就会形成一种条件反射。程序员宁可读源码也不愿意读数学公式，这是普遍现象。笔者认为，随着机器学习一步步走向工程实践，这部分人在读者群中应占绝大部分。

为了最大限度地降低学习的难度，首先在内容上，我们以大量的文字描述来说明重

要的定理和公式，尽可能在数学推导过程中增加充分的文字解释，消除初学者的理解障碍。其次，我们将源码、公式和文字解释对照起来，使初学者在阅读源码和文字解释的同时，也能够轻松理解算法的数学原理，使他们认识到数学分析并不遥远，理解起来并不困难。最后，我们使用矢量编程的设计方式，这种方式的优势是可以部分地将数学公式直接映射到代码上，代码简洁，思路清晰，学习效率很高。三管齐下，使初学者能多角度加深算法概念的理解，在实践应用中做到举一反三。

第二条路是从数学入手，一般针对研究所或科研院校的研究人员。他们喜欢那种有一定的理论高度，看明白了拿来就可以讲课或写论文用的书籍。这部分读者的特点是比较重理论，缺点是实践能力不强。本书可以通过丰富的算法代码弥补他们在此方面的不足。

最后，本书由本土作者编写。笔者翻译过几本国外的专业论文和书籍，也看过不少的本土经典。如果内容差异不大，从效率和接受程度上，看本土书籍要快很多，时间成本对任何一个人都是重要的；本土书籍的另一个优势是作者与大多数的读者都有相似的背景知识结构，因而没有文化差异性，思路上很好理解，容易被读者接受。本书内容多取材于实践，目标明确，针对性强，对读者而言学习效率高。

本书内容及体系结构

本书的特点之一是从结构上阐明了研究机器学习理论和算法的方法。最重要的不是数学，也不是这些算法本身，而是思想的发展过程，这与之前所有的书籍有所不同。全书分为三条主线。

第一条主线是从第一代神经网络（线性分类器）、第二代神经网络（非线性）及其在预测领域的应用，到支持向量机，最后是深度学习。

从第5章开始我们深入讲解了感知器网络及Logisitic网络的算法及相关的理论基础。第6章，我们详细介绍了三种典型的神经网络：BP网络、SOM网络、玻尔兹曼机网络。这两章的内容主要集中在第二代神经网络的模型上。

- 第8章我们从统计学习理论开始，深入探讨了支持向量机的模型，并给出了文本分类的实例。支持向量机的出现结束了浅层机器学习算法的大多数问题，使人工智能走向了一个新阶段。
- 第9、10章我们详细介绍了认知分层理论，并探讨了人类神经系统的两大重要机制：迭代和分层。由此引入了深度神经网络框架（深度学习），并以 Theano

框架为中心介绍了 GPU 运算的模型。深度学习框架中的算法很多，我们介绍了多层感知器和卷积神经网络两个算法，作为读者入门的基础。

第二条主线是贝叶斯理论，从朴素贝叶斯算法到贝叶斯网，最后是隐马尔科夫模型，这部分属于智能推理的范畴。

- 第 2 章我们详细介绍了朴素贝叶斯算法在文本分类中的应用。由于文本处理的大多数算法都是以贝叶斯网为基础的，而朴素贝叶斯是其中最简单的算法，所以以此开篇。
- 第 11 章，我们从随机过程开始，层层深入，相继介绍了马尔可夫链、贝叶斯网络、隐马尔科夫模型。最后，我们给出了隐马尔科夫模型的重要应用——自然语言处理的词性标注模块，并给出详细的代码讲解和结巴分词及词性标注应用。

最后一条主线是矩阵降维、奇异值分解（SVD）和 PCA 算法，因为算法简单，本书都使用真实案例进行讲解。

- 第 4 章，我们通过一个实例介绍了推荐系统的内容，并分析介绍了协同过滤理论中的两个重要算法：KMeans 和 SVD 隐语义分析。我们不仅讲解了 SVD 的数学推导，而且给出了手工计算的代码。
- 第 9 章，我们讲解了主成分分析（PCA）的基本原理和算法，并通过实例讲解，列出了 PCA 的算法实现和监测评估。
- 第 3、9 章，我们介绍了决策时算法的发展历史，以及各个历史时期的代表算法——ID3、C4.5、CART、AdaBoost，并给出基本原理和代码实现。

目前机器学习主要由这三条主线贯穿始终，本书着力于讲解这三条主线的理论发展、思想变迁、数学原理，而具体算法就是其上的一颗颗明珠。希望读者在学习完本书之后，能够将机器学习的各种理论融会贯通。

目　录