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

相关图书推荐


机器学习
MACHINE LEARNING
Machine Learning


Machine Learning for Hackers
机器学习
实用案例解析
O'REILLY


The Art of R Programming
R语言编程艺术


数据挖掘与R语言
Data Mining with R


算法导论
Introduction to Algorithms
ALGORITHMS


算法心得
高效算法的奥秘
Hacker's Delight
(Second Edition)

大数据技术丛书

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：
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

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

10.2 余弦相似度分类 …… 306
10.2.1 中文分词 …… 306
10.2.2 停用词清理 …… 308
10.2.3 算法实战 …… 310
10.3 朴素贝叶斯分类 …… 315
10.3.1 算法描述 …… 316
10.3.2 先验概率计算 …… 316
10.3.3 最大后验概率 …… 316
10.3.4 算法实现 …… 317
10.4 小结 …… 323
思考题 …… 323

第一部分

准备篇

子贡问为仁。子曰：“工欲善其事，必先利其器。居是邦也，事其大夫之贤者，友其士之仁者。”

——孔子

第1章

机器学习发展及应用前景

纵观国内软件工程师的发展路线，前期多以程序员（“码农”）、测试工程师、数据库管理员、多媒体技术员、网页与信息技术员等职业为主；中期主要是软件设计师、软件评测师、技术支持师等职业；后期是职业发展的黄金阶段，这一阶段对于拥有丰富技术经验的工程师来说十分重要，但这一阶段容易遭遇到技术发展瓶颈，因此，很多人将目光投向了项目管理和系统架构，比如：系统架构师、项目管理师等。

近年来，国内对机器学习的研究日益深入，应用领域不断扩大，催生了新的IT职位——机器学习工程师，百度、搜狗、阿里巴巴、淘宝、奇虎等国内IT巨头纷纷提出了对机器学习工程师的需求，掌握机器学习的人才成为了各大IT厂商争抢的“香饽饽”。机器学习迅速走红成为热门技术，这给软件工程师带来了绝佳的发展机遇，研究与应用机器学习算法成为了突破技术瓶颈的方式，机器学习工程师、项目管理师和系统架构师并称为后期发展的三大黄金职位。

1.1 机器学习概述

机器学习作为一门多领域的交叉学科，在近20年里异军突出。机器学习涉及概率论、统计学、微积分、代数学、算法复杂度理论等多门学科。通过可以让计算机自动“学习”的算法来实现人工智能，是人类在人工智能领域展开的积极探索。

2009年，被誉为人工大脑之父的雨果·德·加里斯教授走进清华大学讲堂，在两小时的演讲时间内，给大家描述了一个人工智能的世界：20年后，人工智能机器可以和人类做朋友，50年后，人工智能将成为人类最大的威胁，世界最终会因人工智能超过人类而爆发一场战争，这场智能战争也许会夺去数十亿人的生命。这样的描述并不是幻想，随着人类在人工智能领域取得的进步，这很有可能成为事实。而这一切主要归功于对机器学习的研究和探索。

1.1.1　什么是机器学习

学习是人类具有的一种重要智能行为。人类一直梦想机器能像人类一样学习，也一直在为这个终极目标努力。那么，什么是机器学习呢？长期以来众说纷纭，Langley（1996）定义机器学习为：“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”（Machine learning is a science of the artificial. The field's main objects of study are artifacts，specifically algorithms that improve their performance with experience.）。Mitchell（1997）在《*Machine Learning*》中写到：“机器学习是计算机算法的研究，并通过经验提高其自动进行改善”（Machine Learning is the study of computer algorithms that improve automatically through experience.）。Alpaydin（2004）提出自己对机器学习的定义：“机器学习是用数据或以往的经验，来优化计算机程序的性能标准”（Machine learning is programming computers to optimize a performance criterion using example data or past experience.）。

笔者综合维基百科和百度百科的定义，尝试着将机器学习定义如下：“机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，它是人工智能的核心，是使计算机具有智能的根本途径。机器学习的研究方法通常是根据生理学、认知科学等对人类学习机理的了解，建立人类学习过程的计算模型或认识模型，发展各种学习理论和学习方法，研究通用的学习算法并进行理论上的分析，建立面向任务的具有特定应用的学习系统。”

1.1.2　机器学习的发展

早在古代，人类就萌生了制造出智能机器的想法。中国人在4500年前发明的指南车，以及三国时期诸葛亮发明的尽人皆知的木牛流马；日本人在几百年前制造过靠机械装置驱动的玩偶；1770年英国公使给中国皇帝进贡了一个能写“八方向化，九土来王”8个汉字的机器玩偶（这个机器人至今还保存在故宫博物院），等等。这些例子，都只是人类早期对机器学习的一种认识和尝试。

真正的机器学习研究起步较晚，它的发展过程大体上可分为以下4个时期：

第一阶段是在20世纪50年代中叶到20世纪60年代中叶，属于热烈时期。

第二阶段是在20世纪60年代中叶至20世纪70年代中叶，被称为机器学习冷静期。

第三阶段是从20世纪70年代中叶至20世纪80年代中叶，称为机器学习复兴期。

最新的阶段起始于1986年。当时，机器学习综合应用了心理学、生物学和神经生理学以及数学、自动化和计算机科学，并形成了机器学习理论基础，同时还结合各种学习方法取长补短，形成集成学习系统。此外，机器学习与人工智能各种基础问题的统一性观点正在形成，各种学习方法的应用范围不断扩大，同时出现了商业化的机器学习产品，还积极开展了与机器学习有关的学术活动。

1989 年，Carbonell 指出机器学习有 4 个研究方向：连接机器学习、基于符号的归纳机器学习、遗传机器学习与分析机器学习。1997 年，Dietterich 再次提出了另外 4 个新的研究方向：分类器的集成（Ensembles of classifiers）、海量数据的有教师学习算法（Methods for scaling up supervised learning algorithm）、增强机器学习（Reinforcement learning）与学习复杂统计模型（Learning complex stochastic models）。

在机器学习的发展道路上，值得一提的是世界人工大脑之父雨果·德·加里斯教授。他创造的 CBM 大脑制造机器可以在几秒钟内进化成一个神经网络，可以处理将近 1 亿个人工神经元，它的计算能力相当于 10000 台个人计算机。在 2000 年，人工大脑可以控制“小猫机器人”的数百个行为能力。

2010 年以来，Google、Microsoft 等国际 IT 巨头加快了对机器学习的研究，已经尝到了机器学习商业化带来的甜头，国内很多知名的公司也纷纷效仿。阿里巴巴、淘宝为应付大数据时代带来的挑战，已经在自己的产品中大量应用机器学习算法。百度、搜狗等已拥有能与 Google 竞争的搜索引擎，其产品中也早已融合了机器学习知识，360 安全卫士的奇虎公司也意识到了机器学习的意义所在，这些大公司纷纷表现出对机器学习研发工程师的渴求。近几年正是机器学习知识在国内软件工程师群体中普及的黄金时代，也给软件工程师们进入机器学习这一金领行业带来了机遇。

1.1.3 机器学习的未来

展望未来，出现在《终结者》等系列电影上的场景终将成为现实，并将在未来的人类世界中频频上演。

目前人类已经进入了与高智能机器共同参与战争的新时代。据美国《航空周刊与空间技术》报道，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”，它意味着在未来的海空战场，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的作战模式。这将彻底改变人类战争的方式。未来人类在机器学习研究领域的发展将会进一步推动机器人军队在战场上的应用。

此外，智能机器已经深入到人类的生活、工作中。在民用领域，能从医疗记录中学习的机器将会出现，它们能分析和获取治疗新疾病最有效的方法；智能家居高度发展，分析住户的用电模式、居住习惯后，打造动态家居，从而降低能源消耗、提高居住舒适度；个人智能助理跟踪分析用户的职业和生活细节，协助用户高效完成工作和享受健康生活。所有这些都将有智能机器的功劳。

不久的将来，人类也许该思考：在未来的世界里，机器人将充当什么样的角色，会不会代替人类呢？人类与智能机器之间应如何相处？

人类开始着手研究，如何才能更好地实现下面三大准则：

第一，机器人不可伤害人；

第二，机器人必须服从人的命令；

第三，机器人可以在不违背上述原则的情况下保护自己。

1.2　机器学习应用前景

机器学习应用广泛，无论是在军事领域还是民用领域，都有机器学习算法施展的机会。

1.2.1　数据分析与挖掘

“数据挖掘”和“数据分析”通常被相提并论，并在许多场合被认为是可以相互替代的术语。关于数据挖掘，现在已有多种文字不同但含义接近的定义，例如“识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”；百度百科将数据分析定义为：“数据分析是指用适当的统计方法对收集来的大量第一手资料和第二手资料进行分析，以求最大化地开发数据资料的功能，发挥数据的作用，它是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。”无论是数据分析还是数据挖掘，都是帮助人们收集、分析数据，使之成为信息，并作出判断，因此可以将这两项合称为“数据分析与挖掘”。

数据分析与挖掘技术是机器学习算法和数据存取技术的结合，利用机器学习提供的统计分析、知识发现等手段分析海量数据，同时利用数据存取机制实现数据的高效读写。机器学习在数据分析与挖掘领域中拥有无可取代的地位，2012 年 Hadoop 进军机器学习领域就是一个很好的例子。

2012 年，Cloudera 收购 Myrrix 共创 Big Learning，从此，机器学习俱乐部多了一名新会员。Hadoop 和便宜的硬件使得大数据分析更加容易，随着硬盘和 CPU 越来越便宜，以及开源数据库和计算框架的成熟，创业公司甚至个人都可以进行 TB 级以上的复杂计算。Myrrix 从 Apache Mahout 项目演变而来，是一个基于机器学习的实时可扩展的集群和推荐系统。

Myrrix 创始人 Owen 在其文章中提到：机器学习已经是一个有几十年历史的领域了，为什么大家现在这么热衷于这项技术？因为大数据环境下，更多的数据使机器学习算法表现得更好，机器学习算法能从数据海洋提取更多有用的信息；Hadoop 使收集和分析数据的成本降低，学习的价值提高。Myrrix 与 Hadoop 的结合是机器学习、分布式计算和数据分析与挖掘的联姻，这三大技术的结合让机器学习应用场景呈爆炸式的增长，这对机器学习来说是一个千载难逢的好机会。

1.2.2　模式识别

模式识别起源于工程领域，而机器学习起源于计算机科学，这两个不同学科的结合带来了模式识别领域的调整和发展。模式识别研究主要集中在两个方面：一是研究生物体（包括人）是如何感知对象的，属于认识科学的范畴；二是在给定的任务下，如何用计算机实现模式识别的理论和方法，这些是机器学习的长项，也是机器学习研究的内容之一。

模式识别的应用领域广泛，包括计算机视觉、医学图像分析、光学文字识别、自然语言处理、语音识别、手写识别、生物特征识别、文件分类、搜索引擎等，而这些领域也正是机器学习的大展身手的舞台，因此模式识别与机器学习的关系越来越密切，以至于国外很多书籍把模式识别与机器学习综合在一本书里讲述。

1.2.3 更广阔的领域

目前国外的 IT 巨头正在深入研究和应用机器学习，他们把目标定位于全面模仿人类大脑，试图创造出拥有人类智慧的机器大脑。

2012 年 Google 在人工智能领域发布了一个划时代的产品——人脑模拟软件，这个软件具备自我学习功能，模拟脑细胞的相互交流，可以通过看 YouTube 视频学习识别猫、人以及其他事物。当有数据被送达这个神经网络的时候，不同神经元之间的关系就会发生改变。而这也使得神经网络能够得到对某些特定数据的反应机制，据悉这个网络现在已经学到了一些东西，Google 将有望在多个领域使用这一新技术，最先获益的可能是语音识别。

与此同时，Google 研制的自动驾驶汽车于 2012 年 5 月获得了美国首个自动驾驶车辆许可证，将于 2015 年至 2017 年进入市场销售，如图 1-1 所示。

图 1-1 Google 研制的自动驾驶汽车

自动驾驶汽车依靠人工智能、视觉计算、雷达、监控装置和全球定位系统协同合作，让电脑可以在没有任何人类主动操作的情况下，通过计算机自动安全地操作机动车辆，Google 认为：这将是一种“比人更聪明的”汽车，不仅能预防交通事故，还能节省行驶时间、降低碳排放量。

2013 年，Microsoft CEO 高级顾问 Craig Mundie 在北京航空航天大学学术交流厅发表“科技改变未来”的主题演讲，Mundie 在演讲中谈到了当今 IT 科技的三大挑战：大数据、人工智能和人机互动。他认为随着大数据时代的到来，人们的各种互动、设备、社交网络和传感器正在生成海量的数据，而机器学习可以更好地处理这些数据，挖掘其中的潜在价值。与此同时，他展示了微软研究院在机器学习方面的新产品——英语转汉语实时拟原声翻译，研究过计算语言学的朋友都知道自然语言理解与处理属于机器学习的问题，让计算机理解人

类语言可以视同创造出一个机器，该机器拥有与人类一样聪明的智慧。

机器学习在军事上的应用更加广泛，智能无人机、智能无人舰艇、智能无人潜艇陆续研究成功或已投放战场，其他军事领域也有机器学习研究成果的应用，如：美国国防部高级研究计划局的电子战专家正在尝试推出利用机器学习技术对抗敌方的无线自适应通信威胁，其发布了一份概括性机构通告（DARPA-BAA-10-79），内容为“自适应电子战行为学习”计划（BLADE），以研发确保美国电子战系统能够在战场上学习自动干扰新式射频威胁的算法和技术。

1.3　小结

机器学习作为一门多领域交叉学科，该领域的主要研究对象是人工智能，专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，使之不断改善自身的性能，它是人工智能的核心，是使计算机具有智能的根本途径。

近年来，机器学习的研究与应用在国内外越来越重视。机器学习已经广泛应用于语音识别、图像识别、数据挖掘等领域。大数据时代的到来，使机器学习有了新的应用领域，从包含设备维护、借贷申请、金融交易、医疗记录、广告点击、用户消费、客户网络行为等数据中发现有价值的信息已经成为其研究与应用的热点。

我们以记者与雨果·德·加里斯教授的部分专访内容来结束这一章。

记者：为什么选择这个研究工作？

雨果·德·加里斯：对人类大脑的好奇心和人脑的想象力的好奇心。人类只是一个个分子构成的机器，像计算机的芯片一样，像编制程序一样。另一方面，作为生物人都会死亡消失的，但人工智能机器就不会。所以说，这个研究就像制造神一样。

当人类促使技术进步，让具有人工智能的机器人得以诞生和发展，但总有一天人工智能机器会实现自己进化，当这种技术达到一个奇点的时候，就不需要人类来推动了。比人类聪明得多的人工智能机器将在以年为单位的短时间里产生。

记者：预测一下未来人工智能机器的前景。

雨果·德·加里斯：下一个20年，它们很有可能出现在我们的家里，为我们打扫房间，照顾小孩，和我们聊天，给我们来自地球上知识库里面的无限知识。我们还将可以和它们有性关系，被它们教育，从它们那里得到娱乐和开怀大笑。20年后的大脑制造业，每年全球范围内将可能创造万亿美元的价值。人工智能机器有比我们聪明万亿倍的可能性，不夸张地说，人工智能机器和人类交流，就像人类试图和岩石交流一样艰难。不过，真正的人工智能在我死后的三四十年内不会被制造出来。我活着看不到工作的真正结果，这是让我沮丧和失望的一个根源。

第2章

科学计算平台

机器学习算法具有坚实的数学理论支持，机器学习的应用建立在科学计算的基础上，而数学计算又是科学计算的主要组成部分。计算机技术的飞速发展和计算数学方法及理论的日益成熟，使解决复杂的数学计算问题成为可能。这些问题在以前用一般的计算工具来解决非常困难，而现在用计算机来处理却非常容易。目前用计算机处理得较多的数学计算主要分为以下两类：

第一类是数值计算，它以数值数组作为运算对象，给出数值解；计算过程中可能会产生误差累积问题，影响了计算结果的精确性；计算速度快，占用资源少。

第二类是符号计算，它以符号对象和符号表达式作为运算对象，给出解析解；运算不受计算误差累积问题的影响；计算指令简单；占用资源多，计算耗时长。

数值计算方法成为了科学计算的重要手段，它研究怎样利用计算工具来求出数学问题的数值解。数值计算方法的计算对象是微积分、线性代数、插值与逼近及最小二乘拟合、数值积分与数值微分、矩阵的特征值与特征向量求解、线性方程组与非线性方程求根，以及微分方程数值解法等数学问题，这些是模式识别、数据分析及自动制造等机器学习领域需要应用的数学。

符号计算是专家系统等机器学习领域需要应用的数学，在符号计算中，计算机处理的数据和得到的结果都是符号。符号既可以是字母和公式，也可以是数值，其运算以推理解析的方式进行，不受计算误差积累问题困扰，计算结果为完全正确的封闭解或任意精度的数值解，这意味着符号计算给出的结果能避免因舍入误差而引起的问题。还有更多的数学分支正在进入机器学习领域，复杂的数学计算需要强大的科学计算平台。科学计算平台提供了机器学习算法应用的底层支持。

2.1 科学计算软件平台概述

现代科学研究的方法主要有三种：理论论证、科学实验、科学计算。近年来，科学计算

方法逐步成为科学研究的主流方法，在金融工程、信息检索、基因研究、环境模拟、数值计算、数据分析、决策支持等领域得到了广泛使用。由于计算机技术的发展及其在各技术科学领域的应用推广与深化，这些应用领域不论其背景与含义如何，都要用计算机进行科学计算，都必须建立相应的数学模型，并研究其适合于计算机编程的计算方法。科学计算平台已经成为科学研究必要的基础条件平台，有力地推动了科学研究的发展和工程技术的进步。

机器学习应用需要科学计算的支持。大部分科学计算应用的领域都需要用到机器学习算法，科学计算平台与机器学习之间的关系就像鱼与水的关系。现代机器学习研究与应用早已经离不开科学计算平台的支撑，科学计算平台也因为机器学习的迅猛发展而进入了全新的百家争鸣时代。

2.1.1　常用的科学计算软件

目前常用的科学计算软件有以下几种：

1. MATLAB

MATLAB 是一种用于数值计算、可视化及编程的高级语言和交互式环境。使用 MATLAB，可以分析数据、开发算法、创建模型和应用程序，通过矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言等方式完成计算，比电子表格或传统编程语言（如 C/C++ 或 Java）更方便快捷。MATLAB 具有强大的数值计算功能，可完成矩阵分析、线性代数、多元函数分析、数值微积分、方程求解、边值问题求解、数理统计等常见的数值计算，同时它也能进行符号计算。

2. GNU Octave

GNU Octave 与 MATLAB 相似，它是自由软件基金会开发的一个自由再发布软件，以 John W. Eaton 为首的一些志愿者共同开发了叫作 GNU Octave 的高级语言，这种语言与 MATLAB 兼容，主要用于数值计算，同时它还提供了一个方便的命令行方式，可以数值求解线性和非线性问题，以及做一些数值模拟。

3. Mathematica

Mathematica 系统是美国 Wolfram 研究公司开发的一个功能强大的计算机数学系统。它提供了范围广泛的数学计算功能，支持在各个领域工作的人们做科学研究的过程中的各种计算。这个系统是一个集成化的计算机软件系统，它的主要功能包括符号演算、数值计算和图形三个方面，可以帮助人们解决各种领域里比较复杂的符号计算和数值计算的理论和实际问题。

4. Maple

1980 年 9 月，加拿大滑铁卢大学的符号计算研究小组研制出一种计算机代数系统，取名为 Maple，如今 Maple 已演变成为优秀的数学软件，它具有良好的使用环境、强有力的符号计算能力、高精度的数字计算、灵活的图形显示和高效的可编程功能。Maple 在符号计算

方面功能强大，符号计算式可以直接以数学的形式来输入和输出，直观方便。

5. SPSS

SPSS 预测分析是 IBM 公司的产品，它提供了统计分析、数据和文本挖掘、预测模型和决策优化等功能。IBM 宣称，使用 SPSS 可获得 5 大优势：商业智能，利用强大而简单的分析功能，控制数据爆炸，满足组织灵活部署商业智能的需求，提升用户期望值；绩效管理，指导管理战略，使其朝着最能盈利的方向发展，并提供及时准确的数据、场景建模、浅显易懂的报告等；预测分析，通过发现细微的模式关联，开发和部署预测模型，以优化决策制定；分析决策管理，一线业务员工可利用该系统，与每位客户进行互动，从中获取丰富信息，提高业务成绩；风险管理，在合理的前提下，利用智能的风险管理程序和技术，制定规避风险的决策。

6. R

R 语言是主要用于统计分析、绘图的语言和操作环境。 R 目前由“R 开发核心团队”负责开发，它是基于 S 语言的一个 GNU 项目，语法来自 Scheme，所以也可以当作 S 语言的一种实现，虽然 R 主要用于统计分析或者开发统计相关的软件，但也可用作矩阵计算，其分析速度堪比 GNU Octave 甚至 MATLAB。R 主要是以命令行操作，网上也有几种图形用户界面可供下载。R 内建多种统计学及数字分析功能，还能透过安装套件（Packages）增强。

7. NumPy、SciPy、matplotlib 等 Python 科学计算平台

Python 是一种面向对象的、动态的程序设计语言，它具有非常简洁而清晰的语法，既可以用于快速开发程序脚本，也可以用于开发大规模的软件，特别适合于完成各种高层任务。随着 NumPy、SciPy、matplotlib 等众多程序库的开发，Python 越来越适合用于科学计算。NumPy 是一个基础科学的计算包，包括：一个强大的 N 维数组对象封装了 C++ 和 Fortran 代码的工具、线性代数、傅立叶转换和随机数生成函数等其他复杂功能的计算包。SciPy 是一个开源的数学、科学和工程计算包，能完成最优化、线性代数、积分、插值、特殊函数、快速傅里叶变换、信号处理和图像处理、常微分方程求解等计算。matplotlib 是 Python 最著名的绘图库，它提供了一整套和 MATLAB 相似的命令 API，十分适合交互式制图，它也可以方便地用作绘图控件，嵌入 GUI 应用程序中。

2.1.2 本书使用的工程计算平台

MATLAB、Mathematica、Maple、SPSS 等软件功能齐全、界面友好，同时内含多种强大的软件包，但价格昂贵，它们都是商业化软件，GNU Octave、R 和 Python 科学计算包作为开源免费的工程计算平台，会是不错的选择。

本书选择 R 和 Python 科学计算包作为工程计算平台，Python 和 R 都能在多种操作系统平台上运行。Python 是一种非常流行的脚本语言，用户较多，容易掌握，也有成熟并行计算框架 dispy 等，测试成功的单机机器学习算法稍加修改就能应用于大规模分布式计算的工程之中。正是因为 Python 具有如此之多的优点，Google 内部也经常使用它。R 语言内置

大量统计分析包，能访问部分系统函数，核心为解释执行的语言，大部分用户可见的 R 函数由 R 语言本身编写，出于效率原理，计算密集型任务通过在运行时链接与调用 C、C++、FORTRAN 代码完成。此外，通过 Rcpp 能把丰富的 R 环境与 C/C++ 等结合，将 R 的 API 与数据对象封装成类以及类的方法，供外部 C++ 程序调用。

2.2　计算平台的配置

本章将以 Windows 平台和 Linux 平台为例，讲解 R 和 Python 科学计算平台的配置。Python 和 R 具有跨平台运行的特点，Windows 平台编写的 Python 和 R 代码只需修正兼容性问题即可正常运行在类 UNIX 平台上，如：中文字符的 UTF8 与 GBK 转换、Windows 系统与类 UNIX 平台的文件路径差异等。

2.2.1　Numpy 等 Python 科学计算包的安装与配置

Python 科学计算包有两种安装方式，即：分别安装科学计算平台内的软件包和安装 WinPython 集成计算包。

1. 分别安装科学计算平台内的软件包

先安装 Python，关于它的版本，推荐使用 2.7 版本，然后安装 NumPy、SciPy、matplotlib 等 Python 软件包，它们都有 Windows 系统下的安装包。

Python 安装包的下载页面为 http://www.python.org/download/，选择 2.7 版本的 Windows 安装可执行文件下载即可。

NumPy 安装包下载页面为 https://pypi.python.org/pypi/numpy，下载 Windows 版本的安装可执行文件即可。

SciPy 安装包下载页面为 https://pypi.python.org/pypi/scipy/，该软件包目前没有 Windows 版本的安装执行文件，要用传统的 Python 安装第三方软件包的方式安装，将安装包下载解压，然后在命令行进入解压目录，输入以下命令：

```
python setup.py install
```

Matplotlib 软件包的下载页面为 http://matplotlib.org/downloads.html，下载 Windows 版本的安装可执行文件即可，注意应下载 Latest stable version 对应的软件包。Windows 版本的安装可执行文件通常命名格式为：产品名称 + 平台名称 +CPU 型号 + 版本号。以 Matplotlib 为例，打开其下载页面，如图 2-1 所示。

图 2-1　Matplotlib 下载页面

假设计算机的 CPU 是 32 位，Python 版本号为 2.7，则下载安装 matplotlib-1.3.0.win32-py2.7.exe，如

果 CPU 是 64 位的，Python 版本号为 2.7，则下载安装 matplotlib-1.3.0.win-amd64-py2.7.exe。

在类 UNIX 平台上（以 UBUNTU 为例），可使用下面的命令安装 Python 及相关科学计算包：

```
sudo apt-get install python-numpy python-scipy python-matplotlib ipython ipython-notebook python-pandas python-sympy python-nose
```

2. 安装 WinPython 集成计算包

WinPython 集成计算包集成了 Numpy 等第三方 Python 科学计算库，安装 WinPython 后，Numpy 等计算库和 Python 2.7 会一同被安装。此外，WinPython 附带一款非常不错的 IDE 开发调试环境：Spyder，如图 2-2 所示是 Spyder 的界面截图。

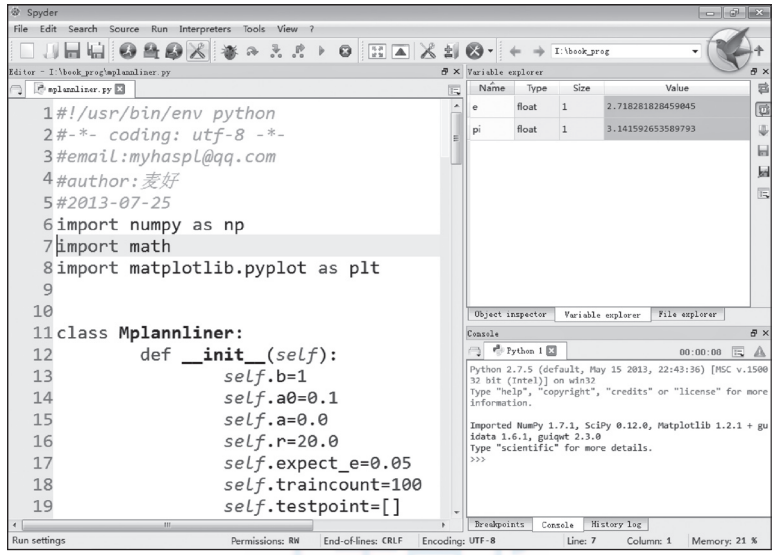

图 2-2 Spyder 界面

在图 2-2 所示的界面中，右上角是类似于 MATLAB 的“工作空间”，可很方便地观察和修改变量（包含多维数组）的值，同时还拥有方便用户的智能代码（Call-Tips 和 Auto-Complete）功能，如图 2-3 所示。

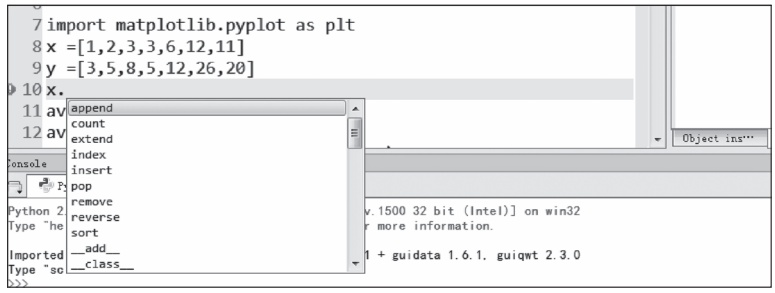

图 2-3 智能代码功能

在 IDE 开发窗口下方的 Console 栏可以使用 pdb（类似于 C 语言的 GDB 调试工具）调试 Python 代码，也可以通过 Spyder 的调试菜单进行调试。下面是 pdb 调试工具的使用帮助：

```
>>> debugfile(r'K:\book_prog\zxecf.py', wdir=r'K:\book_prog')
> k:\book_prog\zxecf.py(7)<module>()
-> import matplotlib.pyplot as plt
(pdb) help
Documented commands (type help <topic>):
========================================
EOF    bt         cont      enable  jump  pp       run      unt
a      c          continue  exit    l     q        s        until
alias  cl         d         h       list  quit     step     up
args   clear      debug     help    n     r        tbreak   w
b      commands   disable   ignore  next  restart  u        whatis
break  condition  down      j       p     return   unalias  where
```

常用的 pdb 调试命令如下：

- h(elp)：打印当前版本 pdb 可用的命令。
- disable/enable：禁用 / 启用断点。
- n(ext)：让程序运行下一行。
- c(ont(inue))：让程序正常运行，直到遇到断点。
- j(ump)：让程序跳转到指定的行数。
- b(reak)：设置断点，例如“b 23”，就是在当前脚本的 23 行打上断点，函数名也可作为参数。
- condition：设置条件断点。下面语句就是对第 5 个断点加上条件 x>=8：

```
(Pdb) condition 5 x>=8
```

- cl(ear)：清除指定参数的断点或所有断点。
- p：打印某个变量。比如：

```
(Pdb) p _file
u'./pic/dog.jpg'
```

- ！：感叹号后面跟着语句，可以直接改变某个变量。
- q(uit)：退出调试。

综上所述，在 Spyder 的帮助下，能更高效地开发与调试 Python 代码，因此笔者推荐在开发环境中安装 WinPython，方便快捷，有利于机器学习算法代码的编写。

2.2.2　OpenCV 安装与配置

OpenCV 是 Intel 开源计算机视觉库，它由一系列 C 函数和少量 C++ 类构成，实现了图像处理和计算机视觉方面的很多通用算法。OpenCV 拥有包括 300 多个 C 函数的跨平台的中高层 API。它不依赖于其他的外部库（尽管也可以使用某些外部库），对工程应用来说，OpenCV 是一个非常好的计算平台，因为它遵守 BSD 开源协议，对非商业应用和商业应用都是免费。

OpenCV 的主要功能有：图像数据操作，图像 / 视频的输入输出，矩阵 / 向量数据操作及线性代数运算，支持多种动态数据结构、基本图像处理、结构分析、摄像头定标、运动分析、目标识别、基本的 GUI 和图像标注。而且，OpenCV 提供了官方的 Python 接口，其使用方法和 C 语言接口基本一致，只是一些函数和结构体可能会有不同。另外，函数通过参数来返回值时一次会返回多个值。

在 Windows 上下载安装 OpenCV 的可执行文件后可直接运行，下载页面为 http://opencv.org/downloads.html。

其在 Linux 平台上的安装方式在 OpenCV 官网上有介绍，具体安装顺序如下：

（1）安装基本软件包。GCC 4.4.x 或更高版本、CMake 或更高版本、Git、GTK+2.x 或更高版本、including headers (libgtk2.0-dev)、pkgconfig、Python 2.6 或更高版本、Numpy 1.5 或更高版本、python-dev、python-numpy、ffmpeg 或 libav 开发包、ibavcodec-dev、libavformat-dev、libswscale-dev。

（2）安装可选软件包。libdc1394 2.x、libjpeg-dev、libpng-dev、libtiff-dev、libjasper-dev。

（3）在 http://opencv.org/downloads.html 下载其源代码，解压后，进入目录以源代码编译方式安装 OpenCV。

```
$cd ~/opencv
$mkdir release
$cd release
$cmake -D CMAKE_BUILD_TYPE=RELEASE -D  CMAKE_INSTALL_PREFIX=/usr/local ..
$make
$sudo make install
```

2.2.3 mlpy 安装与配置

mlpy 是基于 NumPy/SciPy 和 GSL 构建的 Python 模块，它提供了高层函数和类，允许使用少量代码来完成复杂的分类、特征提取、回归、聚类等任务。mply 为免费软件，建立在 GPL3 开源协议之上。

mlpy 在 Windows 下的安装方式较简单，可以直接在下面网址下载可执行文件安装：

http://sourceforge.net/projects/mlpy/files/

在类 Linux 平台上，其安装方法稍稍复杂一些，以 Linux、OSX 和 FreeBSD 为例，安装配置 mlpy，需要先安装配置好以下软件：

- GCC
- Python 且版本 >= 2.6 或为 3.X
- NumPy 且版本 >= 1.3.0
- SciPy 且版本 >= 0.7.0
- GSL 且版本 >= 1.11

然后，在上面网址中找到 mlpy 源代码包下载，并解压安装。假设 GSL 头文件和库文件没有安装在系统的标准位置，在这种情况下，mply 的安装方式如下：

```
$python setup.py build_ext --include-dirs=/path/to/header --rpath=/path/to/lib
$python setup.py install
```

如果 GSL 安装在标准位置，则只需要运行上述命令中的最后一行。

OpenCV 官方提供了 Python 绑定库，以 Python2.7 为例讲述了安装绑定库的方法，在 Windows 下，将它复制到 Python 的目录下，将 opencv\build\python\2.7 下的 cv2.pyd 文件复制到 python-2.7.5\Lib\site-packages 目录下即可。在 Linux 下安装了 Python 后，要确保 usr/lib/python2.7/site-packages 下有 cv.py 和 cv2.so 文件，如果没有，将这两个文件复制过来即可。

2.2.4　BeautifulSoup 安装与配置

BeautifulSoup 是用 Python 写的一个 HTML/XML 的解析器，它可以很好地处理不规范标记，并生成剖析树，通常用来分析“爬虫”抓取的 Web 文档，或者直接充当部分“爬虫”的角色。对于不规则的 HTML 文档，有补全功能。有了它，解析与分类网页就方便多了，节省了开发者的很多时间和精力。

安装 BeautifulSoup 很简单，在 Windows 平台和 Linux 平台上都是使用的传统第三方库安装方式。首先下载 BeautifulSoup 源码，其官网为：http://www.crummy.com/software/BeautifulSoup/。

然后解压后运行以下命令：

```
python setup.py install
```

此外，在 UBUNTU 下还可以使用系统包管理器安装。

```
$ apt-get install python-bs4
```

2.2.5　Neurolab 安装与配置

NeuroLab 是一个简单而强大的、用 Python 编写的神经网络库，包括基础神经网络、训练算法，并具有弹性的构架，可创建其他网络，它用纯 Python 和 numpy 写成。API 的使用与 MATLAB 的神经网络工具箱类似，具有弹性的网络配置和学习算法，可以改变神经网络和学习算法的类型、训练、误差、初始函数和激活函数等神经网络参数。

Windows 和 Linux 下的安装方式如下。

首先在下面的页面下载 Neurolab：

http://code.google.com/p/neurolab/downloads/list

然后解压后运行如下命令：

```
python setup.py install
```

2.2.6　R 安装与配置

R 的原始码可自由下载使用，也有已编译的执行档版本可以下载，可在多种平台下运行，包括类 UNIX（包含 FreeBSD 和 Linux）、Windows 和 MacOS。

WINDOWS 安装方式如下。

首先访问其官网下载页面：

http://ftp.ctex.org/mirrors/CRAN/

然后下载安装可执行文件安装即可。

UBUNTU 下的安装方式如下：

```
$  sudo apt-get update
$  sudo apt-get install r-base
$  sudo apt-get install r-base-dev
```

2.3 小结

“不要重复造轮子”（Stop Trying to Reinvent the Wheel），这可能是每个软件工程师入行时被告知的第一条准则。在轮子适合“机器学习”这台车的情况下，机器学习算法才能跑得更好，适合的科学计算平台就是机器学习的“轮子”。

笔者认为，作为机器学习这驾马车的“轮子”应该具备以下特征：

- 开源免费，且开源协议友好，例如：LGPL 协议或 BSD 协议，这样更有利于商业应用。
- 平台有文档，接口规范，能实际代码用例最好。
- 配置简单灵活，支持的操作系统平台多，运行速度快。
- 代码结构清晰、简单，便于使用者修改这个“轮子”，通俗地说：移植性强。

本书采用的计算平台在本章都一一列出其安装和配置方法。算法是一种计算思维的描述，万变不离其宗，好的工具原理都差不多。

也许随着时间的推移，算法在改进，更好的“轮子”将出现，所以不一定采用本书上所写的这些平台作为机器学习实验和应用的工具，但有一条原则：功能强大的计算平台不一定适合所有的工程，一切以适用为准。

第二部分

基础篇

合抱之木，生于毫末；九层之台，
起于累土；千里之行，始于足下。

——老子

第3章

机器学习数学基础

美国麻省理工学院的约翰·麦卡锡在1955年的达特矛斯会议上提出：人工智能就是要让机器的行为看起来就像人所表现出的智能行为一样。现代有一种观点，把人工智能分为了弱人工智能和强人工智能。维基百科是这样解释这两种智能的。

强人工智能：强人工智能观点认为有可能制造出真正能推理（Reasoning）和解决问题（Problem solving）的智能机器，并且，这样的机器能将被认为是有知觉的，有自我意识的。强人工智能可以有两类，类人的人工智能，即机器的思考和推理就像人的思维一样；非类人的人工智能，即机器产生了和人完全不一样的知觉和意识，使用和人完全不一样的推理方式。

弱人工智能：弱人工智能观点认为不可能制造出真正能推理和解决问题的智能机器，这些机器只不过看起来像是智能的，但是并非真正拥有智能，也不会有自主意识。

目前人类主要的精力放在了弱人工智能研究方面，而弱人工智能则主要依托数学理论来解决机器学习的问题。其实，部分所谓的强人工智能也建立在数学分析的基础上。 因此，大凡说到机器学习，总能看到一堆的数学公式来解说其算法，但讲解数据结构的书很少能看到数学公式，如果说数据结构描述了软件设计思维，那么机器学习就是描述了数学思维。

3.1 数学对我们有用吗

机器学习算法具有坚实的数学理论支持。机器学习的应用建立在科学计算的基础上，而数学计算则是科学计算的主要组成部分。在机器学习研究的各个领域（包括模式识别、数据分析、自动制造、专家系统等）都要应用到数学，数学是机器学习算法的基础。

数学是一切哲学、科学的基础，数学与软件是永远分不开的话题。我们应清醒地认识到，大数据时代已经到来，商业智能等机器学习技术开始普及，这些技术从大学的研究室和讲台走入了社会，应用到实际工程中了。因此，从某种意义上说：数学架起了从软件设计到智能计算的桥梁。

下面看看Common Lisp专家Peter Seibel对Google公司首席Java架构师Joshua Bloch

的部分访谈（Peter Seibel 写，郝培强翻译），里面谈到了数学与软件之间的关系，笔者认为这是目前网上流传的最权威的关于数学和软件关系的定义：

Seibel：你认识有哪位伟大的程序员不会数学或者没有接受过良好的数学教育吗？要成为一个程序员，学习微积分、离散数学和其他的数学知识真的那么重要吗？还是做程序员只需要一种思想方式，即使没有受过这些数字训练，也能拥有？

Bloch：我觉得是思想方式，学不学数学都能拥有这种思想。但是学一下确实有好处。我曾有个同事叫 Madbot Mike McCloskey。他很懂数学，但是没有学过数论。他重写了 BigInteger 的实现。原来的实现是 C 语言函数包的封装，他发誓用 Java 重写，要达到基于 C 语言版本的速度。后来他做到了。为此他学了大量的数论知识。如果他的数学不行，他肯定搞不定这个项目，而如果他本来就精通数论，就无需费力去学习了。

Seibel：但是，这本来就是个数学问题啊。

Bloch：对，这个例子不恰当。但是，我相信即使是跟数学无关的问题，学习数学培养出的思维方式对编程来说也是必不可少的。例如，归纳证明法和递归编程的关系非常紧密，你不理解其中一个，就不可能真正理解另外一个。你可能不知道术语的基本情况和归纳假设，但是如果你不能理解这些概念，你就没有办法写出正确的递归程序。所以，即使是在与数学无关的领域内，不理解这些数学概念的程序员也会遇到很多困难。

你刚才提到了微积分，我觉得它不那么重要。可笑的是这么多年来似乎已经成为了一种思维定势了，只要你受过大学教育，那么人们就认为你应该懂微积分。因为微积分中有很多美妙的思想，可以让人展开无穷的想象。

但是，你可以以连续或者离散这两种不同的方式思维。我觉得对程序员来说，精通离散思维更为重要。例如我刚提到的归纳证明法。你可以证明一种假设对所有整数都成立。证明过程就像施魔法一样。首先证明它对一个整数成立，然后证明针对这个整数成立意味着针对下一个整数也成立，这样就能证明它适用于全部整数。我认为对程序员来说这比理解极限的概念要重要得多。

好在我们无需选择。大学课程里这两样都教得不少。所以即使你用微积分用得没离散数学那么多，学校里还是会教授微积分的。但是我认为离散的东西比连续的东西更重要。

综上所述，程序员是以程序设计语言为工具，编程解决特定问题的，而编程的基础是计算机科学，计算机科学的基础是数学。数学对程序员的作用具体表现在以下方面：

1. 培养编程思想

目前流行的编程语言呈现高度同质化的趋势，只要学会某种语言，其他语言只是语法和接口的变化，但软件设计的编程思想却不尽相同。编写计算机程序的目的是为了解决实际问题，严谨的思维模式是高效解决问题的关键。数学天生具有严谨性，其基本要素是：逻辑和直观、分析和推理、共性和个性，这些也是编程思想的精髓。程序对问题的解决过程蕴含着数学思想方法。

美国著名数学教育家波利亚说过，用数学思维解决问题就意味着要善于解题，当我们解题时遇到一个新问题，总想用熟悉的题型去套，这只是满足于解出来，只有对数学思想、数

学方法理解透彻及融会贯通时，才能提出新看法和巧解法。因此，我们在软件编码时要有意识地应用数学思想去分析问题、解决问题，培养数学头脑。

2. 提高算法效率

数学与软件算法相辅相成，数学是软件算法的灵魂，软件算法是数学的工具。数学的证明与求解过程往往是推导计算的过程，很多推导计算过于复杂，必须用程序算法验证，数学推导和算法设计有着密切关系；计算机算法的过程需要数学语言进行描述，算法效果的评估也需要数学方法，特别复杂的算法效率评测甚至需要建立专门的数学模型。

《纽约时报》2011 年报道，著名的摩尔定律归纳了硬件技术进步的速度，而软件开发的突飞猛进推翻了这一定律。德国科学家和数学家马丁·格罗斯彻对 15 年之久的重要生产任务进展进行了调查，结果表明，在这 15 年里，运算完成速度提高了 4300 万倍，其中 1000 倍来自于处理器速度提高，43000 倍则来自软件算法效率的改进。

唐纳德·克努特（经典著作《计算机程序设计艺术》的作者，此书被认为与数学著作的《几何学原理》相当）是算法和程序设计技术的先驱者。他大学一年级的暑假接触到 IBM650，钻研使用手册后，他对数学产生了深厚的兴趣。一年后，他改学数学，从此与计算机结缘。他的第一个计算机程序也与数学相关：为他所在的校篮球队设计公式，根据球员在每场比赛中的得分、助攻、抢断、篮板球、盖帽等多项统计数字，对球员进行综合评估，并编写程序实现这个公式。

3.2 机器学习需要哪些数学知识

掌握机器学习算法至少需要以下几种数学的基本知识。

1. 微积分

微积分的诞生是继欧几里得几何体系建立之后的一项重要理论，它的产生和发展被誉为“近代技术文明产生的关键之一，它引入了若干极其成功的、对以后许多数学的发展起决定性作用的思想”。微积分学在科学、经济学和工程学领域有广泛的应用，解决了仅依靠代数学不能有效解决的问题。微积分学建立在代数学、三角学和解析几何学的基础上，包括微分学、积分学两大分支，包括连续、极限、多元函数的微积分、高斯定理等内容。

微积分在天文学、力学、化学、生物学、工程学、经济学、计算机科学等领域有着越来越广泛的应用，比如：在医疗领域，微积分能计算血管最优支角，将血流最大化；在经济学中，微积分可以通过计算边际成本和边际利润来确定最大收益；微积分可用于寻找方程的近似值；通过微积分解微分方程，计算相关的应用，比如，宇宙飞船利用欧拉方法来求得零重力环境下的近似曲线等。

在机器学习和数据分析领域，微积分是很多算法的理论基础，如：多层感知器神经网络算法。多层感知器是一种前馈人工神经网络模型，算法分为两个阶段：正向传播信号、反向传播误差。

正向传播信号阶段是对样本的学习阶段，输入的信息从输入层传入，经各个隐层计算后

传至输出层，计算每个单元的实际值，向各层各单元分摊产生的误差；反向传播误差阶段通过网络输出与目标输出的误差对网络进行修改审查，将正向输出的误差再传播回各层进行权重值调整，直到误差最小化或达到规定的计算次数。微积分理论在多层感知器模型中运用较多，下面是 3 个应用的例子。

（1）非线性激活函数是中间隐藏层的精髓，由于这些非线性函数的帮助，神经网络才能对线性和非线性模型进行学习，训练成功的网络能对待解决问题进行拟合和仿真。非线性激活函数要求处处可微，主要有 Logistic 函数和双曲正切函数。

Logistic 函数定义为：

$$\varphi(t) = \frac{1}{1+\exp(-a \cdot t)}$$

双曲正切函数定义为 :

$$\varphi(t) = \tanh(a \cdot t)\ \frac{e^{a \cdot t}-e^{-a \cdot t}}{e^{a \cdot t}+e^{-a \cdot t}}$$

（2）权值更新规则。权值更新规则定义为：

$$w(n+1)=w(n)+\eta \cdot \Delta w$$

其中 ΔW 基于最速下降法，定义为：

$$\Delta w=-\frac{\partial \varepsilon(w)}{\partial w}=-e(n)\ \frac{\partial e(n)}{\partial w}$$

（3）神经元局部梯度。梯度是一个向量场，标量场中某一点上的梯度指向标量场增长最快的方向，梯度的长度是这个最大的变化率。神经元局部梯度定义为：

$$\delta_j(n)=-\ \frac{\partial \varepsilon(n)}{\partial v_j(n)}$$

2. 线性代数

线性代数是高等数学中的一门成熟的基础学科，它内容广泛，不但包含行列式、矩阵、线性方程组等初等部分，而且包括线性空间、欧式空间、酉空间、线性变换和线性函数、λ–矩阵、矩阵特征值等更深入的理论，线性代数在数学、物理学、社会科学、工程学等领域也有广泛的应用。

线性代数理论是计算技术的基础，在机器学习、数据分析、数学建模领域有着重要的地位，这些领域往往需要应用线性方程组、矩阵、行列式等理论，并通过计算机完成计算。下面是几个应用线性代数的例子。

（1）人口模型描述人口系统中人的出生、死亡和迁移随时间变化的情况，以及它们之间定量关系的数学方程式或方程组，分为连续模型和离散模型。其中离散模型适合于计算机仿真。在人口离散模型中，用 $x_0(t), x_1(t), x_2(t), \cdots, x_m(t)$ 表示 t 时刻的年龄构成，其中 $x_i(t)$ 表示 t 年代年满 i 周岁但不到 i+1 周岁的人口数，写成向量形式如下：

$$x(t)=\begin{Bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_m(t) \end{Bmatrix}$$

则离散人口模型可写成：

$$\left.\begin{aligned} x(t+1)=H(t)x(t)+\beta(t)B(t)x(t)+g(t) \\ x(t_0)=x_0 \end{aligned}\right\}$$

式中 $H(t)$，$B(t)$ 为以下相应维数的矩阵：

$$H(t)=\begin{Bmatrix} 0 & & & & \\ 1-\mu_1(t) & 0 & 0 & & \\ & 1-\mu_2(t) & 0 & & \\ 0 & & \ddots & \ddots & \\ & & & 1-\mu_{m-1}(t) & 0 \end{Bmatrix}$$

$$B(t)=\begin{Bmatrix} 0\cdots b_{a1}(t) & b_{a1+1}(t)\cdots b_{a2}(t) & 0\cdots 0 \\ 0 & & \end{Bmatrix}$$

式中 $\mu_i(t)(i=0, 1, \cdots, m-1)$ 称为按龄死亡率，m 为人类能活到的最高年龄。

在这个模型中，通过矩阵的形式，将时间、出生、死亡和迁移 4 个因素及它们之间的定量关系进行完全描述。

（2）投入产出技术是研究一个经济系统各部门间的“投入”与“产出”关系的数学模型，该方法最早由美国著名的经济学家瓦·列昂捷夫提出，是目前比较成熟的经济分析方法。投入产出数学模型根据投入产出原理建立的经济数学模型，揭示国民经济各部门、再生产各环节之间的内在联系，进行经济分析、预测和安排预算计划。

投入产出分析通常从投入产出表分析开始。投入产出表以数学方程式的形式来反映客观经济运行过程和经济结构，它是根据投入产出表所反映的经济内容，利用线性关系而建立起来的两组线性方程组。其中，行模型根据投入产出表的横行关系建立经济数学模型，其经济含义是揭示国民经济各部门生产的货物和服务的使用去向，研究产出分配问题；列模型根据投入产出表的纵列建立经济数学模型，其经济含义是揭示国民经济各部门生产经营过程中发生的各种投入，研究国民经济各部门生产货物和服务的价值形成问题。

（3）自回归模型是统计上一种处理时间序列的方法，从回归分析中的线性回归发展而来，用同一变量例如 x 的前期进行预测（即 x_1 至 x_{t-1} 预测本期 x_t 的表现），并假设它们为线性关系，模型中 X 的当期值等于若干个后期值的线性组合，加常数项，加随机误差，其公式定义为：

$$X_t=c+\sum_{i=1}^{p}\varphi_i X_{t-i}+\varepsilon_t$$

其中：c 是常数项；ε_t 被假设为平均数等于 0、标准差等于 σ 的随机误差值；σ 被假设为对于任何的 t 都不变。

（4）支持向量机 SVM（Support Vector Machine）是一种小样本的机器学习方法，它通过非线性映射，把样本空间映射到一个高维乃至无穷维的 Hilbert 特征空间中，使得非线性可分转化为在特征空间中的线性可分。SVM 方法的理论证明、核函数设计等都需要线性代

数理论的支持。

SVM 定义了以下两个超平面：

$$w \bullet x - b = 1$$

$$w \bullet x - b = -1$$

试图使它们之间没有任何样本点，且这两个超平面之间的距离最大，这样分类问题转变为二次规划最优化问题，在约束条件下最小化 $|w|$；然后用标准二次规划技术标准和程序解决，最终表示为以下训练向量的线性组合：

$$w = \sum_{i=1}^{n} \alpha_i c_i x_i$$

式中大于 0 的 α_i 对应的 x_i 就是支持向量。

3. 概率论

概率论是研究随机性或不确定性现象的数学，用来模拟实验在同一环境下会产生不同结果的情况。下面这些概率理论是概率论的基础。

（1）古典概率。拉普拉斯试验中，事件 A 在事件空间 S 中的概率 $P(A)$ 为：

$$P(A) = \frac{\text{构成事件 } A \text{ 的元素数目}}{\text{构成事件空间 } S \text{ 的所有元素数目}}$$

■ 条件概率

一事件 A 在一事件 B 确定发生后会发生的概率称为 B 给之 A 的条件概率，定义为：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

■ 概率公理

公理 1：$0 \leqslant P(A) \leqslant 1 (A \in S)$

公理 2：$P(S)=1$

公理 3：$P(A \cup B)=P(A)+P(B)$，如果 $A \cap B=0$

（2）概率分布包括二项分布、几何分布、伯努利分布、泊松分布、均匀分布、正态分布、指数分布等。样本空间随机变量的概率分布可用累积分布函数和概率密度函数进行分析。

随机变量 X 的累积分布函数定义为：

$F(x)=P(X \leqslant x)$ 其中，x 为任意实数

设 X 为连续型随机变量，其累积分布函数为 $F_X(x)$，若存在 $f_X(x)$，满足：

$$\forall -\infty < a < \infty, F_X(a) = \int_{-\infty}^{a} F_X(x)\mathrm{d}x$$

则 $f_X(x)$ 是它的概率密度函数。

概率论在机器学习和数据分析领域有举足轻重的地位，比如马尔可夫链理论。马尔可夫链对于现实世界的很多现象都给出了解释，泊松过程是连续时间离散状态的马尔可夫链，布朗运动是连续时间连续状态的马尔可夫链等。

马尔可夫链在计算数学、金融经济、机器学习、数据分析等领域都有重要的应用，马尔

可夫链是数学中具有马尔可夫性质的离散时间随机过程，在给定当前知识或信息的情况下，仅使用当前的状态预测将来，在马尔可夫链的每一步，系统根据概率分布，从一个状态变到另一个状态或保持当前状态。

马尔可夫链是随机变量 $X_1, X_2, X_3, \cdots$ 的一个数列，这些变量所有可能取值的集合为状态空间，X_n 的值则是在时间 n 的状态。设 X_{n+1} 对于过去状态的条件概率分布可定义如下：

$$P(X_{n+1}=x|X_0, X_1, X_2, \cdots, X_n)=P(X_{n+1}=x|X_n)$$

同理，可计算更多步的转移概率：

$$p(X_{n+2}|X_n)=\int P(X_{n+2}, X_{n+1}|X_n)\mathrm{d}X_{n+1}=\int P(X_{n+2},|X_{n+1}|)P(X_{n+1},|X_n)\mathrm{d}X_{n+1}$$

$$p(X_{n+3}|X_n)=\int P(X_{n+3}, X_{n+2})\int P(X_{n+2}|X_{n+1}|)P(X_{n+1}|X_n)\mathrm{d}X_{n+1}\mathrm{d}X_{n+2}$$

4. 统计学

统计学是收集、分析、表述和解释数据的科学，作为数据分析的一种有效工具，统计方法已广泛应用于社会科学和自然科学的各个领域。统计学与概率论联系紧密，前者以后者为理论基础。统计学主要分为描述统计学和推断统计学。描述统计学描绘或总结观察量的集中和离散情形，基础的数学描述包括了平均数和标准差等；推断统计学将资料中的数据模型化，计算它的机率并且做出对于母群体的推论，主要包括假设检定、对于数字特征量的估计、对于未来观察的预测、相关性预测、回归、变异数分析、时间序列、数据挖掘等。

无论是描述统计学还是推断统计学都是数据分析技术的基础。通过描述统计学方法，数据分析专家能对数据资料进行图像化处理，将资料摘要变为图表，分析数据分布特征。此外，还可以分析数据资料，以了解各变量内的观察值集中与分散的情况等。通过推断统计学方法，对数据未知特征做出以概率形式表述的推断，在随机抽样的基础上推论有关总体数量特征。

5. 离散数学

离散数学是数学的几个分支的总称，研究基于离散空间而不是连续的数学结构，其研究内容非常广泛，主要包括数理逻辑、集合论、信息论、数论、组合数学、图论、抽象代数、理论计算机科学、拓扑学、运筹学、博弈论、决策论等。

离散数学广泛应用于机器学习、算法设计、信息安全、数据分析等领域，比如：数理逻辑和集合论是专家系统的基础，专家系统是一类具有专门知识和经验的计算机智能程序系统，一般采用人工智能中的知识表示和知识推理技术，模拟通常由领域专家才能解决的复杂问题；信息论、数论、抽象代数用于信息安全领域；与信息论密切相关的编码理论可用来设计高效可靠的数据传输和数据储存方法；数论在密码学和密码分析中有广泛应用，现代密码学的 DES、RSA 等算法技术（包括因子分解、离散对数、素数测试等）依赖于数论、抽象代数理论基础；运筹学、博弈论、决策论为解决很多经济、金融和其他数据分析领域的问题提供了实用方法，这些问题包括资源合理分配、风险防控、决策评估、商品供求分析等。

以上是机器学习需要的核心数学知识，但不是全部知识。随着今后人类对机器学习的深入研究，将有更多的数学分支进入机器学习领域。因此，仅掌握大学数学知识是不够的，还需要向更高层次进军，对于非数学专业毕业的朋友来说，还应该学习其他数学分支理论，比

如说泛函分析、复变函数、偏微分方程、抽象代数、约束优化、模糊数学、数值计算等。

建议读者购买以下数学书籍，随时翻阅参考。

Finney，Weir，Giordano.《托马斯微积分》. 叶其孝，王耀东 , 唐兢译 . 第 10 版 . 北京：高等教育出版社 2003-1

Steven J.Leon.《线性代数》. 张文博，张丽静译 . 第 8 版 . 北京：机械工业出版社

William Mendenhall 等 .《统计学》. 梁冯珍，关静译 . 第 5 版 . 北京：机械工业出版社

Dimitri P. Bertsekas 等 .《概率导论》. 郑忠国，童行伟译 . 第 2 版 . 北京：人民邮电出版社

Kenneth H.Rosen 等 .《离散数学及其应用》. 袁崇义，屈婉玲，张桂芸译 . 第 6 版 . 北京：机械工业出版社

Eberhard Zeidler 等 .《数学指南：实用数学手册》. 李文林译 . 北京：科学出版社

它们都是机器学习所涉及的经典数学书，可以考虑将它们和《设计模式》、《算法导论》、《深入理解计算机系统》等经典算法书放在一起，作为案头必备书。

3.3　小结

机器学习以数学理论为基础，这里的数学理论主要是应用数学。应用数学是应用性较强的数学学科或分支的统称，数学本来起源于实际应用的需要，应用一直是数学的发展动力之一，一种数学理论和一门数学学科的生命力的强弱，在很大程度上依赖于它有无应用需求，机器学习就是数学的应用领域之一。

应用数学是应用目的明确的数学理论和方法的总称，是纯数学（纯数学研究数学本身，不以应用为目的，以其严格、抽象和美丽著称，主要研究空间形式的几何类、离散系统的代数类、连续现象的分析类）的相反，包括微分方程、向量分析、矩阵、拉普拉斯变换、傅里叶变换、复变分析、数值方法、概率论、数理统计、运筹学、控制理论、组合数学、信息论等许多数学分支，也包括从各种应用领域中提出的数学问题的研究。随着计算机技术的发展，现在计算数学也加入了应用数学的行列。

一位 MIT 的牛人在 BLOG 中曾提到，数学似乎总是不够的，为了解决和研究工程中的一些问题，不得不在工作后，重新回到图书馆捧起了数学教科书。他深深感到，从大学到工作，课堂上学的和自学的数学其实不算少了，可是在机器学习领域总是发现需要补充新的数学知识。看来，要精通机器学习知识，必须在数学领域学习、学习、再学习，这一切都是很艰苦的。要学好机器学习必须做好艰苦奋斗的准备，坚持对数学知识的追求。

本章对机器学习中需要掌握的相关数学知识提出了要求，同时也推荐了有关数学书籍。不要因为学习数学麻烦、难度大就不去接触它，数学才是工程师软实力的体现。古人云：“磨刀不误砍柴工”，这个“刀”就是数学知识。

对于初学者的入门教材而言，本书是一个不错的选择，由浅入深，通俗易懂。本书不在理论上做过多的纠结，讲究“make hands dirty”,在对于机器学习几个主要问题有接触的基础上，引入实际工程上的样例，使得读者能较快地在机器学习的相关职位中上手。从这个方面来说，此书是在快速发展的大数据时代中一本值得一读的书。

—— **吴炜** 淘宝 资深算法工程师

本书从耐心帮助读者了解“什么是机器学习”开始，由浅入深地讲解机器学习算法，最后用大量篇幅讲述了算法的应用，并附有完整代码，代码能实际运行。这本书适合不同层次的读者，它能切实地帮助程序员理解机器学习，并将相关算法应用于实践工作中。

—— **樊恒光** 杭州阿里科技有限公司 一搜产品线工程师

本书是机器学习实践爱好者的盛宴。作者利用计算机上的数学计算工具，循序渐进地展示了与之相关的基本算法及其实现过程。本书的特色在于，书中提供的大量与机器学习应用相关的实践案例，包括统计分析和模式识别的基本方法，都具有较强的代表性，正所谓实践出真知，本书是机器学习实践的宝典，为涉足统计方法和机器学习实践的爱好者指明了一条简单易行的求知之路。

—— **宋翼** 中国科学院自动化研究所 在读博士

全书把机器学习的编程语言、数学理论和实例有机结合，是一本值得你拥有的参考书，极具工程价值。书中的朴素贝叶斯分类方法对我来说特别有感触，之前参加过数据挖掘比赛，将此方法应用于推荐系统，具有较高的准确率和鲁棒性等特点。此外，机器学习与嵌入式技术的关系越来越密切，智能物联网正在走近，相信未来的几年会走近千家万户。

—— **方家挺** 自仪股份 嵌入智能工程师

机器学习是一门复杂且难以找到突破口的学科，此书结合实际应用，让读者更容易踏入机器学习的大门，参考本书的算法，应用在实际工作中，帮助实现金融智能化。

—— **欧阳星** 日信证券 软件工程师

这本书使难懂的机器学习理论变得浅显易懂，书中没有其他教科书上枯燥无味的理论，而是通过实践与应用讲解，让工程师们很快进入数据分析与机器学习领域。

—— **张恒** 多贝网 运营经理

上架指导：计算机/机器学习
ISBN 978-7-111-46207-1
9 787111 462071 >
定价：69.00元

投稿热线：(010) 88379604
客服热线：(010) 88378991 88361066
购书热线：(010) 68326294 88379649 68995259

华章网站：www.hzbook.com
网上购书：www.china-pub.com
数字阅读：www.hzmedia.com.cn