Springer

AN INTRODUCTION TO MACHINE LEARNING

机器学习导论

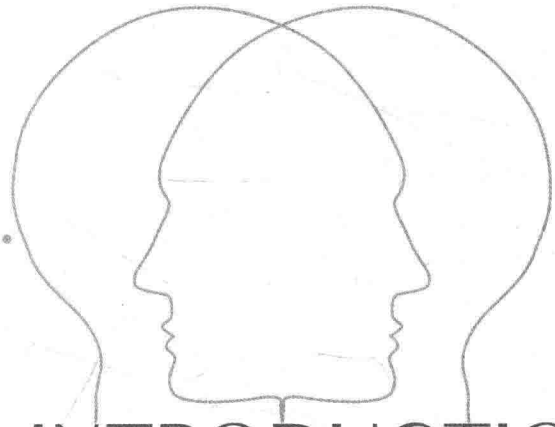
[美] 米罗斯拉夫·库巴特（Miroslav Kubat）◎著
王勇　仲国强　孙鑫◎译

人工智能专家**米罗斯拉夫·库巴特教授**25年倾心打造

系统解读了有关机器学习的14个方面，快速读懂机器学习

全面揭开机器学习的奥秘

机械工业出版社
CHINA MACHINE PRESS



AN INTRODUCTION TO MACHINE LEARNING

机器学习导论

[美] 米罗斯拉夫·库巴特（Miroslav Kubat）◎著
王勇　仲国强　孙鑫◎译

机械工业出版社
CHINA MACHINE PRESS

本书通过给出易操作的实践指导、采用简单的案例、激励学生讨论有兴趣的应用问题，用一种易于理解的方式介绍了机器学习的基本思想。本书主题包括贝叶斯分类器、最近邻分类器、线性和多项式分类器、决策树、神经网络以及支持向量机，且展示了如何把这些简单工具通过“提升”（Boosting）的方式结合起来，怎样将它们应用于更加复杂的领域，以及如何处理各种高级的实践问题。书中对广为人知的遗传算法也做了介绍。

图书在版编目（CIP）数据

机器学习导论／（美）米罗斯拉夫·库巴特（Miroslav Kubat）著；王勇，仲国强，孙鑫译. —北京：机械工业出版社，2016.9
书名原文：An Introduction to Machine Learning
ISBN 978－7－111－54868－3

Ⅰ. ①机… Ⅱ. ①米… ②王… ③仲… ④孙… Ⅲ. ①机器学习—研究 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字（2016）第 222726 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：坚喜斌　　责任编辑：坚喜斌　陈瑞文
版式设计：张文贵　　责任校对：赵蕊
责任印制：常天培
涿州市京南印刷厂印刷
2016 年 11 月第 1 版·第 1 次印刷
170mm×240mm·20.5 印张·303 千字
标准书号：ISBN 978－7－111－54868－3
定价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
电话服务　　网络服务
服务咨询热线：（010）88361066　　机 工 官 网：www.cmpbook.com
读者购书热线：（010）68326294　　机 工 官 博：weibo.com/cmp1952
（010）88379203　　教育服务网：www.cmpedu.com
封面无防伪标均为盗版　　金 书 网：www.golden-book.com

献给我的妻子，瓦卢卡（Verunka）

推荐序

机器学习是人工智能领域的一个重要分支，其研究涉及代数、几何、概率统计、优化、泛函分析、图论、信息论、算法、认知计算等多个学科的知识，其应用不仅仅限于模式识别、计算机视觉、数据挖掘、生物信息学、智能控制等科学和工程领域，甚至在社会科学领域的研究中也有应用，如管理学、经济学和历史学等。目前，随着计算机科学和智能科学技术的进步，机器学习得到了快速发展，其方法被广泛应用到了各个领域。尤其是近些年，深度学习方法快速发展并在多个领域展示出优异性能，使机器学习和整个人工智能领域受到极大的关注。

机器学习是基于已有数据、知识或经验来设计模型或发现新知识的一个研究领域。20 世纪 50 ~ 70 年代是机器学习研究的初期，人们基于逻辑知识表示试图给机器赋予逻辑推理能力，取得了很多振奋人心的成果；20 世纪 80 年代，专家系统受到高度重视，为专家系统获取知识成为一个重要方向。20 世纪 80 年代中后期，人工神经网络由于误差反向传播（BP）算法的重新提出和广泛应用而形成一股热潮，但其地位在 90 年代后期被以支持向量机为核心的统计学习理论所取代。20 世纪 90 年代以后，受重视的机器学习方法还有集成学习、概率图模型、半监督学习、迁移学习等。2006 年，以加拿大多伦多大学的 G. Hinton 教授为代表的几位研究人员在深度学习方面取得巨大突破，在 Google、Microsoft、Facebook 等科技公司的推动下，深度学习借助于大数据和高性能计算的有利条件得到了广泛应用和高度关注。目前，搜索引擎、机器人、无人驾驶汽车等高科技产品都依赖于机器学习技术。机器学习，特别是深度学习，在语音识别、人脸识别、围棋、游戏等方面已经超过了人类水平，可以想象机器学习与人类的生产、生活之间的关系将会越来越紧密。

过去几十年，机器学习领域也出现了一些经典的著作或教材。1983 年，R. Michalski、J. Carbonell 和 T. Mitchell 主编的《机器学习：一种人工智能方法》一书出版，标志着机器学习成为人工智能的一个独立研究领域。《Machine Learning》期刊创刊于 1986 年，目前依然是机器学习领域的顶级期刊。1990 年，J. Carbonell

主编的《机器学习：范式与方法》对归纳学习、基于解释的学习、遗传算法和连接主义学习等机器学习范式及方法进行了深入探讨。T. Mitchell 于 1997 年出版的《机器学习》是一本经典的机器学习教材，其中文版已于 2003 年由机械工业出版社出版，但因为出版年限较早，许多内容已没有时效性。1998 年，V. Vapnik 出版的《统计学习理论》是一本完整阐述统计机器学习思想的名著。2001 年出版、2009 年再版的《统计学习基础：数据挖掘，推理和预测》是美国斯坦福大学教授 T. Hastie，R. Tibshirani 和 J. Friedman 的一部力作，其中对最为流行的机器学习方法进行了全面而深入的介绍，因其严谨的数学推导，该书不失为机器学习研究进阶的很好的读物。E. Alpaydin 所著的《机器学习导论》出版于 2004 年并于 2010 年再版，书中对基础的机器学习方法进行了介绍，是一本机器学习入门的很好的教材。C. Bishop 所著的《模式识别与机器学习》和 K. Murphy 所著的《机器学习：一个概率的视角》分别于 2006 年和 2012 年出版，两本书都从概率的角度全面而细致地介绍了许多经典的机器学习模型。C. Bishop 的《模式识别与机器学习》可帮助读者打下坚实的机器学习基础，而 K. Murphy 的书则相对介绍了更多较新的机器学习算法，甚至有一章专门介绍了深度学习方法。2012 年，李航老师出版了《统计学习方法》，2016 年，周志华老师出版了《机器学习》。这两本书中，《统计学习方法》主要集中于几种重要机器学习模型的介绍，而《机器学习》内容相对更加全面，深入浅出，堪称机器学习的中文经典著作。相对于以上这些机器学习书籍，M. Kubat 所写的这本《机器学习导论》更像是一本科普性质的读物，作者尽量避开复杂的数学公式，用生动形象的方式介绍机器学习算法，而且本书篇幅适当，又涵盖了几乎所有基本的机器学习方法，使得本书不仅适合作为本科学生机器学习课程的教材，也适合于想了解机器学习入门知识的普通读者。

本书的译者都是工作在机器学习教学与研究第一线的年轻老师，其中仲国强副教授过去是我的博士研究生，在模式识别和机器学习领域都有很扎实的研究基础。相信本书的中译本对于国内机器学习的教学和研究都会有所裨益，也为更多的人，尤其是初学者了解机器学习打开一扇门。

中国科学院自动化研究所副所长、模式识别国家重点实验室主任

刘成林

前　言

目前，机器学习慢慢走向成熟。你可能觉得这只是老生常谈，请让我做一个详细说明。

人们希望机器某一天能够自己学习，这个梦想几乎在计算机出现时就有了，也许更早。不过，长久以来，这仅仅是一个想象而已。罗森布拉特（Rosenblatt）感知器的提出曾经掀起过一股热潮，但是现在回想起来，这股热潮没能持续很长的时间。至于接下来的尝试，使情况发展得更糟糕，这个领域甚至没有再引起人们的注意，长期被忽视，因而无法取得重大突破，也没有这一类的软件公司，后续研究寥寥无几且得到的资金支持也不多。这个阶段，机器学习一直不被看好，像进入休眠期一样，在其他成功学科的阴影里生存。

然而，接下来发生的一切使这些颓势彻底改变了。

一群有识之士指出，在20世纪70年代的人工智能领域，基于知识的系统曾经风靡一时，但它们有一个弱点："知识"从哪里来？当时主流的观点认为，应该让工程师和领域专家合作，用if-then的形式表示出来。但是实际情况差强人意，专家们发现很难把掌握的知识表达给工程师。反过来，工程师也不知道该问什么问题以及如何表示答案。尽管有几个广为人知的成功案例，但是其他大多数研究都试图建立知识库，并且成千上万的规则令人沮丧。

这些有识之士主张简单和直接的操作。如果难以准确地告诉机器如何处理某个问题，那么为什么不间接地给出指令，通过例子展示所需要的技能，计算机将通过这些例子来学习！

当然，这必须要有能够进行学习的算法才有意义，这也是困难所在。无论是罗森布拉特的感知器还是后来出现的技术都不太管用。然而，机器学习在技术方面的缺乏算不上是障碍，相反是一个挑战，并激发出了很多绝妙的点子。其中，使计算机有学习能力这个想法开创了一个激动人心的新领域，并引起了世人的

关注。

这一想法在1983年爆发了。一卷很厚的论文集——《机器学习：人工智能的方法》[㊀]中提出了很多各式各样的方法来求解这个谜题。在它的影响下，几乎一夜之间一个新的学科诞生了。3年后，后续著作一本接一本地出现。相关学术刊物也很快被创立，有着巨大影响力的年度学术会议相继召开。几十、或许是几百篇博士论文完成并通过答辩。

早期阶段，问题不仅是如何学习，而是学什么和为什么学。这段充满创造力的岁月让人难以忘怀，唯一有些遗憾的是很多非常好的想法后来被放弃了。实用主义占了上风，资源都被投向那些最有希望的方向。经过一段时间的发展，具体研究基本成形：知识系统if-then规则的归纳，分类归纳，程序基于经验来提高技能，Prolog程序自动调优，以及其他方面。相关的研究方向非常多，一些知名学者希望通过写书来引领未来的发展，这其中有些人做得很成功。

机器学习发展的一个重要的转折点是汤姆·米切尔（Tom Mitchell）的传奇教科书[㊁]。该书向博士生和科学家们总结了该领域的发展现状，慢慢地大学也用这本书作为研究生的教材。同时，研究方法也变得更加系统化。大量机器学习测试库被建立起来，用于比较性能或者学习算法的优劣。统计评估方法也被广泛地使用在评估过程中。相关流行程序的公开版本很容易获得，从事这个学科的人数增至数千，甚至更多。

现在，到了很多大学都为本科生开设机器学习课程的阶段。通常这些课程需要不同类型的教材。除了掌握基本技术以外，学生还需要了解不同方法的优点和缺点，以及不同情况下每种方法的独特之处。最重要的是，他们需要理解在特定情况下，哪些技术是可行的，哪些是不可行的。只有这样才能在解决具体问题时做出正确的选择。一本教材除了满足以上的各项要求外，还应该介绍一些数学概念，多包括一些实用的建议。

㊀ 米切尔斯基（R. Michalski），卡波内尔（J. Carbonell），米切尔（T. Mitchell）编辑。

㊁ T. Mitchell. Machine Learning [M]. New York: McGraw-Hill, 1997.

关于教材，还要考虑材料的多少、结构以及风格，以便能够支持一个学期的导论课程。

第一个问题是材料的选择。当高科技公司准备成立机器学习研究团队时，大学就要向学生传授相应的知识和技能，以及对有关行业需求的理解。出于这个原因，本书重点介绍了贝叶斯分类器，最近邻分类器，线性和多项式分类器，决策树，神经网络的基础，以及提升（Boosting）算法的原理。本书用很大篇幅来描述具体应用的典型特征。在现实中，当面对有一定难度的任务时，一些基本方法和老师在实验环境下演示的结果可能不完全一样。因此在学习过程中，学生必须知道每种方法会发生什么。

本书共包括 14 章，每章覆盖一个专题。各章分成很多个小节，每节介绍一个关键问题。建议学生在做完每一节后面的 2 ~ 4 个“控制问题”后再学习下一节。这些问题用来帮助检查学生对学习材料的掌握情况。如果不会做这些题，则有必要重新阅读相关内容。

俗话说，实践出真知。每章结尾安排了必要的练习用于实际操作。如果接下来的思考实验能够全部完成，将有助于更深入地理解所学内容的各个方面。不过这些实验难度较大，只有付出很大努力才能获得正确的答案。所学的知识在上机实验中可被进一步巩固。编程对于学习同样也很重要。现在，人们都习惯从网上下载所需的程序，这是捷径，但本书不建议这样做，因为只有具体实现了程序的全部细节，才能领会机器学习技术的精妙之处。

目　录

机器学习导论

第1章　一个简单的机器学习任务

你会发现精确地描述你母亲的相貌，让朋友能在超市里认出她，是很难的。但如果你给他看几张你母亲的照片，他就能立刻找出所需要的特征。这就是人们常说的，一张图片，也就是一个样例，胜过千言万语。

这就是我们希望用技术去实现的。当不能足够精确地定义某些事物或概念时，我们希望以样例的方式把它们传输给机器。然而，计算机必须能将样例转换成知识才能奏效。所以，我们的兴趣在于机器学习（machine learning）的算法和技术，也是这本书的主题。

第1章将任务表示为一个搜索问题，并介绍了爬山搜索算法。爬山搜索算法不仅是我们解决机器学习任务的初步尝试，而且也为解决后面几章中的一些辅助问题提供便利的工具。在这些基础上，我们将继续探索一些能使学习过程苦中带乐的问题，包括性能标准、实验方法以及其他方面。

1.1 训练集和分类器

让我们首先介绍一些贯穿全书的问题和基本概念。

预分类训练样例。图1.1展示了6种乔尼（Johnny）喜欢和不喜欢的派。这些正例（positive examples）和负例（negative examples）的基本概念构成了一个训练集（training set），并以此由机器归纳出一个分类器（classifier）——一种能将今后任何的派归为正、负两个类别之一的算法。

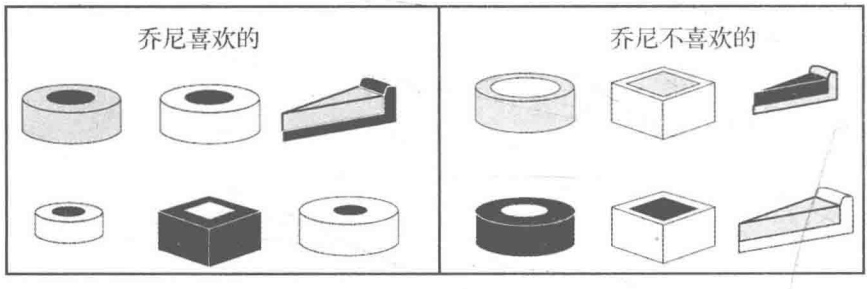


图1.1 一个简单的机器学习任务：归纳学习出一个能将今后任何派归类为正例和负例的分类器，如归纳学习一个“乔尼喜欢的派”的分类器