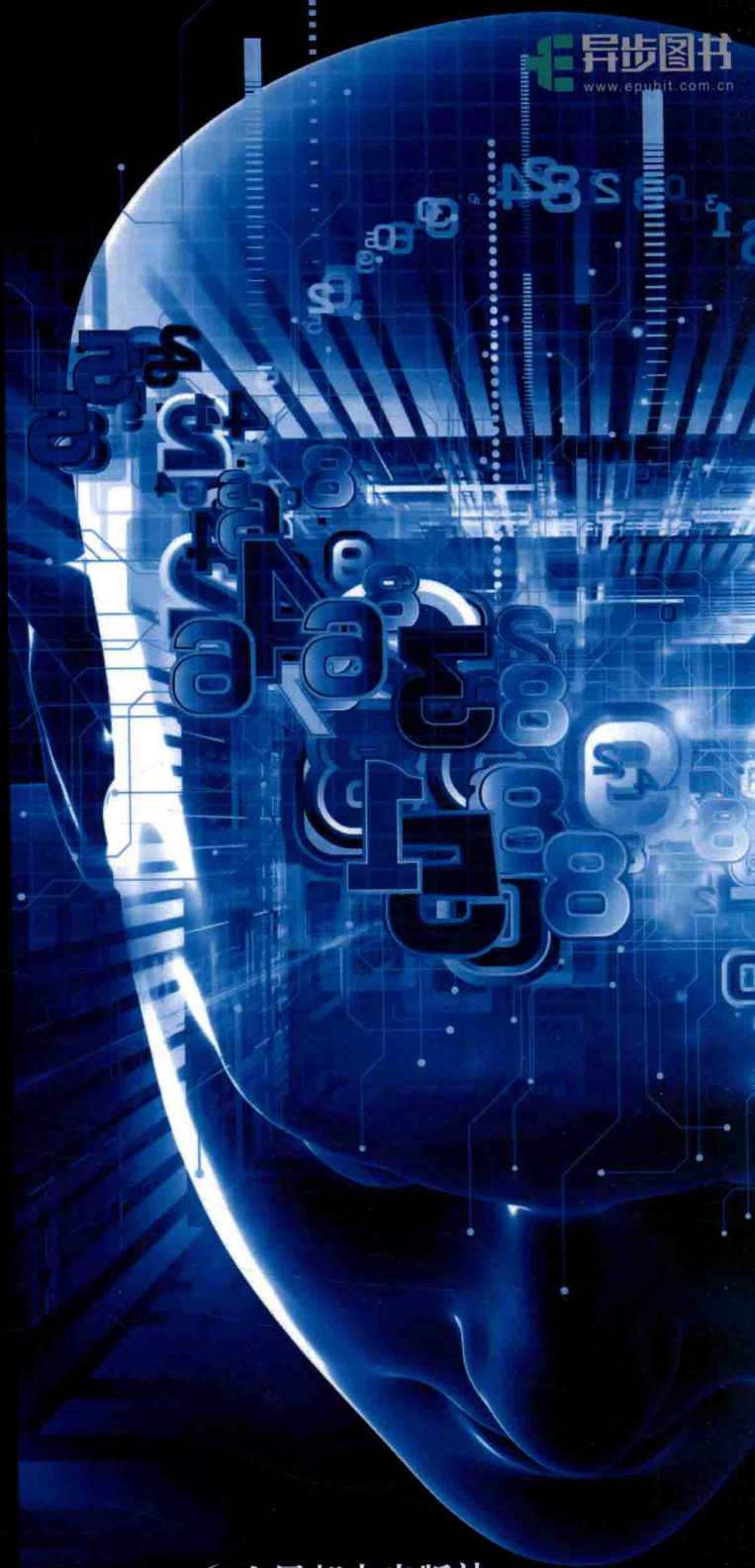
机器学习实践应用

李博 著

人工智能，触手可及。
让数据起舞，用算法扩展业务边界。



中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

机器学习实践应用

李博 著

人民邮电出版社
北京

图书在版编目（CIP）数据

机器学习实践应用 / 李博著. -- 北京 : 人民邮电出版社, 2017.7
ISBN 978-7-115-46041-7

Ⅰ. ①机… Ⅱ. ①李… Ⅲ. ①机器学习－研究 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2017)第114073号

内 容 提 要

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度等多门学科，专门研究计算机怎样模拟或实现人类的学习行为。机器学习是人工智能的核心，是使计算机具有智能的根本途径。

本书通过对机器学习的背景知识、算法流程、相关工具、实践案例以及知识图谱等内容的讲解，全面介绍了机器学习的理论基础和实践应用。书中涉及机器学习领域的多个典型算法，并详细给出了机器学习的算法流程。

本书适合任何有一定数据功底和编程基础的读者阅读。通过阅读本书，读者不仅可以了解机器学习的理论基础，也可以参照一些典型的应用案例拓展自己的专业技能。同时，本书也适合计算机相关专业的学生以及对人工智能和机器学习感兴趣的读者阅读。

◆ 著　　李　博
责任编辑　胡俊英
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：17.5
字数：328 千字　　2017 年 7 月第 1 版
印数：1－3 000 册　　2017 年 7 月北京第 1 次印刷

定价：69.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

推荐序

近年来，在 IT 圈大家谈论最多的就是人工智能。AlphaGo 与围棋选手的人机大战更是让我们领略到人工智能技术巨大潜力的同时，又将人工智能推向了一个新的制高点。

人工智能的发展得益于云计算和大数据技术的成熟与普及。和人工智能相关的还有两个核心词汇——机器学习和深度学习。这三者有着什么样的关系？所谓人工智能，通俗地讲是指由人工制造出来的系统所表现出来的智能。人工智能研究的核心问题包括推理、知识、交流、感知、移动和操作物体的能力。而机器学习是人工智能的一个分支，很多时候机器学习几乎成为人工智能的代名词。机器学习简单来讲就是通过算法，使机器能从大量历史数据中学习规律，从而对新的样本做出智能识别或对未来做预测。深度学习是机器学习的一个新领域。之所以称为“深度”，是因为前面说的机器学习是浅层的学习，主要基于概率统计、矩阵或图模型而得出的分析结论。深度学习的概念源于人工神经网络的研究，它基于神经网络框架，通过模拟人脑学习的方式来处理数据。在人工智能实践中，数据是载体和基础，智能是追求的目标，而机器学习则是从数据通往智能的技术桥梁。因此，在人工智能领域，机器学习才是核心，是现代人工智能的本质。

人工智能的火热使市场上对机器学习人才的需求不断提高，很多从事软件开发的程序员纷纷转行投向机器学习领域。但机器学习对人才的技术和理论水平要求都非常高，除了要掌握统计学中各种复杂的机器学习算法的理论推导外，还要懂计算机算法的实现逻辑以及分布式、并行化等架构理论。

本书是以应用场景为导向，以代码实现为样例贯穿始终，并融入了通俗易懂的理论知识。对于机器学习爱好者和想进入相关领域的从业者来说，是一本值得推荐的好书。

从 2015 年开始，我有幸与作者在同一个团队工作，一起设计并研发阿里云的机器学习平台——PAI。作者对机器学习的理解以及产品上的设计思想都在本书中完美地呈现，值得准备进入机器学习领域的爱好者和从业者好好品读。

感谢作者让我在新书出版之前先睹为快。

——刘吉哲

阿里云高级专家

致谢

感谢我的父母这些年对我的鼓励，感谢我的女朋友，家人的支持永远是我的源动力，让你们生活得幸福是我奋斗的目标。感谢我的大学同学，特别是本科宿舍的室友，你们是我心中的一股清流。最后我要特别感谢我的同事，感谢楚巍、不老、吉哲、云郎、贾总、品道等人以及 UED 小团队，感谢你们对我工作上的支持和帮助。在阿里云大家庭中，我工作得很快乐，个人成长也非常迅速。同时，我也非常感谢出版社的编辑胡俊英在本书写作期间为我提供建议和帮助。

最后对自己这段时间的写作过程做一个总结，最大的感触是，在这样快速紧张的生活和工作节奏下，连续 8 个月坚持做一件事情是非常需要毅力的。每天下班之后坚持学习和写作 2 小时，常常熬到凌晨才关灯睡觉，但是这份坚持换来了将近 500 小时的时间用来“充电”。在这段时间中，写作已经成为我的一种生活方式，在飞机上、在高铁上、在出租车上、在厕所中……很多地方都留下了思考和回忆。无论最终能做到什么程度，都希望自己可以继续把这样的激情保持下去。最后感谢所有在工作和学习中给过我帮助的人，也感谢所有拒绝我、批评过我的人，因为有你们才有了这本书。

前言

人工智能是近年来非常火的话题，人们似乎看到了在某些领域内机器智能取代人力的可能性。之所以人们可以得到这样的判断，主要是基于以下几方面原因：随着互联网的发展，人类社会积累了大量的数据可供分析；机器学习的算法不断迭代，特别是近年来随着深度学习的发展，人们从理论层面取得了实质性突破；随着分布式计算的成熟，云计算让计算资源不再成为瓶颈。我们可以把人工智能看作一个数据挖掘体系，在这个体系当中，机器学习的作用主要是学习历史数据中的经验，把这些经验构建成数学模型。人类利用机器学习算法生成的模型，就可以解决日常的一些问题，如商品推荐和对股票涨跌的预测等。

以上谈到了机器学习的主要作用，我们再来了解机器学习在业务中的应用，其实机器学习算法正在逐步向“平民化”演变。早些时候，只有一些规模比较大的公司会投入资源在智能算法的研究上，因为这些算法需要大量的数据积累以及计算资源，而且整个业务框架跟算法的结合也需要耗费很大人力，所以只有少数数据业务量达到一定规模的公司会在这方面投入。但是随着各种开源算法框架的发展以及计算资源的价格走低，机器学习不再是“奢侈品”，很多规模不大的公司也开始尝试用机器学习算法生成的模型来指导自身业务，用数据来解决业务问题是代价最小的方式，而且效果会随着数据量的积累变得越来越明显。机器学习算法正在帮助越来越多的企业实现转型，从传统的商业智能（Business Intelligence，BI）驱动到人工智能（Artificial Intelligence，AI）驱动。通过平日里与客户打交道，我们可以了解到，现在不只是互联网公司，更多传统行业，如教育、地产和医疗等，也在尝试把自己的业务数据上传到云，通过机器学习算法来提升自己的业务竞争力。

综上所述，业务与机器学习算法的结合很有可能是下一阶段行业变革的驱动力，如果固守原来的传统技术，不尝试提升业务的数据驱动力，企业很有可能在这一波新的浪潮中被淘汰。本书尝试将算法与实际的业务实战相结合，将对机器学习的全链路逐一进行介绍。在描述算法理论的时候，本书尽可能用更直白易懂的语句和图示来替代公式。另外，为了帮助读者更有成效地理解机器学习算法的使用逻辑，书中不单介绍了算法，还对整个数据挖掘的全流程，包括数据预处理、特征工程、训练以及预测、评估进行了介绍。而且本书还通过真实案例的数据，在各种不同业务场景下对整个数据挖掘流程进行了详细介绍。此外，书中还简单地介绍了深度学习和知识图谱这两个未来可能被更多关注的领域。总之，本书不是一本理论教程，而是一本推动算法与业务实践相结合的指南。

写作本书的目的

我从研究生阶段开始接触机器学习算法，在硕士研究生期间主要从事算法的理论研究和代码实现，当时参与了一些开源算法库的开发和算法大赛，那时对机器学习的理解更多的是停留在数学公式推导层面。那时候理解的机器学习就是一门统计科学，需要把公式研究透彻。直到入职阿里云，从事了机器学习平台相关的工作，我对机器学习的看法发生了很大改变。根据平日里与客户的沟通，我认识到，对绝大部分中小企业用户而言，机器学习算法只是帮助大家提升业务成效的工具，很多用户对机器学习的理解还处于比较初级的阶段，与这种现状相矛盾的是目前市面上部分机器学习相关的图书都更偏向于理论研究，而比较缺乏实际应用的场景。

写这本书的目的就是希望可以提供这样一本素材，能够让渴望了解机器学习的人快速了解整个数据挖掘体系的轮廓，可以用最小的成本帮助用户把算法迁移到机器学习云服务上去。至于算法的精密度和深度的探索，那是数学家需要考虑的事情，对绝大部分的机器学习算法用户而言，这样一本能帮助大家快速理解算法并能够将其在业务上实践的教程可能会更加有效。

对我而言，本书也是我对自己学习成果的总结。从 2013 年起，我陆陆续续在 CSDN、GitHub 和云栖社区上分享过一些自己在 IT 领域的学习笔记和代码，收到了很多朋友的反馈，也有一些出版社的朋友找到我希望可以把这些内容整理成书，但是一直没有特别笃定的想法——什么样的书是有价值的。通过近一年来的机器学习平台产品建设以及与客户的不断接触，我心中的想法逐渐清晰，很多机器学习爱好者最关心的是如何使用算法而不是这些算法背后的推理，于是本书就应运而生了。虽然我才疏学浅，书中内容未免有描述不足之处，但是我真心希望这本书可以在读者探索机器学习的道路上为其提供助力。

读者对象

本书的读者对象如下：

- 有一定数学基础，希望了解机器学习算法的人；

- 有编程基础，希望自己搭建机器学习服务解决业务场景的工程师；
- 数据仓库工程师；
- 与数据挖掘相关的高校学生；
- 寻求数据驱动业务的企业决策者。

如何阅读本书

本书的结构是按照读者对机器学习的认知过程和数据挖掘的算法流程来组织的，一共分为 5 个部分，共 9 章内容。

第 1 部分是机器学习的背景知识介绍，包括第 1 章。这一部分主要介绍机器学习的发展历史以及现状，另外，也介绍了机器学习的一些基本概念，为接下来的内容做准备。

第 2 部分介绍机器学习的算法流程，包括第 2～6 章，分别介绍了场景解析、数据预处理、特征工程、机器学习常规算法和深度学习算法。在第 5 章的算法部分，对常见的分类算法、聚类算法、回归算法、文本分析算法、推荐算法和关系图算法都进行了介绍，从这一章可以了解到不同业务场景下不同算法的区别和用法。第 6 章对深度学习相关内容进行了讲解，包括常用的 3 种模型 DNN、CNN 和 RNN 的介绍。

第 3 部分介绍机器学习的相关工具，包括第 7 章的内容。这里的工具是一个广泛的概念，包括了 SPSS 和 R 语言这样的单机统计分析环境，也包括了分布式的算法框架 Spark MLib 和 TensorFlow，还有企业级的云算法服务 AWS ML 和阿里云 PAI。通过阅读这一章，读者可以根据自身的业务特点，选择适合自己的算法工具。

第 4 部分介绍机器学习算法的实践案例，包括第 8 章，帮助读者理解整个数据挖掘流程。这一章针对不同行业和不同场景搭建了实验，分别介绍了如何通过机器学习算法应对心脏病预测、商品推荐、金融风控、新闻分类、贷款预测、雾霾天气预报和图片识别等业务场景，因此也是本书的核心章节。

第 5 部分主要针对知识图谱这个热点话题进行介绍，包括第 9 章，知识图谱的介绍主要是从图谱的概念以及实现的角度来说明。

尽管读者可以根据自己的侧重点来选择阅读顺序，但我强烈建议读者按照顺序来阅读，这样对理解书中的概念并能够循序渐进地掌握相关知识更有帮助。

勘误和服务

虽然花了很多时间去反复检查和核实书中的文字、图片和代码，但是因为认知能力有限，书中难免会有一些纰漏，如果大家发现书中的不足之处，恳请反馈给我，我一定会努力修正问题，我的个人邮箱是 garvin.libo@gmail.com。如果大家在阅读本书的时候遇到什么问题，也欢迎通过各种方式与我取得联系，个人网站为 www.garvinli.com，另外本人的博客地址是 http://blog.csdn.net/buptgshengod。读者也可以到异步社区的页面内提交勘误，网址详见 http://www.epubit.com.cn/book/detail/4757。因为工作繁忙，可能来不及一一回复，但是我会尽力与读者保持沟通，谢谢大家的支持。

目录

第 5 部分 知识图谱

第 1 部分

背景知识

第 1 章
机器学习概述

在本章中，笔者会以对于人工智能发展历史的回顾作为开篇，进而介绍一些人工智能的发展现状，还会引出对于机器学习的基本概念的一些讲解。这一章作为全书的开篇，希望给各位读者一个宏观的概念——什么是机器学习？它会给我们的生活带来哪些改变？

1.1 背景

正如爱因斯坦所说："从希腊哲学到现代物理学的整个科学史中，不断有人试图把表面上极为复杂的自然现象归结为几个简单的基本概念和关系，这就是整个自然哲学的基本原理。"人类进化的发展史，从某种意义上来讲就是不断归纳经验进而演绎的过程。从刀耕火种的新石器时代到近代的工业革命以及现代科技的发展，人类已经积累了大量的经验。这些经验既是"种瓜得瓜，种豆得豆"这样的常识，也是例如相对论这样的定理公式。人类文明正沿着时间这条坐标轴不断前进，如何利用过往的经验来推动人类社会的再一次飞跃，人工智能或许是我们需要的答案。

人工智能的起源应该可以追溯到 17 世纪甚至更早，当时人们对于人工智能的定义是基于推理的。人们畅想着如果两个哲学家或者历史学家的观点出现矛盾，两个人不必再进行无休止的争吵，世界上的所有理论会抽象成类似于数学符号的语言，人们只需要拿出笔来计算就可以解决矛盾。这种抽象逻辑给了后人引导，如今，机器学习在行业上的应用也是将业务逻辑抽象成数字来进行计算，从而解决业务问题。但是在远古时代，这些逻辑还只是科学家脑中的想法。实际上，直到有机器的出现，人工智能才真正作为一门学科而受到广泛关注。

谈到近代人工智能的起源就不得不提到一个名字——图灵（见图 1-1）。

随着第二次世界大战的爆发，越来越多的机械开始替代手工，人们开始幻想什么时候

机器能代替人类来进行思考。在20世纪40年代，关于人工智能的讨论开始兴起。但是，机器做到什么程度才算人工智能，这需要一个标准来判定。图灵用了最直白的话语描述了人工智能，这就是图灵测试（见图1-2）。

图1-1 阿兰•图灵

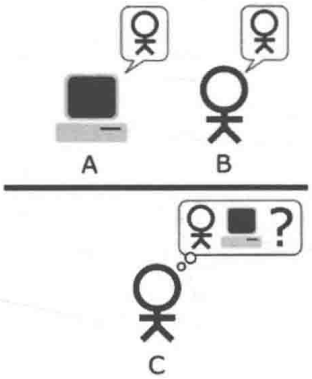


图1-2 图灵测试

1950年，计算机科学和密码学的先驱阿兰•麦席森•图灵发表了一篇名为《计算机器与智能》的论文，文中定义了人工智能测试的方法，让被测试人和一个声称自己有人类智力的机器在一起做一个实验。测试时，测试人与被测试人是分开的，测试人只有通过一些装置（如键盘）向被测试人问一些问题，随便是什么问题都可以。问过一些问题后，如果测试人能够正确地分出谁是人、谁是机器，那机器就没有通过图灵测试，如果测试人没有分出谁是机器、谁是人，那这个机器就是有人类智能的。

人工智能的另一个重要标志是人工智能这一学科的诞生，故事发生在1956年达特茅斯会议。会议上提出了这样的理论："学习或者智能的任何其他特性都能被精确地描述，使得机器可以对其进行模拟。"这个论调很像机器学习算法在今日的应用，我们需要提取可以表示业务的特征，然后通过算法来训练模型，用这些模型对于未知结果的预测集进行预测。这次会议对于人工智能在更广阔的领域发展起到了推动作用。在之后的20年里，人类在人工智能，特别是相关的一些统计学算法的研究上取得了突破进展，比较有代表性的如神经网络算法，就是在这个时期诞生的。有了这些智能算法作支撑，更多的真实场景才可以在数学层面进行模拟，人类慢慢学会通过数据和算法的结合来进行预测，从而实现某种程度上的智能化应用。

人工智能在发展过程中也遇到过非常多的挑战。20世纪70年代，随着理论算法的逐步成熟，人工智能的发展遇到了计算资源上的瓶颈。随着计算复杂度的指数性增长，20世纪70年代的大型机器无法负担这一切。同时，当时的互联网还处于发展初期，在数据

积累方面也才刚刚起步。科学家往往没有足够的数据去训练模型，以图像印刷文字识别（Optical Character Recognition，OCR）为例。如果想针对某一场景训练一套精度较高的OCR模型，需要千万级的数据样本，这样的数据无论从数据获取、存储和计算成本来看，在当时都是不可能实现的。所以人工智能在之后很长的一段时间内都受限于计算能力以及数据量的不足。

虽然经历了近 20 年的消沉时期，但是数据科学家对于人工智能的探索从未停止过。在21世纪，随着互联网的井喷式发展，越来越多的图像和文本数据被分享到网页上，停留在互联网巨头的服务器中，随之而来的是用户在网上的浏览记录和购物记录的收集。互联网已经变成了一个大数据仓库，许多网络大咖们纷纷将注意力投向数据挖掘领域，数据库成为了一座座金矿，数据科学家们开始用一行行公式和代码挖掘数据背后的价值，越来越多的公司做起了数据买卖。这些代码和公式就是本书的主角——机器学习算法。马云先生在很多年前的公开演讲上就已经明确表示过“阿里巴巴是一家数据公司”。数据的积累就像是一块块肥沃的土地，需要机器学习算法来在上面耕种，云计算就是挥舞在土地上的“锄头”。PB级数据的积累使得人们不得不将单机计算迁移到多机，并行计算理论开始得到了广泛的应用，这就催生了云计算的概念。云计算，就是分布式计算，简单来讲就是将一个很复杂的任务进行拆解，由成百上千的机器各自执行任务的一个小模块，然后将结果汇总。

以 Hadoop 为代表的开源分布式计算架构为更多的企业提供了分布式计算的技术支持。随着 Caffe 和 Tensorflow 等高效率的深度学习架构被开源，许多小型企业也具备了自主研发改进算法模型的能力。人工智能的应用开始普及，并且逐渐融入我们的生活当中。人们开始习惯了在 Google 上输入一个词条马上就能返回上千万条信息，通过刷脸或者指纹识别来进行支付，在淘宝购物时获得智能商品推荐。图像识别、文本识别和语音识别的发展给我们的生活带来了颠覆式的影响。2016 年，Google 关于人工智能的一场秀将人工智能产业带到了一个新高度。机器智能战胜人类围棋选手一直以来被认为是不可能实现的任务，但是 AlphaGo 成功地实现了这一点。AlphaGo 的成功不仅仅验证了深度学习和蒙特卡洛搜索算法的实践性，更加再一次印证了这样的事实，即人类不再是产生智能的唯一载体。任何机器，只要能够进行信息的接收、存储和分析，都是可以产生智能的。而这里面的关键因素是信息的量级以及算法的深度。

人工智能的发展史，就是对于过往经验的收集和分析方法不断演绎的历史。在机器出现之前，人类只能通过别人的分享和自己的实践在很小的信息量级上来对事物进行判断，这种对于外界事物的认知受限于人的脑力和知识量。不同于人类的脑力，抽象意义上的机