The Quest for Machine Learning

百面机器学习

算法工程师带你去面试

100+ Interview Questions for Algorithm Engineer

诸葛越　主编

葫芦娃　著

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

The Quest for Machine Learning

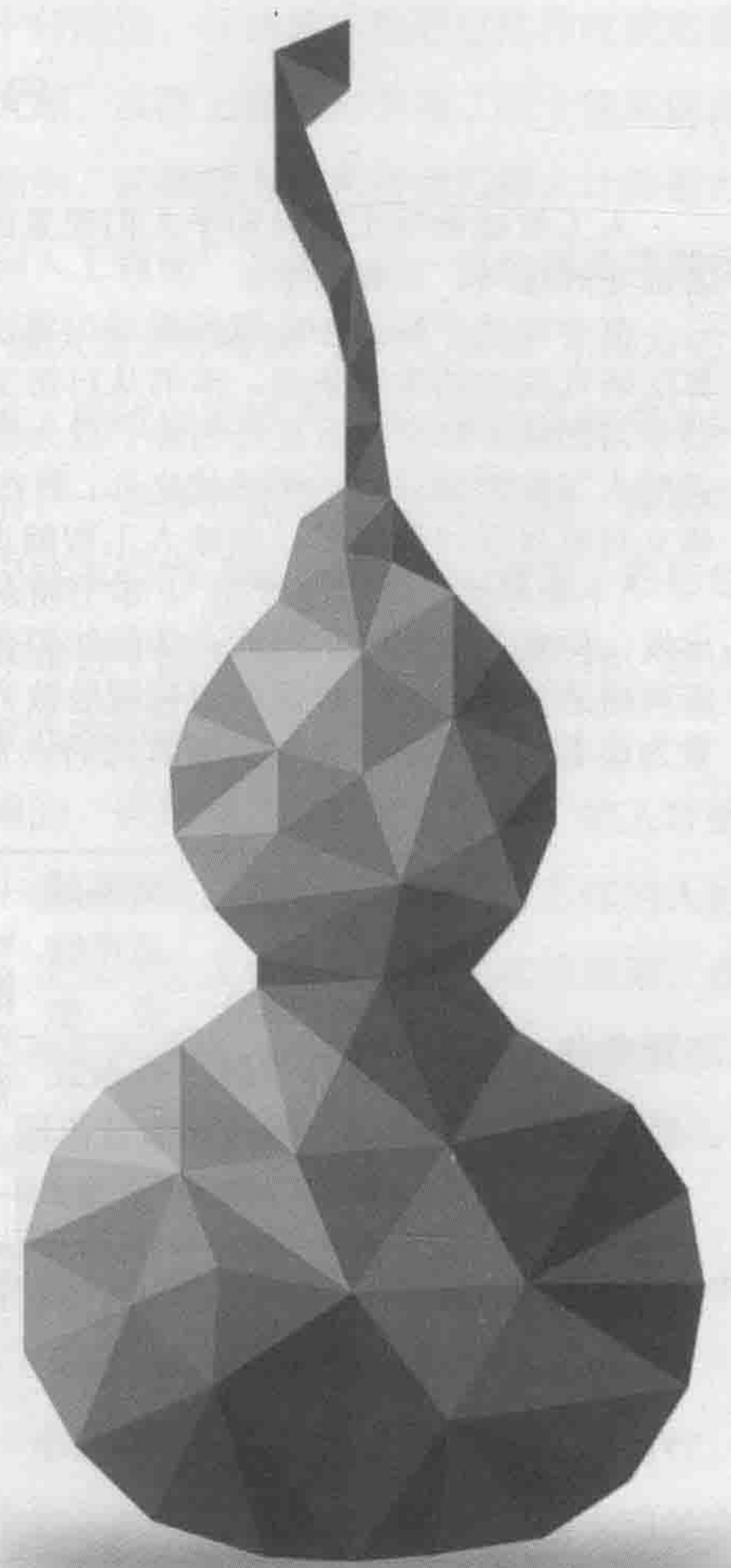
百面机器学习

算法工程师带你去面试

100+ Interview Questions for Algorithm Engineer

诸葛越 主编

葫芦娃 著



人民邮电出版社

北京

图书在版编目（CIP）数据

百面机器学习 ：算法工程师带你去面试 / 诸葛越主编 ；葫芦娃著. -- 北京 ：人民邮电出版社，2018.8
ISBN 978-7-115-48736-0

Ⅰ. ①百… Ⅱ. ①诸… ②葫… Ⅲ. ①机器学习－算法 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2018)第154485号

内 容 提 要

人工智能领域正在以超乎人们想象的速度发展，本书赶在人工智能彻底“占领”世界之前完成编写，实属万幸。

书中收录了超过 100 道机器学习算法工程师的面试题目和解答，其中大部分源于 Hulu 算法研究岗位的真实场景。本书从日常工作、生活中各种有趣的现象出发，不仅囊括了机器学习的基本知识，而且还包含了成为优秀算法工程师的相关技能，更重要的是凝聚了笔者对人工智能领域的一颗热忱之心，旨在培养读者发现问题、解决问题、扩展问题的能力，建立对机器学习的热爱，共绘人工智能世界的宏伟蓝图。

“不积跬步，无以至千里”，本书将从特征工程、模型评估、降维等经典机器学习领域出发，构建一个算法工程师必备的知识体系；见神经网络、强化学习、生成对抗网络等最新科研进展之微，知机器学习领域胜败兴衰之著；“博观而约取，厚积而薄发”，在最后一章为读者展示生活中各种引领时代的人工智能应用。

◆ 主　　编　诸葛越
著　　葫芦娃
责任编辑　俞　彬
责任印制　马振武

◆ 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
北京缤索印刷有限公司印刷

◆ 开本：720×960　1/16
印张：26.75
字数：480 千字　　2018 年 8 月第 1 版
印数：1－4 000 册　　2018 年 8 月北京第 1 次印刷

定价：89.00 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

很荣幸有机会推荐清华大学计算机系 1991 级校友诸葛越和她的团队写的新书《百面机器学习：算法工程师带你去面试》。

毋庸置疑，人工智能现在正在蓬勃兴起，就像生机勃发的春天，就其热度而言，说它处在夏天也十分贴切，但我更愿意把它比作收获的金秋。目前席卷全球的人工智能大潮，实际上是机器学习二三十年来理论和算法研究厚积薄发的结果（当然，还要加上与大数据和强大计算能力的风云际会），其本质属于“弱人工智能”范畴。这一波大潮恣肆到极致后一旦消退，我们期望的下一波大潮必然将是“强人工智能”所催发的，但由于其理论探索的高度困难性，尚难以设想下一波大潮什么时候才会再次奔涌而至。所以当下的我们，一定要把握住这难得的机遇，抓紧收获“弱人工智能”慷慨馈赠的足够丰硕的“果实”。可以想象，形形色色的人工智能应用将在近一两年走进千家万户，会像互联网一样，给人们的生活，给社会和经济带来深远的影响。

然而，收获并不是唾手可得的，只有有能耐摘取“果实”的人才能尽享丰收的喜悦——这就是在一线从事人工智能和机器学习工作的人们（通常也被称作算法工程师）。正是这些人，针对不同的实际应用，在不断地尝试新的方法，不断地实现新的算法。他们了解需求、收集数据、设计算法、反复实验并持续优化。他们是人工智能新一代技术的“弄潮儿”和推动者。

那么，你是否想成为他们中的一员呢？你又如何能快速成为他们中的一员呢？

也许这本书可以帮你前进一步。在人工智能技术如火如荼的时代，大批优秀的研究员和程序员正辛勤致力于解决人工智能和机器学习的实际应用问题，市场上急需这方面的技术实操书。而本书刚好填补了这方面的空白。它的内容由简至繁依次展开，涵盖了机器学习各个实用领域，并采取了举例和问答的形式，生动活泼，使每个读者既能了解人工智能

从业者所需要的技能，又能学会掌握这些技能。

我从事人工智能研究已有三十余年了，研究兴趣比较广泛，涵盖了自然语言理解、机器学习、社会人文计算等，与这个领域相知相行。我认识诸葛越多年，她是我们系有名的“学霸”，曾经获得美国计算机学会数据库专业委员会十年最佳论文奖（*ACM SIGMOD Test of Time Award*）。回国后她也常常来系里参加活动。我了解到她的团队中的每一位成员都有非常优秀的背景。本书是工业界每天从事机器学习工作的数据科学家一起撰写的著作，它一定不会让你失望。

希望更多的朋友通过读这本书，成为更好的算法工程师、数据科学家和人工智能的实践者。我带领的研究小组最近研制了一个“九歌”古诗自动写作系统，2017 年登录央视大型科学挑战类节目《机智过人》，它在节目中的表现初步达到了与人类诗人难分伯仲的程度，而其基本框架正是得益于本书讲述了的长短期记忆网络和 Seq2Seq 模型。这里我姑且借用“九歌”写作的一首五绝集句诗，祝本书的所有读者都能在这个激动人心的技术新时代更上一层楼：

更上一层楼《登鹳雀楼》唐·王之涣

蝉声满树头《闲二首》唐·元稹

春光无限好《感皇恩·春水满池塘》宋·叶景山

月涌大江流《旅夜书怀》唐·杜甫

孙茂松

清华大学计算机系教授，博导，前系主任，前党委书记

2018 年 6 月 2 日于清华园

前言

Preface

人工智能的三次浪潮

2018 年年初，招聘季正如火如荼地进行，而“数据科学家”和“算法工程师”绝对算得上热门职业。

“人工智能”“机器学习”“深度学习”“建模”“卷积神经网络”等关键词，不仅仅是人们茶余饭后的谈资，而且更会像“数据结构”“排序”和“链表”一样，成为软件工程师的必备技能。

人工智能技术正在对社会结构、职场、教育等带来革命性的变化。未来几年是人工智能技术全面普及化的时期，也是该技术的相关人才最为稀缺的时期。所以，我们希望能够通过这本书，帮助对人工智能和机器学习感兴趣的朋友更加深入地了解这个领域的基本技能，帮助已经有计算机技术基础的同行们，成为驾驭人工智能和机器学习的高手。

写在书的前面，我先简单介绍一下我了解的人工智能和机器学习的背景和历史，解释为什么现在是学习机器学习算法的大好时机。

■ 我与人工智能

我的本科专业是人工智能。当年我上大学时，清华大学的计算机系每个年级有 6 个班，入学的时候就把每个班的专业分好。我们三班的专业是人工智能。所以在本科的时候，我就接触到许多当时人工智能领域的前沿技术。我的人工智能入门课的导师是可亲可敬的林尧瑞教授，也是《人工智能导论》的作者。这门课被我们戏谑为“猴子摘香蕉”，因为最开始的问题就是一只智能的猴子，如何自己组合积木去拿到天花板上挂着的香蕉。

当时清华大学的本科是 5 年制，正要开始改革，有少部分学生可以在四年级的时候开始接触研究生的一些活动，6 年可以拿到硕士学位。我有幸被选为这几个学生之一，在本科四年级的时候，我进入了清华大学的人工智能实验室，师从张钹老师，做一些简单的研究。从张老师和

高年级的同学们那里，我学到人工智能领域不少当时国际先进的知识。

刚刚进入斯坦福的时候，去听一个小型的午餐讲座 (Brown Bag)，也就是一二十个人吧。那位同学讲到一半，教室门突然被打开，大胡子的约翰·麦卡锡 (John McCarthy) 教授走了进来，大声地问："听说这里有不要钱的午饭？" 然后他走到房间的前面，抓了两个三明治，大摇大摆地走出去了。主持讲座的老师愣了一下，说："欢迎大家来到斯坦福——世界上最著名的科学家会走进你们的教室来抢你们食物的地方！"

或许你不知道，"人工智能" (Artificial Intelligence) 这个词，就来自约翰·麦卡锡。

因为本科是人工智能专业，所以我对人工智能一直比较感兴趣，在斯坦福又去学了一次人工智能课 CS140。当时教这个课的是尼尔斯·尼尔森（Nils Nilsson）教授。他是另外一位人工智能的学科创始人和世界级专家，写作了被广泛引用的经典之作——《对人工智能的探索》(*The Quest for Artificial Intelligence*）。尼尔森教授的课非常有趣，我还跟他做了一个小的项目，规划一个扫地机器人的路径。至今，我还保留了这门课的笔记。

说实话，我年轻的时候每天做作业、做课题，没有意识到，能和这些顶级科学家同堂是多么幸运的事，也未必知道自己正在见证某个技术领域的世界前沿。最顶尖的技术，开始都是只有小众才能理解和欣赏的。

然而，我的博士论文并没有专攻人工智能，反而做的是大数据方向，做了最早的数据仓库和数据挖掘工作。现在看来，我这几次和人工智能以及人工智能大咖的偶遇，刚好和人工智能的三次浪潮有关。第一次人工智能的浪潮就是约翰·麦卡锡那一代人。他们从 20 世纪 50 年代开始，打下了计算机学科和人工智能的理论基础。第二次是我在清华大学期间，研究者们看到了一些人工智能应用的可能性，比如机械手、机器人、专家系统。最近，基于大数据、机器学习的人工智能再次兴起，可以称为人工智能的第三次浪潮 。

人工智能的三次浪潮

我来简单定义和解释一下本书用到的概念。

人工智能泛指让机器具有人的智力的技术。这项技术的目的是使机器像人一样感知、思考、做事、解决问题 。人工智能是一个宽泛的技术领域，包括自然语言理解、计算机视觉、机器人、逻辑和规划等，它可以被看作计算机专业的子领域，除了和计算机相关，它还和心理学、认知科学、社会学等有不少交叉。

机器学习指计算机通过观察环境，与环境交互，在吸取信息中学习、自我更新和进步。大家都了解计算机程序是怎么回事，一个程序是计算机可以执行的一系列的指令，比如打印一张图。那么机器学习跟我们熟知的程序的本质区别是什么呢？你可以想象，某个程序是机器写的，而不是一个程序员写的。那么机器怎么知道如何写这个程序呢？这个机器就是从大量的数据当中学到的。

简单地说，大多数机器学习算法可以分成**训练 (training)** 和**测试 (testing)** 两个步骤，这两个步骤可以重叠进行。训练，一般需要训练数据，就是告诉机器前人的经验，比如什么是猫、什么是狗、看到什么该停车。训练学习的结果，可以认为是机器写的程序或者存储的数据，叫**模型 (model)**。总体上来说，训练包括**有监督 (supervised learning)** 和**无监督 (unsupervised learning)** 两类。有监督好比有老师告诉你正确答案；无监督仅靠观察自学，机器自己在数据里找模式和特征。深度学习（deep learning）是机器学习的一种方法，它基于神经元网络，适用于音频、视频、语言理解等多个方面。

我们先来短暂地回顾一下人工智能的三次浪潮。它们有什么特点？又有什么不同？它们又是怎样互相联系，如何在前一次的基础之上建立的？

第一次人工智能浪潮大约在 20 世纪 50 年代。1956 年，在达特茅斯的人工智能研讨会上，约翰·麦卡锡正式提出“人工智能”这个概念，被公认是现代人工智能学科的起始。麦卡锡与麻省理工学院的马文·明斯基（Marvin Minsky）被誉为“人工智能之父”。

在计算机被发明的早期，许多计算机科学家们就认真地思考和讨论这个人类发明出来的机器，和人类有什么根本区别。图灵机和图灵测试，就是这个思考的一个最典型结果。最初的那批思考人工智能的专

家，从思想和理论上走得非常前沿，内行的专家很早就看到了计算机的潜力。我们现在所问的这些问题，他们其实都问过了。比如，什么叫“推理”(reasoning)，机器如何推理；什么叫“懂得”(understanding)，机器如何懂得；什么叫知识 (knowledge)，机器如何获取和表达知识；什么时候，我们无法分辨出机器和人。这个阶段产生了许多基础理论，不仅是人工智能的基础理论，也是计算机专业的基石。

从技术上来说，第一次人工智能的大发展，主要是基于逻辑的。1958 年麦卡锡提出了逻辑语言 LISP。从 20 世纪 50 年代到 20 世纪 80 年代，研究者们证明了计算机可以玩游戏，可以进行一定程度上的自然语言理解。 在实验室里，机器人可以进行逻辑判断、搭积木；机器老鼠可以针对不同的路径和障碍做出决定；小车可以在有限的环境下自己驾驶。研究者们发明了神经网络，可以做简单的语言理解和物体识别。

然而，在人工智能的前二三十年里，它虽然是一个硕果累累的科研领域，人们实际生活中的用处却几乎没有。20 世纪 80 年代初，人工智能因为缺乏应用而进入“冬季”。到 80 年代末和 90 年代初，在我刚入大学的那段时间里，人工智能科学家们决定另辟蹊径，**从解决大的普适智能问题，转向解决某些领域的单一问题。**“专家系统”这个概念被提了出来，它让这些研究成果找到了第一个可能的商业出路。

计算机技术经过了 30 年左右的发展，数据存储和应用有了一定的基础。研究者们看到人工智能和数据结合的可能性，而结合得最好的应用就是“专家系统”。如果我们能把某一个行业的数据，比如说关于心脏病的所有数据，都告诉一个机器，再给它一些逻辑，那这个机器岂不是就成了“心脏病专家”，如果我们要看病，是否就可以问它?

看病、预报天气等各行各业的专家系统，听起来非常有希望、有意义，也确实有实际的应用场景，所以当时学术界对人工智能又掀起了一阵热潮。然而，比较有意思的是，当我们想要用这些专家系统来做一些聪明的诊断的时候，我们发现遇到的问题并不是如何诊断，而是大部分的数据在当时还不是数字化的。病人的诊断历史还停留在看不懂的医生手写处方上。有些信息就算是已经开始数字化，也都是在一些表格里面，或者是在一些不互相连接的机器里面，拿不到，用不了。

于是，我们这一批想去做自动诊断的人，反而去做了一些基础的工作。这个基础的工作用一句话说，就是把世界上所有的信息数字化。

在一批人致力于把世界上每一本书、每一张图、每一个处方都变成电子版的时候，互联网的广泛应用，又把这些信息相互联接了起来，成了真正的大数据。同时，摩尔定律（Moore's law）预测的计算性能增加一直在起作用。随着计算能力的指数增长，那些只能在实验室里或有限场景下实现的应用，离现实生活越来越近了。1997 年，“深蓝”打败当时的世界象棋冠军 Garry Kasparov，和 2017 年 AlphaGo 围棋打败李世石一样，被公认是一个里程碑 。其实，随着计算能力的提高，在这些单一的、有确定目标的事情上机器打败人，都只是个时间问题。

第三次的人工智能浪潮就是基于另外两个技术领域的大发展，一个是巨大的计算能力，一个是海量的数据。巨大的计算能力来自于硬件、分布式系统、云计算技术的发展。最近，专门为神经网络制作的硬件系统（neural-network-based computing）又一次推动了人工智能软硬件结合的大进步。海量的数据来源于前几十年的数据积累和互联网技术的发展。比如，2001 年上市的 GPS 系统，带来前所未有的大量出行数据；智能手机带来了前所未有的人们生活习性的数据，等等。计算能力和数据的结合，促进、催化了机器学习算法的飞跃成长。

这次的人工智能浪潮起始于近 10 年。技术的飞跃发展，带来了应用前所未有的可能性。**最近这次人工智能浪潮和前两次最基本的不同是它的普遍应用和对普通人生活的影响。也就是说，人工智能离开了学术实验室，真正走进大众的视野。**

■ 人工智能全面逼近人类能力?

为什么这次人工智能浪潮如此凶猛？人工智能真的全面逼近了人类的能力吗？人工智能技术现在发展到什么阶段？我们先来看 3 个简单的事实。

首先，历史上第一次，计算机在很多复杂任务的执行上超过人类或者即将超过人类，比如图像识别、视频理解、机器翻译、汽车驾驶、下围棋，等等。这些都是人们容易理解的，一直由人类完成的任务。所以，人工智能取代人类的话题开始出现在各种头条。

其实，在单一技术方面，许多计算相关的技术早已超过人类的能力，而且被广泛应用，比如导航、搜索、搜图、股票交易。不少人已经习惯于用语音给简单指令操作。但是，这些相对单纯的技术主要是“完成一个任务”，计算机没有过多地涉猎人的感知、思考、复杂判断，甚至于情感。

然而，近几年来机器完成的任务，从复杂性和形式越来越逼近人类。比如，基于机器学习的自动驾驶技术已经趋于成熟，这项技术不仅会对人们的出行方式有革命性的影响，而且会影响到城市建设、个人消费、生活方式。人们也许再也不需要拥有汽车，再也不需要会开车。大家对这类新技术的快速到来既兴奋又恐惧，一方面享受技术带来的便利，另一方面又对太快的变化有些手足无措。

另外，计算机的自学习能力不断增强。现代机器学习算法，尤其深度学习类机器学习算法的发展，使机器的行为不再是相对可预测的“程序”或者“逻辑”，而更像“黑盒思考”，有了近乎人类的难以解释的思考能力。

然而，仔细看来，虽然在不少特殊领域中，人工智能有了突飞猛进的发展，但是距离人工智能的鼻祖们在第一次浪潮时研究的通用智能（general purpose intelligence）其实还相差非常远。这是第二个事实。机器还是被放在特定情况下完成特定任务，只不过任务更复杂了。机器还是缺少一些最基本的人的智能，比如常识。人工智能仍然无法理解哪怕是简单的情感，比如害怕。对两三岁的孩子来说非常简单的帮忙、合作，机器都是做不到的。好比有人开玩笑说：“它们还是不会炒鸡蛋。”

第三个事实，是这次人工智能和机器学习的应用场景非常宽广。近几年人工智能和机器学习应用的大发展，这个曾经是学术研究领域的概念一时间进入大众视野，成为和未来相关的必谈话题。计算机视觉、深度学习、机器人技术、自然语言理解，都被提到应用层。算法类的应用走出学术界，深入社会的各个角落，渗入人们生活的方方面面。大家熟知的有人脸识别、自动驾驶、医疗诊断、机器助手、智慧城市、新媒体、游戏、教育等，还有并不常被谈论的比如农业生产的自动化、老人和儿童的护理、危险情景的操作、交通调度，等等。我们很难想象社会的哪

一个方面，不会被这次浪潮所波及。

向前看十年，人工智能和机器学习的大发展，在于这些技术的普及和应用。大批的新应用将会被开发，人工智能基础设施会迅速完善，原有的传统软件和应用需要被迁移使用新的算法。所以，**现在是成为一个人工智能和机器学习专家的良机。**

■ 这本书是如何写成的

无论海内海外，媒体行业一直都走在人工智能应用的最前沿，因为媒体往往接触上千万甚至上亿的用户；有千变万化的用户每天离不开的内容，比如新闻、体育、电影；有丰富多彩的内容与用户的结合场景；还有丰厚的有创意的商机。

Hulu 是一家国际领先的视频媒体公司，提供优质电影、电视剧点播和直播节目。Hulu 技术架构最为先进的一点是人工智能和机器学习算法的广泛应用，用在个性化内容推荐、搜索、视频内容理解、视频传输和播放、广告预测和定向、安全检测、决策支持，甚至视频编辑和客服系统。机器学习算法的背后是专门打造的大规模数据处理系统。“算法无处不在”是 Hulu 当今和未来技术架构的定位。可以说，Hulu 是未来的互联网技术公司，全面“算法化”的一家带头公司。

为了支持各类的人工智能算法应用，Hulu 在北京的创新实验室集合了大批人工智能和机器学习的顶尖人才。Hulu 的数据科学家、算法工程师和软件工程师都工作在同一个团队，每天解决用户的实际问题，积累了大量实用的经验。Hulu 北京的学习气氛也相当浓厚。除了定期的机器学习专题研讨和大数据及机器学习公开课，Hulu 也在内部开设了深度学习课程。

2017 年年底，人民邮电出版社的俞彬编辑问我能否写一本关于人工智能和机器学习算法实操的书。目前市场上有关人工智能的书可以分为两类，一类是非常系统的教科书，还有一类是关于人工智能和人类未来的社科类图书。我们能否写一本实操类的书，介绍一个真正的计算机从业人员需要掌握的技能呢？

抱着试一试的心理，我让公司里的同事自愿报名参加这个集体项目。一共有 15 位资深研究员和算法工程师参与了这本书的内容创作，这是个

成功的合作案例。我们先学习了一下现有的相关书籍，然后头脑风暴了一番，觉得我们可以做一个问答集，以比较有趣的问答形式，集中当前算法工程师和研究员感兴趣的话题，用问答引出这个行业的基本概念。

在互联网行业，敏捷开发都是以最快的速度，做一个“最小化产品”，让用户的反馈来带领产品的方向。我们写这本书也是如此。为了让大家能够落笔写出没有错误、通俗易懂的问答，为了收集读者的反馈，也为了不把写一本大部头书列为第一天的目标，我们先在 Hulu 的微信公众号上，以每周发两个问答的形式，从 2017 年 11 月到 2018 年 3 月期间，一共发出了 30 篇“机器学习问答”系列文章。这些文章受到了业界好评，也收到各种问题和反馈，成了我们这本书的核心内容。

关于书的章节组织，我们也进行过仔细的讨论。人工智能和机器学习算法范围很大，我们的理念是要涵盖该领域最基本的内容，介绍基本概念，同时，跟上算法发展的最新步伐。所以本书介绍了传统机器学习算法，比如逻辑回归、决策树等，同时花了比较大的篇幅介绍近几年流行的最新算法，包括各种神经网络（深度学习）、强化学习、集成学习等，还会涉猎学术界正在讨论中的新领域和新算法。同时，本书强调了实现一个企业里真正实用的算法系统所需要的技能，比如采样、特征工程、模型评估。因为机器学习算法往往需要比较深的背景知识，所以在每个问题和解答之前，会对该领域做简单的背景介绍。每个问答有不同的难度，以供读者自我衡量。

在核心的机器学习算法问答内容之外，我们增加了两个部分，一是“机器学习算法工程师的自我修养”，介绍业界典型的算法工程师的工作内容和要求。这些实例可以帮助广大的读者了解掌握机器学习技能以后的工作和去向。二是“人工智能热门应用”，相信不少读者都听说过这些应用的故事，比如无人驾驶车、AlphaGo 等。我们希望从内行人的角度，解释一下这些超级应用背后的原理是什么。当你读完本书，掌握了机器学习技能以后，你也可以在幕后操作这些热门的智能应用了。

本书信息量很大，涉猎人工智能和机器学习的各个子领域。每个公司、每个业务、每个职位，不一定会用到全部的技能。所以关于阅读这本书，我有以下几个建议。

（1）顺读法：从头至尾阅读。如果你能读懂全部内容，所有的题目都会解答，欢迎你到 Hulu 来申请工作吧！

（2）由简至难法： 每道题的旁边都标明了难度。一星最简单，五星最难。在本书中，还提供了一个题目的列表。一颗星的题目，主要是介绍基本概念，或者是为什么要做某一件事，比如 “什么是 ROC 曲线？” “为什么需要对数值类型的特征做归一化？” 。如果你是机器学习的入门学习者，可以从背景知识和简单的题目出发，循序渐进。

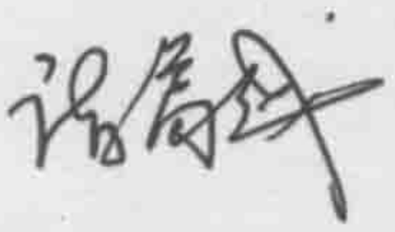
（3）目标工作法：不是所有的公司、所有的职位都需要懂得各类算法。如果你目前的工作或者想去的工作在某个领域，它们可能会用到某几类算法。如果你对某个新的领域很感兴趣，比如循环神经网络，那你可以专攻这些章节。不过无论用哪类算法，特征工程、模型评估等基本技能都是很重要的。

（4）互联网阅读法：一本书很难把广泛的领域讲得面面俱到，尤其是题目和解答，可以举一反三有很多花样。所以，我们在很多章节后都有总结和扩展。对某个领域感兴趣的朋友们，可以以这本书为起点，深入到扩展阅读，成为这一方面的专家。

（5）老板读书法：如果你是一个技术管理者，你需要解决的问题是算法可能对你现有的技术体系有什么帮助，和怎么找到合适的人，帮你做出智能的产品。建议你可以粗略地浏览一下本书，了解机器学习的各个技术领域，找到合适的解决方案。然后，你就可以用本书作面试宝典了。

这本书出版的目的，是让更多的人练习和掌握机器学习相关的知识，帮助计算机行业人员了解算法工程师需要的实际技能，帮助软件工程师成为出色的数据科学家，帮助公司的管理者了解人工智能系统需要的人才和技能，帮助所有对人工智能和机器学习感兴趣的朋友们走在技术和时代的前沿。

人工智能和机器学习的算法还在日新月异地发展中，这本书也会不断更新，不断地出新版本。希望得到读者朋友们的悉心指正，让我们一起跟上这个技术领域的进步步伐。



2018 年 4 月 10 日

机器学习算法工程师的自我修养

通往机器学习算法工程师的进阶之路是崎岖险阻的。《线性代数》《统计学习方法》《机器学习》《模式识别》《深度学习》，以及《颈椎病康复指南》，这些书籍将长久地伴随着你的工作生涯。

除了拥有全面、有条理的知识储备，我认为，想成为一名优秀的算法工程师，更重要的是对算法模型有着发自心底的热忱，对研究工作有一种匠心精神。这种匠心精神，直白来讲，可以概括为：发现问题的眼光、解决问题的探索精神，以及对问题究原竟委的执着追求。这里，我想给大家分享一个小故事，也是发生在本书作者身边真实的情景。

在微信红包占领家家户户年夜饭的那个时代，我们的小伙伴也没有例外。一群心有猛虎、细嗅蔷薇的算法研究员深切意识到自己不仅手速慢，运气也可谓糟糕。在埋头疯点手机屏幕的间隙，他们查阅了抢红包策略的相关文献，发现国内外对这一理论框架的探究极度匮乏。知识拯救命运，他们决定将红包机制的公平性提升到理论高度。通过大量的模拟实验，统计在不同顺位领到红包的大小。数据分析显示，越后面领到红包的人，虽然红包金额的期望（均值）和前面的人相同，但方差会更大，这也意味着他们更容易获得一些大额红包。从此，掌握这一规律的研究员们在各个群中“屡试不爽”，再也没有抢到过红包，留下的只有“手慢了，红包派完了”几个大字。

新年钟声敲响的时分临近，Boss 级别的人物往往会在群里发一些超级大额的红包。最夸张的一次有一位幸运儿在 10 人红包中领到 2 角钱，还没来得及在心中完成“老板真抠门”的碎碎念，抬头定睛一看，最佳手气 500 多元。判若云泥的手气虽没有埋下同事关系间的芥蒂，却让这帮算法工程师们产生了新的思考——如果把大额红包分成多份给大家抢，会减小“人品”因素带来的“贫富差距”吗？理论结合实际，他们不仅通过数学推导确认这一结论，还设计了一系列实验证明了多个红包的确会缩小不同人领到红包金额之间的差异性（方差）。从此，他们

组的 Leader 在发大红包的时候都会刻意平均分成几份，既增加了大家抢红包的乐趣，又避免了有人因运气不佳而扼腕兴叹的愤懑。

当然，故事不止于此。他们还利用红包的特性编写了一系列面试题，筛选着一批又一批的机器学习算法工程师，例如，“用红包产生随机数”“用红包随机选出 n 个候选人”，诸如此类源自生活的小问题在本书后续章节中亦不难寻其踪迹。

这种探究问题的匠心精神充斥着他们生活的各个角落。每天下楼吃饭等电梯的时候，因担心上厕所错过电梯，他们建立多个模型分析不同时段电梯平均等待时间对应厕所时机的最优选择；在夕阳的余晖下欣赏湖光塔影时，他们会思考为何粼粼波光成了图像编码中的棘手难题；打开购物 APP 看着目不暇接的喜欢抑或不喜欢的商品，他们反思自己搭建的推荐系统是否也会让用户有着相同的无奈或是欣喜。每一件小事，因为对研究有了热爱，都可以成为工作的一部分，成为开启机器学习大门的钥匙。

工作中的算法工程师，很多时候，会将生活中转瞬即逝的灵感，付诸产品化。组里的一位同事在看某国产剧的时候，发现可以非常方便地跳过片头和片尾。从消费者的角度出发，这的确是一个大有裨益的产品特征，于是他仔细统计了我们自己平台的视频源数据，发现只有一部分视频含有片头、片尾的时间点信息，而且都是人为标记的。试想，对于一家具有百万量级内容源的视频公司，在所有的剧集上人为标记片头、片尾信息有如天方夜谭。通过广泛的背景调研、方法尝试，攫取前人工作之精华，不断加以创新，依据自己的数据特点量体裁衣，他们的团队设计出了一种基于深度神经网络与浅层特征融合的片尾自动检测模型。经过反复的迭代与充分的实验，得到了令人满意的结果。这一工作也申请了美国发明专利，并一步步走向产品化。

将算法研究应用到工作中，与纯粹的学术研究有着一点最大的不同，即需要从用户的角度思考问题。很多时候，你需要明确设计的产品特征、提升的数据指标，是不是能真正迎合用户的需求，这便要求算法工程师能在多个模型间选择出最合适的那个，然后通过快速迭代达到一个可以走向产品化的结果。这种创新精神与尝试精神便是“匠心”一词在工作

中的体现。

当然，匠心精神诚可贵，知识储备作为成功的根底亦必不可少，这也是我们写作这本书的初衷。扎实的数学基础、完整的算法体系、深入的模型理解，是我们想传达给读者的精华之所在。本书前几章内容，如特征工程、模型评估、经典模型等，是机器学习领域的基石，是每个算法工程师应该融会贯通，内化于自己知识体系中的。而想成为一个研究专业或是应用领域的专家，则需要在技能树中的某几个分支不断生长发展。或许大家都听过啤酒与尿布的小故事，但搭建一个成熟、稳定的推荐系统，不仅需要通晓降维（第 4 章）、优化算法（第 7 章），更要对神经网络（第 9 章、第 10 章）、强化学习（第 11 章）等新生代模型不断钻研、深入理解，将学术前沿与产品形态紧密结合。例如，若是在技能树中专攻马尔可夫模型、主题模型（第 6 章），建立完整的概率图模型知识网络，并将循环神经网络（第 10 章）的理论体系融会贯通，形成自己独到的理解和感悟，便可以在机器翻译、语音聊天助手等自然语言处理的应用场景中驾轻就熟，游刃有余。

成为机器学习算法工程师的道路固然崎岖，却充满着旖旎和壮阔。你需要做的只是，想清自己真正想成为的那个角色，踏踏实实地在本书中汲取足够多的养分，然后，静静合上书页，在生活中体会种种细节，感受机器学习的璀璨多姿。

葫芦娃

2018 年 4 月