THE MASTER ALGORITHM

HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD

终极算法

机器学习和人工智能如何重塑世界

(Pedro Domingos)

[美] 佩德罗·多明戈斯◎著

黄芳萍◎译

出版集团

终极算法

机器学习和人工智能如何重塑世界

[美] 佩德罗·多明戈斯◎著 (Pedro Domingos) 黄芳萍◎译

中信出版集团 · 北京

图书在版编目（CIP）数据

终极算法：机器学习和人工智能如何重塑世界 /（美）佩德罗·多明戈斯著；黄芳萍译. -- 北京：中信出版社，2017.1 (2017.3重印)

书名原文：The Master Algorithm:How the Quest for the Ultimate Learning Machine Will Remake Our World

ISBN 978-7-5086-6867-3

I. ①终… II. ①佩… ②黄… III. ①机器学习－研究②人工智能－研究 IV. ①TP18

中国版本图书馆CIP数据核字（2016）第256013号

终极算法：机器学习和人工智能如何重塑世界

著　　者：[美]佩德罗·多明戈斯
译　　者：黄芳萍
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座　邮编　100029）
承 印 者：北京通州皇家印刷厂

开　　本：880mm×1230mm　1/32　　印　　张：13.5　　字　　数：298千字
版　　次：2017年1月第1版　　印　　次：2017年3月第4次印刷
京权图字：01-2014-7284　　广告经营许可证：京朝工商广字第8087号
书　　号：ISBN 978-7-5086-6867-3
定　　价：68.00元

献给我的姐姐瑞塔

在我写作本书的过程中，她没能战胜病魔

所有科学中最重大的目标就是，从最少数量的假设和公理出发，用逻辑演绎推理的方法解释最大量的经验事实。

——阿尔伯特·爱因斯坦

通过拓展我们不假思索就能运算的能力，文明就会大为进步。

——阿尔弗雷德·诺思·怀特海

作为一位机器学习领域研习10年以上的专业技术人员，我当初入行的时候没有想到，短短的10年间，这项技术会如此快速地改变众多行业，并影响全球数十亿用户生活的方方面面。在今天，当你用今日头条浏览新闻资讯的时候，当你用网易云音乐查看推荐歌单的时候，当你在百度搜索信息的时候，当你在互联网金融平台申请借款的时候，甚至在你调戏Siri和小冰的时候，其实都是其背后的机器学习算法在云端服务器中为你默默服务。但对于这样一种重要技术，市面上一直缺少一本适合普通读者的入门科普读物，而众多的专业书籍要求读者具备一定的高等数学和计算机基础算法知识，并不适合科普的需要。直到中信出版社的朋友将这本书的翻译稿推荐给我时，我欣慰地发现，这正是想了解一点机器学习的普通读者所需要的啊。本书的作者多明戈斯是华盛顿大学的终身教授，也是一位在机器学习领域具有20年研究经历的资深科学家。多明

戈斯一直致力于融合各种机器学习算法的优势，提出一种可以解决所有应用问题的通用算法，即终极算法。在这本书里，作者详细地阐述了他的思路。其实我个人在阅读本书的过程中，始终对“终极算法”的提法充满怀疑。在我看来，机器学习作为人工智能领域的主流技术，在现实社会中一直以技术工具的面目为人所知。不同的技术流派和相应算法往往可以很好地解决一些问题，却对另一些问题一筹莫展。所谓的终极算法真的存在吗？如果存在，有价值吗？

可以拿内燃机举个例子，就我这个外行来说，也知道存在活塞式发动机、涡喷发动机、涡轴发动机、涡扇发动机、涡桨发动机、冲压发动机等不同种类的内燃机。不同的内燃机特性迥异，适用的工况也不尽相同。小到家用小汽车，大到导弹驱逐舰，人类制造的各种机动设备，都可以根据自己的效率需求、动力需求、寿命需求，乃至启动速度等多种需求维度选择发动机种类。如果有人非要搞个终极内燃机，并企图用这种内燃机替代现存的各类内燃机，为所有大大小小、需求不同的机动设备提供统一动力，估计大概率是要失败的。这种通用的终极内燃机如果能搞出来，在大部分领域肯定竞争不过各领域的专用内燃机，或者成本太高，或者能效太低。

带着这种疑问，我通篇读下来之后才发现作者的另一层用意。诚如作者所说，很多普通人可能没有意识到自己的生活中机器学习算法的影响已经无处不在，机器学习已经在逐渐接管现实世界。大众对这样一种技术的认知程度和该技术的重要性相比显得远远不够，在不远的未来，了解机器学习并有能力利用机器学习改进自己工作的人在职业发展上会具备巨大的优势。“不要和人工智能对抗，

要让人工智能为你服务”是作者诚挚的忠告。而要利用好机器学习这个工具，并不一定需要读一个计算机博士学位，但有必要了解一些基本的概念，了解各种技术的优缺点和能力边界。正如一位称职的驾驶员不必了解具体怎么制造汽车发动机，但是对发动机的工作原理和种类还是需要略知一二的。因此，相比一板一眼地介绍机器学习的典型算法，作者设计了一个更引人入胜的套路：先抛出一个“是否存在一种终极算法”的问题，然后带着读者一章一章地回顾机器学习发展史上的重要流派和代表算法。每回顾一派，就鼓励读者思考终极算法应该如何借鉴这类算法的优点。好奇的普通读者带着疑问读完本书后，不论其是否相信终极算法的存在，至少对各类算法都会有一定的印象。以讨论终极算法为名，行科普之实，到这一步，我觉得作者的目的已经达到一半了。

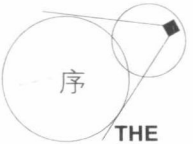
另外，在文末作者还提到，无论终极算法是否存在，他希望这个大胆的问题能够激发部分读者的好奇，甚至被这个问题吸引成为机器学习的专业研究人员。确实，每一种学科都需要至高的理想驱动向前，就如同物理的大一统理论，当无数杰出的天才为一个终极问题孜孜以求时，就算这个问题本身在这些人的有生之年可能没有答案，但是这个学科一定会因为这些伟大的探索历程取得辉煌的进步。我想，这也许是作者因为对机器学习的热爱夹带的另一个私货吧。

作为今日头条的一位算法架构师，我倒是希望头条用户都能陷入作者的“圈套”，带着好奇心，好好读读这本书。如果大多数用户都能了解一些机器学习的基础知识，应该就能够更好地和推荐算

法互动，不断把算法调教得更好，更符合自己真正的兴趣，而不会因为算法一开始推荐的内容不好就放弃这个产品。诚如作者所说，也许在未来，对应人类的心理学，也会出现机器心理学，了解一点机器人的心理，会让你和机器的互动更有效率，也会让机器更快地成为你忠实、不知疲倦的助手。

曹欢欢

今日头条首席算法构架师



你也许不知道，但机器学习就在你身边。当你把查询信息输入搜索引擎时，它确定该向你显示哪些搜索结果（包括显示哪些广告）。当你打开邮箱时，大部分垃圾邮件你无法看到，因为计算机已经把这些垃圾邮件过滤了。登录亚马逊网站购买一本书，或登录网飞（Netflix）公司网站观看视频，机器学习系统会推荐一些你可能喜欢的产品。脸书（Facebook）利用机器学习决定该向你展示哪些更新，推特（Twitter）也同样会决定显示哪些文章。你使用计算机的任何时候，都有可能涉及机器学习。

传统上认为，让计算机完成某件事情的唯一方法（从把两个数相加到驾驶飞机），就是非常详细地记录某个算法并解释其如何运行。但机器学习算法就不一样：通过从数据中推断，它们自己会弄明白做事方法。掌握的数据越多，它们的工作就越顺利。现在我们不用给计算机编程，它们自己给自己编程。

机器学习不仅存在于网络空间，它还存在于你每天的生活中：从你醒来到入睡，每时每刻无所不在。

早上7点你的收音机闹钟响起，播放的是你之前从未听过的歌曲，但你的确很喜欢这首歌。Pandora电台（可免费根据你的喜好播放歌曲）的优势在于，根据你听的音乐，电台掌握了你的品位，就像你自己的radio jock账号一样。这些歌曲本身可能借助机器学习来播放。接下来你吃早餐，阅读早报。早报在几个小时前印好，利用学习算法，印刷过程经过仔细调整，以免报纸出现折痕。你房间的温度刚刚好，电费明显少了很多，因为你安装了Nest智能温控器。

你开车去上班，车持续调整燃油喷射和排气再循环，以达到最佳的油耗。你利用一个交通预报系统（Inrix）来缩短高峰时段上下班的时间，这当然能减缓你的压力。上班时，机器学习帮你克服信息超载。你利用数据立方体来汇总大量数据，从每个角度观察该立方体，获取最有用的信息。你要决定是采用布局方案A，还是采用布局方案B，以便为网站带来更多的业务。网络学习系统会尝试两种布局方案，并给予反馈。你得对潜在供应商的网站进行调查，但网站的语言是外语。没关系，谷歌会自动为你翻译。E-mail会自动分类并归入相应的文件夹，只把最重要的信息留在邮箱里，非常方便。文字处理软件帮你查找语法和拼写错误。你为即将到来的行程查找到一个航班，但决定推迟购买机票，因为必应旅行（Bing Travel）预测票价很快会下降。也许你没有意识到以上这些，要不是机器学习帮助你，你可能要马不停蹄地亲自做

很多事情。

你在休息时间查看自己的共同基金，大部分基金利用学习算法来选股，其中的某些基金完全由学习系统运作。午餐时间到了，你走在大街上，想找个吃饭的地方，这时候用手机上的Yelp点评应用程序来帮助你。你的手机充满了学习算法，它们努力工作，改正拼写错误、理解口头指令、减少传输误差、识别条形码，还有其他很多事情。手机甚至可以预测你接下来会做什么，然后依此给出建议。例如，当你吃完午餐后，它会小心翼翼地提示你，下午和外地来访者的会面要推迟，因为她的航班延误了。

下班时夜幕已降临，你走向自己的车，机器学习会保证你的安全，监测停车场监控摄像头的录像，如果探测到可疑人的行动，它会提示不在场的安保人员。在回家路上，你在超市门口停车，走向超市货物通道，通道借助学习算法进行布置：该摆放哪些货物，通道末尾该展示哪些产品，洋葱番茄辣酱是否该放在调味酱区域，或是放在墨西哥玉米片旁边。你用信用卡付款。学习算法会向你发送信用卡支付提示，并在得到你的确认后完成支付。另外一个算法持续寻找可疑交易，如果它觉得你的卡号被盗，则会提示你。还有一种算法尝试评估你对这张卡的满意度，如果你是理想的客户但对服务不太满意，银行会在你决定换卡之前，为你提供更贴心的服务。

你回到家，走到信箱旁，发现有朋友的一封来信，这是通过能阅读手写地址的学习算法派送的。当然也会有垃圾来信，由另外的学习算法进行选择。你停留了一会儿，呼吸夜晚清新凉爽的空气。你所在城市的犯罪率明显下降了，因为警察开始使用统计

算法来预测哪里的犯罪率最高，并在那里集中巡警力量。你和家人共享晚餐。市长出现在新闻里，你为他投票，因为选举那天，学习算法确定你为“关键未投票选民”之后，他亲自给你打了电话。吃完晚餐，你观看球赛，两支球队都借助统计学习来挑选队员。你也可能和孩子们在Xbox上玩游戏，Kinect[①]学习算法确定你在哪里、在做什么。你在睡前吃药，医生通过学习算法的辅助来设定和检测吃药的最佳时间。医生也可能利用机器学习来帮你诊断疾病，例如，分析X射线结果并弄明白一系列非正常症状。

机器学习参与了你人生的每个阶段。如果你为了参加SAT大学入学考试（美国学术能力评估测试）而在网上学习，某学习算法会给你的练习短文打分。如果你申请商学院，且最近要参加GMAT（经企管理研究生入学考试），其中的一个文章打分工具就是一个学习系统。可能当你求职时，某学习算法会从虚拟文件中挑选出你的简历，并告诉未来的雇主：这位是很不错的人选，看看吧。最近公司给你加薪可能还多亏另外的学习算法。如果想买套房子，Zillow.com网站会估算你看中的每套房的价值，接着房子就有了着落。之后申请住房贷款，某学习算法会研究你的申请，并建议是否可以通过申请。最重要的是，如果你使用在线约会服务，机器学习甚至可能帮你找到人生挚爱。

社会在不断变化，学习算法也是如此。机器学习正在重塑科学、技术、商业、政治以及战争。卫星、DNA（脱氧核糖核酸）

① Kinect是微软对Xbox360体感周边外设正式发布的名字。——编者注

测序仪以及粒子加速器以前所未有的精细程度探索自然，同时，学习算法将庞大的数据转变成新的科学知识。企业从未像现在这样了解自己的用户。在美国大选中，拥有最佳选举模型的候选人奥巴马最终战胜了对手罗姆尼，获得了竞选胜利。无人驾驶汽车、轮船、飞机分别在陆地、海面、空中进行生产前测试。没有人把你的喜好编入亚马逊的推荐系统，学习算法通过汇总你过去的购买经历就能确定你的喜好。谷歌的无人驾驶汽车通过自学，懂得如何在公路上平稳行驶，没有哪个工程师会编写算法，一步一步指导它该怎么走、如何从A地到达B地——这也没必要，因为配有学习算法的汽车能通过观察司机的操作来掌握开车技能。

机器学习是“太阳底下的新鲜事”：一种能够构建自我的技术。从远古祖先学会打磨石头开始，人类就一直在设计工具，无论这些工具是手工完成的，还是大批量生产的。学习算法本身也属于工具，可以用它们来设计其他工具。“计算机毫无用处，”毕加索说，“它们只能给你提供答案。”计算机并没有创造性，它们只能做你让它们做的事。如果你告诉它们要做的事涉及创造力，那么就要用到机器学习。学习算法就像技艺精湛的工匠，它生产的每个产品都不一样，而且专门根据顾客的需要精细定制。但是不像把石头变成砖、把金子变成珠宝，学习算法是把数据变成算法。它们掌握的数据越多，算法也就越精准。

现代人希望让世界来适应自己，而不是改变自己来适应世界。机器学习是100万年传奇中最新的篇章：有了它，不费吹灰之力，世界就能感知你想要的东西，并依此做出改变。就像身处魔法林，

在你通过时，周围的环境（今天虚拟，明天现实）会进行自我重组。你在树木和灌木中选出的路线会变成一条路，迷路的地方还会出现指路标志。

这些看似有魔力的技术十分有用，因为机器学习的核心就是预测：预测我们想要什么，预测我们行为的结果，预测如何能实现我们的目标，预测世界将如何改变。从前，我们依赖巫医和占卜师进行预测，但他们太不可靠；科学的预测就更值得信赖，但也仅限于我们能系统观察和易于模仿的事物，大数据和机器学习却大大超出这个范围。我们可通过独立的思维来预测一些常见的事情，包括接球和与人对话，但有些事情，即便我们很努力，也无法预测。可预测与难以预测之间的巨大鸿沟，可以交给机器学习来填补。

矛盾的是，尽管学习算法在自然和人类行为领域开辟了新天地，但它们仍笼罩在神秘之中。媒体每天都报道涉及机器学习的新闻：苹果公司发布Siri个人助理，IBM[①]沃森（IBM的超级计算机）在《危险边缘》游戏中战胜了人类，塔吉特（Target）能在未成年妈妈的父母发现之前通知她怀孕，美国国家安全局在寻找信息连接点……在这些新闻事件中，学习算法如何起作用仍不得而知。计算机“吞入”数以万亿的字节，并神奇地产生新的观点，关于大数据的书籍甚至也避谈“这个过程到底发生了什么”。我们一般认为学习算法就是找到两个事件之间的联结点，例如，用谷歌搜索“感冒

① IBM，国际商业机器公司。——编者注

药”和患感冒之间的联系。然而，寻找联结点与机器学习的关系就像是砖与房子的关系，房子是由砖组成的，但一堆砖头肯定不能称之为“房子”。

当一项新技术同机器学习一样流行且具有革命性时，不弄明白其中的奥妙实在太可惜。模棱两可会导致误差和滥用。亚马逊的算法能断定当今世界人们在读什么书，这一点比谁都强。美国国家安全局的算法能断定你是否为潜在恐怖分子。气候模型可以判定大气中二氧化碳的安全水平。选股模型比我们当中的多数人更能推动经济发展。你无法控制自己理解不了的东西，这也是追求幸福的公民、专家或普通人需要了解机器学习的原因。

本书的第一个目标就是揭示机器学习的秘密。只有工程师和机修工有必要知道汽车发动机如何运作，但每位司机都必须明白转动方向盘会改变汽车的方向、踩刹车会让车停下。当今极少有人知道学习算法对应的原理是什么，更不用说如何使用学习算法。心理学家丹·诺曼（Don Norman）创造了“概念模型”（conceptual model）这个新词，代指为了有效利用某项技术而需粗略掌握的知识。本书就将介绍机器学习的概念模型。

并不是所有算法的工作原理都相同，这些差异会产生不同的结果，比如亚马逊和网飞的推荐系统。假设这两个系统试着根据“你喜欢的东西”来对你进行引导，亚马逊很有可能会把你带到你之前常浏览的书籍类别，网飞则可能会把你带到你不熟悉且似乎有点奇怪的区域，并引导你爱上那里。在本书当中，我们会看到诸如亚马逊、网飞之类的公司使用的各式各样的算法。与亚马

逊相比，网飞公司的算法对你的爱好理解得更深（尽管还是很有限），然而具有讽刺意味的是，这并非意味着亚马逊也应该利用这个算法。网飞的商业模式是依靠晦涩的电影、电视节目的长尾效应来推动需求，这些电影和节目的成本很低。它一般不推荐大片，因为你的会员订阅费可能有限。亚马逊则没有这样的问题：尽管擅长利用长尾效应，但它同样乐意把更昂贵的热销商品卖给你，这也会简化其物流工作。对于那些奇怪的产品，如果是订阅会员可免费享用的，我们可能会乐意去尝试，而如果需要另外掏钱，我们去选择它们的可能性就小得多。

每年都会出现上百种新的算法，但它们都是基于几个相似的基本思路。为了明白机器学习如何改变世界，你有必要理解这些思路。本书就将对此进行介绍。学习算法并不是那么深奥难懂，除了运用在计算机上，对于我们来说很重要的问题都可以通过学习算法找到答案，比如：我们如何学习？有没有更好的方法？我们能预测什么？我们能信任所学的知识吗？对这些问题，机器学习的各个学派有不同的答案。

机器学习主要有5个学派，我们会对每个学派分别介绍：符号学派将学习看作逆向演绎，并从哲学、心理学、逻辑学中寻求洞见；联结学派对大脑进行逆向分析，灵感来源于神经科学和物理学；进化学派在计算机上模拟进化，并利用遗传学和进化生物学知识；贝叶斯学派认为学习是一种概率推理形式，理论根基在于统计学；类推学派通过对相似性判断的外推来进行学习，并受心理学和数学最优化的影响。在构建机器学习的目标推动下，我