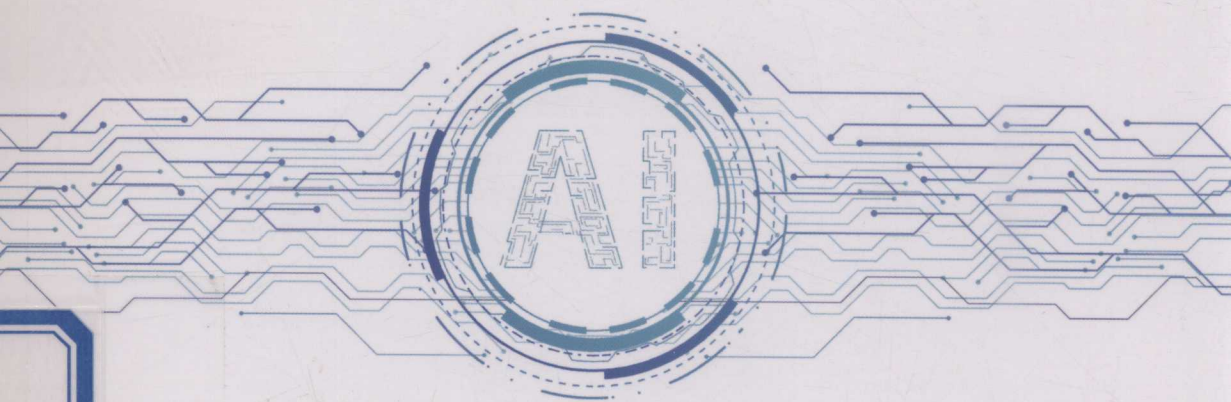
From the Statistical World to Artificial Intelligence

Practical Cases and Algorithms

从统计世界走向人工智能

——实战案例与算法

陆培丽 著



科学出版社

从统计世界走向人工智能

——实战案例与算法

陆培丽　著

科学出版社

北京

内 容 简 介

本书叙述了从数学到统计、从统计到人工智能的发展，结合大量的实际商业应用案例介绍了诸多经典的机器学习算法，比如 LASSO 回归、MCMC、决策树、随机森林和神经网络等。本书将案例与算法结合，基于人工智能的场景，从理论到实际操作层层递进，读者从中可以学习从需求到分析，再到结论的实际编程方法。当读者阅读完本书后，不仅可以了解实际问题的需求，而且可以学习到解决问题的算法。

本书适合作为统计学、应用统计、人工智能、大数据、金融、经济与管理等专业大学生的教学用书，开拓他们不同维度的学习思路，以及在理论学习中灵活应用人工智能模型的知识与 Python 的能力。金融从业人员可以通过阅读本书免去烦琐的数据整理等工作，提高工作效率，包括在财报分析、银行信用画像以及投资等领域。

图书在版编目（CIP）数据

从统计世界走向人工智能：实战案例与算法/陆培丽著. —北京：科学出版社，2020. 3

ISBN 978-7-03-063624-9

Ⅰ. ①从… Ⅱ. ①陆… Ⅲ. ①人工智能–算法–研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字（2019）第 273877 号

责任编辑：李静科 郭学雯 / 责任校对：彭珍珍

责任印制：吴兆东 / 封面设计：无极书装

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京虎彩文化传播有限公司 印刷

科学出版社发行 各地新华书店经销

*

2020 年 3 月第 一 版 开本：720×1000 B5

2020 年 3 月第一次印刷 印张：10 3/4 插页：3

字数：207 000

定价：68.00 元

（如有印装质量问题，我社负责调换）

序

如同蒸汽时代的蒸汽机、电气时代的发电机、信息时代的计算机和互联网一样，人工智能正成为推动人类进入智能时代的决定性力量。当前，新一轮科技革命和产业变革正在萌发，大数据的形成、理论算法的革新、计算能力的提升及网络设施的演进驱动人工智能发展进入新阶段，智能化成为技术和产业发展的重要方向。人工智能具有显著的溢出效应，将进一步带动其他技术的进步，推动战略性新兴产业总体突破，正在成为推进供给侧结构性改革的新动能、振兴实体经济的新机遇、建设制造强国和网络强国的新引擎。错失一个机遇，就有可能错过一整个时代。

2017 年 7 月，国务院印发并实施《新一代人工智能发展规划》，标志着将人工智能提升到了国家发展战略层面。同年 11 月，上海市政府制订并发布了《关于本市推动新一代人工智能发展的实施意见》，以便更好地对接国家重大战略，牢牢把握、充分发挥上海在人工智能发展上的优势。该意见聚焦应用驱动、科技引领、产业协同、生态培育，全面实施智能上海 (AI@SH) 行动，计划于 2020 年基本建成国家人工智能发展高地。

在人工智能产业蒸蒸日上的发展之时，我收到了《从统计世界走向人工智能 —— 实战案例与算法》的书稿。陆培丽是上海交通大学数学科学学院的优秀校友，曾经受邀在上海交通大学数学科学学院建院 90 周年纪念大会上做主题演讲。她将自己在量化金融领域将近 20 年的积累与理解，同人工智能的算法与应用相结合，汇聚成了本书的一个个生动的案例，助推了金融科技的发展。这些案例都是她和她的研究团队在商业领域的实战经历，是人工智能产业化、市场化、商业化的生动写照。该书从数学讲述到统计，再从统计到人工智能的发展，结合大量的实际的应用案例，其中包括能源价格预测、财务分析、生物科技案例、银行证券金融、医学等背景领域，介绍了诸多经典的机器学习算法如何解决这些领域的问题，从易到难，逐步深入。该书出发点即为了将高校学生在课堂中学习的理论知识应用于具有商业模式背景的课题，适合作为统计学专业、应用统计专业、人工智能专业、大数据专业、金融专业、管理专业与经济专业大学生的教学用书，开拓他们不同维度的学习思路，并培养他们在理论学习中灵活应用人工智能模型的知识与编程的能力。人工智能、机器学习方向的学者，可以通过该书了解到商业模式，更明晰地了解如何将自己的理论研究成果应用于实践中。金融和实体业的应用者，通过该书可以学习到相关的人工智能技术实战程序操作方法，因此该书可以为金融机构和实体企业提供方法论和初步的解决方案。

该书对于大学生来说，是他们了解人工智能领域的敲门砖，为他们在人工智能领域的研究，以及未来就业的选择，都提供了有力的指导。任何领域的发展都离不开青年才俊的贡献，陆培丽和她的研究团队在该书中展现出的严谨的学术态度和产学结合的创新精神，着实令我欣慰。相信在他们的努力与奋斗之下，中国的人工智能产业会越来越好。

我们也相信，该书不仅仅适用于关注人工智能领域发展的各界人士，也将为大家打开人工智能在金融业和其他企业实战运用的大门，引领更多有识之士加入到人工智能的时代浪潮当中来。长风破浪会有时，直挂云帆济沧海。我们拭目以待。

毛军发

中国科学院院士、上海交通大学人工智能研究院院长

前　言

20 年前，我进入了上海交通大学数学系，在懵懂中选择了数学专业。大数据、神经网络与建模等名词是我在大学时就已经听到的。从那时候在讨论班上有几个研究生做数学理论的推导，到现在每个人都知道的人工智能，这个积累过程，我们走了 20 年。

我的工作生涯一直在金融投资领域，并且绝大多数时间和数学、量化、程序化交易相关。我的第一份工作是在高盛，部门就叫作 program trading，隶属高盛东京。这份工作一干就是十几年。我工作生涯最初是从和程序化交易打交道开始的；除此之外，我的工作也包括长期看盘投资和研究金融领域的二级市场。在我的职业生涯中，除了金融，数学和统计占了很重要的一部分。

目前，我主要致力于金融量化的投资领域，并且发展了金融和科技交叉领域。从我的工作中，我越来越感受到科技在金融中发挥的力量，尤其是人工智能在金融研究和投资领域发挥出的神奇力量。复杂的深度算法超越了一般的统计计量方法，在大数据的领域发挥了无可替代的作用。

我能有今天的成就，不只是自己的努力和坚持，更要感谢家人一直以来对我的支持、所有一起共事过的合伙人对我的帮助、老师对我的精心栽培。感谢我的导师叶中行教授一路以来对我的学习和工作的指导。感谢上海交通大学研究生院王亚光老师从当年学习上对我的指导到如今工作上的指导。感谢在本书出版过程中上海交通大学数学科学学院的老师和领导在各方面的大力支持，让我感受到上海交通大学数学科学学院几十年如一日对院友的支持。感谢在读或者已经在各个工作岗位上我带过的学生。如果没有研究小组成员的共同努力，本书是不可能完成的。这些成员是我见过的最优秀的学生，他们以极大的热情投入到研究实践中。他们孜孜不倦，在某些项目上花费了很多的时间。他们仿佛站在我的背后，让我感觉责任重大，激励我更加努力，快速进步。

希望我们研究团队的成员十年后再回顾他们的这个起点和这十年的职业生涯时，每个人都能感受到所取得的长足进步。我希望这本书能让更多的学生看到人工智能和商业背景结合的案例，并为他们的职业生涯起步奠定良好的基础。这里是我们研究团队的名单 (按姓氏拼音排序)：曹晓芳、郭强、金衍瑞、刘彬鑫、刘逊知、马海乾、秦浩洋、沈嘉琪、孙晓军、涂一辉、王琰驹、王奕能、杨佳敏、易超、周游、左文婷。

曹晓芳，她的数学功底超群，现为上海交通大学数学科学学院统计系二年级研

究生，她总会把数学转化成直观的语言，令团队成员受益匪浅。郭强，拥有理科生的身份，却有着胜于文科生的文笔和口才，极具才华。金衍瑞，我们团队的博士，他踏实、认真、自律，虽然内敛，但骨子里充满了冲劲，并在非结构化数据领域带给团队拓展性思路。刘彬鑫是我们团队最风趣幽默的小伙子，他主修化学，但是也在编程和人工智能的技术上得到了提高。刘逊知，我认识他的时候，他还是化学系本科三年级学生，他主动参加了由上海交通大学数学科学学院统计系教授们和我一起组织的金融实战讨论班，他对金融科技充满热情，在他身上我看到了年轻人的冲劲和热情。马海乾，对研究的领域刻苦钻研，善于分析与解决研究过程中产生的问题。秦浩洋，复旦大学数学科学学院硕士，他不仅在人工智能领域有深入的研究，同样能将研究成果清晰地表达、展现出来，综合能力强。沈嘉琪，学生工作的积极参与者，协调组织能力极强，擅于思考。孙晓军，在一家全球著名的互联网金融公司就业，他的贡献在神经网络的图像识别方面。涂一辉，上海交通大学“致远荣誉计划”数学博士，他不仅学业出色，而且工作能力极佳，他自发组织了学校范围内的机器学习讨论班，至今已坚持两年，我相信这种坚持的精神会伴随他一路的成功。王琰驹，将多年在统计专业的学习应用到人工智能领域的研究中，想法独特。王奕能，经管专业的学生，设计领域是他的专长。杨佳敏，思维敏捷、执行力强，对研究内容追求极致。易超，研究团队中唯一当过兵的人，具备坚毅的素质和说到做到的精神。周游，用支持向量机做了高频领域的研究。左文婷，毕业于上海交通大学数学科学学院，擅于用量化分析手段解决金融科技领域的问题。

本书的研究团队虽然是年轻的，但是具有极高的素质。在本书即将完成时，我们团队的所有研究人员来到了历届互联网大会召开的圣地 —— 乌镇，开展团建。我也希望研究团队里的每个人能够借此机会结合自己的理想和兴趣，树立远大的目标，将来能够在以数据算法为特点的人工智能金融分析及企业分析领域崭露头角，为中国金融科技领域的革新做出自己的贡献。

本书的初衷来自于大学教学案例和大学生理论联系实际的迫切需求。与众多市面上的人工智能的书籍不同，本书的特点是把案例与算法结合在一起。从案例引入到提出、解决问题的人工智能算法，由浅入深地介绍了算法，再从总结算法到提出衍生应用，形成了本书的独特风格。通过阅读本书，读者不仅能够看到人工智能场景，还可以学习从需求到分析再到结论的实际编程方法。本书的另一大特点就是从大学生必修的数学课 —— 统计分析 (最简单的回归分析) 一直到深度学习的算法，深入浅出地带读者领略大学的必修课程是如何作为基础在实际问题中发挥作用的。

我们撰写本书的目的是给大学生提供必要的教育、培训与支持，帮助大学生从普通的统计分析衍生学习人工智能的案例，从实战的角度为学生展示人工智能给商业、企业带来的变化，并且鼓励大学生用自己所能掌握的基本科学知识来解决各

个领域的实际问题，把课堂上学习的数学统计知识和人工智能实践联系起来，为踏上职业生涯做好准备，同时也在实战项目的学习中不断地探索自己的兴趣，为未来的职业规划做好充分的准备。

本书作为大学生的实战案例课程，从最简单的案例出发，一步一步引导学生进行研究开发，调整参数，并最终得到准确率较高的结果。其中所有的案例分析均是基于课堂上案例分析的成果。

本书是受上海交通大学研究生院院长王亚光教授委托，上海交通大学数学科学学院统计系主任韩东教授支持，在大家的鼓励和支持下完成的。作为上海交通大学的业界导师，我觉得为更多的愿意努力的学生提供职业前景分析，并带领他们走向更加光明的前程是我责无旁贷的使命。饮水思源，爱国荣校，我愿意在这条路上践行校训！

书中不妥之处在所难免，恳请读者批评指正。

陆培丽

2019 年 6 月于上海

目　录

第1章　数学 → 统计 → 人工智能

1.1　数学与统计

数学知识和数学家就好像浩瀚夜空中的繁星，璀璨夺目，光彩熠熠。数学源自于古希腊语，是研究数量、结构、变化以及空间模型等概念的一门学科。通过抽象化和逻辑推理的使用，数学在计数、计算、量度和对物体形状及运动的观察中产生。数学的基本要素是：逻辑和直观、分析和推理、共性和个性。古希腊数学家欧几里得用公理方法建立起一套完整的数学体系，这在数学的发展历史上有着里程碑式的意义，其中的理性思辨和严谨逻辑对于几何学、数学和科学的未来发展都有极大的影响。数学相关专业的同学对吉米多维奇那本著名的《数学分析习题集》应该都不陌生。

统计学与数学有着某种有趣而奇特的关系。统计学是应用数学的一个分支，主要通过概率论建立数学模型，收集所观察系统的数据，进行量化分析、总结，做出推断和预测，为相关决策提供依据和参考。它被广泛应用在各门学科，从物理和社会科学到人文科学，甚至被用在工商业及政府的情报决策上。从伯努利到贝叶斯，一代代科学家都在不断拓宽统计的领域。

随着数字化的进程不断加快，人们越来越多地希望能够从大量的数据中总结出一些经验规律从而为后面的决策提供一些依据。统计学不仅仅是对数据的统计，而且包含了调查、收集、分析、预测等。应用的范围十分广泛。统计数据包含不确定性，但这不意味着我们要忽略错误。每一个估计值都有一个置信区间，在 95%的时间内可以预期它是正确的，但我们永远不可能 100%地确定任何东西。但只要有足够多的数据，正确的模型就可以从噪声中分离出信号。这使得统计学在处理有许多未知的混杂因素的事物 (如模拟社会学现象或任何涉及人类决策的事物) 时成为一个强有力的工具。

1.2　数据与统计

1.2.1　动态的数据

我们通常把大数据和人工智能连接在一起讨论。现在的数据已经不是我们以往所认为的静态数据，而是动态的数据。思考数据的全新方式就是把数据想象成一

条“端到端”的供应链。数据的处理应该是获得、清洗、整合、策划和存储的动态过程。

人工智能需要数据，同时要在反馈中训练数据；不断地发展和吸收实时数据的算法才能同时提高数据的数量和质量。要实现互动，就必须让数据和我们的工作生活无缝衔接，这就需要数据在线，实时记录，不断更新数据，从新生的数据中不停地归纳演绎。并且，所采集的数据在实际业务中可以被随时应用，驱动下一个算法或者策略的产生。

数据最好是由互不相干的、易于访问的数据组成。例如，要识别社交网站的情绪走向，需要根据天气、购物者的特征、新闻事件甚至几乎任何可以想象得到的新数据维度来进行数据的追踪。数据可以是由数据服务提供商提供的数据或是任何人以免费方式获取的开源性数据。

提升数据的处理速度是因为数据有时效性，我们需要对整个数据处理链条给予加速处理。低频的数据，可能对于及时决策的权重比较低，被存储在较远的服务器上，而高频的数据是我们优先考虑的对象。这些数据由于被访问的频率高，被存储在高性能的系统中以便快速检索。

在用人工智能处理数据的过程中，如果我们提高人工智能的效用，简化数据访问的流程，利用好未使用但是可能有用的数据，那么如何挖掘一切有价值的“沉默数据”会是一个需要长期研究且极具价值的课题。

1.2.2　非结构化的数据

目前新的研究前沿是我们所使用到的非结构化信息，而信息的组成，10%来自结构化的数据，而 90%来自互联网中的非结构化数据，以文本为主。非结构化的数据分析对于我们来说是最前沿的研究。把这些非结构化的数据抽象出来，变成我们能够用的数据信息，关键在于构建以知识为节点的自动分析平台。

十年前，企业数据存储以百万兆字节为单位，这在当时被视为大的飞跃。而今，大数据远远超出了传统的结构化数据，包括了各种非结构化数据 —— 文本、音频、视频等。十年前我们得到的是已知的和准确的结构化数据，今日，我们得到的大量数据的来源是未知和不可靠的。

非结构化数据举例，该案例来自于本公司的实际金融科技项目研究之一。这是我们从公司财务报表中的财务附注信息中提取出来的重要文字结构。重要文字结构的算法基础来源于深度学习的算法，这里只论述非结构化数据。我们得到了重要的文本信息后，再从相应的财务附注的文本信息中提取到结构化数据，见图 1.1。

Keyword	Year	Code	sheet	i	j	Data	rol_head1	rol_head2	col_head1	col_head2	col_head3
应付账款	2016	002386		74	0	0 项目					
应付账款	2016	002386		74	0	1 期末余额					
应付账款	2016	002386		74	0	2 期初余额					
应付账款	2016	002386		74	1	0 流动资产：					
应付账款	2016	002386		74	1	1	期末余额		流动资产：		
应付账款	2016	002386		74	1	2	期初余额		流动资产：		
应付账款	2016	002386		74	2	0 货币资金					
应付账款	2016	002386		74	2	1 2,417,390,	期末余额		货币资金		
应付账款	2016	002386		74	2	2 1,422,717,	期初余额		货币资金		
应付账款	2016	002386		74	3	0 结算备付金					
应付账款	2016	002386		74	3	1	期末余额		结算备付金		
应付账款	2016	002386		74	3	2	期初余额		结算备付金		
应付账款	2016	002386		74	4	0 拆出资金					
应付账款	2016	002386		74	4	1	期末余额		拆出资金		
应付账款	2016	002386		74	4	2	期初余额		拆出资金		
应付账款	2016	002386		74	5	0 以公允价值计量且其变动计入当期损益的金融资产					
应付账款	2016	002386		74	5	1 6,145,470.	期末余额		以公允价值计量且其变动计入当期损益的金融资产		
应付账款	2016	002386		74	5	2	期初余额		以公允价值计量且其变动计入当期损益的金融资产		

图 1.1 非结构化数据处理展示图

1.2.3 商业场景的数据初始化

所谓数据初始化，不仅包括分析对象，例如，客户的经营数据、财务数据，还有更多维度的数据的记录、分析和融入，构成对分析对象全方位的描述。上述的结构化数据和非结构化数据的数据初始化就不是一个简单的课题，而是一个高成本和困难的事情。我们不仅需要数字被数据化，还需要文字信息被数据化，地理方位被数据化，图片被数据化，情绪指标被数据化。举例，最简单的分析对象，客户年龄数据就包含几十套标准，包括身份证上登记的年龄、实际经营者的年龄、心理分析出的年龄等。这些数据各有价值，但传统的方法难以对它们进行融合分析，而用某些深度学习方法就可以分析出比较有用的结果。

同时，数据初始化是一件非常重要，且能产生高收益的事情。它能把我们所看到的事物从现象变为可量化的形式，这是我们做任何分析的前提之一。从数据的初始化开始，再到数据的存储、提取、分析、反馈、输出、再反馈、表达，此过程中，数据作为一种媒介，贯穿始终。

1.2.4 统计中的数据与商业中的数据

在传统统计学中，一个随机样本可以在很大程度上推导出全局的特征。从样本到总体，并利用假设检验证明其可靠性是传统统计学的常用方法。传统统计学在证明其收敛性和收敛速度的基础上，为实践奠定基础，以静态数据为主。

而商业环境是动态的，始终处于不断变化的过程中。因为满足商业环境的数据是全样本记录，并且是不断动态收集和变化的，量更大，形式更复杂，因此难度也更大，更具研究价值[1]。

1.3 统计与人工智能

统计和人工智能都是从数据中创建模型，但目的不同。统计学家非常注重使用