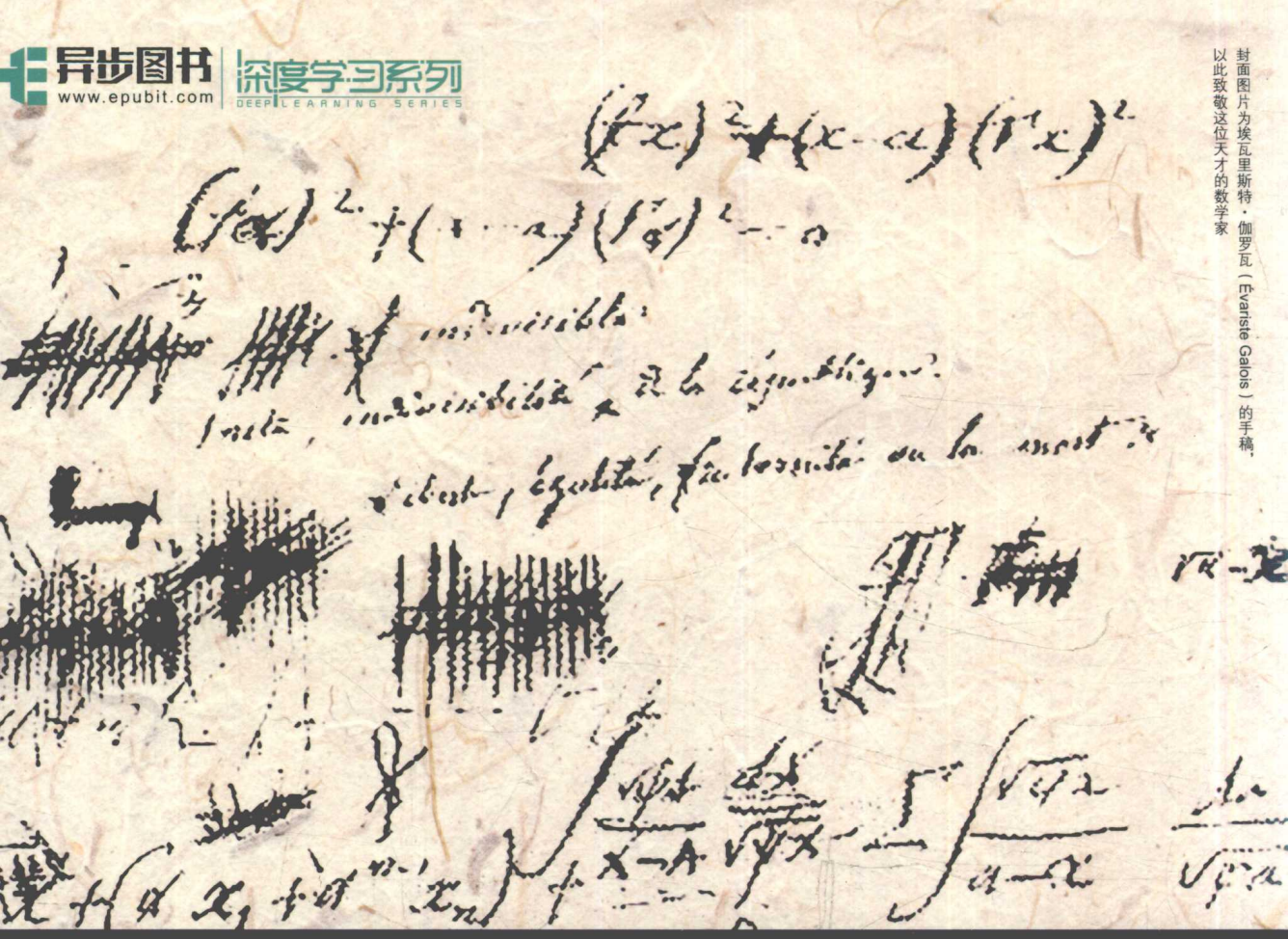
异步图书 www.epubit.com | 深度学习系列 DEEP LEARNING SERIES



封面图片为埃瓦里斯特·伽罗瓦（Évariste Galois）的手稿，以此致敬这位天才的数学家

精通数据科学：从线性回归到深度学习

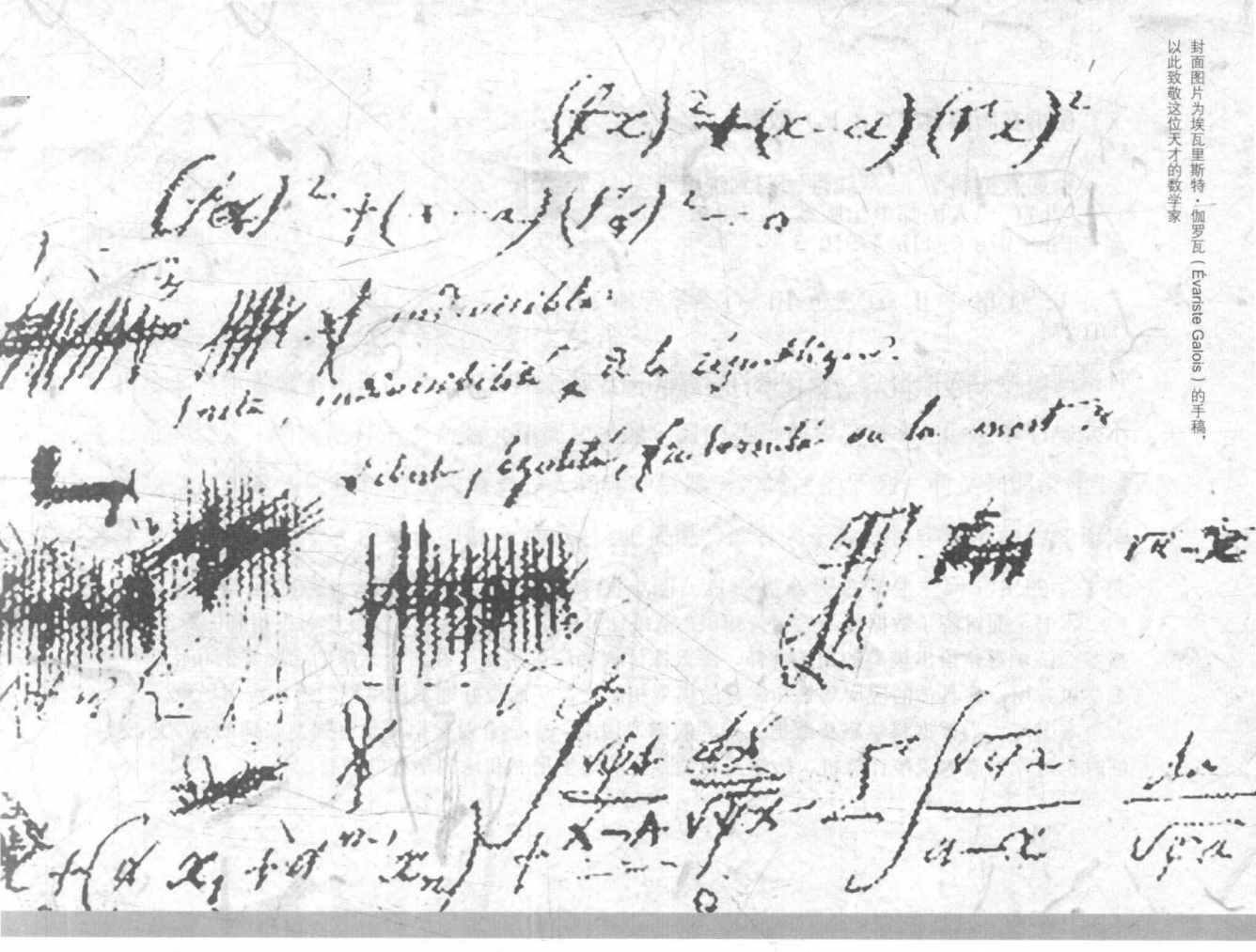
唐亘◎著

★将统计学、机器学习和计算机科学融会贯通，为读者搭建系统的知识体系

★以 Python 为工具，教会读者如何建模

★详解分布式机器学习、神经网络、深度学习等大数据和人工智能的前沿方向

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



封面图片为埃瓦里斯特·伽罗瓦（Évariste Galois）的手稿，以此致敬这位天才的数学家

精通数据科学：从线性回归到深度学习

唐亘 著

人民邮电出版社
北京

图书在版编目（CIP）数据

精通数据科学 : 从线性回归到深度学习 / 唐亘著
. -- 北京 : 人民邮电出版社, 2018.6
ISBN 978-7-115-47910-5

Ⅰ. ①精… Ⅱ. ①唐… Ⅲ. ①数据管理 Ⅳ.
①TP274

中国版本图书馆CIP数据核字(2018)第030929号

内 容 提 要

本书全面讲解了数据科学的相关知识，从统计分析学到机器学习、深度学习中用到的算法及模型，借鉴经济学视角给出模型的相关解释，深入探讨模型的可用性，并结合大量的实际案例和代码帮助读者学以致用，将具体的应用场景和现有的模型相结合，从而更好地发现模型的潜在应用场景。

本书可作为数据科学家和数据工程师的学习用书，也适合对数据科学有强烈兴趣的初学者使用，同时也可作为高等院校计算机、数学及相关专业的师生用书和培训学校的教材。

◆ 著　　唐　亘
责任编辑　张　爽
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
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：27
字数：549 千字　　2018 年 6 月第 1 版
印数：1 - 4 000 册　　2018 年 6 月北京第 1 次印刷

定价：99.00 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

序一

我与本书作者素不相识，读完作者发来的电子书稿后，感受到了以往在读技术类书籍时从未有过的惊喜。国内已有不少介绍大数据和机器学习的教科书和参考书，但这本书与众不同，它的重点不是传统教科书式的概念导入和各种机器学习算法的罗列，而是强调统计学、机器学习和计算机科学 3 门学科的融会贯通，试图呈现给读者关于数据科学较全面的知识体系。特别是对常用的统计和机器学习软件的详细说明，对提高在校大学生、研究生的动手能力和企业科技人员解决实际问题的能力大有裨益。

中国工程院院士，第三世界科学院院士，曾任中国科学院计算技术研究所所长

李国杰

序二

回首30年的新兴产业的来路，我们看到的或许是遵循着摩尔定律飞速增长的集成电路，或许是从互联网到移动互联网，再到物联网的更广泛的互联互通。但其背后，数据作为新兴产业的血液，其价值得到了广泛的认知和关注。早在2011年，我们完成了4篇大数据行业的前瞻报告，撰写《大数据时代的历史机遇》分析了大数据时代的产业机会与变革。后来又同申万宏源的计算机首席分析师刘洋一起勾勒了大数据产业的版图和发展路径。

如今作为一个大数据产业的实践者，我们看到大数据产业正如我们所预期的那样，成为了人工智能、虚拟现实以及区块链等新一轮新兴产业浪潮的核心，成为了传统产业转型升级的必备资源，成为了企业保持领先抑或实现弯道超车的必争之地。然而数据资源怎么用，数据模型如何建立，算法模型如何运用，成为了学界、产业界、资本界都在关注的关键问题。

本书站在数据学科的角度，融合了数学、计算机科学、计量经济学的精髓。不仅从“道”的层面为读者阐释了数据科学所要解决的核心问题——数据模型、算法模型的理论内涵和适用范围，而且从“术”的层面以常用的IT工具——Python为基础，教会读者如何建模以及通过算法实现数据模型，具有很强的实操性。在此基础之上，本书还为读者详解了分布式机器学习、神经网络、深度学习等大数据和人工智能的前沿技术。相信本书将成为数据科学工作者、数据工程师、数据产业实践者的必备手册，以及想要了解和学习数据科学的人员的首选教材。

易选股金融智能证券董事长，键桥通讯董事

易欢欢

前　言

和武侠世界里有少林和武当两大门派一样，数据科学领域也有两个不同的学派：以统计分析为基础的统计学派，以及以机器学习为基础的人工智能派。虽然这两个学派的目的都是从数据中挖掘价值，但彼此“都不服气”。注重模型预测效果的人工智能派认为统计学派“固步自封”，研究和使用的模型都只是一些线性模型，太过简单，根本无法处理复杂的现实数据。而注重假设和模型解释的统计学派则认为人工智能派搭建的模型缺乏理论依据、无法解释，很难帮助我们通过模型去理解数据。

从历史上来看，一门学科出现相互对立的学派通常意味着这门学科处于爆发的前夜，比如 20 世纪初的经济学，凯恩斯学派和新古典经济学派的长期论战极大地促进了宏观经济学的发展，并深刻地影响了各国政府的经济政策，并由此改变了人们的生活方式。现在数据科学也处在这样相似的位置和时间节点，它已经开始并将继续改变我们的世界。

抛开这些学术上的纷争，在实际工作中应该采用哪个学派的方法来解决数据挖掘的问题呢？答案是两者都需要，而且两者都重要。在某些应用场景中，比如图像识别领域，人工智能模型有非常惊艳的表现。虽然人们还没弄清楚这些模型的工作原理，但并不妨碍它们在现实中发挥作用。事实上，人类在很多其他领域里也是这种实践先行的状态。

但在更多的应用场景中，统计学派的方法则显得更为重要。我曾在欧洲的一家保险公司里参与过一个车险定价的项目，在这个项目里，数据科学家们主要尝试了两类模型，一类是很容易解释的逻辑回归和决策树模型，另一类是较为复杂的随机森林模型。随机森林模型的预测效果更好，如果将其投入生产中，仅在法国每年就能产生数千万欧元的利润。但问题是随机森林模型难以解释，监管部门根本不接受，所以只能退而求其次，使用效果较差但更易解释的决策树模型。抛开监管层的要求不说，模型的可解释性也是非常重要的。试想一下，顾客去保险公司购买车险时，被告知需要比别人花更多的钱，而对方提供的理由是，有一个不好解释的模型预测出顾客需要付更多钱，我想大部分顾客会难以接受这样的理由和做法吧。

上述的两种建模方式虽然在处理数据的方法上有很大差异，但它们有一个共同的“物质基础”——计算机。只有借助计算机强大的运算能力，我们才能在工程上实现搭建好的模型，使之发挥作用。因此，数据科学是统计学、机器学习以及计算机科学 3 门学科的交叉，涉及的知识点和技能点很庞大且复杂。如果能将这 3 门学科融会贯通，那么就能描绘出有关数据科学的全景图，进而搭建起一个完整的知识体系，而这正是我编写本书的初衷。

本书内容

本书按照结构共分为 13 章，主要内容如下。

第 1~3 章主要介绍数据科学要解决的问题、常用的 IT 工具 Python 以及数据科学所涉及的数学基础。

第 4~7 章主要讨论数据模型，包含 3 部分内容：一是统计中经典的线性回归和逻辑回归模型；二是计算机估算模型参数的随机梯度下降法，这是模型工程实现的基础；三是来自计量经济学的启示，主要涉及特征提取的方法以及模型的稳定性。

第 8~10 章主要讨论算法模型，也就是机器学习领域比较经典的模型，各章依次讨论了监督式学习、生成式模型以及非监督式学习。

目前数据科学的两个前沿领域分别是大数据和人工智能。

第 11 章介绍大数据中很重要的分布式机器学习。

第 12~13 章讨论人工智能领域的神经网络和深度学习。

本书除基础知识外，按照主题亦可分为 3 部分。

第 1 部分主要讨论统计学派的模型和对数据的处理方法，涉及第 4、5、7 章。

第 2 部分主要讨论人工智能学派的方法，涉及第 8、9、10、12、13 章。

第 3 部分主要介绍数据科学的工程实现，涉及第 6、11 章。

在每一章的讨论中，一般会通过一个简单的例子引出模型，然后讲解模型的理论基础，接着展示模型实现的核心代码，最后讨论模型的优缺点以及与其他模型的比较。这样既能很直观地展示模型，也能结合实际代码较深入地讨论它的细节，从而帮助读者更好地掌握和使用模型。

配套代码

针对本书，最好的阅读方法是对照每一章的示例代码，动手实现所讨论的模型。这样会极大加深读者对模型的理解，提高实践能力，否则就会像读小说一样，阅读时感觉不错，但实际使用时就无从下手了。

本书配套代码的下载地址为 https://github.com/GenTang/intro_ds，供读者借鉴和使用。

需要注意的是，为了在正文中节省篇幅，突出重点，本书所展示的代码是基于 Linux 系统下的 Python 2.7，而提供下载的配套代码则是兼容 Python 3 和 Windows 系统的。

本书内容

本书按照结构共分为 13 章，主要内容如下。

第 1~3 章主要介绍数据科学要解决的问题、常用的 IT 工具 Python 以及数据科学所涉及的数学基础。

第 4~7 章主要讨论数据模型，包含 3 部分内容：一是统计中经典的线性回归和逻辑回归模型；二是计算机估算模型参数的随机梯度下降法，这是模型工程实现的基础；三是来自计量经济学的启示，主要涉及特征提取的方法以及模型的稳定性。

第 8~10 章主要讨论算法模型，也就是机器学习领域比较经典的模型，各章依次讨论了监督式学习、生成式模型以及非监督式学习。

目前数据科学的两个前沿领域分别是大数据和人工智能。

第 11 章介绍大数据中很重要的分布式机器学习。

第 12~13 章讨论人工智能领域的神经网络和深度学习。

本书除基础知识外，按照主题亦可分为 3 部分。

第 1 部分主要讨论统计学派的模型和对数据的处理方法，涉及第 4、5、7 章。

第 2 部分主要讨论人工智能学派的方法，涉及第 8、9、10、12、13 章。

第 3 部分主要介绍数据科学的工程实现，涉及第 6、11 章。

在每一章的讨论中，一般会通过一个简单的例子引出模型，然后讲解模型的理论基础，接着展示模型实现的核心代码，最后讨论模型的优缺点以及与其他模型的比较。这样既能很直观地展示模型，也能结合实际代码较深入地讨论它的细节，从而帮助读者更好地掌握和使用模型。

配套代码

针对本书，最好的阅读方法是对照每一章的示例代码，动手实现所讨论的模型。这样会极大加深读者对模型的理解，提高实践能力，否则就会像读小说一样，阅读时感觉不错，但实际使用时就无从下手了。

本书配套代码的下载地址为 https://github.com/GenTang/intro_ds，供读者借鉴和使用。

需要注意的是，为了在正文中节省篇幅，突出重点，本书所展示的代码是基于 Linux 系统下的 Python 2.7，而提供下载的配套代码则是兼容 Python 3 和 Windows 系统的。

说明

本书中部分插图中含有未翻译的英文专有名词，原因如下。

一方面，目前相关的参考文献中没有明确且权威的中文名称与之对应，如强行翻译，难保准确，且易给读者造成误解。另一方面，对于数据科学这门学科，英文名词可能是大家更加熟悉的称呼，翻译为中文后也许会使读者在理解上更加困难。

在此说明，望各位读者理解和支持。

读者反馈

由于作者知识水平有限，书中难免存在纰漏之处，敬请各位读者朋友批评指正。请发送邮件到作者的电子邮箱 tgbaggio@hotmail.com 或本书编辑的电子邮箱 zhangshuang@ptpress.com.cn。

致谢

感谢潘健辉博士，他从行文风格和数学细节上为我提出了很多宝贵的意见。感谢我的太太安恺业女士以及我的父母，他们在本书的编写期间给了我很多鼓励。感谢李国杰院士、林晓东教授、杨卫东教授、张溪梦（Simon Zhang）先生、易欢欢先生、贾真先生、张益军先生、彭耀先生、谢佳女士以及赵甘晶女士为本书提供的帮助。感谢我的初中数学老师吴献女士对我的谆谆教诲。感谢本书的编辑张爽女士为本书的顺利出版所做的工作。需要感谢的人还有很多，限于篇幅，这里就不一一列举了。

资源与支持

本书为异步社区出品的图书，在社区（https://www.epubit.com/）上为您提供以下资源与服务。

配套资源

本书源代码请到异步社区的本书购买页面中下载。

请在异步社区本书页面中点击 配套资源 ，跳转到下载界面，按提示进行操作即可。注意：为保证正常购书用户的权益，会要求您输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免还会存在差错。欢迎您将发现的问题告诉我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区主页 https://www.epubit.com/，搜索到本书页面，点击“提交勘误”，输入勘误信息，最后单击“提交”按钮即可。之后本书的作者和编辑会对您提交的勘误进行审核。确认并接受后，您将获赠异步社区的 100 积分。积分可用于在社区兑换优惠券，以及兑换样书或奖品之用。

扫码关注本书

请扫描下方二维码关注本书，即可在异步社区微信服务号中看到本书和进一步的服务信息。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请发邮件到此邮箱，邮件标题中请注明本书书名。

如果您有兴趣出版图书、录制教学视频，或参与图书翻译、技术审校等工作，可以发邮件，有意出版图书的作者还可以到异步社区在线提交投稿：

www.epubit.com/selfpublish/submission

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，请发邮件联系我们。

如果您在网上发现有针对异步图书的各种形式的盗版行为，包括图书或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的举动是对作者权益的保护，我们也由此才能继续为您带来有价值的内容。

关于异步社区和异步图书

异步社区是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务，社区创办于 2015 年 8 月，提供超过 1000 种图书、近 1000 种电子书，以及众多技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

异步图书是由异步社区编辑团队策划出版的精品 IT 专业图书品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，在封面上印有异步图书的 LOGO。我们的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目 录