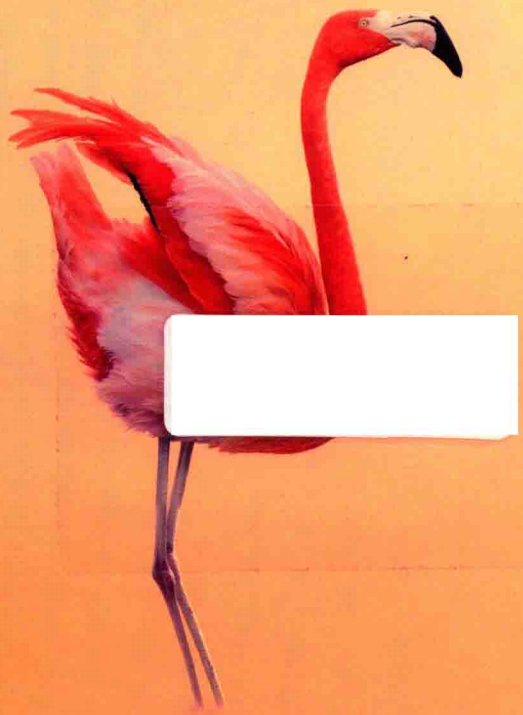
艺术的R语言，秒天秒地秒统计
Get！女神学姐数据科学技能包

Broadview®
www.broadview.com.cn

套路！机器学习

林荟 / 著

北美数据科学家的私房课



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

套路！机器学习

北美数据科学家的私房课

林荟 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

数据科学家目前是北美最热门的职业之一，平均年薪突破 10 万美元。但数据科学并不是一个低门槛的行业，除了对数学、统计、计算机等相关领域的技术要求以外，还要掌握相关应用领域的知识。本书的写作对象是那些现在从事数据分析相关行业，或者之后想从事数据分析行业的人，意在为实践者提供数据科学家这种职业的相关信息。读者可以从阅读中了解到数据科学能解决的问题、数据科学家需要的技能、及背后的“分析哲学”。对于新手而言，一开始就直奔艰深的理论，很容易因为困难而失去兴趣，最终放弃。因此本书倡导的是一种循序渐进的启发性教学路径，着重于数据科学的实际应用，让读者能够重复书中的结果。学习数据分析技能最好的方式是实践！为了平衡理论和应用，书中包括了一些选学小节，用来介绍更多的模型数理背景或给出必要的参考资料来源。本书抽丝剥茧介绍技术内核，帮助大家知其然，同时知其所以然。

希望笔者在北美从事数据科学工作多年踏遍大大小小不计其数的“坑”换来的经验，能够帮助读者更加顺利地成为数据科学家！

图书在版编目（CIP）数据

套路！机器学习：北美数据科学家的私房课 / 林荟著. —北京：电子工业出版社，2017.10
ISBN 978-7-121-32658-5

Ⅰ. ①套… Ⅱ. ①林… Ⅲ. ①数据管理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 221018 号

策划编辑：张月萍
责任编辑：牛　勇
特约编辑：顾慧芳
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×980　1/16　印张：20.75　字数：445 千字
版　　次：2017 年 10 月第 1 版
印　　次：2017 年 10 月第 1 次印刷
印　　数：3000 册　　定价：68.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐序一

伴随着计算机硬件、数据获取和存储技术、分布式算法的飞速发展，以及海量数据的积累，数据科学成为近年来飞速发展的学科。但确切地说，数据科学还不是一门定义完善的学科。直到最近两年，大学里才慢慢开始设立数据科学相关的项目和学位。林荟博士的著作及时地填补了“如何成为成功的数据科学家”领域的空白。由于数据科学家的就业市场非常火热，很多领域的人才都想通过提升自身技术水平和经验成为真正的数据科学家。但正如林博士在书中指出的“数据科学家=数据+科学+艺术家”一样，想成为成功的数据科学家，各个领域的人才需要通过大量的学习和实践来弥补自身的欠缺。比如传统的统计学家和计量经济师需要熟悉编程、数据库操作和大数据分布式计算架构。对于刚刚毕业的理工科硕士和博士，积累利用真实数据解决实际问题的经验，提高书面和口头表达能力，提升团队协作能力和自身的影响力是至关重要的。

林博士的著作首先系统地阐述了什么是数据科学以及成为成功数据科学家的必要条件，然后通过具体的数据和例子来引导读者一步步地理解和学习如何获取这些必要的条件成为真正的数据科学家。本书中各个章节的数据和具体操作都由开源系统的 R 语言来实现。读者可以下载所有的数据和代码，通过自己运行这些代码来加深对每个章节知识的理解，并且可以很快、灵活地学以致用来解决学习和工作中遇到的数据科学相关的项目。对数据科学家而言，很大一部分精力要花在数据的理解、整合和预处理上面。林博士通过自己在数据科学领域多年的经验来仔细讲解如何理解和预处理数据，这是本书的亮点之一。没有很好地理解数据，没有透彻地了解具体要解决的问题，就不可能找到好的解决方法。接着林博士用语言生动诙谐的例子介绍了在数据科学中常见的模型和方法。读者可以通过相关例子和代码来高效理解这些模型和方法，并可以快速地学以致用。虽然几乎所有的算法都有相应的程序包来实现，但作为成功的数据科学家，理解模型的理论背景和基础是必需的。因为只有理解了这些程序包的理论基础，才能有效地对不同数据、不同问题来选择解决的方法并且设置合理的参数。本书对常用模型和方法进行了介绍和引申，可以帮助读者了解

各个模型和方法背后的理论。简言之，本书系统地阐述了如何成为成功的数据科学家，读者可以通过本书的数据和代码，高效学习并能很快应用到实际项目中去。

伴随着大数据应用从互联网科技公司普及到传统商业领域，诸如零售、制造、交通、电力和能源、航空航天、金融、医疗保健，以及大数据在各级政府部门政策制定和实施中的应用，对数据科学家的需求还会逐年增高。尤其是大数据在新兴领域如工业互联网、物联网、智能家居和传感器网络的重要应用，很多相应的数据科学家的职位也会有新的需求。比如在制造业工业物联网领域的数据科学家岗位，除了上面提到的知识和经验，通常还会要求对制造业背后的物理和工程原理有所了解。具备了相应工业的基础知识和原理，数据科学家才能更好地理解数据并建立有效的模型和应用。这也对各理工科背景的人才敞开了数据科学的大门。同时通过大量用户数据的积累，数据科学家也对人文学科的人才敞开了大门。数据科学是一个飞速发展的学科，它通过数据和模型来影响各个学科和领域从而产生价值。数据科学家使得采集的数据有了真正的用武之地。对数据科学感兴趣的人才们，请从本书开始，不断提升自己的技术和经验，真正成为成功的数据科学家，为各行各业带来颠覆性的创新吧！

李明写于美国西雅图，默瑟岛

2017 年 6 月

李明博士，毕业于美国爱荷华州立大学（Iowa State University），拥有物理和统计学背景。曾任通用电气全球研发中心（GE Global Research Center）统计方向负责人（Statistical Leader）、沃尔玛技术部（Walmart Technology）数据科学家（Data Scientist），现任美国亚马逊（Amazon）资深数据科学家（Senior Data Scientist）。李博士还担任美国统计学会（American Statistical Association）质量和生产力分会（Quality and Productivity Section）2017 年度主席，以及统计在物理和工程中的应用年度奖评选委员会主席（SPES Award，美国统计学会年度奖项之一）。李博士的职业生涯中曾涉及金融、零售、制造、电力和能源、交通、医疗保健和航空航天等多个产业及相关跨产业领域。

推荐序二

又来一个找我写序的……感觉自己都快成写序专业户了，惭愧惭愧。以前叫我写序的作者我一般都不熟，但这次这位我还算熟，所以终于可以说点电视上不让播的内容了。八年前林博士和我一同进入爱荷华州立大学（俗称 Ames 村办大学）统计系读博，当时我们的背景完全相反：我在测度论课上奄奄一息，在 R 里如鱼得水，林荟在 R 入门课上死去活来，在理论课上羽化登仙。毫不脸红地吹个牛：要不是我当年的提携，她早就能写出这本书了。

玩笑归玩笑。总的来说，看到这本书时我还是吃了一惊。看来我读博的时候一定是遇到了一个“假”林荟。尽管上学的时候她在村办大学的牲口学院（好吧，兽医学院）有一些科研经历，但我记得也就是画画 ROC 曲线、跑跑逻辑回归而已。士别三年，竟然已经成了一名 R 语言老司机，而且还写出一本主题这么宏大的书。书里举的例子都是种子、生猪、农业论坛，鬼知道她这几年都经历了些什么。以前她抗拒写代码，主要原因是对着电脑时间长了怕脸上长痘，看来后来还是决定为（数据）科学献身了。我们假装感动三秒钟。

书的内容我大致看了一遍，因为都是熟人，我评价起来也就不客套了；按书的内容，分两方面说：R 语言和数据科学。

一般来说，我不在乎别人的 R 代码写得好不好，因为反正写得再好也没我写得好（明年请在我的坟头多烧两张纸）。我对计算机相关书籍的最低标准是不要把“阈值”写成“阀值”，我仔细看过了，本书作者写的是对的。看 R 相关的书籍时，我也有个怪癖，就是找有没有 if (x == TRUE) 或者 y[which(y > 3)]这样的语句，其实语句都没错，只是看看作者的强迫癌是不是到了晚期（if (x)和 y[y > 3]就已足够）。本书作者似乎没有患强迫癌。不过这也无妨，很多时候我觉得对代码吹毛求疵反而影响效率，而且不太老的司机分享的经验对新司机可能更有用。在我眼中，这本书在 R 方面有两个亮点：一是里面介绍了很多 R 的附加包，例如 caret，读者拿起来应该能很快上手；二是几乎以假乱真地模拟数据，这一点可能会为人诟病（不是真实数据），但我觉得模拟数据有其独特的价值，就是你掌控着整个

小宇宙，数据从生成到建模到解释，一路的过程你都可以看清楚，而且可以变着法子变换新数据玩，学习模型使用方法。

数据科学我就不敢妄言了，毕竟我毕业之后已经转向纯码农，很少做有关统计或数据的一线工作。就我的快速粗读来看，我感觉话题的覆盖范围很广，但深度也比较适宜。广度和深度通常只能二选一，也没有优劣之分。我读书少，也限于篇幅，就随意翻两页点评两个例子，从我自己的视角管窥一下本书的价值。比如多年前我就坚信，讲主成分分析的人如果不马上讲偏最小二乘就是瞎掰，尤其是主成分回归，而本书作者很明确地指出了主成分回归的弊病。再比如 Bootstrap 方法，作者讲，“假如你只有一个样本，难道你不停地采用放回抽样的方法就能得到大样本了？”这是很漂亮的一拳。很多方法因为实施简单，所以很容易让人忘了它们的先决条件。我非常反对迷信模型或方法甚至软件，世上没那么多万金油。基于同样的原因，我很欣慰地看到本书不是清一色的 ggplot2 图形（虽然有些图可能长得略丑，但想得美就好了）。

仔细看完本书的话，应该能看出作者是苹果粉（某一页上画图时字体用的是 Songti SC）以及“段子狗”。都读完了博士，选电脑还这么看脸，还整天为各种段子操碎了心，所以这位数据科学家也是蛮拼的。

谢益辉写于奥马哈

作者自序

首先，感谢你翻开这本书！

这是一本什么书？

这是一本关于数据的科学和艺术的书。书中介绍了数据科学这个行业、数据科学家需要的技能，以及“分析哲学”。书中对最常用、最有效的模型进行了展开。数据科学这个行业的本质是通过分析数据解决实际问题，所以本书很看重读者能够真正将书中介绍的知识付诸实践。书中的数据全部都是公开的，书中的代码，建模过程都可以重复。一切不能重复的分析都是瞎掰！

- **为什么写这本书**

当前关于大数据、人工智能的炒作着实令人眼花缭乱，如大数据平台（如 Hadoop、Spark），以及一些黑箱模型（如神经网络，深度学习“实际上就是多层神经网络”）。各路媒体和“砖家”深谙吃瓜群众“不明觉厉”的心态，所以就像个妓院头牌似地越发摆谱。曾经的我也是吃瓜群众中的一员，妥妥地迷失在这信息时代造成的漫天泡沫中，仿佛卡在一扇旋转门里，转了很久不知道去哪。了解一件事情最有效的方法就是实践。很幸运的是，在过去的 4 年里，我主导了大大小小各种分析项目。正是这些实践经验造就了这本书。我并没有打算写一本数据科学的圣经，告诉你关于数据科学的一切。只想尽我所能地给大家还原一个真实的数据科学和数据科学家。希望能为后来者提供一些信息，使得你们能够少走弯路。

- **为什么学习数据科学**

这个问题的答案因人而异。从事某个行业和同某人结婚一样，都有很大的随机性和主观性。所以下面只是我个人喜欢这个行业的理由。

1. 我把数据科学家定义为匠人。个人很享受作为一个匠人，统帅三军之能不如薄技在身。当你相信自己在某些领域有专长并且因此产生自我价值感时，就会有激情。激情是有

吸引力的，就像爱一样，这是一种值得为之奋斗的感觉。

2. 这个世界上的手艺很多，为什么我做的是数据科学？因为我觉得数据科学这门手艺能够帮你培养在当今信息海啸中独善其身的技能——独立思考的能力。用数据进行决策能够让你看问题更清晰，有逻辑，理性客观。这种能力不是只有数据分析师才需要掌握的，理性思考是贯穿很多人一生的必修课，尤其是在互联网时代，通过理性思考甄别过滤信息比之前任何时候都重要。此外，人的大脑是有连贯性的，已经习得某项技能的人，再学另外一项技能的时候，学得会比上一次快一些，因为学习经验在起作用。而若是习得的基础知识是可积累、可扩展的，那么随后可能习得的技能可变现价值就会越来越高。通过数据分析进行决策就是一门可扩展性极高的技能，几乎可以扩展到这个数据时代的方方面面，而且随着社会的数据化趋势，这种可扩展性产生的“复利效应”将越来越大——有着可怕的潜力。

3. 数据科学是美的，美只有爱知道，所以热爱是选择这个行业的主要理由。不知道从什么时候开始，中国互联网上开始流传一句话：生活不止眼前的苟且，还有诗和远方。其实问题不在于缺少诗和远方，而在于你以为眼前的是苟且。如果你热爱自己当前所做的事情，那就是诗，就是远方。如果你不热爱自己所做的事情，在你找到自己真正热爱的事情之前，到哪里都是苟且。我希望阅读这本书的所有人都能够在数据分析中找到乐趣。归根结底，快乐并不是什么深奥的事情，无非是猫吃鱼，狗吃肉，奥特曼打小怪兽。

最后，感谢父母的爱和支持，感谢你们帮助我找到自己热爱的东西。感谢 Scott Iverson，他是我在市场营销领域的导师，没有他，我无法将数据科学很好地应用于市场营销。感谢王正林以及所有为本书出版做出努力的人，没有你们就没有本书的问世。再次感谢你选择本书！

前　言

数据科学家目前是北美最热门的职业之一，平均年薪突破 10 万美元。但数据科学并不是一个低门槛的行业，除了对数学、统计、计算机等相关学科技术的要求以外，还需要相关应用领域的知识。这个职业听起来很酷，但如果你对数据分析没有兴趣的话，你也会觉得这个行业很苦。这里我默认本书的读者都至少是对这个行业有兴趣和激情的。本书的写作对象是那些现在从事数据分析相关行业或者之后想从事数据分析行业的人，意在为实践者提供数据科学家这种职业的相关信息。读者可以从阅读中了解到数据科学家需要的技能，及背后的“分析哲学”。书中会对部分最常用、有效的模型加以展开。关于模型技术部分，我希望读者有初步统计知识，最好知道线性回归。

数据科学家这个行业的本质是应用。市面上有很多文章、出版物介绍各种数据模型，大多数此类书籍并不能让读者重复书中所述的分析过程，对于书中介绍的知识，读者真正实践起来会遇到很多困难。本书着重于数据科学的实际应用，让读者能够重复书中的结果，这也用到了统计软件 R 的自动化报告功能。可能有读者会问，为什么要可重复？根据个人经验，学习数据分析技能最好的方式是实践：动手重复分析的过程，检查分析结果，发现问题后再去查询相关模型的背景技术知识。这一过程得到的学习效果远远超过死磕一本大部头的技术理论书籍，但磕了一年之后发现碰到实际问题不知道该用什么工具实践这些书中讲到的模型方法。而且对于新手而言，一开始就直奔艰深的理论，很容易因为困难而失去兴趣，最终放弃。本书倡导的是一种循序渐进的启发性教学路径，从实际问题入手，抽丝剥茧进入技术内核。

本书主要部分将避免过多的数学公式，但难免有例外。我们在一些地方提到方法背后的技术细节是为了帮助读者理解模型的长处和弱点，而非单纯地介绍数理统计知识。这并不意味着这些数理背景知识不重要，相反尽可能多地了解模型背后的数学很重要且有意义，为了平衡理论和应用，我们会在有的章中加一些选学小节，用来介绍更多的模型数理背景或给出必要的参考资料来源，如果不感兴趣的读者可以跳过这些小节，不会影

响本书主要部分的阅读。书中的每一章都只是冰山一角，我并不试图彻底地介绍模型，而是选择性地解释其中部分我觉得重要的地方。我会尽量将想要强调的概念和内容在分析数据的过程中体现出来，而不仅仅是数学公式符号表达。想要成为数据科学家，仅靠阅读本书是远远不够的，读者需要进一步查阅书中提到的参考资料，或者选修相关课程。

随着计算机科学的发展，不仅收集存储的数据增加了，分析数据的软件包也不断推陈出新，这极大地降低了应用统计学习方法的壁垒。现在不管会建模的不会建模的，大都听过线性回归，这个经典统计模型可追根溯源至 19 世纪 Legendre 和 Gauss 发表的若干关于最小二乘的论文。现在你要通过最小二乘拟合一个线性模型那就是动动手指头两秒钟的事情。可在那个计算器都没有的时代，能优化误差平方和这样的东西的大牛都会被认为是火星人。那个年代美国宪法规定每十年必须进行一次人口普查，1880 年排山倒海的普查资料花了 8 年时间处理分析，一个名叫 Herman Hollerith 的品学兼优的美国少年跳出来，在 1890 年发明了一种排序机，利用打孔卡存储资料，再由机器感测卡片，协助人口调查局对统计资料进行自动化制表，结果不出 3 年就完成了人口普查工作，Herman 同学也顺带用这个发明拿个了工程学博士学位。你可能要问，计算能力这么落后，那这伙数学家捣鼓出来的方法谁用？天文学家用。线性模型最早用在天文学研究中。研究中使用统计方法的，那时绝对是小众边缘群体，全都可以贴上火星制造的标签。然后，盼星星盼月亮我们终于在 1912 年 6 月等到了图灵这个天才的降临，如图 1 所示。

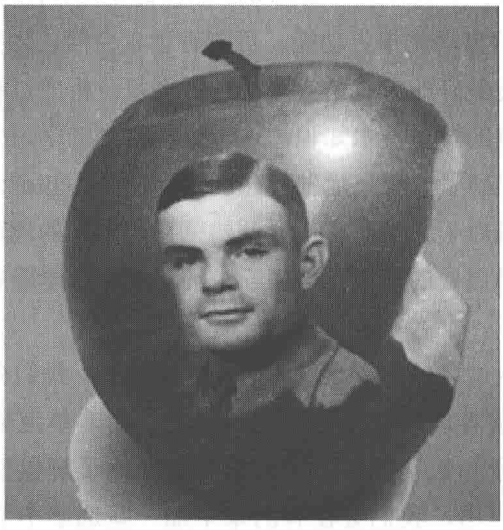


图 1

若不是图灵这个孩子被性取向拖了后腿，数据科学家这个行业早几十年可能就火了。当然，统计泰斗们也没有闲着，Fisher 在 1936 年提出了线性判别分析。在 20 世纪 40 年代，又一家喻户晓的经典统计模型——逻辑回归——问世了！在 20 世纪 70 年代早期，Nelder 和 Wedderburn 发明了广义线性模型这个词，这是一个更大的统计模型框架，它将随机分布函数和系统效应（非随机效应）通过一个连接函数（link function）连起来，之前的线性模型和逻辑回归都是该框架下的特例。到 20 世纪 70 年代末，可以用来分析数据的方法已经有好多了，但这些方法几乎都是线性模型，因为在那时，拟合非线性关系的计算量相对当时的计算机水平来说还是太大了。等到 20 世纪 80 年代，计算机技术终于发展到可以使用非线性模型了。Breiman、Fridman、Olshen 和 Stone 提出了分类回归树。随后的一些机器学习方法进一步丰富了数据科学家可以使用的工具集。计算机软件的飞速发展使得这些方法模型得以应用在更加广泛的领域，应用涵盖了商业、健康、基因、社会心理学研究和政策分析，等等。数据科学家这个行业随着数据量的增加和分析软件的进步不断地向前发展。

关于分析软件，本书使用 R。选择 R 语言的原因如下：

1. R 免费，且可以在不同操作系统上使用。
2. R 开源、可扩展：它在通用公共许可（General Public License）下发行，在此构架下任何人可以检查修改源程序，并且 R 语言含有很多最新的模型。
3. R 有强大图形可视化和自动化报告功能。
4. 笔者 10 年使用 R 的经验证明：无论在学术界还是业界，这都是非常有效的工具。

网上有大量的 R 入门教程，关于用 R 进行数据分析的书也有好多，所以这里就不重复造轮子了，不熟悉 R 语言的读者可以先学习相关资料，这里我假设读者已经有一定的 R 语言基础。

本书布局如下，先介绍数据科学家这个行业的“分析哲学”和数据分析的一般流程。这是非技术的部分，但对于从业者来说非常重要，它帮助你对这个职业设定一个合理的预期。其中会讨论数据科学家需要的技能。之后的章节会对这里提到的部分我觉得重要的技能进一步展开讨论，由于篇幅所限，不可能详细讨论开始这几章中提到的所有技能。随后开始进入技术部分，讲分析环节的第一步——数据预处理，这一步虽然不是正式建模，但却是整个分析过程中最耗时的一个环节。这步没有到位将严重影响模型质量。也正是因为预处理重要，所以单独作为一个章节，没有和其他建模技术合并起来。第 6 章“基础建模技术”介绍的是一些在建模过程中需要的辅助性的技术以及建模需要注意的问题。之后正

式介绍各种笔者在从业过程中经常用到的模型。

本书用来展示模型的数据大部分是通过 R 得到的模拟数据集。为什么用模拟数据而不是真实数据呢？原因如下：

1. 你可以控制数据生成过程，免去了传输下载数据的麻烦。
2. 你可以根据需要改变生成数据的代码，得到新的数据，观察数据变化对模型结果的影响。
3. 对于自己创建的数据，我们知道数据要表达的真实信息，那么就可以评估分析使用的模型的准确性，然后再用于真实数据。
4. 可以通过使用模拟数据在拿到真实数据前准备好代码模板，这样，当你有真实数据时就可以迅速进行分析。
5. 通过重复数据模拟的过程可以加深对模型假设的理解。

同一章后面的代码通常建立在之前代码上，但每章的代码自成系统，也就是说你不需要以其他章节代码运行结果为前提重复某章的代码。有一定 R 语言基础的读者可以通过学习生成数据的代码了解数据的结构以及模型假设。R 语言的新手学习这些代码可能会觉得太困难，没有关系，你们可以跳过生成数据的细节，只需要了解数据的语境，都有哪些变量以及变量类型。你可以直接从网站上读取这些数据。书中的代码和数据可以在这个 GitHub 页面上找到：https://github.com/happyrabbit/DataScientistR。

现在开始我们的旅程吧！

目　录