O'REILLY®

TURING 图灵程序设计丛书

精通特征工程

Feature Engineering for Machine Learning

通过Python示例掌握特征工程基本原则和实际应用，
增强机器学习算法效果

[美] 爱丽丝·郑 阿曼达·卡萨丽 著
陈光欣 译

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

TURING 图灵程序设计丛书

精通特征工程

Feature Engineering for Machine Learning

[美] 爱丽丝·郑　阿曼达·卡萨丽 著
陈光欣 译

Beijing • Boston • Farnham • Sebastopol • Tokyo　O'REILLY®

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北　京

图书在版编目（CIP）数据

精通特征工程 /（美）爱丽丝·郑（Alice Zheng），（美）阿曼达·卡萨丽（Amanda Casari）著；陈光欣译. -- 北京：人民邮电出版社，2019.4
（图灵程序设计丛书）
ISBN 978-7-115-50968-0

Ⅰ. ①精… Ⅱ. ①爱… ②阿… ③陈… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2019)第045692号

内 容 提 要

本书介绍大量特征工程技术，阐明特征工程的基本原则。主要内容包括：机器学习流程中的基本概念，数值型数据的基础特征工程，自然文本的特征工程，词频－逆文档频率，高效的分类变量编码技术，主成分分析，模型堆叠，图像处理，等等。

本书适合机器学习相关从业者和数据科学家阅读。

◆ 著　[美] 爱丽丝·郑　阿曼达·卡萨丽
译　陈光欣
责任编辑　岳新欣
责任印制　周昇亮

◆ 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
三河市祥达印刷包装有限公司印刷

◆ 开本：800×1000　1/16
印张：10.75
字数：254千字　2019年4月第1版
印数：1－3 500册　2019年4月河北第1次印刷
著作权合同登记号　图字：01-2018-8085号

定价：59.00元

读者服务热线：(010)51095186转600　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

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上
Standing on Shoulders of Giants

iTuring.cn

版权声明

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 *Make* 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版、在线服务还是面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列非凡想法（真希望当初我也想到了）建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

前言

简介

为了提取知识和做出预测，机器学习使用数学模型来拟合数据。这些模型将特征作为输入。**特征**就是原始数据某个方面的数值表示。在机器学习流程中，特征是数据和模型之间的纽带。**特征工程**是指从原始数据中提取特征并将其转换为适合机器学习模型的格式。它是机器学习流程中一个极其关键的环节，因为正确的特征可以减轻构建模型的难度，从而使机器学习流程输出更高质量的结果。机器学习从业者有一个共识，那就是建立机器学习流程的绝大部分时间都耗费在特征工程和数据清洗上。然而，尽管特征工程非常重要，专门讨论这个话题的著作却很少。究其原因，可能是正确的特征要视模型和数据的具体情况而定，而模型和数据千差万别，很难从各种项目中归纳出特征工程的实践原则。

然而，特征工程并不只是针对具体项目的行为，它有一些基本原则，而且最好结合具体情境进行解释说明。在本书中，每一章都集中阐述一个数据问题：如何表示文本数据或图像数据，如何为自动生成的特征降低维度，何时以及如何对特征进行标准化，等等。你可以将本书看作内容互有联系的短篇小说集，而不是一部长篇小说。每一章都对大量现有特征工程技术进行了简单介绍，它们综合在一起，阐明了特征工程的基本原则。

掌握一门学科不仅仅是要了解其中的定义以及能够推导公式。仅知道它的工作机制和用途是不够的，你还必须理解它为什么这样设计，它与其他技术有何联系，以及每种方法的优点和缺点。只有清楚地知道事情是如何完成的，对其中的基本原理有直观的理解，并能将知识融会贯通，才称得上精通。尽管一本好书可以让你初窥门径，但只靠读书不能登堂入室，你必须动手实践，将你的想法变成实际的应用，这是一个不断迭代的过程。在每次迭代中，我们都能将想法理解得更加透彻，并逐渐找到更巧妙、更有创造性的实现方法。本书的目的就是帮助你更好地实现想法。

本书力求追本溯源，数学方法倒在其次。我们不仅讨论**如何**去做，还尽力探究**为什么**这样

做。我们的目的是获得想法背后的**直觉**，这样就能知道何时以及如何去应用这些方法。每个人的学习方式各不相同，因此本书使用了大量的文字描述和图片来进行讲解。使用数学公式则是为了精确地表示直觉，并充当本书与其他资料之间的沟通桥梁。

本书的代码示例是用 Python 编写的，使用了大量免费和开源的程序包。NumPy 库提供了数值向量和矩阵操作；Pandas 提供了数据框，这是 Python 数据科学的基础数据结构；scikit-learn 是一个通用的机器学习包，包含了大量的模型和特征转换功能；Matplotlib 和样式库 Seaborn 提供了绘图和可视化支持。你可以在本书的 GitHub 仓库（https://github.com/alicezheng/feature-engineering-book）中找到 .ipynb 格式的示例。

前 4 章的节奏不快，因为要照顾一下那些刚刚接触数据科学和机器学习的读者。第 1 章介绍机器学习流程中的基本概念（数据、模型、特征等）。第 2 章研究数值型数据的基础特征工程：过滤、分箱、缩放、对数变换和幂次变换，以及交互特征。第 3 章开始介绍自然文本的特征工程，并研究词袋、*n*-gram 和短语检测等技术。第 4 章介绍 tf-idf（词频 – 逆文档频率），并将其作为特征缩放的一个例子，说明特征缩放为什么会有效。从第 5 章开始，节奏开始加快，我们要讨论高效的分类变量编码技术，包括特征散列化和分箱计数。第 6 章介绍主成分分析（PCA），此时我们已经深入到机器学习的腹地了。第 7 章将 *k*-均值聚类作为一种特征化技术，说明了模型堆叠这一重要概念。第 8 章专门讲解图像处理，图像数据的特征提取要比文本数据困难得多。我们先介绍两种手动提取特征的技术：SIFT 和 HOG，然后再介绍深度学习这种最新的图像特征提取技术。最后，第 9 章通过一个完整的例子（为一个学术论文数据集创建推荐器）演示几种技术的实际应用。

特征工程是一个非常宽泛的话题，每天都有新的方法被发明出来，尤其是在自动特征学习领域。为了控制本书的篇幅，我们必须做一些取舍。本书不讨论音频数据的傅里叶分析，尽管这是一个与线性代数中的特征分析（第 4 章和第 6 章会介绍）联系得非常紧密的美妙主题。我们也不讨论随机特征，它与傅里叶分析密切相关。通过讲解图像数据的深度学习，本书对特征学习做了简要的介绍，但没有深入介绍目前正在蓬勃发展的大量深度学习模型。还有一些高级研究方法也不在本书讨论范围内，比如随机投影、复杂的文本特征化模型（如词向量和布朗聚类）、隐含空间模型（如隐含狄利克雷分布和矩阵分解）。如果这些名词对你来说根本无所谓，那么恭喜你；如果你的兴趣所在是特征学习的最前沿，那本书可能不适合你。

本书假定你具有机器学习的基础知识，比如知道什么是模型、什么是向量，但我们会通过一个简单的复习使所有读者处于同一起点。如果你具有线性代数、概率分布和最优化方面的经验，那将会很有帮助，但这些经验不是必需的。

排版约定

本书使用了下列排版约定。

- 黑体字
  表示新术语和重点强调的内容。
- 等宽字体（`constant width`）
  表示程序片段，以及正文中出现的程序元素，比如变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 加粗等宽字体（**`constant width bold`**）
  表示应该由用户输入的命令或其他文本。
- 等宽斜体（*`constant width italic`*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

本书还包含一些线性代数公式。我们使用以下惯例：斜体小写字母表示标量（如 a），加粗斜体小写字母表示向量（如 $\boldsymbol{v}$），加粗斜体大写字母表示矩阵（如 $\boldsymbol{U}$）。

该图标表示提示或建议。

该图标表示一般注记。

该图标表示警告或警示。

使用示例代码

本书的附加资料（示例代码、练习题等）可以从 https://github.com/alicezheng/feature-engineering-book 下载。

本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无须联系我们获得许可。比如，用本书的几个代码片段写一个程序就无须获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无须获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN。比如："*Feature Engineering for Maching Learning* by Alice Zheng and Amanda Casari (O'Reilly). Copyright 2018 Alice Zheng and Amanda Casari, 978-1-491-95324-2."

如果你觉得自己对示例代码的使用超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

O'Reilly Safari

Safari（原来叫 Safari Books Online）是一个会员制的培训和参考平台，面向企业、政府、教育从业者和个人。

会员可以访问几千种图书、培训视频、学习路径、互动式教程和精选播放列表，提供这些资源的出版商超过 250 家，包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology，等等。

要获得更多信息，请访问 http://oreilly.com/safari。

联系我们

请把对本书的评价和问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

本书有一个网页，上面列出了勘误[1]、示例和所有附加信息，网页地址为：
http://bit.ly/featureEngineering_for_ML

注 1：本书中文版勘误请到 www.ituring.com.cn/book/2050 查看和提交。——编者注

对于本书的评论和技术性问题，请发送电子邮件到：

bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

http://www.oreilly.com

我们在 Facebook 的地址如下：

http://facebook.com/oreilly

请关注我们的 Twitter 动态：

http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：

http://www.youtube.com/oreillymedia

致谢

首先，最应该感谢的是本书的编辑 Shannon Cutt 和 Jeff Bleiel，他们在本书漫长的出版过程（对此我们并不了解）中给予了两位初出茅庐的作者细致的指导。没有他们的努力工作，本书根本不会面世。还要感谢 O'Reilly 的图书策划编辑 Ben Lorica，是他的鼓励和肯定将本书从一个不切实际的想法变成了现实。感谢 Kristen Brown 和 O'Reilly 的制作团队，感谢他们对细节的极度关注和等待我们回应时的超级耐心。

对于数据科学家来说，出版一本书要比养个孩子付出更多的辛劳。对于所有帮助提高本书质量或使内容更加清晰的建议，我们都非常感激。Andreas Müller、Sethu Raman 和 Antoine Atallah 在百忙之中抽出宝贵的时间，帮我们进行了技术审阅。Antoine 不但神速地提供了反馈，还允许我们使用他强大的计算机来做实验。Ted Dunning 在统计学和应用机器学习方面的技术真是炉火纯青，他还极其慷慨地分享了自己的时间和想法，*k*-均值那一章中的方法和示例就是他贡献出来的。Owen Zhang 公开了他在 Kaggle 竞赛中使用响应速率特征的一些心得，并被 Misha Bilenko 收集进了机器学习的分箱计数方法集中。最后，还要感谢 Alex Ott、Francisco Martin 和 David Garrison 的反馈意见。

Alice的特别感谢

我要感谢 Graph/Data/Turi 团队在本书第一阶段中提供的大力支持。在与用户的交流中，我们产生了写作本书的想法。在为数据科学家建立一个新机器学习平台的过程中，我们发现人们需要更加系统化地理解特征工程。感谢 Carlos Guestrin，他允许我从繁忙的创业工作中抽身，集中精力进行写作。

感谢 Amanda，她先是作为技术审阅人，后来加入到了本书的写作过程中。你是最棒的终结者！本书已经完成，如果还想一边喝着茶和咖啡、吃着三明治和外卖，一边进行写作的话，我们就得找个新项目了。

特别感谢 Daisy Thompson，她是我的医生，也是我的朋友，感谢她在本书写作过程中给予我始终如一的支持。没有你的帮助，我需要更多时间才能下定决心，甚至会半途而废。就像在所有工作中一样，你给我的写作带来了光明与希望。

Amanda的特别感谢

因为只是写了一本书，不是获得了终身成就奖，所以我尽量只感谢那些与本书相关的人。

非常感谢 Alice 让我担任本书的技术编辑和合著者。从你的身上我学到了很多，包括如何写出好笑的数学笑话，以及如何将复杂概念解释清楚。

最后要特别感谢我的丈夫 Matthew，他尽了最大的努力来支持我、鼓励我向目标前进，从不让我蒙混过关。你是我最好的搭档，也是我最喜爱的伙伴。哪怕我只取得了一点点成就，你也以此为荣，并不断激励我。

电子书

扫描如下二维码，即可购买本书电子版。

目录