使用无监督学习建立自动化的预测和分类模型

深度学习精要（基于R语言）

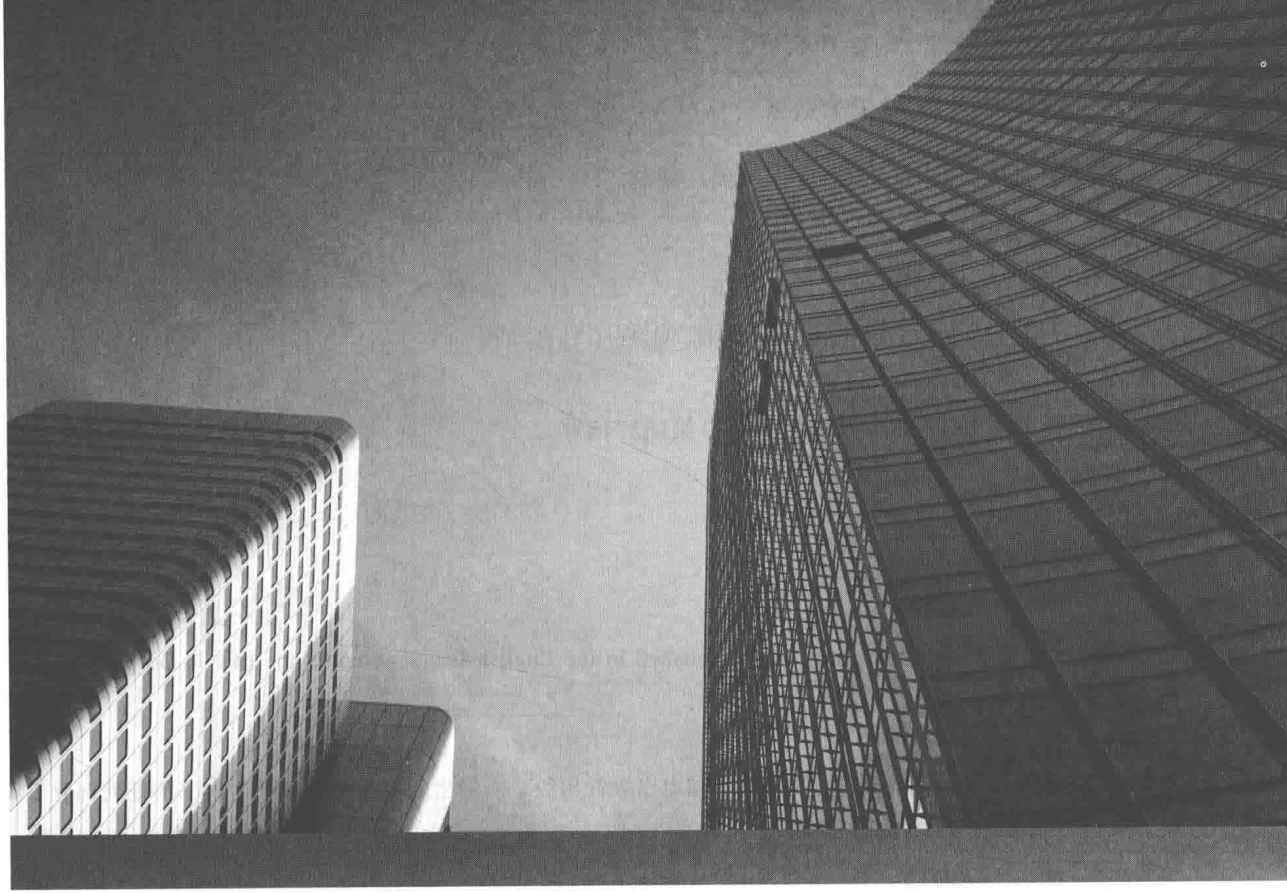
R Deep Learning Essentials

[美] Joshua F. Wiley 著

高蓉 译

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS



深度学习精要（基于R语言）

[美] Joshua F. Wiley 著
高蓉 译

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

深度学习精要 : 基于R语言 / (美) 威利(Joshua F. Wiley) 著 ; 高蓉译. -- 北京 : 人民邮电出版社, 2017.9
ISBN 978-7-115-46415-6

Ⅰ. ①深… Ⅱ. ①威… ②高… Ⅲ. ①程序语言－程序设计 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2017)第183745号

版权声明

◆ 著 [美] Joshua F. Wiley
译 高 蓉
责任编辑 陈冀康
责任印制 焦志炜

◆ 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
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：10.75
字数：207 千字 2017 年 9 月第 1 版
印数：1－2 400 册 2017 年 9 月河北第 1 次印刷

著作权合同登记号 图字：01-2016-7607 号

定价：49.00 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

内容提要

本书重点介绍如何将 R 语言和深度学习模型或深度神经网络结合起来，解决实际的应用需求。全书共 6 章，分别介绍了深度这习基础知识、训练预测模型、如何防止过拟合、识别异常数据、训练深度预测模型以及调节和优化模型等内容。

本书适合了解机器学习概念和R语言并想要使用R提供的包来探索深度学习应用的读者学习参考。

作者简介

Dr. Joshua F. Wiley 是莫纳什大学的讲师，也是统计咨询公司 Elkhart 集团有限公司的资深合伙人。他从位于洛杉矶的加利福尼亚大学获得了博士学位。他的研究集中于使用高级数量方法来理解社会、心理，以及与心理和生理健康有关的生理过程之间的复杂的相互影响。在统计和社会科学方面，Joshua 关注生物统计并且对可重复性研究以及数据和统计模型的图形显示非常有兴趣。通过在 Elkhart 集团有限公司的顾问工作以及他之前在 UCLA 统计顾问集团的工作，Joshua 已经帮助过各种各样的客户，从试验研究者到生物技术公司。他开发或者共同开发了许多 R 包，包括 `varian`，一个用来构建贝叶斯尺度位置结构方程模型的包，以及 `MplusAutomation`，一个将 R 链接到商业软件 `Mplus` 的热门 R 包。

我要感谢我的妻子和家人多年来的支持与鼓励，使我对工作始终抱有热忱。

审阅人简介

Vincenzo Lomonaco，1991 年出生于意大利的圣乔瓦尼-罗通多。他在巴西利卡塔度过了童年时代，在获得了科学学院文凭之后，他搬到了摩德纳。之后不到 3 年，他以优异的成绩从计算机科学专业毕业。由于受到博洛尼亚盛名和研究活动的吸引，他决定在那里开始计算机硕士的学习。2015 年，他以优异成绩毕业，毕业论文是《用于计算机视觉的深度学习：卷积神经网络和分层时间记忆在目标识别任务中的比较》。目前，他是博洛尼亚大学的博士研究生，研究深度学习和生物启发模式识别。

前言

本书主要介绍如何在 R 编程语言和环境当中训练并使用深度学习模型或深度神经网络。本书无意于提供有关深度神经网络的深入的理论覆盖，但它将给你足够的理论背景，帮助你理解深度神经网络的基础、应用以及结果的解释。本书还将提供一些包和函数，用来训练深度神经网络，优化它们的超参数来提升模型的准确度、生成预测或者建立模型的其他应用。为了着手处理现实生活中的例子和应用，本书将提供关于深度学习要领的易于阅读的全面介绍。

本书的内容

第 1 章“深度学习入门”，展示如何创建 R 和 H2O 包并安装在计算机或服务器上，内容涉及所有和深度学习有关的基本概念。

第 2 章“训练预测模型”，涉及如何训练一个浅层无监督的神经网络预测模型。

第 3 章“防止过拟合”，解释了可用于防止模型过拟合数据的不同方法，为了

提升泛化能力，叫作无监督数据上的正则化。

第 4 章“识别异常数据”，涉及识别异常数据，比如欺诈活动或者离群点，和如何执行无监督深度学习。

第 5 章“训练深度预测模型”，展示了如何训练深度神经网络来解决预测或分类问题，比如图像识别。

第 6 章“调节和优化模型”，解释了如何调整模型的调节参数来提升并优化深度学习模型的准确度和性能。

附录即文献包含了本书所有引用的参考书目。

预备知识

使用这本书，你不需要掌握太多的知识。你所需要的软件的主要部分是 R，它是开源的，可以在 Windows、Mac OS 和多种 Linux 上运行。你还需要最新版本的 Java。当你安装好了 R 和 Java，你还需要安装一些 R 包，所有这些 R 包都可以在主流的操作系统上工作。

或许，更具有挑战性的要求是，对于真正的深度学习应用，哪怕是探索非常小的例子，都要求有现代的硬件。在本书中，笔者主要使用的台式机，配置为 2.50 GHz 的 Intel Xeon E5-2670 v2（10 个物理核，20 个逻辑核），32GB 的内存和三星 850 PRO 512GB SSD。你不一定需要一个相同的系统，但是笔者发现在 16GB 内存、双核 i7 处理器的笔记本电脑上运行某些例子是很耗费时间的。

目标读者

本书适合那些有追求的数据科学家，他们熟知机器学习概念和 R，并且正在使

用 R 提供的包来探索深度学习范式。你最好对 R 语言有一个基本的理解而且对统计算法和机器学习技术运用自如，但你并不需要精通深度学习的概念。

排版约定

在本书中，你会发现许多文本样式，它们区分了不同类型的信息。这里是一些有关这些样式的例子，解释了它们的含义。

文本中的代码、数据库表名称、文件夹名称、文件名称、文件扩展名、路径名称、虚拟 URLs、用户输入以及推特用户定位显示如下所示。

“当然，我们无法真正使用 `library()` 函数，除非我们安装了这些包。”

代码块的设置如下所示。

```
## uncomment to install the checkpoint package
## install.packages("checkpoint")
library(checkpoint)

checkpoint("2016-02-20", R.version = "3.2.3")
```

当我们希望把你的注意力吸引到代码块的一个特别部分的时候，我们会将有关的行或项目设置成粗体，如下所示。

```
performance.outsample[,-4]
  Size Maxit Shuffle Accuracy AccuracyLower AccuracyUpper
1   40    60   FALSE     0.93          0.92          0.94
2   20   100   FALSE     0.92          0.91          0.93
3   20   100    TRUE     0.92          0.91          0.93
4   50   100   FALSE     0.91          0.90          0.92
5   50   100   FALSE     0.92          0.91          0.93
```

命令行的输入和输出形式如下所示。

```
h2oiris <- as.h2o(
  droplevels(iris[1:100, ]))
```

新术语或者重要的词汇用粗体显示。

警告或者重要的注释会出现在类似这样的方框中。

提示或技巧会类似这样出现。

读者反馈

我们始终欢迎来自读者的反馈，请让我们知道你对这本书的看法——喜欢哪些内容，不喜欢哪些内容。读者的反馈对我们来说十分重要，这样我们才能出版读者最需要的书。

常规的反馈请通过电子邮件发送到 feedback@packtpub.com，在邮件中请注明书名。

如果你是某方面的专家，有兴趣写书，或者为某一本书投稿，请阅读我们的作者指南，地址是 www.packtpub.com/authors。

客户支持

现在你已经是一本 Packt 图书的光荣的拥有者，为了让你的付出得到最大的回报，我们还将为你提供其他方面的服务。

下载示例代码

你可以登录 http://www.packtpub.com 的账户，下载本书的示例代码文件。如果你是从其他地方购买的本书，可以访问 http://www.packtpub.com/support，注册之后，我们会为你发送一封附有文件的电子邮件。

你可以根据下面的步骤下载代码文件：

（1）使用你的电子邮件地址和密码在我们的网站登录或者注册；

（2）将鼠标指针悬停在顶部的 SUPPORT 选项卡上；

（3）单击 Code Downloads & Errata；

（4）在搜索框中输入该书的名称；

（5）选择你要找的书来下载代码文件；

（6）在下拉菜单中选择你从何处购买了这本书；

（7）单击 Code Download（代码下载）。

如果下载了文件，请你确保使用下列软件的最新版本来解压缩或者提取文件夹。

- Windows：WinRAR / 7-Zip
- Mac：Zipeg / iZip / UnRarX
- Linux：7-Zip / PeaZip

下载本书的彩色图像

我们还为你提供了本书中所用的屏幕截图/示意图彩色图像的 PDF 文件。彩色图像能更好地帮助你理解输出中的变化。你可以从 https://www.packtpub.com/sites/default/files/downloads/RDeepLearningEssentials_ColorImages.pdf. 下载这些文件。

勘误

尽管我们会尽全力确保书中内容的准确性，但是错误仍然在所难免。如果你在我们的某本书中发现了错误——文字错误或者代码错误，而且愿意为我们报告这些错误，我们将感激不尽。这样不仅可以消除其他读者的挫败感，而且能帮助我们改进这本书的后续版本。如果你发现了任何的错误，可以访问 http://www.packtpub.com/submit-errata 来提交，选择你的书，单击 Errata Submission Form（勘误表提交表单），输入勘误详情。勘误通过验证之后，你提交的内容会被接受而且勘误会上传到我们的网站，或者添加到这本书勘误部分现有的勘误列表中。

如果你想查看之前提交的勘误，请访问 https://www.packtpub.com/books/content/support，在搜索栏中输入书名，所查询的信息会出现在勘误部分。

盗版举报

对所有的媒体来说，在互联网上剽窃版权材料都是一个棘手的问题。Packt 很重视保护我们的版权和许可。如果你在互联网上发现我们产品的任何形式的非法复制品，请立即告知我们的网址和网站名称，这样我们可以采取补救措施。

如果你发现可疑的盗版材料，请通过 copyright@packtpub.com 联系我们。

你的举报有助于保护作者的权益以及我们为你提供有价值内容的能力，我们对此深表感谢。

问题解答

如果你对本书的任何方面有疑问，可以通过 questions@packtpub.com 联系我们，我们会尽力解决问题。

目录

第 1 章
深度学习入门

本章讨论深度学习，这是一种强大的多层架构，可以用于模式识别、信号检测以及分类或预测等多个领域。深度学习并不新鲜，但在过去十年它获得了极高的关注，这部分归功于计算能力的不断发展和训练模型不断涌现出更有效的新方法，也源于可使用的数据量不断增加。在本章中，我们将学习深度学习是什么，训练这种模型有哪些 R 包，如何建立分析系统以及如何连接 R 和 H2O。在随后的章节，我们会将 H2O 用于许多案例，这些案例探讨如何真正训练和使用一个深度学习模型。

本章包括以下内容。

- 什么是深度学习？
- 使用 R 包来训练深度学习模型，如深度信念网络或深度神经网络。
- 连接 R 和 H2O，深度学习使用 H2O。

1.1 什么是深度学习

为了理解深度学习是什么，最简单的方式也许是首先理解常规机器学习是什么。一般来说，机器学习主要用于开发和使用那些从原始数据中学习、总结出来的

用于进行预测的算法。预测是个非常笼统的术语。例如，机器学习中的预测可以包括预测某位消费者将会在一家给定的公司花费是多少，或者预测一笔特殊的信用卡消费中是否存在欺诈。预测也包括更一般的模式识别，如给定的图片显示了什么字母，或者这张照片中是否有马、狗、人、脸、建筑等。深度学习是机器学习的一个分支，其中的深度（多层）架构用于映射输入或观测特征与输出之间的联系。这种深度架构使得深度学习特别适合处理含有大量变量的问题，同时可以把深度学习生成的特征当作学习算法整体的一部分，而不是把特征生成当作一个单独步骤。现已证明，深度学习在图像识别（包括笔迹以及图片或者物体的识别）和自然语言处理（如语音识别）领域非常有效。

现在已有许多类型的机器学习算法。在本书中，我们主要讨论神经网络，因为它在深度学习中非常流行。但是，这种侧重并不意味着这就是用于机器学习甚至深度学习的唯一技术，也不是说其他的技术没有价值或者不适合，技术的选择取决于具体的任务。我们将在 1.2 节从概念上更深入地讨论神经网络和深度神经网络是什么。

1.2 神经网络的概念综述

神经网络正如其名所示，命名的灵感源于身体中的神经过程和神经元。神经网络包括一系列的神经元，或者叫作节点，它们彼此连结并处理输入。神经元之间的连结经过加权处理，权重取决于从数据中学习、总结出的使用函数。一组神经元的激活和权重（从数据中自适应地学习）可以提供给其他的神经元，其中一些最终神经元的激活就是预测。

为了将这个过程刻画得更具体，我们借助于一个来自人类视觉感知的例子来加强理解。祖母细胞这个术语用于指这样一个概念，在大脑的某个地方有一个细胞或者神经元，它专门只对某个复杂的特定对象有反应，如我们的祖母。这种特性需要数千个细胞来代表我们遇到的每个独特实体或对象。相反的观点是，人们通过汇集更多的基本片断来建立复杂的表达方式从而形成了视觉感知。图 1-1 是一张正方形