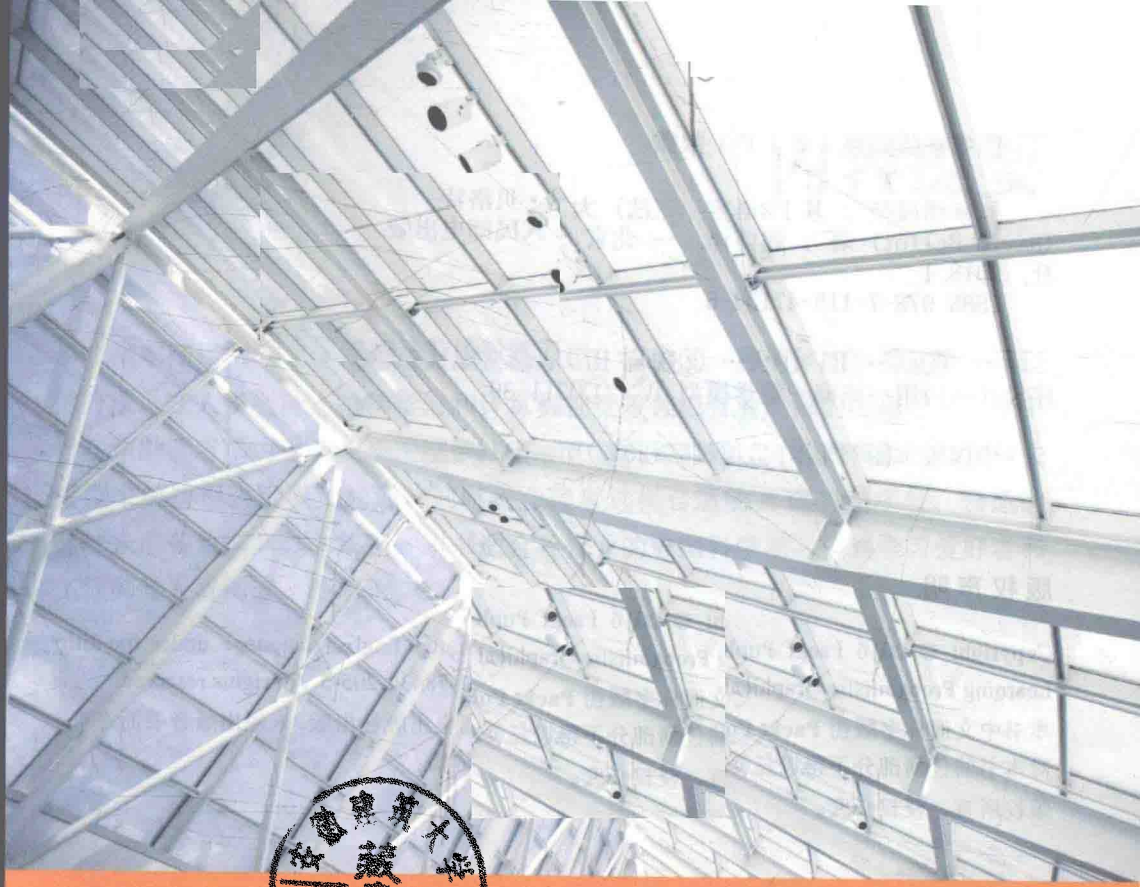
通过现实世界的问题和R语言示例代码，掌握概率图模型

概率图模型
基于R语言

Learning Probabilistic Graphical Models in R

[法] 大卫·贝洛特（David Bellot）著　　魏博 译

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



概率图模型
基于R语言

Learning Probabilistic Graphical Models in R

[法] 大卫·贝洛特（David Bellot）著　　魏博 译

人民邮电出版社
北京

图书在版编目（CIP）数据

概率图模型 : 基于R语言 / (法) 大卫·贝洛特(David Bellot) 著 ; 魏博译. -- 北京 : 人民邮电出版社, 2018.1
ISBN 978-7-115-47134-5

Ⅰ. ①概… Ⅱ. ①大… ②魏… Ⅲ. ①程序语言－程序设计－应用－概率－数学模型 Ⅳ. ①O211-39

中国版本图书馆CIP数据核字(2017)第275114号

版权声明

◆ 著　[法] David Bellot
译　魏　博
责任编辑　王峰松
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
北京瑞禾彩色印刷有限公司印刷

◆ 开本：720×960　1/16
印张：12.75
字数：205 千字　2018 年 1 月第 1 版
印数：1－3 000 册　2018 年 1 月北京第 1 次印刷

著作权合同登记号　图字：01-2017-3665 号

定价：59.00 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

概率图模型结合了概率论与图论的知识，提供了一种简单的可视化概率模型的方法，在人工智能、机器学习和计算机视觉等领域有着广阔的应用前景。本书旨在帮助读者学习使用概率图模型，理解计算机如何通过贝叶斯模型和马尔科夫模型来解决现实世界的问题，同时教会读者选择合适的 R 语言程序包、合适的算法来准备数据并建立模型。本书适合各行业的数据科学家、机器学习爱好者和工程师等人群阅读、使用。

作者简介

David Bellot 是法国国家信息与自动化研究所（INRIA）计算机科学专业的博士，致力于贝叶斯机器学习。他也是美国加州大学伯克利分校的博士后，为英特尔、Orange 电信和巴克莱银行等公司工作过。他现在财经行业工作，使用机器学习技术开发财经市场的预测算法，同时也是开源项目，如 Boost C++ 库的贡献者。

译者简介

魏博，志诺维思（北京）基因科技有限公司高级算法工程师。本科毕业于武汉大学数学系，博士毕业于中国科学院数学与系统科学研究院计算机软件与理论专业。前阿里巴巴优酷事业部视频搜索算法专家，欧普拉软件科技（北京）有限公司新闻推荐高级算法工程师。长期关注于用户需求建模、行为建模和自动推理。数据挖掘、机器学习和数据可视化爱好者，尤其热衷于海量数据中用户视角和用户行为模式的刻画和推断，以及自然语言处理问题。

审稿者简介

Mzabalazo Z. Ngwenya 拥有开普敦大学数学统计专业的研究生学历。他在统计咨询行业有广泛的业务，并有大量的 R 开发经验。他的兴趣主要在统计计算方面。他之前审阅了 Packt 出版社的书籍 *R Studio for R Statistical Computing*（Mark P.J. van der Loo 和 Edwin de Jonge）；*R Statistical Application Development by Example Beginner's Guide*（Prabhanjan Narayanachar Tattar）；*Machine Learning with R*（Brett Lantz）；*R Graph Essentials*（David Alexandra Lillis）；*R Object-oriented Programming*（Kelly Black）；*Mastering Scientific Computing with R*（Paul Gerrard 和 Radia Johnson）；*Mastering Data Analysis with R*（Gergely Daróczi）。

Prabhanjan Tattar 现在是 Fractal Analytics 公司的高级数据科学家。他拥有 8 年的统计分析经验。生存分析和统计推断是他的主要科研和兴趣方向。他在同行评审的期刊上发表了多篇科研论文，并撰写了两本 R 语言书籍：《R 语言统计应用开发实例》（*R Statistical Application Development by Example*，Packt Publishing）和《R 语言统计教程》（*A Course in Statistics with R*, Wiley）。R 程序包 gpk、RSADBE 和 ACSWR 也是由他维护的。

前言

概率图模型是机器学习领域表示现实世界带有概率信息的数据和模型的最先进技术之一。在许多场景中，概率图模型使用贝叶斯方法来描述算法，以便可以从带有噪声和不确定性的现实世界中得出结论。

本书介绍了一些相关话题，例如推断（自动推理和学习），可以自动从原始数据中构建模型。它解释了算法是如何逐步运行的，并使用诸多示例展示了即时可用的 R 语言解决方案。介绍完概率和贝叶斯公式的基本原理之后，本书给出了概率图模型（Probabilistic Graphical Models，PGM），以及几种类型的推断和学习算法。读者会从算法设计过渡到模型的自动拟合。

本书关注在解决数据科学问题上有成功案例的有用模型，例如贝叶斯分类器、混合模型、贝叶斯线性回归，以及用于构建复杂模型的基本模型组件。

主要内容

第 1 章，概率推理，介绍了概率论和概率图模型的基本概念，并通过贝叶斯公式的表示，为概率模型提供一种易用、高效、简单的建模方法。

第 2 章，精确推断，介绍了如何通过简单图形的组合和模型查询构建概率图模型。该查询使用一种叫作联结树算法的精确推断算法。

第 3 章，学习参数，包括从数据集中使用最大似然法，拟合和学习概率图模型。

第 4 章，贝叶斯建模——基础模型，介绍了简单而强大的贝叶斯模型，其可以作为更加复杂模型的基础模块，以及如何使用自适应算法来拟合和查询贝叶斯模型。

第 5 章，近似推断，介绍了概率图模型上的第二种推断方法，同时介绍了主要的采样算法，例如马尔科夫链蒙特卡洛（MCMC）。

第 6 章，贝叶斯建模——线性模型，介绍了更高级贝叶斯视角的标准线性回

归算法，并给出了解决过拟合问题的解决方案。

第 7 章，概率混合模型，介绍了更加复杂的概率模型，其中的数据来自于几种简单模型的混合。

附录，介绍了本书所引用的所有书籍和文献。

环境准备

本书的所有例子都需要版本在 3.0 以上的 R 环境中运行。

本书受众

本书面向需要处理海量数据，并从中得出结论的读者，尤其是当数据有噪声或者存在不确定性的读者。数据科学家、机器学习爱好者、工程师和其他对机器学习最新技术感兴趣的人会觉得概率图模型很有意思。

读者反馈

欢迎读者反馈。让我们知道你关于这本书的想法——喜欢什么，不喜欢什么。读者反馈对于我们很重要，它可以帮助我们开发读者真正需要的话题。想给我们发送反馈，只需要发电子邮件至 feedback@packtpub.com，并在邮件主题中告知书名。如果你是某个话题的专家，并且有兴趣编写书籍或者给予贡献，请查看我们的作者指导：www.packtpub.com/authors。

客户支持

你现在已经是 Packt 书籍的荣誉所有者。你还拥有以下权利。

下载示例代码

你可以从 http://www.packtpub.com 的个人账户中下载本书的示例代码文件。

如果你是从别的地方购买的本书，你可以访问 http://www.packtpub.com/support，在此网站注册后，会直接发邮件给你代码文件。你可以通过下列步骤下载代码文件：

1. 使用你的邮箱和密码在我们的网站登录并注册；
2. 在顶部的 SUPPORT 标签上悬停光标；
3. 单击 Code Downloads & Errata ；
4. 在 Search 框中输入书名；
5. 选取代码文件所在的书籍；
6. 选择购书途径的下拉菜单；
7. 单击 Code Download。

你也可以在本书网站的页面单击 Code Files 按钮下载代码文件。这本书的网页可以通过 Search 搜索框输入书名找到。你需要登录自己的 Packt 账户。

文件下载完成之后，确保使用下列软件的最新版解压或抽取文件：

- Windows 系统使用 WinRAR / 7-Zip。
- Mac 系统使用 Zipeg / iZip / UnRarX。
- Linux 系统使用 7-Zip / PeaZip。

勘误

尽管我们已经非常细心地保证内容的正确性，但是错误还是会发生。如果你在我们的书中找到一处错误并告诉我们，不管是文本错误或是代码错误，我们都会非常感激。你的善举会省去其他用户的烦恼，并帮助我们改进本书的后续版本。如果你找到了任何勘误，请访问 https://www.packtpub.com/submit-errata 报告给我们。你只须选取书名，单击 Errata Submission Form 链接，输入勘误的具体信息。一旦勘误确定之后，我们会接受你的提交。勘误会上传到我们的网站，或者添加到书籍勘误部分已有的勘误列表下。要查看以前提交的勘误，访问 https://www.packtpub.com/books/content/support，在搜索框输入书名、所需信息会出现在 Errata 部分下。

版权

互联网上版权资料的盗版问题一直是所有媒介无法避免的问题。在 Packt，我们一直严肃对待版权和许可的保护问题。如果你在互联网上遇到任何形式的我社出版物的非法副本，请立即把具体地址或者网站名称提供给我们，我们可以采取补救措施。请联系 copyright@packtpub.com，附上可疑的盗版材料的链接。我们非常感谢你在保护作者方面的努力，也会注重提升自我能力，给你带来更有价值的内容。

疑问

如果你对本书有任何疑问，可以联系我们 questions@packtpub.com。我们会尽全力解决您的问题。

目 录

第 1 章
概率推理

在有关 21 世纪的所有预测中，最不希望的一个也许是我们需要每天收集世界上任何地方、关于任何事情的海量数据。近几年来，人们见证了关于世界、生活和技术方面难以置信的数据爆炸，这也是我们确信引发变革的源动力。虽然我们生活在信息时代，但是仅仅收集数据而不发掘价值和抽取知识是没有任何意义的。

在 20 世纪开始的时候，随着统计学的诞生，世界都在收集数据和生成统计。那个时候，唯一可靠的工具是铅笔和纸张，当然还有观察者的眼睛和耳朵。虽然在 19 世纪取得了长足的发展，但是科学观察依然处在新生阶段。

100 多年后，我们有了计算机、电子感应器以及大规模数据存储。我们不但可以持续地保存物理世界的数据，还可以通过社交网络、因特网和移动电话保存我们的生活数据。而且，存储技术水准的极大提高也使得以很小的容量存储月度数据成为可能，甚至可以将其放进手掌中。

但是存储数据不是获取知识。存储数据只是把数据放在某个地方以便后用。同样，随着存储容量的快速演化，现代计算机的容量甚至在以难以置信的速度提升。在读博士期间，我记得当我收到一个崭新、耀眼的全功能 PC 来开展科研工作时，我在试验室是多么的骄傲。而今天，我口袋里老旧的智能手机，还要比当时的 PC 快 20 倍。

在本书中，你会学到把数据转化为知识的最先进的技术之一：**机器学习**。这项技术用在当今生活的方方面面，从搜索引擎到股市预测，从语音识别到自动驾驶。而且，机器学习还用在了人们深信不疑的领域，从产品链的质量保障到移动手机网络的天线阵列优化。

机器学习是计算机科学、概率论和统计学相互融合的领域。机器学习的核心

问题是推断问题或者说是如何使用数据和例子生成知识或预测。这也给我们带来了机器学习的两个基础问题：从大量数据中抽取模以及高层级知识的算法设计，和使用这些知识的算法设计——或者说得更科学一些：学习和推断。

皮埃尔 - 西蒙·拉普拉斯（Pierre-Simon Laplace，1749—1827），法国数学家，也是有史以来最伟大的科学家之一，被认为是第一批理解数据收集重要性的人：他发现了数据不可靠，有不确定性，也就是今天说的有噪声。他也是第一个研究使用概率来处理不确定性等问题，并表示事件或信息信念度的人。

在他的论文《概率的哲学》（*Essai philosophique sur les probabilités*，1814）中，拉普拉斯给出了最初的支持新老数据推理的数学系统，其中的用户信念会在新数据可用的时候得到更新和改进。今天我们称之为贝叶斯推理。事实上，托马斯·贝叶斯确实是第一个、早在 18 世纪末就发现这个定理的人。如果没有贝叶斯工作的铺垫，皮埃尔 - 西蒙·拉普拉斯就需要重新发现同一个定理，并形成贝叶斯理论的现代形式。有意思的是，拉普拉斯最终发现了贝叶斯过世之后发表的文章，并承认了贝叶斯是第一个描述归纳推理系统原理的人。今天，我们会提及拉普拉斯推理，而不是贝叶斯推理，并称之为贝叶斯 - 普莱斯 - 拉普拉斯定理（Bayes-Price-Laplace Theorem）。

一个多世纪以后，这项数学技术多亏了计算概率论的新发现而得到重生，并诞生了机器学习中一个最重要、最常用的技术：概率图模型。

从此刻开始，我们需要记住，概率图模型中的术语**图**指的是图论，也就是带有边和点的数学对象，而不是图片或者图画。众所周知，当你想给别人解释不同对象或者实体之间的关系时，你需要拿纸画出带有连线或箭头的方框。这是一种简明易懂的方法，可以来介绍任何不同元素之间的关系。

确切地说，**概率图模型**（Probabilistic Graphical Models，PGM）是指：你想描述不同变量之间的关系，但是，你又对这些变量不太确定，只有一定程度的相信或者一些不确定的知识。现在我们知道，概率是表示和处理不确定性的严密的数学方法。

概率图模型是使用概率来表示关于事实和事件的信念和不确定知识的一种工具。它也是现在最先进的机器学习技术之一，并有很多行业成功的案例。

概率图模型可以处理关于世界的不完整的知识，因为我们的知识总是有限的。我们不可能观察到所有的事情，不可能用一台计算机表示整个宇宙。和计算机相比，我们作为人类从根本上是受限的。有了概率图模型，我们可以构建简单的学习算法，或者复杂的专家系统。有了新的数据，我们可以改进这些模型，尽全力优化模型，也可以对未知的局势和事件做出推断或预测。

在第 1 章中，你会学到关于概率图模型的基础知识，也就是概率知识和简单的计算规则。我们会提供一个概率图模型的能力概览，以及相关的 R 程序包。这些程序包都很成功，我们只需要探讨最重要的 R 程序包。

我们会看到如何一步一步地开发简单模型，就像方块游戏一样，以及如何把这行模型连接在一起开发出更加复杂的专家系统。我们会介绍下列概念和应用。每一部分都包含几个可以直接用 R 语言上手的示例：

- 机器学习。
- 使用概率表示不确定性。
- 概率专家系统的思想。
- 使用图来表示知识。
- 概率图模型。
- 示例和应用。

1.1 机器学习

本书是关于机器学习领域的书籍，或者更广义地叫作人工智能。为了完成任务，或者从数据中得出结论，计算机以及其他生物需要观察和处理自然世界的各种信息。从长期来看，我们一直在设计和发明各种算法和系统，来非常精准地并以非凡的速度解决问题。但是所有的算法都受限于所面向的具体任务本身。另外，一般生物和人类（以及许多其他动物）展现了在通过经验、错误和对世界的观察等方式取得适应和进化方面令人不可思议的能力。

试图理解如何从经验中学习，并适应变化的环境一直是科学界的伟大课题。自从计算机发明之后，一个主要的目标是在机器上重复生成这些技能。

机器学习是关于从数据和观察中学习和适应的算法研究，并实现推理和借