释放贝叶斯框架的灵活性与力量

Python贝叶斯分析

[阿根廷] 奥斯瓦尔多·马丁（Osvaldo Martin） 著
田俊 译

Bayesian Analysis with Python

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

Bayesian Analysis with Python

Python贝叶斯分析

[阿根廷] 奥斯瓦尔多·马丁（Osvaldo Martin） 著
田俊 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Python贝叶斯分析 / (阿根廷) 奥斯瓦尔多·马丁(Osvaldo Martin) 著 ; 田俊译. -- 北京 : 人民邮电出版社, 2018.2
ISBN 978-7-115-47617-3

Ⅰ. ①P… Ⅱ. ①奥… ②田… Ⅲ. ①软件工具－程序设计②贝叶斯理论 Ⅳ. ①TP311.561②O225

中国版本图书馆CIP数据核字(2017)第324763号

版权声明

◆ 著 [阿根廷]奥斯瓦尔多·马丁（Osvaldo Martin）
译 田 俊
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
北京捷迅佳彩印刷有限公司印刷

◆ 开本：720×960 1/16
印张：14.75
字数：241 千字 2018 年 2 月第 1 版
印数：1－3 000 册 2018 年 2 月北京第 1 次印刷

著作权合同登记号 图字：01-2017-3662 号

定价：69.00 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

本书从务实和编程的角度讲解了贝叶斯统计中的主要概念，并介绍了如何使用流行的 PyMC3 来构建概率模型。阅读本书，读者将掌握实现、检查和扩展贝叶斯统计模型，从而提升解决一系列数据分析问题的能力。本书不要求读者有任何统计学方面的基础，但需要读者有使用 Python 编程方面的经验。

作译者简介

作者简介

Osvaldo Martin 是阿根廷国家科学与技术理事会（CONICET）的一名研究员。该理事会是负责阿根廷科技进步的主要组织。Osvaldo Martin 曾从事结构生物信息学和计算生物学方面的研究，特别是在如何验证结构蛋白质的模型方面有着深入研究；此外，在应用马尔科夫—蒙特卡洛方法模拟分子方向上有着丰富的经验，尤其喜欢用 Python 解决数据分析问题。他曾讲授结构生物信息学、Python 编程等课程，最近还开设了贝叶斯数据分析的课程。Python 和贝叶斯统计改变了他对科学的认知和对问题的思考方式。他写本书的最大动力是希望借助 Python 帮助大家理解概率模型，同时，他也是 PyMOL 社区（一个基于 C/Python 的分子可视化社区）的活跃成员，最近他也对 PyMC3 社区做了一些贡献。

我要感谢我的妻子 Romina 在我写作本书的过程中对我的支持，以及对我所有“疯狂”项目的支持。我还要感谢 Walter Lapadula、Juan Manuel Alonso 和 Romina Torres-Astorga，他们为本书提供了宝贵的反馈和建议。

此外，我还要特别感谢 PyMC3 的核心开发者们，本书正是因为他们的辛勤奉献才得以成为可能，希望本书能够促进 PyMC3 的传播和使用。

技术审校者简介

Austin Rochford 是 Monetate Labs 的首席数据科学家，他开发的产品用于帮助零售商在每年数十亿的活动中进行个性化销售。同时他还是一位积极倡导贝叶斯方法的数学家。

译者简介

田俊，计算机专业硕士。2016 年毕业于中国科学院自动化研究所，主要研究方向为自然语言处理中的短文本分类，毕业后曾在滴滴出行担任算法工程师，目前在微软从事自然语言处理方面的工作。

中文版审校者简介

劳俊鹏，心理学博士，PyMC 团队成员。2014 年毕业于英国格拉斯哥大学，主要研究认知神经心理学。2013 年至今在瑞士弗里堡大学从事心理学研究，专攻数据建模分析和神经计算模型。

前言

贝叶斯统计距今已经有超过250年的历史，其间该方法既饱受赞誉又备受轻视。直到近几十年，得益于理论进步和计算能力的提升，贝叶斯统计才越来越多地受到来自统计学以及其他学科乃至学术圈以外工业界的重视。现代的贝叶斯统计主要是计算统计学，人们对模型的灵活性、透明性以及统计分析结果的可解释性的追求最终造就了该趋势。

本书将从实用的角度来学习贝叶斯统计，不会过多地考虑统计学范例及其与贝叶斯统计之间的关系。本书的目的是借助Python做贝叶斯数据分析，尽管与之相关的哲学讨论也很有趣，不过受限于篇幅，这部分内容并不在本书的讨论范围之内，有兴趣的读者可以通过其他方式深入了解。

这里我们采用建模的方式学习统计学，学习如何从概率模型的角度思考问题，结合模型和数据利用贝叶斯理论推导出逻辑结论。这种建模方式使用的是一种数值计算的方法，其中，模型部分会由基于PyMC3的代码构建。PyMC3是用于贝叶斯统计的Python库，它为用户封装了大量的数学细节和计算过程。贝叶斯方法在理论上源于概率论，这也是为什么许多讲贝叶斯方法的书中都充斥着需要一定数学基础的公式。学习统计学方面的数学知识显然有利于构建更好的模型，同时还能让你对问题、模型和结果有更好的直觉。不过类似PyMC3的库能够帮助你在有限的数学知识水平下学习并掌握贝叶斯统计。在阅读本书的过程中，你将亲自见证这一过程。

本书结构

第1章，**概率思维——贝叶斯推断指南**，介绍了贝叶斯理论及其在数据分析中的意义，并进一步阐述了贝叶斯思维方式的定义以及为什么和如何使用概率来处理不确定性。本章还包含本书其余章节中的一些基本概念。

第2章，**概率编程——PyMC3编程指南**，从计算的角度重新回顾了前一章

提到的概念。这一章中我们将引入 PyMC3，并学习如何用它来构建概率模型、对后验进行采样、判断采样是否正确以及分析和解释贝叶斯结果。

第 3 章，**多参和分层模型**，介绍了贝叶斯模型中最基础的内容，并在此基础上加入了一些更复杂的内容。我们将学习如何利用多个参数构建并分析模型，以及如何利用分层模型的优势往模型中添加结构。

第 4 章，**利用线性回归模型理解并预测数据**，介绍了线性回归模型的广泛应用以及如何将其应用于更复杂的模型。在本章中，我们将学习如何利用线性模型解决回归问题，以及如何处理异常值和多变量的问题。

第 5 章，**利用逻辑回归对结果进行分类**，在前一章的基础上对线性模型做了进一步推广，将其应用于解决多输入 / 多输出分类问题。

第 6 章，**模型比较**，讨论了统计和机器学习中一些常见的模型比较难点。我们将学习一些信息测准和贝叶斯因子方面的理论知识，以及如何将其应用于模型比较。

第 7 章，**混合模型**，讨论了如何将简单的模型混合在一起构建出更复杂的模型，这种方法将引导我们认识新的模型，并从混合模型的角度重新回顾前面几章中学到的模型。此外，本章还讨论了如何进行数据聚类和如何处理计数类型数据等问题。

第 8 章，**高斯过程**，简要讨论了一些非参数统计方面的高级概念作为本书的结束，包括什么是核函数、如何使用线性核回归以及如何将高斯过程用于回归问题。

准备工作

本书代码部分使用的是 Python 3.5 以上版本，因此，建议你使用 Python 3 的最新版，尽管本书的大部分代码都能在更早的版本上运行（包括 Python 2.7，不过可能需要稍微修改）。

安装 Python 和 Python 库最简单的方法是使用 Anaconda（一个用于科学计算的软件），你可以通过网站 https://www.continuum.io/downloads 了解和下载 Anaconda。安装好 Anaconda 之后，可以使用 `conda install` 库的名称来安装 Python 库。

本书会用到以下 Python 库：

- Ipython 5.0；
- NumPy 1.11.1；
- SciPy 0.18.1；
- Pandas 0.18.1；
- Matplotlib 1.5.3；
- Seaborn 0.7.1；
- PyMC3 3.0。

读者对象

本书的阅读对象为不熟悉贝叶斯统计方法，同时又希望学习如何进行贝叶斯数据分析的本科生、研究生或数据科学家。本书不要求读者有统计学或贝叶斯分析方面的背景，书中尽可能地减少了数学公式的使用，只在我们认为有利于读者理解相关概念的地方用到。此外，所有的概念都通过代码、图表以及文字进行了详细描述。本书假设读者知道如何使用 Python 进行编程，最好熟悉 NumPy、Matplotlib 或者 Pandas。

惯例

在阅读本书过程中，你会看到一些不同的排版方式用于区分不同的信息，以下是这些排版的例子及其解释。

文字中的代码单词、数据表的名字、文件夹名、文件名、文件扩展名、路径、链接、用户输入都按以下方式排版：“为了准确计算 HPD，我们将使用 `plot_post` 函数”。

代码片段的排版方式如下：

```
n_params = [1, 2, 4]
p_params = [0.25, 0.5, 0.75]
x = np.arange(0, max(n_params)+1)
f, ax = plt.subplots(len(n_params), len(p_params), sharex=True, sharey=True)
```

```
    for i in range(3):
        for j in range(3):
            n = n_params[i]
            p = p_params[j]
            y = stats.binom(n=n, p=p).pmf(x)
            ax[i,j].vlines(x, 0, y, colors='b', lw=5)
            ax[i,j].set_ylim(0, 1)
            ax[i,j].plot(0, 0, label="n = {:3.2f}\np = {:3.2f}".format(n,
p), alpha=0)
            ax[i,j].legend(fontsize=12)
    ax[2,1].set_xlabel('$\\theta$', fontsize=14)
    ax[1,0].set_ylabel('$p(y|\\theta)$', fontsize=14)
    ax[0,0].set_xticks(x)
```

所有命令行的输入或输出都按以下方式排版：

conda install NamePackage

读者反馈

欢迎读者对本书的反馈，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，只需要发送电子邮件至 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 的个人账户下载本书的示例代码文件。如果是从别的地方购买的本书，可以访问 http://www.packtpub.com/support，注册后，代码文件会直接通过电子邮件发送给你。你也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 按钮。

（4）在 Search 文本框中输入书名。

（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 按钮。

你也可以在本书网站的页面单击 Code Files 按钮下载代码文件。可以通过在 Search 文本框中输入书名找到本书的网页。你需要登录自己的 Packt 账户。

文件下载完成后，确保使用下列软件的最新版解压或抽取文件：

- WinRAR / 7-Zip for Windows；
- Zipeg / iZip / UnRarX for Mac；
- 7-Zip / PeaZip for Linux。

本书涉及的代码也可以在 GitHub 上下载 https://github.com/ PacktPublishing/ Bayesian-Analysis-with-Python。此外，在 https://github. com/PacktPublishing/ 中还可以查看一系列其他书籍和视频的代码。

下载本书的彩图

我们还提供了本书中所有彩色图表的 PDF 文件，从而帮助你理解印刷造成的差异。你可以从 https://www.packtpub. com/sites/default/files/downloads/ BayesianAnalysiswithPython_ColorImages.pdf 下载。

勘误

尽管我们已经非常细心地努力保证内容的正确性，但是错误还是不可避免。如果你在本书中发现错误，不管是文本错误或是代码错误，请告诉我们。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。如果你发现了任何错误，请访问 http://www.packtpub.com/submit-errata 报告给我们。你只需选取书

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，在搜索框中输入书名，所需信息便会出现在 Errata 部分下。此外，你也可以在以下地址查看和提交勘误信息：https://github.com/aloctavodia/BAP。

版权

互联网上版权资料的盗版问题一直是所有媒介关心的问题。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，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所做的努力，我们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，我们会尽全力解决你的问题。

目 录

第1章
概率思维——贝叶斯推断指南

归根到底，概率论不过是把常识化作计算而已。

——皮埃尔—西蒙·拉普拉斯

本章我们将学习贝叶斯统计中的核心概念以及一些用于贝叶斯分析的基本工具。大部分内容都是一些理论介绍，其中会涉及一些 Python 代码，绝大多数概念会在本书其余章节中反复提到。尽管本章内容有点偏理论，可能会让习惯代码的你感到有点不安，不过这会让你在后面应用贝叶斯统计方法解决问题时容易一些。

本章包含以下主题：

- 统计模型；
- 概率及不确定性；
- 贝叶斯理论及统计推断；
- 单参数推断以及经典的抛硬币问题；
- 如何选择先验；
- 如何报告贝叶斯分析结果；
- 安装所有相关的 Python 库。

1.1 以建模为中心的统计学

统计学主要是收集、组织、分析并解释数据，因此，统计学的基础知识对数据分析来说至关重要。分析数据时一个非常有用的技巧是知道如何运用某种编程语言（如 Python）编写代码。真实世界里充斥着复杂而杂乱的数据，因此对数据做一些预处理操作必不可少。即便你的数据已经是整理好的了，掌握一定的编程技巧仍然会给你带来很大帮助，因为如今的贝叶斯统计绝大多数都是