Pearson

BAYESIAN METHODS FOR Hackers

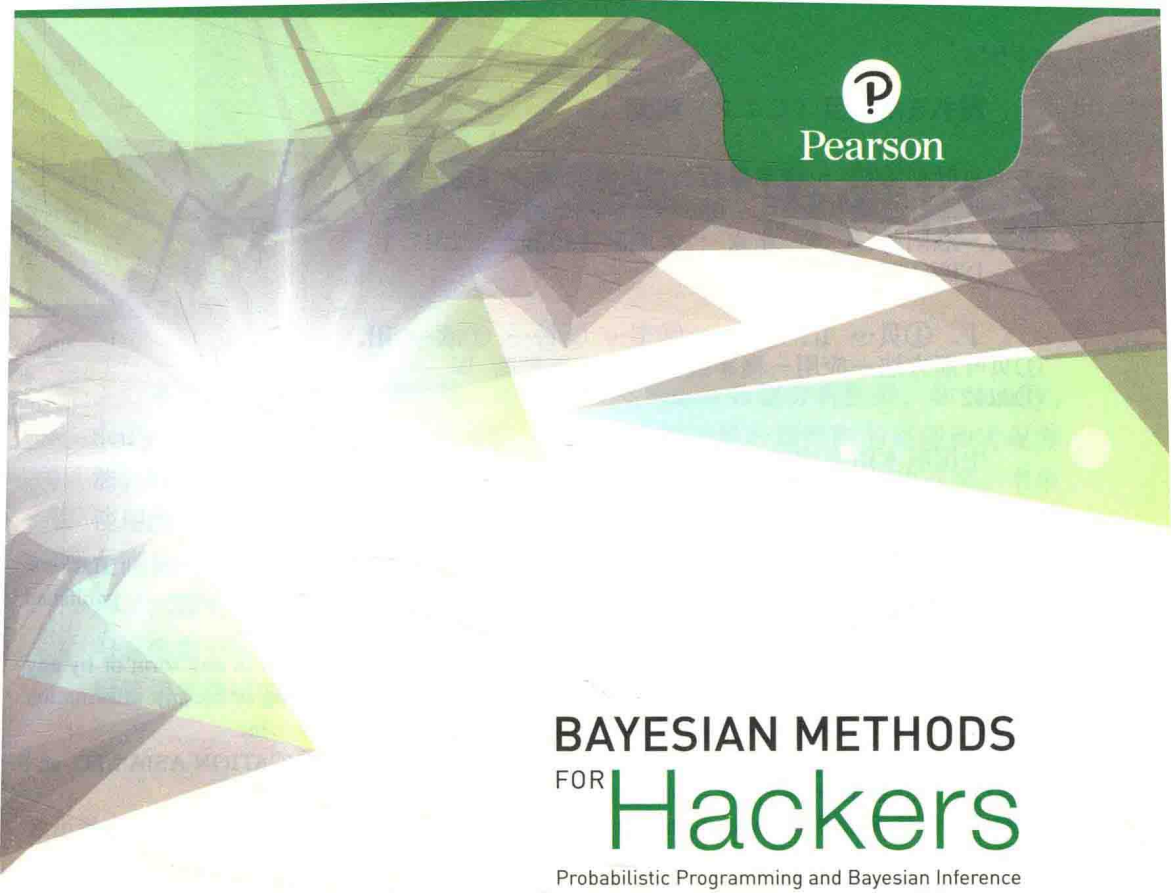
Probabilistic Programming and Bayesian Inference

贝叶斯方法

概率编程与贝叶斯推断

[加] Cameron Davidson-Pilon 著
辛愿 钟黎 欧阳婷 译
余凯 岳亚丁 审校

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



BAYESIAN METHODS
FOR Hackers
Probabilistic Programming and Bayesian Inference

贝叶斯方法

概率编程与贝叶斯推断

[加] Cameron Davidson-Pilon 著
辛愿 钟黎 欧阳婷 译
余凯 岳亚丁 审校

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

贝叶斯方法 : 概率编程与贝叶斯推断 / (加) 卡梅隆 戴维森-皮隆(Cameron Davidson-Pilon)著 ; 辛愿, 钟黎, 欧阳婷译. — 北京 : 人民邮电出版社, 2017.1
ISBN 978-7-115-43880-5

Ⅰ. ①贝… Ⅱ. ①卡… ②辛… ③钟… ④欧… Ⅲ. ①贝叶斯方法－应用－概率统计②贝叶斯推断 Ⅳ. ①O212

中国版本图书馆CIP数据核字(2016)第274840号

◆ 著 [加] Cameron Davidson-Pilon
译 辛 愿 钟 黎 欧阳婷
审 校 余 凯 岳亚丁
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
北京方嘉彩色印刷有限责任公司印刷

◆ 开本：720×960 1/16
印张：14.5
字数：247 千字 2017 年 1 月第 1 版
印数：1－3 500 册 2017 年 1 月北京第 1 次印刷

著作权合同登记号 图字：01-2016-5335 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

内容提要

本书基于 PyMC 语言以及一系列常用的 Python 数据分析框架，如 NumPy、SciPy 和 Matplotlib，通过概率编程的方式，讲解了贝叶斯推断的原理和实现方法。该方法常常可以在避免引入大量数学分析的前提下，有效地解决问题。书中使用的案例往往是工作中遇到的实际问题，有趣并且实用。作者的阐述也尽量避免冗长的数学分析，而让读者可以动手解决一个个的具体问题。通过对本书的学习，读者可以对贝叶斯思维、概率编程有较为深入的了解，为将来从事机器学习、数据分析相关的工作打下基础。本书适用于机器学习、贝叶斯推断、概率编程等相关领域的从业者和爱好者，也适合普通开发人员了解贝叶斯统计而使用。

中文版推荐序

从20世纪80年代末到90年代，人工智能领域出现了3个最重要的进展：深度神经网络、贝叶斯概率图模型和统计学习理论。从2010年以来，由于深度神经网络在语音和图像等应用领域的巨大成功，其重要性被学术界和工业界广泛接受和推崇。相对而言，同样具有巨大实用价值的贝叶斯学习远没有受到充分的重视。在这个背景下，本书的出版对于推动贝叶斯学习和推断的实践具有非常积极的意义。本书通过浅显易懂的方式介绍了各种典型贝叶斯机器学习算法，并结合具体应用给出代码示例，无论是对于在各个公司中工作的工程师，还是从事机器学习研究的学者，在实践方面都有很强的指导价值。我个人相信，在下一个10年里，工程师掌握贝叶斯学习和推断，就像今天掌握C/C++、Python等编程语言一样重要和普遍。

原书序

贝叶斯方法是现代数据科学家运用的众多工具集中的一种，可以用来解决预测、分类、垃圾邮件检测、排序、推断等诸多问题。然而，目前大多数关于贝叶斯统计和推断的资料都注重于数学细节，而较少从更加实用的工程角度进行考虑。因此我很乐意将本书加入到丛书（Addison-Wesley 数据分析丛书）里，带给实践者一本关于贝叶斯方法的必备书籍。

Cameron（本书作者）在该主题上的知识背景，以及他对采用切实可行的例子进行实验的专注，使得本书对于想要学习贝叶斯方法的数据科学家和普通程序员来说，都是一本非常好的入门书籍。本书充满了实例、图表和可运行的 Python 代码，因此你能很容易地开始解决实际问题。如果你对数据科学、贝叶斯方法并不熟悉，或没有用 Python 执行过数据科学任务，本书将是一本帮你起步的无价之宝。

Paul Dix
丛书编辑

前　言

贝叶斯方法是一种常用的推断方法，然而对读者来说它通常隐藏在乏味的数学分析章节背后。关于贝叶斯推断的书通常包含两到三章关于概率论的内容，然后才会阐述什么是贝叶斯推断。不幸的是，由于大多数贝叶斯模型在数学上难以处理，这些书只会为读者展示简单、人造的例子。这会导致贝叶斯推断给读者留下“那又如何？”的印象。实际上，这曾是我自己的先验观点。

最近贝叶斯方法在一些机器学习竞赛上取得了成功，让我决定再次研究这一主题。然而即便以我的数学功底，我也花了整整 3 天时间来阅读范例，并试图将它们汇总起来以便理解这一方法。那时并没有足够的文献将理论和实际结合起来。而让我产生理解偏差的正是由于没能将贝叶斯数学理论和概率编程实践结合起来。当然，如今读者已经无需再遭遇我当时的情景。本书就是为了填补这一空缺而编写的。

如果我们最终是要进行贝叶斯推断，那么一方面我们可以采用数学分析来实现这一目的，而另一方面，随着计算成本的下降，我们已经可以通过概率编程来完成这一任务。后一种方法更加有用，因为它避免了在每一步介入数学干预，而这也使得进行贝叶斯推断不再以通常很棘手的数学分析为前提。简而言之，后一种计算途径，是从问题起点经过小幅中间步骤到达问题终点，而前一种途径则大幅跃进，并通常最后远离目标。此外，如果没有深厚的数学功底，也根本无法完成前一种途径所需要的数学分析。

本书首先从计算和理解的角度，而后从数学分析的角度对贝叶斯推断进行了介绍。当然，作为一本入门书籍，本书将停留在入门阶段。对于受过数学训练的人来说，本书产生的疑问可通过其他偏重数学分析的书来解答。对于缺少数学背景的爱好者，或是仅对贝叶斯方法的实践而非数学理论感兴趣的读者来说，本书足以胜任且蕴含趣味。

选择 PyMC 作为概率编程语言有两方面原因。首先，在写本书之时，并没有集中的关于 PyMC 的说明和实例等资料。官方文档面向具有贝叶斯推断和概率编程背景知识的人。而我们希望本书可以鼓励各个层次的人了解 PyMC。其次，随着近来用 Python 实现科学计算框架的流行及其核心进展，PyMC 可能很快会成为核心组件之一。

PyMC 的运行需要一些依赖库，包括 NumPy 以及可选的 SciPy。为了不产生限制，本书的实例只依赖 PyMC、NumPy、SciPy 和 Matplotlib。

本书内容安排如下。第 1 章介绍贝叶斯推断方法以及与其他推断方法的比较。我们会看到第一个贝叶斯模型，并对其进行建立和训练。第 2 章以实例为重点，讲述如何用 PyMC 构建模型。第 3 章介绍计算推断背后的一个强大算法——马尔科夫链蒙特卡洛，以及一些贝叶斯模型的调试技术。在第 4 章里，我们再次回到推断的样本量问题上，并解释为何样本量大小如此重要。第 5 章介绍强大的损失函数，它将在真实世界的问题与数学推断之间建立连接。我们将在第 6 章回顾贝叶斯先验，并通过启发式的方法找到先验的更优解。最后，我们在第 7 章探索如何将贝叶斯推断用于 A/B 测试。

本书用到的所有数据集都可以从这里获得：https:// github.com/CamDavidsonPilon/Probabilistic-Programming-and- Bayesian-Methods-for-Hackers。

致　谢

谨将本书献给许多重要的人：我的父母、兄弟和我最好的朋友。此外，本书要献给开源社区，是他们每天都在为我们默默地做出贡献。

我要感谢参与到本书写作里的人们。首先要感谢的是为这本书的网络版做出贡献的人。这些作者里很多人提交的代码、思路或文章使本书得以完成。然后，我要感谢对本书进行审校的 Robert Mauriello 和 Tobi Bosede，他们牺牲自己的时间来把一些难以理解的抽象概念变得浅显易懂，并缩减内容以便更好地阅读体验。最后，我要感谢我的朋友以及同事，他们在整个过程中给与我支持。

关于作者

Cameron Davidson-Pilon，接触过数学在多个领域的应用——从基因和疾病的动态演化，到金融价格的随机模型。他对于开源社区最主要的贡献包括这本书以及 lifelines 项目。Cameron 成长于加拿大的安大略省圭尔夫市，而就读于滑铁卢大学以及莫斯科独立大学。如今他住在安大略省渥太华市，并在电商领军者 Shopify 工作。

关于译者

辛愿，浙江大学硕士毕业，腾讯公司基础研究高级工程师，舆情系统开发经理。曾在百度从事推荐系统、用户画像、数据采集等相关研究工作，拥有多项专利，组织过上海大数据技术沙龙。目前专注于文本挖掘、舆情分析、智能聊天机器人等相关领域。

钟黎，腾讯公司研究员。曾在中国科学院、微软亚洲研究院、IBM 研究院（新加坡）从事图像处理、语音处理、机器学习等相关研究工作，拥有多项专利，目前聚焦在自然语言处理、深度学习和人工智能等相关领域。

欧阳婷，华南理工大学硕士毕业，腾讯公司后台策略工程师。在电信、互联网行业参与过推荐系统、资源优化、KPI 预测、用户画像等相关项目，拥有多项专利，目前聚焦在欺诈检测、时序分析、业务安全等相关领域。

关于审校者

余凯博士，地平线机器人技术创始人、CEO，国际著名机器学习专家，中组部国家“千人计划”专家，中国人工智能学会副秘书长。余博士是前百度研究院执行院长，创建了百度深度学习研究院。他在百度所领导的团队在广告变现、搜索排序、语音识别、计算机视觉等领域做出杰出贡献，创纪录地连续三次获得公司最高荣誉——“百度最高奖”。他还创建了中国公司第一个自动驾驶项目，后发展为百度自动驾驶事业部。回国前，余博士在德国和美国的工业界工作了 12 年，服务于西门子、微软、NEC 硅谷实验室等机构。他发表的学术论文被国际同行引用超过 12 000 次，2011 年在斯坦福大学计算机系主讲课程“CS121: Introduction to Artificial Intelligence”。余博士在南京大学获得学士和硕士学位，在德国慕尼黑大学获得计算机科学博士学位。

岳亚丁博士，腾讯公司专家研究员，腾讯技术职级评委会基础研究岗位的负责委员。岳博士拥有 19 年在金融、电信、互联网行业的数据挖掘经验，主导或参与过用户画像、在线广告、推荐系统、CRM、欺诈检测、KPI 预测等多种项目。他曾在微软（加拿大）从事行为定向广告的模型研发，另有 11 年的工程结构、海洋水文气象的力学研究及应用的工作经历。岳博士在华中科技大学获得力学博士学位，在美国圣约瑟夫大学获得计算机科学硕士学位。

目 录