机器学习实践手册

用R语言解决你的机器学习问题

关于作者

Atul Tripathi在机器学习和量化投资领域有超过11年的从业经验，并拥有14年的软件开发和研究经验。他一直致力于先进的机器学习技术，如神经网络和马尔可夫模型的研究。在研究机器学习技术的同时，他利用其解决了与图像处理、电信、语音识别和自然语言处理等领域相关的诸多问题。他还利用神经网络模型开发了文本挖掘工具。在量化投资领域，他利用蒙特卡洛仿真开发了价值风险，极值定理，期权定价和能源衍生品等模型 。

关于审稿人

Ryota Kamoshida是Python库MALSS (MAchine Learning Support System), (<https://github.com/canard0328/malss)>的开发者，目前在Hitachi, Ltd公司担任计算机科学领域的高级研究员。

前言

当今世界，数据已经成为新的“黑金”并以指数级的速度增长着。这种增长既包括现有数据的增长，也包括新数据的增长，新的数据以结构化和非结构化的形式来源于多种多样的数据源，例如社交媒体，互联网，文档以及物联网。数据流必须及时地被收集、处理、分析、并最终展示以确保数据的使用者能够在如今快速变化的环境中作出理性的决定。机器学习技术利用待解决问题的上下文应用于这些数据上，用统计学技术确保不断快速到达的复杂数据能够以科学的方式分析。利用机器学习算法从数据中进行迭代学习，可以发现数据中的隐藏模式和规律。机器学习的这种迭代学习的模式是非常重要的，正因如此，当机器学习模型被暴露在新数据中时，他们才能从新的数据集中独立地适应并学习以产出可靠的结论。

我们将首先对本书中包含的多种不同的机器学习主题进行简介，随后，我们将会基于现实世界的问题在不同的章节中对各个主题进行一一探讨，例如分类、聚类、模型选择和正则化、非线性问题、监督学习、无监督学习、强化学习、结构化预测、神经网络、深度学习，还有最后的实例研究 。本书的机器学习算法以R语言作为编程语言。本书适用于R语言的初学者，但是熟悉R语言对理解和使用本书的代码肯定是会有所帮助的。

您将学习如何合理地决定使用哪类算法以及如何应用这些算法得到最佳的效果。如果您想要对图像、文字、语音、或者其他形式的数据都建立有意义的多功能应用，这本书绝对会成为您的得力助手。

这本书包括什么

第1章，机器学习导言，涵盖了机器学习的各种概念。本章使读者了解本书涵盖的各个主题 。

第2章，分类，包括以下算法：判别分析、多元逻辑回归、Tobit回归，泊松回归。

第3章，聚类，包括以下主题和算法：层次聚类，二元聚类，k-均值聚类。

第4章，模型选择和正则化，包括以下主题和算法：收缩方法（shrinkage methods），降维方法和主成分分析。

第5章，非线性，包括以下主题和算法：广义加性模型，平滑样条，局部回归。

第6章，监督学习，包括以下主题和算法：决策树学习，朴素贝叶斯，随机森林，支持向量机，随机梯度下降。

第7章，无监督学习，包括以下主题和算法：自组织映射和矢量量化。

第8章，强化学习，包括以下主题和算法：马尔可夫链，蒙特卡洛模拟。

第9章，结构化预测，包括以下主题和算法：隐马尔可夫模型。

第10章，神经网络，包括以下主题和算法：神经网络。

第11章，深度学习，包括以下主题和算法：递归神经网络。

第12章，案例研究-探索世界银行数据，包括世界银行数据分析。

第13章，案例研究-再保险合同定价，包括再保险合同定价。

第14章，案例研究-用电量预测，包括用电量预测。

这本书你需要什么？

本书的的重点是用R语言构建基于机器学习的应用。我们已经使用R语言构建过各种解决方案。我们的重点是利用R语言库和函数以最佳的方式来克服现实世界的挑战。我们尽量保持所有代码的友好和可读性。我们认为这将使我们的读者能够很容易地理解代码， 并在不同的场景中随时使用它。

这本书的目标读者是谁？

这本书是为想构建实用的基于机器学习的应用的，从事于统计，数据分析，机器学习，计算机科学，或其他专业的学生和专业人士准备的 。本书适用于R语言的初学者，但是熟悉R语言对理解和使用本书的代码肯定是会有所帮助的。对于那些希望在现有技术栈中探索机器学习技术的有经验的 R语言程序员来说，本书也将是非常有用的。

特殊段落

在本书中，您讲频繁看到如下标题（准备工作和怎样做）。

我们使用如下这些特殊段落来更加清晰的说明怎样完成一个流程指南。

准备工作

这部分告诉你该流程指南需要哪些准备，并描述了如何设置该流程指南所要求的软件或其他先决条件。

怎样做……

该部分包含了流程指南的各个步骤。

公约

在本书中，您将发现许多用于区别不同信息类型的文本样式。下面是一些文本样式的例子和它们的含义解释。

文本中的代码、数据库表名、文件夹名称、文件名、文件扩展名、路径名、虚拟URL、用户输入和 Twitter 句柄用下面的格式表示：“我们将把数据保存到 fitbit\_details 帧中"

任何命令行输入和输入以如下格式表示：



**新名词**和**重要的词语**用加粗字体表示。你在屏幕上看到的文字，例如菜单或对话框中的文字，以这样的方式显示：“**Monte Carlo v/s Market n Zero Rates** ”

 警告或者重要的提示

 小提示和技巧

读者反馈

我们永远欢迎读者的反馈。让我们知道你对这本书的看法——你喜欢或者不喜欢的内容。读者反馈对我们来说很重要, 因为它能帮助我们了解最能让您受益的主题 。

发送反馈给我们，可以发送电子邮件到 feedback@packtpub.com 并请在邮件中注明该书的标题 。

如果你有, 你有兴趣在写作或贡献一本书, 见我们的作者指南在 www.packtpub.com/authors。