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 。

如果你擅长某个话题并有兴趣写作或贡献一本书，请见我们的作者指南[www.packtpub.com/authors](http://www.packtpub.com/authors)。

客户支持

现在你已经是Packt book的一名骄傲的拥有者，我们有许多帮助你从这次购买中获益更多的渠道。

下载示例代码

您能从您在<http://www.packtpub.com>的账户中下载这本书的示例代码文件。如果您是在其他地方购买的这本书，您可以访问<http://www.packtpub.com/surpport> 并注册，这些文件就会通过email的方式直接发送给你 。

您可以通过以下步骤下载代码文件：

1、使用您的电子邮件地址和密码登录或注册我们的网站 。

2、将鼠标悬停在顶部的**SUPPORT**选项卡上。

3、点击**Code Downloads & Errata**。

4、在搜索框中输入该书的名称。

5、选择您要下载代码文件的那本书。

6、从下拉菜单中选择您购买这本书的地方。

7、点击**Code Download**。

您也可以通过点击Packt出版网站中这本书页面的**Code Files**按钮来下载代码文件。通过在搜索框中输入这本书的书名可以访问到该页面。请注意您需要登录您的Packt帐号。

文件被下载后，请使用最新版本的解压软件解压或提取文件夹：

Windows: WinRAR / 7-Zip

Mac: Zipeg / iZip / UnRarX

Linux: 7-Zip / PeaZip

这本书的代码包也被托管在GitHub上，https://github.com/PacktPubl ishing/Practical-Machine-Learning-Cookbook。我们种类丰富的书和录像中的代码包同样被托管在GitHub上，地址是<https://github.com/PacktPublishing/> 。快去查看吧！

下载这本书的彩色图片

我们还为您提供了一个PDF文件，其中包含了本书的彩色截图和图表。这些彩色图片将帮助您更好的理解输出的变化。您能够从下面的地址下载该文件<https://www.packtpub.com/sites/default/files/down loads/PracticalMachineLearningCookbook_ColorImages.pdf>。

勘误表

虽然我们已经尽己所能来保证内容的精确性，但错误难免发生。如果你在我们的书籍中发现了错误，可能是文字中的错误或代码中的错误，并报告给我们的话，我们将不胜感激。通过这样做，您可以帮助其他读者免受错误的影响，并帮助我们改进本书的后续版本。如果您发现了任何错误，请通过如下方式报告给我们。访问<http://www.packtpub.com/submit-errata>，选择你的书名，点击**Errata Submission Form**链接，并输入错误的详细内容。一旦您的勘误被验证，您的提交将被接受，勘误将上传到我们的网站或者被添加到标题为勘误表的勘误表章节下。

如果您想查看以前提交的勘误，请访问https://www.packtpub.com/books/conten t/support，并在搜索框中输入书名。所需的信息将出现勘误部分下。

盗版

互联网上的各种媒体形式的盗版版权材料是一个屡禁不止的问题。在Packt，我们采取的版权和许可证保护非常严肃。如果您在互联网上遇到我们作品的任何形式的非法复制品，请立即向我们提供地址或网站名称，以便我们采取措施。

请通过copyright@packtpub.com联系我们，并附上疑似盗版材料的链接。

我们感谢您的帮助，帮助保护我们的作者，以及我们为您带来有价值内容的能力。

问题

如果您有任何关于本书的问题，请通过 [questions@packtpub.com](mailto:questions@packtpub.com) 联系我们，我们将尽力处理您的问题。

1 机器学习引言

在本章中，我们将介绍机器学习及其涵盖的多个话题。您将了解以下内容：

•什么是机器学习？

•分类方法概述

•聚类方法概述

•模型的选择和正则化概述

•非线性方法概述

•监督学习概述

•无监督学习概述

•增强学习概述

•结构预测概述

•神经网络概述

•深度学习概述

什么是机器学习？

人类自出生起就被暴露在各种数据中 。眼睛、耳朵、鼻子、皮肤、舌头不断收集着各种形式的数据，然后大脑将其转换成视觉、听觉、嗅觉、触觉和味觉。大脑处理各种形式的感觉器官收到的原始数据，并将其转换成语音，进而用语音表达对于这些原始数据的观点。

当今世界，我们用与机器相连的传感器收集数据。从各式各样的网站和社交网络收集数据。之前的手写材料也在电子化，数字化后被加入数据集中。 这些形式丰富的从不用数据源中采集的数据需要经过处理才能得到更有洞察力的，更有意义的结果。

机器学习算法有助于从不同数据源收集数据，转换富数据集的形式，并基于分析结果，帮助我们采取明智的行动。机器学习算法也被设计成高效，准确的方法，并提供通用的学习方法来解决以下问题：

•处理大规模问题

•做出准确的预测

•解决各种不同的学习问题

•学习得到结果并得到在何种条件下这些问题能够被学习

机器学习算法的一些应用领域如下：

•基于销售数据的价格预测

•预测药物的分子反应

•检测汽车保险欺诈

•分析股市回报

•识别高风险贷款

•风力发电厂产量预测

•跟踪和监测医疗保健设备的利用率和位置

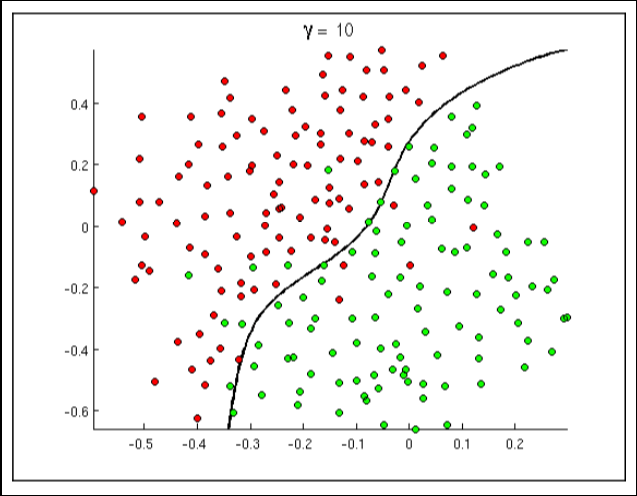
•计算能源有效利用率

•分析智能城市交通运输增长趋势

•采矿业矿产储量估算

分类方法概述

线性回归模型本质上是进行量化的响应，但是这样的响应本质上是定性的。类似于态度（强烈不同意，不同意，中立，同意和强烈同意）这样的响应，其本质上是定性的。对于一个观察来说，预测一个定性的响应可以被当作对这个观察进行分类，因为这涉及到把这个观察分配给一个类别或种类。分类器对于今天的许多问题来说是非常有价值的工具，比如药物或基因组学预测，垃圾邮件检测，面部识别和财务问题。



聚类方法概述

聚类是将相似对象聚合起来的过程。每一个类由彼此之间相似并且与其他组的对象不相似的对象组成。 聚类的目标是确定一组未标记数据的内在分组。 聚类可用于数据挖掘（DNA分析，营销研究，保险研究等），文本挖掘，信息检索，统计计算语言学和基于语料库的计算词典学等不同应用领域。 聚类算法必须满足的要求如下：

可扩展性

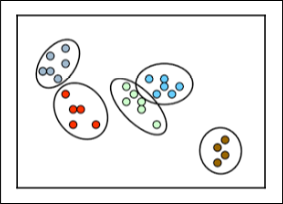
处理各种类型的属性

发现任意形状的集群

处理噪音和异常值的能力

可解释性和可用性

下图显示了聚类的表示：



监督学习概述

监督学习需要学习一组输入变量（通常为向量）和输出变量（也称为监控信号）之间的映射，并应用此映射来预测未知数据的输出。监督学习的方法尝试发现输入变量和目标变量之间的关系。发现的关系在被称为“模型”的结构中表示。通常隐藏在数据集中的模型描述和现象解释，在知道输入属性后，可以用于预测目标属性的值。

监督学习是从有监督的训练数据（训练样本集）推测出预测函数的机器学习任务。训练数据由一组训练样本组成。在监督学习中，每个例子是一个由输入对象和期望输出值组成的。监督学习算法分析训练数据并学习出预测函数。

为了解决监督学习问题，必须执行以下步骤：

确定训练样本的类型。

收集训练集。

确定预测函数的输入变量。

确定预测函数的结构和相应的学习算法。

完成设计

评估预测函数的准确性。

监督学习的方法可以被应用在各个领域，如市场营销，财务和制造业。

在监督学习中要考虑的一些问题如下：

权衡有偏变量

函数复杂性和训练数据量

输入空间的维度

输出数据的噪声

数据的异构性

数据冗余

存在相互作用和非线性问题

无监督学习概述

无监督学习针对输入样本的整体，学习出一种特定的模型来表征输入样本整体的统计结构。无监督学习是重要的，因为它在大脑的学习过程中比在监督学习更常见。例如，眼睛中光感受器的活动是随着视觉世界而不断变化的。他们持续提供着可用于显示世界上有什么对象，如何呈现，照明条件等等的所有信息。然而，基本上没有关于场景内容的信息在视觉学习期间是可用的。这就使得无监督的方法至关重要，并被用作适合神经突触的计算模型。

在无监督学习中，机器接收输入，但是既没有有监督的目标输出，也没有从环境中获得奖励或者反馈。想象一下机器在没有得到坏境的任何反馈时可能会学到什么，这似乎有些神秘。

然而，建立一个正式的无监督学习框架是可行的，因为基于这样的概念——机器学习的目标是建立一种输入的表示用于决策制定，预测未来输入，高效传输输入到其他机器，等等。 某种意义上来说，无监督学习可被认为是在上述数据中构建模式，并且考虑噪声的影响。

无监督学习的一些目标如下：

在大型数据集中发现有用的结构，而不需要目标输出

提高的对于输入数据集的学习速度

通过为每个可能的数据向量分配分数或概率来构建数据向量模型

增强学习概述

增强学习是训练一个会自我行动的“机器人”来最大化它从世界中获取收益的问题。它是关于如何行动以及如何将事件情况映射到动作，以最大化量化的奖励信号的问题。 像大多数机器学习一样，学习者刚开始不知道要采取哪些行动，就而是要通过尝试来发现哪些行为能够产生最大的回报。强化学习的两个最重要的区别特征是“尝试、错误，搜索”的过程和延迟奖励。 强化学习的一些例子如下：

当一名棋手思考下一步棋时，他是通过计划下一步可能的反馈以及计数的反馈来做决策的。

自适应控制器实时调整炼油厂的操作参数。控制器在指定的边际成本的基础上权衡优化收益/成本/质量，而不是严格遵守工程师最初建议的设定参数。

一只瞪羚在出生后几分钟就挣扎地站起来。半小时后，它已经能够每小时奔跑20英里。

教一只狗一个新的技巧 - 不告诉它做什么，但是如果它做正确/错误的事情，就给它可以奖励/惩罚。它不得不弄清楚它是如何得到奖励/惩罚，这被称为信用分配问题。

增强学习就像试错学习一样。代理人应该从环境经验中发现良好的策略，以至在过程中不失去太多的回报。“探索”是关于寻找有关环境的更多信息，而“利用”是利用已知信息来最大化奖励。例如：

餐厅选择：利用：去你最喜欢的餐厅。探索：尝试一个新的餐厅。

石油钻探：利用：在最着名的地点进行钻探。探索：在新的位置钻探。

增强学习的主要内容如下：

策略：策略是代理的行为函数。它确定从环境的感知状态到所采取的行为之间的映射。它对应于心理学所称的一组刺激反应规则或关联。

价值函数：价值函数是对未来奖励的预测。状态的价值是从该状态开始代理人可以期望在未来积累的奖励总额。而奖励决定了环境状态的即时、内在的可取性，价值表示了状态长期的可取性，其中考虑了状态之后的状态序列以及伴随这些状态的奖励。

模型：模型预测下一步环境将会如何变化。它预测下一个状态和下一个状态的即时奖励。

结构化预测概述

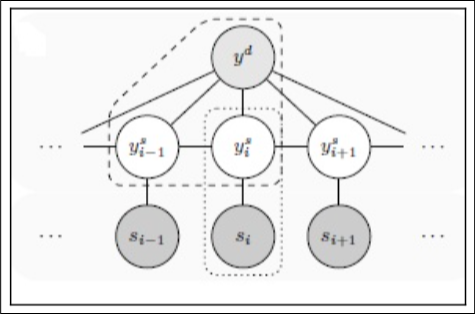
结构化预测是机器学习问题的重要应用领域。考虑输入x和输出y，x和y可能是如下类型的数据：一组标注过的时间序列，一张图片的一组属性，句子的解析成分，或者从一张图片上被分割出的一组对象。结构化预测问题很有挑战性的，因为y‘s是包含它的输出变量的数量的指数级别。由于预测需要搜索巨大的参数以及统计学因素的考虑，结构预测问题的计算上是具有挑战性的。因为从有限的数据学习精确的模型需要推导不同结构化输出之间的共同点。结构化预测本质上是一个映射问题，其中映射必须捕捉x和y之间的区别性交互，并且还允许在y上进行有效的组合优化。结构化预测是关于从输入数据中预测结构化输出，不同于只预测预测一个数值的分类或回归问题。例如：

自然语言处理 - 自动翻译（输出：句子）或句子解析（输出：解析树）

生物信息学 - 二级结构预测（输出：二分图）或酶功能预测（输出：树中的路径）

语音处理 - 自动转录（输出：句子）或文本到语音（输出：音频信号）•

机器人 - 规划（输出：动作顺序）



神经网络概述

神经网络代表了进行信息处理的仿大脑结构。这些模型受生物学的启发，而不是大脑实际功能结构的精确复制。神经网络能够从数据中学习的能力很强，已被证明是许多预测和分析应用场景中非常有应有价值的系统。

人工神经网络通过更新网络架构和连接权重来学习，使网络能够有效地执行任务。它可以从可用的训练模式中学习，或者从训练样本或输入-输出关系中自动学习。学习过程由以下之一设计：

了解可用信息

学习范例 - 从环境中获取模型

学习规则 - 找出更新权重的过程

学习算法 - 通过学习规则确定调整权重的过程

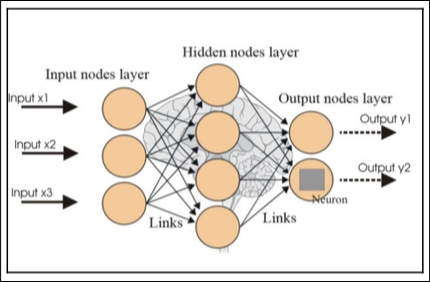
学习规则有四种基本类型：

纠错规则

玻尔兹曼

赫布

竞争学习



输入节点层 隐含节点层 输出节点层

输入x1 输出y1

输入x2 输入y2

输入x3 神经元

连接 连接

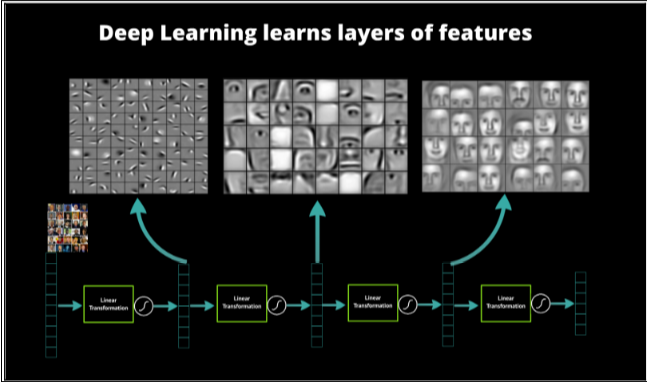
深度学习概述

深度学习是指一组相当广泛的机器学习技术和架构，他们的特点是 使用了多层非线性信息处理结构化模型架构。 深入学习架构有三大类：

无监督或生成式深度学习网络

有监督深层学习网络

混合深度学习网络



深度学习模型学习各特征层