独家策划：

IBM

Machine Learning

for dummies®

了解机器学习的基本原理

了解机器学习算法

建立您的数据科学团队

Judith Hurwitz
Daniel Kirsch

IBM 限量版

Machine Learning

IBM 限量版

作者
Judith Hurwitz、Daniel Kirsch

for dummies®

Machine Learning For Dummies®，IBM 限量版

出版商：
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN: 978-1-119-53567-6 (pbk); ISBN: 978-1-119-53569-0 (ebk)

美国制造

10 9 8 7 6 5 4 3 2 1

出版商致谢

帮助出版本著作的部分人员名单如下：

项目编辑：Carrie A. Burchfield

编辑经理：Rev Mengle

策划编辑：Steve Hayes

业务发展代表：Sue Blessing

IBM 编著者：Jean-Francois Puget、Nancy Hensley、Brad Murphy、Troy Hernandez

目录

简介

机器学习对软件设计的影响越来越大，因此，软件设计必须考虑机器学习，才能跟上业务变化的步伐。机器学习的强大功能有助于您使用数据推动业务规则和逻辑。区别在哪里？在传统软件开发模式下，程序员基于当前业务状态编写逻辑，之后再添加相关数据。然而，业务变化现已成为常态，想要预测哪些变化会推动市场转型的可能性微乎其微。

机器学习的价值在于，您可以不断从数据中获取信息，进而预测未来。这一强大的算法和模型体系广泛应用于各个行业，便于使用者改善流程并洞察数据中的模式和异常。

但机器学习并非孤立的工作，而是一个团队过程，需要数据科学家、数据工程师、业务分析师和业务主管协同合作。机器学习需要依靠合作发挥作用，这样团队才能将工作重心放在解决业务问题上。

关于本书

机器学习傻瓜书 IBM 限量版提供机器学习的各方面知识，以及机器学习如何影响您利用数据获得全新洞察。数据的价值体现在您对数据的使用和管理方面。本书介绍可帮助您为公司创造业绩的各类机器学习技术、模型和算法。业务主管和技术主管都能从中获益，学会如何使用机器学习预测未来。

大胆的假设

本书的信息对很多人都有用，但我们还是要承认，我们的确对您的身份作了一些假设：

- 您已经非常熟悉在公司应用机器学习算法开发新软件的情况。您需要做足准备，将团队引导至正确的方向，以便公司能借助这些强大的算法和模型获取最大价值。
- 您将制定一项长期策略，以便开发经得起时间考验的软件。管理层希望能够充分利用客户、员工、市场前景和业务趋势等所有相关的重要数据。您的目标是为未来做好打算。
- 您了解公司内部数据的巨大潜在价值。
- 您了解机器学习的优点及其对公司的影响，并且您想要确保团队已准备好借助这一强大工具在全新业务模式出现时保持竞争力。
- 作为业务主管，您想要应用这一最重要的新兴技术发挥最大的创新能力。

本书所用图标

以下图标旨在凸显书中的重要信息：

提示有助于识别需要特别注意的信息。

警告图标指出您需要重视的内容。我们重点列出了利用机器学习模型和算法时的常见缺点。

此图标强调您应当牢记的重要信息。

本章要点

- 机器学习和大数据的定义
- 信任数据
- 了解混合云之所以重要的原因
- 使用机器学习和人工智能
- 了解机器学习的方法

第 1 章
认识机器学习

关于新兴高级分析以何种方式为企业提供竞争优势的讨论方兴未艾，其中机器学习、人工智能 (AI) 和认知计算正在占据主导权。毫无疑问，当前的行业领导者正面临全新、难以预料的竞争者。这些企业正寻求新的策略，为未来做准备。尽管企业可以尝试不同的策略，但归根结底，所有的策略都要靠数据说话。在本章中，我们深入探讨了机器学习能在企业策略的制定方面发挥怎样的价值。您应当如何看待机器学习呢？借助可能发挥决定性作用的先进分析技术，您能为公司贡献哪些力量？

什么是机器学习?

对于希望通过创新的方式利用数据资产来重塑业务模式的开发企业而言，机器学习已成为其最重要的话题之一。机器学习为何如此重要？借助适当的机器学习模型，企业能够不断预测业务方面的变化，以便进一步预测未来的趋势。随着数据不断丰富，机器学习模型会确保解决方案不断更新。价值非常明显：如果您利用机器学习时使用最合适且不断变化的数据源，便有机会预测未来。

机器学习是 AI 的一种形式，它支持系统从数据中获取信息，无需进行显式编程。然而，机器学习的过程也并不简单。


谨记

机器学习会用到各种算法，并反复从数据中获取信息以改善结果、描述数据和预测结果。随着这些算法吸收训练数据，之后便可以生成基于这些数据的更精准模型。当您使用数据不断训练机器学习算法，便会获得相应的机器学习模型。训练后，当您往模型中输入数据时，便会获得相应结果。例如，预测算法将创建出预测模型。当您为预测模型提供数据时，您将收到基于训练模型数据的预测结果。如今，机器学习在创建分析模型时必不可少。

您很可能无意识地用到机器学习应用程序。例如，当您访问某电子商务网站查看产品并阅读评价时，可能会看到您感兴趣的其他类似产品。诸如此类的相似推荐并不是一大群开发人员通过硬编码实现的，而是站点实施了机器学习模型的结果。该模型吸收了您的浏览历史以及其他用户的浏览和购买数据，从而为您展示您可能想要购买的其他类似产品。

从数据中迭代学习

机器学习首先用数据集训练模型，之后才将其用于部署。一些机器学习模型为*在线模型*，会不断吸收新数据并随之更新。其他的模型称为“*离线机器学习模型*”。这些模型由机器学习算法生成，部署后不可更改。在线模型的迭代过程会改善数据元素之间的关联类型。由于这些模型的复杂性和规模，这些模式和关联会很容易被人类忽视。模型训练后，便可实时使用模型从数据中学习。


提示

此外，复杂算法可基于变量的快速变化自动进行调整，这些变量包括传感器数据、时间、天气数据和客户情绪指标等。例如，机器学习模型可以做出推断：如果天气变化较快，天气预测模型将能预测飓风，以便及时拉响警报。机器学习经过训练过程和自动调整后，准确度将得到提升。通过近乎实时地处理新数据并训练系统适应数据中不断变化的模式和相关性，在线机器学习算法不断精炼模型。

古老技术的复兴

AI 和机器学习算法并不是新鲜事物。AI 领域可以追溯至二十世纪五十年代。IBM 的研究员 Arthur Lee Samuels 开发了最早的机器学习程序——一款下西洋跳棋的自主学习程序。事实上，是他创造了*机器学习*这一术语。他的机器学习方法在 1959 年发表于 *IBM Journal of Research and Development* 的论文中有详细解释。

过去几十年中，AI 技术已广泛用于提升基础代码的表现。过去几年中，随着分布式计算模型以及计算和存储的成本降低，人们对 AI 和机器学习的兴趣高涨，进而投入大量资金在初创软件公司上。如今，

我们实现了重大进步，开发出许多商业解决方案。为什么市场需求变得明确起来？有 6 个关键推动因素：

- 现代处理器的功能已逐渐增强，密集度也越来越大。密集度与性能的比值大大提升。
- 存储和管理大量数据的成本大大降低。此外，创新的存储技术也使得运行速度更快，并且能够分析更大规模的数据集。
- 跨计算机集群分布计算处理的能力大大提升了分析复杂数据的能力，且用时极短。
- 有更多的业务数据集可用于支持分析，其中包括天气数据、社交媒体数据和医疗数据集。很多此类数据都以云服务和定义明确的应用程序编程接口 (API) 的形式提供。
- 机器学习算法已在拥有庞大用户群的开源社区公布。因此，更多资源、框架和库将使开发变得更加容易。
- 可视化更易使用。您无需成为数据科学家，就能解读结果，并将机器学习广泛应用于诸多行业。

大数据的定义

*大数据*是拥有以下四个共同特点（又称为“4V”）中任意一个的数据源：

- 极大的数据*量级 (Volumes)*
- 以极快的*速度 (Velocity)* 移动数据
- 极广泛的数据源*类型 (Variety)*
- 极高的*准确性 (Veracity)*，确保数据源的真实性

机器学习模型用大数据训练后，准确性会大大提升。当数据不足时，只能依靠少量数据做出判断，可能导致错误解读趋势，或者错过刚开始出现的模式。而大数据对于训练机器学习模型非常有用，企业仅需几千个数据点便可使用机器学习。


警告

不要低估您正在处理的任务。数据的准确性和背景信息必须能够得到证实。在迅速变化的市场中，创新企业需要部署灵活的模型，能在几毫秒内做出推断，为面临风险的客户快速评估最佳解决方案，让客户满意。这就必须确定用以分析的数据的正确数量和类型，进而分析对业务结果的影响。大数据包括所有数据，如来自电子邮件、社交媒体、文本流、图片和机器用传感器的结构化、非结构化及半结构化数据。


警告

传统的商业智能 (BI) 产品并不能很好地处理不断变化的数据源带来的复杂性。BI 工具一般是为处理易于理解的高度结构化数据而设计，这些数据通常存储在关系数据库中。这些传统 BI 工具一般只分析数据的快照而不是全部数据集。对大数据进行分析需要能够收集、存储、管理和操作大规模数据的技术，以实现高速运行，及时获得准确的洞察。随着计算技术的发展和混合云架构的出现，以前只能靠成本高昂的超级计算机处理的海量数据，如今已经可能以较低的成本管理。

与机器学习相关的大数据

机器学习需要获得正确的数据集，以便将其应用于学习过程。虽然企业不一定要有大数据才能使用机器学习，但大数据能帮助提升机器学习模型的准确性。将数据虚拟化应用于大数据现已成为可能，这样数据就能以最高效且最节约成本的方式存储在本地或云上。此外，还有其他一些物理限制会妨碍以可接受的速度处理海量数据，而这些物理

限制也由于网速和可靠性提高被消除了。加之计算机内存的成本和成熟度带来的影响，完成所有上述技术转型后，企业现在能以各类特定方式充分利用数据，这些方式放在五年前简直是天方夜谭。

任何技术转型都不能孤立地实现；变化往往伴随着无法解决的业务问题以及技术的成熟。重要技术已经足够成熟到支持机器学习的复兴，这样的例子数不胜数。不断成熟的大数据技术包括数据虚拟化、并行处理、分布式文件系统、内存数据库、容器化技术和微服务。这些技术优势的组合能帮助企业解决重大业务难题。企业从不缺少大量的数据。行业领导者几十年来无法使用其丰富的数据源获取可付诸实践的洞察，这让他们非常沮丧。

借助大数据技术和机器学习模型，企业能够预测未来，更好地为颠覆做好准备。

理解和信任数据的必要性

要想提供准确的机器学习模型，只是获取大量数据还远远不够，确保数据源的准确性和意义也很重要。此外，这些数据源相互之间应有一定关联，这样建立的模型才能准确可靠。您需要理解数据源的源头，以及数据源之间的关系是否合理。

除了相信数据之外，执行数据清理或整理也至关重要。清理数据是指您将数据转化为能被机器学习算法理解的形式。例如，算法使用的是数字，但数据通常是文字形式。您需要将文字转化为数字。此外，您必须确保这些数字生成的方式合乎情理且具有内部一致性。您需要决定您将如何处理缺失数据和其他数据的不规则变化。

数据精化为创建能生成可靠结果的分析模型奠定了基础。数据精化过程有助于确保数据及时、干净、易懂。

混合云的重要性

处理机器学习和大数据的过程中，不少企业发现结合公共和私有云服务，是确保可扩展性、安全和合规的最有效方式。例如，一家公司可能想要充分利用云上的图形处理器 (GPU)，而不是自行构建基于 GPU 的环境来进行深度学习。这就是一种混合式方法。

混合云是将私有云和公共云服务合二为一，旨在和谐地发挥作用。混合云环境能基于诸如成本、安全和性能等关键因素，采用灵活方式针对特定工作量为企业提供最适合的服务。

云计算支持企业测试新技术，无需承担本地硬件巨大的预付成本。团队能迅速使用机器学习技术开展工作，无需经过采购和集成过程。随着企业不断成熟，安全、控制或云计算成本会急剧上升，企业可能会选择在本地配置部分硬件。

充分利用机器学习的力量

过去 30 年间，分析在企业运营过程中发挥的作用已发生显著变化。不少公司的分析成熟度级别都经历了从描述性分析、到预测分析、再到机器学习和认知计算的提升。有些公司已成功通过分析理解了公司过去的情况，并藉此预测未来。这些公司能够描述各类行动和事件将如何影响结果。虽然分析得出的结果能用于做出预测，但一般而言，这些预测都带着预先期望的色彩。

数据科学家和业务分析师以前受到分析模式的限制，这些分析模式基于历史数据做出预测。然而，总是存在未知因素会对未来结果产生重大影响。当业务环境发生变化时，企业需要寻求办法，建立能够做出反应和变化的预测模型。

在本节中，我们将为您提供两种先进分析方法。

描述性分析

描述性分析帮助分析师了解业务现状。您需要了解历史数据的背景信息，以便了解当前业务的实际情况。这种方法有助于企业回答诸如哪种产品风格本季度更畅销，以及哪 些地区增长最大 / 最小等问题。

预测分析

预测分析能发现数据中的模式和异常，由此来预测变化。采用这种模型，分析师可以处理多种相关数据源，进而预测出结果。预测分析充分利用成熟的机器学习算法，来获得持续洞察。

预测分析工具要求不断为模型提供反映业务变化的新数据。这种方法能帮助企业预料到客户偏好、价格侵蚀、市场变化和其他会对业务成果造成影响的因素发生的细微变化。

借助预测模型，您可以窥见未来。例如，您能够回答以下类型的问题：

- 如何改变网络体验以吸引客户频繁消费？
- 根据国际新闻和内部财务因素，您将如何预测某股票或投资组合的表现？
- 根据肿瘤的具体特点和基因序列，确定哪种药物配方能为癌症患者提供最佳疗效？

统计学和数据挖掘结合机器学习能发挥的作用

统计学、数据挖掘和机器学习有助于了解数据、描述数据集的特点，以及发现数据内部的相关性和模式并建立模型。这些技术和学科工具应用于解决业务难题时会出现大量重叠的情况。

许多广泛使用的数据挖掘和机器学习算法植根于传统的统计分析。数据科学家将统计学、数据挖掘和机器学习专业知识与技术背景相结合，以实现所有学科的协作。除了结合用于预测结果的上述能力和技术，了解业务问题、业务目标和专业知识也必不可少。单纯依靠统计学而不考虑业务方面，就无法获得好的结果。

以下观点阐述了这些学科之间的相互关系：

- **统计学**是对数据进行分析的科学。经典或传统统计学本质上是依靠推理，这种方法可用于得出数据（各类参数）的相关结论。统计建模的重点主要是做出推断以及理解变量的特点。机器学习模型能利用统计算法并将其应用于预测分析。在统计模型中，可以通过假设的方式确认特定算法的有效性。
- **数据挖掘**基于统计学原理探索并分析大量数据，以发现其中的模式。在此过程中，会利用算法发现数据中的相关性和模式，然后利用发现的相关模式信息做出预测。数据挖掘可用于解决各类业务问题，如欺诈检测、购物车分析和客户流失分析。一般而言，企业会使用数据挖掘工具处理较大量级的结构化数据，如客户关系管理数据库或飞机零部件库存。数据挖掘的目标是解释并理解数据。数据挖掘并不是为了做出预测或支持假设。

一些分析供应商提供特定软件解决方案，以实现结构化和非结构化数据的数据挖掘。一般而言，数据挖掘的目标是从较大的数据集提取数据，用于分类或预测用途。在数据挖掘中，数据呈集群形式。例如，营销人员可能对回应优惠促销的人群特征感兴趣，而并不关心对促销无回应的人群。在本例中，数据挖掘将根据两个不同类别提取数据，并分析每个类别的特点。营销人员可能有兴趣预测哪些人群会回应促销活动。数据挖掘工具旨在支持人工决策流程。因此，数据挖掘应该展示能被人类利用的模式。相比而言，机器学习能自动发现用于进行预测的模式。

鉴于机器学习算法对高级分析的重要性，下一节“机器学习的背景知识”中会详细介绍这一学科。

机器学习的背景知识

要了解机器学习所扮演的角色，我们需要为您提供一些背景知识。言及大数据、分析和先进技术时，经常会提到 AI、机器学习和深度学习等术语。AI 可以理解为是描述可以“思考”系统的广义范畴。例如，能了解您偏好的恒温器或者能识别照片中不同个体及其动作的应用程序都可以看作是 AI 系统。

如图 1-1 所示，AI 有四个主要分支。在本书中，我们重点介绍机器学习。然而，为了理解机器学习，了解其来龙去脉非常重要。

图 1-1：AI 是包含机器学习和自然语言处理的总体范畴。

当我们探究机器学习时，我们侧重于学习能力，以及基于数据而不是显式编程来训练模型的能力。在第 6 章，我们重点介绍了应用机器学习解决业务难题。

深入探究机器学习的类型之前，很有必要了解 AI 的其他分支：

» **推理：**机器推理可以使系统基于数据做出推算。其实，推理有助于填补数据不完整导致的空缺。机器推理有助于得出数据之间的相互关联。例如，如果某系统数据充足，问“食用鸡腿的安全内部温度是多少度？”，系统能够告诉您答案为“165°C”。其中的逻辑链如下：鸡腿是用来吃的，是指鸡的腿部（而不是鼓槌），鸡腿包括黑肉，而黑肉需要在 165°C 下烹制，因此答案是“165°C”。***注意：***在此例中，系统并未显式训练过有关鸡腿安全内部温度的数据，而是依靠现有知识填补了数据空缺。

- **自然语言处理（NLP）：**NLP 是训练计算机理解书写文字和人类讲话的能力。NLP 技术对于捕捉文档中非结构化文本或用户交流的含义非常重要。因此，NLP 是系统解读文本和口头语言的主要方法。此外，NLP 还是非技术人员使用先进技术的基础性技术。例如，NLP 无需编码就能帮助用户询问系统有关复杂数据集的问题。不同于结构化的数据库信息依靠架构赋予数据相关性和含义，非结构化的信息必须经过解析和标注才能找到文本含义。NLP 的必要工具包括分类、本体论、测试、目录、字典和语言模型。
- **规划：**自动规划是智能系统以自主且灵活的方式构建操作序列以实现特定最终目标的能力。与预编程的决策过程（从 A 到 B 再到 C，以得到最终结果）不同，自动规划更为复杂，需要系统基于所给难题的相关性做出调整。

机器学习的方法

机器学习技术在提高预测模型的准确性方面必不可少。根据待解决业务问题的性质，针对不同类型和量级的数据存在不同的方法。在本节中，我们将讨论机器学习的类型。

监督式学习

监督式学习通常以现有数据集和该数据分类方式的特定理解为开端。监督式学习的目的在于发现数据中的模式，以便应用于分析过程。这些数据具有定义数据意义的标记特性。例如，一组数据包括数以百万计的动物图片以及每种动物的介绍，这样您就能开发一种机器学习应用程序将不同动物区分开来。通过给数据中各种动物类别添加标记，

您可以创建数百个不同物种的分类。数据的属性和意义确定之后，训练模型数据的用户自然能充分理解，以便将这些数据与标记的详情相匹配。标记为连续时，称之为回归；数据来自有限值集时，称之为分类。其实，监督式学习用到的回归有助于您理解变量之间的相关性。天气预报就是监督式学习的示例之一。天气预报使用回归分析将已知的历史天气变化模式和目前天气条件考虑在内，进而预测天气状况。


提示

当用预处理的示例训练算法时，算法的性能使用测试数据进行评估。数据子集中发现的模式有时在更大量级的数据中无法检测出来。如果此模型只能表示训练子集中存在的模式，则表示遇到了被称为过度拟合的问题。过度拟合意味着您的模型仅适合您的训练数据，可能无法应用于较大量级的未知数据。为避免出现过度拟合，需要参照未预见或未知的标记数据进行测试。使用未预见的数据作为测试集能够帮助您评估模型在预测结果方面的准确性。监督式训练模型广泛适用于解决各类业务问题，包括欺诈检测、推荐解决方案、语音识别或风险分析。

无监督学习

无监督学习非常适合用于处理大量无标记的数据。例如，社交媒体应用程序（如 Twitter、Instagram、Snapchat 等）都拥有大量无标记数据。要理解这些数据代表的意义，需要算法能基于其发现的模式或群集对数据分类，并以此为根据理解数据的意义。因此，监督式学习执行分析数据的迭代过程，无需人为干预。无监督学习常结合垃圾邮件检测技术一起使用。正常邮件和垃圾邮件中过多的变量，让分析师无法标记来路不明的批量电子邮件。因此，基于聚类和相关性机器学习分类器被用于识别垃圾邮件。

无监督学习算法将数据分割成示例群组（集群）或功能群组。无标记的数据创建数据的参数值和分类。实际上，这一过程会向数据添加标记，使其成为监督后的数据。无监督学习能够决定数据量级较大时的结果。在这种情况下，开发人员不知道数据分析的相关情况，因此在这一步无法添加标记。也因此，无监督学习能作为将数据交给监督式学习过程之前的第一步骤。

无监督学习算法能够帮助企业理解大量未标记的全新数据。与监督式学习类似（详见上一节），此类算法也会寻找数据中的模式。然而，不同之处在于，这一算法处理的数据还未被理解。例如，在医疗行业，收集特定疾病的大量数据能帮助医师获得症状模式洞察，并将其与患者的结果建立关联。要标记与疾病（如糖尿病）相关联的所有数据来源将花费大量时间。因此，无监督学习方法能比监督式学习方法更快获得结果。

强化学习

强化学习是一种行为学习模型。该算法接收数据分析的反馈，进而方便用户得到最佳结果。强化学习与监督式学习的其他类型不同，该系统并不会使用样本数据集进行训练，而是通过试错法进行学习。因此，一连串能解决当前问题的正确决策将促使过程“强化”。

强化学习最普遍的应用领域之一是机器人或博弈。以训练机器人爬楼梯为例。机器人根据它的行动后果改变爬楼梯的方式。机器人坠落时，数据会重新校准，以便采取其他方法尝试爬楼梯，直到机器人通过试错训练了解其中的规律。换言之，即机器人从成功的动作序列中总结规律。这种学习算法需要能发现成功爬楼梯而不坠落的目标以及能实现这一结果的事件序列之间的相关性。

强化学习算法也应用于自动驾驶汽车。在许多方面，训练自动驾驶汽车的过程因存在诸多潜在障碍而变得极为复杂。如果道路上的所有汽车都是自动驾驶的，试错将更容易进行。然而，实际情况是，人类司机的行为通常是不可预测的。即便面临这种最复杂的情形，这种算法也能不断进行优化，从而使行动有所收获。简而言之，可以把强化学习理解为训练动物做动作并给予其相应奖励的方式。如果狗狗每次听到命令后坐下能得到奖励，它每次都会这么做。

神经网络和深度学习

深度学习是机器学习的一种特定方法，融合连续层中的神经网络，以便通过迭代方式总结数据中的模式。当您试图了解非结构化数据中的模式，深度学习尤其实用。

深度学习（复杂的神经网络）旨在模仿人类大脑的运作机理，以训练计算机处理定义模糊的抽象概念和难题。一般的 5 岁儿童能轻松区分老师的脸孔和学校安全员的脸孔。相比之下，计算机需要执行大量操作才能完成区分出哪位是老师哪位是学校安全员。神经网络和深度学习常用于图像识别、语音识别和计算机视觉应用。

神经网络由三个或三个以上的层组成：一个输入层、一个或多个隐藏层以及一个输出层。数据的吸收在输入层完成。然后数据将基于应用于这些节点的权重在隐藏层和输出层得到改进。典型的神经网络可能包括几千乃至几百万个密集相连的简单处理节点。深度学习一词适用于神经网络中具备多个隐藏层的情况。使用迭代方法，神经网络能不断调整并做出推断，直到找到特定的停止点。神经网络常用于图像识别和计算机视觉应用。

深度学习是机器学习技术的一种，它使用分层神经网络从无监督算法和监督算法中总结规律。深度学习通常被称为机器学习的子学科。一般而言，深度学习从未标记的非结构化数据总结规律。尽管深度学习和传统的神经网络十分相似，但这一技术可以拥有更多的隐藏层。问题的复杂度越高，模型中的隐藏层也越多。

深度学习会对许多领域的业务带来影响。例如，从汽车到客户管理，语音识别几乎渗透了各行各业。在物联网（IoT）制造应用中，深度学习可用于预测机器何时发生故障。深度学习算法能帮助执法人员追踪已知嫌疑人的一举一动。

本章要点

- 着手制定策略
- 认识机器学习技术在处理业务问题中的应用
- 建立机器学习与结果之间的联系
- 了解机器学习的业务用途

第 2 章
应用机器学习

借助机器学习，您有机会使用业务数据预测业务变化，为未来做好准备。机器学习显然是一套复杂技术，只有找到技术与结果之间的联系才能实现其价值。业务不是静态的，因此，从数据中了解的信息越多，应对业务变化的能力就越强。

着手制定策略

在制定策略之前，先要了解您想解决的问题。当企业经历重大策略过渡时，某些难题会自己浮出水面。当前业务和现有客户互动处于何种状态？客户未来的购买需求是什么以及您能为他们提供什么？答案很明显是要询问客户是否满意，以及他们将来想要购买怎样的产品。这是个不错的出发点，但还远远不够。当出现变革性的新技术时，满意的客户也可能瞬间变得不满意。如果采用传统的商业智能 (BI) 分析，可以充分了解您业务的过往情况，但无法预测业务的未来发展前景。


谨记

业务不是静态的，结构化、非结构化和半结构化的数据中隐藏着关于客户的诸多细节和信息。机器学习技术的价值在于能够揭示庞大数据中的模式和反常现象。选择正确的机器学习算法，并结合适合的数据源，能帮您预测未来。

使用机器学习消除策略偏差

一般而言，策略计划和策略执行的起点，是获取对客户满意度和未来要求的洞察。前景市场怎样？哪些竞争威胁可能对公司造成影响？但这些都远远不够。即使是能力最强的策略顾问也无法预见突然出现的新发现或新趋势。


警告

公司领导层经常受到自身假设和偏见的牵绊。公司管理层总是通过自己的视角看待出现的数据并解读其结果。当出现采用未预见的业务模式的新竞争者时，业务可否持续？尽管很多人对变化缺乏洞察力，但变化随时可能会发生。然而，其主要指标通常隐藏在大量的非结构化或半结构化数据中。

要从大量的非结构化数据中获益，理清这些数据源的确切意义至关重要。数据的来源是什么？哪些人操作过这些数据？数据源是否可靠？先进分析在初期常会导致令人失望的结果，原因在于分析师拿到数据源就用，未先对其进行审查。在采取行动之前，必须确保数据是干净且准确的。确定数据的准确性后，即可将其应用于解决您的业务问题，机器学习技术能提供重要洞察。同时，您还要确保拥有足够的数据，以便发现数据中的模式或异常。


提示

确认数据的质量后，还要对数据应用于解决问题的背景有所了解。例如，一棵树盛夏时节掉叶子，这预示着这棵树可能生病了。如果这棵树在隆冬时节落叶则属于正常现象。因此，如果不了解背景信息，您

很可能会错误地解读结果。同时，还应重点关注数据元素之间的关联。条件之间有什么样的关系？在树木健康的示例中，季节、树叶的颜色和落叶的数量之间有直接的联系。但您还是要在相关性方面多加小心。背景信息错误可能会导致您发现的相关性不合情理。看似树木落叶和网上购买的外套数量之间有一定联系。虽然这两种事件都是因为天气转凉，但树木和外套之间并无联系。

为了让公司实现高效使用机器学习支持业务策略，您需要使用这些统计方法找到数据集内部的模式和异常。当提供的数据源可靠、数据量合适且干净度足够，便可根据需要解决的业务问题使用最合适的机器学习算法创建一种模型。这个模型只是机器学习工作流的开端。

通过使用大量数据，可以建立数据模型、训练数据，进而开始从数据中学习，从而提高制定决策的能力。从数据中学习的价值在于，机器学习系统能够发现可能不那么明显的模式或异常。客户的购买时间和维修时间之间有什么联系吗？某段时间内的天气状况会对销量产生影响吗？社交媒体数据中有能够揭示客户认知和客户购买模式细微变化的标志吗？通过对来自不同数据源的大量数据建模可以提供任何人类个体单纯依靠孤立的数据所无法发现的重要洞察。

警告

将数据的相关性作为分析方法的讨论由来已久。尽管数据相关性非常重要，但有时也可能具有误导性。6 月份橙汁的消耗量和该月交通事故的发生率增长之间或许存在某种联系，但这两者之间没有因果关系。因此，在某些情况下，尽管相关性可能有一定用处，但可能会导致不准确。这就说明了背景信息的重要性。如果橙汁和交通事故之间存在有用的背景信息，则相关性也会有用。因此，开始将机器学习应用于规划和战略过程时，需要将机器学习和高级分析作为不可或缺的工具。

更多数据可以使规划更准确

机器学习在业务策略方面能发挥怎样的作用？以一家对客户满意度进行传统数据分析的公司为例。在分析数据的过程中，该公司发现了数据中的一些异常。鉴于所使用的数据集，分析师剔除了不一致的数据，认为这些数据不准确。然而，如果存在更多数据，那么可能就会从这些被认为出错的异常数据中得到客户购买方式或客户满意度发生了变化的提示。采用最合适的机器学习算法，并向模型添加更多数据，对其进行训练和分析，就会更加清楚地看出，确实存在将直接影响到业务未来发展的变化。

例如，数据科学家发现一些细微差别，于是加入新的数据源，以加强业务变化或发展有关的统计分析或暴露这种分析的缺点。随着时间的推移，模型吸收了更多数据。系统能获得更多洞察和难察觉的规律，进而更准确地预测未来。因此，机器学习是战略规划的重要伙伴。

认识机器学习技术

为确保数据科学家使用正确的机器学习技术实现业务目标，必须了解企业如何更好地利用这些先进技术发展业务和持续专注于新机遇。

机器学习是一种系统性方法，它利用先进算法和模型不断训练数据，并使用额外数据进行验证，目的是应用最适合的机器学习算法来解决问题（我们在第 1 章详细探讨过这一主题）。机器学习的优势在于，它能够利用算法和模型预测结果。正确的做法是数据科学家使用正确的算法，吸收最合适的数据（准确且干净的数据），并使用表现最好的模型。如果上述所有元素都具备，就能不断训练模型，发现数据中的规律，进而得到想要的结果。建模、训练模型和验证过程自动化能实现准确的预测，进而应对业务变化。

建立机器学习与结果之间的联系

机器学习技术有潜力重塑整个市场和业务策略。例如，使用机器学习技术实现的自动驾驶汽车将改变整个汽车行业。机器学习算法和模型正在改变 X 射线图像的分析方式。机器学习能主动预测安全隐患，在发生损坏之前进行维修。依靠机器学习技术可以创建出数百种不同的解决方案，这些方案能彻底改变各行各业。

机器学习有多种不同的方法和算法，具体取决于需要解决的问题。您需要了解自己想解决的问题。您设计的模型将反映出对数据的理解和您基于数据预测结果的能力。

应用机器学习满足业务需求

机器学习能为想要利用大数据的企业提供潜在价值，帮助这些企业更好地理解行为、偏好或客户满意度方面的细微变化。企业领导者开始重视其企业及行业中很多通过查询无法理解的事情。对您有益或有害的并不是您知道的问题，而是数据中隐藏的模式和异常。本节提供公司如何使用机器学习技术实现业务差异化的一些示例。

了解客户流失的原因

您是否听到过“留住现有客户比获取新客户成本更高”？客户流失是某些行业长期存在的问题，例如电信、零售和金融服务行业。

了解如何防止客户流失比以往更重要。我们处在新兴公司提供创新业务模式的时代。例如，手机服务提供商过去常设立两年合约，每次变更服务时签约期延长。随着竞争格局的改变，许多公司发现他们不得不抛弃这些合约。这一改变对客户有好处，但导致了客户流失骤升。没有了客户合约的保障，手机公司必须寻求留住客户的新方法。

为防止客户不断流失，必须获得客户历史记录、偏好、过去购买的服务和投诉的相关数据。在高度稳定的市场中，这种方法或许可以用于预测未来。而对于剧烈波动的市场，这种方法并不奏效。您必须能够预测市场变化及客户购买模式的变化。使用机器学习模型能帮助您预测影响收益的变化。实际上，手机提供商需要能够发现数据中的模式和异常。手机提供商具有通过许多不同的客户获取大量数据的优势。通过使用正确的算法，供应商能够创建一种适当的模型，将产品类型与能留住老客户并吸引新客户的促销活动对应起来。留住老客户并吸引新客户的成本有多高？新计划是否会导致收入严重下滑？能带来的收益是否值得所花费的成本？机器学习技术能提供各种类型的预测。

在客户流失方面，传统的 BI 方法和机器学习方法之间有何区别？企业借助传统 BI 方法能够了解过去的情况并对客户忠诚度趋势进行评估。相比而言，机器学习算法则能创建一种吸收大量内部和外部数据的模型。对数据训练和验证后，分析师可以开始预测客户偏好方面的变化。这种模型可以预测客户购买模式的未来变化。

机器学习以统计算法为基础，来创建具有学习和预测能力的模型。用于流失分析的最常见预测模型为分类统计算法，如逻辑回归和神经网络。

识别罪犯

警方在跟踪罪犯方面任务艰巨。因此，街道上出现了越来越多的摄像头，以帮助发现非法活动。但是如何鉴定罪犯呢？虽然一张图片可能胜过千言万语，但如果没有证人指认罪犯，破案并不容易。执法部门正试图通过机器学习利用图像数据。

需要特别指出的是，深度学习算法和基于神经网络的算法非常适合处理面部识别。实际上，神经网络能够模仿人类大脑。借助神经网络算法，研究人员能够识别图像中集群和模式。图像分析可以将对象分类为不同类别（如人、车、道路或路灯），从而建立索引并搜索视频事件。此外，面部识别算法能用于数字化人像照片的各部分，并消除无用的不相关数据。识别人脸所需的重要元素包括眼睛、鼻子、嘴巴以及伤疤等特征。通过收大量集面部图片数据，此算法就能识别人脸中的模式。测试是帮助模型区别两张不同人脸的核心技术。有些新兴神经网络技术能通过稀疏数据实现此类训练，从而使这些系统对警方更实用。

警方如何利用这类的神经网络呢？这一解决方案需要借助已知罪犯的图像数据，其中包括监控摄像头收集的数据，以及可能在当地参与犯罪的嫌疑犯的图片。当犯罪活动发生时（如本地商店发生抢劫），摄像头捕获的图像能识别出劫匪的脸部。这些图像可以与大量数据进行匹配。大致来说，这一模型试图将特定脸部特征与图片库进行对比，以查看是否存在匹配。如果警方发现匹配，他们将快速展开逮捕，无需先花时间与目击者谈话，也不用花费数小时查看商店的监控录像。

预防事故发生

许多行业依靠成熟的预防性维修方法确保流程和系统按预期安全、正常地运作。制造业、石油和天然气以及公用事业等行业的成败与他们预防事故的能力息息相关。常见的做法是制定维修计划，但这还远远不够。例如，环境条件可能会影响机器或系统的运作，供暖或空调系统可能会出现故障。天气条件巨变也会对机器产生影响。


提示

机器学习算法能以多种方式应用于预防性维修。例如，回归算法可以作为特定模型的基础，来预测机器发生故障的时间。各种分类算法能用于为机器故障有关的模式建模。传感器生成的大量半结构化数据，能用于建模和比较性能模式，从而检测出异常。

本章要点

» 通过机器学习改造应用程序

» 了解您的数据

» 认识机器学习周期

第 3 章
深入了解机器学习

机器学习是一套强大的技术，能帮助企业改变他们对数据的理解。这一技术方法与公司利用数据的传统方法区别非常大。机器学习技术并非先从业务逻辑入手而后应用数据，而是使用数据创建逻辑。这种方法的最大好处之一，在于消除了业务假设和偏见，避免领导者采取并非最好的策略。

机器学习需要专注于管理准备充分的正确数据。企业也必须能够选择正确的算法，以提供设计精良的模型。但这项工作并非到此结束。机器学习需要实施数据管理、建模、训练和测试周期。在本章中，我们重点介绍支持机器学习解决方案的技术基础。

机器学习对应用程序的影响

我们敢大胆断言，借助机器学习，您能够以数据为基础，并从数据中获得逻辑。公司如何朝着目标前进？与其他复杂应用程序的开发和部署一样，这需要一个规划过程，以了解需要解决的业务问题，并收集正确的数据源。

这种创建应用程序的方法对公司有何影响？根据逻辑构建应用程序时，您假设业务流程保持不变。然而，实际情况是流程会变。如果您能够首先进行数据建模，便可以发现流程和逻辑中的变化。因此，机器学习能够创建更加动态、高效的应用程序。

算法的作用

有关机器学习的探讨如果缺失专门介绍算法的章节，则不够完整。

算法是计算机遵循的处理、操作和转化数据的一套指令。算法可以很简单，如添加一列数字的技术；算法也可以很复杂，如识别图片中的人脸。

算法要想发挥作用，必须处理成计算机能够理解的程序。大多数机器学习算法通常用以下语言编写：Java、Python 或 R 语言。每种语言都包含支持各类机器学习算法的机器学习库。此外，这些语言拥有活跃的用户社区，会定期提供代码，探讨想法、问题以及解决业务问题的方法。

机器学习算法与其他算法不同。使用大多数算法时，程序员需首先要输入算法。然而，使用机器学习时，省去了这一过程。借助机器学习，数据本身可以创建模型。向算法添加的数据越多，算法就越成熟。当机器学习算法接触到越来越多的数据时，就能创建更加准确的算法。

机器学习算法的类型

选择正确的算法一半是科学，一半是艺术。两位需要解决同一个业务问题的数据科学家可能会选择不同的算法来解决这一问题。然而，了解不同种类的机器学习算法有助于数据科学家找到最佳算法类型。本节简单介绍机器学习算法的主要类型。

贝叶斯算法

*贝叶斯算法*支持数据科学家将对于模型的预设概念进行编码，不受数据代表的意义所限制。前文用大量的篇幅介绍了用于定义模型的数据，您可能会发问，为什么人们对贝叶斯算法感兴趣。原因就是当您不具备大量数据来有把握地训练模型时，此类算法尤其有用。

例如，如果您对模型的一部分已经有所了解，就可以直接借助贝叶斯算法进行编码。以医学图像诊断系统发现肺部疾病为例。如果一份已发表的期刊研究基于生活方式估算不同肺部疾病的概率，这些概率可以编码到模型中。

聚类

*聚类*是一种非常直接、易于理解的技术，即具有相似参数的对象分组在一起（即一个簇）。一个簇中的所有对象之间比其他簇中对象的相似性更高。聚类是一种无监督学习，因为数据是未标记的。这类算法将解读构成每个项目的参数，然后相应地将它们分组在一起。

决策树

*决策树算法*使用分支结构来呈现决策的结果。决策树可以用于映射决策的可能结果。决策树的每个节点都代表一种可能的结果。每个节点会基于结果发生的可能性被赋以一定百分比。

提示

决策树有时用于营销活动。您可能想预测为客户和潜在客户发送八折优惠券的结果。您可以将客户分成四种类型：

- 收到优惠券可能购买的人群
- 无论如何都会购买的人群
- 绝不会购买的人群
- 可能对促销活动有抵触情绪的敏感人群

如果您将营销活动发送出去，显然您想要避免将其发送给上述群体中的三个，因为他们要么不回应，要么无论如何都会买，要么会消极应对。将目标定位在*可能购买的人群*上会给您带来最佳投资回报率(ROI)。决策树能帮助您找出这四类客户群，并基于对营销活动的反应划分潜在客户和客户。

降维

*降维*帮助系统删除对分析无用的数据。这类算法用于删除冗余数据、异常值和其他无用数据。降维在分析传感器和其他物联网 (IoT) 用例的数据时非常有用。在物联网系统中，可能有数千个数据点只是表示传感器处于开启状态而已。存储并分析“开启”数据意义不大，而且会占用重要的存储空间。此外，删除这些冗余数据后，机器学习系统的性能也会得到改进。最后，降维还能帮助分析师实现数据可视化。

基于实例的算法

*基于实例的算法*根据新数据点与训练数据的相似性对其进行分类。这套算法又是被称为“*消极学习法*”，因为没有训练阶段。基于实例的算法只是把新数据和训练数据相匹配，并基于新数据点与训练数据的相似性对其分类。

基于实例的学习并不适合处理拥有随机变量、无关数据或数值缺失数据的数据集。基于实例的算法在模式识别方面非常有用。例如，实例学习可用于化学和生物结构分析以及空间分析。生物、制药、化学和工程领域的分析通常会用到各类基于实例的算法。

神经网络和深度学习

神经网络尝试模仿人类大脑处理问题的方式，并使用互联单位层基于观测数据学习并推断相关性。神经网络拥有若干互通的层。当神经网络的隐藏层数量多于一个时，有时也称为*深度学习*。神经网络模型能够随着数据变化进行调整并总结规律。神经网络通常用于数据未标记或处理非结构化数据的场合。计算机视觉是神经网络的一个关键用例。(如需了解有关神经网络的更多信息，请参见第 1 章)。

深度学习在各类应用领域中都得到了充分利用。自动驾驶汽车使用深度学习帮助车辆了解周边环境。摄像头捕捉到周围环境的图像之后，深度学习算法解读这些非结构化的数据，帮助系统做出近乎实时的决策。同样地，深度学习嵌入放射科医师使用的应用程序，以帮助解读医学图像。

图 3-1 描述了神经网络的架构。神经网络的每一层都会在将数据传输至下一层之前进行过滤并转化。

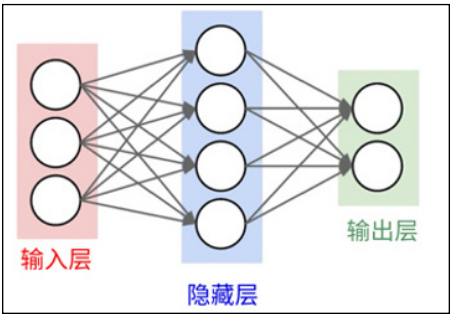

图 3-1： 神经网络架构。

线性回归

*回归算法*常用于统计分析，是机器学习使用的重要算法。回归算法帮助分析师模拟数据点之间的关系。


提示

回归算法能量化数据集中变量之间的关系强度。此外，回归分析可以基于数据的历史值预测其未来的值。但必须记住，回归分析假定相关性与因果关系有关。在没有了解数据相关背景时，回归分析可能产生不准确的预测。

正则化以避免过度拟合

*正则化*是修改模型以避免出现过度拟合问题的一种技术。您可以将正则化应用于任何机器学习模型。例如，您可以对决策树模型正则化。正则化大大简化了易于发生过度拟合的复杂模型。如果模型出现过度拟合，在获得新数据集之后将产生不准确的预测。


谨记

模型的构建与特定数据集太过贴合即发生过度拟合，这类模型对于泛化数据集的预测能力较差。

基于规则的机器学习

*基于规则的机器学习算法*使用关联规则描述数据。基于规则的系统与机器学习系统形成鲜明的对比，后者会创建能普遍适用于所有输入数据的模型。理论上，基于规则的系统非常容易理解：如果输入了 X 数据，则执行 Y。然而，随着系统投入使用，基于规则的机器学习方法会变得非常复杂。

例如，一个系统中可能包含 100 个预定义规则。随着该系统吸收越来越多的数据并得到训练，可能出现数百种规则的例外情况。创建基于规则的方法需要格外小心，确保其不要变得太复杂而失去透明性。试想，要创建一种基于规则的算法应用到税法会多么复杂。

训练机器学习系统

经过开发和改善模型的迭代过程，选择正确的算法并进行训练，之后就可以开始测试系统。训练是机器学习过程的一个重要步骤。

当您训练机器学习系统时，您了解输入的数据（如客户收入、购买历史、位置等），而且了解您想要的结果（预测客户流失的可能性）。然而，原始数据转化成客户流失预测的数学函数很大程度上未知。随着该学习算法不断吸收更多的客户数据，系统能更加准确地预测客户流失的可能性。

训练机器学习算法以创建准确模型可以分为三个步骤：

1. **表示。**

   该算法创建模型，并将输入的数据变成想要的结果。随着该学习算法接触的数据越来越多，就会开始获知原始数据之间的关系，找出哪些数据点可以作为所需结果的有效预测因子。

2. **评估。**

   算法创建多个模型后，无论是人类还是算法都需要根据何种模型的预测最为准确来做出评估，并对模型打分。模型投入使用后，要牢记该模型将处理未知数据。因此，确保对模型进行泛化，避免其与您的训练数据过度拟合。

3. **优化。**

   算法创建多个模型并进行打分之后，即可选择表现最好的算法。由于算法将处理更多元的输入数据集，因此要选择最泛化的模型。

训练过程最重要的部分是要拥有足够的数据，这样才能有效地测试模型。通常来说，第一次训练会产生混杂的结果。这意味着您可能需要改进模型，或者是提供更多数据。这个过程和学习任何新学科都不同，首先要基于不完整的信息做出假设。随着您了解的信息越来越多，就可以确定是否需要从更多数据源获取更多数据。当您从数据中获得更多洞察时，您的假设也可能随之改变。机器学习的价值之一，在于您无需在学习过程的一开始就提前确定问题的答案。

完成训练过程后，即可验证您对于该领域的了解，进而发现您是否拥有足够的知识，或是否需要收集更多数据和了解更多信息。这正是设计机器学习系统时自动实现的过程。

数据准备

人们讨论机器学习时，机器学习算法往往得到极大关注，不过成功却取决于好的数据。

了解数据对成功至关重要。如果您基于错误的数据创建了模型，预测显然不会准确。此外，您需要思考哪些数据应该用于机器学习应用程序。

识别相关数据

在制定业务决策时，需要用到基于各类来源的不断变化的数据。您的数据源可能既包括记录数据（如客户、产品、交易和财务数据）的传统系统，又包括外部数据（如社交媒体、新闻报道、天气数据、图片数据或地理空间数据）。此外，许多数据结构对于分析信息也非常重要，包括结构化和非结构化数据。

结构化数据源

结构化数据通常存储在传统的关系数据库中，指的是拥有明确长度和格式的数据。多数企业的本地数据中心拥有大量结构化数据。结构化数据的示例包括以下几种：

- **传感器数据：**示例包括射频识别 (RFID) 标签、智能仪表、医疗器械和全球定位系统 (GPS) 数据。
- **网络日志数据：**服务器、应用程序、网络等工作时，会捕获有关其活动的各类数据。
- **销售点数据：**当收银员扫描所购产品的条码时，将生成与产品有关的所有数据。
- **财务数据：**如今许多财务系统依靠编程实现，其运行基于自动化流程的预定义规则。
- **天气数据：**将收集天气数据的传感器部署于各城镇、城市、地区，收集温度、风力、气压和降雨等有关的数据。此类数据有助于气象学家创建超本地化的预测。
- **点击流数据：**每次点击网站上的链接时，便产生此类数据。分析这些数据后可确定客户行为和购买模式。

非结构化数据源

非结构化数据尽管有一些隐含结构，但并不遵循特定格式。它们还未得到企业的充分利用，有可能带来巨大的利润。云、移动端和社交媒体使得非结构化的数据飞速增长。非结构化数据的示例包括以下几种：

- **公司内部文本：**如所有文档、日志、调查结果和电子邮件中的文本。实际上，企业信息占据当今世界文本信息的比例非常大。
- **社交媒体数据：**此类数据由 YouTube、Facebook、Twitter、LinkedIn 和 Flickr 等社交媒体平台生成。
- **移动数据：**此类数据包括文本信息、笔记、日历项、图片、视频和输入到第三方移动应用程序的数据。
- **卫星图像：**此类数据包括天气数据或政府在其卫星监控图像上捕获的数据。
- **照片和视频：**此类数据包括安保、监控和交通数据。
- **雷达或声纳数据：**此类数据包括车辆、气象和海洋数据。

数据监管

了解并监管数据是高效使用机器学习解决实际业务问题的先决条件。训练数据时的监管程度，与在生产环境中使用数据时的不同。在传统的数据仓库或关系数据库管理中，公司很可能有明确规定数据需要如何处理和保护的原则。例如，在零售行业，围绕客户个人身份信息制定特定安全条例非常关键。必须确保未授权的人无法访问私密或受限数据。您还可以控制能够查看数据和更改数据的人员。

警告

企业在使用基于机器学习的解决方案预测结果时，必须考虑数据治理影响。构建机器学习应用程序时，请思考以下三项数据监管要求：

- **确保私人数据不受损害。**在项目的开始阶段，了解机器学习应用程序会用到哪些数据类型。例如，应用程序是否会处理受行业规则或政府法规保护的客户或员工数据？如果机器学习算法结果会生成额外客户数据，这些结果可能需要得到相应保护。

- **数据存放必须受到监管规则的保护。**了解数据的实际存放位置，以及机器学习的发生位置。部分国家/地区要求公民数据保存在本国境内。另一些条例和法规可能禁止某些数据迁移至公共云。如果应用程序会将数据迁移至其他位置以执行机器学习任务，必须遵照上述数据存放要求。
- **确保敏感数据的私密性。**了解哪些人有权查看输入至机器学习应用程序的数据。

机器学习周期

创建机器学习应用程序或实施机器学习算法是一种迭代过程。您不能只训练模型一次就置之不理，数据会变化，偏好会改变，竞争对手也会随之出现。因此，模型投产之后您需要不断对其进行更新。尽管您不必进行建模时所采用的同等程度的训练，但您不能认为模型可以自给自足。


谨记

机器学习周期是连续的，选择正确的机器学习算法只是其中一个步骤。机器学习周期的步骤如下：

- **识别数据：**识别相关数据源是周期中的第一步。此外，在开发自己的机器学习算法时，请考虑扩展目标数据以优化系统。
- **准备数据：**确保数据的清洁、安全，并进行妥善监管。如果您基于不准确的数据创建机器学习应用程序，该应用程序将以失败告终。
- **选择机器学习算法：**可能存在多种机器学习算法适用于您的数据和业务挑战。
- **训练：**您需要训练算法以创建模型。根据数据和算法的类型，训练过程可以是监督式、无监督或强化学习。

- **评估：**评估模型，以找到表现最佳的算法。
- **部署：**机器学习算法创建的模型能部署云上和本地应用程序上。
- **预测：**部署完成后，可以基于新输入的数据做出预测。
- **评估预测：**评估预测的准确性。分析预测准确性所得到的信息会反馈回机器学习周期，以帮助提升准确性。


提示

您的模型开始做出预测后，对数据进行评估并重新开始这一过程。所有数据是否都相关？是否存在能够帮助提升预测准确性的新数据集？通过不断改进模型和评估新方法，可确保基于机器学习的应用程序的相关性。

本章要点

- 了解机器学习如何支持您的目标
- 专注于业务问题
- 协作的重要性
- 选择试点项目
- 确定最佳学习模型

第 4 章
机器学习初体验

要使用机器学习技术帮助贵公司达到先进分析水平，这离不开计划和路线图。您不能只是招聘一群数据科学家，而后寄望于他们为公司创造成果。

在本章中，我们重点介绍采用最佳方法开始机器学习过程以支持您的业务目标。思考如何开始从公司产生的数据中获得洞察。如果有系统地采用机器学习技术，就能很好地预测市场变化和客户期望与您开展业务的方式变化。

了解机器学习的作用

在挑选目标项目之前，帮助公司管理层了解机器学习的作用十分重要。机器学习并非无所不能。机器学习方法支持您使用算法创建基于数据的模型。因此，设定期望值非常重要。在第 5 章中，我们将

探讨团队需要的技能类型。您必须拥有专家（如数据科学家），但业务分析师和业务策划师更了解如何将机器学习应用到业务中，从而解决一些十分复杂的问题。数据的充足性和多样性能为业务带来重要优势，能帮助您实现业务增长和变革。

专注于业务问题

开始应用机器学习技术支持业务策略时，必须理解三个基本问题：

- **您想解决的业务问题是什么？**

  确保充分理解了您想解决的问题性质。您可能会发现收入改变或者客户购买的产品类型发生了变化。您了解客户购买的原因吗？您了解市场变化对您满足客户的能力有什么影响吗？一般而言，您虽然拥有大量关于客户、产品组合和市场的信息，但需要对这些信息进行深层分析，以便未雨绸缪。或许您正在考虑向传统客户群推出一款新产品。您需要了解新产品在接下来这一年将给您的收入带来哪些影响。

- **您可以利用哪些隐藏数据源以便更好地了解机遇和威胁？**

  贵公司拥有的业务信息量可能远超想象。从客户支持日志中可以洞察客户所遇到的问题。这些数据能让您洞见修复问题所需的时间。某些数据也以文本形式存储，其中隐藏着客户未来的期望。尽管这些数据都存在，但可能从未被用于了解您的业务。讽刺的是，您其实早已拥有能够预测未来的所有必要数据。这些数据能够帮助您透过表象预测未来。

» **如何整理您的数据?**

问题在于要确保数据已满足特定条件，可以执行相应分析，以便从中发现模式。您使用的数据源是否正确且处于最新状态？您是否已将数据转化为可用格式？您是否保护了客户隐私数据中的身份信息？您是否选择了能将数据放到相关行业背景中的最佳第三方数据源?

尽管机器学习已得到技术和业务市场的关注，但仍要确保所选方法和工具与需要解决的问题最匹配。行业不同，所处理的数据类型不同，以及您期望获得的结果类型不同，那么可以采用的方法也不同。


谨记

对于多数企业，能够理解数据中隐藏的模式具有很大的潜在优势。多数公司拥有孤立地存储在不同业务部门的重要数据。社交媒体源中也有一些重要数据。数据可能是从非结构化的数据源中被发现，如与新的研究发现有关的文档。数据也可能存在于半结构化的数据源中，如传感器或基于物联网的系统。

您的首要任务是确定哪些数据源和数据类型最适合解决您的问题。明白这一点后，便能确定使用怎样的算法来建立最合适的模型。诠释机器学习算法解决具体问题的用例成百上千，本节仅介绍其中三个示例。

统筹数据孤岛

您所处的市场竞争激烈，新兴公司层出不穷，势必会颠覆市场。因此，必须找到了解客户偏好和要求方面细微变化的方式。尽管您不辞辛苦地开展客户调查并回应客户抱怨，但这些信息常常散布于各个业务部门。与客户打交道的每个部门对客户有着不同视角的理解。如果

可以获得所有这些接触点，与客户互动方面会有什么不同？这个角度会告诉您哪些您不曾知道的客户偏好呢？上述很多业务部门与不同客户处理不同的产品线。


提示

借助机器学习，您可以将各类内部和外部数据源统筹起来，并建立模型，以帮您找出会影响您提供的产品及其提供方式的模式和异常。

以某服装连锁店为例，该机构拥有数据并使用最合适的算法了解客户期望的改变，即客户满意的方面和不满的方面。这些数据随之提供购买模式改变方面的洞察。客户群会增加吗？现有客户会流失吗？新客户的人口统计数据有哪些？新客户与现有客户是否以同样的方式购买相同的产品？成功的公司能够打破部门界限之间的数据孤岛，充分利用其数据。颠覆型企业非常敏捷，他们了解数据在发展客户群和增加收入方面的价值。从数据中及早获取洞察和特定模式，可将问题转化为机遇。

防患于未然

大城市能用于解决复杂问题的资源往往较为有限。一些问题会削弱他们应对问题的能力，这些问题甚至有可能会压垮政府。交通问题可能导致交通阻塞、引发事故、造成污染，让城市变得不宜居。当发生洪水或桥梁坍塌事故时，城市支持服务需要在人们受到巨大影响前做好行动准备。不宜居的城市很难吸引新公司落户。


谨记

利用相关数据（如天气数据、备选交通路线数据、社交媒体等）模拟交通模式，有助于城市管理部门提醒市民远离危险地区改道而行。赶在各类事件发生前预测问题可改善状况，使城市更具活力，避免发生生命和财产损失。如何做到这一点？机器学习能够以人类大脑难以企及的速度学习可以改变交通方式的相关模式和条件。

让客户更专注

公司发现有创造业务机会的更好方式时，就会进行创新。为变化做好准备的唯一方式是获得必要的数据和分析，以帮您确定取得所需结果需要采取的最有效的下一步行动。只有对问题了然于心，才能找到问题的最佳处理方式。借助机器学习，当您无法预测答案或结果时，也能找到解决方案。

了解不断变化的客户期望，便可帮助客户在明确需求之前了解他们的需求。了解客户购买模式的细微变化能帮助简化业务，不断变化套餐和优惠策略。讽刺的是，企业通常能从各类公共数据源获得这些数据。将这些数据与客户的相关信息进行匹配，可能获得一些致胜的方法。

机器学习需要协作

机器学习大多关注点在于数据科学家所建模型的可行性。虽然这些模型能够预测业务结果，但与企业严密合作也必不可少。业务线 (LoB) 领导最好能够了解用于分析相关业务的重要数据。然而，他们通常对客户最关心的方面和最重要的数据有偏见。数据科学和数据分析团队发现能提升企业能力的新数据源，从而揭示隐藏的模式和趋势，这一点十分重要。业务部门之间、企业领导层之间和数据科学家之间的有效合作能够创造重大价值，实现真正的差异化和意义重大改变。

开展试点项目

在了解可使用机器学习加以解决的问题类型之后，可能需要进行一些实验。不要寄望于依靠招聘一些数据科学家后让他们分别进行实验。您需要让业务分析师、管理人员、策划师、数据科学家和分析师相互协作。

本节介绍帮您成功开展试点项目的步骤。

第 1 步：确定增长机会

首先，建立问题与业务结果之间的联系。但不要超出您的能力范围。请确保选择小问题，以便轻松找到您拥有的数据以及可以获得的数据源。

您的公司可以利用哪些机会发展业务？或许您已发现，曾受到大众欢迎的产品已经不再畅销。例如，您的产品采用哪种最优包装能增加未来的销量？通过了解数据并进行建模，您可以了解拥有的数据如何帮您预测最佳方案，进而满足客户不断变化的需要。您所选的试点项目也是一种营销工具，可用于向公司证明您有能力预测客户未来的需求。

第 2 步：执行试点项目

用第 1 步得到的具体想法开始实施您的试点项目。确保您明白项目目标和您将使用的数据类型。好的试点项目应当是您力求解决的较大问题的一部分。如果试点项目成功，则表示您设定的目标很合理。您将获得如何确定后续步骤的洞察。

您能从试点项目中学到很多。虽然从成功的试点项目中能学到很多，但从失败的试点项目中可能学到更多。我们能从客户购买模式中得知什么？您能确定客户目前购买产品的方式以及客户购买方式如何开始发生变化吗？这些从数据中得出的新模式以及了解这些模式对企业的价值，可帮助您转变策略。

第 3 步：评估

假设您在执行试点项目（见第 2 步）时发现了一些有趣的模式。您发现客户及其未来需求出现了有趣的变化。您的结果与目前开展业务的方式有什么不同？公司管理层已就客户和客户需求做出特定假设。试点项目的结果是否表明您的假设与结果不符？

摒弃偏见之后，您可能会对结果与设想之间存在的巨大差异十分惊讶。这就是利用机器学习解决业务问题的一大好处：您能以不同方式了解您的公司。您可能会发现试点项目表明客户期望与您的假设和偏见大相径庭。随着您不断添加新数据源，客户需求的变化更加明显。然后，这些答案将回馈到业务规划中，帮助公司在使用新方法提高收入的道路上发展更快。

第 4 步：后续行动

试点项目的好处在于能够为您提供利用机器学习和预测分析方面的洞察，更好地了解您的公司。如果您将试点计划作为一系列项目的第一步，便能从数据中发现模式。在这一阶段，您想进行扩展以纳入更多数据，以使更多业务领导者加入这一过程。选择来自公司的诸多不同领域的更多数据源，能帮助改进分析流程。借助机器学习，您能应用于项目的数据越多，就越有机会获得能应用于业务策略的洞察。

确定最佳学习模型

在应用机器学习来解决业务问题时，最复杂的任务之一是选择最合适的模型。为了使机器学习成为预测业务结果不可或缺的工具，选择最合适的模型是这一过程的最佳出发点。选择模型时最复杂的问题之一，是要确保模型能够在未来引入新数据时也能正常运行。


警告

所选算法必须足够泛化，以便使用新数据时也能保持准确。如果算法与现有数据集的关系过于紧密，此类过度拟合将在未来引发问题。因此，当您选择算法时，首先要确保使用的数据集是信息的代表性样本。如果您的数据集是重点业务领域的代表性样本，试点项目将会更成功。例如，在选择算法时，您可能首先选择公司内部熟知的样品数据集。下一步，您可以添加与您的假设相关但来源完全不同的数据集。您所选的算法如何从人们熟知的数据集和新数据集预测结果呢？

确定算法选择的工具

毫无疑问，为您的数据和业务问题选择最佳算法并非易事。幸运的是，市场开始意识到我们需要有利于算法选择的工具来推动发展。您如何选择正确的模型？这是个很难的问题。


警告

虽然过度拟合或许是一个问题，但更严重的问题在于模型随着时间的推移准确性会降低。因此，随着数据的变化，您必须不断对模型进行重新训练。选择正确的算法最好通过自动选择算法来完成。以分类算法为例。目前存在 40 多种不同的分类器算法。根据数据科学家所使用的数据，您可以结合使用这些不同的算法。于是，您就可以得到数百种可供选择的组合。如果您的数据科学家需要测试可能合适的算法，或许需要花费很长时间才能找到最佳算法。但通过使用自动化工具，科学家能更快速确定评分最高且最适合您数据的最佳算法组合。

自动化工具之所以重要，不仅是因为算法复杂，还因为您需要确保选择用于建立模型的算法不会影响数据延迟和数据一致性。

工具选择方法

各类开源工具都可以用于帮助数据科学家选择正确算法。这些工具通常与所用语言（Python、R、Java 等）有直接关系。为什么数据科学家应使用工具来选择算法呢？许多不同的机器学习模型可能都有助于解决问题。如果数据科学家能使用不同的算法进行实验，便可改进模型预测结果的能力，并创建出能够扩展的模型。

本章要点

» 确定您的团队需要的技能

» 寻找资源以便深入了解机器学习

第 5 章
了解机器学习技能

如果您已从头阅读到本书的这一章，对利用机器学习解决业务问题的复杂性和优势已经有了大致了解。您知道需要为团队配置正确的技能，包括语言和工具。本章中介绍利用机器学习优势帮助企业取得成功的技术知识，以便实现您企业的业务目标。

确定您需要的技能

要将机器学习成功应用于解决极为复杂的业务问题，您的团队需要各种工具。初看上去，您可能会想您能聘用一大批数据科学家。然而，现实情况是很难找到您需要且能够推动公司快速发展的数据科学家。熟练的数据科学家并不多。此外，因为这类人才非常紧缺，您不得不为他们提供很高的薪酬。解决办法是您需要以不同角度思考问题，建立一支专门利用机器学习进行创新的团队。您可以让数据科学家集中精力建立有经验的数据分析师能够使用的模型。同时可以开始选择下

一代工具，以便精明的分析师能够获得诸多机器学习技术。可以利用各类在线培训课程为您的团队提供培训。

本节介绍我们建议您重点关注的十大技能。每项技能都包含多种元素。因此，要确保团队成员在各领域都学到较深入的知识，以提升公司支持业务的能力。

了解可用的工具有哪些

领导层在支持使用机器学习实现公司目标方面发挥着怎样的作用？可用于机器学习的工具或技术不止一种，您可以使用各种各样的工具。您应该花些时间使用不同的方法进行实验，以找到最适合解决问题的方法。您可以寻找有助于工具选择的最佳实践。

学习语言

使用机器学习推动业务时，多种受欢迎的语言都很有帮助。语言的受欢迎程度会随着时间的推移而变化，因此学习一种以上语言通常是非常有用的。诸如 Python、R、Java 和 C++ 等语言是利用机器学习推动业务的基石。在机器学习的环境中运行少不了诸如 Linux、Hadoop、Spark 和云服务等工具。

探索算法

您需要了解在机器学习中有用的各种算法。出色的数据科学家需要对概率论和统计方法有深入了解，在创建有效的机器学习模型时常会用到这些工具。机器学习需要依靠关键算法，包括创建能确定数据中的模式和相关性并确定集群的模型。如需了解机器学习算法的更多详情，请参见第 3 章。

选择合适的模型

将正确的机器算法模型应用于解决棘手的问题十分重要。越来越多的机器学习算法包以 API 的形式存在，包括 Spark MLlib、H2O 和 TensorFlow。开发人员最重要的技能之一是了解哪种算法最合适解决问题。例如，当您想要了解两个点之间的关系时，线性回归模型是适合的解决办法。但如果您要读懂图片的内容，可能需要用到 TensorFlow。许多机器学习技术与多种问题相匹配。数据科学家需要能够确定哪种算法和库最适合解决相关问题。

了解概率论和统计学的价值

大量的学习算法是基于概率论和统计学建立的。朴素贝叶斯、高斯混合模型和隐马尔可夫模型是需要了解的三种重要方法。

了解数据管理

数据科学家还需要对所使用的数据有所了解。数据的来源是什么？这一来源是否可靠且可追溯？用于解决问题的数据源是否合乎情理？在这种情况下，程序员或数据科学家需要与公司联手验证数据源。

评估数据源的干净度

数据源的好坏决定着您的机器学习项目能否成功。您需要了解数据来源，确保其可靠性。还需要确定选择的几种数据源合并后是否合乎情理。

备忘录：建立团队

应如何计划数据科学团队？无论公司规模如何，能助您成功的团队都具备若干共同特征。请牢记，建立团队是为了解决业务问题。该团队或许是由一两名员工负责所有事情，或者公司较大，需要为每项技能配备一名专员。很可能找不到具备所有这些技能的员工（我们称之为“*全才*”），但下方的备忘录能帮助您着手计划：

- 组建一支具有多种技能的团队。要确保平衡技术团队成员与业务成员。
- 选出一位数据科学家主管，该员工应当精通编程和架构原则。此外，该员工必须具备可靠的领导才能，以引导团队实现业务目标。
- 邀请一位了解您的行业和公司的业务分析师。
- 确保团队中有成员能够从数据中发现模式。这一技能与解读数据或理解数据不同，而是使用数据形成讨论或引发行动。
- 选择有代表性的业务领导者，这些领导者了解他们需要从项目获得怎样的成果。
- 为团队配置主题专家，这名专家需要对工作流程和数据性质的细节有深刻认识。专家需要与懂得如何捕获和处理数据的数据工程师合作。
- 必要时寻找能够为团队提供培训的顾问，教他们了解能支持项目目标的新语言或新工具。
- 当公司内部没有所需人才时，邀请特定技术领域的专家。

如果您在一家大型公司工作，公司内部可能有各类人员能完成上述任务。这种情况下，要确保领导能创造较好的协作环境。如果您在小公司，选择团队中真正了解公司基础情况和相关目标的成员。使用您的现有技术团队成员，必要时邀请行业专家为员工提供培训。

了解如何将工作组合到一起

借助机器学习，您最起码可以建立一款基于业务结果的应用程序。因此，您需要了解软件的所有元素和基础设施是如何支持这些结果的。这些元素如何相互拼凑和交流以形成一个系统？当系统加入更多数据和逻辑后，您如何创建可扩展的环境？必须了解，您建立的系统需要测试、管理和记录等等。

了解数据的生命周期

机器学习的最大优点之一是它需要不断吸收新数据，以作出准确的预测。因此，您需要认识到机器学习不是一次性任务，而是连续的过程。您提供的数据越准确，数据量越大，获得的结果也就越可靠。

找到新的应用领域

机器学习在许多不同行业和许多不同领域都能发挥作用。探索机器学习从试点项目到开发产品的过程，有助于获得新应用领域方面的见解。公司的许多其他领域可能受益于机器学习提供的预测分析类型。

学习新知识

因为机器学习是一个新兴市场，迫切需要技术人员帮助企业实现目标。很显然，企业迫不及待地想要找到他们需要的所有熟练专业人才。这对于 IT 专业人员是一个很好的机会，他们可以转变成数据科学和机器学习技术领域的专家。幸运的是，能帮助您学习的资源不计其数。在本节中，我们为您提供了一些资源，愿您有好的开始。

Medium: Inside Machine Learning

此站点为您提供有关机器学习话题的深入分析文章。从天气预测到机器人，您可以探索热门机器学习案例并从行业专家的意见中获得见解。如需了解更多信息，请访问：`medium.com/inside-machine-learning`。

CognitiveClass.ai

立即访问 https://cognitiveclass.ai，免费获取数据科学和认知计算技能。基于 IBM 社区计划进行分类。课程包括“Machine Learning with Apache SystemML”（利用 Apache SystemML 进行机器学习）。

Coursera 在线学习

Coursera 是一个在线学习平台，提供多个领域的课程和学位，包括机器学习。该平台与众多大学合作，提供超过 2,000 种课程。立即访问以下网址报名：www.coursera.org/learn/machine-learning。

Udacity 机器学习相关课程

Udacity 是一家营利性的教育机构，在线提供大型开放式网络课程 (MOOC)。您可以访问以下网址，了解相关信息：www.udacity.com/course/intro-to-machine-learning--ud120。

Galvanize

沉浸式数据科学课程包括深入学习机器学习以及借助结构化和非结构化数据集解决分类、回归和聚类方面的实际问题。学员可以找到 Scikit-learn、NumPy 和 SciPy 等库，并使用实际案例研究将对这些库的理解应用于实际应用程序。如需了解更多相关信息，请访问 www.galvanize.com/san-francisco/data-science。

edX 课程

edX 是一家 MOOC 供应商。该供应商提供大学级别的在线课程。部分课程免费提供。如需了解有关“Machine Learning for Data Science and Analytics”（用于数据科学与分析的机器学习）在线

课程的更多内容，请访问 www.edx.org/course/machine-learning-data-science-analytics-columbiax-ds102x-1。

MITOpenCourseware

MIT 已建立包含其所有课程的网站。这些课程免费提供给参与者。您可以访问以下网址了解有关机器学习的更多信息：http://bit.ly/1tP7pPU。

Google Research Blog

谷歌的研究人员在该网站发表了机器学习和深度学习相关话题的论文。您可以访问以下网址了解有关深度学习的更多信息：research.googleblog.com/2016/01/teach-yourself-deep-learning-with.html。

Kaggle Wiki

Kaggle Public Wiki 是学习统计学、机器学习和其他数据科学概念的资源。该网站提供各类教程以及数据科学竞赛平台。立即访问 www.kaggle.com/wiki/Home。

KDnuggets

KDnuggets 是提供关于分析的海量信息和关于数据科学的各类信息的流行站点。请访问 www.kdnuggets.com/about/index.html 查看相关内容。

Data Science Central

Data Science Central 是大数据从业者的在线站点。该站点包含一个社区平台，该平台设有一个供网友交换信息和提供技术支持的技术论坛。前往 www.datasciencecentral.com 了解更多信息。

IBM 推荐的资源

IBM 机器学习社区能为您提供各类资源，供您补充机器学习方面的知识。如需了解更多信息，请访问以下站点：

- **ibm.com/machinelearning**：了解企业如何使用机器学习解决难题并寻求新机遇。
- **ibm-ml-hub.com**：获得实际专业知识，快速高效地将机器学习应用于改变您的业务。
- **ibm.com/datascience**：研究满足您需求的最佳工具，了解数据科学团队如何合作以便在短时间内带来创新价值。
- **datascienceforall.com**：无论您是对最近开源工具的“码农”还是寻求拖放工具以更好地实现数据科学项目的协作和快速发现的分析师，请访问该数据科学社区，发现最近的最佳实践和能助您成功的资源。
- **datasciencemeetups.com**：了解您所在领域的最近聚会，或加入有数据科学专家参与且拥有丰富共享资源的虚拟聚会。

您还可以使用社交媒体，关注数据科学世界的动态。请访问以下社区：

- **Facebook:** www.facebook.com/IBMDataScience
- **Twitter:** twitter.com/IBMDataScience 或 @IBMDataScience

本章要点

- 了解机器学习如何应用于患者健康
- 利用物联网做出预测
- 响应潜在 IT 问题
- 预防欺诈行为

第 6 章 使用机器学习提供业务问题的解决方案

从社交媒体到复杂的财务应用领域，机器学习在不断渗透到计算的各个方面。机器学习能用于提升客户体验，更好地处理复杂数据并预测结果，甚至可能改变不同企业的运营方式。通过关联数据以检测模式和异常，有助于企业预测结果并改善运营。几乎各行各业都有无数示例。在本章中，我们为您提供了一些应用机器学习解决复杂业务问题的示例。

将机器学习应用于患者健康

患者治疗的最大问题之一是药物通常对每个人的影响都不同。有些药物对某位患者能引起严重的副作用，却可能是另一位患者的有效治疗方式。患者可能患有其他疾病，这些疾病可能引发药物不良反应。年

龄和性别也可能影响药物的疗效。医生经常需要采用试错法找到最合适的疗法。


提示

选择最有效疗法的解决方案之一是基于分类或回归算法建立机器学习模型。分类模型可以基于患者检查和病情的已知结果预测药物的影响。然后利用回归模型可以预测患者服用特定药物时的病情变化。利用数据创建此类模型有助于研究人员了解患者过去对各类药物的反应。构建模型并对其进行训练后，就能够确定某种药物是否对患者最有效。

如果模型为在线模型，随着吸收的患者数据增多，模型还会持续改进。此时，借助应用程序编程接口 (API) 可以建立包括对话界面的解决方案。采用这种方式，医生能与模型互动并询问各类问题，以确保提供副作用较少的合适疗法。

利用物联网获得更多可预测的结果


提示

机器学习模型是物联网 (IoT) 的理想应用。要理解有关物联网数据的分析，首先要知道其处理的是传感器生成的数据集。这些传感器现在既便宜又足够成熟，能支持几乎所有的应用领域。传感器生成的数据包含特定结构，因此是应用机器学习技术的理想选择。尽管数据本身并不复杂，但生成的数据数量通常极其庞大。借助此类传感器数据以及已知中断，机器学习算法能够创建模型以预测未来的机械问题。此模型包括有关运行良好机器的最优指标底线数据以及之前发生故障的数据点。随着对此模型进行训练，模型将能发现异常，进而预测故障发生的概率。

过去使用的方法

为确保质量控制和高效的性能，机器需要定期进行管理、维修和监控。将设备断电进行不必要的维修会造成停机。同样地，运行设备直至其出现故障将导致意外中断甚至灾难性后果。因此，企业需要具备找到潜在问题的能力并及时修复问题，以免导致停机。

达到这一级别的预防性维护并不容易。借助传统的诊断方法，您能够了解近一个月的情况或近一天的情况。制造业公司是较早采用传感器技术的先行者，他们用该技术监控设备运作的好坏。企业过去监控传感器输出的典型方法是确定传感器的输出是否与预期输出匹配。然而，为了预防故障的发生，在故障带来损失之前对其进行预测非常重要。

尽管设备几十年来都已配备传感器，但要聚合传感器生成的数据并不容易。随着网络的进步以及低成本云计算和存储的出现，现在可以聚合此类传感器数据了。随着先进分析技术的出现，捕获传感器生成的信息并应用机器学习技术预测机器可能发生故障的时间成为可能。

主动响应 IT 问题

由于存在不同网络设备、服务器、应用程序、存储系统、端点等，IT 运营一直较为复杂。每个系统都有管理其组件的特定方式。随着软件新版本的实施，要保证系统按预期运行，可能需要更新配置。这是系统交互以维持稳定状态的常见方式。通常来说，某个区域的一个错误就能导致大规模中断，尽管数据中心有重要的仪表，但很难找到问题的根本原因。

一般来说，为追踪系统的运行状况，企业可能部署十多种不同的监控工具。这些监控工具能捕获所监控系统的大量相关数据。然而，由于日志中包含数据，解读如此之多的系统数据非常困难。要理解这些数据，必须要理解上述日志。除了此类日志和系统数据之外，还可以在故障单中找到有价值的数据，其中包含描述问题的文本或来自应用程序性能管理系统的数据。


提示

对这种复杂的 IT 运营数据应用机器学习算法有助于企业主动响应潜在 IT 问题。过去通过事件相关性来寻找性能数据中的模式。然而，有时单凭相关性可能造成误导。因此，为了获得更准确的结果，数据科学家开始将多种机器学习算法聚集起来，以发现事件中的异常。应用机器学习的价值在于它能够基于数据中心产生的复杂数据集（警报、日志和仪表数据或传感器）创建模型。机器学习算法能够基于所有相关数据创建模型。该模型能够理解组成环境的各类元素之间的相关性，还能帮助发现理想性能指标的模式，并将其与现有环境状态进行比较。随着更多数据的加入，模型也会不断得到更新。

防止欺诈行为


警告

检测欺诈行为与猫抓老鼠的游戏类似。不法之徒在实施欺诈时方面变得越来狡猾。随着越来越多的客户使用在线服务，欺诈的可能性也大幅增长。此外，支付服务供应商希望确保为客户提供顺畅的交易，不愿阻止合法支付。许多公司发现，有助于阻止欺诈的唯一方式是使用基于机器学习算法的软件。训练后的模型能在欺诈事件发生前发现异常。事实上，该模型能识别与入侵或未授权行为有关的操作，进而在损害发生前限制入侵者。

打击欺诈已成为一项需要结合多种技术的复杂挑战。为发现欺诈行为，通常会将线性技术、神经网络和深度学习结合起来使用（更多详细信息请参见第 1 章和第 3 章）。长期以来，线性算法被用于区分正当活动和欺诈活动。然而，简单的算法无法预测，因为犯罪分子会经常改变技术。因而，也无法提前预测犯罪活动的发生。

因为单凭线性算法无法发现先进的欺诈技术，所以更加先进的机器学习算法开始投入使用。例如，神经网络和深度学习目前得到支付服务供应商的广泛使用。深度学习模型能处理成千上万的数据点，以了解交易的相关情况。


提示

公司不会孤立地使用神经网络或深度学习。相反，为实现*集成建模*，通常会三种技术结合使用以发挥优势。例如，虽然线性算法可能检测不到某些欺诈活动，但它可能非常擅长捕捉最常见且直观的骗局。最终模型将参考各机器学习模型的结果，并批准或阻止交易。此类评估与患者获取多个医生的意见非常相似。最终的目标是结合多种意见获得更准确的结果。

本章要点

- 将机器学习嵌入应用程序
- 预训练数据即服务将成为必备条件
- 机器学习即服务恐成“爆款”
- 简化机器学习流水线
- 自动化算法选择
- 透明度和信任要求更严格
- 使机器学习成为端到端流程

第 7 章
关于机器学习前景的十大预测

机器学习正逐渐成为软件行业最重要的开发方法。尽管这一先进技术已存在数十年，但到现在才得以商业化。我们正步入机器学习技术成为能够为那些想要了解数据中隐藏价值的企业创造价值的必要工具的时代。机器学习的未来发现前景如何？在本章中，您可以探索我们的十大预测。

机器学习将嵌入大多数应用领域

如今，机器学习技术在各类专业领域都开始流行起来。企业采用机器学习技术以期帮助他们预测未来，并带来具有竞争力的差异化。

在接下来的几年中，您将见证机器学习模型嵌入几乎所有应用程序和各类设备，包括移动设备和物联网中心。在很多情况下，用户将不会意识到他们正在使用机器学习模型。零售网站和在线广告是机器学习模型嵌入的两个典型日常应用领域。在上述例子中，机器学习模型通常用于为用户提供更加定制化的体验。

机器学习将会对各个行业造成巨大且颠覆性的影响。因此，机器学习也将明显改变我们的工作方式。例如，医院可以利用机器学习模型依据社区条件预测入院率。入院可能与天气情况、传染性疾病的爆发和诸如城市中发生的大事件等其他情况有关。

我们正开始见证越来越多的机器学习模型嵌入成套解决方案，如客户管理解决方案和工厂管理系统等。嵌入机器学习模型后，同样的系统会变得更智能，且能够提供预测能力，便于为公司带来价值。

预训练数据即服务将成为必备条件

开发认知和机器学习模型的主要障碍之一是训练数据。一般而言，数据科学家必须承担收集、标记、训练数据的工作。另一个方法是使用面向公众的数据集或众包工具收集并标记数据。虽然上述方法都能解决问题，但其耗时巨大，且执行起来较为复杂。

提示

为克服上述困难，不少供应商提供预训练的数据模型。例如，某公司可提供成百上千个预标记医学图像，以帮助客户开发一款能够筛选医学图像并找到潜在健康问题的应用程序。

模型将需要不断重新训练

目前，大多数机器学习模型都是离线的。这些离线模型使用训练过的数据进行训练，然后进行部署。离线模型部署后，底层模型在吸收更多数据时不会发生变化。离线模型的问题在于其假定新的数据将保持较高的一致性。

在接下来的几年中，您将看到更多可用的机器学习模型。随着这些模型不断使用新数据更新，模型能作出更加准确的预测分析。然而，偏好和趋势也会发生改变，但离线模型无法随着这些新数据的改变而作出调整。例如，机器学习模型基于客户流失的可能性作出预测。模型在部署时可能很准确，但随着更加灵活的新竞争对手出现，客户有了更多选择，流失的可能性也就会增大。但由于原始模型使用较老的数据进行训练，此时市场还没有新进入者，因此这个模型无法再为企业提供准确的预测。另一方面，如果该模型为在线模型，并基于新数据不断训练，其对流失的预测就会更准确。即使客户偏好变化，市场格局改变，也是如此。

机器学习即服务将盛行

随着模型和算法不断推动着机器学习的成熟，您将注意到机器学习即服务 (MLaaS) 越来越受欢迎。MLaaS 描述了通过云进行交付的各类机器学习功能。MLaaS 市场中的供应商提供诸如图像识别、语音识别、数据可视化和深度学习等工具。用户通常上传数据至供应商的云，之后进行机器学习计算。

警告

将大量数据集迁移至云所面临的困难包括网络成本、合规和监管风险以及性能。然而，借助云服务，企业能够使用机器学习，省去了与采购硬件有关的前期时间和成本。

此外，MLaaS 降低了与机器学习有关的复杂度。例如，团队可以使用自然语言处理 (NLP)——一种用于解读文本或识别图像的工具，以实现人类和机器之间的对话。NLP 和图像识别都非常适合应用于擅长处理特定计算密集型任务的云服务。许多模型的训练和迭代过程的性能差异尤其显著。大型图形处理器 (GPU) 专为提高图像的渲染速度而设计，以便大大减少周期时间。

NLP 的成熟

我们期望在下一个十年中，NLP 将更加成熟，成为用户通过书面或口头接口与系统交流的规范。NLP 是能让机器理解人类的口头或书面语言的结构和意义的技术。此外，NLP 技术能让机器输出人类能够理解的口头语言信息。研究人员已专注于 NLP 技术几十年，机器学习正帮助加快 NLP 系统的实施。目前，机器要理解单词和句子的上下文非常困难。通过将机器学习应用于 NLP，系统能够了解单词和句子的语境和意义。以句子“A bat flew toward the crowd”为例。这个句子可能是指击球员不小心放开的棒球棒或者飞行的哺乳动物朝着一群人飞去。要理解这句话的意思，系统需要理解这句话的语境。

自动化的发展将简化机器学习流水线

自动化机器学习过程将使缺乏技术的员工也能享受机器学习的便利。此外，借助自动化，技术用户将能够集中精力攻克艰巨任务，而不仅仅是自动化重复的任务。机器学习中存在许多冗长乏味的重要细节，这些细节具备自动化的条件（例如数据清理）。数据可视化是自动化帮助简化机器学习流程的另一个领域。系统可基于所提供的数据集选择最合适的可视化，使得理解数据点之间的关系不再困难。

专业硬件将提升机器学习的性能

我们将步入买得起成熟硬件的时代。因此，不少企业能够购买足够强大的硬件以快速处理机器学习算法。此外，这些强大的硬件消除了机器学习的处理瓶颈，因此能够将机器学习嵌入更多的应用领域。

一般而言，CPU 可用于支持深度学习训练流程，并产生不同的结果。这些 CPU 因其在类神经网络中处理步骤的繁琐方式可能会导致出错。相比而言，GPU 拥有数百种能够实现数千个并行硬件线程的更简单的核心。由于 GPU 在深度学习应用领域的重要性，该技术目前出现大量研究，以提供更强大的芯片。云计算供应商也看到了 GPU 的价值，越来越多的供应商开始在云上提供 GPU 环境。

除了 GPU，研究人员正使用现场可编程门阵列 (FPGA) 来成功运行机器学习工作。当运行类神经网络和深度学习操作时，FPGA 有时比 GPU 的性能更高。

自动化算法选择和算法测试

一般而言，数据科学家需要了解如何使用数十种特定的机器学习算法。在第 3 章中，我们探讨了机器学习算法的主要类型。为处理不同类型的数据或解决您试图回答的不同类型的问题，我们会用到各类算法。

选择正确的算法创建机器学习模型并非总是一蹴而就的。数据科学家在找到能创建最佳模型的算法之前，可能尝试过若干不同的算法。这一过程需要花费时间，对专业知识的要求也很高。自动化正应用于帮助加快算法选择。借助自动化，数据科学家能够快速找到较合适的一两种算法，无需手动测试更多算法。此外，自动化还有助于经验不那么丰富的开发人员和分析师轻松使用机器学习算法。

透明度和信任必不可少

只有了解了机器学习推荐特定结果的方式和原理，才能相信这些结果。用于医学图像扫描的深度学习模型可能标记出可能发生癌细胞生长的图像。然而，单纯识别出图像远远不够。医生需要了解机器学习模型认为生长是癌性的原因。经过分析哪些信息才使模型得出这一诊断结论？医生必须确定结果能得到数据的证实。

机器学习将成为端对端流程

步入机器学习商业化的时代，我们将从开发和运营的角度将机器学习视为端对端流程。这意味着，该流程包括发现解决复杂问题的正确数据，确保数据不断得到适当的训练、模型化和管理。机器学习的生命周期很重要，因为其关系重大。机器学习模型可以成为预测未来的强大工具。

机器学习能为您创造未来

新的竞争者不断出现，客户期望达到前所未有的高度。您的公司和声誉要求您顺应新趋势，预测客户需求。机器学习技术贯穿企业的各个领域，可用于提高性能、提升客户满意度、减少客户流失和增加收入。在本书中，您会了解到何为机器学习，如何在您的公司中应用机器学习，以及机器学习能为您的公司带来哪些好处。

内容一览

- 什么是机器学习?
- 解释商业需要
- 机器学习的重要算法
- 您的数据科学团队需要的技能
- 企业如何利用机器学习
- 机器学习的前景

IBM

Learn more at:
IBM.com/machinelearning

IBM.com/datascience

Hurwitz & Associates 的总裁 **Judith Hurwitz** 是一名顾问和思想领袖。Hurwitz & Associates 的原理分析师 **Daniel Kirsch** 是机器学习、云、安全领域的一名研究员和顾问。

Go to Dummies.com®
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-53567-6
Part #: IMM14209USEN-00
Not for resale

for dummies®

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.