特征工程训练营

[美] 希南·奥兹德米尔(Sinan Ozdemir) 著

殷海英 译

清華大學出版社

北 京

北京市版权局著作权合同登记号　图字：01-2024-2611

Sinan Ozdemir
Feature Engineering Bookcamp
EISBN: 978-1-61729-979-7

图书在版编目(CIP)数据

特征工程训练营 /(美) 希南·奥兹德米尔(Sinan Ozdemir)著 ；殷海英译. -- 北京 ：清华大学出版社, 2024. 8. -- ISBN 978-7-302-66909-8

Ⅰ. TP181

中国国家版本馆 CIP 数据核字第 2024TV7702 号

责任编辑：王　军
封面设计：孔祥峰
版式设计：思创景点
责任校对：马遥遥
责任印制：刘　菲

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市人民印务有限公司
经　　销：全国新华书店
开　　本：148mm×210mm　　**印　　张**：10.125　　**字　　数**：282 千字
版　　次：2024 年 9 月第 1 版　　**印　　次**：2024 年 9 月第 1 次印刷
定　　价：69.80 元

产品编号：101492-01

作 者 简 介

Sinan Ozdemir 是 Shiba 公司的创始人兼首席技术官(CTO)，目前负责管理支持公司社交商务平台的 Web3 组件和机器学习模型。Sinan 曾是约翰·霍普金斯大学的数据科学讲师，是多种关于数据科学和机器学习的教材的作者。此外，他是已被收购的 Kylie.ai 公司的创始人，该公司开发了具备 RPA(机器人流程自动化)功能的企业级对话式人工智能平台。Sinan 拥有约翰·霍普金斯大学纯数学(pure mathematics)专业硕士学位，目前居住在加利福尼亚州的旧金山市。

致　　谢

本书需要大量人士的辛勤工作；我坚信，所有付出成就了一部卓越之作。我真心希望你也有同感！我想要感谢很多人，因为他们的鼓励和帮助，支持我走到了今天。

首先，我要感谢我的爱人 Elizabeth。你一直支持我，当我围着厨房踱步，试图为复杂的主题找到最佳类比时，你耐心地听我讲述；轮到我遛狗时，你也会帮我遛；我太专注于写作了，完全忘记了这件事。我对你的爱胜过一切。

接下来，要感谢 Manning 的整个团队，因为是你们使得本书得以出版。我明白整个出版过程确实花费了一些时间，但正是你们的不断支持让我能走出困境。你们对本书质量的承诺让每个阅读它的人都受益匪浅。

我还想感谢所有在本书不同阶段阅读了我的稿件并给出建议的审稿人。感谢 Aleksei Agarkov、Alexander Klyanchin、Amaresh Rajasekharan、Bhagvan Kommadi、Bob Quintus、Harveen Singh、Igor Dudchenko、Jim Amrhein、Jiri Pik、John Williams、Joshua A. McAdams、Krzysztof Jędrzejewski、Krzysztof Kamyczek、Lavanya Mysuru Krishnamurthy、Lokesh Kumar、Maria Ana、Maxim Volgin、Mikael Dautrey、Oliver Korten、Prashant Nair、Richard Vaughan、Sadhana Ganapathiraju、Satej Kumar Sahu、Seongjin Kim、Sergio Govoni、Shaksham Kapoor、Shweta Mohan Joshi、Subhash Talluri、Swapna Yeleswarapu 和 Vishwesh Ravi Shrimaland，你们的建议使这

本书变得更好。

最后，我要特别感谢技术校对人员，他们督促我检查了所有细节，并审阅了我的代码！

总之，许多人的努力使本书成为可能。感谢所有参与者！

自　序

与许多数据科学家和机器学习工程师一样，我的职业培训和学习主要源于实际经验，而非传统的学术教育。我在约翰·霍普金斯大学攻读纯数学专业，并未系统学习过回归和分类模型。在获得硕士学位后，我决定从攻读博士转向加入硅谷的初创公司，自学机器学习和人工智能的基础知识。

我阅读免费的在线资源和参考书籍，开始了我的数据科学学习之旅，并成立了一家专注于打造企业级 AI 的公司。我接触到的几乎所有资料都集中在用于对“数据”和“预测”建模的模型和算法类型。我通过阅读书籍学习理论知识，并在类似 Medium 的网站上阅读文章，了解人们如何将理论应用于实际生活。

直到几年后，我才开始意识到，在学习和了解模型、训练和参数调整等主题方面，我只能走这么远。当时我正在处理原始文本数据，构建企业级的聊天机器人，我注意到 NLP(自然语言处理)方面的书籍和文章在内容上的巨大差异。它们确实介绍了很多我可以使用的分类和回归模型，但更关注如何处理原始文本，以便让模型能够很好地使用这些数据。它们更多地讨论了如何调整数据参数，而不是调整模型本身的参数。

为什么人们不对表格数据进行与文本数据相同的严格处理呢？原因不可能是它不必要或者没有帮助，因为几乎所有关于数据科学流程所花费时间的调查都显示，人们将大部分时间用在获取和清洗数据上。我决定填补这个空白，并将其变成一本书。

在写本书的几年前，我撰写了另一本关于特征工程的书。我的第一本特征工程书聚焦于基础特征工程，强调解释工具和算法，而

非展示它们如何在日常应用中使用。而本书采用了更贴近实际的方式。每一章都专注于特定领域的一个应用案例，配有一个数据集，以应用不同的特征工程技术。

我努力以简单易懂、简洁明了的形式勾勒我在特征工程方面的思考过程。在我的数据科学和机器学习职业生涯中，特征工程一直是其中的一个重要组成部分。我希望本书能够让你更深入地了解数据处理，并成为你与同事在工作内容方面的谈资。同时，本书也为你提供工具和窍门，帮助你明确在何时应用哪些特征工程技术。

前　言

本书旨在介绍流行的特征工程技术，讨论何时以及如何运用这些技术的框架。我发现，有些书籍只关注其中一方面，有时可能显得有些单薄。专注于概述的书籍往往忽略了实际应用的一面，而专注于框架的书籍可能让读者产生疑问：“为什么这样做有效呢？”我希望读者在理解和应用这些技术方面都能充满信心。

本书目标读者

本书面向已经踏入机器学习领域并寻求提升能力与技能的机器学习工程师和数据科学家。假设读者已经掌握机器学习、交叉验证、参数调优以及使用 Python 和 scikit-learn 进行模型训练的基础知识。本书在此基础上进一步拓展，将特征工程流程直接融入现有的机器学习框架中，以提供更深入的学习体验。

本书的学习路线图

本书包含两个导论性章节(第 1～2 章)，涵盖了特征工程的基础知识，包括如何识别不同类型的数据以及特征工程的不同类别。第 3～8 章的每一章都专注于一个具体的案例研究，使用不同的数据集和目标。每章都为读者提供一个新的视角、一个新的数据集以及特定于我们处理的数据类型的新的特征工程技术。本书的目标是提供

关于特征工程技术种类的广泛而全面的知识，同时展示各种数据集和数据类型。

关于代码

本书涵盖了许多源代码示例，它们以编号的代码清单和正常文本行的形式呈现。在两种情况下，源代码都采用等宽字体的格式，以便与普通文本区分开来。有时，代码也以粗体显示，用于突出显示在相应章中与之前步骤不同的代码，例如当新特性添加到现有代码行时。

许多情况下，源代码经过重新格式化；我们添加了换行符并重新调整了缩进，以适应书中可用的页面空间。某些情况下，这样做仍不够，代码清单中会包含续行标记(➥)。代码清单中附带了许多注释，用于突出显示重要的概念。

可扫描封底二维码下载代码。

关于本书封面

本书封面上的图案标题为“Homme du Thibet”，即“来自中国西藏的人”，摘自 Jacques Grasset de Saint-Sauveur 于 1797 年出版的作品集。每幅插图都经过精心手绘和上色。

在那些日子里，很容易通过人们的穿着来确定他们的居住区域，他们的职业或社会地位。Manning 出版社通过几个世纪前代表地区文化丰富多样性的插图来表现电脑行业的开创性，并让这些珍贵的插图重新焕发生机和光彩。

目　录

第 1 章

特征工程简介

本章主要内容：

- 理解特征工程和机器学习流程
- 探讨特征工程在机器学习过程中的重要性
- 了解特征工程的类型
- 了解本书的结构以及我们将关注的案例研究类型

当前围绕人工智能(AI)和机器学习(ML)展开的许多讨论往往天生以模型为中心，聚焦于 ML 和深度学习(DL)的最新进展。这种模型优先的方法往往对用于训练这些模型的数据关注不足，甚至完全忽视。类似 MLOps 的领域正迅速发展，通过系统性地训练和利用 ML 模型，尽量减少人为干预，以“释放”工程师的时间。

许多知名的 AI 专家正在敦促数据科学家更关注以数据为中心的机器学习视角，而不是过于关注模型选择和超参数调整过程。这种视角更侧重于提高我们所摄取并用于训练模型的数据质量。Andrew Ng 曾公开表示：“机器学习基本上就是特征工程”，我们需要更加倾向于以数据为中心的方法。在以下场景中，采用以数据为中心的方法

尤为有效：

- 数据集的观察值较少(小于 10KB)，因此我们可以从更少的行中提取尽可能多的信息。
- 数据集的列数相对于观察值的数量很大。这可能导致所谓的维度诅咒，这是一种描述数据空间极度稀疏的现象，使得机器学习模型难以从中学习。
- 数据和模型的可解释性至关重要。
- 数据领域本质上是复杂的(例如，在数据不完整和不干净的情况下，几乎无法实现精确的金融建模)。

我们应该将注意力集中在机器学习流程中最需要细致和谨慎考虑的部分：特征工程。

在本书中，我们将深入研究用于识别最强特征、创建新特征的不同算法和统计测试程序。在我们的语境中，将特征定义为对 ML 模型有实际意义的数据属性或列。我们将通过几个案例来展开深入研究，每个案例研究都属于不同领域，包括医疗保健和金融，并将涉及多种类型的数据，如表格数据、文本数据、图像数据和时间序列数据。

1.1 特征工程是什么，为什么它如此重要

对于不同的数据科学家，“特征工程”这个术语可能唤起不同的联想。对于一些数据科学家来说，特征工程是我们缩小监督模型所需特征范围的手段(例如，试图预测响应或结果变量)。对于其他人来说，它是从非结构化数据中提取数值表示，以用于非监督模型的方法学(例如，试图从先前非结构化的数据集内提取结构化数据)。特征工程不仅包括这些，还涵盖了更多内容。

在本书中，特征工程是一门艺术，其目的在于对数据进行操纵和转换，使其以最佳方式呈现出 ML 算法试图建模的底层问题，并减少数据中固有的复杂性和偏差。

数据实践者通常依赖机器学习和深度学习算法从数据中提取和学习模式，即使他们使用的数据格式不佳和非最优。原因包括实践者过于信任他们的机器学习模型，或者根本不知道处理混乱和不一致数据的最佳实践，希望机器学习模型能够为他们“解决问题”。这种做法甚至没有给机器学习模型从适当的数据中学习的机会，从一开始就注定了数据科学家的失败。

关键在于数据科学家是否愿意或能够尽可能充分利用他们的数据(为机器学习任务精心设计最佳特征)。如果我们不进行合适的特征工程，而依赖复杂而缓慢的机器学习模型来解决问题，可能最终得到性能较差的机器学习模型。相反，如果我们花时间了解数据，并为机器学习模型制定精心设计的特征，我们就可能得到更小、更迅速的模型，其性能与较大模型相当，甚至更出色。

我们期望机器学习模型在我们选择的评价指标上表现尽可能出色。为达成这一目标，可对数据和模型进行调整(见图 1-1)。

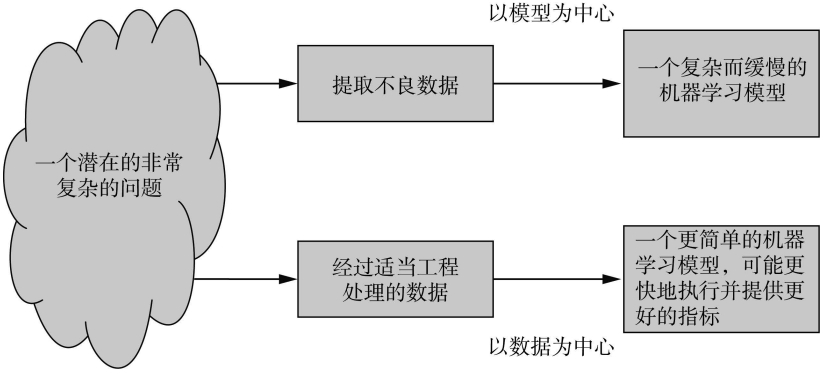


图 1-1　在数据为中心的机器学习方法中，我们并不太关心改进机器学习代码，而是关心以某种方式处理填充数据，使得机器学习模型更容易发现和利用数据中的模式，从而在整个流程中实现更好的性能

本书的焦点不在于如何优化机器学习模型，而在于转换和操纵数据的技术，使得机器学习模型更容易处理和学习数据集。我们将展示一系列特征工程技术，这些技术可帮助整个机器学习流程，而

不仅是选择具有更好超参数的模型。

1.1.1 谁需要特征工程

根据 Anaconda 在 2020 年进行的“数据科学现状”调查(参见 https://www.anaconda.com/state-of-data-science-2020)，数据整理(我们可以将其视为特征工程的代名词，额外增加了数据加载步骤)占用的时间过长，因此数据整理是每位数据科学家都关心的问题。调查显示数据管理仍然占据了数据科学家大量的时间。报告中显示有将近一半的时间用于数据加载和“整理”。报告声称这一情况“令人失望”且“数据准备和整理挤占了宝贵的、真正的数据科学工作时间”。需要注意，“整理”数据是一个相当模糊的术语，很可能被用作对探索性数据分析以及所有特征工程工作的概括性描述。我们认为数据准备和特征工程是数据科学家工作中真实、重要且几乎总是不可避免的部分，应该受到与专注于数据建模的流程部分同等的重视。

本书致力于展示强大的特征工程过程，包括模型公平性评估(见第 4 章)、基于深度学习的表示学习(见第 5 章)、假设检验(见第 3 章)等。这些特征工程技术对模型性能的影响不亚于模型选择和训练过程。

1.1.2 特征工程的局限性

值得强调的是，良好的特征工程并非灵丹妙药。例如，特征工程无法解决机器学习模型数据量过少的问题。虽然并没有确定数据规模何时算过小的确切阈值，但大多数情况下，当处理少于 1000 行的数据集时，特征工程只能尽力从这些观测中提取尽可能多的信息。当然，也存在例外情况。在我们的自然语言处理和图像案例研究中，当涉及迁移学习时，我们将看到预训练的机器学习模型如何从仅有的几百个观测中学习，但这也仅是因为它们已经在数十万个观测上进行了预训练。

特征工程也无法在特征与响应变量之间本来就没有联系的情况下创造这种关联。如果最初的特征并不隐含地具有对响应变量的预测能力，那么再多的特征工程都无法建立这种关系。我们可能能够在性能上取得小幅提升，但不能指望特征工程或机器学习模型会神奇地为我们创造特征与响应变量之间的关系。

1.1.3　出色的数据，出色的模型

出色的模型离不开出色的数据。如果没有深刻反映问题本质的良好结构化数据，几乎不可能确保获得准确且公平的模型。

我在机器学习领域的职业生涯中，大部分时间都在研究自然语言处理(NLP)。具体来说，我专注于构建能够自动从非结构化的历史记录和知识库中提取和优化对话 AI 架构的机器学习流程。在初期，我主要关注从原始的人类对话记录中提炼和实施知识图谱，并利用最先进的迁移学习和序列到序列模型，开发出能够学习新主题的对话 AI 流程，以适应不断更新的话题。

我最近结识了一位名叫 Lauren Senna 的对话架构设计师和语言学家。她向我介绍了在对话中使用的深层结构，这些结构是她和她的团队用来构建能够在一周内胜过我自动推导的所有机器人的关键。Lauren 分析了关于人们与机器人交流互动的心理，以及为什么这与知识库文章的写作方式不同。那时我终于意识到，我需要花更多时间将机器学习努力集中在预处理上，以展现这些潜在的模式和结构，使预测系统能抓住它们并变得比以往更准确。她和我在某些情况下负责提高机器人性能，并且我们获得了高达 50%的性能提升；我在各种会议上分享了数据科学家如何利用类似技术来解锁他们自己数据中的模式。

若不了解和尊重数据，我永远无法发挥模型的潜力，这些模型致力于捕捉、学习并放大数据中蕴含的模式。

1.2 特征工程流程

在深入研究特征工程流程之前，我们需要回顾一下整个机器学习流程。这是很重要的，特征工程流程本身是更大机器学习流程的一部分，因此这将为我们提供理解特征工程步骤所需的全局视角。

机器学习流程

机器学习流程通常包括五个步骤(图 1-2)：

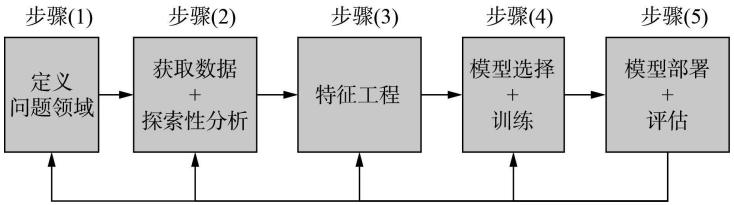


图 1-2 机器学习流程。从左到右：我们必须深入了解问题领域，获取并深入理解数据，进行特征工程(显然是本书的主要关注点)，选择并训练模型，然后在部署模型时要明白，如果模型评估显示任何形式的数据或概念漂移，我们可能需要回溯到过去的每一步，因为这可能是模型衰减的表现——随着时间推移，机器学习模型性能下降

(1) 定义问题领域——我们试图通过机器学习解决何种问题？这是定义我们要优先考虑的特征的时刻，例如模型预测速度或可解释性。这些考虑事项在进行模型评估时将变得至关重要。

(2) 获取数据并进行探索性分析——考虑并实施收集数据的方法，确保数据的公平性、安全性，并尊重数据提供者的隐私。这也是执行探索性数据分析(EDA)的绝佳时机，以对我们正在处理的数据有深刻了解。假设你已经对数据进行了充分的 EDA 工作，而我将在本书中尽力帮助你更全面地了解数据。如果这是一个监督学习问题，我们是否需要处理类别不平衡的情况？如果这是一个无监督学习问题，我们是否有足够代表总体的数据样本，以获得足够深刻

的洞察？

(3) 特征工程——这是本书的重点，也是机器学习流程的关键步骤。这一步创建能输入 ML 模型的数据的最优表示。

(4) 模型选择和训练——这是机器学习流程的重要组成部分，应该认真而谨慎地执行。在这个阶段，我们选择最适合数据和步骤(1)中考虑事项的模型。如果模型的可解释性被强调为首要考虑因素，或许我们会选择树模型而不是深度学习驱动的模型。

(5) 模型部署和评估——在这个阶段，数据已经准备好，模型已经训练好，现在是将模型投入生产的时刻。此时，数据科学家可考虑模型版本控制和预测速度(作为评估模型就绪情况的因素)。例如，是否需要某种用户界面以同步获取预测，还是我们可以脱机执行预测？必须使用评估过程来追踪模型的性能，并密切关注模型的衰减。

提示 谈到问题领域时，成为在该领域解决问题的数据科学家并不需要成为该领域的专家。话虽如此，我强烈建议你至少与该领域的专家联系并进行一些研究。

在机器学习流程的最后一步，我们还需要留意概念漂移(当对数据的解释发生变化)和数据漂移(当数据的基本分布发生变化)；这两个概念指的是数据随时间可能发生变化的情况。

概念漂移是指特征或响应的统计特性随时间发生变化的现象。如果在某个时间点在一个数据集上训练一个模型，我们就定义了一个函数的快照，这个函数将特征与响应关联起来。随着时间的推移，数据表示的环境可能发生变化，我们对这些特征和响应的看法也可能发生改变。这个概念最常应用于响应变量，但也可以用于特征。

假设我们是流媒体平台的数据科学家。我们的任务是构建一个模型，预测何时向用户展示一个进度限制，并询问他们是否仍在观看。可基于用户按下按钮后的分钟数或他们当前观看的节目的平均长度等指标构建一个基本模型，我们的响应将是一个简单的 True 或 False，

表示是否应该显示进度限制。在模型创建时，我们的团队齐聚一堂，并作为领域专家，考虑了展示这个进度限制的所有可能方式。也许观看视频的人睡着了，也许他们因为办事而不小心离开了。因此，我们构建了一个模型并部署了它。两个月后，我们开始收到延迟显示进度限制的请求，我们的团队再次汇聚在一起来阅读这些请求。原来，有一大群人(包括本文作者)使用流媒体应用为他们的狗和猫播放宁静的纪录片，以帮助宠物缓解长时间离开主人时的分离焦虑。这是我们的模型没有考虑到的概念。现在，必须添加观察和特征，如“节目是否为动物类节目”，以帮助解释这个新概念。

数据漂移是指由于某种原因我们数据的基础分布发生了变化，但我们对该特征的解释仍然保持不变。这在发生模型未考虑到的行为变化时很常见。考虑一下过去的流媒体平台。在 2019 年底，我们构建了一个模型，以预测某人观看节目的小时数，考虑了他们过去的观看习惯、喜欢的节目类型等变量，效果良好。突然间，一场全球性的疫情暴发，一些人开始更频繁地在线观看视频，甚至可能在工作时进行观看，以制造出即使独自一人在家中，也仿佛有人在身边的感觉。响应变量的分布(以观看的小时数衡量)将戏剧性地向右偏移，考虑到这种分布变化，模型可能无法保持过去的性能水平。这就是数据漂移。观看小时数的概念没有改变，但该响应的基础分布发生了变化。

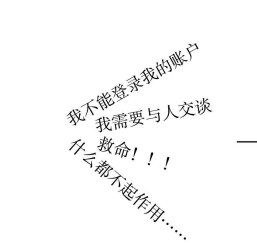
这个理念同样适用于特征。如果我们针对新的响应变量“如果我们提供下一集，这个人会观看吗？”并将观看时长作为特征，那么模型以前未遇到的这种分布的戏剧性变化依然成立。

如果我们放大机器学习流程的中间部分，会看到特征工程。特征工程作为较大机器学习流程的一部分，可被看作拥有独立步骤的独立流程。如果我们双击并打开机器学习流程中的特征工程部分，将看到以下步骤。

(1) 特征理解：识别我们正在处理的数据级别至关重要，这将影响我们可以使用哪些类型的特征工程。在这个阶段，我们将不得

不完成诸如“确定我们的数据属于哪个级别”的工作。别担心，我们将在下一章详细介绍数据的级别。

(2) 特征结构化：假设我们的数据中存在一些非结构化的数据(如文本、图像、视频等；见图 1-3)。



I	Can't	...	Help
0	1	...	0
0	0	...	9
2	0	...	1
5	0	...	1
0	1	...	1

特征 1 (例如，戴着王冠)	特征2 (例如，正在躺下)	···	特征*n*
0.324	1	...	0
0	0	...	9.234
2	0	...	0.421
8.4	0	...	0.961
0	1	...	1

图 1-3　原始数据(如文本、音频、图像和视频)只有经过转换，变成数值向量的形式，才能被机器学习算法处理。我们将这个过程称为特征结构化，可以通过提取技术来实现，比如应用词袋算法或采用自动编码器等非参数特征学习方法(第 5 章的案例研究将详细介绍词袋算法和自动编码器的应用)

我们必须将它们转换为结构化格式，以便我们的机器学习模型可以理解它们。例如，将文本片段转换为向量表示或将图像转换为矩阵形式。可使用特征提取或特征学习来实现这个目标。

(3) 特征优化：为数据建立了结构化表示后，可应用各种优化技术，如特征改进、提取、构建和选择，以获取最适合模型的数据。日常特征工程工作的主要内容通常涉及这一方面。本书中的绝大多

数代码示例都将围绕特征优化展开。每个案例研究都将包含一些特征优化的实例，其中我们需要创建新的特征或者对现有特征进行加工，使其更适用于我们的机器学习模型。

(4) 特征评估：当我们修改特征工程流程以尝试不同的场景时，我们希望了解应用的特征工程技术是否有效。为了实现这一目标，我们可以选择一个单一的学习算法，可能还要选择一些参数选项进行快速调整。然后，可将不同特征工程流程的应用与一个恒定模型进行比较，以评估在个体变化的情况下哪些流程步骤表现更好。如果没有看到期望的性能，将返回到之前的优化和结构化步骤，尝试获得更好的数据表示(见图 1-4)。

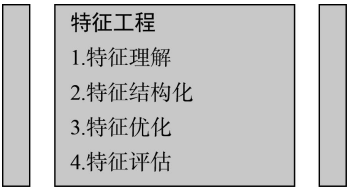


图 1-4 聚焦于机器学习流程的特征工程阶段，我们可以看到它所执行的步骤，以开发出正确且成功的特征工程流程

1.3 本书的编排方式

一本涵盖众多案例研究的书籍可能在编排上面临一些挑战。一方面，我们希望提供足够的背景和洞察，解释我们将要用于特征工程的技术。另一方面，我们也充分认识到通过实例和代码样本来巩固概念的重要性。

为此，我们将共同努力，以一种引人入胜的方式叙述每个案例研究，展示解决特定领域问题的端到端代码。我们将逐步解析代码，解释为什么那样做以及接下来我们将要做什么。我希望这样既能为读者呈现实际操作的代码，又能提供对问题的高层次思考，达到两者兼顾的效果。

1.3.1　特征工程的五种类型

本书的主要关注点是特征工程的五个主要类别。第 2 章将介绍这五个类别，并在整本书中不断回顾它们。

(1) 特征改进：通过数学变换提升现有特征的可用性。

例子：通过从其他列进行推断，填充天气数据集中的缺失温度值。

(2) 特征构建：通过从现有可解释特征中创造新的可解释特征来丰富数据集。

例子：在房屋估值数据集中，通过将房屋总价特征除以房屋面积特征，创建一个每平方英尺价格的特征。

(3) 特征选择：从现有特征集合中选择最佳的特征子集。

例子：在创建每平方英尺价格特征后，如果之前的两个特征对机器学习模型没有更多价值，可能会将它们移除。

(4) 特征提取：依赖算法自动创建新的、有时是不可解释的特征，通常基于对数据进行参数化的假设。

例子：依赖预训练的迁移学习模型，如 Google 的 BERT，将非结构化文本映射到结构化且通常是不可解释的向量空间。

(5) 特征学习：通常利用深度学习，通过从原始的非结构化数据(如文本、图像和视频)中提取结构和学习表示，自动生成全新的特征集。

例子：训练生成对抗网络(GAN)以解构和重构图像，以便学习针对特定任务的最优表示。

现在，应该注意两件事情。首先，无论我们是使用监督学习模型还是无监督学习模型，都没有关系。这是因为我们所定义的特征是对机器学习模型有意义的属性。因此，无论目标是将观测结果聚类在一起还是预测股票在几小时内的价格变动，我们设计特征的方式都将迥然不同。其次，通常人们会在数据上执行与特征工程一致的操作，但并不打算将数据馈送到机器学习模型中。例如，有人可能希望将文本向量化为词袋表示，以创建词云可视化，或者一家公司可能需要填补客户数据中的缺失值以计算流失统计数据。这当然

是有效的，但它并不符合特征工程在机器学习中的相对严格的定义。

如果我们看一下特征工程的四个步骤以及五种特征工程如何融入其中，我们将得到如图 1-5 所示的流程图，其中展示了一个端到端的流程，说明了如何为工程化特征摄取和操作数据，从而更好地帮助机器学习模型解决手头的问题。

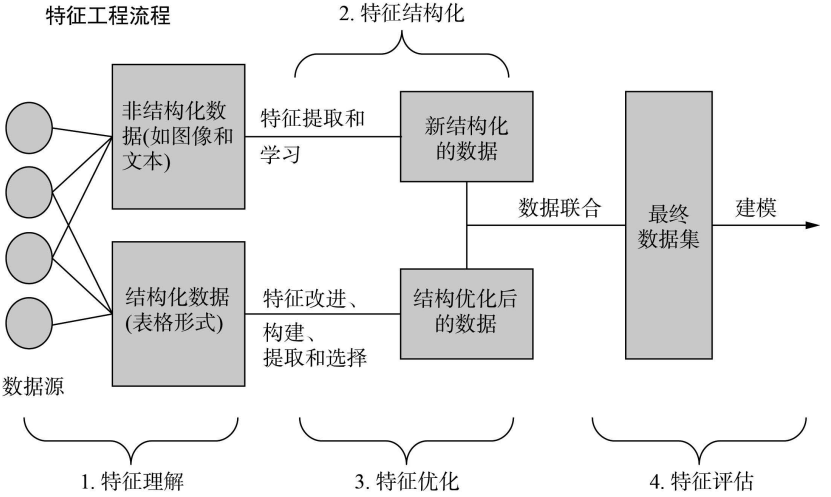


图 1-5 特征工程流程有四个阶段，包括理解我们的数据、对数据进行结构化和优化，然后使用机器学习模型对数据进行评估。注意，将原始结构化数据和新结构化数据合并的数据联合是可选的，由数据科学家和手头的任务决定

1.3.2 本书案例研究的概述

本书旨在展示日益复杂的特征工程过程，这些过程是分步构建的，并通过示例、代码样本和案例研究提供使用这些过程的基础知识。本书的前几个案例研究主要关注核心的特征工程过程，这是所有数据科学家都应该掌握的，且几乎适用于所有数据集。随着我们在本书中呈现更多案例研究，将使用更先进的技术，并更聚焦于不同类型的数据。

你可以自由地直接跳转到任何使用你想要掌握的特征工程技术的具体案例研究，并立即开始学习。本书包括六个案例研究，涵盖不同领域，使用各种数据类型。每个案例研究都会在前一个基础上引入越来越先进的特征工程技术。

我们的第一个案例研究是医疗保健/COVID-19 诊断，其中将使用与全球 COVID-19 大流行相关的结构化数据。在这个案例研究中，将尝试对 COVID-19 进行预测性诊断，利用表格形式的结构化数据。将深入了解数据的不同层次：特征改进、特征构建和特征选择。

第二个案例研究是公平性，强调偏见和伦理。这个案例研究专注于超越传统的机器学习度量标准，深入探讨在人们的正当利益受损时，盲目追随算法建议可能带来的危害。我们将研究如何通过引入不同的公平定义并识别数据中的受保护特征来保护模型免受数据中固有偏见的影响。与缓解数据中的偏见相关的特征选择和特征构建将发挥关键作用。

第三个案例将关注 NLP/分类推文情感，我们将开始看到更先进的特征工程技术(如特征提取和特征学习)的实际应用。这里的问题陈述相对简单：推文的作者是快乐的、中立的还是不快乐的？我们将深入研究传统的参数化特征提取方法(如主成分分析)，以及更现代的特征学习方法(如迁移学习和自编码器)，并对这些方法进行比较。

第四个案例将深入研究图像/物体识别。我们将使用两个不同的图像数据集，尝试教会模型识别各种物体。我们将再次看到传统的参数化特征提取方法(如梯度方向直方图)与现代的特征学习方法(如生成对抗网络)之间的较量，以及不同的特征工程技术在模型性能和可解释性之间的权衡。

第五个是时间序列/短线交易案例研究，我们将寻找 alpha(跑赢指数)，并尝试部署深度学习来执行最基本的短线交易问题：在接下来的几小时内，这只股票价格会显著下跌、上升或保持相对稳定？这似乎很简单，但涉及股市时，没有什么是简单的。在这个案例研究中，时间序列技术将占据主导地位，特征选择、改进、构建和提

取也发挥了作用。

第六个案例研究将走上一条风景优美但常被忽视的小路。使用 Flask 进行特征存储/流数据的研究，我们将探讨如何将特征工程技术部署到 Flask 服务中，以使特征工程工作更高效，并广泛地提供给更多的工程师群体。我们将在 Flask 中设置一个 Web 服务，创建一个特征存储，用于存储和提供来自短线交易案例研究的实时数据。

在每一个案例研究中，将遵循相同的学习模式：

(1) 将介绍数据集，通常伴随一个简短的数据探索分析步骤，以帮助我们了解原始数据集。

(2) 然后将制定问题陈述，以帮助我们了解哪些特征工程技术将是适用的。

(3) 接下来将按照特征工程的类型分组，实施特征工程过程。

(4) 代码块和视觉元素将贯穿整个流程，帮助我们更清晰地了解特征工程技术对机器学习模型的影响。

(5) 我们将以“本章小结”结束，以概括每个案例研究的要点。

1.4 本章小结

- 作为机器学习流程的一部分，特征工程是对数据进行转化以提高机器学习性能的艺术。
- 当前有关机器学习的讨论主要以模型为中心。更应该关注以数据为中心的机器学习方法。
- 特征工程的四个步骤包括特征理解、特征结构化、特征优化和特征评估。
  - 特征理解——为了更好地解释数据。
  - 特征结构化——为了在机器学习中有效组织数据。
  - 特征优化——为了尽可能地从数据中提取信号和模式。
  - 特征评估——根据机器学习调整特征工程。
- 数据科学家将超过一半的时间都花费在整理和操作数据

上；值得花费充分的时间来整理数据集，以使所有下游任务更加轻松和有效。

- 优秀的特征工程能够产生更高效的数据集，使我们能够采用更快速、更小的模型，而不是依赖于通过混乱数据训练出来的缓慢而复杂的模型。
- 本书提供了许多案例研究，帮助读者真正学习和运用特征工程技术。