Packt

Hands-On Automated Machine Learning

自动机器学习入门与实践：使用Python

A beginner's guide to building automated machine learning systems using AutoML and Python

[美] Sibanjan Das　Umit Mert Cakmak 著　谢琼娟 译　马勇 审校

華中科技大學出版社
http://www.hustp.com

Hands-On Automated Machine Learning

自动机器学习入门与实践：使用Python

[美] Sibanjan Das　Umit Mert Cakmak 著
谢琼娟 译　　马勇 审校

华中科技大学出版社
中国 · 武汉

内容简介

AutoML可以将部分机器学习过程自动化，减轻数据科学从业者的工作负担，深受高级分析人员的喜爱。本书介绍搭建AutoML模块的基础知识，并通过练习帮助读者消化这些知识。读者将学习使用机器学习流水线自动实现数据预处理、特征选择、模型训练、模型优化等任务，学习应用auto-sklearn和MLBox等已有的自动化库，并且创建和扩展自定义的AutoML环节。阅读本书，你将对AutoML有更清晰的认识，能利用真实数据集完成自动化任务。书中知识可运用到实际的机器学习项目中，或者在机器学习竞赛中助你一臂之力。

图书在版编目(CIP)数据

自动机器学习入门与实践 : 使用Python / (美) 西班扬 · 达斯, (美) 乌米特 · 卡卡马克著 ; 谢琼娟译. --武汉 : 华中科技大学出版社, 2019.12

ISBN 978-7-5680-4952-8

Ⅰ. ①自… Ⅱ. ①西… ②乌… ③谢… Ⅲ. ①软件工具 - 程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2019)第262461号

湖北省版权局著作权合同登记 图字：17-2019-267号

书　　名 自动机器学习入门与实践 : 使用Python
　　Zidong Jiqi Xuexi Rumen yu Shijian: Shiyong Python
作　　者 [美] Sibanjan Das　Umit Mert Cakmak
译　　者 谢琼娟
审　　校 马　勇

策划编辑 徐定翔
责任编辑 陈元玉
责任监印 徐　露

出版发行 华中科技大学出版社（中国 · 武汉）
　　武汉市东湖新技术开发区华工科技园（邮编 430223 电话027-81321913）
录　　排 武汉东橙品牌策划设计有限公司
印　　刷 湖北新华印务有限公司
开　　本 787mm × 960mm 1/16
印　　张 15.5
字　　数 383千字
版　　次 2019年12月第1版第1次印刷
定　　价 72.90元

本书若有印装质量问题, 请向出版社营销中心调换
全国免费服务热线400-6679-118竭诚为您服务

前言
Preface

亲爱的读者，欢迎来到自动**机器学习**（machine learning，ML）的世界。**自动机器学习**（automated ML，AutoML）的使命是将部分机器学习过程自动化。现成的 AutoML 工具可减轻数据科学从业者的工作负担，在高级分析领域中接受度很高。本书介绍搭建AutoML模块的基础知识，并通过实践促进读者快速吸收这些知识。

读者将学习使用机器学习流水线自动实现数据预处理、特征选择、模型训练、模型优化等任务。书中会讲解如何应用 auto-sklearn 和 MLBox 等已有自动化库，以及创建和扩展自定义 AutoML 学习组件。

阅读本书，你会对AutoML的各个方面有更清晰的认识，能利用真实数据集完成自动化任务。书中的知识可运用到机器学习项目实践中，或在机器学习竞赛中助你一臂之力。我们希望购买本书的读者都觉得物有所值，干货十足。

目标读者
Who This Book is For

本书尤其适合机器学习初学者（包括新晋数据科学家、数据分析师、机器学习爱好者）学习，同时也适合对搭建高速机器学习流水线感兴趣的机器学习工程师和数据专业人员阅读。

大纲
What This Book Covers

第 1 章　AutoML 简介。为理解 AutoML 打基础，介绍各种自动化学习库。

第 2 章　Python 机器学习简介。介绍机器学习概念，便于理解 AutoML 方法。

第 3 章　数据预处理。深入诠释各种数据预处理方法、自动化对象、如何自动化，也会介绍特征工具和 sklearn 预处理方法。

第 4 章　自动化算法选择。指出哪些算法适用于哪类数据集。介绍不同算法的计算难度和可扩展性，也会接触到一些依据训练和推理时间来确定使用哪种算法的方法。本章会演示 auto-sklearn，以及如何扩展引入新算法。

第 5 章　超参数优化。讲解自动化超参数优化的基础知识。

第 6 章　创建 AutoML 流水线。阐述如何将不同组件组合起来构建一个端到端的 AutoML 流水线。

第 7 章　深度学习探究。介绍诸多深度学习概念及其对 AutoML 的贡献。

第 8 章　机器学习和数据科学项目的重点。总结全文，并分享一些从多方面权衡 AutoML 复杂性和成本的信息。

充分利用本书
To Get the Most Out of This Book

阅读本书唯一需要准备的是对机器学习的求知欲。除此之外，如果你以前接触过 Python 编程和机器学习基础知识，则能更好地利用本书，但这并非必备前提。学习本书，请提前安装 Python 3.5 和 Jupyter Notebook。

若具体章节中有特别要求，则会在该章第一节中提出。

下载代码示例
Download the Example Code Files

你可通过账号登录 www.packtpub.com，下载本书的代码示例。如果你在其他渠道购买本书，可访问 www.packtpub.com/support 进行注册，通过邮件接收代码示例。

下载代码示例的步骤如下。

（1）在 www.packtpub.com 登录或注册。

（2）选择 Support 页签。

（3）点击 Code Downloads & Errata。

（4）在搜索框中输入本书名称，并根据提示进行操作。

下载代码示例后，使用以下工具最新版进行解压或提取。

- Windows 系统：WinRAR/7-Zip。
- Mac 系统：Zipeg/iZip/UnRarX。
- Linux 系统：7-Zip/PeaZip。

本书涉及的代码包已上传到 GitHub，地址：https://github.com/PacktPublishing/Hands-On-Automated-Machine-Learning。如果代码有更新，则会在当前 GitHub 仓库中进行更新。

在 https://github.com/PacktPublishing/还可找到丰富的其他书目和视频中用到的代码包。去看看吧！

下载彩色图片
Download the Color Images

本书提供 PDF 版，包含本书中用到的彩色版截图或图表，下载链接为：https://www.packtpub.com/sites/default/files/downloads/HandsOnAutomatedMachineLearning_ColorImages.pdf。

文本约定
Conventions Used

本书中用到了一些文本约定。

`CodeInText`（文本代码）：表示代码文本、数据库表格名称、文件夹名称、文件名、文件扩展名、路径、虚拟 URL、用户输入及 Twitter 账号。比如：“使用 `sklearn.preprocessing` 模块中的 `StandardScaler` 对 `statisfaction_level` 一栏的值进行标准化处理。”

代码片段如下：

```
{'algorithm': 'auto',
 'copy_x': True,
 'init': 'k-means++',
 'max_iter': 300,
 'n_clusters': 2,
 'n_init': 10,
 'n_jobs': 1,
 'precompute_distances': 'auto',
 'random_state': None,
 'tol': 0.0001,
 'verbose': 0}
```

命令行输入或输出如下：

```
pip install nltk
```

黑体：表示新术语、重要词汇，或在屏幕上看到的词语。比如，菜单或对话框中的内容会显示为**黑体**。例如：“出现一个 NLTK Downloader 下载弹窗。在 Identifier 中选择**全部**，等待安装完成。”

警告或重要的注释，用此图标表示。

提示或技巧，用此图标表示。

目录

Table of Contents

第1章

AutoML 简介

Introduction to AutoML

过去十年，科学和技术领域发生了翻天覆地的变化。2007 年，第一部 iPhone 面世，在那之前，手机都用物理键盘。触摸屏并不是最新的发明，因为 Apple 公司已经有过类似的原型产品，而且 IBM 公司早在 1994 年就推出了最早的智能手机 Simon 个人通信器（Simon personal communicator）。Apple 公司的想法是制造一部集听音乐、看视频、浏览网站和 GPS 导航等全部多媒体功能于一体的设备，第一代 iPhone 的计算能力已经足以支撑以上这些功能。技术发展到今天的程度，速度惊人。第一代 iPhone 诞生十年后，现在的 iPhone 除了基本功能，已经可以识别人脸，可以认出动物、车辆和食品等实物，还能理解自然语言，跟人对话。

现在又有了可打印器官的 3D 打印机、自动驾驶汽车、可编组飞行的无人机、基因编辑、可重复利用的火箭和会后空翻的机器人。这些都不再是科幻小说里读到的情景了，而是实实在在发生在现实生活中的例子。过去的想象，已变成如今的现实。大家甚至开始谈论**人工智能**（artificial intelligence，AI）对人类的威胁。许多前沿科学家，如斯蒂芬·霍金，都在警告人类可能被 AI 之类的生命体毁灭。

AI 和**机器学习**（machine learning，ML）在过去几年里成了最热门的话题。过去十多年里，机器学习算法取得诸多成果和重大进步，如 Google 公司的 AlphaGo 击败了全球第一的人类围棋手柯杰。这并不是机器学习算法第一次击败人类。有

些精细场景，比如识别不同动物的物种，机器算法常常比人类更厉害。

这种显著进步引起了商业界的浓厚兴趣。这些技术虽然听起来是独立的学术研究，但其实有着巨大的商业意义。各行业的企业都想要利用这些算法，努力适应不断变革的技术场景。大家都意识到，谁能把这些技术运用到业务中，谁就能拔得头筹。本章讲解什么是机器学习和 AutoML，内容包括以下几个方面。

- 机器学习的范围。
- 什么是 AutoML。
- 为什么和怎么用 AutoML。
- 何时需要 AutoML。
- AutoML 库概览。

1.1 机器学习的范围
Scope of Machine Learning

机器学习和预测分析有助于企业关注重点领域、提前预知问题、节约成本、提高收入。这是继**商业智能**（business intelligence，BI）之后的自然演变。商业智能多采用仪表盘，展示各种**关键表现指标**（key performance indicators，KPI）和性能指标，以便系统地监控业务流程，支持企业做出更好的决策。

商业智能工具可深入挖掘组织中的历史数据，发现趋势，理解季节效应，调查非常规事件，等等。这类工具也提供实时分析，支持设置警告和提示，以便精准地管理事件。然而，商业智能工具已经无法满足现代商业的需求。为什么？商业智能工具主要处理历史数据和近实时数据，但它解答不了未来的问题，比如它无法回答如下问题。

- 生产线中的哪台机器可能出故障？
- 哪个客户可能投奔竞争对手？
- 哪个公司的股价明天会上涨？

解答企业现在想知道的这类问题，机器学习和预测分析就能派上用场了。

不过要小心！机器学习模型并不一定能保证获得准确的结果。尽管机器学习的发展速度超乎想象，但我们还是需要找到正确的应用方向和领域，才能让机器学习发挥真正的作用，从而创造实用价值。

要发挥机器学习的作用，最好先从具备以下条件的小项目入手。

- 有相对容易的决策流程。
- 你熟悉各种假设条件。
- 你熟悉已有的数据。

这里的关键是项目范围和执行步骤要有明确的定义。不过，从小项目做起，不代表愿景也小。要始终考虑未来的可扩展性，支持慢慢加码到大数据源。

机器学习算法有很多种，每一种都为解决特定问题而设计，各有利弊。机器学习领域的研究一直在发展，从业者每天都在提出新方法，不断扩张新边界。结果是，研究人员很容易迷失在爆炸的信息中，尤其在开发机器学习应用的时候，建模的每个阶段都有许多可选的工具和技术。为了更容易构建机器学习模型，需要将整个过程分解成多个小阶段。**自动机器学习**（automated ML，AutoML）流水线中有许多动态的部分，如特征预处理、特征选择、模型选择和超参数优化。其中每个部分都要悉心处理，才能成功交付项目。

稍后我们会详细介绍机器学习的概念，现在先讲讲为什么要关注 AutoML。

解决问题的工具和技术越来越多，有时反而成了困扰，因为调研和探索某个问题的解决方法会耗费大量时间。机器学习的问题也一样。高性能机器学习模型的搭建包含若干精巧的小步骤，步步相连，构成机器学习流水线，然后合理泛化到生产环境中。

流水线中涉及许多步骤，因而可能变得冗长繁杂。每一步都有许多方法可选，而且这些方法还能排列组合，所以必须系统性地验证机器学习流水线中的环节。

这时候就该 AutoML 出场了！

1.2 什么是 AutoML
What is AutoML?

自动机器学习（AutoML）将特征预处理、模型选择和超参数优化等常用步骤自动化，以简化机器学习的建模流程。接下来的章节会详细介绍这些步骤，并且会教读者动手构建一套 AutoML 系统，从而对 AutoML 工具和库有更深刻的理解。

在开始之前，有必要回顾一下什么是机器学习模型，以及如何训练模型。

机器学习算法对数据进行处理，识别特定的模式，这一学习过程称为**模型训练**（model training）。模型训练的结果是机器学习模型。有了机器学习模型，你不用制定明确的规则，它就可针对数据提出见解或解答。

在实际应用机器学习模型时，需要输入大量数据，用于算法训练。训练后的成果是可用于预测的机器学习模型。这种预测可根据服务器当前状态来确定它未来四个小时是否需要维护，或者判断客户会不会投向竞争对手。

有时待解决的问题本身都没有明确定义，甚至我们都不知道需要什么样的答案。在这种情况下，机器学习模型可帮助探索数据集，比如识别行为相似的客户群，或者根据不同股票之间的关联关系发现股票的层级结构。

模型划分出客户群后，有什么用？至少可以知道：同一群体的客户有哪些相似的特征，比如年龄、职业、婚姻状况、性别、喜好、日常消费习惯、总消费额等。不同群体的客户是彼此不同的。有了这些信息，我们就可以针对每个群体推送不同的广告。

可以使用简单的数学术语说明这一流程。设有数据集X，包含n个样本。样本可代表客户或不同的动物。通常，每个样本都是一个实数集，称为**特征**（feature），比如，一位 35 岁的女性客户在商店消费了 12000 美元，可以用向量（0.0，35.0，12000.0）表示。注意，这里性别是用*0.0*表示的，男性客户可以用*1.0*表示。向量的大小称为维度，通常用m表示。这是一个大小为 3 的向量，即三维数据集。

根据问题的类型不同，需要为每个样本添加标签。假如这是一个二分类问

题，可以用 1.0 和 0.0 标识样本，那这个新变量就叫**标签**（label）或**目标**（target）变量。目标变量一般用y表示。

有了x和y，机器学习模型就可以看成是一个权重为w（模型参数）的函数f：

$$f(x;w)$$

模型参数是在训练过程中学到的，还有些参数是在训练开始前设置好的，这种参数称为**超参数**（hyperparameter），稍后会对这个概念进行解释。

数据集的特征应该先进行预处理，再用于模型训练。比如，有些机器学习模型隐含的假设是特征呈正态分布。在许多真实场景中，特征并不是正态分布的，那么就需要借助一些特征转换法，如对数转换，把特征调整为正态分布。

特征处理完成，且模型超参数也设置好后，就开始模型训练了。模型训练结束后，会学到模型参数，可以用来预测新数据的目标变量。模型做出的预测通常表示为$\hat{y}$：

$$\hat{y} = f(x;w)$$

训练过程中发生了什么？由于我们已知训练数据集的标签，可以将当前模型预测出的结果与原标签做对比，据此反复调整模型参数。

这种对比是基于**损失函数**（也称代价函数，loss function 或 cost function）进行的，表示为$L(\hat{y},y)$。损失函数可表示预测的失真度。常见的损失函数有平方损失（square loss）、合页损失（hinge loss）、逻辑损失（logistic loss）和交叉熵损失（cross-entropy loss）。

模型训练完成后，可以用 test 测试数据测试机器学习模型的性能。test 数据是训练过程中没有用过的数据集，用于检验模型的泛化能力。评估模型的表现可以用不同的指标；根据得到的结果再重复前面的步骤，反复调整，得到更好的性能。

现在，你应该对训练机器学习模型的工作原理有了大概的认识。

那么，AutoML 是什么呢？说起 AutoML，大多数情况下是指自动化数据准备（即特征预处理、特征提取、特征选择）和模型训练（模型选择、超参数优化）。这个过程中的每一步有多少可选项，根据问题类型不同而有很大差别。