人工智能 人才培养系列

机器学习案例实战

◎ 赵卫东 著

15个机器学习大案例，涵盖银行、保险、证券、零售、电力等多个行业

强调实战，与实际紧密结合，解决目前掌握机器学习技术的痛点问题

使用多种业界主流开源机器学习平台，包括谷歌 TensorFlow、腾讯 TI-ONE 以及百度 PaddlePaddle 等

中国工信出版集团
人民邮电出版社 POSTS & TELECOM PRESS

人工智能人才培养系列

机器学习案例实战

中国科学技术大学 图书馆藏

◎ 赵卫东 著

人民邮电出版社
北京

图书在版编目（CIP）数据

机器学习案例实战 / 赵卫东著. -- 北京 : 人民邮电出版社, 2019.9
ISBN 978-7-115-51410-3

Ⅰ. ①机… Ⅱ. ①赵… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2019)第111444号

内 容 提 要

机器学习已经广泛地应用于各行各业，深度学习的兴起再次推动了人工智能的热潮。本书结合项目实践，首先讨论了 TensorFlow、PySpark、TI-ONE 等主流机器学习平台的主要特点；然后结合 Tableau 介绍了数据可视化在银行客户行为分析中的应用。在此基础上，利用上述介绍的这些平台，通过多个项目案例，详细地分析了决策树、随机森林、支持向量机、逻辑回归、贝叶斯网络、卷积神经网络、循环神经网络、生成对抗网络等机器学习算法在金融、商业、汽车、电力等领域的应用。

本书内容深入浅出，提供了详细的 Python 代码，既可以作为从事机器学习、数据挖掘工作的相关研究人员和技术人员的参考书，也可以作为高校相关专业机器学习、数据挖掘等课程的实验和实训教材。

◆ 著　　赵卫东
责任编辑　张　斌
责任印制　陈　犇

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市中晟雅豪印务有限公司印刷

◆ 开本：787×1092　1/16
印张：18.25　　2019 年 9 月第 1 版
字数：431 千字　　2019 年 9 月河北第 1 次印刷

定价：59.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前言 FOREWORD

机器学习是人工智能的核心技术。本书针对典型的实际应用情景，结合作者最近几年在科研、教学和企业培训中的成果，基于 TensorFlow、PySpark 和 TI-ONE 等主流的开源机器学习平台，使用真实的业务数据和企业应用问题，详细、深入地介绍了机器学习实施的基本思路、关键步骤和难点。本书通过这些实际应用案例介绍了数据可视化、典型的机器学习算法以及电子推荐技术的具体应用，使得读者能够深刻地理解机器学习的专业知识和解决问题的思路，提升对实际应用领域问题的分析和动手能力。

本书可以与 2018 年出版的《机器学习》一书配套学习。通过对本书的学习，读者不仅可以模仿书中的案例实践基于开源机器学习平台的实际数据分析应用，也很容易举一反三，对新的数据分析问题提出合理的分析思路。

本书不是简单地介绍机器学习理论，而是通过分析目前机器学习技术的痛点（即与实际应用结合不紧密等问题）而编写的实战案例集。当前，国内机器学习实战方面的资料存在着以下 3 个明显问题：一是机器学习的应用案例比较粗略，问题也比较简单，分析过程不具体，难以支撑机器学习技术的学习，而这方面又是培养人工智能应用人才非常重要、不可或缺的环节；二是数据量比较小，分析的问题仅仅是实际问题的简化，数据的分析深度、算法的复杂度还达不到机器学习的要求；三是内容分散，主流的机器学习开源平台非常多，各有千秋，而实际上机器学习的工作使用 TensorFlow、PySpark 等几种主流的机器学习平台就可以胜任。

本书通过精心地选择实际数据和应用问题，突出使用机器学习解决数据分析过程中常见的问题，使读者不仅能理解几种主流机器学习平台的原理，还能针对实际问题设计可视化分析、机器学习等分析程序，具有较强的实战性。

使用本书的读者需要有一定的 Python 编程基础，如果对 Spark 有一定的了解更佳。对于学习者而言，定义机器学习模型是一项复杂而又有难度的工作，而借助良好的机器学习框架，可以降低应用门槛。为了便于学习机器学习的分析过程，本书使用了多种业界主流的开源机器学习平台，包括 TensorFlow、PySpark 和 TI-ONE 等，这些平台一般注册账号后就可以使用，使读者在数据分析的过程中可以把主要的精力放在数据分析的思路上，降低了应用开发的难度。

本书可作为对机器学习感兴趣的研究人员和工程技术人员的参考资料，也可作为高等院校本科生、研究生的机器学习、数据分析、数据挖掘等课程的实验或实训教材。

感谢腾讯、谷歌、百度等公司资助的产学合作协同育人项目的支持。在本书写作的过程中，研究生蒲实、耿甲、于召鑫、袁雪如、陈伯宇、胡远文等在资料收集方面做了一些工作，在此特表示感谢。

2019年5月

于复旦大学

试读结束，需要全本请在线购买：www.ertongbook.com

目 录 CONTENTS

第1章　常用机器学习平台

一个功能强大且易学、易用的机器学习平台对于开展机器学习项目非常重要。良好的机器学习框架提供了丰富的预制组件，可以方便机器学习模型的设计和实现。目前存在以下几类基本的机器学习平台：一类是开源的机器学习平台，API（Application Programming Interface，应用程序编程接口）丰富且不用付费，但学习成本高，例如 R、Python、Mahout、Spark MLlib 等。还有一类是商业化的机器学习平台，这类平台算法有限，但经过了长期的实践检验，系统问题比较少，学习成本低，很少编程甚至不用编程，但系统内的分析模型不够丰富，例如 IBM SPSS Modeler。此外，还有一类机器学习平台综合了以上两类平台的优点，既提供了丰富的算法调用接口，可以通过图形化的人机接口快速搭建机器学习的工作流，又可以减少编程的工作量。目前微软、谷歌以及国内的 BAT（百度、阿里巴巴、腾讯）等公司都提供了这样的机器学习平台。

1.1　常用机器学习工具

Rapid Miner 是一个用于机器学习和数据挖掘实验的工具。该工具用 Java 编程语言编写，通过基于模板的框架提供高级分析。它使得实验可以由大量的可任意嵌套的操作符组成，这些操作符在 XML 文件中描述较详细，并且是由 Rapid Miner 的图形用户界面完成的，用户不需要编写代码。它包含许多模板和其他工具，可以轻松地分析数据。

Apache Mahout 是 Apache 软件基金会的一个项目，用于协同过滤、聚类和分类领域的分布式或其他可伸缩机器学习算法的实现。Apache Mahout 主要支持三种用例：建议挖掘采取用户行为，并尝试查找用户可能喜欢的项目；集群需要文本文档，并将它们分组为局部相关的文档；分类从现有的分类文档中学习特定类别文档的特点，并能够将未标记的文档分配给正确的类别。

TensorFlow 是被广泛使用的实现机器学习以及其他涉及大量数学运算的算法库之一。TensorFlow 由谷歌开发并开源，是 GitHub 上最受欢迎

的机器学习库之一。TensorFlow 采用数据流图进行数值计算。其中 Tensor 是可以代表 n 维数据集的张量，Flow 使用计算图进行计算。数据流图是用节点和边组成的有向图来描述数学运算。节点一般对应数学操作或状态，并对应节点之间的输入/输出关系。在 TensorFlow 中，所有不同的变量和运算都储存在计算图中。因此在构建完模型所需要的图之后，需要开启一个 Session 来运行整个计算图。TensorFlow 的模型构建的基本流程包括构建计算图、馈送输入张量、更新权重并且返回输出值。使用 TensorFlow 可以方便地搭建各种常见的神经网络，也可以模拟多种回归算法，并且在此基础上对模型中的参数进行训练，得到训练好的模型可用于后续实验。但 TensorFlow 内部概念众多、结构复杂，繁杂的 API 导致新用户上手困难，冗长的代码使得工程实现比较费力。

PaddlePaddle 是由百度开源的一款全功能的深度学习框架，其架构历经多次迭代，为开发者提供易学、易用、安全、高效的深度学习研发体验。PaddlePaddle 对开发者非常友好，所有的 API 都提供详尽的中文文档，并且提供了 Jupyter 文稿。PaddlePaddle 的代码易于理解，方便用户理解框架和提出问题。PaddlePaddle 的 API 中对算法原理进行了概括，方便用户学习理解深度学习算法。PaddlePaddle 支持 Windows、Linux 和 macOS 等多种操作系统，具有非常好的可拓展性，用户无须配置第三方库即可完成整个 PaddlePaddle 框架的编译。PaddlePaddle 提供了全面的深度学习 API，支持 Python 调用。同时 PaddlePaddle 对于图像分类、目标检测、图像语义分割、图像生成、场景文字识别、度量学习、视频分类、语音识别、机器翻译、强化学习、中文词法分析、情感倾向分析、语义匹配、机器阅读理解和个性化推荐等具体的深度学习问题提供了训练好的模型库，用户可以直接调用模型。PaddlePaddle 还有一个基于 Web 的 IDE，支持使用者在浏览器中使用 Jupyter Notebook 编程来开发 AI 应用，随后发送到云端调试或者运行，程序运行时的输出会实时地显示在浏览器里。PaddlePaddle 底层使用 C++编写，运行速度快，占用内存少。PaddlePaddle 在分布式计算上也表现优异，可通过与 Kubernetes 合作实现弹性作业调度。

Caffe2 是面向工业级应用的框架，应用广泛。但是从安装部署角度来说，Caffe2 的用户体验并不是非常友好，官方文档和教程支持也不是十分充足。而且 Caffe2 只支持 Python 2，这限制了其未来的拓展。

MXNet 是一款灵活高效的深度学习框架，并行计算性能好、运行速度快，并且程序节省内存，支持 R、Julia、Python、Scala、C++等多种语言。MXNet 支持命令式和声明式两种编程方式，代码更加灵活。但 MXNet 是由社区推动的深度学习框架，很多问题出现后还需要用户去查阅源码，而且模型库支持不够，需要开发者写代码实现。

PyTorch 是 Facebook 开发的面向学术界的一个框架，安装方便，使用简单，构建网络也比较容易。PyTorch 运行后立刻出结果，不同于 TensorFlow 必须把程序写完之后才知道结果是什么。但 PyTorch 不适合工业级应用。

VS Tools for AI 和 VS Code Tools for AI 是微软公司发布的一系列人工智能工具，建立在微软多年的旗舰产品之上，提供了强大的前端集成式编程环境，支持多种平台。在公有云、私有云上都提供了可扩展的 GPU 集群管理和调度工具，可以自动生成并优选神经网络模型，支持不同框架训练出来的机器学习模型。

此外还有 Amazon Machine Learning（AML）、Theano 等，有兴趣的读者可以查询相关资料。

1.2 TI-ONE 平台概述

智能钛机器学习平台是腾讯公司实现机器学习模型训练和运行的一站式平台化解决方案。该平台主要为模型训练、运行、评估与优化提供支持。用户可以上传标注的数据，利用平台切分成训练集、验证集以及测试集。训练模型的算法可以自行编写，也可以使用平台提供的，然后，在平台上设置相关参数，计算资源参数，并训练模型，模型的可用性也可以在平台上进行检测。

TI-ONE 机器学习平台是智能钛机器学习平台的子平台之一，适合有一定机器学习经验的建模人员使用，TI-ONE 平台支持使用编程语言实现数据处理、特征获取，可以使用可视化、模块化的建模工具，通过配置参数的方式构建机器学习模型训练工程，平台可以提供基本的机器学习和深度学习算法，计算资源由平台管理，用户只需要专注于业务场景相关的模型。

TI-ONE 平台提供云端的具备高可用性的 GPU 分布式集群服务器，可以满足大规模深度学习模型训练的性能要求；平台内部兼容 TensorFlow、Torch、Caffe 等多种主流的机器学习框架，从而可以支持用户自编程代码的上传和运行，为用户提供了灵活性。

TI-ONE 平台对 GPU 分布式集群服务器上的深度学习模型训练算法做了优化，能够大幅度地提升训练速度，从而大大地减少模型训练所花费的时间；平台提供了搭建好的机器学习开发环境，并且为用户管理计算资源，可以为用户节省这部分的时间，使用户的精力可以集中在业务相关的工作中。平台提供的沙箱能够帮助用户在保证数据安全和稳定的环境中，整合多方数据进行建模。

TI-ONE 平台适合应用在所有需要使用机器学习或深度学习平台进行定制建模的场景中，典型的场景有风控、营销推荐、预测、非结构化数据处理、文本分析和关系挖掘等。平台可以通过接收原始数据的输入，训练各个场景下的不同模型，应用到对应的业务场景中。

TI-ONE 平台的架构可以分为六个层次，从上到下依次是产品层、交互层、算法层、框架层、调度层以及资源层。产品层表示用户所接触的 TI-ONE 平台。交互层表示用户的交互方式，也就是图形化界面。算法层是平台开发团队实现的算法并且以组件的形式提供给用户使用，提供的算法有机器学习、深度学习以及图算法。框架层包含 TI-ONE 平台内部算法、实现所依赖的框架以及提供给用户的自编程功能可运行的框架：Spark、TensorFlow、Angel、Mariana、Caffe、Scikit-Learn、MXNet、PyTorch。调度层采用新一代的企业级容器平台 GaiaStack，用于资源管理和调度。资源层可以提供计算资源以及存储资源，供用户自编程调用和各类组件调用。

TI-ONE 是一站式机器学习平台，是专为 AI 初学者设计的机器学习平台，具有可视化操作界面、具象化的算法结果、拖曳式的任务流、可灵活自定义的特性以及内置的丰富模型算法与案例。该机器学习平台的特性如下。

（1）拖曳式任务流：拖曳式设计，各个元素可以自由地组合，以一种搭积木的方式绘制任务流。

（2）多实例调度：支持手工、定时、批量参数、重跑，可以方便用户在各个应用场景下的灵活需求。

（3）支持多机器学习语言和框架：Python、R、Spark、TensorFlow 以及腾讯的 Angel 都可使用。

试读结束：需要全本请在线购买：www.ertongbook.com

（4）内置机器学习算法：算法包括特征工程、机器学习、深度学习、图算法等，充分满足不同场景下的使用需求。

（5）数据可视化：提供可视化服务，模型训练效果可以悬浮呈现，用户无须烦琐操作就可以方便地辨别模型质量。

（6）模型的完整闭环："一站式"机器学习平台体验，从模型训练、评估、服务部署到在线推理，覆盖全工作流程，形成机器学习训练的完整闭环。

在开始使用 TI-ONE 服务之前，首先需要开通 TI-ONE 与 COS（Cloud Object Storage，云对象存储）服务，COS 服务已接入 TI-ONE 产品，用于工程中的各环节。TI-ONE 系列产品目前开放免费试用。

TI-ONE 申请的流程如下：在产品介绍页单击"立即申请"按钮填写申请单后提交，进行线上白名单申请（需要到腾讯云平台）。接到服务申请后，腾讯云平台进行需求审核，并安排相应的工作人员进行初步需求确认、洽谈。审核通过后会发送审核结果给用户，用户可以根据指引在产品页进行试用体验。

TI-ONE 平台提供了五大类的组件，如图 1.1 所示。从上至下依次是输入、组件、算法、模型以及输出。其中输入包括数据源、数据转换、公共数据集，数据源有 COS 数据集以及本地数据；组件下有三个选项，分别是统计分析、机器学习、深度学习，机器学习包括 Spark 组件和 PySpark 组件，深度学习包括 PyCaffe 组件、PyCaffe 定制版组件、PyTorch 组件、TensorFlow 组件以及 TensorFlow 多机版组件；算法包含 27 个机器学习算法以及 16 个深度学习算法；模型即算法相关的组件；输出是机器学习输出用到的功能组件。

图 1.1　TI-ONE 平台的组件

PySpark 组件面向使用 Python 的 Spark 用户，用户可以使用 Python 编写 Spark 程序，通过该组件来完成部署，这个组件支持 Python 脚本上传与实时修改，还支持 PySpark 的 SQL 功能，灵活性很强，很适合数据预处理，也适合偏好 PySpark 的 ML 库的使用者和 Python 使用者。在使用 PySpark 组件时，推荐使用 PySpark 中的 DataFrame 来替代 Pandas 中的 DataFrame，这是由于前者是分布式执行的，而后者则是单机执行的。

1.3　PySpark 介绍

Spark 是一种分布式计算框架，并且有一套生态系统，其中包括 Spark Core、Spark SQL、Spark

MLlib、Spark Streaming 和 Spark Graphx，支持进行离线计算、交互式查询、机器学习、流计算以及图计算。PySpark 是 Spark 为 Python 开发者提供的 API。子模块包括 pyspark.sql 模块、pyspark.streaming 模块、pyspark.ml 模块、pyspark.mllib 模块；核心类包括 pyspark.SparkContext、pyspark.RDD 、pyspark.sql.SQLContext 、pyspark.streaming.StreamingContext 、pyspark.streaming.DStream 和 pyspark.sql.DataFrame。

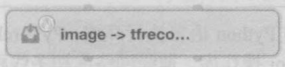
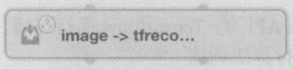
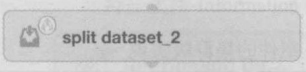
PySpark 的 ML 包和 MLlib 包都是机器学习包，可以应用于分类、回归等常见的机器学习问题。两者内部集成的具体算法有一些差别，模型的训练、预测和评估的细节上有所差别，但对于常用的机器学习功能，都是可以满足需求的。

1.4 TI-ONE 机器学习平台主要的组件

1.4.1 数据源组件

TI-ONE 平台提供了两种外来的数据源组件：COS 数据集和本地数据，可在控制台左侧导航的数据源分类下找到。此外，还有数据转换组件以及公共数据集，如表 1.1 所示。

表 1.1 数据转换组件

组件	说明
数据源-COS 数据集 	将 COS 数据集组件拖曳至画布中，为下游算法提供输入数据，指定 COS 数据路径即可通过数据流自动传递。 可以通过“是否检查数据”开关来判断数据是否存在后再启动后续节点处理
数据源-本地数据 	将本地数据组件拖曳至画布，把本地文件上传到目标 COS 路径（需要有上传目标目录的写权限，文件大小不能超过 256MB），为下游算法提供输入数据
数据转换-image→tfrecord （Image Classification） 	将图像分类数据转换成 TFRecord 格式，TFRecord 是谷歌推荐的一种二进制文件格式，理论上它可以保存任何格式的信息。TensorFlow 提供了丰富的 API 可以帮助用户轻松读写 TFRecord 文件
数据转换-image→tfrecord （Object Detection） 	将物体检测数据集转换成 TFRecord 格式，指定图像列表路径、图像路径、标签路径以及转换后的 TFRecord 训练集和测试集的路径
数据转换-split dataset 	将数据集按比例分成训练集和测试集，可以选择是分类问题还是检测问题以及测试集所占比例
公共数据集- AI Annotation Dataset 	本数据集为 Pascal VOC 2007 目标检测标注数据集

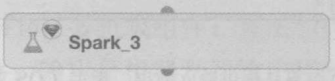
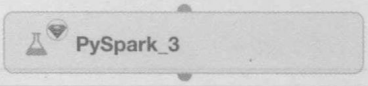
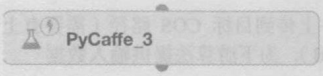
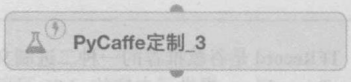
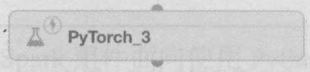
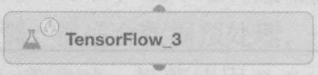
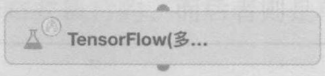
续表

组件	说明
公共数据集- AI Image Dataset 	本数据集为 Pascal VOC 2007 目标检测图片
公共数据集-大赛公共数据集 	本数据集是各类大赛中所使用的公共数据集

1.4.2 机器学习组件

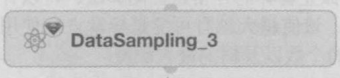
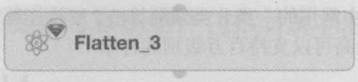
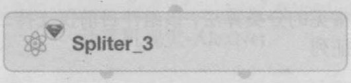
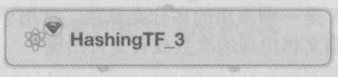
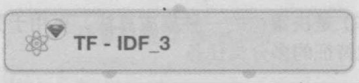
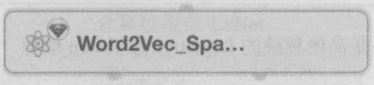
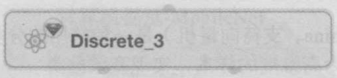
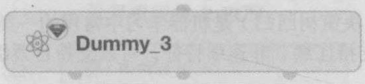
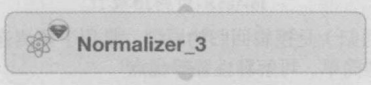
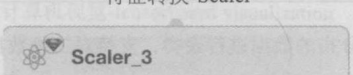
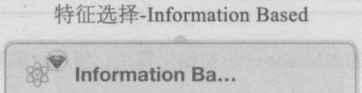
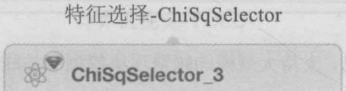
机器学习组件提供了常用的机器学习开发框架、使用的库以及对应的计算资源，如表 1.2 所示。

表 1.2　　机器学习组件

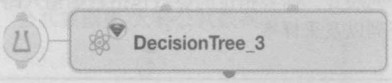
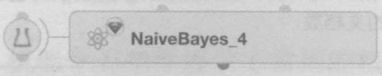
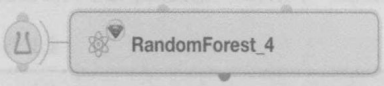
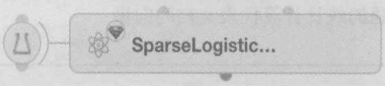
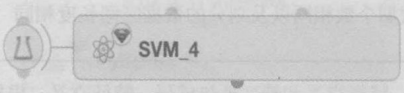
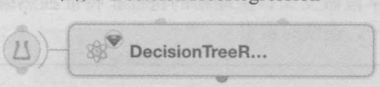
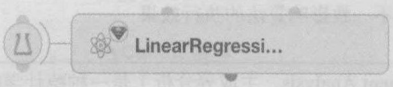
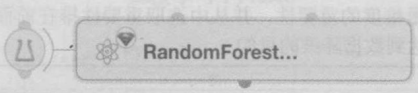
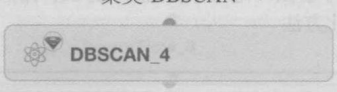
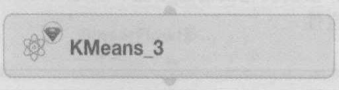
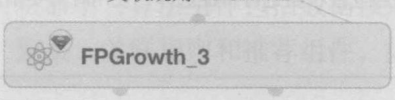
组件	说明
机器学习-Spark 	Spark 面向使用 Scala/Java 的 Spark 用户，用户编写 Spark 应用程序，编译打包成 jar 后，通过 Spark 组件完成部署。使用方法：从左侧组件列表里拖曳出一个 Spark 节点，单击任务节点，会从右侧弹出配置框，配置完毕后，鼠标右键单击启动运行
机器学习-PySpark 	PySpark 面向使用 Python 的 Spark 用户，用户通过 Python 编写 Spark 应用程序，上传 Python 脚本和实时修改，比较灵活，所以比较适合用来做数据预处理
深度学习-PyCaffe 	Caffe 是一个清晰而高效的深度学习框架，具有入门快、运行速度快、模块化、开放性等特点。PyCaffe 组件的内核是 Caffe 1.0 版本
深度学习-PyCaffe 定制 	支持引入 PyCaffe 类库后编写的自定义脚本，Caffe 的 Python 接口不仅支持加载模型、对模型进行前向和反向训练，还能实现控制 I/O、网络虚拟化等多种功能，所有的模型数据、派生类和变量都可以通过 Python 接口访问
深度学习-PyTorch 	PyTorch 实现了机器学习框架 Torch 的 Python 语言执行环境。PyTorch 运行在 GPU、CPU 之上，支持基础 Tensor 操作库、神经网络库和多线程并发库，可以与 NumPy、SciPy 和 CPython 一起运行
深度学习-TensorFlow 	TensorFlow 为用户提供了基于 Python API 的 TensorFlow 运行环境，用户可将编写的脚本和依赖文件上传进行算法训练
深度学习-TensorFlow（多机版） 	TensorFlow（多机版）是 TensorFlow 组件的集群版实现

机器学习算法组件包括数据预处理、特征提取、特征转换、特征选择、异常检测、分类、回归、聚类、关联规则和推荐组件，如表 1.3 所示。

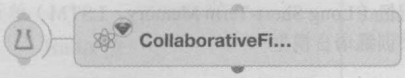
表 1.3 机器学习算法组件

组件	说明
数据预处理-DataSampling 	DataSampling 提供了从原数据集随机抽取特定比例或者特定数量的小样本的方法，该模块常用于抽取小样本用于数据的可视化
数据预处理-Flatten 	Flatten 将特征按照某个列展开，需要指定的参数包括数据输入路径、ID 字段、展开列、特征列以及采样率
数据预处理-Spliter 	Spliter 对数据按比例进行随机划分，使原始样本数据被划分成训练集和测试集，并且可以指定划分的数据集所占的比例
特征提取-HashingTF 	HashingTF 完成文本中词频率的计算，可求得需要统计的词在文档中出现的次数，以及包含词的文档数
特征提取-TF-IDF 	TF-IDF 是文本特征处理算法，可以得到一篇文档中每个词汇在该文档中的相对重要程度
特征提取-Word2Vec_Spark 	Word2Vec_Spark 是一种基于 Spark 实现的 Word2Vec 算法，通过词向量以及文档中出现的词语的统计计算代表文档的向量
特征转换-Discrete 	Discrete 算法对属性进行离散化。离散化的方法包括等频和等值等方式，分别使得划分的数据个数相同以及划分的数据区间长度相同
特征转换-Dummy 	Dummy 包含两个阶段：特征交叉和特征独热编码。特征交叉：根据配置文件，对指定的特征字段做交叉，生成新的特征；特征独热编码：将特征名编码成全局统一、连续的索引
特征转换-Normalizer 	Normalizer 提供了特征标准化，将各维特征转化为[0, 1]的值，可以避免不同特征值的分布域不一致影响算法的执行效果
特征转换-PCA 	PCA（Principal Component Analysis，主成分分析）是一种统计学的特征降维方法，它将数据从原来的坐标系投影到新的坐标系，通过每个维度的方差大小来衡量维度的重要性，并从中选取重要性排在前面的特征作为新的特征，达到数据降维的目的
特征转换-Scaler 	Scaler 模块集成了最大/最小值归一化和标准化两种方式，用户可通过特征配置文件来指定归一化方法
特征选择-Information Based 	基于信息计算的特征选择，该模块包括信息增益、基尼（Gini）系数、信息增益率以及对称不确定性
特征选择-ChiSqSelector 	ChiSqSelector 基于卡方独立性检验进行特征选择，使用时需要指定特征所在列和标签列

续表

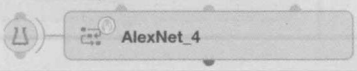
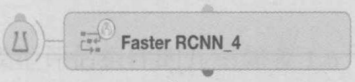
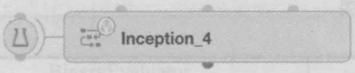
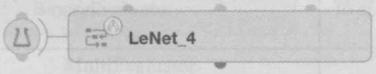
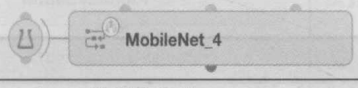
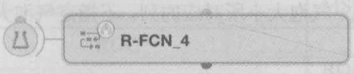
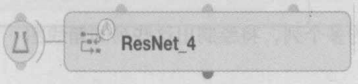
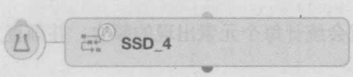
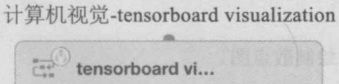
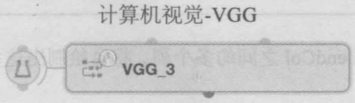
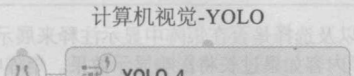
组件	说明
异常检测-IsolationForest 	IsolationForest 是一种基于孤立森林的异常点检测算法，可以计算出每个样本成为异常点的概率，该值越大越有可能是异常点。使用时需要指定每棵树的样本数、树的个数以及树的最大深度
分类-DecisionTree 	DecisionTree 是机器学习中常用的一类分类预测算法，支持连续、非连续特征的多分类任务，最高可以支持百万级别的样本
分类-LogisticRegression 	LogisticRegression 是一种常见的分类算法，该组件目前仅支持二分类。使用时需要指定选择的特征列
分类-NaiveBayes 	NaiveBayes（朴素贝叶斯）是一种常用的多分类算法，常用于文本分类，每个特征表示词在一篇文档出现的次数或者是否出现
分类-RandomForest 	RandomForest（随机森林）是决策树的一种集成算法，可用于分类和回归，支持连续、非连续特征的多分类任务
分类-SparseLogisticRegression 	SparseLogisticRegression 是高维稀疏的 LogisticRegression 模型
分类-SVM 	SVM（Support Vector Machine，支持向量机）是一种常用的分类算法，可以把低维度的样本转化为高维度的样本，实现高效分类
回归-DecisionTreeRegression 	DecisionTreeRegression（决策树回归）是机器学习中常用的一类分类/回归算法。该回归算法支持连续、非连续特征。可以支持百万级别的样本
回归-LinearRegression 	LinearRegression（线性回归）是逻辑回归的原型，常用于预测连续的目标值，该算法具有模型简单、可解释性强等优点
回归-RandomForestRegression 	RandomForestRegression（随机森林回归）是决策树的一种集成算法，可用于回归，支持连续特征
聚类-DBSCAN 	DBSCAN 算法对于任意分布的数据进行聚类，支持对二维数据进行聚类
聚类-KMeans 	KMeans 算法实现了并行的 KMeans++的初始化算法，使用时需要指定属性
关联规则-FPGrowth 	FPGrowth 算法的并行实现，支持大规模的频繁项集挖掘和关联规则的生成

续表

组件	说明
推荐-CollaborativeFiltering 	CollaborativeFiltering（协同过滤）是经典的基于邻居的推荐算法，平台上的协同过滤算法通过 ALS 的矩阵分解优化求解

深度学习算法组件包括计算机视觉、自然语言处理、表示学习组件，如表 1.4 所示。

表 1.4　　深度学习算法组件

组件	说明
计算机视觉-AlexNet 	AlexNet 是来自 DCNN 的图像分类，在没有 LRN 层的情况下实现
计算机视觉-Faster RCNN 	Faster RCNN 将特征抽取、proposal 提取、bounding box regression 和分类都整合在了一个网络中，使得网络综合性能有较大提高，尤其是在检测速度方面。使用时需要配置的参数有输入数据路径、日志存储路径、主网络、类别个数、训练步长以及计算资源参数
计算机视觉-Inception 	Inception 网络增加了网络的宽度，同时增加了网络对尺度的适应性，而且有降低特征图厚度的作用，网络的非线性也得到了增加
计算机视觉-LeNet 	LeNet 是最早的卷积神经网络（Convolutional Neural Networks，CNN）之一，由 Yann LeCun 提出，利用卷积、权值共享、池化等操作提取特征，避免了大量的计算成本，最后使用全连接神经网络进行分类识别
计算机视觉-MobileNet 	MobileNet 的基本单元是深度级可分离卷积，是一种可分解的卷积操作，可以分解为两个更小的操作：Depthwise 卷积和 Pointwise 卷积。MobileNet 在计算量和参数量上比较有优势
计算机视觉-R-FCN 	R-FCN 提出 Position-sensitive score maps 来解决目标检测的位置敏感性问题；是以区域为基础的、全卷积网络的二阶段目标检测框架；比 Faster-RCNN 快 2.5～20 倍
计算机视觉-ResNet 	ResNet 使用的是 shortcut connection 的连接方式，它提出了两种映射：恒等映射（Identity Mapping）和残差映射（Residual Mapping）。通过这两种方式，ResNet 解决了随着网络加深而准确率下降的问题
计算机视觉-SSD 	SSD 采用 CNN 来进行检测，采用了多尺度的特征图，利用卷积进行检测，并且设置先验框。主要思路是均匀地在图片的不同位置进行密集抽样，抽样时可以采用不同尺度和长宽比，然后利用 CNN 提取特征后直接进行分类与回归，整个过程只需要一步，优势是速度快
计算机视觉-tensorboard visualization 	TensorFlow 是可用于训练大规模深度神经网络的计算框架。为了更方便 TensorFlow 程序的理解、调试与优化，可以使用 TensorBoard 来展现 TensorFlow 图
计算机视觉-VGG 	VGG 网络模型可以应用在人脸识别、图像分类等方面，VGG 在加深网络层数的同时为了避免参数过多，在所有层都采用 3×3 的小卷积核，卷积层步长为 1。与 AlexNet 相比，VGG 对图片有更精确的估值而且更省空间
计算机视觉-YOLO 	YOLO 将物体检测作为回归问题求解。基于一个单独的端对端网络，完成从原始图像的输入到物体位置和类别的输出。YOLO 检测网络包括 24 个卷积层和 2 个全连接层。YOLO 网络的整个检测网络的流程简单，速度快；背景误查率低；通用性强

试读结束：需要全本请在线购买：www.ertongbook.com