HZ BOOKS
华章IT

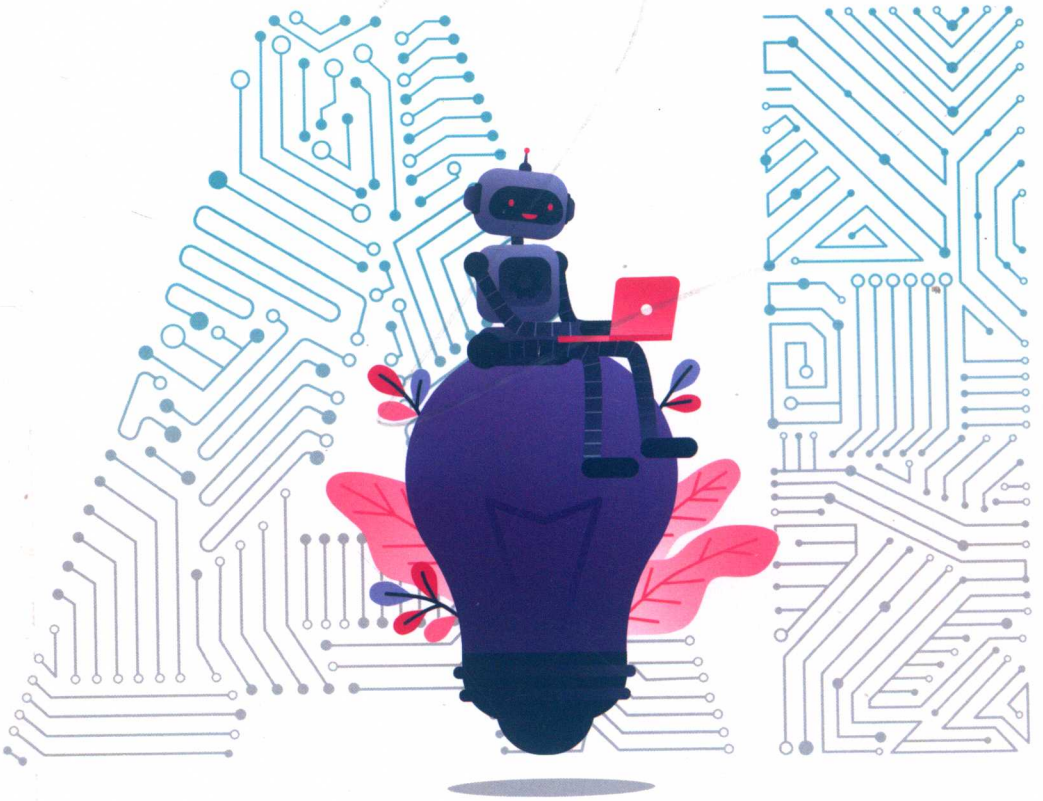
智能系统与技术丛书

知名互联网公司资深工程师撰写，打通高效机器学习脉络，掌握竞赛神器XGBoost
以机器学习基础知识做铺垫，深入剖析XGBoost原理、分布式实现、模型优化、深度应用等

深入理解XGBoost

高效机器学习算法与进阶

何龙 著



机械工业出版社
China Machine Press

随着企业数据规模的不断扩大，基于规则解决业务问题的传统方法已无法满足企业需求，迫切需要一种高效、准确的方式提升业务指标，如何快速训练出高准确率的模型成为许多企业面对的挑战。

XGBoost（eXtreme Gradient Boosting），初惊艳于Kaggle竞赛，后以出众的效率和较高的准确度得到广泛应用。正如其名，它是Gradient Boosting思想的实现，最大的特点是能够通过并行实现加速计算，同时在算法上加以改进提高了精度。

XGBoost存在一定的学习门槛，资料又较为分散，缺乏一个系统、完整的学习教程可以参考。作者作为XGBoost开源社区贡献者，经长期积累与实践，潜心研思，总结成书，以飨读者。

- **理论先行：**以机器学习常用算法为铺垫，打下良好的理论基础。
- **夯实基础：**全面覆盖了决策树、Gradient Tree Boosting、目标函数近似、切分点查找算法等主要内容。
- **重视深度：**深入阐述了分布式实现、排序学习、模型解释性等算法和技术。
- **案例丰富：**算法讲解与应用案例相辅相成，以便读者灵活运用、融会贯通。

此外，本书也可作为算法开发人员的手边工具书，在学习和工作中随时查阅参考。

上架指导：计算机/人工智能
ISBN 978-7-111-64262-6
9 787111 642626 >
定价：99.00元

投稿热线：(010) 88379604
读者信箱：hzit@hzbook.com
客服电话：(010) 88361066 88379833 68326294

华章网站：www.hzbook.com
网上购书：www.china-pub.com
数字阅读：www.hzmedia.com.cn

深入理解XGBoost

高效机器学习算法与进阶

何龙 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

深入理解 XGBoost：高效机器学习算法与进阶 / 何龙著 . —北京：机械工业出版社，2020.1
（2020.5 重印）
（智能系统与技术丛书）

ISBN 978-7-111-64262-6

I. 深… II. 何… III. 机器学习 – 算法 IV. TP181

中国版本图书馆 CIP 数据核字（2019）第 262402 号

深入理解 XGBoost：高效机器学习算法与进阶

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：高婧雅 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2020 年 5 月第 1 版第 2 次印刷
开 本：186mm × 240mm 1/16 印 张：23.75
书 号：ISBN 978-7-111-64262-6 定 价：99.00 元

客服电话：（010）88361066 88379833 68326294 投稿热线：（010）88379604
华章网站：www.hzbook.com 读者信箱：hzit@hzbook.com

内容简介

本书以机器学习基础知识做铺垫，深入剖析XGBoost的原理、分布式实现、模型优化、深度应用等。

第1～3章使读者对机器学习算法形成整体认知，了解如何优化模型以及评估预测结果，并熟悉常用机器学习算法的实现原理和应用，如线性回归、逻辑回归、决策树、神经网络、支持向量机等。

第4章借助实际案例，讲解如何通过XGBoost解决分类、回归、排序等问题，并介绍了XGBoost常用功能的使用方法。

第5～7章是本书的重点，从理论推导与源码层面深入剖析XGBoost，涵盖XGBoost原理与理论证明、分布式XGBoost的实现、XGBoost各组件的源码解析。

第8～9章为进阶内容，着重解析算法实践与工程应用中的难点，进而帮助读者更好地解决实际问题。

第10章介绍了一些将树模型与其他模型融合较为前沿的研究方法，以开拓眼界，拓展思路。

作者简介

何龙

现就职于滴滴出行，XGBoost开源社区贡献者，专注于人工智能和机器学习领域，从底层算法原理到上层应用实践都有广泛的兴趣和研究。较早接触XGBoost，熟悉XGBoost应用开发，深入阅读源码，具有丰富的项目开发经验。

写作投稿
微信：xishuihua2008　邮箱：165075460@qq.com

封面设计　锡　彬

前言

大数据时代的今天，基于规则解决具体业务问题的传统方式已无法满足企业需求，机器学习与人工智能逐渐走入人们的视野，并迅速得到了众多企业的广泛关注。各大互联网公司相继成立了自己的机器学习研究院，或建立机器学习团队。然而，随着企业业务规模、数据规模日益扩大，业务类型越来越复杂，怎样在短时间内训练出高准确率的模型，成为许多企业面对的挑战。

在机器学习与人工智能的浪潮中，XGBoost 凭借高效、便捷、扩展性强等优势，在众多开源机器学习库中脱颖而出，广受各大企业青睐。目前 XGBoost 已成为热门的机器学习开源项目之一，拥有强大的社区支持，技术也日趋成熟。

为什么要写这本书

最初写这本书的想法萌生于两年前。当时，一些刚接触 XGBoost 的同事让我推荐学习资料，但我发现除了英文论文和官方文档外，竟找不到一本 XGBoost 的入门书籍。当然，论文和官方文档是学习 XGBoost 的重要参考资料，但对于刚接触机器学习的初学者而言，学习这些资料的成本相对较高。如果没有足够的理论基础，初学者容易一开始就被细节和难点缠住，降低学习的积极性。

XGBoost 涉及的相关知识较多，资料比较分散，苦于缺乏一个系统、完整的学习教程可以参考，学习者不得不在搜集资料上耗费大量时间。此外，对于 XGBoost 的应用也少有完整的案例剖析。想深入理解 XGBoost 的学习者，只能通过研究项目源码的方式进行学习，这显然不是一个特别高效的学习方式。

为了能够深入理解 XGBoost 中各个组件的实现原理，笔者也花费了很多时间和精力。在阅读了相关论文文档、深入研究源码并多次实践后，积累了很多学习笔记，对 XGBoost 也有了自己的理解，由此便萌生了将其整理成书的想法。这样既可以帮助更多的人快速了解和学习 XGBoost，使自己的学习所得发挥更大的价值，也可以在梳理所学知识的过程中进一步提升。

本书特色

本书是国内少有的系统、全面地介绍 XGBoost 技术原理的书籍，以通俗易懂的方式对 XGBoost 的原理和应用进行介绍，力求帮助读者深入理解 XGBoost。

（1）**讲授循次而进，符合初学者的认知规律。**本书首先介绍机器学习中的常用算法，帮助读者直观地理解算法的基本原理，打下良好的理论基础。然后由浅入深，鞭辟向里，带领读者深入探索机器学习前沿技术。

（2）**内容涵盖全面，重视理解深度。**本书不仅全面覆盖了决策树、Gradient Tree Boosting、目标函数近似、切分点查找算法等常见内容，还详细讲解了分布式实现、排序学习、模型解释性、DART 等内容。

（3）**案例实用丰富，帮助读者解决实际遇到的机器学习问题。**本书在每个算法讲解之后都配有相应的编程示例，不仅使读者理解算法原理，还有助于提升灵活运用算法的能力。

另外，本书可以作为算法开发人员手边的工具书，在学习和工作的过程中随时查阅参考。

读者对象

- 人工智能领域的算法工程师
- 人工智能领域的架构师
- 其他对机器学习感兴趣的人

如何阅读本书

本书共有 10 章，具体内容如下。

第 1 章介绍了何谓机器学习和机器学习中的一些基本概念，以及机器学习应用开发的步骤，并对集成学习的历史发展、XGBoost 的应用场景及其优良特性进行了概述。

第 2 章详细讲解了 Python 机器学习环境的搭建及常用开源工具包的安装和使用，并以一个简单的示例展示 XGBoost 的使用方法。

第 3 章讲述了常用机器学习算法的实现原理和应用，如线性回归、逻辑回归、决策树、神经网络、支持向量机等，使读者对机器学习算法有一个整体认知，同时了解如何在模型训练过程中进行优化、如何对模型结果进行评估。

第 4 章通过 scikit-learn 与 XGBoost 相结合，以实际的案例向读者说明如何通过 XGBoost 解决分类、回归、排序等问题，并介绍了 XGBoost 中常用功能的使用方法。

第 5 章深入介绍 XGBoost 的实现原理，包括 Boosting 算法思想、XGBoost 目标函数近

似、切分点查找算法、排序学习、模型可解释性等内容，从理论上保证 XGBoost 的有效性，使读者深入理解 XGBoost。

第 6 章详细介绍了分布式 XGBoost 的实现原理，包括分布式机器学习框架 Rabit，XGBoost 在 Spark、Flink 平台的实现，GPU 版本的实现等。

第 7 章从源码的角度深入剖析了 XGBoost 中各个组件的实现原理，详细介绍了模型训练、预测、解析，以及不同目标函数和评估函数的实现过程。

第 8 章详细介绍了如何通过模型选择与优化提高模型的泛化能力，从偏差和方差的角度进行了解释，通过交叉验证和 Bootstrap 等方法来说明模型选择过程，并介绍了常用的超参数优化方法。

第 9 章通过实际案例分析，使读者能够深入理解 XGBoost 的特性并灵活运用，能够依据不同场景将 XGBoost 与其他模型融合，更好地解决实际问题。

第 10 章介绍了业界或学术界中树模型与其他模型融合的一些研究方法，包括 GBDT+LR，多层 GBDT 模型结构 mGBDT，树模型与深度学习、强化学习的融合等。

如果你是具有一定机器学习理论基础的从业者，可以跳过前 4 章的基础部分，直接从第 5 章开始阅读。如果你是一位机器学习的初学者，并打算在未来深入钻研这个方向，建议系统地阅读本书，并动手实践书中的所有实例。

勘误和支持

由于笔者水平有限，编写时间仓促，书中难免存在一些错误或表达不准确之处，恳请读者批评指正。如果你有更多的宝贵意见，可通过电子邮箱 xgbbook2019@163.com 联系我，期待得到你们的真挚反馈，让我们在技术之路上互勉共进。

感谢

感谢中国人民大学的杜小勇老师、陈晋川老师，是他们帮我打开了科研的大门，让我体会到钻研知识的乐趣。感谢滴滴出行的李奘、卓呈祥、杜龙志以及所有在工作中帮助过我的同事。感谢机械工业出版社华章公司的高婧雅编辑在本书写作过程中给予我的支持和帮助。

特别感谢本书的第一读者——我的爱人高阳，因为写作，我牺牲了很多陪伴她的时间，感谢她的支持和理解，也感谢她帮我进行文稿校对。感谢她一路的支持和陪伴。

何龙

目录

第 1 章
机器学习概述

本章首先介绍何谓机器学习，以及与机器学习相关的基本概念，这是学习和理解机器学习的基础。按照学习方式的不同，机器学习可以分为不同类型，如监督学习、无监督学习、强化学习等，1.1.2 节详细介绍了它们各自的特点和使用场景。其次，借助本章对机器学习应用开发步骤的详细说明，读者能够清晰地了解机器学习应用的开发流程。最后，1.2 节概述了集成学习，以及 XGBoost 如何被提出并在业界广泛应用。

1.1 何谓机器学习

近几年，机器学习可谓是业界最热门的领域之一，AlphaGo 以 4 : 1 的比分击败李世石，人工智能和机器学习一夜火遍世界各地。机器学习离我们并不遥远，甚至可以说已经渗透到我们生活的方方面面。例如，网上购物时，电商网站根据用户偏好为用户推荐商品；Siri 手机语音助手可查询天气、播放音乐；打车时，打车软件帮我们预估行程时间、规划行程路线；点外卖时，外卖 App 将订单分配给附近空闲的骑手等。这些无一不是通过机器学习技术来实现的。

机器学习领域知名学者 Tom M.Mitchell 曾给机器学习做如下定义：

如果计算机程序针对某类任务 T 的性能（用 P 来衡量）能通过经验 E 来自我改善，则认为关于 T 和 P，程序对 E 进行了学习。

通俗来讲，机器学习是计算机针对某一任务，从经验中学习，并且能越做越好的过程。一般情况下，“经验”都是以数据的方式存在的，计算机程序从这些数据中学习。学习的关键是模型算法，它可以学习已有的经验数据，用以预测未知数据。

在很多领域，仅仅靠人很难从诸多信息中将有效信息提取出来的。例如，我们想知道一个人是否会去购买某个电影的电影票。想要知道这个答案，最直接、有效的方法就是去问他

本人，因为他本人的回答是与结果最接近的，也就是相关性最强的一个特征。假如我们并不认识这个人，或并没有条件直接与他本人沟通，那么还有另外一种思路——问他的朋友，他的朋友可能对他比较了解，知道他喜欢哪种类型的影片。但往往这个条件也不一定能达到，因为对于这样的需求场景，更多的可能是影院想知道他的顾客会不会购买某个电影的电影票。而影院所拥有的顾客信息通常是用户的性别、年龄、以往观影记录、消费记录等基本信息。对于普通人来说，通过这些原始数据预测该顾客未来的行为，很难给出一个比较准确的答案。此时便需要机器学习把无序的数据转换成有用的信息，从而解决相关问题。

机器学习横跨了多个学科，包括计算机科学、统计学等，而从事机器学习的人不仅需要扎实的计算机知识和数学知识，还需要对机器学习应用场景下的业务知识非常了解。因此，很多人觉得机器学习门槛很高，还没有开始学习就望而却步了。其实机器学习的入门并没有想象中那么难，当然也不意味着机器学习的技术含量低。机器学习的特点是：**入门门槛低，学习曲线陡**。很多人入门之后容易陷入一种瓶颈状态，很难有更高的突破，所以学习机器学习一定要有耐心和毅力。

学习机器学习所需的基础知识有以下几类：

1）数学：线性代数（矩阵变换）、高等数学；

2）概率分布、回归分析等统计学基础知识；

3）Python、NumPy、Pandas 等数据处理工具；

4）Hadoop、Spark 等分布式计算平台。

读者不要被上面所罗列的知识吓到，因为即使你不具备这些知识，也可以学习机器学习，在学习的过程中随用随查即可。当然，如果已经事先具备了这些知识，那你学习起来一定事半功倍。下面介绍机器学习相关的基本概念。

1.1.1 机器学习常用基本概念

假如我们有一批房屋特征数据，其中包括卧室数量、房屋面积等信息，如表 1-1 所示。

表 1-1 房屋特征数据表

ID	卧室数量	房屋面积（m^2）	房价（万元）
1	2	77	165
2	3	100	190
3	2	90	180
4	3	122	269
5	3	136	295
6	4	173	405

其中，每一条记录称为样本，样本的集合称为一个数据集（data set）。类似卧室数量、房屋面积等列（不包括房价列）称为特征（feature）。房价是比较特殊的一列，它是我们需要预测的目标列。在已知的数据集中，目标列称为标签（label），它可以在模型学习过程中进行指导。并非所有的数据集均包含标签，是否包含标签决定了采用何种类型的机器学习方法（后续会对不同类型的机器学习方法进行介绍）。数据集一般可以分为训练集、验证集和测试集，三者是相互独立的。

- 训练集用于训练和确定模型参数；
- 验证集用于模型选择，帮助选出最好的模型；
- 测试集用于评估模型，测试模型用于新样本的能力（即泛化能力）。

如果机器学习任务的预测目标值是离散值，则称此类任务为分类任务。比如比较常见的垃圾邮件分类系统，类别只有垃圾邮件、非垃圾邮件两类，这是一个分类任务，并且是一个二分类任务（类别只有 2 种）。若类别有多种，则称这类任务为多分类任务。例如预测电影所属类型，其包括动作片、爱情片、喜剧片等多个类别。如果预测值是连续值，则称为回归任务，如表 1-1 中的预测房价。另外，还可以对数据进行聚类，即找到数据的内在结构，发现其中隐藏的规律。例如我们以前看过的电影，即使没有人告诉我们每部电影的类型，我们也可以自己归纳出哪些影片属于喜剧片、哪些属于动作片。

1.1.2　机器学习类型

按照学习方式的不同，可以将机器学习划分为几种类型：监督学习（supervised learning）、无监督学习（unsupervised learing）、半监督学习（semi-supervised learning）、强化学习（reinforcement learning）。

1）**监督学习**，顾名思义，即有“监督”的学习，这里的“监督”指的是输入的数据样本均包含一个明确的标签或输出结果（label），如前述购票预测系统中的“购票”与“未购票”，即监督学习知道需要预测的目标是什么。监督学习如同有一个监督员，监督员知道每个输入值对应的输出值是什么，在模型学习的过程中，监督员会时刻进行正确的指导。监督学习输入的数据称为训练数据。训练数据中的每个样本都由一个输入对象（特征）和一个期望的输出值（目标值）组成，监督学习的主要任务是寻找输入值与输出值之间的规律，例如预测房屋价格系统，输入值是房屋的面积、房间数量等，输出值是房屋价格。监督学习通过当前数据找出房屋面积、房间数量等输入值与房屋价格之间的内在规律，从而根据新的房屋样本的输入值预测房屋价格。

2）**无监督学习**，与监督学习相反，即无监督学习输入的数据样本不包含标签，只能在输入数据中找到其内在结构，发现数据中的隐藏模式。在实际应用中，并非所有的数据都是

可标注的，有可能因为各种原因无法实现人工标注或标注成本太高，此时便可采用无监督学习。无监督学习最典型的例子是聚类。

3）**半监督学习**是一种介于监督学习和无监督学习之间的学习方式。半监督学习是训练数据中有少部分样本是被标记的，其他大部分样本并未被标记。半监督学习可以用来进行预测，模型需要先学习数据的内在结构，以便得到更好的预测效果。

4）**强化学习**是智能体（agent）采取不同的动作（action），通过与环境的交互不断获得奖励指导，从而最终获得最大的奖励。监督学习中数据标记的标签用于检验模型的对错，并不足以在交互的环境中学习。而在强化学习下，交互数据可以直接反馈到模型，模型可以根据需要立即做出调整。强化学习不同于无监督学习，因为无监督学习旨在学习未标记数据间的内在结构，而强化学习的目标是最大化奖励。

1.1.3 机器学习应用开发步骤

开发机器学习应用时，读者可以尝试不同的模型算法，采用不同的方法对数据进行处理，这个过程十分灵活，但也并非无章可循。本节会对机器学习应用开发中的经典步骤进行逐一介绍。

（1）定义问题

在开发机器学习应用之前，先要明确需要解决的是什么问题。在实际应用中，很多时候我们得到的并非是一个明确的机器学习任务，而只是一个需要解决的问题。首先要将实际问题转化为机器学习问题，例如解决公司员工不断收到垃圾邮件的问题，可以先对邮件进行分类，通过机器学习算法将垃圾邮件识别出来，然后对其进行过滤。由此，我们将一个过滤垃圾邮件的现实问题转化为了机器学习的二分类问题（判断是否是垃圾邮件）。

（2）数据采集

数据对于机器学习是至关重要的，数据采集是机器学习应用开发的基础。数据采集有很多种方法，最简单的就是人工收集数据，例如预测房屋价格，可以从和房屋相关的网站上获取数据、提取特征并进行标记（如果需要）。人工收集数据耗时较长且非常容易出错，所以通常是其他方法都无法实现时才会采用。除人工收集数据外，还可以通过网络爬虫从相关网站收集数据，从传感器收集实测数据（如压力传感器的压力数据），从某些 API 获取数据（如交易所的交易数据），从 App 或 Web 端收集数据等。对于某些领域，也可直接采用业界的公开数据集，从而节省时间和精力。

（3）数据清洗

通过数据采集得到的原始数据可能并不规范，需对数据进行清洗才能满足使用需求。例如，去掉数据集中的重复数据、噪声数据，修正错误数据等，最后将数据转换为需要的格