

“工业和信息化领域急需紧缺人才（大数据和人工智能）培养工程”系列丛书



机器学习实战

基于 **Sophon** 平台的机器学习理论与实践

星环科技人工智能平台团队 编著

MACHINE LEARNING IN ACTION

PRINCIPLES AND PRACTICE
BASED ON THE SOPHON PLATFORM



机械工业出版社
China Machine Press

机器学习实战

基于Sophon平台的机器学习理论与实践

人工智能技术的快速发展，带来了相关技术平台和行业应用的繁荣。从 Caffe、CNTK、CoreML 到 TensorFlow 和 TensorRT，从 CPU、GPU 到 TPU、FPGA，从图形处理、视觉识别再到自然语言处理，随着技术体系越来越复杂，应用场景日益广泛，开发门槛也越来越高。在这种情况下，我们非常需要一本关于机器学习的专业图书，不仅讲解关于整体知识体系的理论基础，也可以提供基于人工智能研发实战平台环境的操作指导。

星环科技作为国内大数据和人工智能平台的领航者，其人工智能平台（Sophon）可以让用户快速完成从特征工程、模型训练再到模型上线应用的机器学习全生命周期的开发工作。

星环科技人工智能平台团队编写的这本实战书籍，立意就是阐述团队在使用 Sophon 机器学习平台解决多个业务难题和探索隐藏在其中的数据之美的过程中的积累和沉淀。全书针对各种类型的机器学习问题，基于星环自研的人工智能平台（Sophon）环境，结合实际案例，依照数据、分析、建模的演进历程进行系统化阐述并进行实战讲解。通过阅读和学习本书，读者既能了解人工智能相关的算法原理，也能了解人工智能一些可落地的具体应用场景，此外，本书还适合作为 Sophon 平台工具的使用手册，供平台用户参考。



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

上架指导: 计算机/机器学习

ISBN 978-7-111-64265-7



9 787111 642657

定价: 79.00元

MACHINE LEARNING IN ACTION

PRINCIPLES AND PRACTICE
BASED ON THE SOPHON PLATFORM

机器学习实战

基于 **Sophon** 平台的机器学习理论与实践

星环科技人工智能平台团队 编著



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

机器学习实战: 基于 Sophon 平台的机器学习理论与实践 / 星环科技人工智能平台团队编著.

—北京: 机械工业出版社, 2020.1

(“工业和信息化领域急需紧缺人才(大数据和人工智能)培养工程”系列丛书)

ISBN 978-7-111-64265-7

I. 机… II. 星… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字 (2019) 第 272556 号

机器学习实战

基于 Sophon 平台的机器学习理论与实践

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: 张志铭

责任校对: 殷虹

印刷: 三河市宏图印务有限公司

版次: 2020 年 1 月第 1 版第 1 次印刷

开本: 186mm×240mm 1/16

印张: 14

书号: ISBN 978-7-111-64265-7

定价: 79.00 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

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

作者简介

星环科技人工智能平台团队

星环科技人工智能平台团队由五十多位优秀的研发工程师和算法工程师组成，逾八成具有国内外名校硕士及以上学历。其中研发子团队的工作重心为一站式人工智能建模平台 Sophon；算法子团队则负责基础算法的研发及改进，并在数据挖掘、传统机器学习、计算机视觉、自然语言处理、知识图谱等领域进行多项前瞻性研究以及项目实施落地。星环科技人工智能平台团队一直致力于“把中国人自主研发的领先创新技术赋能全世界各行各业，促进社会可持续发展，通过科技让人类的生活更美好”。目前产品应用已覆盖金融、安防、电力、交通、教育等数十个行业和领域，申请专利近三十个。

本书内容覆盖了机器学习领域从理论到实践的多个主题，总共分为 10 章。

第 1 章为导论，介绍机器学习的背景、定义和任务类型，构建机器学习应用的步骤，以及开发机器学习工作流的方式。

第 2 章详细介绍数据预处理和特征工程技术，并辅以实例进行验证。

第 3~6 章介绍回归模型、分类模型、模型融合和聚类模型，这些内容是机器学习理论和实践中的传统重点。其中不仅介绍各种常见数据类型的处理方法，还针对删失数据进行了专门的综述和实践。

第 7 章介绍机器学习领域较难的图计算话题，并从工业界视角解读如何将图计算落地。

第 8 章针对特征工程、建模过程中大量调参的场景介绍自动机器学习的理论和应用，并细致比较和测试了各种自动特征工程算法在不同数据上的表现。

第 9 章介绍自然语言处理（词向量、序列标注、关键词抽取、自动摘要和情感分析）技术，使用新闻文本数据搭建文本分类的流程。

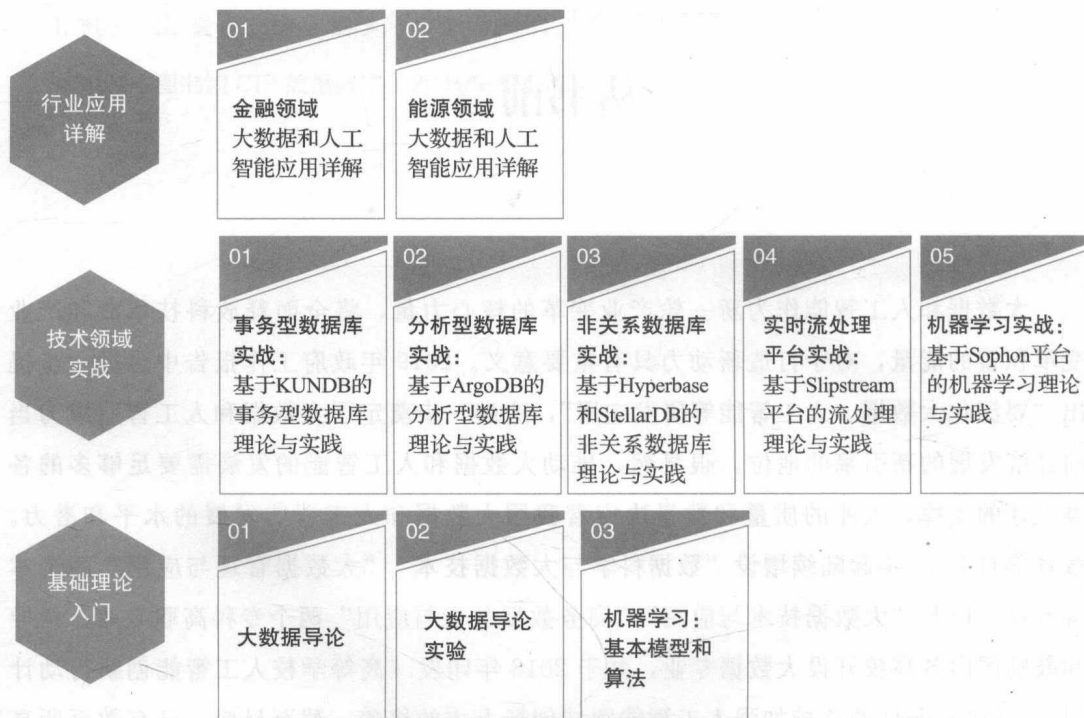
第 10 章介绍计算机视觉中图像分类和目标检测的应用以及车辆检测的落地案例。

本书适合作为高等院校计算机、软件工程、人工智能等相关专业的教学用书，同时也可供从事机器学习相关领域的工程技术人员阅读和参考，帮助他们掌握机器学习相关的算法原理，并能通过专业工具平台快速搭建各类模型，构建机器学习的行业应用。

丛书前言

大数据和人工智能作为新一轮产业变革的核心力量，将全面释放科技革命和产业变革积蓄的能量，对于打造新动力具有重要意义。2019 年政府工作报告中也进一步提出“要深化大数据、人工智能等研发应用”，这进一步奠定了大数据和人工智能成为当前经济发展的新引擎的地位。很显然，推动大数据和人工智能的发展需要足够多的各类人才的支撑，人才的质量和数量决定着我国大数据和人工智能发展的水平和潜力。教育部自 2016 年起陆续增设“数据科学与大数据技术”“大数据管理与应用”两个本科专业，以及“大数据技术与应用”“商务数据分析与应用”两个专科高职专业，指导和鼓励国内各高校开设大数据专业，并于 2018 年印发《高等学校人工智能创新行动计划》，鼓励有条件的高校加强人工智能领域创新人才的培养。截至目前，已有数百所高校获批开设大数据和人工智能相关专业。我们认为，当前大部分大数据和人工智能的人才培养都应该紧贴行业和面向实际应用。从这个意义上说，主要依靠学校力量可能并不能完全满足多样化人才培养的需求。事实上，从大数据和人工智能整个生态来看，尤其在一些相对高端的应用领域，工业界往往走得较快，相应地，相当一部分教育机构则处在追赶的状态。其次，因为广受业界关注，大数据和人工智能相关领域的方法、技术和工具众多，同时也在快速演变，这无疑会让很多应用型初学者无所适从。从这个意义上说，有必要删繁就简，突出主干。最后，我们注意到，不管是大数据技术还是人工智能技术，在整个信息化的体系当中，都不能孤立存在。比如，当前很多大数据和人工智能厂商都逐渐开始将产品放到云端，逐渐推出一些云服务，我们称之为大数据的 3.0 时代。为了支撑逐渐云化的大数据或者人工智能，技术栈中最好也应该包含 DevOps 方法、Kubernetes 容器管理引擎等内容。鉴于此，我们规划了“工业和信

息化领域急需紧缺人才（大数据和人工智能）培养工程”系列教材，按照三个层次来组织大数据和人工智能相关内容。



第一层次主要目标是打基础，因此，主要提供导论性质的基础课程，建设三本教材，即《大数据导论》和《大数据导论实验》以及对应机器学习的《机器学习：基本模型和算法》。

第二层次主要目标是从技术视角提供当前大数据和人工智能相关技术的深入介绍。建设五本教材，内容涉及事务型数据库、分析型数据库、非关系数据库、实时流处理以及机器学习等。

第三层次主要目标是从行业视角提供具体领域的大数据和人工智能应用案例详解，目前规划建设两本教材，分别是《能源领域大数据和人工智能应用详解》和《金融领域大数据和人工智能应用详解》。

本丛书的主要特点如下：

- **面向工业界应用型人才培养的需求来规划教材内容** 不管是丛书顾问团队还是编写人员，我们都优先考虑以一线从业人员为主，试图尽最大可能还原大数据和人工智能应用场景中的各类实际问题，并以问题驱动的方式来组织教材的大部分内容。便于读者加深对于大数据和人工智能相关技术和方法及其在实际业务场景中的应用的~~理解~~。
- **依托现有生态，删繁就简的同时围绕大数据和人工智能主题** 系统化组织教材内容。星环公司作为国内大数据和人工智能另一产品最丰富的供应商之一（2019 信通院大数据产品能力评测），在大数据和人工智能的技术应用、培训、竞赛组织、研究与开发等环境经验丰富，同时形成了闭环以及自有生态。首先，从实践中抽取和提炼问题以及解决问题的经验，形成教材的主要内容，同时也为培训和竞赛组织提供素材。其次，教材可以辅助培训和竞赛的开展，同时，培训和竞赛的组织也为完善教材提供必要的反馈。最后，实践中面临着大量实际问题，则为完善和增强产品提供了研究课题；反之，研究成果也会反哺产品线，更好地服务客户。因此，本系列教材在内容选择，尤其是技术路线选择中，依托了星环公司产品线作为技术主干，同时适当兼顾其他广受好评的一些开源工具和相关技术。
- **按照大数据 3.0 的要求来规划内容** 我们认为，未来大数据和人工智能产品往云端逐步迁移是不可避免的趋势，因此，适当引入云计算、DevOps 以及容器技术等内容是非常有必要的。

本丛书面向多层次的读者，既可以作为高校大数据相关专业的教材，也可以作为一般社会培训教材。特别是，我们联合工信部中国信息通信研究院中国人工智能产业发展联盟、数据中心联盟共同建立大数据人才发展中心，进行大数据人才的培养认证工作，因此，本系列丛书也可以作为“工业和信息化领域急需紧缺人才培养工程”中大数据和人工智能方向的培训教材。

SERIES EDITORIAL BOARD

丛书编委会

“工业和信息化领域急需紧缺人才（大数据和人工智能）培养工程”系列丛书编委组成如下：

□ 顾问团队（按姓氏拼音排序）

- 郭正彪 博士 长江证券互联网金融总部数据总监
刘 速 博士 中国石油规划总院信息技术专家、信息中心数据分析技术部经理
李秀生 四川新网银行股份有限公司 CIO
李昱见 郑州地铁集团有限公司信息管理部副部长
李 喆 武汉农村商业银行信息科技部总经理
邵 栋 副教授 南京大学软件学院副院长
孙建军 教授 南京大学信息管理学院院长
孙煜华 中国南方电网广州供电局高级信息技术专家
孙元浩 星环信息科技（上海）有限公司创始人兼 CEO
王浩学 博士 河南省信息产业投资集团信息中心 CTO
王新民 北京大学大数据分析与应用技术国家工程实验室研究员
王志军 联通公司信息化部副总经理、中国计算机学会大数据专家委员会委员
夏 永 湖北省农村信用社联合社数据中心总经理
张平文 中国科学院院士 北京大学数学科学学院教授

□ 执行主编

荣国平

□ 编委委员（按姓氏拼音排序）

代克 范颖捷 何铁科 李传艺 林晨 刘汪根 吕程 孙乐飞 徐杰 杨俊
杨一帆 张贺 周洪明 朱珺辰 等

P R E F A C E

前 言

人工智能技术的快速发展，带来了技术平台和行业应用的繁荣，从 Caffe、CNTK、CoreML 到 TensorFlow、TensorRT，从 CPU、GPU 到 TPU、FPGA、ARM，从图形处理、视觉识别到自然语言处理，技术体系越来越复杂，开发门槛越来越高；大量的技术人员需要不断授受技术更新，更多的应用需要考虑额外的迁移成本，更多的市场需要投入大量的资源以充分体现人工智能赋予的价值。

目前产业界开始出现少量技术使用门槛低、应用开发方便的机器学习平台（Machine Learning Platform，MLP）或者数据科学平台（Data Science Platform，DSP），但这些平台大部分还局限在特定行业的有限算法应用，需要不断进行架构优化、模型扩展和算法增强，提供多种场景下的应用迁移工具，才能形成较为成熟的产品化平台。

星环科技作为国内大数据和人工智能平台的领航者，自 2013 年成立以来，专注于企业级容器云计算、大数据和机器学习核心平台的研发和服务，拥有一批来自国内外著名高科技企业和科研院校的优秀专业人才，是国内大数据领域最早掌握核心技术的企业，也是最早开展机器学习平台理论与实践的公司之一，产品在政府、金融、公安等行业得到大规模应用。

星环科技人工智能平台 Sophon 是从大数据到人工智能演进过程中诞生的一款创新性机器学习技术平台。用户可以基于该平台快速完成从特征工程、模型训练到模型上线的机器学习全生命周期开发工作。

Sophon 平台具有以下技术特点:

- 采用去中心化的全分布式架构、性能线性扩展, 满足海量数据处理模式下的快速训练和精准推理要求。
- 一站式的机器学习集成开发平台, 支持自动化开发、图形化操作及可视化建模, 可快速构建行业应用解决方案。
- 支持多种复杂算法, 支持自定义模型和算法导入, 可适应多种特定应用场景的复杂建模和模型迁移要求。
- 集成大量面向行业领域的分析工具, 如实体画像、视频分析、自然语言处理等, 便于第三方应用快速定制开发。
- 支持深度学习的知识图谱, 能够便捷实现含图结构的应用建模, 支持实体间多关系图的分析展示和演进变化, 发现更有价值的图谱关系。

随着使用机器学习平台的用户越来越多, 应用场景日益广泛, 非常需要一本关于机器学习理论总结和实践指导的专业图书, 不仅可以讲解整体知识体系的理论基础, 也可以作为使用星环人工智能平台 (Sophon) 工具的指导手册。

目前市面上销售的机器学习相关书籍, 要么偏重原理介绍和公式推导, 要么重点描述开源算法的实现调用, 无法满足二者兼顾的要求。为此, 我们结合理论分析和实践指导要求, 编写了这本面向机器学习一线工程技术人员的专业书籍。它既能帮助读者深入理解相关算法原理, 也有助于读者学会利用专业工具平台快速搭建模型, 构建机器学习的行业应用。

本书内容覆盖了机器学习领域从理论到实践的多个课题, 总共分为 10 章。

第 1 章为导论, 介绍机器学习的背景、定义和任务类型, 构建机器学习应用的步骤, 以及开发机器学习工作流的方式。

第 2 章详细介绍数据预处理和特征工程, 并辅以实例进行验证。

第 3~6 章介绍回归模型、分类模型、模型融合、聚类模型, 这些内容是机器学习

理论和实践中的传统重点。其中不仅介绍对各种常见数据类型的处理方法，还针对缺失数据进行了专门的综述和实践。

第7章介绍机器学习领域较难的图计算，并从工业界视角解读如何将图计算落地。

第8章针对特征工程、建模过程中大量调参的场景介绍自动机器学习的理论和应用，并细致比较和测试了各种自动特征工程算法在不同数据上的表现。

第9章介绍自然语言处理（词向量、序列标注、关键词抽取、自动摘要和情感分析），使用新闻文本数据搭建文本分类的流程。

第10章介绍计算机视觉中图像分类和目标检测的应用以及落地案例（车辆检测）。

书中的第1~2章是基础内容，建议读者认真阅读，其他章节则可根据需要选择性地阅读。

全书由孙元浩和杨俊统一主持和整理，参与编写的作者还包括杨一帆、裴瑞光、林木丰、乐向楠、陆增翔、蒲瑜琪、李祥祥、曾宪宇、赵文谦、林晨、浦锦毅、安磊、许凯琪、孙乐飞和吴香莲。

本书从雏形到定稿，历时近一年，非常感谢参与本书编纂校对工作的算法工程师和架构师，没有他们无私的理论分享和实践指导，本书是难以高质量完成的。在此我们对所有编者表示衷心的感谢和敬意。

孙元浩

2019年7月

CONTENTS

目 录

丛书前言	2.1.1 探索性数据分析	14
丛书编委会	2.1.2 数值特征	18
前言	2.1.3 类别特征	22
	2.1.4 时间特征	24
	2.1.5 文本特征	24
	2.1.6 过滤方法	26
第1章 机器学习导论	2.1.7 封装方法	29
1.1 什么是机器学习	2.1.8 嵌入方法	30
1.1.1 机器学习的背景	2.1.9 自动特征工程	31
1.1.2 机器学习的定义	2.2 交互式数据预处理	31
1.1.3 机器学习的任务	2.3 本章小结	32
类型		
3		
1.1.4 构建机器学习应用	第3章 回归模型	35
的步骤	3.1 回归任务概述	35
5	3.2 回归算法原理	36
1.2 开发机器学习工作流的	3.2.1 线性回归	36
方式	3.2.2 决策树回归	38
7	3.2.3 生存回归	41
1.2.1 数据导入	3.3 Sophon 案例	42
7	3.4 本章小结	45
1.2.2 流程搭建		
9		
第2章 数据预处理与特征工程		
2.1 特征提取		
13		

第4章 分类模型	47	5.3.2 建模过程	71
4.1 分类任务概述	47	5.3.3 结果分析	74
4.2 分类算法原理	48	5.4 本章小结	75
4.2.1 逻辑回归	48	第6章 聚类模型	77
4.2.2 因子分解机	51	6.1 聚类任务概述	77
4.2.3 XGBoost	52	6.2 聚类算法原理	78
4.3 使用 Sophon 建立分类模型	52	6.2.1 K-Means	78
4.3.1 场景介绍	52	6.2.2 模糊 C 均值	79
4.3.2 建模过程	54	6.2.3 Canopy	81
4.3.3 结果分析	56	6.2.4 高斯混合	83
4.4 本章小结	61	6.3 聚类模型实例	85
第5章 模型融合	63	6.3.1 场景介绍	85
5.1 集成学习理论	63	6.3.2 建模过程	87
5.1.1 集成学习基本概念	63	6.3.3 结果分析	91
5.1.2 个体学习器	63	6.4 本章小结	91
5.1.3 基学习器集成	64	第7章 图计算	93
5.1.4 常用的集成学习方法	65	7.1 背景和问题描述	93
5.2 常用融合方法	70	7.2 常用算法介绍	95
5.2.1 平均法	70	7.2.1 PageRank	95
5.2.2 学习法	70	7.2.2 标签传播	97
5.3 使用 Sophon 进行模型融合	71	7.2.3 中心性检测	98
5.3.1 场景与数据集介绍	71	7.2.4 图嵌入	99
		7.3 落地案例	102
		7.3.1 场景介绍	102
		7.3.2 建模过程	103
		7.3.3 结果分析	107
		7.4 本章小结	109

第8章 自动机器学习	111	9.2 使用 Sophon 建立自然语言 处理模型	154
8.1 场景介绍	111	9.2.1 场景介绍	154
8.2 自动特征工程	113	9.2.2 建模流程	154
8.2.1 自动多表特征 扩展	114	9.2.3 模型评估	155
8.2.2 自动特征构建	122	9.3 落地案例	156
8.3 建模过程	127	9.4 本章小结	159
8.4 结果分析	129	第10章 计算机视觉	161
8.5 真实测试案例	135	10.1 计算机视觉概述	161
8.5.1 数据集	135	10.2 计算机视觉算法原理	161
8.5.2 前置设置	136	10.2.1 图像分类	161
8.5.3 测试结果分析	137	10.2.2 目标检测	163
8.5.4 Abalone 和 Airfoil Self-Noise 数据集 的增强测试	138	10.3 计算机视觉模型示例	166
8.5.5 结论	139	10.3.1 图像预处理	166
8.6 本章小结	140	10.3.2 图像分类算法 建模	166
第9章 自然语言处理	141	10.3.3 目标检测算法 建模	169
9.1 自然语言处理算法 原理	141	10.4 落地案例	171
9.1.1 词向量	141	10.5 本章小结	173
9.1.2 序列标注	143	附录 A 企业级人工智能应用 平台 Sophon	175
9.1.3 关键词抽取	147	参考文献	206
9.1.4 文本自动摘要	148		
9.1.5 文本情感分析	152		