英特尔® FPGA中国创新中心系列丛书

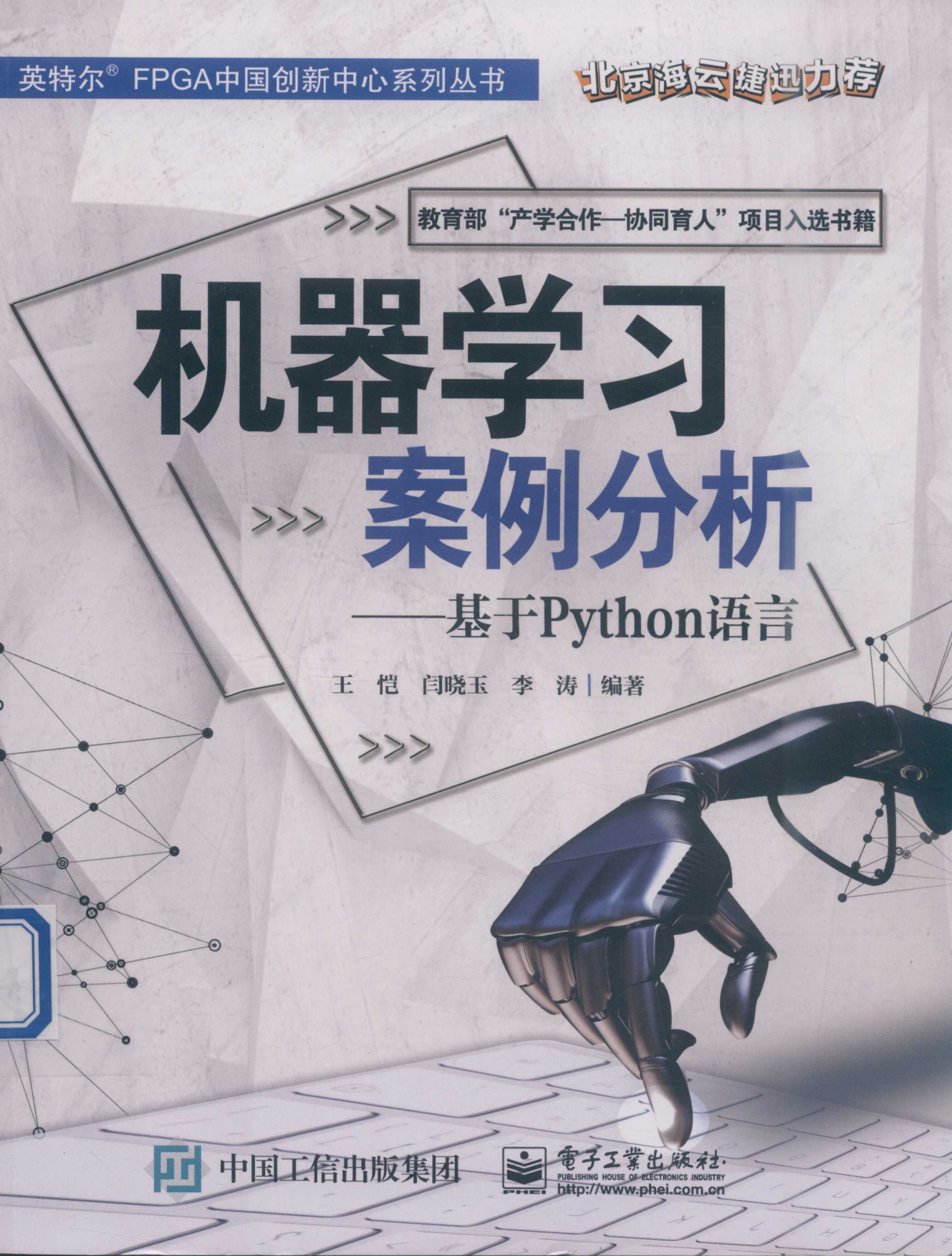
北京海云捷迅力荐

教育部“产学合作—协同育人”项目入选书籍

机器学习案例分析

——基于Python语言

王 恺 闫晓玉 李 涛 | 编著



中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

英特尔®FPGA 中国创新中心系列丛书

机器学习案例分析

——基于Python语言

王 恺 闫晓玉 李 涛 | 编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书共 5 章内容，主要结合目前流行的人工智能编程语言 Python 对机器学习案例进行分析，介绍机器学习的相关理论，并展示使用机器学习方法解决实际应用问题的具体过程。本书包括基础知识、分类案例、聚类案例、回归预测案例和综合案例，力争通过通俗易懂的案例和代码分析使读者快速掌握机器学习的具体应用方法。本书既适合计算机相关专业人员，也适合非计算机相关专业人员阅读。理论性强，较难理解的内容统一放在了附录 A 中，这部分内容适合具备一定理论基础、对机器学习理论推导有兴趣的读者。

本书可以作为我国高校计算机专业学生和非计算机专业理工科学生机器学习入门课程的教材。

图书在版编目（CIP）数据

机器学习案例分析：基于 Python 语言 / 王恺，闫晓玉，李涛编著. —北京：电子工业出版社，2020.3
（英特尔®FPGA 中国创新中心系列丛书）

ISBN 978-7-121-38181-2

Ⅰ. ①机… Ⅱ. ①王… ②闫… ③李… Ⅲ. ①机器学习 ②软件工具—程序设计 Ⅳ. ①TP181 ②TP311.561

中国版本图书馆 CIP 数据核字（2019）第 279358 号

责任编辑：刘志红（lzhmails@phei.com.cn）　　特约编辑：王　纲
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×980　1/16　印张：20.5　字数：448.7 千字
版　　次：2020 年 3 月第 1 版
印　　次：2020 年 3 月第 1 次印刷
定　　价：98.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254479，lzhmails@phei.com.cn。

指导委员会

张征宇　李　华　田　亮　张　瑞

工作委员会

王　恺　闫晓玉　李　涛

序

众所周知，我们正在进入一个全面科技创新的时代。科技创新驱动并引领着人类社会的发展，从人工智能、自动驾驶、5G，到精准医疗、机器人等，所有这些领域的突破都离不开科技的创新，也离不开计算的创新。从 CPU、GPU，到 FPGA、ASIC，再到未来的神经拟态计算、量子计算等，英特尔正在全面布局未来的端到端计算创新，以充分释放数据的价值。中国拥有巨大的市场和引领全球创新的需求，其产业生态的全面性及企业创新的实力、活力和速度都令人瞩目。英特尔始终放眼长远，以丰富的生态经验和广阔的全球视野，持续推动与中国产业生态的合作共赢。以此为前提，英特尔在 2018 年建立了英特尔® FPGA 中国创新中心，与 Dell、海云捷迅等合作伙伴携手共建 AI 和 FPGA 生态，并通过组织智能大赛、产学研对接及培训认证等方式，发掘优秀团队，培养专业人才，孵化应用创新，加速智能产业在中国的发展。

该系列丛书是英特尔® FPGA 中国创新中心专为 AI 和 FPGA 领域的人才培养和认证而设计编撰的系列丛书，非常高兴作为英特尔® FPGA 中国创新中心总经理为丛书写序。同时也希望该系列丛书能为中国 AI 和 FPGA 相关产业的生态建设和人才培养添砖加瓦！

英特尔® FPGA 中国创新中心总经理　张 瑞

2019 年秋

张 瑞

张瑞先生现任英特尔® FPGA 中国创新中心总经理，总体负责中国区芯片对外合作，以及自动驾驶和 FPGA 等领域的生态建设。同时也兼任（中国）汽车电子产业联盟副理事长和副秘书长的职务，致力于推动包括 5G、机器视觉、传感器融合和自主决策等多项关键自动驾驶相关技术在中国的落地和合作。

张瑞先生拥有多年世界领先半导体公司的从业经历。在加入英特尔之前，曾在瑞萨电子和飞思卡尔半导体担任多个关键技术和管理职务。

张瑞先生曾于 2008 年编写并出版过科学技术类图书《Coldfire 处理器深入浅出》一书。

FOREWORD

前言

机器学习（Machine Learning，ML）是人工智能的一个分支，它是一门多领域交叉学科，专门研究计算机怎样模拟或实现人类的学习行为，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习方法可以根据经验数据自动完成模型参数学习，而不需要人为设定规则，大幅降低了人工分析的工作量和难度，已成为目前解决人工智能相关问题的主要方式。另一方面，作为目前流行的人工智能编程语言，Python具有简单易学、免费开源、跨平台性、高层语言、面向对象、丰富的库、胶水语言等优点，不仅大量计算机专业人员使用Python进行人工智能算法快速开发，而且非计算机专业人员也利用Python结合封装好的人工智能算法解决其专业问题。

本书由南开大学计算机学院的教师结合多年教学经验和人工智能教育的发展需要编著而成，可作为我国高校计算机专业学生和非计算机专业理工科学生机器学习入门课程的教材。本书从案例出发，通过具体问题向读者直观展示了利用机器学习方法解决人工智能问题的详细步骤，以及利用Python程序设计语言快速应用机器学习方法解决人工智能问题的具体过程，力争使读者在有限时间内快速掌握每种机器学习方法适合解决的人工智能问题。我们也提供了一些机器学习的理论分析和推导过程，使对机器学习理论有兴趣的读者能够对相关知识有一个初步认识和掌握，为读者学习更深层次的机器学习理论打下了一个良好的基础。

在利用本书学习机器学习相关知识时，建议读者一定要多思考、多分析、多动手实践。当阅读一个具体案例分析时，要认真思考每一个案例的具体解决步骤，从中学习利用机器学习方法解决人工智能问题的一般过程。当阅读案例代码时，要自己梳理程序结构，在计算机上重现该程序的运行结果，通过逐语句执行，并查看变量状态的方式分析各语句的作用。只有这样，才能真正掌握利用机器学习解决人工智能问题的具体方法和流程，也才能真正做到熟练运用机器学习方法解决实际遇到的应用问题。

本书的特色包括：（1）以案例为主线，引入相关知识点，使读者在具体应用中快速掌握机器学习解决人工智能问题的具体方法和流程。（2）强调应用性，同时也给出了必要的机器学习理论及推导，既适合作为计算机相关专业人员进行机器学习的入门读物，也适合对“利用机器学习方法解决人工智能问题”有兴趣的非计算机相关专业人员阅读。

（3）将简单易懂的案例代码分析和理论性强、较难理解的内容分开，方便读者根据实际需求进行相关章节的阅读。

本书包括 5 章和附录 A，下面简单介绍各部分内容。

第 1 章，首先给出了机器学习的基本概念及分类。其次，从 Python 编程环境、基本数据类型、分支语句和循环语句、函数、类和对象、文件读写、异常处理等方面使读者快速掌握 Python 程序设计语言的入门知识。再次，介绍了应用机器学习解决人工智能问题时常用的 Python 第三方库，包括 NumPy、SciPy、Pandas、Matplotlib 和 Scikit-learn。最后，给出了网络爬虫及信息提取、股票数据图表绘制两个案例分析，使读者快速掌握使用 Python 解决实际问题的方法。

第 2 章给出了 4 个分类案例。首先是员工离职预测案例，分别使用基本线性分类器、最小二乘分类器、感知器和逻辑回归分类器，根据员工对公司满意度、最新考核评估等特征对员工是否离职进行了预测。其次是 Iris（鸢尾花）数据分类案例，分别使用 k 近邻分类器和决策树分类器，根据花萼长度、花萼宽度等特征对鸢尾花的种类进行了预测。再次是新闻文本数据分类案例，介绍了文本分词、去停用词、文本表示与特征选择等，介绍了文本数据预处理的方法和具体实现，并分别使用朴素贝叶斯分类器、支持向量机分类器和 Adaboost 分类器，对搜狐新闻数据（SogouCS）完成了国内、国际、体育、社会、娱乐等 12 个频道的分类预测。最后是手写数字图像识别案例，使用 BP 神经网络，基于 MNIST 数据集完成了对神经网络模型的训练和测试。

第 3 章给出了 2 个聚类案例。首先是人脸图像聚类案例，结合 k 均值聚类和 PCA 降维，对 ORL 人脸数据集的部分类别数据进行了聚类分析。然后是文本聚类案例，介绍了极大似然估计、隐变量和高斯混合模型（GMM）的基础知识，并实现 GMM 算法完成两类搜狐新闻的聚类分析。

第 4 章给出了 2 个回归预测案例。首先是房价预测案例，分别使用线性回归和岭回归模型，对 Kaggle 上的 housing 数据集完成了房价预测分析，同时也通过比较展示了不同数据预处理方法和特征选取方法对模型性能的影响。然后是股票走势预测案例，介绍了长短周期记忆网络（LSTM）的基本原理，并利用 TensorFlow 搭建 LSTM 网络，完成了股票开盘价、收盘价、最高价、最低价和成交量的预测。

第 5 章给出了 2 个综合案例。首先是场景文本检测案例，使用传统文本检测的方法

和适当的文本识别库，实现一个能在较复杂的街景中提取文字信息的简易 Demo 程序。作为一个场景文本检测的入门级案例，本案例各处理步骤所使用的方法都比较简单。对场景文本检测问题感兴趣的读者，可阅读近几年 CVPR、ICCV 等顶级会议上关于场景文本检测的论文，以获取相关问题的最新方法。然后是面部认证案例，介绍了 Siamese（孪生）网络的基本原理，基于 TensorFlow 实现了该网络，基于 LFW 人脸数据库完成了模型训练和测试，并搭建面部认证 Demo 程序进行了模型的具体应用方法。通过本章内容，读者应对基于机器学习的人工智能软件系统的构建过程有一个基本的认识。

附录 A 给出了理论性强、较难理解的内容。具体包括逻辑回归分类器原理介绍、自己编程实现决策树分类器、支持向量机的数学推导、Adaboost 的数学推导和代码实现、神经网络的数学推导和代码实现、期望最大化算法和高斯混合模型，以及基于波士顿房价数据集的房价预测代码实现。读者可根据自己的实际需求选择部分内容进行学习。

本书的编写分工如下：王恺负责 1.1 节、第 5 章及附录 A 的编写，并完成全书统稿和定稿工作；闫晓玉负责 1.2～1.6 节及第 2 章的编写；李涛负责第 3、4 章的编写。

在本书的编写过程中，南开大学计算机学院 2019 级研究生马志、卜旺、周可可帮助收集整理了第 2～4 章的案例，南开大学计算机学院 2015 级本科生周睿、龚航提供了场景文本检测和面部认证两个综合案例，电子工业出版社有限公司的刘志红编辑给予了大力支持，在此表示真诚的感谢！

本书还参考了国内外的一些机器学习方面的书籍及大量的网上资料，力求有所突破和创新。由于能力和水平所限，书中出现的不妥甚至错误之处，恳请读者指正。

作者

2019 年 12 月于南开园

反侵权盗版声明

CONTENTS

目录

第 1 章

基 础 知 识

1.1 机器学习简介

1.1.1 基本概念

机器学习（Machine Learning，ML）是人工智能的一个分支，它是一门多领域交叉学科，专门研究计算机怎样模拟或实现人类的学习行为，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。利用机器学习方法解决实际问题时，涉及模型结构设计、学习目标（也称优化目标、目标函数或损失函数）设计、优化算法设计等方面的工作。机器学习的目标是根据已有数据（训练数据，也称训练样本）设计模型并学习模型参数，使得学习后的模型能够在未知数据（测试数据，也称测试样本）上展现出较好的性能（具有较低的泛化误差，或具有较强的泛化能力）。需要注意，在进行模型设计和参数学习时只能使用训练数据，而不能使用任何测试数据。

机器学习模型可简单表示为

$$\boldsymbol{y} = f(\boldsymbol{x};\boldsymbol{\theta}) \tag{1.1}$$

其中，f 是机器学习模型的数学表示（一个映射函数），$\boldsymbol{x}$ 是模型的输入，$\boldsymbol{y}$ 是模型的输出，$\boldsymbol{\theta}$ 是模型的参数。模型设计和参数学习过程，实际上就是根据训练数据进行映射函数 f 的设计，并按预先定义的优化目标（如预测输出与目标输出之间的平方误差）进行参数 $\boldsymbol{\theta}$ 的学习。模型应用过程，实际上就是根据设计好的映射函数 f 及学习好的参数 $\boldsymbol{\theta}$，对于一个数据通过模型给出其预测输出。例如，对于 2.2 节将要介绍的鸢尾花分类问题，输入数据 $\boldsymbol{x}$ 是由花萼长度、花萼宽度、花瓣长度和花瓣宽度组成的一个包含 4 个元素的特征向量（此时称该数据的特征维度为 4），而目标输出数据 $\boldsymbol{t}$ 则是某个鸢尾花子类（山鸢尾、变色鸢尾或维吉尼亚鸢尾，通常用整数表示不同类别，如 0、1、2 等）；通过设计模型及基于训练数据的模型参数学习，使模型能够根据输入的测试数据 $\boldsymbol{x}'$，得到预测输出数据 $\boldsymbol{y}'$，并且 $\boldsymbol{y}'$ 与目标输出数据 $\boldsymbol{t}'$ 应尽可能接近。

在机器学习模型的设计中，需要避免两种情况，即欠拟合和过拟合。如图 1-1 所示，是欠拟合和过拟合的一个简单示例。所谓欠拟合，是指所设计的机器学习模型过于简单，无法表示数据中蕴含的复杂规律。出现欠拟合情况时，机器学习模型在训练数据和测试

数据上的性能相近，但均表现较差。所谓过拟合，是指所设计的机器学习模型过于复杂，其能够完美地对训练数据进行拟合，但在训练过程中未使用的测试数据上表现则很差。出现过拟合情况时，机器学习模型在训练数据上性能很好，但在测试数据上性能很差。无论是欠拟合，还是过拟合，都会使得模型在测试数据上表现出不好的性能（较高的泛化误差，或较差的泛化能力），无法满足实际应用需要。因此，如何设计复杂度适中的机器学习模型，使其具有较强的泛化能力（模型在测试集上有较好的表现），是机器学习中一个非常重要的问题。

为了能够在不使用任何测试数据的情况下，设计出复杂度适中的机器学习模型，在实际应用中通常会将可用于训练的数据进一步分为两部分，分别是训练数据和验证数据。**验证数据**仅用于预测模型的泛化能力，而不参与模型的参数学习过程。当可用于训练的数据本身就很少时，通常采用 K 折交叉验证方法来进行模型的设计。所谓 K 折交叉验证，是指将可用于训练的数据近似等分为 K 份，每次训练时使用其中 K-1 份作为训练数据进行模型参数学习，而没有参与训练的那一份作为验证数据，用于进行模型泛化能力的预测。K 份数据中的每一份都用作一次验证数据后，K 次实验结果的平均值即该模型泛化能力的预测依据。

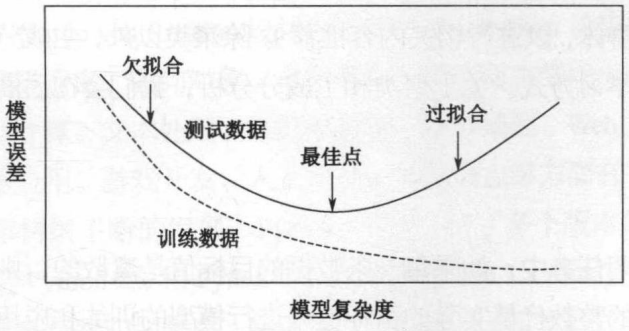


图 1-1 欠拟合和过拟合示例

1.1.2 机器学习分类

从不同的角度，可以对机器学习方法进行不同的分类。从训练数据是否包含目标值的角度，机器学习可以分为有监督学习方法和无监督学习方法；从目标值是不是连续值的角度，机器学习可以分为用于分类任务的方法和用于回归任务的方法。这里简单介绍

一下这些方法的基本概念。

1. 有监督学习

在有监督学习中，每一条训练数据既包含特征向量 $\boldsymbol{x}$，也包含目标值 $\boldsymbol{t}$（目标值可以是单个数值，也可以是一个向量）。通过预先设计好的机器学习模型和优化目标函数，根据这些训练数据进行模型参数学习，使得每一条训练数据的特征向量输入模型后，模型能够给出与目标值尽可能接近的预测值（当然，这里也要注意避免前面所提到的过拟合问题），即使得

$$\sum_{s \in S} D(y^s, t^s) \tag{1.2}$$

尽可能小。其中，S 是训练数据集合，y^s 是机器学习模型对训练数据 s 的预测输出，t^s 是训练数据 s 的目标值，D 是某种距离度量函数（如欧氏距离等）。

2. 无监督学习

在无监督学习中，每一条训练数据仅包含特征向量 $\boldsymbol{x}$，而没有目标值 $\boldsymbol{t}$。聚类（Clustering）是无监督学习的一个重要应用，其自动根据数据之间的相似度，对数据进行分类，从而发掘数据之间的关联关系（如通过分析社交网站上用户与用户之间的关系，将用户分成不同的群体，以进行相关内容推荐）。除聚类以外，主成分分析这种特征降维方法也采用无监督学习方式。关于聚类和主成分分析，我们会在后面介绍更详细的信息和具体使用方法。

3. 分类

在一个机器学习任务中，如果每一条数据的目标值是离散的，则该任务是一个分类任务。通常用不同的整数代替实际的目标值来进行模型的训练和应用。例如，假设有若干物体的图片，每一幅图片的目标值是狗、猫、轮船、飞机中的一个，则可以将目标值编码为 0、1、2、3，其对应关系是 0→狗、1→猫、2→轮船、3→飞机。我们的任务就是设计并训练模型，使其可以对输入的图片产生 0～3 的整数输出，而 0～3 这 4 个整数分别对应 4 种不同的物体。

需要注意，对于分类任务，通常也使用 One-Hot（独热）向量编码形式表示目标值，One-Hot 是指向量中只有一个元素的值为 1，其余元素的值均为 0。例如，对于前面提到

的图片分类的例子，可以将狗、猫、轮船、飞机这 4 个目标值分别编码为(1,0,0,0)、(0,1,0,0)、(0,0,1,0)和(0,0,0,1)。

4. 回归

在一个机器学习任务中，如果每一条数据的目标值是连续的，则该任务是一个回归任务。例如，假设要对某种产品的价格进行预测，该产品的价格是连续值，因此，该问题是一个回归问题。需要注意，因为计算机通常用有限的二进制数来表示数据，所以计算机中任何类型的数据实质上都是可数的。通常来说，如果一个任务的目标值在一定精度下可以连续取值，则认为该目标值是连续的。

可见，回归任务和分类任务的区别就在于目标值是连续的，还是离散的。在实际设计机器学习模型时，很多机器学习模型既可以用于回归任务，也可以用于分类任务。

1.2 Python 基础

Python 语言诞生于 1990 年，由荷兰 CWI（Centrum Wiskunde & Informatica，数学和计算机研究所）的 Guido van Rossum 设计并领导开发。Python 语言具有简单易学、免费开源、跨平台、高层语言、面向对象、丰富的库、胶水语言等优点，已在系统编程、图形界面开发、科学计算、文本处理、数据库编程、网络编程、Web 开发、自动化运维、金融分析、多媒体应用、游戏开发、人工智能、网络爬虫等方面有着广泛的应用。

经过 20 多年持续不断的发展，Python 语言经历了多个版本的更迭。目前使用的 Python 版本主要是 Python 2.x 和 Python 3.x。但是 Python 3.x 并不完全兼容 Python 2.x 的语法，所以如果没有特殊应用需求，建议使用 Python 3.x 版本。

1.2.1 Python 编程环境

在 Linux、Windows、MacOS 等平台上，都可以安装 Python 语言环境以支持 Python 程序的运行。但由于每个人使用 Python 的应用场景不一样，设置 Python、安装附加包，并没有一个统一的解决方案，这里将给出 Windows 和 MacOS 系统中详细的 Python 安装