“十三五”普通高等教育规划教材

机器学习及其应用

汪荣贵 杨 娟 薛丽霞 编著

提供电子课件

http://www.cmpedu.com

机械工业出版社
CHINA MACHINE PRESS

“十三五”普通高等教育规划教材

机器学习及其应用

汪荣贵　杨　娟　薛丽霞　编著

机 械 工 业 出 版 社

本书比较系统地介绍机器学习的基础理论与应用技术。首先，介绍掌握机器学习理论和方法所必须具备的基础知识，包括机器学习的基本概念与发展历程、模型构造与优化的基本方法；然后，介绍和讨论监督学习、无监督学习、集成学习、强化学习等传统机器学习理论与方法；最后，在详细探讨神经网络与深度学习基本理论的基础上，介绍深度卷积网络、深度循环网络、生成对抗网络等若干典型深度学习模型的基本理论与训练范式，分析讨论深度强化学习的基本理论与方法。本书站在高年级本科生和低年级硕士研究生的思维角度编写，尽可能用朴实的语言深入浅出地准确表达知识内容，着重突出机器学习方法的思想内涵和本质，使得广大读者能够掌握全书主要内容。

本书每章均配有一定数量的习题，适合作为智能科学与技术、数据科学与大数据技术、计算机类相关专业的本科生或研究生的机器学习入门级教材，也可供工程技术人员和自学的读者学习参考。

本书配套授课电子课件，需要的教师可登录 www.cmpedu.com 免费注册，审核通过后下载，或联系编辑索取（QQ：2850823885；电话：010-88379739）。

图书在版编目(CIP)数据

机器学习及其应用/汪荣贵，杨娟，薛丽霞编著．--北京：机械工业出版社，2019.8
“十三五”普通高等教育规划教材
ISBN 978-7-111-63202-3

Ⅰ.①机… Ⅱ.①汪… ②杨… ③薛… Ⅲ.①机器学习-高等学校-教材 Ⅳ.①TP181

中国版本图书馆 CIP 数据核字(2019)第 147229 号

机械工业出版社(北京市百万庄大街 22 号 邮政编码 100037)
策划编辑：郝建伟 责任编辑：郝建伟 陈崇昱
责任校对：张艳霞 责任印制：孙 炜

天津嘉恒印务有限公司印刷

2019 年 8 月第 1 版·第 1 次印刷
184mm×260mm·24.75 印张·615 千字
1-2500 册
标准书号：ISBN 978-7-111-63202-3
定价：79.00 元

电话服务
客服电话：010-88361066
010-88379833
010-68326294
封底无防伪标均为盗版

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
金 书 网：www.golden-book.com
机工教育服务网：www.cmpedu.com

前　言

人类文明由早期农耕时代经历漫长演化进入工业时代，再由工业时代进一步发展迈入当今的信息时代。数字化、网络化和智能化是信息时代的基本特征，将给人类文明带来科学技术水平上的巨大提升，令社会的方方面面产生深刻的变革，使得当代人们的生活和工作更加舒适、便捷。目前，作为引领信息社会发展动力的信息技术已经历了数字化和网络化阶段，正朝着智能化方向快速发展，人工智能技术在全社会得到前所未有的重视和广泛应用，并以前所未有的速度向前飞跃发展。为顺应时代发展潮流和把握发展机遇，我国及时制定并推出了新一代人工智能发展规划，把人工智能发展放在国家战略层面进行系统布局，使得人工智能成为新一轮产业变革的核心驱动力。目前，人工智能的理论研究和应用开发是一个非常重要的优先发展方向。

人工智能作为人脑器官的延伸，主要目标是通过计算机模拟人类大脑的某些思维方式或智能行为，如推理、证明、识别、感知、认知、理解、学习等思维方式或活动，使得计算机能够像人类一样进行思考。从外部环境中获得知识和经验的学习能力是人类的一项基本思维能力，机器学习要解决的问题就是如何使得机器具有与人类类似的学习能力，使得机器系统能够较好地了解外部环境并能够适应外部环境的变化。机器学习为人工智能系统提供了基础性的核心算法支撑，人工智能系统主要使用机器学习技术解析外部环境数据，从数据中获取知识和模型参数，获得可用于决策或预测的数学模型。要想学好人工智能，首先必须牢固掌握机器学习的基础理论与应用技术。

机器学习的主要目标是通过计算手段从经验数据等先验信息中获得一个具有较好泛化性能的数学模型，并使用该模型完成预测、分类和聚类等机器学习任务。因此，机器学习的研究对象主要是从经验数据等先验信息中产生或构造模型的训练学习算法，或者说机器学习是一门关于训练学习算法设计理论与应用技术的学问。我们知道，算法设计是一种思维的艺术，需要一定的抽象思维能力和数学知识。机器学习算法更是如此，不仅涉及微积分、数理逻辑、数理统计、矩阵计算、图论等数学知识，而且涉及众多最优化理论与方法，这些为广大初学者掌握机器学习知识带来一定困难。为较好地满足广大读者系统地掌握机器学习入门性基础理论与应用技术的需要，本书的编写着重考虑如下两个要点：

第一，注重知识体系的完备性。作为人工智能的核心技术，机器学习随着人工智能的产生而产生并随着人工智能理论的发展而发展，目前已形成一个非常庞大且正在快速延伸发展的知识体系，众多学习算法精彩纷呈、目不暇接、不胜枚举。本书通过深度凝练机器学习的现有知识体系，构造一套相对完备的入门级机器学习基础理论与应用技术，在基本涵盖连接学习、符号学习和统计学习这三种基本学习类型的基础上，注重突出对基本理论与关键技术的介绍和讨论。

第二，强调可读性和可理解性。本书站在高年级本科生和低年级硕士研究生的思维角度编写，在保证表达准确的前提下，尽可能用朴实的语言深入浅出地介绍机器学习理论及相关算法设计技术，尽可能细致地阐述理论与算法的思想内涵和本质。通过学习书中实际算例的具体演算过程，读者能够对机器学习理论与算法有更加清晰、全面的理解。需要说明的是，本书并没有为了增加可读性而降低应有的内容深度，只是通过比较恰当的方式把相关知识表达得更加清

楚明白，使得广大读者能够通过自己的努力就可以比较轻松地掌握机器学习的基本理论与应用技术。

本书比较系统地介绍机器学习的入门性基础理论与应用技术，内容主要包括机器学习的基本知识、模型估计与优化的基本方法、监督学习和无监督学习方法、集成学习、强化学习方法、神经网络与深度学习方法，将机器学习的经典内容与深度学习等前沿内容有机地结合在一起，形成一套相对完整、统一的知识体系，并在每个章节穿插相应的应用实例，使得广大读者不但能够较好地掌握机器学习的基本理论，而且能够比较系统地掌握其应用技术，为今后的工作和进一步学习打下扎实的理论与应用基础。全书共包含如下 9 章内容：

第 1 章和第 2 章是全书最基础的知识内容，主要为后续机器学习具体方法的介绍提供必备的理论和技术基础。第 1 章主要介绍机器学习的基本知识，包括机器学习基本概念、误差分析、发展历程及需要解决的基本问题；第 2 章主要介绍模型估计与优化的基本方法，包括模型的参数估计、模型优化的基本概念与方法，以及若干模型正则化策略。

第 3 章至第 6 章比较系统地介绍传统机器学习理论与方法。第 3 章主要介绍监督学习模型与算法，包括线性模型、决策树模型、贝叶斯模型和支持向量机模型；第 4 章主要介绍聚类分析、主分量分析、稀疏编码等无监督学习的基本理论和方法；第 5 章主要介绍集成学习方法，包括 Bagging 集成学习和 Boosting 集成学习方法；第 6 章主要介绍强化学习方法，包括基本强化学习和示范强化学习方法。

第 7 章至第 9 章比较系统地介绍神经网络与深度学习方法。第 7 章主要介绍神经网络与深度学习的基本知识，包括神经网络的基本概念、基本模型和常用模型，以及深度学习的基本理论和模型训练方法；第 8 章主要介绍几种常用的深度网络模型与训练范式，包括深度卷积网络、深度循环网络和生成对抗网络；第 9 章主要介绍深度强化学习理论与方法，包括基于价值的学习和基于策略的学习。

限于篇幅，本书未将半监督学习、多示例学习、流形学习、迁移学习、度量学习、元学习、分布式学习等相对比较专门的机器学习前沿研究内容纳入介绍范围，读者可以查阅相关专著、学术论文或技术报告。事实上，如果牢固掌握了本书所介绍的机器学习基本知识内容，那么进一步学习和研究这些前沿知识就不是一件很难的事情。

本书由汪荣贵、杨娟、薛丽霞编著。感谢研究生叶萌、朱正发、汤明空、李文静、俞鹏飞、姚旭晨、陈龙、江迪、郑岩、韩梦雅、邓韬、王静、龚毓秀、李明熹、董博文、麻可可、李懂、刘兵，以及本科生孙旭等同学提供的帮助，感谢合肥工业大学计算机与信息学院、合肥工业大学人工智能学院、机械工业出版社的大力支持。

由于时间仓促，书中难免存在不妥之处，敬请读者不吝指正。

编者

2019 年 6 月

目　录

第 1 章　机器学习概述

近年来，人工智能和机器学习的发展十分迅速，并且已经对社会生活产生越来越多的影响，专门从事这方面理论研究或应用开发的人员也越来越多。人工智能、机器学习这些名词或概念究竟表达怎样的含义？实现机器学习究竟具有哪些基本方式或方法？机器学习在解决实际问题时究竟能够发挥怎样的作用？任何一个从事机器学习理论研究或系统开发的专业人员都应对这些问题具备明确而正确的认识。本章将对这些问题展开讨论，为读者提供机器学习最基本的概念和知识框架。首先介绍机器学习的基本概念，包括人工智能与机器学习的关系、机器学习的基本术语及误差分析；然后以机器学习的发展历程为主线简要介绍连接学习、符号学习和统计学习的基本思想，它们分别代表机器学习的三种不同基本类型；最后分析讨论机器学习的三个基本问题，即特征提取、规则构造和模型评估。

1.1　机器学习基本概念

从外部环境中学习所需知识或技能是人类的一项重要能力，机器学习要解决的问题就是如何使机器也能像人类一样具有这种学习能力。目前，机器学习作为实现人工智能的一项核心技术，已在数据挖掘、计算机视觉、搜索引擎、语音识别、游戏博弈、经济预测与投资分析等众多领域得到了广泛应用。本节主要介绍机器学习的基本概念，包括人工智能与机器学习的关系、机器学习的定义、有关机器学习的若干基本术语以及机器学习的误差分析。

1.1.1　人工智能与机器学习

发明创造某种工具来延伸人类器官功能是实现人类科技进步的一种重要手段。例如，汽车、轮船和飞机等工具的发明延伸了人腿的功能，极大提升了人类的交通能力；摄像机、望远镜和显微镜等工具的发明延伸了人眼的功能，极大提升了人类的视觉能力。为摆脱复杂繁重的科学与工程计算任务，人们发明了计算机代替人脑进行计算。实践证明计算机不仅能够胜任科学与工程计算工作，而且算得比人脑更快、更准确。那么计算机是否可以进一步承担人脑的推理或思维等智能任务呢？受此启发，以麦卡赛、明斯基、罗切斯特和申农等一批具有远见卓识的科学家共同探究使用机器模拟人类思维或人类智能的一系列问题，并在 1956 年夏季首次提出人工智能的概念，标志着人工智能学科的诞生。

人工智能的主要目标是通过计算机来模拟人的某些思维能力或智能行为，如推理、证明、识别、感知、认知、理解、学习等思维能力或活动，让计算机能够像人类一样进行思考。六十多年来，人工智能取得了长足的发展，目前在机器翻译、智能控制、图像理解、语音识别、游戏博弈等领域有着广泛应用。纵观人工智能的发展历程，可依据所用核心技术的不同将其大致分为逻辑推理、知识工程和机器学习这三个基本阶段。

20 世纪 50 年代至 70 年代是人工智能发展的早期阶段，那时人们普遍认为实现人工智能的关键技术在于自动逻辑推理，只要机器被赋予逻辑推理能力就可以实现人工智能。因此，早

期人工智能主要通过谓词逻辑演算来模拟人类智能。这个阶段的人工智能的主流核心技术是符号逻辑计算，在数学定理自动证明等领域获得了一定成功。

然而，人们逐步意识到如果没有一定数量的专业领域知识支撑，则很难实现对复杂实际问题的逻辑推理。因此，以知识工程为核心技术的专家系统在 20 世纪 70 年代至 90 年代逐步成为人工智能的主流。专家系统使用基于专家知识库的知识推理取代纯粹的符号逻辑计算，在故障诊断、游戏博弈等领域取得了巨大成功。

专家系统需要针对具体问题的专业领域特点建立相应的专家知识库，利用这些知识来完成推理和决策。例如，如果让专家系统做疾病诊断，就必须把医生的诊断知识建成一个知识库，然后使用该库中的知识对病情进行推断。然而，把专家知识总结出来并以适当的方式告诉计算机程序有时非常困难，通常需要针对每个具体任务手工建立相应的知识库。例如在图像识别领域，为识别图像中目标是否为猫而建立的知识库并不能用于对目标是否为狗的识别，若要实现对图像中狗的识别，就必须专门建立用于识别狗的知识库。因此，专家知识的人工获取和表示方式严重制约了人工智能的进一步发展。

俗话说，授人以鱼不如授人以渔。既然把专家知识总结出来再灌输给计算机的知识工程方式非常困难甚至在很多场合不可行，那么可以考虑让人工智能系统自己从数据中学习领域知识。从外部环境中学习所需知识或技能是人类的一项重要能力，机器学习要解决的问题就是如何使得机器也能够像人类一样具有这种学习能力。事实上，机器学习的思想可以追溯到 20 世纪 50 年代的感知机数学模型，该模型可以通过使用样本数据调整连接权重的方式保持模型对外部环境变化的自适应性。专家系统的知识工程困境使得机器学习思想和技术逐步得到重视，并在 20 世纪 80 年代初步形成一套相对完备的机器学习理论体系。

自 20 世纪 90 年代中期以来，机器学习得到迅速发展并逐步取代传统专家系统成为人工智能的主流核心技术，使得人工智能逐步进入机器学习时代。特别是近十几年来，数据量爆发式增长、计算机运算能力的巨大提升和机器学习新算法（深度学习）的出现，使得人工智能获得飞跃式迅猛发展。目前，以机器学习为主流核心技术的人工智能在多个领域取得的巨大成功已使其成为社会各界关注的焦点和引领社会未来的战略性技术。

图 1-1 表示一种典型的人工智能系统计算框架，其中机器学习模块通过适当的算法解析数据，从数据中获取知识和模型参数，输出可用于决策或预测的数学模型，为人工智能系统提供核心算法支撑。计算机视觉、语音工程等专业应用模块使用机器学习算法提供的数学模型完成对相关对象的识别、合成、分析、理解、决策等信息处理任务。

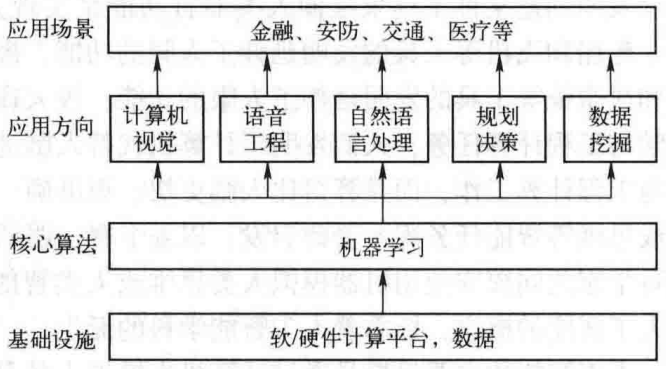

图 1-1　人工智能系统典型框架

由以上分析可知，机器学习为人工智能系统提供基础性的模型和算法支撑，是实现人工智能系统必备的核心技术。下面具体讨论机器学习的基本含义。

在使用计算机解决实际问题时，通常需要对实际问题建立数学模型，将对实际问题的求解转化为对数学模型的求解。此时不可避免地会出现一些模型参数，这些参数的取值情况往往会对模型及其求解结果产生很大的影响，一般需要调整参数以便取得更好的结果。然而，当模型

参数较多或者取值状态比较复杂时，手工调整参数就会变得非常困难和费时。为解决这个问题，可以考虑从实际问题中采集适当的样本数据，通过对这些样本数据进行解析自动计算出所需的模型参数，并随着样本数据变化而自动调整参数取值，使得数学模型和求解算法具有良好的普适性和自适应性。上述做法类似于人类向周围环境学习知识或规则的行为，样本数据相当于周围环境，模型参数相当于学习获得的知识或规则，由此产生机器学习理论和算法的基本思想。

从外部环境中学习所需知识或技能是人类的一项重要能力，获取知识或技能的根本目的在于提高自身的判断、推理、决策或识别等思维水平。因此，从本质上说，机器学习就是通过样本数据等适当的经验信息来改善模型的性能。例如，在使用模型 M_0 识别猫或狗的图片时，可采用适当方式将一些关于猫或狗的带标注图片输入模型 M_0 中，通过改进 M_0 参数或结构产生一个新的模型 M_p，使得模型 M_p 的识别正确率高于 M_0。这就是一个机器学习的过程。此时，经验信息表现为猫或狗的带标注图片，模型的性能即为识别的正确率。

机器学习对初始模型 M_0 的改善不仅体现在模型参数方面，有时还会对模型结构进行改进。因此，通常用改进模型 M_p 泛指机器学习的输出结果。由此得到如下机器学习定义：

机器学习是一种通过先验信息来提升模型能力的方式。具体地说，对于给定的任务和性能度量标准，使用先验信息 E，通过某种计算方式 T 改进初始模型 M_0，获得一个性能更好的改进模型 M_p，即有 $M_p=T(M_0,E)$。

机器学习定义中的任务所界定的范畴非常广泛，在不同应用领域有着不同的具体含义。例如，如果编写一个机器学习程序让机器人能够行走，那么机器人行走就是一个任务。但是机器学习本身不是任务，因为机器学习是获取或提升完成某项任务所需能力的一种途径。

从上述机器学习概念可知，机器学习的目标就是通过计算手段从经验数据等先验信息 E 中产生一个性能改善的新模型 M_p。因此，机器学习的研究内容就是使用计算机从经验数据等先验信息中产生模型的算法，即学习算法。如果说计算机科学是一门关于算法的学问，那么机器学习就是一门关于学习算法的学问。

【例题 1.1】 已知样本数据集为

$$D=\{(x,y)\mid(1.1,1.9),(2.7,2.3),(3.2,3.4),(3.6,2.9),(4.7,3.4),(5.1,4.3)\}$$

D 中数据点在坐标系中的分布如图 1-2 所示。令初始模型 $M_0: y=ax+b$，试根据数据集 D 优化 M_0 并计算 $x=6$ 时的模型输出 $\hat{y}(6)$。

【解】 对于初始模型 M_0，令 M_0 对每个样本 x_i 的预测输出 $\hat{y}_i$ 与其真实值 y_i 之间的误差平方为 e_i，即 $e_i=(\hat{y}_i-y_i)^2$，则模型 M_0 对所有样本的累计误差为

$$Q(a,b)=\sum_{i=1}^{6}e_i=\sum_{i=1}^{6}(\hat{y}_i-y_i)^2=\sum_{i=1}^{6}(\hat{y}_i-ax_i-b)^2$$

由于对模型 M_0 进行优化的依据是 D 中所有样本的真实取值，故当模型对所有样本预测值与真实值之间的累计误差最小时，模型对样本的预测输出最准确，此时的模型就是所求的优化模型，即机器学习定义中具有更好性能的新模型。基于以上分析，可将模型 M_0 的参数求解转化为计算累计误差最小值的优化问题。

将模型 M_0 的参数 a,b 看作累计误差函数 Q 的变量，由于在多元函数极值点处，函数对其所有变量的偏导数均为 0，故分别对参数 a,b 求偏导，并令偏导数为 0，得到

$$\frac{\partial Q}{\partial a}=2\sum_{i=1}^{6}(y_i-ax_i-b)(-x_i)=0$$

$$\frac{\partial Q}{\partial b}=2\sum_{i=1}^{6}(y_i-ax_i-b)(-1)=0$$

联立上述等式并代入 D 中样本数据，解得：$a\approx 0.66, b\approx 0.789$。

由此得到优化模型 M_p：$y=0.66x+0.789$，图 1-3 中的实线表示 M_p 的函数图像，将 $x=6$ 代入 M_p，可求得优化模型的预测值：$\hat{y}(6)=4.749$。□

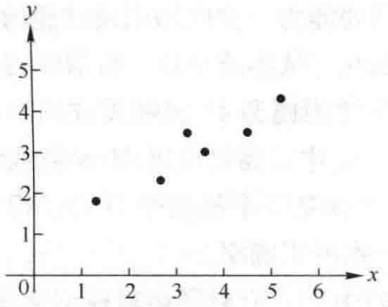

图 1-2　样本分布图

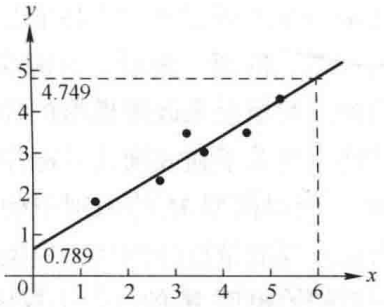

图 1-3　优化模型 M_p

例题 1.1 所示的机器学习实例主要通过对初始模型参数的优化估计求出具有更好性能的新模型。事实上，机器学习有时还可以根据实际需要改变初始模型的结构。例如，对于样本数据集 $D=\{(x_i,y_i)\mid i=1,2,\cdots,n\}$，假设初始模型为如下 k 次多项式

$$N_0: y=\theta_0+\theta_1 x+\theta_2 x^2+\cdots+\theta_k x^k\ (k>2)$$

则可以通过适当方式将 N_0 简化为如下二次多项式模型 $N_p: y=\hat{\theta}_0+\hat{\theta}_1 x+\hat{\theta}_2 x^2$。

事实上，要将初始模型 N_0 变为简化模型 N_p 的形式，只需通过适当方式调整 N_0 的参数，使得参数 $\theta_3,\theta_4,\cdots,\theta_k$ 的取值趋向于 0 即可。可用均方误差最小化的思想实现这个效果，为此构造如下以 $\theta_0,\theta_1,\cdots,\theta_k$ 为自变量的函数

$$Q(\theta_0,\theta_1,\cdots,\theta_k)=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2+C\sum_{j=3}^{k}\theta_j^2 \tag{1-1}$$

其中，$\hat{y}_i=\theta_0+\theta_1 x_i+\theta_2 x_i^2+\cdots+\theta_k x_i^k$。

函数 $Q(\theta_0,\theta_1,\cdots,\theta_k)$ 中的第一项为模型 M_0 对所有样本取值的累计误差，第二项是对参数 $\theta_3,\theta_4,\cdots,\theta_k$ 添加的限制条件。在使函数 $Q(\theta_0,\theta_1,\cdots,\theta_k)$ 最小化的过程中，可将 C 定义为一个非常大的取值，使得参数 $\theta_3,\theta_4,\cdots,\theta_k$ 的取值趋向于 0，以尽量消除 C 值对函数 $Q(\theta_0,\theta_1,\cdots,\theta_k)$ 最小化取值的影响。此时用函数 $Q(\theta_0,\theta_1,\cdots,\theta_k)$ 代替累计误差函数求最小值，相当于在对参数 $\theta_3,\theta_4,\cdots,\theta_k$ 做趋向于 0 的限制条件下解出优化模型的全部参数 $\hat{\theta}_0,\hat{\theta}_1,\hat{\theta}_2$。具体地说，就是由于函数 Q 中第二项权重过大，为使函数 Q 整体取值最小，该项所涉及参数 $\theta_3,\theta_4,\cdots,\theta_k$ 均会趋向于 0，由此即可获得结构调整后的优化模型 $y=\hat{\theta}_0+\hat{\theta}_1 x+\hat{\theta}_2 x^2$。

1.1.2　机器学习基本术语

如前所述，机器学习主要通过样本提供的信息来提升模型性能以完成给定的学习任务，即从样本中学习。对于任意一个给定的样本对象 ξ，一般需要对其提取若干属性形成对该样本的数据描述或表征，并将这些属性值作为机器学习模型的输入。令

$$x_1=\psi_1(\xi), x_2=\psi_2(\xi), \cdots, x_m=\psi_m(\xi)$$

为样本 ξ 的 m 个属性提取函数，则可通过这些函数将样本 ξ 映射成一个 m 元**表征向量** $\boldsymbol{X}$，即

$$\boldsymbol{X}=X(\xi)=(x_1,x_2,\cdots,x_m)^{\mathrm{T}}$$

其中，x_i为样本ξ的第i个属性值，$i=1,2,\cdots,m$。

显然，表征向量$\boldsymbol{X}$是对样本对象ξ的一个数据抽象，从数学的角度看，两者并没有本质上的差异。因此，为方便表达，在不产生混淆的情况下，通常将表征向量为$\boldsymbol{X}$的样本ξ简称为样本X，即不加区分地使用表征向量$\boldsymbol{X}$和表征向量为$\boldsymbol{X}$的样本ξ这两个没有本质差异的概念。

机器学习的任务是指所要解决的问题，主要包括回归、分类和聚类等。回归任务是通过若干带有标注的样本数据构造出一个预测模型$R(X)$，使得$R(X)$的预测输出尽可能符合真实值，并称$R(X)$为**回归模型**。通常将用于构造模型的样本称为**训练样本**，用于测试模型效果的样本称为**测试样本**。一般使用两组不同的样本集合分别作为训练样本集和测试样本集。

设$\xi_1,\xi_2,\cdots,\xi_n$是任意给定的n个训练样本，$\boldsymbol{X}_k=(x_{1k},x_{2k},\cdots,x_{mk})^{\mathrm{T}}$和$y_k$（$k=1,2,\cdots,n$）分别表示$\xi_k$的表征向量和标注值，则由这$n$个样本构成的训练样本集$D$可以表示为

$$D=\{\langle \boldsymbol{X}_k,y_k\rangle \mid k=1,2,\cdots,n\}$$

回归模型$R(X)$的初始模型是一个带有参数的计算模型，机器学习的模型训练算法使用训练样本集D中的数据信息计算出$R(X)$的全部参数，得到具体的回归模型。有了回归模型的具体参数，就可以使用该模型完成回归任务。例如，例题 1.1 解决的就是一个机器学习回归任务，通过训练样本数据计算所得的模型$y=0.66x+0.789$就是一个具体的回归模型。

日常生活和工作中会经常遇到一些分类问题，例如有时需要将产品按质量分为优等品、合格品和次品，将公司客户分为贵宾客户和普通客户等。可以使用机器学习方式实现这种分类任务，即根据带标注训练样本构造相应的分类模型，然后根据分类模型实现对目标的自动分类。显然，如果回归模型的预测输出是离散值，则机器学习的回归任务就转化为分类任务。也就是说，分类其实是预测输入样本所在类别的一类特殊回归任务，特殊性在于要求预测结果为离散的类别值而不是连续值。

用于分类任务的机器学习模型称为**分类模型**或**分类器**，分类任务的目标是通过训练样本构造合适的分类器$C(X)$，完成对目标的分类。分类类别只有两类的分类任务称为**二值分类或二分类**，这两个类别分别称为**正类**和**负类**，通常用+1 和−1 分别指代。分类类别多于两类的分类任务通常称为**多值分类**。

对于一个具体的回归或分类任务，所有可能的模型输入数据组成的集合称为**输入空间**，所有可能的模型输出数据构成的集合称为**输出空间**。显然，回归或分类机器学习任务的本质就是寻找一个从输入空间到输出空间的映射，并将该映射作为预测模型。从输入空间到输出空间的所有可能映射组成的集合称为**假设空间**。

回归或分类模型的训练计算可以看成是一个在假设空间中搜索所需模型的过程，模型训练算法在假设空间中搜索合适的映射，使得该映射的预测效果与训练样本所含的先验信息相一致。事实上，满足条件的映射通常不止一个，此时需要对多个满足条件的映射做出选择。在没有足够依据进行唯一性选择的情况下，有时需要做出具有主观倾向性的选择，即更愿意选择某个映射作为预测模型。这种选择的主观倾向性称为机器学习算法的**模型偏好**。例如，当多个映射与训练样本所包含的先验信息一致时，可选最简单的映射作为预测模型，此时模型偏好为最简单的映射。这种在同等条件下选择简单事物的倾向性原则称为**奥卡姆剃刀原则**。

自然界和社会生活中经常会出现物以类聚、人以群分的现象，例如，羊、狼等动物总是以群居的方式聚集在一起，志趣相同的人们通常会组成特定的兴趣群体。机器学习的**聚类任务**就是对样本数据实现物以类聚的效果。显然，聚类的类别由不同样本之间的某种相似性确定，因而聚类类别所表达的含义通常是不确定的，聚类样本也不带特定的标注来表示样本所属的类

别。这是聚类与回归或分类任务之间的本质区别。通常将不带标注的样本称为**示例**，而将带标注的样本称为**样例**。

在聚类任务中，所有输入示例的集合称为**示例集**，被划分为同一类别的示例所构成的集合称为一个**簇**。图 1-4 表示一个具有两个簇的示例集。

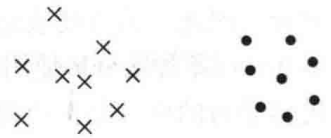

图 1-4　包含两个簇的示例集

对于任意给定的一个 n 元示例集 $S=\{x_1,x_2,\cdots,x_n\}$，假设 Δ 是一个对 S 的划分，则有

$$\Delta=\{S_1,S_2,\cdots,S_t\} \tag{1-2}$$

其中，$S_1,S_2,\cdots,S_t$ 均是由 S 中示例构成的簇，且满足

$$S=S_1\cup S_2\cup\cdots\cup S_t$$

聚类的目标是寻找一个对 S 的适当划分 Δ，使得划分的各簇内部的示例之间的相似性尽可能地小。通常采用欧式距离或余弦距离等作为样本之间相似性的度量标准。图 1-5 表示一个依据相似性度量标准被划分到与之相似的簇中的示例。

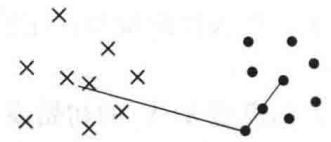

图 1-5　聚类示意图

聚类任务使用的先验信息与回归或分类任务有着很大差别。聚类任务的先验信息为示例，即不带标注的样本，而回归和分类任务的先验信息均为带标注的样本。事实上，除了带标注样本和不带标注样本之外，先验信息有时还以某种反馈信息的形式存在。可根据先验信息的不同形式，将机器学习分为监督学习、无监督学习和强化学习三种基本方式。

监督学习是指利用一组带标注样本来调整模型参数，以提升模型性能的学习方式。监督学习的基本思想是通过标注值告诉模型在给定输入的情况下应该输出什么值，由此获得尽可能接近真实映射方式的优化模型。监督学习不像传统计算机问题求解那样需要根据实际问题的具体情况设计一个固定流程进行计算，而是由计算机根据带标记的样本集自动获得一个问题的求解模型并由此实现对问题的求解。图 1-6 表示监督学习的基本流程。

无监督学习通过比较样本之间的某种联系实现对样本的数据分析。相比于监督学习，无监督学习的最大特点是学习算法的输入是无标记样本。例如，现有一些图片，其中每张图片内容是两类不知名花卉之一，通过观察花卉特点将同类的花卉图片放到一起，这便是无监督学习。在实际问题中遇到样本缺失标记或者人工标注成本过高的情况，可以使用无监督学习方式实现对这些数据自动分析，将所得到的分析结果作为参考信息。

强化学习是根据反馈信息来调整机器行为以实现自动决策的一种机器学习方式。一个强化学习系统主要由智能体和环境两个部分组成。智能体是行为的实施者，由基于环境信息的评价函数对智能体的行为做出评价，若智能体的行为正确，则由相应的回报函数给予智能体正向反馈信息以示奖励，反之则给予智能体负向反馈信息以示惩罚。强化学习的基本流程如图 1-7 所示，智能体根据环境的当前状态选择下一个动作，环境对这个动作做出评价并反馈给智能体，同时更新环境状态，不断重复这一过程直至达到某种设定，选取累计奖励值最大的一组动作作为所求的最终策略。

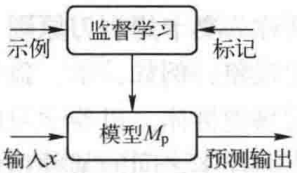

图 1-6　监督学习流程图

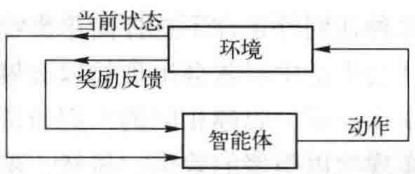

图 1-7　强化学习流程图

1.1.3 机器学习误差分析

机器学习模型是对实际求解问题的一种数学抽象，模型的输出结果与其对应的真实值之间往往会存在一定的差异，这种差异称为该模型的**输出误差**，简称为**误差**。机器学习的一个重要手段就是以模型输出误差为基本依据不断优化或校正模型，使得模型的输出误差尽可能变小。因此，对机器学习模型进行误差分析，从误差分析角度分析并寻找影响机器学习模型性能的关键因素，是机器学习的重要研究内容。

为便于误差分析，通常需要构造某种函数用于度量模型对单个样本的输出误差，这样的函数称为**损失函数**。具体地说，对于给定的机器学习模型 f，假设该模型对应于输入样本 X 的输出为 $\hat{y}=f(X)$，与 X 对应的实际真实值为 y，则可用以 y 和 $f(X)$ 为自变量的某个函数 $L(y,f(X))$ 作为损失函数来度量模型 f 在输入样本 X 下的输出误差。

损失函数的具体形式有很多种，可根据实际问题需要来构造或选用适当的损失函数进行误差分析。例如，$L(y,f(X))=[y-f(X)]^2$ 和 $L(y,f(X))=|y-f(X)|$ 是两种经常用于度量回归模型输出误差的损失函数，分别称为**平方损失函数**和**绝对值损失函数**。

在机器学习中，面向单个样本的损失函数所度量的只是模型在某个特定样本下的输出误差，不能很好地反映模型在某个样本集上对所有样本的整体计算准确度。因此，需要进一步定义面向某个特定样本集的综合误差，通常称为该样本集上的**整体误差**。

对于任意给定的一个 n 元样本集 $S=\{(X_1,y_1),(X_2,y_2),\cdots,(X_n,y_n)\}$，模型 f 在 S 上的整体误差 $R_S(f)$ 的定义为

$$R_S(f)=E[L(y,f(X))]=\frac{1}{n}\sum_{i=1}^{n}L(y_i,f(X_i)) \tag{1-3}$$

即将 $R_S(f)$ 定义为 S 中所有单个样本所对应损失函数值的平均值。

对于某个给定的机器学习任务，假设与该任务相关的所有样本构成的样本集合为 D，则机器学习模型在样本集合 D 上的整体误差称为该模型关于该学习任务的**泛化误差**。具体地说，令样本集合 D 中所有样本的概率分布为 $P(D)$，模型 f 对输入样本 X 的输出为 $\hat{y}=f(X)$，X 所对应的实际真实值为 y，则可将模型 f 的泛化误差定义为

$$R_{\exp}(f)=E_{P(D)}[L(y,f(X))] \tag{1-4}$$

泛化误差表示机器学习模型在整个样本集合 D 上的平均误差，是刻画机器学习模型普适性的重要指标，作为模型求解和模型评估的基本依据，它在机器学习的过程中发挥着极为重要的作用。然而，精确计算模型的泛化误差需要知道整个样本集合 D 中所有样本的真实取值和概率分布，这通常是不可行的。因此，一般无法计算泛化误差的精确值，需要采用某些便于计算的度量指标作为泛化误差的近似代替值。

机器学习模型训练的目标是尽可能获得普适性或泛化性最好的模型，理论上要求模型的泛化误差达到最小。然而，通常无法直接计算模型的泛化误差，更难以直接对泛化误差进行优化分析。由于训练样本通常采样自整个样本集合 D，训练样本集通常与 D 有着比较相似的样本概率分布，故一般采用训练误差近似替代泛化误差来对模型进行训练。

所谓**训练误差**，是指模型在训练样本集上的整体误差，也称为**经验风险**。具体地说，对于任意给定的 n 元训练样本集合 $G=\{(X_1,y_1),(X_2,y_2),\cdots,(X_n,y_n)\}$，假设模型 f 对输入样本 X 的预测输出为 $\hat{y}=f(X)$，则该模型关于训练样本集 G 的训练误差定义为

$$R_{\text{emp}}(f)=\frac{1}{n}\sum_{k=1}^{n}L(y_k,f(X_k)) \tag{1-5}$$

其中，X_k表示训练集中的第 k 个样本；$f(X_k)$表示模型对输入样本 X_k的输出$\hat{y}_k$；y_k为机器学习任务中与输入 X_k对应的实际真实取值。

因此，机器学习中的模型训练或优化通常使用最小化训练误差的方法来完成。该方法称为**经验风险最小化**方法，由此得到的优化模型为

$$\hat{f}=\arg\min_{f\in F} R_{\text{emp}}(f) \tag{1-6}$$

其中，F 为假设空间。

对于已训练出的模型，通常使用测试误差近似替代泛化误差的方法来对该模型进行测试。所谓**测试误差**，是指模型在测试样本集上的整体误差。具体地说，对于任意给定的 v 元测试样本集合 $T=\{(X_1^t,y_1^t),(X_2^t,y_2^t),\cdots,(X_v^t,y_v^t)\}$，该模型关于 T 的测试误差的定义为

$$R_{\text{test}}=\frac{1}{v}\sum_{k=1}^{v}L(y_k^t,f(X_k^t)) \tag{1-7}$$

其中，X_k^t表示测试集中的第 k 个样本；$f(X_k^t)$表示模型对输入 X_k^t的输出$\hat{y}_k^t$；y_k^t为机器学习任务中与输入 X_k^t对应的实际真实值。

对于训练样本集合中的每个样本，该样本都会存在一些普适于整个样本集 D 的共性特征和一些仅仅适合于特定训练样本集的个性特征。在机器学习中，模型训练的最理想效果就是充分提取训练样本的共性特征而尽量避免提取其个性特征，使得训练出来的模型具有尽可能广泛的普适性，即具有尽可能好的泛化性能。

然而，模型的训练通常以最小化训练误差为标准，此时对于固定数量的训练样本，随着训练的不断进行，训练误差会不断降低，甚至趋向于零。如果模型训练误差过小，就会使训练出来的模型基本上完全适应于训练样本的特点。此时，训练模型不仅拟合了训练样本的共性特征而且也拟合了训练样本的个性特征，反而降低了训练模型的泛化性能，使得泛化误差不断增大。这种同时拟合训练样本的共性特征和个性特征的现象，在机器学习领域通常称为模型训练的**过拟合**现象。

避免过拟合现象的一个有效措施是尽可能扩大训练样本的数量，尽可能降低样本在训练样本集与整个样本集上概率分布的差异，以充分增强训练样本的共性特征，弱化训练样本的个性特征。近年来计算机运算能力的巨大提升以及在各行各业中不断涌现的大数据，使得通过扩大训练样本数量以避免过拟合现象的措施变得可行，这正是机器学习在如今互联网和大数据时代得到迅猛发展的重要原因。

由以上分析可知，在机器学习的模型训练中，随着训练过程的进行，训练误差会一直不断降低，但泛化误差则会先减小，然后因产生过度拟合现象而导致不断增大，具体如图 1-8 所示。在训练的初始阶段，由于模型尚未充分拟合训练样本的共性特征，故此时模型的泛化误差较大。这种由于未能充分拟合训练样本共性特征造成模型泛化误差较大而导致模型泛化能力较弱的现象称为模型训练的**欠拟合**现象。随着训练过程的不断进行，训练误差和泛

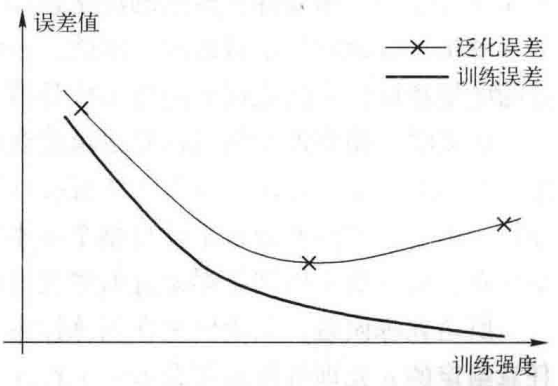

图 1-8　训练误差和泛化误差关系图

化误差均不断减少，欠拟合现象通常会逐渐消失。

对于给定的训练样本集合，如果对模型训练强度不做适当控制，就会在模型训练的后期将训练样本的个性特征引入模型当中，从而引起泛化误差的增大，产生过拟合现象。因此，泛化误差由下降变为上升的转折点处对应的训练模型具有最好的泛化性能。也就是说，对于给定的训练样本集合，可以在适当训练强度下获得具有最好泛化性能的训练模型。

现在进一步分析讨论不同训练样本集合的差异对模型训练结果的影响，考察训练模型对训练样本集合变化的稳定性。

对于任意给定的一个初始模型f，假设$D_1,D_2,\cdots,D_s$是s个不同的训练样本集合，其中每个训练样本均采样自整个样本集合D，通过训练样本集合D_i训练初始模型f所得到的优化模型记为f_i，$i\in(1,2,\cdots,s)$，$\hat{y}_i=f_i(X)$表示第i个模型对于输入样本X的模型输出，X所对应的实际真实值为y，则这s个优化模型对于输入样本X的期望输出为

$$E[F(X)]=\frac{1}{s}\sum_{i=1}^{s}f_i(X) \tag{1-8}$$

其中，$F(X)=(f_1(X),f_2(X),\cdots,f_s(X))^{\mathrm{T}}$，可将其看成一个关于$f(X)$的离散随机变量。

此时，模型$f(X)$对于测试样本集合变化的稳定性可用相应的方差指标进行度量。模型$f(X)$在训练样本集$D_1,D_2,\cdots,D_s$下所得优化模型$f_1(X),f_2(X),\cdots,f_s(X)$输出的方差为

$$\mathrm{Var}[F(X)]=E\{[F(X)-E[F(X)]]^2\}=\frac{1}{s}\sum_{i=1}^{s}[f_i(X)-E[F(X)]]^2$$

对于任意一个给定的初始模型f，如果该模型变化的自由度较大，例如模型参数的数目较多或者参数的取值范围较大，则能够更好地适应训练样本数据的变化，能对多种不同的训练样本集合获得较好的拟合效果；反之，如果该模型参数的变化自由度较小，则模型适应训练数据变化的能力就比较差，可以有效拟合的训练数据范围也就比较有限。机器学习模型这种适应训练数据变化的能力，称为模型的**学习能力**或**模型的容量**。

显然，模型的容量主要反映该模型对数据的拟合能力。模型的容量越大其对数据的拟合能力就越强，越能够适应训练样本数据的变化。可以使用模型输出在不同训练样本集合下的综合偏差对其进行度量，这种综合偏差称为**模型输出的偏差**，简称为**偏差**。

对于模型$f(X)$在训练样本集$D_1,D_2,\cdots,D_s$下的优化模型$F(X)=(f_1(X),f_2(X),\cdots,f_s(X))^{\mathrm{T}}$，$F(X)$作为一个离散随机变量与$X$所对应实际真实值$y$之间的偏差$\mathrm{Bias}[F(X)]$为

$$\mathrm{Bias}[F(X)]=E[F(X)]-y \tag{1-9}$$

对基于平方损失函数的泛化误差$R_{\exp}(f)=E[L(y,F(X))]=E\{[F(X)-y]^2\}$，对其进行偏差-方差分解，可得

$$\begin{aligned}E\{[F(X)-y]^2\}&=E\{[F(X)-E[F(X)]+E[F(X)]-y]^2\}=E\{[F(X)-E[F(X)]]^2\}+\\&\quad E\{[E[F(X)]-y]^2\}+2E\{F(X)-E[F(X)]\}\{E[F(X)]-y\}\end{aligned}$$

由于$E\{F(X)-E[F(X)]\}=E[F(X)]-E[F(X)]=0$，故有：

$$\begin{aligned}E\{[F(X)-y]^2\}&=E\{[F(X)-E[F(X)]]^2\}+E\{[E[F(X)]-y]^2\}\\&=E\{[F(X)-E[F(X)]]^2\}+\{E[F(X)]-y\}^2\\&=\mathrm{Var}[F(X)]+\{\mathrm{Bias}[F(X)]\}^2\end{aligned}$$

即有

$$E\{[F(X)-y]^2\}=\mathrm{Var}[F(X)]+\{\mathrm{Bias}[F(X)]\}^2 \tag{1-10}$$

由式（1-10）可知，模型的泛化误差等于模型输出的方差与模型输出的偏差平方之和。

如前所述，模型输出的偏差反映模型容量的大小或者说模型学习能力的强弱，模型输出的方差则反映模型对训练样本变化的敏感程度。一般而言，对于容量较大的模型，由于其拟合能力较强，因而会使得模型输出的偏差相对较小。然而，大容量模型的变化自由度通常较大，会导致模型参数对样本数据的变化比较敏感，使得模型输出的方差较大。因此，同时减小模型输出的方差和偏差是不可行的，图 1-9 表示泛化误差与偏差及方差之间的关系。

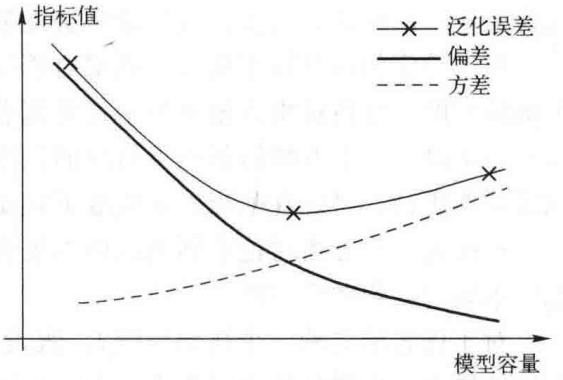

图 1-9　泛化误差与方差及偏差的关系

从图 1-9 中可以看出，随着模型容量的增加，模型输出的偏差随之减小，模型输出的方差却随之增大。因此，模型的泛化误差会出现先减后增的情况。当模型容量较低时，其拟合能力较弱，难以对训练样本的共性特征进行有效拟合，故欠拟合现象会较为严重。当模型容量过高时，模型对数据的变化太过敏感，具有过强的拟合能力，对训练样本的个性特征也进行了拟合，此时过拟合现象较为严重。由此可知，对于具体的机器学习任务而言，模型容量并非越高越好，一个容量适中的机器学习模型通常更能满足任务需求。

1.2　机器学习发展历程

机器学习作为人工智能的一个重要研究领域随着人工智能的产生而产生，并且随着人工智能理论的发展而发展。目前，机器学习理论大致分为连接学习、符号学习和统计学习这三种基本类型。符号学习和连接学习分别源自人工智能的符号主义和连接主义，统计学习则源自符号学习中的归纳学习。从历史上看，连接学习是机器学习最初采用的策略，感知机和神经网络是机器学习初创时期的代表性成果。20 世纪 80 年代，随着人工智能符号主义的发展，符号学习逐步成为机器学习的主流技术。20 世纪 90 年代以来，统计学习方法逐步走向成熟，并以其巨大理论创新和良好应用效果逐步取代符号学习成为机器学习的研究热点。近年来，得益于计算机运算能力的巨大提升和数据量的快速增长，以深度学习为代表的连接学习再次兴起，涌现出一大批优秀的理论和应用成果。统计学习和深度学习的巨大成功使得人工智能全面进入机器学习时代，并成为引领社会未来的战略性技术。

1.2.1　感知机与连接学习

人工智能最早期的探索是从模仿人类或动物大脑的生物结构开始的。人类或动物大脑神经系统最基本的组成结构是神经元，相互连接的多个神经元通过相互传送某些化学物质以改变电位的方式来实现信息传递与交互。麦卡洛克和皮茨在 1943 年发表论文《神经活动中内在思想的逻辑演算》（A Logical Calculus of the Ideas Immanent in Nervous Activity），首次提出模拟生物神经元的数学模型，名为**M-P 模型**。图 1-10 表示该模型的基本结构，其中 $\{x_1,\cdots,x_m\}$ 为 m 个模型输入变量，$x_i \in (0,1)$，θ 为阈值，模型输出 y 有两种可能的取值状态：当 $\sum_{i=1}^{m} x_i > \theta$ 时，$y=1$；否则，$y=0$。

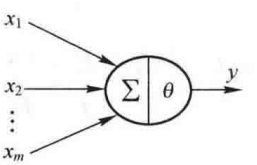

图 1-10　M-P 模型示意图

M-P 模型是对单个神经元的简单模拟，模型的输出值仅为 0 或 1，没有区分 m 个输入在