大数据创新人才培养系列

机器学习与大数据技术

MACHINE LEARNING AND BIG DATA

◎ 牟少敏 著

全面论述经典的**机器学习**理论和算法、**大数据技术**基本原理和方法

兼顾**深度学习**和**人工智能**技术，与**生产实际**的应用场景相结合

详细阐述**大数据技术**与**图形图像处理**技术应用

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

大数据创新人才培养系列

机器学习与大数据技术

MACHINE LEARNING
AND BIG DATA

◎ 牟少敏 著

人民邮电出版社
北京

图书在版编目（CIP）数据

机器学习与大数据技术 / 牟少敏著. -- 北京 : 人民邮电出版社, 2018.6（2018.9重印）
（大数据创新人才培养系列）
ISBN 978-7-115-48771-1

Ⅰ. ①机… Ⅱ. ①牟… Ⅲ. ①机器学习②数据处理
Ⅳ. ①TP181②TP274

中国版本图书馆CIP数据核字(2018)第187605号

内 容 提 要

机器学习、大数据技术是计算机科学与技术的重要研究内容。本书比较全面地论述了机器学习与大数据技术的基本概念、基础原理和基本方法，力求通俗易懂，深入浅出。本书的主要内容包括聚类、遗传算法、粒子群算法、人工神经网络和支持向量机等常见的机器学习算法，重点讲解了深度学习常见的模型、大数据相关内容和大数据技术的具体应用、常见的图像处理技术、Python 语言的编程基础，以及基于 Python 的科学计算和机器学习算法，并配有大量的源代码。书中介绍了作者近年来取得的部分相关研究成果，涉及机器学习、大数据技术等多个领域。

本书适合计算机科学与技术、数据科学与技术的研究生和本科生使用，也可供从事农业大数据等领域的相关人员参考。

◆ 著　　牟少敏
责任编辑　张　斌
责任印制　沈　蓉　彭志环

◆ 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
北京七彩京通数码快印有限公司印刷

◆ 开本：787×1092　1/16
印张：13.25　　2018 年 6 月第 1 版
字数：353 千字　　2018 年 9 月北京第 2 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前言 FOREWORD

机器学习是近 20 多年来兴起的涉及计算机科学与技术、概率论与数理统计和认知科学等多领域交叉的学科，主要研究机器模仿人类的学习过程，以进行知识和技能的获取。作为人工智能领域中一个重要的组成部分，机器学习广泛运用于数据挖掘、计算机视觉、自然语言处理，以及机器人研发等领域。

本书是作者在多年讲授“机器学习”和“大数据技术”课程，以及长期从事机器学习和农业大数据研究工作的基础上编写的。全书共分 9 章，第 1 章简要介绍机器学习、大数据、人工智能和图像处理技术的基础知识，第 2 章和第 3 章主要介绍机器学习和深度学习的理论与方法，第 4 章和第 5 章主要介绍大数据和农业智能的相关知识，第 6 章主要介绍图像处理与分析技术，第 7 章是作者近年来取得的与机器学习、大数据和图像处理技术相关的部分科研成果，第 8 章和第 9 章主要介绍机器学习和大数据的编程基础。

本书的编写得到了王秀美、林中琦、曹旨昊、苏婷婷、孙肖肖、郭嘉和张烁的大力支持和帮助，在此表示感谢。

本书吸收当前微课版教材优点，在书中放置了二维码，读者可以通过扫描二维码获取部分编程源码，以方便使用。

由于作者水平有限，写作时间仓促，书中难免存在错误，敬请读者批评指正。

牟少敏

2018 年 3 月于山东农业大学

目 录 CONTENTS

01 第1章 绪论

目前，云计算、物联网、大数据、机器学习、人工智能、芯片技术和移动网络等新一代信息技术不断涌现，掀起了新一轮技术革命和产业革命的浪潮，新一代信息技术受到了政府、学术界、媒体和企业的广泛关注，同时也带来了巨大的市场机遇，具有广阔的应用前景。

人工智能不是一个新名词，在 1956 年达特茅斯会议上计算机专家约翰·麦卡锡首先提出了“人工智能”的概念。1980 年美国卡耐基·梅隆大学设计并实现了具有知识库和推理功能的专家系统；1997 年 IBM 公司的“深蓝”战胜了国际象棋世界冠军卡斯帕罗夫；2016 年谷歌公司的“阿尔法狗”（AlphaGO）战胜了韩国棋手李世石和我国的围棋天才柯洁。这些里程碑式的标志使得人们对人工智能未来的发展充满了渴望和期待。

人工智能至今尚没有一个统一的定义。专家和学者们从不同的角度出发，给出了各自的定义：畅销书《人工智能》的作者伊莱恩·里奇（Elaine Rich）认为人工智能是研究如何利用计算机模拟人脑从事推理、规划、设计和学习等思维活动，协助人类解决复杂的工程问题；麻省理工学院教授温斯顿（Winston）认为人工智能是那些使知觉、推理和行为成为可能的计算的研究；加州大学伯克利分校教授斯图尔特·罗素（Stuart Russell）则把人工智能定义为：像人一样思考的系统，像人一样行动的系统。

机器学习的发展可以追溯到 1950 年，其发展过程大体经历了 3 个重要时期，即推理期、知识期和学习期。1970 年前称为推理期，主要标志是让机器具有简单的逻辑推理能力；1970 年后称为知识期，主要标志是 1965 年斯坦福大学教授费根鲍姆（E. A. Feigenbaum）等人研制了世界上首个专家系统。20 世纪 80 年代至今称为学习期，主要标志是让机器从样本中学习。1983 年，美国加州理工学院霍普菲尔德（J. J. Hopfield）教授提出了著名的 Hopfield 反馈神经网络；1986 年，斯坦福大学教授鲁姆哈特（D. E. Rumelhart）等人提出了 BP 神经网络；1995 年，美国工程院院士瓦普尼克（Vapnik）教授提出了基于统计学习理论的支持向量机，产生了以支持向量机为代表的核机器学习方法，如核聚类和核主分量分析等。深度学习是机器学习和人工智能的一个重要组成部分，来源于人工神经网络研究和发展，最早由加拿

大多伦多大学的辛顿（Geoffrey E. Hinton）教授于 2006 年提出，辛顿通过 pre-training 较好地解决了多层网络难以训练的问题。深度学习近年来在图像识别和语音识别上取得了突破性的进展，深度学习的成功主要归功于 3 大因素，即大数据、大模型和大算力。深度学习的优越性能将人工智能推向了新的高潮。

目前，大数据背景下机器学习的研究又成为人们研究和关注的热点。传统机器学习的分类算法很难直接应用到大数据环境下，不同的分类算法面临着不同的挑战。大数据环境下的并行分类算法的研究成为一个重要的研究方向。目前，针对并行机器学习的研究方法主要有：基于多核与众核的并行机器学习、基于集群或云的并行机器学习、基于超算的机器学习和基于混合体系结构的并行机器学习。

“数据仓库之父”比尔·恩门（Bill Inmon）早在 20 世纪 90 年代就经常提起大数据。自 2008 年 9 月国际著名的期刊《自然》（*Nature*）出版了大数据专刊以来，大数据的处理、分析和利用已经成为各行各业和科研人员关注的焦点。美国把大数据视为“未来的新石油”，我国将大数据上升为国家战略，大数据产业正在逐步地进入成熟期。目前，大数据几乎是家喻户晓，成为当今非常热门的话题。从电视上经常可以看到有关大数据的新闻，比如：中央电视台将大数据分析技术应用于新闻报道中，推出了两会大数据、春运大数据等相关栏目。

当今世界是一个“数据为王”的时代，数据的重要性已经引起各个国家政府、企业和科研人员的高度重视，大数据背后的价值也在发挥着重要的作用。IBM 智力竞赛机器人沃森（Watson）收集了 2 亿页知识文本数据，并采用并行处理集群，利用大数据处理技术进行数据分析，可在 1 秒内完成对大量非结构化信息的检索。目前，软硬件技术与行业需求正在极大地推动大数据的发展。

大数据首先要有数据，因此大数据的采集技术是非常重要的。物联网技术、电商平台等各种采集技术和方法为大数据的采集提供了有力的支撑。另外，数据采集的完整性、准确性和稳定性，决定了数据采集的质量及数据是否能真实可靠地发挥作用。例如：传统农业田间数据的采集有时必须采用人工手段来进行，由于环境的复杂性等原因，往往存在数据采集不完整和不准确等问题。利用物联网技术进行农业数据的采集具有实时性、多样性和可靠性，又如：农业小气候站采集的气象数据具有实时性、多样性和可靠性的特点，为农业的辅助决策提供较为准确的依据。

研究大数据不仅仅是各种数据的采集和存储，更重要的是如何利用好大数据，通过分析和挖掘海量数据，发现其内在有价值和有规律的知识，并服务于各个领域。大数据的分析挖掘技术又为机器学习的发展和应用提供了广阔的空间。

目前，深度学习成为机器学习热点的同时，又为人工智能的发展提供了巨大的发展空间，例如：利用深度学习感知、识别周围环境，以及各种对车辆有用的信息，使得无人驾驶汽车成为可能；微软和谷歌利用深度置信网络，将语音识别的错误率降低了 20%～30%。

深度学习在云计算和大数据背景下取得实质性进展，云计算为深度学习提供了平台。云计算平台服务的优点：搭建快速、操作简捷、智能管理、运行稳定、安全可靠和弹性扩展。国内云计算平台有很多，如著名的阿里巴巴公司和百度公司等。

物联网（Internet of Things）的概念是由麻省理工学院自动识别（MIT Auto-ID）中心阿什顿（Ashton）教授 1999 年提出的，其原理是利用各种传感设备，如射频识别装置、红外感应器、全球定位系统、激光扫描器等种种装置与互联网结合起来从而形成的一个巨大网络。《传感器通用术语》

（GB7665—87）对传感器的定义是："能感受规定的被测量并按照一定的规律转换成可用信号的器件或装置，通常由敏感元件和转换元件组成"。通俗地讲，物联网就是物与物相连的互联网。目前，各种传感器广泛地应用到我们的衣食住行等日常生活中，如湿度传感器、气体烟雾传感器、超声波传感器和空气质量传感器等。传感器正在朝着微型化、智能化、多功能化和无线网络化的方向发展。与发达国家相比，我国自主传感器核心技术仍需不断提高，高端传感器芯片以进口为主，市场竞争较为激烈。

当前，新一代信息技术革命已经成为全球关注的重点。同时，新产品、新应用和新模式不断涌现，改变了传统经济发展方式，极大地推动了新兴产业的发展壮大。这也给研究计算机技术的专业人员和企业带来新的机遇和挑战，这就需要加速学科深度交叉和融合，需要学术界和企业界深度交叉和融合，需要充分利用各行各业大数据，学习和研究人工智能、深度学习和大数据等新技术的基本概念、基本思想、基本理论和技术，掌握常用的相关开发工具，需要挖掘大数据背后的价值，发现规律、预测趋势，并辅助决策。

大数据必须和具体的领域、行业相结合，才能真正地为政府和企业决策提供帮助，才能产生巨大的实用价值和应用前景。本书以农业为应用背景，重点研究机器学习、深度学习、图像处理技术，以及大数据技术在农业领域中的应用。

1.1 机器学习

1.1.1 概述

机器学习简单地讲就是让机器模拟人类的学习过程，来获取新的知识或技能，并通过自身的学习完成指定的工作或任务，目标是让机器能像人一样具有学习能力。

机器学习的本质是样本空间的搜索和模型的泛化能力。目前，机器学习研究的主要内容有3类，分别是模式识别（Pattern Recognition）、回归分析（Regression Analysis）和概率密度估计（Probability Density Estimation）。模式识别又称为模式分类，是利用计算机对物理对象进行分类的过程，目的是在错误概率最小的情况下，尽可能地使结果与客观物体相一致。显然，模式识别的方法离不开机器学习。回归分析是研究两个或两个以上的变量和自变量之间的相互依赖关系，是数据分析的重要方法之一。概率密度估计是机器学习挖掘数据规律的重要方法。

机器学习与统计学习、数据挖掘、计算机视觉、大数据和人工智能等学科有着密不可分的联系。人工智能的发展离不开机器学习的支撑，机器学习逐渐成为人工智能研究的核心之一。大数据的核心是利用数据的价值，机器学习是利用数据挖掘价值的关键技术，数据量的增加有利于提升机器学习算法的精度，大数据背景下的机器学习算法也迫切需要大数据处理技术。大数据与机器学习两者是互相促进、相互依存的关系。

1.1.2 评价准则

评价指标是机器学习非常重要的一个环节。机器学习的任务不同，评价指标可能就不同。同一种机器学习算法针对不同的应用，可以采用不同的评价指标，每个指标的侧重点不一样。下面介绍常用的机器学习评价指标。

1. 准确率

样本分类时，被正确分类的样本数与样本总数之比称为准确率（Accuracy）。与准确率对应的是错误率，错误率是错分样本数与总样本数之比。

显然，准确率并没有反映出不同类别错分样本的情况。例如：对于一个二类分类问题，准确率并不能反映出第一类和第二类分别对应的错分样本的个数。但是，在实际应用中，因为不同类别下错分样本的代价或成本不同，往往需要知道不同类别错分样本的情况。例如：在医学影像分类过程中，未患有乳腺癌被错分类为患有乳腺癌，与患有乳腺癌被错分类为未患有乳腺癌的重要性显然是不一样的。另外，数据分布不平衡时，样本占大多数的类主导了准确率的计算等情况，这就需要求出不同类别的准确率。

2. 召回率

召回率（Precision-Recall）指分类正确的正样本个数占所有的正样本个数的比例。它表示的是数据集中的正样本有多少被预测正确。

3. ROC 曲线

ROC（Receiver Operating Characteristic）曲线是分类器的一种性能指标，可以实现不同分类器性能比较。不同的分类器比较时，画出每个分类器的 ROC 曲线，将曲线下方面积作为判断模型好坏的指标。ROC 曲线的纵轴是“真正例率”（True Positive Rate，TPR），横轴是“假正例率”（False Positive Rate，FPR）。ROC 曲线下方面积（The Area Under The ROC Curve，AUC）是指 ROC 曲线与 x 轴、点（1，0）和点（1，1）围绕的面积。ROC 曲线如图 1-1 所示。显然，$0 \leqslant AUC \leqslant 1$。假设阈值以上是阳性，以下是阴性，若随机抽取一个阳性样本和一个阴性样本，分类器正确判断阳性样本的值高于阴性样本的概率。在图 1-1 示例中，有 3 类分类器，AUC 值分为 0.80、0.78 和 0.80，AUC 值越大的分类器正确率越高。

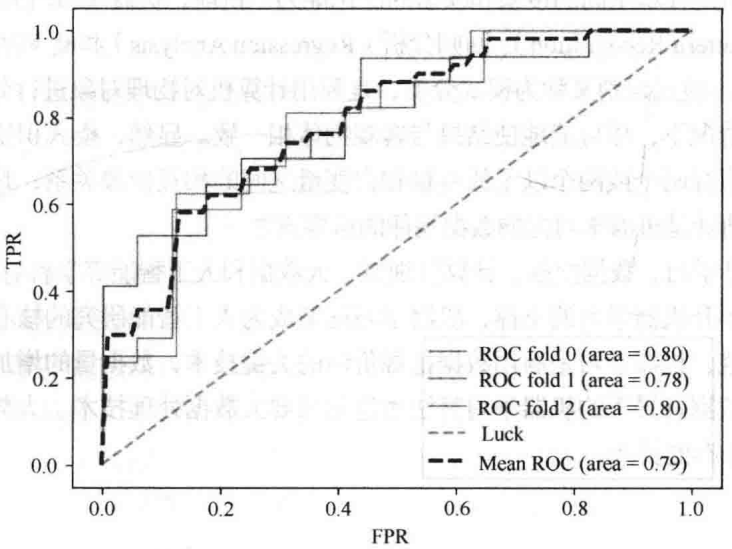

图 1-1　ROC 曲线

4. 交叉验证

交叉验证（Cross-Validation）的基本思想是将数据分成训练集和测试集。在训练集上训练模型，

然后利用测试集模拟实际的数据，对训练模型进行调整或评价，最后选择在验证数据上表现最好的模型。

交叉验证法的优点是可以在一定程度上减小过拟合，还可以从有限的数据中获取尽可能多的有效信息。常用的交叉验证的方法如下。

（1）*K* 折交叉验证

K 折交叉验证的基本思想：将数据随机地划分为 *K* 等份，将其中的 *K*-1 份作为训练集，剩余的 1 份作为测试集，计算 *K* 组测试结果的平均值作为模型精度的估计，并作为当前 *K* 折交叉验证下模型的性能指标。

K 折交叉验证实现了数据的重复利用。一般情况下，*K* 的取值为 10。针对不同的应用场景，可以根据实际情况确定 *K* 值，数据量或样本数较大时，*K* 的取值可以大于 10。数据量或样本数较小时，*K* 的取值可以小于 10。

（2）留一交叉验证

留一交叉验证（Leave One Out Cross Validation）的基本思想：假设有 *N* 个样本，将每一个样本作为测试样本，其他 *N*-1 个样本作为训练样本，得到 *N* 个分类器和 *N* 个测试结果。用这 *N* 个结果的平均值来衡量模型的性能。留一交叉验证是 *K* 折交叉验证的特例。

5. 过拟合与欠拟合问题

机器学习过程中，模型对未知数据的预测能力称为泛化能力（Generalization Ability），是评估算法性能的重要评价指标（Evaluation Metrics）。泛化指的是训练模型对未知样本的适应能力。优秀的机器学习模型其泛化能力强。

过拟合（Over-fitting）是由于训练模型中涉及的参数过多，或参加训练的数据量太小等原因，导致了微小的数据扰动都会产生较大的变化或影响，造成了模型对已知数据预测精度很高，而对未知数据预测精度较低的现象，即测试样本输出和期望的值相差较大，也称为泛化误差较大。

通常情况下，解决过拟合问题的方法有以下两种。

（1）利用正则化来控制模型的复杂度，改善或减少过度拟合的问题。

（2）根据实际问题增加足够的训练数据。

欠拟合（Under-fitting）是模型在训练和预测时，其准确率都较低的现象。产生的原因可能是模型过于简单，没有充分地拟合所有的数据。解决欠拟合问题的方法是优化和改进模型，或采用其他的机器学习算法。

1.1.3 分类

根据机器学习算法的学习方式，机器学习分为以下 3 种。

1. 有监督学习

有监督学习（Supervised Learning）是利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的学习过程，也称为有老师的学习。有监督学习的过程是：首先利用有标号的样本进行训练，构建相应的学习模型。然后，再利用这个模型对未知样本数据进行分类和预测。这个学习过程与人类认识事物的过程非常相似。常用有监督学习的算法有：贝叶斯分类、决策树和支持向量机等。

2. 无监督学习

无监督学习（Unsupervised Learning）是对无标号样本的学习，以发现训练样本集中的结构性知

识的学习过程，也称为无老师的学习。无监督学习事先并不需要知道样本的类别，而是通过某种方法，按照相似度的大小进行分类的过程。它与监督学习的不同之处在于，事先并没有任何训练样本，而是直接对数据进行建模。常用无监督学习的算法有：聚类算法和期望最大化算法。

3. 半监督学习

半监督学习（Semi-Supervised Learning）是有监督学习和无监督学习相结合的学习，是利用有类标号的数据和无类标号的数据进行学习的过程。其特点是利用少量有标号样本和大量无标号样本进行机器学习。在数据采集过程中，采集海量的无标号数据相对容易，而采集海量的有标号样本则相对困难，因为对无标号样本的标记工作可能会耗费大量的人力、物力和财力。例如，利用计算机辅助医学图像分析和判读的过程中，可以从医院获得海量的医学图像作为训练数据，但如果要求把这些海量图像中的病灶都标注出来，则是不现实的。现实世界中通常存在大量的未标注样本，但有标记样本则比较少，因此半监督学习的研究是非常重要的。

此外，根据算法的功能和形式可把机器学习算法分为：决策树学习、增量学习、强化学习、回归学习、关联规则学习、进化学习、神经网络学习、主动学习和集成学习等。

1.1.4 常用工具

1. WEKA

WEKA 是一款常用的、开源的机器学习和数据挖掘工具，主要功能有数据预处理、分类、回归和关联规则等。WEKA 内集成了决策树和贝叶斯分类等众多机器学习算法，是数据分析和挖掘的技术人员常用的工具之一。

2. Python 语言

Python 是一种面向对象的编程语言，由荷兰人吉多·范罗苏姆（Guido van Rossum）发明，最早的公开发行版诞生于 1991 年。Python 提供了大量的基础代码库，极大地方便了用户进行程序编写。Python 语言在数据挖掘和分析、机器学习和数据可视化等方面发挥了巨大的作用。目前，Python 是最热门的人工智能和机器学习的编程语言。

3. Matlab

Matlab 是美国 MathWorks 公司出品的一款商用软件，是科研工作者、工程师和大学生必备的数据分析工具之一，主要用于科学计算，如数值计算、数据分析、数据可视化、数字图像处理和数字信号处理等。

4. R 语言

R 语言是一种为统计计算和图形显示而设计的语言环境，是贝尔实验室开发的 S 语言的一种实现。它提供了有弹性的、互动的环境分析，也提供了若干统计程序包，以及一系列统计和图形显示工具，用户只需根据统计模型，指定相应的数据库及相关的参数，便可灵活机动地进行数据分析等工作。目前，R 语言在数据挖掘和分析、机器学习和数据可视化方面发挥了巨大的作用。

5. 深度学习框架

深度学习的发展离不开高性能的框架与硬件的支持。随着半导体工艺和微电子等技术的飞速发

展，支持深度学习的硬件环境也在飞速发展，出现了以多核 CPU（Central Processing Unit）、高性能图形处理器 GPU（Graphics Processing Unit）、APU（Accelerated Processing Unit）等处理器为代表的高性能并行计算系统，为深度学习分析和挖掘奠定了硬件基础。目前，深度学习大都使用 GPU 在各种框架上进行模型训练，深层神经网络在 GPU 上运算的速度要比 CPU 快一个数量级。

随着深度学习研究和应用的不断深入，各种开源的深度学习框架不断涌现，目前常用的深度学习框架有 Caffe、TensorFlow、Theano、Torch 和 CNTK 等。下面简单介绍几种常用的深度学习框架。

（1）Caffe

Caffe 是一种被广泛使用的开源深度学习框架，由加州大学伯克利分校的贾扬清开发。Caffe 是首个主流的工业级深度学习工具，运行稳定，代码质量高，适用对稳定性要求高的生产环境。目前在计算机视觉领域 Caffe 依然是最流行的工具包，并且有很多扩展。Caffe 最开始设计时的目标只针对图像，没有考虑文本、语音等数据，因此对卷积神经网络的支持非常好，但对时间序列 RNN、LSTM 等的支持不是特别充分。许多研究人员采用 Caffe 做人脸识别、位置检测和目标追踪等，很多深度学习的论文也都是使用 Caffe 来实现其模型的。

（2）TensorFlow

Google 公司开源的 TensorFlow 框架是相对高阶的机器学习库，用户可以方便地用它设计各种神经网络结构，是理想的深度学习开发平台。TensorFlow 使用了向量运算的符号图方法，使指定新网络变得比较容易，但是不支持双向 RNN 和 3D 卷积。TensorFlow 移植性高，一份代码几乎不经过修改就可轻松地部署到有任意数量 CPU 或 GPU 的 PC、服务器或者移动设备上。TensorFlow 框架针对生产环境高度优化，产品级的高质量代码和设计可以保证其在生产环境中稳定运行。

（3）Theano

Theano 由 Lab 团队开发并维护，是一个高性能的符号计算及深度学习库。Theano 因其出现时间早，一度被认为是深度学习研究和应用的重要标准之一。Theano 专门为处理大规模神经网络训练的计算而设计，其核心是一个数学表达式的编译器，可以链接各种可以加速的库，将用户定义的各种计算编译为高效的底层代码。

（4）Torch

Torch 是一个高效的科学计算库，含有大量的机器学习、计算机视觉、信号处理和网络的库算法。Torch 对卷积网络的支持非常好，通过很多非官方的扩展支持大量的 RNN 模型。

（5）CNTK

CNTK 是由微软公司推出的开源深度学习工具包，性能优于 Caffe、Theano、TensoFlow，支持 CPU 和 GPU 两种模式。

各种框架的底层语言和操作语言的比较，详见表 1-1 所示。

表 1-1　　各种深度学习框架的比较

框架名称	底层语言	操作语言
Caffe	C++	Python，C++，Matlab
TensorFlow	C++，Python	Python，C++
Theano	Python，C	Python

续表

框架名称	底层语言	操作语言
Torch	C，Lua	C，Lua，C++
Keras	Python	Python
MXNet	C++，Python	Python，C++
CNTK	C++	C++，Python

1.2 大数据

大数据迅速发展成为当今科技界和企业界甚至世界各国政府关注的热点。《自然》（*Nature*）和《科学》（*Science*）等国际顶尖学术期刊相继出版专刊探讨大数据带来的机遇和挑战。美国把大数据视为“未来的新石油”，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为国家间和企业间新的争夺焦点。“大数据时代”已然来临。

迄今为止并没有公认的关于“大数据”的定义。一般认为大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。从宏观世界角度看，大数据是融合物理世界、信息空间和人类社会三元世界的纽带。从信息产业角度看，作为新一代信息技术重要组成部分的大数据已成为经济增长的新引擎。从社会经济角度看，大数据已成为第二经济的核心和支撑。第二经济是指处理器、传感器和执行器等，以及运行在其上的经济活动。

相较于传统数据，人们将大数据的特征总结成“4V”，即数据量大（Volume）、多样性（Variety）、价值密度低（Value）和高速度（Velocity）。大数据的主要难点并不在于数据量大，因为通过对计算机系统的扩展可以在一定程度上缓解数据量大带来的挑战。大数据真正难点来自数据多样性和高速度。数据类型多样使得系统不仅要处理结构化数据，还要处理文本和视频等非结构化数据。在金融分析、航空航天等行业，数据处理速度要求非常高，时间就是效益。传统的数据处理算法无法满足快速响应的需求，因此迫切需要新型算法的支持。为了应对大数据面临的挑战，以 Google 为代表的互联网企业近几年推出了各种不同类型的大数据处理系统，推进了深度学习、知识计算和可视化等技术在大数据背景下的发展。

1.3 人工智能

人工智能（Artificial Intelligence，AI）定义为：一门融合了计算机科学、统计学、脑神经学和社会科学的综合性学科，其目标是希望计算机拥有像人一样的智力，可以替代人类实现识别、认知和决策等多种能力。

在发展过程中，人工智能主要形成了 3 大学术流派，即符号主义（Symbolicism）、连接主义（Connectionism）和行为主义（Actionism）。

（1）符号主义又称逻辑主义或计算机学派。符号主义最早在 1956 年提出“人工智能”的概念，学派的代表人物有纽厄尔（Newell）和西蒙（Simon）等。符号主义认为，人工智能起源于数学逻辑，人的过程就是符号操作的过程，通过了解和分析人的认知过程，让计算机来模拟实现人所具有的相

应功能。符号主义的发展大概经历了 2 个阶段：推理期（20 世纪 50～70 年代）和知识期（20 世纪 70 年代以后）。在“推理期”，人们基于符号知识表示，通过演绎推理技术取得了很大的成就；在“知识期”，人们基于符号表示，通过获取和利用领域知识来建立专家系统，在人工智能走向工程应用中取得了很大的成功。

（2）连接主义又称仿生学派或生理学派。连接主义认为，人工智能源于仿生学，特别是对人脑模型的研究，人的思维基元是神经元，而不是符号处理过程。20 世纪 60～70 年代，连接主义（尤其是对以感知机（Perceptron）为代表的脑模型的研究）出现过热潮，由于受到当时的理论模型、生物原型和技术条件的限制，脑模型研究在 20 世纪 70 年代后期至 80 年代初期落入低潮。直到霍普菲尔德（Hopfield）教授在 1982 年和 1984 年发表 2 篇重要论文，提出用硬件模拟神经网络以后，连接主义再次焕发生机。1986 年，鲁梅尔哈特（Rumelhart）等人提出多层网络中的反向传播算法（BP）算法。进入 21 世纪后，连接主义卷土重来，提出了“深度学习”的概念。

（3）行为主义又称进化主义或控制论学派。行为主义认为，人工智能源于控制论，早在 20 世纪 40～50 年代，控制论思想就成为时代思潮的重要内容，对早期人工智能工作者有较大的影响。早期的研究工作重点是在研究模拟人在控制过程中的智能行为和作用，例如：对自适应、自寻优、自组织，以及自学习等控制论体系的基础上，进行对“控制论动物”的研制。20 世纪 60～70 年代，基于上述控制论体系的研究取得了一定的进展，为 80 年代出现的智能控制和智能机器人奠定了基础。在 20 世纪末，行为主义以人工智能新学派的面孔出现，麻省理工学院教授布鲁克斯（Brooks）的六足行走机器人是典型代表，该机器人是一个基于感知-动作模式模拟昆虫行为的控制系统，被认为是新一代的“控制论动物”。

20 世纪 80 年代，机器学习成为一个独立的科学领域，各种机器学习技术百花初绽。费根鲍姆等人在著名的《人工智能手册》一书中，把机器学习分为机械学习、示教学习、类比学习和归纳学习。机械学习将外界的输入信息全部存储下来，等到需要时原封不动地取出来；示教学习和类比学习就是“从指令中学习”和“通过观察和发现学习”；归纳学习就是“从样例中学习”。80 年代后研究最多的就是归纳学习，它包括：监督学习、无监督学习、半监督学习、强化学习等。

归纳学习有两大主流：符号主义学习和连接主义学习。前者代表算法有决策树和基于逻辑的学习，后者代表算法有基于神经网络的学习。

20 世纪 90 年代中期，统计学习闪亮登场，并迅速占据主流舞台，代表性技术是支持向量机，以及更一般的“核方法”。目前所说的机器学习方法，一般认为是统计机器学习方法。

人工智能的“智能”之处主要体现在计算智能、感知智能和认知智能 3 个方面。计算智能是机器可以智能化存储和运算的能力，感知智能是使机器具有像人类一样的“听、看、说、认”的能力，认知能力是使机器具有思考和理解的能力。推动人工智能发展的 3 大要素是数据资源、核心算法和计算能力。当前人工智能领域技术主要包括语言识别、机器人、自然语言处理、图像识别和专家系统等。

人工智能、机器学习和深度学习三者之间是包含关系，人工智能的研究最早包含了机器学习，或者说机器学习是其核心组成部分，人工智能与机器学习密不可分。目前，人工智能的热点是深度学习，深度学习是机器学习的一种方法或技术。深度学习在图像识别和语音识别中识别精度的大幅提高，加速了人脸识别、无人驾驶、电影推荐、机器人问答系统和机器翻译等各个领域的应用进程，逐步形成了“人工智能+”的趋势。