## 1.1学习的机器

过去十年，人们对机器学习的兴趣呈爆炸式增长。几乎每天都会在计算机科学课程、行业会议以及《华尔街日报》中遇到“机器学习”这个词。在所有关于机器学习的谈论中，人们常常把机器学习能够做到的事情和他们期望它做到的事情混为一谈。实际上，机器学习是使用算法从原始数据中提取信息，并用某类模型表示信息的一门学问。我们使用模型从那些未建模的数据中推断某些信息。神经网络是机器学习的一种模型，已经存在至少50年了。

神经网络的基本单元是节点，基本上是仿照哺乳动物大脑的神经元构建的。神经元之间的连接也是仿照生物的大脑构建的，也会随着时间（通过“训练”）进化。接下来的两章会深入探讨这些模型是如何工作的。

在20世纪80年代中期和90年代早期，神经网络在架构上取得了很多重要进展。但神经网络需要大量时间和数据才能取得好的结果，这限制了它的应用场景，磨灭了人们的兴趣。21世纪初，计算机的计算能力呈指数级增长，这个行业经历了之前从未发生过的计算技术的“寒武纪大爆发”。深度学习作为这个领域中一个强有力的竞争者，在计算能力呈爆炸性增长的十年中脱颖而出，赢得了许多重要的机器学习竞赛。这股热度在2017年依然没有消退，如今在机器学习的每个角落都能看到深度学习的身影。

本章将深入探讨深度学习的定义。本书结构精心编排，实践者可以利用本书做以下事情：

•复习线性代数和机器学习相关的基础知识；

•复习神经网络的基础知识；

•学习四大主流深度网络架构；

•使用书中的示例来尝试实现实用的深度网络变体。

希望你觉得这些内容实用且易于理解。我们先快速概览机器学习的基础知识，之后的章节还会介绍一些核心概念，以便你更好地理解本书的剩余内容。

* + 1. 机器如何学习

在定义机器如何学习之前，首先需要定义“学习”。日常生活中，当说到“学习”时，指的是“通过学习、经验或者接受教育来获得知识”。结合我们的主题，可以把机器学习看作使用算法从数据样本中获取其结构描述的做法。计算机学到一些关于结构的信息，这代表原始数据中的信息。结构描述是所构建模型的另一种说法，包含着从原始数据中提取的信息。我们可以使用这些结构或模型来预测未知数据。结构描述（或模型）可以呈现为多种形式，其中包括：

•决策树

•线性回归

•神经网络的权重

每种类型的模型使用不同的方式将规则应用于已知数据，从而预测未知数据。决策树以树结构的形式创建一组规则，而线性模型创建一组表示输入数据的参数。

神经网络有一个所谓的参数向量，用于表示网络中节点之间连接的权重。本章稍后会介绍这类模型的细节。

**机器学习与数据挖掘**

数据挖掘已经出现几十年了，它像机器学习的许多术语一样，常被曲解或错用。基于本书内容，“数据挖掘”的实践被定义为“从数据中提取信息”。机器学习的不同之处在于，它指的是数据挖掘中用于从原始数据获取结构描述的算法。下面是对数据挖掘做法的简要概括。

* 为了学习概念：

◦需要原始数据的样本。

* 从数据选取行或实例的样本：

◦这些样本代表数据中特定的模式。

* 机器从这些数据模式中学习概念：

◦机器通过算法进行学习。

在IBM和斯坦福大学工作的人工智能领域的先驱ArthurSamuel将机器学习定义为：

在不直接针对问题编程的情况下，赋予计算机学习能力的一个研究领域。

Samuel开发了一款可以玩西洋跳棋的软件，它能调整策略，因为它能学到输赢的概率与棋盘上某种棋面之间的关系。这种寻找通往胜利或失败的模式，然后识别并强化成功模式的基本做法，支撑着机器学习和人工智能走到今天。

机器能够通过学习达成其自身目标，对此我们已经憧憬了几十年。这也许主要是受到现代人工智能教父Stuart Russell和Peter Norvig的著作《人工智能：一种现代的方法》的影响，书中写道：

一个缓慢、微小的大脑，无论是生物学上的还是电子学上的，是如何感知、理解、预测和操纵比自己更大、更复杂的世界的？

这句话暗示了学习的概念受到了自然界的过程和算法的启发。图1-1以图形化的方式展示了人工智能、机器学习和深度学习之间的关系。

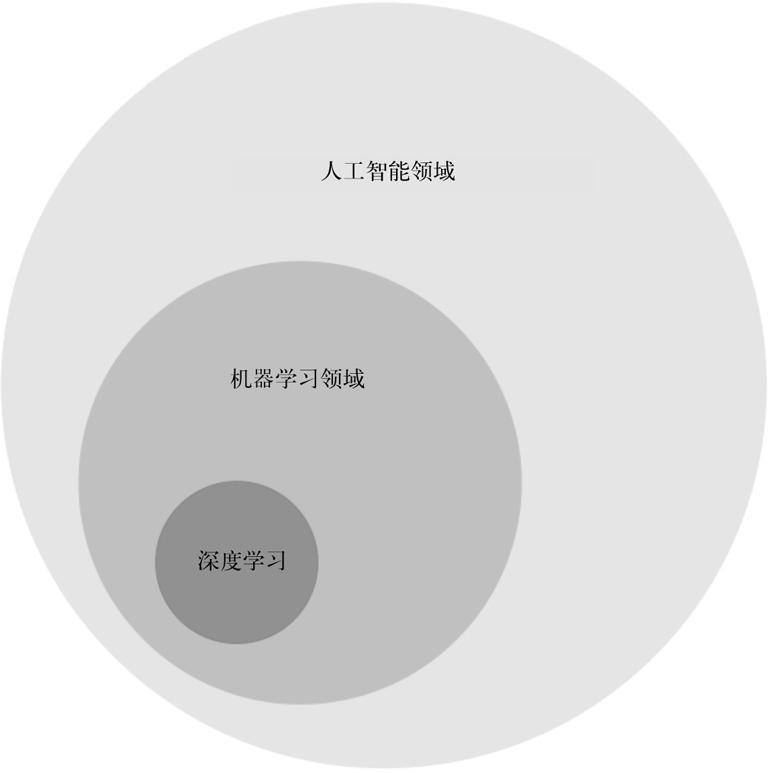


图1-1：人工智能与深度学习之间的关系

人工智能是个广阔的领域，且已经存在了很长时间。深度学习是机器学习领域的一个子集，是人工智能的子领域。下面快速了解一下深度学习的另一个根源：神经网络是如何受到生物学启发的。

* + 1. 生物学的启发

生物学上的神经网络（大脑）大约由860亿个神经元组成，每个神经元与许多其他神经元相连。

人脑中全部的连接

数据研究人员保守估计，人脑中神经元之间有超过500万亿个连接。即使今天最大的人工神经网络离这个数字也相去甚远。

从信息处理的角度看，生物学神经元是一个可兴奋的单元，可以通过电信号和化学信号处理和传输信息。生物大脑中的神经元被视作大脑、中枢神经系统的脊髓和外周神经系统的神经节的主要成分。本章稍后会介绍，人工神经网络的结构要比它简单得多。

生物神经网络和人工神经网络的比较

生物神经网络要比它的人工神经网络版本复杂几个数量级！

人工神经网络有两个主要特性遵循了大脑工作的一般原理。首先，神经网络最基本的单元是人工神经元（或简称为节点）。人工神经元模仿大脑的生物神经元，就像生物神经元一样，它们受到输入的刺激。这些人工神经元传递一些（但不是所有）收到的信息给其他的人工神经元，并通常伴有转换。随着本章内容的推进，将会详细讨论这些神经网络中的转换。

其次，通过训练，大脑中的神经元可以只传递那些有助于大脑达成更大目标的信号。同样，我们可以训练神经网络的神经元只传递有用的信号。本章会在这些特性的基础上，介绍人工神经网络如何通过位（bit）和函数来模拟生物神经网络。

生物学对计算机科学的启发

生物学对计算机科学的启发不限于人工神经网络。过去的50年中，学术研究还探索了自然界中带给计算机科学灵感的其他主题，例如：

* 蚂蚁
* 白蚁
* 蜜蜂
* 遗传算法

蚁群已经被研究者看作一个强大的去中心化计算机，其中没有一只蚂蚁是会导致整个系统失效的中心节点。蚂蚁不断切换任务，通过定量共识协调等元启发式算法，找到接近最优的负载均衡解决方案。蚁群能够执行清洁、防御、筑巢和觅食任务，同时根据相关需求，为每个任务分配接近最佳数量的工蚁，过程中没有个体蚂蚁直接协调工作。

* + 1. 什么是深度学习

对很多人来说，深度学习很难定义，因为它在过去十年中慢慢地改变了形式。一个有用的定义是，深度学习是处理“两层以上的神经网络”的技术。这个定义的存疑之处是它使深度学习听起来像是自20世纪80年代以来一直存在一样。我们认为神经网络在最近几年展现出其辉煌成果之前，就已在架构上超越了早期的网络形式（并具有更强大的处理能力）。下面是神经网络发展的一些方面：

* 比之前的网络拥有更多神经元；
* 神经网络中出现了更复杂的层/神经元之间的连接方式；
* 用于进行训练的计算能力呈现爆炸式增长；
* 自动特征提取。

基于本书的主题，深度学习被定义为具有大量参数和层的神经网络，拥有以下四种基本网络架构之一：

* 无监督预训练网络（unsupervised pretrained network，UPN）
* 卷积神经网络（convolutional neural network，CNN）
* 循环神经网络（recurrent neural network，RNN）
* 递归神经网络

上述架构还存在一些变体，比如混合了卷积和循环的神经网络。本书会聚焦于上面列出的四种架构的网络。

自动特征提取是深度学习相较于传统机器学习算法的另一大优点。特征提取指由网络决定数据集的哪些特征可以可靠地用于标记数据的过程。历史上，机器学习的实践者花费了他们生命中的数月、数年甚至数十年来人工创建用于数据分类的穷举特征集。在2006年深度学习大爆发开始时，最先进的机器学习算法融会了人类数十年努力积累的用于对输入分类的相关特征。对于几乎所有需要人工微调的数据类型，深度学习在精度上都超过传统算法。这些深度网络有助于数据科学团队将他们的时间、汗水和泪水节省下来，去完成更有意义的任务。

* + 1. 钻进奇幻的兔子洞

深度学习对计算机科学观念的渗透，超过了近代历史上的大多数技术。部分原因是这一技术不仅拥有机器学习模型中顶级的精度，而且它的创造能力甚至让非计算机科学家着迷。一个例子是艺术生成演示，即一个深层网络基于某位著名画家的作品进行训练，然后能够以这位画家的独特风格渲染其他照片，如图1-2所示。

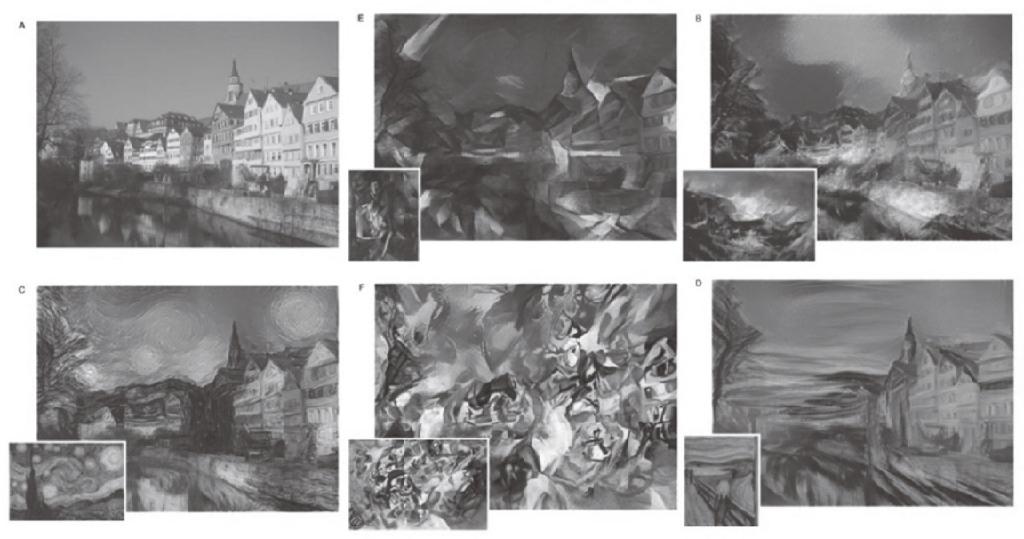


图1-2：Gatys等人2015年的艺术风格图片

这引发了对许多哲学问题的讨论，比如“机器有创造力吗”，以及进一步的“什么是创造力”之类的问题，我们把这些问题留给你后续思考。机器学习已经发展了很多年，它就像季节变化，微妙但稳定，直到有一天你醒来时发现机器已经成为Jeopardy节目（美国著名知识竞答节目）的冠军，或者击败了一位围棋大师。

机器能变得有智慧，并具有和人类同等的智能吗？人工智能是什么，它能变得多么强大？这些问题尚未得到解答，本书也没有所有的答案。我们只是试图展示机器智能的一些侧面。今天可以通过深度学习的实践来充实我们的生活环境。

**关于人工智能的进一步讨论**

如果想了解更多关于人工智能的信息，请阅读附录A。

* 1. 提出问题

要想理解有关机器学习应用的基础知识，最佳方式是从提出正确的问题开始。以下是需要定义的事项。

* 我们想要从中提取信息（模型）的输入数据是什么？
* 哪种模型最适合这个数据？
* 基于这个模型，我们希望从新的数据中探索出什么样的答案？

如果能回答这三个问题，我们就可以建立一个机器学习工作流，它将建立模型并产生我们想要的答案。为了更好地完成这个工作流，首先回顾一下为了实践机器学习所需了解的一些核心概念。稍后再看看在机器学习中如何将它们结合起来，然后利用这些信息加深我们对神经网络和深度学习的理解。

* 1. 机器学习背后的数学：线性代数

线性代数是机器学习和深度学习的基石，为求解用来建立模型的方程提供了数学基础。

线性代数的一本非常好的入门书是James E. Gentle的Matrix Algebra: Theory, Computations and Applications in Statis-tics。

我们从被称为标量的基本概念开始，来了解这个领域的一些核心概念。

* + 1. 标量

在数学中，“标量”一词指的是向量中的元素。标量是用于定义向量空间的实数和字段元素。

在计算中，“标量”与“变量”含义相同，是与符号名配对的存储位置。这个存储位置保存着一个被称为值的未知信息量。

* + 1. 向量

基于使用场景，向量的定义如下：

对于正整数n，向量是n元组、有序（多）集合或者n个数的数组，其中的数被称为元素或标量。

详细说来，即通过一个名为向量化的过程创建一个被称为“向量”的数据结构。向量中元素的数量被称为向量的“模”（或“长度”）。向量也可以用来表示n维空间中的点。空间意义上，从原点到由向量表示的点的欧几里得距离就是向量的“长度”。

在数学书中，向量经常写成下面的形式：