第 1 章　引　　言

好好学习，天天向上。

—— 毛泽东，1951 年题词

大数据时代，人类收集、存储、传输、管理数据的能力日益提高，各行各业已经积累了大量的数据资源，如著名的 *Nature* 杂志于 2008 年 9 月出版了一期大数据专刊 [1]，列举了生物信息、交通运输、金融、互联网等领域的大数据应用。如何有效分析数据并得到有用信息甚至知识成为人们关注的焦点。人们寄希望于智能数据分析来完成该项任务。机器学习是智能数据分析技术的核心理论。*Science* 杂志于 2015 年 7 月组织了一个人工智能专题 [2]，其中有关机器学习的内容依然占据了重要的部分。本章将讨论机器学习的基本目的、基本框架、思想发展以及未来走向。

1.1　机器学习的目的：从数据到知识

人类最重要的一项能力是能够从过去的经验中学习，并形成知识。千百年来，人类不断从学习中积累知识，为人类文明打下了坚实的基础。“学习”是人与生俱来的基本能力，是人类智能（human intelligence）形成的必要条件。自 2000 年以来，随着互联网技术的普及，积累的数据已经超过了人类个体处理的极限，以往人类自己亲自处理数据形成知识的模式已经到了必须改变的地步，人类必须借助于计算机才能处理大数据，更直白地说，我们希望计算机可以像人一样从数据中学到知识。

由此，如何利用计算机从大数据中学到知识成为人工智能研究的热点。“机器学习”（machine learning）是从数据中提取知识的关键技术。其初衷是让计算机具备与人类相似的学习能力。迄今为止，人们尚不知道如何使计算机具有与人类相媲美的学习能力。然而，每年都有大量新的针对特定任务的机器学习算法涌现，帮助人们发现完成这些特定任务的新知识（有时也许仅仅是隐性新知识）。对机器

学习的研究不仅已经为人们提供了许多前所未有的应用服务（如信息搜索、机器翻译、语音识别、无人驾驶等），改善了人们的生活，而且也帮助人们开辟了许多新的学科领域，如计算金融学、计算广告学、计算生物学、计算社会学、计算历史学等，为人类理解这个世界提供了新的工具和视角。可以想见，作为从数据中提取知识的工具，机器学习在未来还会帮助人们进一步开拓新的应用和新的学科。

机器学习存在很多不同的定义，常用的有三个。第一个常用的机器学习定义是“计算机系统能够利用经验提高自身的性能”，更加形式化的论述可见文献 [3]。机器学习名著《统计学习理论的本质》给出了机器学习的第二个常见定义，“学习就是一个基于经验数据的函数估计问题”[4]。在《统计学习基础》这本书的序言里给出了第三个常见的机器学习定义，“提取重要模式、趋势，并理解数据，即从数据中学习”[11]。这三个常见定义各有侧重：第一个聚焦学习效果，第二个的亮点是给出了可操作的学习定义，第三个突出了学习任务的分类。但其共同点是强调了经验或者数据的重要性，即学习需要经验或者数据。注意到提高自身性能需要知识，函数、模式、趋势显然自身是知识，因此，这三个常见的定义也都强调了从经验中提取知识，这意味着这三种定义都认可机器学习提供了从数据中提取知识的方法。众所周知，大数据时代的特点是“信息泛滥成灾但知识依然匮乏”。可以预料，能自动从数据中学到知识的机器学习必将在大数据时代扮演重要的角色。

那么如何构建一个机器学习任务的基本框架呢？

1.2 机器学习的基本框架

考虑到我们希望用机器学习来代替人学习知识，因此，在研究机器学习以前，先回顾一下人类如何学习知识是有益的。对于人来说，要完成一个具体的学习任务，需要学习材料、学习方法以及学习效果评估方法。如学习英语，需要英语课本、英语磁带或者录音等学习材料，明确学习方法是背诵和练习，告知学习效果评估方法是英语评测考试。检测一个人英语学得好不好，就看其利用学习方法从学习材料得到的英语知识是否能通过评测考试。机器学习要完成一个学习任务，也需要解决这三方面的问题，并通过预定的测试。

对应于人类使用的学习材料，机器学习完成一个学习任务需要的学习材料，一般用描述对象的数据集合来表示，有时也用经验来表示。对应于人类完成学习任务的学习方法，机器学习完成一个学习任务需要的学习方法，一般用学习算法来表示。对应于人类完成一个学习任务的学习效果现场评估方法（如老师需要时时观察课堂气氛和学生的注意力情况），机器学习完成一个学习任务也需要对学习效果进行即时评估，一般用学习判据来表示。对于机器学习来说，用来描述数

据对象的数据集合对最终学习任务的完成状况有重要影响，用来指导学习算法设计的学习判据有时也用来评估学习算法的效果，但一般机器学习算法性能的标准评估会不同于学习判据，正如人学习的学习效果即时评估方式与最终的评估方式一般也不同。对于机器学习来说，通常也会有特定的测试指标，如正确率，学习速度等。

可以用一个具体的机器学习任务来说明。给定一个手写体数字字符数据集合，希望机器能够通过这些给定的手写体数字字符，学到正确识别手写数字字符的知识。显然，学习材料是手写体数字字符数据集，学习算法是字符识别算法，学习判据可以是识别正确率，也可以是其他有助于提高识别正确率的指标。

数据集合、学习判据、学习算法对于任何学习任务都是需要讨论的对象。数据集合的不同表示，影响学习判据与学习算法的设计。学习判据与学习算法的设计密切相关，下面分别讨论。

1.2.1　数据集合与对象特性表示

对于一个学习任务来说，我们希望学到特定对象集合的特定知识。无论何种学习任务，学到的知识通常是与这个世界上的对象相关。通过学到的知识，可以对这个世界上的对象有更好的描述，甚至可以预测其具有某种性质、关系或者行为。为此，学习算法需要这些对象的特性信息，这些信息可以客观观测，即关于特定对象的特性信息集合，该集合一般称为对象特性表示，是学习任务作为学习材料的数据集合的组成部分。理论上，用来描述对象的数据集合的表示包括对象特性输入表示、对象特性输出表示。

显然，对象特性输入表示是我们能够得到的对象的观测描述，对象特性输出表示是我们学习得到的对象的特性描述。需要指出的是，对象的特性输入表示或者说对象的输入特征一定要与学习任务相关。根据丑小鸭定理（Ugly Duckling Theorem）[5]，不存在独立于问题而普遍适用的特征表示，特征的有效与否是问题依赖的。丑小鸭定理是由 Satosi Watanabe 于 1969 年提出的，其内容可表述为“如果选定的特征不合理，那么世界上所有事物之间的相似程度都一样，丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大”。该定理表明在没有给定任何假设的情况下，不存在普适的特征表示；相似性的度量是特征依赖的，是主观的、有偏置的，不存在客观的相似性度量标准。因此，对于任何机器学习任务来说，得到与学习任务匹配的特征表示是学习任务成功的首要条件。对于机器学习来说，一般假设对象特征已经给定，特别是对象特性输入表示。

对于对象特性输入表示，通常有三种表示方式。一种是向量表示，对于每个对象，可以相对独立地观察其特有的一些特征。这些特征组成该对象的一个描述，

并代表该对象。第二种表示是网络表示，对于每个对象，由其与其他对象的关系来描述，简单说来，观察得到的是对象之间的彼此关系。第三种是混合表示, 对于每个对象，其向量表示和网络表示同时存在。

不论对于人还是机器，能够提供学习或者训练的对象总是有限的。不妨假设有 N 个对象，对象集合为 $O=\{o_1,o_2,\cdots,o_N\}$，其中 o_k 表示第 k 个对象。其对应的对象特性输入表示用 $X=\{x_1,x_2,\cdots,x_N\}$ 来表示，其中 x_k 表示对象 o_k 的特性输入表示。当每个对象有向量表示时，x_k 可以表示为 $x_k=[x_{1k},x_{2k},\cdots,x_{pk}]^{\mathrm{T}}$。因此，对象特性输入表示 X 可以用矩阵 $[x_{\tau k}]_{p\times N}$ 来表示，其中 p 表示对象输入特征的维数，$x_{\tau k}$ 表示 o_k 的第 τ 个输入特征值，这些特征值可以是名词性属性值，也可以是连续性属性值。

如果对象特性输入表示 X 存在网络表示，即 X 可以用矩阵 $[\mathfrak{N}_{kl}]_{N\times N}$ 来表示，其中 $\mathfrak{N}_{kl}$ 表示对象 o_k 与对象 o_l 的网络关系。如果是相似性关系，则对象特性输入表示 X 为相似性矩阵 $S(X)=[s_{kl}]_{N\times N}$，其中 s_{kl} 表示对象 o_k 与对象 o_l 的相似性。通常，s_{kl} 越大表明对象 o_k 与对象 o_l 的相似性越大。因此，对象 o_k 可以由行向量 $[s_{k1},s_{k2},\cdots,s_{kN}]$ 表示。如果是相异性关系，则对象特性输入表示 X 为相异性矩阵 $D(X)=[D_{kl}]_{N\times N}$，其中 D_{kl} 表示对象 o_k 与对象 o_1 的相异性。类似的，D_{kl} 越大表明对象 o_k 与对象 o_l 的相异性越大。因此，对象 o_k 可以由行向量 $[D_{k1},D_{k2},\cdots,D_{kN}]$ 表示。如果是相邻关系，对象特性输入表示 X 为邻接性矩阵 $A(X)=[a_{kl}]_{N\times N}$，其中 a_{kl} 表示对象 o_k 与对象 o_l 是否相邻，通常其取值为 0 或者 1。

对应的对象特性输出表示用 $Y=\{y_1,y_2,\cdots,y_N\}$ 来表示，其中 y_k 表示对象 o_k 的特性输出表示。具体的表示形式由学习算法决定，通常是对象特性输出表示 Y 可以用矩阵 $[y_{\tau k}]_{d\times N}$ 来表示，其中 d 表示对象输出特征的维数，$y_{\tau k}$ 表示 o_k 的第 τ 个输出特征值，这些特征值通常是连续性属性值。

显然，除去对象特性输入、输出表示，数据集合还有其他部分，这些部分的表示与知识表示有关，通常依赖于知识表示。知识表示不同，学习算法的数据集合输入输出表示也会不同。一个容易想到的公开问题是，适合于机器学习的统一知识表示是否存在? 如果存在，是何形式? 现今的机器学习方法一般是针对具体的学习任务，设定具体的知识表示。因此，本章先不讨论学习算法的输入输出统一表示, 这个问题留待第 2 章讨论。

1.2.2 学习判据

完成一个学习任务，需要一个判据作为选择学习到的知识好坏的评价标准。理论上，符合一个学习任务的具体化知识可以有很多。通常，如何从中选出最好

的具体化知识表示是一个 NP 难问题。因此，需要限定符合一个特定学习任务的具体化知识范围，适当减小知识假设空间的大小，减少学习算法的搜索空间。为了从限定的假设空间选择最优的知识表示，需要根据不同的学习要求来设定学习判据对搜索空间各个元素的不同分值。判据设定的准则有很多，理论上与学习任务相关，本书将在以后的章节中进行讨论。需要指出的是，有时学习判据也被称为目标函数。在本书中，对于这两个术语不再特意区别。

1.2.3　学习算法

在学习判据给出了从知识表示空间搜索最优知识表示的打分函数之后，还需要设计好的优化方法，以便找出对应于打分函数达到最优的知识表示。此时，机器学习问题通常归结为一个最优化问题。选择最优化方法对有效完成学习任务很关键。目前，最优化理论在机器学习问题中已经变得越来越重要。典型的最优化算法有梯度下降算法、共轭梯度算法、伪牛顿算法、线性规划算法、演化算法、群体智能等。如何选择合适的优化技术，得到快速、准确的解是很多机器学习问题的难点所在。这就要求工程技术和数学理论相结合，以便很好地解决优化问题。一般建议初学者先采用已有的最优化算法，之后再设计专门的优化算法。

是否有不依赖于具体问题的最优学习算法呢？如果有的话，只需学一种算法就可以包打天下了。可惜的是，结论是否。著名的没有免费午餐定理已经明确指出：不存在对于所有学习问题都适用的学习算法 [6–8]。

1.3　机器学习思想简论

机器学习作为一个单独的研究方向，应该说是在 20 世纪 80 年代第一届 ICML 召开之后才有的事情。但是，广义上来说，机器学习任务，或者学习任务，一有人类就出现了。在日常生活中，人们每天都面临如何从自己采集的数据中提取知识进行使用的问题。比如，大的方面，需要观察环境的变化来学习如何制定政策使得我们这个地球可持续发展；小的方面，需要根据生活的经验买到一个可口的柚子或者西瓜，选择一个靠谱的理发师，等等。在计算机出现以前，数据采集都是人直接感知或者操作，采集到的数据量较小，人可以直接从数据中提取知识，并不需要机器学习。如对于回归问题，高斯在 19 世纪早期（1809）就发表了最小二乘法；对于数据降维问题，卡尔·皮尔逊在 1901 年就发明了主成分分析（PCA）；对于聚类问题，K-means 算法最早也可追溯到 1953 年 [9]。但是，这些算法和问题被归入机器学习，也只有在机器收集数据能力越来越成熟导致人类直接从数据中提取知识成为不可能之后才变得没有异议。

在过去的 30 年间，机器学习从处理仅包含上百个样本数据的玩具问题（toy-problem）起步，发展到今天，已经成为从科学研究到商业应用的标准数据分析工具。但是其研究热点也几经变迁，本书将从思想史的角度略加总结。

机器学习最早的目标是从数据中发现可以解释的知识，在追求算法性能的同时，强调算法的解释性。早期的线性感知机、决策树和最近邻等算法可以说是这方面的典型代表作。但是，1969 年，Minsky 指出线性感知机算法不能解决异或问题 [10]。由于现实世界的问题大多是非线性问题，而异或问题可以说是最简单的非线性问题，由此可以推断线性感知机算法用处不多。这对于以线性感知机算法为代表的神经网络研究可以说是致命一击，直接导致了神经网络甚至人工智能的第一个冬天。感知机算法的发明人、神经网络先驱 Rosenblatt 于 1971 年因故去世，更加增添了这个冬天的寒意。

需要指出的是，很多实际应用并不要求算法具有可解释性。比如机器翻译、天气预报、卜卦算命等。在这种需求下，如果一个算法的泛化性能能够超过其他同类算法，即使该算法缺少解释性，则该算法依然是优秀的学习算法。20 世纪 80 年代神经网络的复苏，其基本思路即为放弃解释性，一心提高算法的泛化性能。神经网络放弃解释性的最重要标志是其激活函数不再使用线性函数，而是典型的非线性函数如 Sigmoid 函数和双曲函数等，其优点是其表示能力大幅提高，相应的复杂性也极度增长。众所周知，解释性能好的学习算法，其泛化性能也要满足实际需求。如果其泛化性能不佳，即使解释性好，人们也不会选用。在 20 世纪 80 年代，三层神经网络的性能超过了当时的分类算法如决策树、最近邻等，虽然其解释性不佳，神经网络依然成为当时最流行的机器学习模型。在神经网络放弃解释性之后，其对于算法设计者的知识储备要求也降到了最低，因此，神经网络在 20 世纪 80 年代吸引了大批的研究者。

当然，也有很多实际应用要求算法具有可解释性，如因果关系发现、控制等。应该说，同时追求解释性和泛化性能一直是非神经网络机器学习研究者设计学习算法的基本约束。一旦一个算法既具有很好的解释性，其性能又超过神经网络，神经网络研究就将面临极大的困境。这样的事情在历史上也曾真实地发生过。1995 年 Vapnik 提出了支持向量机分类算法，该算法解释性好，其分类性能也超过了当时常见的三层神经网络，尤其需要指出的是，其理论的分类错误率可以通过 Valiant 的 PAC 理论来估计。这导致了神经网络研究的十年沉寂，有人也将其称为人工智能的第二个冬天。在这期间，大批原先的神经网络研究者纷纷选择离开，只有少数人坚持研究神经网络。这个时间段对于机器学习来说，显然不是冬季。在这十年间，人们提出了概率图理论、核方法、流形学习、稀疏学习、排序学习等多种机器学习新方向。特别是在 20 世纪末和 21 世纪初，由于在搜索引擎、字符识别等应用领域取得的巨大进展，机器学习的影响力日益兴旺。其标志事件

有：1997 年 Tom Mitchell 机器学习经典教科书的出现 [3]，2010 年和 2011 年连续两年图灵奖颁发给了机器学习的研究者 Valiant 和 Pearl。

三十年河东，三十年河西。2006 年以后，神经网络突破了三层网络结构限制，大幅提高了模型的表示能力，又逢大数据时代相伴而生的高计算能力，神经网络化身深度学习，再次将分类能力提高到同时代其他模型无法匹敌的程度，有人将其称为人工智能的第三个春天。在机器学习的许多应用领域，深度学习甚至成为机器学习的代名词。虽然如此，时至今日，深度学习只是机器学习的一个分支，无论其沉寂或者过热，都不能逆转而只能加速全部机器学习本身应用越来越普及、理论越来越深入的发展趋势。

如今，机器学习算法每天被用来帮助解决不同学科不同商业应用的各种实际数据分析问题，相关的研究者每年也会针对相同或者不同的学习问题设计成百上千的新学习算法。面对一个学习任务，使用者经常面对十几个甚至几百个学习算法，如何从已有的算法中选择一个适当的方法或者设计一个适合自己问题的算法成为当前机器学习研究者和使用者必须面对的问题。早在 2004 年，周志华在国家自然科学基金委员会秦皇岛会议上做了一个名为“普适机器学习”的学术报告，其中曾明确指出：机器学习“以 Tom Mitchell 的经典教科书（McGraw Hill 出版社，1997）为例，很难看到基础学科（例如数学、物理学）教科书中那种贯穿始终的体系，也许会让人感到这不过是不同方法和技术的堆砌”。因此，已有的机器学习算法是否存在共性，是否存在统一的框架来描述机器学习算法的设计过程，就变成了一个亟待解决的问题。本书将从知识表示的角度出发，来阐述我们对这一问题的研究结果，并据此讨论现存的机器学习算法的适用范围。

延 伸 阅 读

目前有多种不同的视角和观点研究机器学习。例如，可以从概率图角度来看待机器学习 [12,13]，可以从统计角度来讨论机器学习 [11]，还可以从神经网络的观点来阐述机器学习 [16]，也可以调和以上各派观点来阐述机器学习 [17]。客观地说，上述观点都有一定道理，但是也有一个共同而重要的缺陷，那就是没有给出一个统管一切学习（包括机器、人类和生物）的理论。这正是 Jordan 和 Mitchell 在 2015 年在*Science*上发文指出的，机器学习所关注的两大问题之一：是否存在统管一切机器、人类和生物的学习规律 [14]。本书将致力于解决这一个问题。为此，本书采取了不同于以往的观点，从知识表示这一角度来阐述机器学习，并以此为出发点对现在的机器学习方法进行统一研究。

本书的基本出发点是，每个机器学习算法都有自己的知识表示。如果数据中

含有的知识不适合特定机器学习算法的知识表示，期望这种机器学习算法能够学到数据中含有的知识并不现实。因此，知识表示对于机器学习至关重要。但是，众所周知，经典的知识定义是柏拉图提出的，在 2000 多年的时间里未受到严重的挑战。直到 1963 年，盖梯尔写了一生唯一的一篇三页纸论文。这短短的三页纸使盖梯尔成为哲学史上绕不过去的人物，改变了盖梯尔的命运，也改变了知识论的发展进程。这三页纸中提出的盖梯尔难题直接否定了经典的知识定义 [18]。其直接后果是到目前并没有一个统一的知识定义，更不用说知识的统一表示。因此，暂时放弃知识的整体研究，而致力于知识的基本组成单位研究也许是一条更为可行的路径。本书即是这样的一个尝试和努力。

注意到知识的最小组成单位是概念 [15]，而目前的机器学习主要关注于从数据中提取概念。因此，研究概念的表示也将有助于机器学习的研究。正是从这一点出发，本书以一种统一的方式研究了常见的机器学习算法，如密度估计、回归、数据降维、聚类和分类等。

当然，机器学习的发展不仅与知识表示直接相关，也与最优化、统计等密切相关。历史上，计算机、数学、心理学、神经学、生物信息学、哲学等很多学科都曾极大地促进了机器学习的发展。未来是否还有其他学科对机器学习有重要影响，也是一个有趣的话题。

最后，稍微讨论一下与机器学习相关的学习、研究资料。目前，机器学习的发展方兴未艾，特别是学习算法的研究成果日新月异。除了已经列入参考文献的部分经典著作外，还有很多有影响的学术会议、学术期刊和网络资源等，如机器学习相关学术会议 ICML、NIPS、COLT，学术期刊 TPAMI 和 JMLR，网络资源 http://videolectures.net/，有兴趣的读者可以自行查阅。

习　　题

1. 机器学习可以从哪些观点或角度进行研究或者阐述？你比较赞同哪种观点？为什么？
2. 你认为机器学习的发展存在哪些问题？如何有效地解决这些问题？
3. 机器学习综合了很多其他学科的知识，正是由于这些学科的加入，才促使了机器学习的发展。你认为还有必要将哪些学科或领域的知识加入到机器学习中？机器学习未来将何去何从？
4. 请你拿笔任意地在纸上写 10 次“machine”这个单词，再请你一个同学也在纸上写 10 次这个单词。然后你们观察这 20 个单词（可以看成 20 张图片），试着去提取它们的特征，比如笔画、弯曲处和圈的特点，来识别你的笔迹和你同学的笔迹。然后想想如何让计算机做这件事情。

参 考 文 献

[1] Nature. Special Issue. Big Data. http://www.nature.com/news/specials/bigdata/index.html. 2008.

[2] Science. Special Issue. Artificial Intelligence. http://www.sciencemag.org/site/special/artificialintelligence/index.xhtml. 17 July 2015.

[3] Mitchell　T. Machine learning. New York: MaGraw Hill, 1997.

[4] Vapnik V N. The nature of statistical learning theory. 2nd ed. New York: Springer-Verlag. 1999.（其中文版见：统计学习理论的本质，张学工译. 北京：清华大学出版社，2000）

[5] Watanabe S. Knowing and guessing: a quantitative study of inference and information. New York: Wiley. 1969: 376-377.

[6] Wolpert D H, Macready W G. No free lunch theorems for search. Technical Report SFI-TR-95-02-010. Sante Fe, NM, USA: Santa Fe Institute. 1995.

[7] Wolpert D H. The lack of a priori distinctions between learning algorithms. Neural Computation, 1996, 8(7): 1341-1390.

[8] Wolpert D H, Macready W G. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67-82.

[9] Thorndike R L. Who belongs in the family. Psychometrika, 1953, 18(4): 267-276.

[10] Minsky M, Papert S. Perceptons. Cambridge, MA: The MIT Press, 1969.

[11] Hastie T, Tibshirini R, Friedman J H. The elements of statistical learning. Springer, 2003.

[12] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge, MA: The MIT Press, 2009.

[13] Murphy K P. Machine learning: a probabilistic perspective. Cambridge, MA: The MIT Press, 2012.

[14] Jordan M I, Mitchell T. Machine learning: trends, perspectives, and prospects. Science, 2015, 349: 255-260.

[15] Murphy G L. The big book of concepts. Cambridge, MA: The MIT Press, 2004.

[16] Haykin S O. Neural networks and learning machines. Eaglewood Cliff, NJ: Prentice Hall, 2008.

[17] 周志华. 机器学习. 北京：清华大学出版社, 2016.

[18] Gettier E L. Is justified true belief knowledge? Analysis, 1963, 23(6): 121-123.