中国计算机学会学术著作丛书

机器学习

从公理到算法

于剑 著

清华大学出版社

中国计算机学会学术著作丛书

机器学习

从公理到算法

于剑 著

清华大学出版社
北京

内 容 简 介

这是一本基于公理研究学习算法的书。共 17 章，由两部分组成。第一部分是机器学习公理以及部分理论演绎，包括第 1、2、6、8 章，论述学习公理以及相应的聚类、分类理论。第二部分关注如何从公理推出经典学习算法，包括单类、多类和多源问题。第 3～5 章为单类问题，分别论述密度估计、回归和单类数据降维。第 7、9～16 章为多类问题，包括聚类、神经网络、K 近邻、支持向量机、Logistic 回归、贝叶斯分类、决策树、多类降维与升维等经典算法。最后第 17 章研究了多源数据学习问题。

本书可以作为高等院校计算机、自动化、数学、统计学、人工智能及相关专业的研究生教材，也可以供机器学习的爱好者参考。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究. 侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

机器学习：从公理到算法/于剑著.—北京：清华大学出版社，2017 (2017.8 重印)
（中国计算机学会学术著作丛书）
ISBN 978-7-302-47136-3

Ⅰ.①机… Ⅱ.①于… Ⅲ.①机器学习 Ⅳ.①TP181

中国版本图书馆 CIP 数据核字(2017)第 116725 号

责任编辑： 薛　慧
封面设计： 何凤霞
责任校对： 赵丽敏
责任印制： 杨　艳

出版发行： 清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者： 三河市吉祥印务有限公司
经　　销： 全国新华书店
开　　本： 175mm×245mm　　**印　张：** 15.25　　**插　页：** 1　　**字　数：** 301 千字
版　　次： 2017 年 7 月第 1 版　　**印　次：** 2017 年 8 月第 2 次印刷
印　　数： 2001～4000
定　　价： 80.00 元

产品编号：069688-01

自　　序

北大读博之时，蒙先师指点，研究归类之术。其理繁复，致予目眩五色，心力交瘁。然应门之童可辨识诸物，岂懂是理哉？

疑惑日久，遍求诸经。尝读维特根斯坦之哲学研究，知相似性为归类之要，然血指汗颜，不得要领。倏忽十年，访友寻师。一日顿悟，曰：归哪类，像哪类。像哪类，归哪类。此即孔子所谓“君君臣臣父父子子”之意也。周易所谓“水流湿，火就燥，云从龙，风从虎”之意也。

如不然，归哪类，不像哪类；像哪类，不归哪类。所谓君不君，臣不臣，父不父，子不子。长此以往，名不正，言不顺，雌雄莫辨，黑白难分，不亦谬乎！

是为序。

于剑
jianyu@bjtu.edu.cn
2014 年 5 月

前　言

机器学习的主要目的是从有限的数据中学习到知识，而知识的基本单元是概念。借助于概念，人类可以在繁复的思想与多彩的世界之间建立起映射，指认各种对象，发现各种规律，表达各种想法，交流各种观念。一旦缺失相应的概念，人们将无法思考、交流，甚至无法顺利地生活、学习、工作、医疗、娱乐等。哲学家如卡西尔等甚至认为人类的本质特性是能够使用和创造各种符号概念。因此，如何使机器能够像人一样自动发现、运用概念，正是机器学习的基本研究内容。本书将集中讨论这个问题。

所谓的概念发现，是指从一个给定概念（或者概念集合）的有限外延子集提取对应的概念（或者概念集合）表示，又称归类问题。通过自然进化，人类可以从一个概念（或概念集合）的有限外延子集（有限的对象）中轻松提取概念（或概念集合）自身。对于人类如何处理归类问题，人们已经研究了很多年，发明了许多理论，比如经典概念理论、原型理论、样例理论和知识理论等，积累了很多的研究成果。本书借助认知科学的研究成果，提出了类的统一表示数学模型，以及与之相关的归类问题的统一数学表示。由此提出了类表示公理、归类公理和分类测试公理。据此，本书分别研究了归类结果分类、归类算法分类等诸多问题。特别需要提出的是，本书首次归纳了归类算法设计应该遵循的 4 条准则 —— 类一致性准则、类紧致性准则、类分离性准则和奥卡姆剃刀准则。在理论上，任何机器学习算法的目标函数设计都遵循上述 4 条准则的 1 条或者数条。

对于具体的机器学习问题，本书依据奥卡姆剃刀准则，按照归类表示从简单到复杂的顺序，重新进行了组织。本书不仅论述了单类问题比多类问题的归类表示简单，聚类问题比分类问题的归类表示简单，单源数据学习比多源数据学习的归类表示简单，而且对于单类问题、多类问题自身的归类表示复杂度也进行了研究。在此基础上，指出单类问题包括密度估计、回归和单类数据降维等，并借助提出的公理框架以统一的方式演绎推出了在密度估计、回归、数据降维、聚类和分类等问题中常用的机器学习算法。

本书中章节的组织结构都是类似的，特别是与具体学习算法有关的章节。每

试读结束：需要全本请在线购买：　www.ertongbook.com

章有一个简短的开篇词。如果该章是学习算法章节，该开篇词用来简要说明本章算法的主要设计思想。如果该章是理论章节，该开篇词说明该理论问题的主要目标。每章结尾有延伸阅读或者讨论，延伸阅读提供更深入的相关阅读文献，讨论说明本章的相关内容与分析或者尚未解决的问题。

作者讲授机器学习已十数年，有感于当前的机器学习算法理论依据过多过杂，同时也一直羡慕欧氏几何从五条公理出发导出所有结论的风格。撰写本书，既是将欧氏几何风格移植到机器学习的一个尝试，更是试图为机器学习与模式识别提供一个统一但又简单的理论视角。总之，机器学习公理化这个问题在本书中提出，也在本书中解决了。

于剑

2017 年 3 月

目　录

试读结束：需要全本请在线购买： www.ertongbook.com

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，第三个突出了学习的可理解性。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。

试读结束：需要全本请在线购买：www.ertongbook.com