HZ BOOKS
华章IT

智能系统与技术丛书

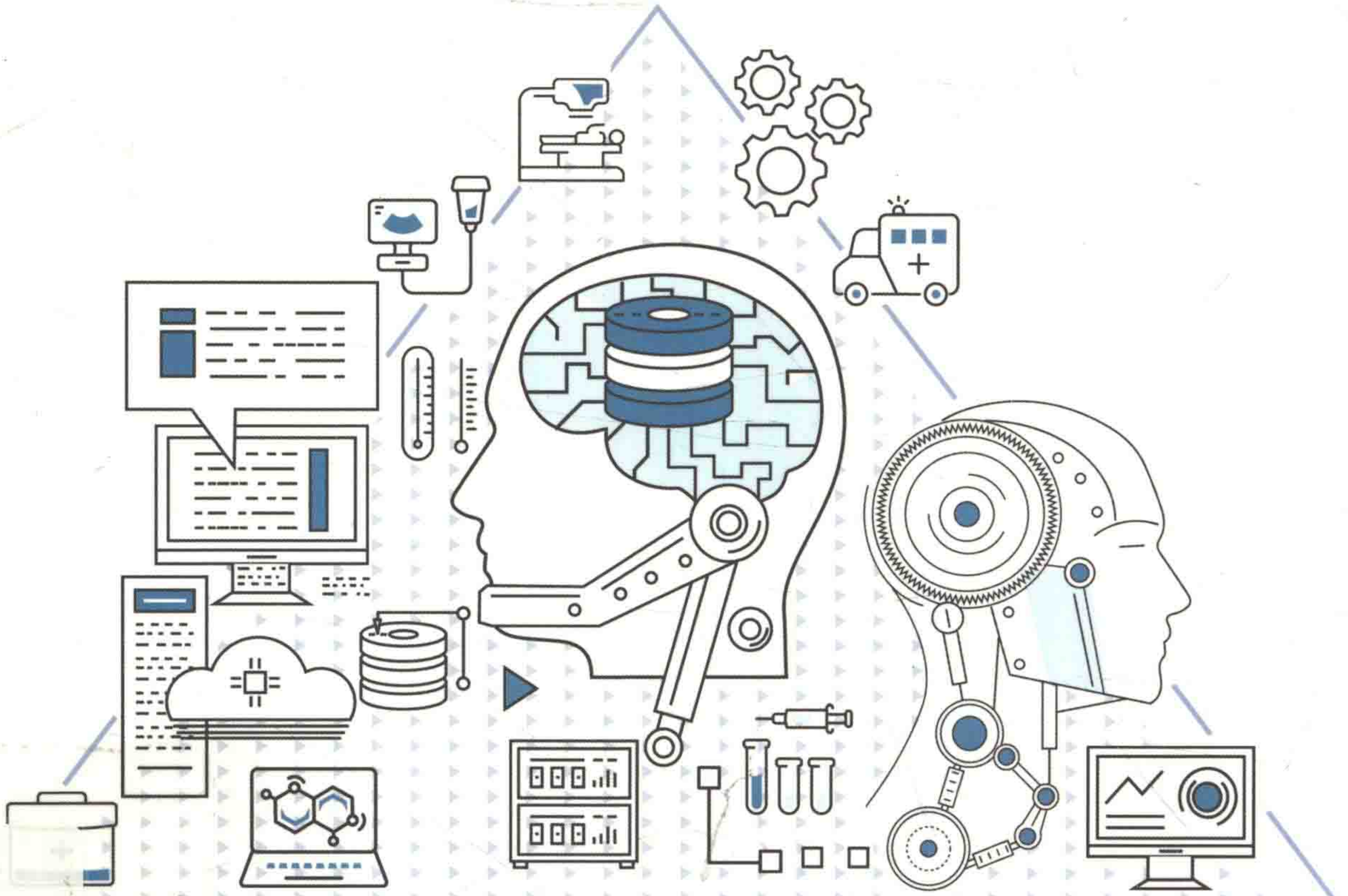
Machine Learning Technology and Practice

Deep Applications of Medical Big Data

机器学习技术与实战

医学大数据深度应用

[加] 洪松林（Hong Song Lin）编著



机械工业出版社
China Machine Press

Machine Learning Technology and Practice
Deep Applications of Medical Big Data

机器学习技术与实战
医学大数据深度应用

[加] 洪松林（Hong Song Lin）编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习技术与实战：医学大数据深度应用 /（加）洪松林（Hong Song Lin）编著 . —北京：机械工业出版社，2018.4
（智能系统与技术丛书）

ISBN 978-7-111-59599-1

I. 机… II. 洪… III. 机器学习 – 应用 – 医学 – 数据处理 – 研究 IV. R319

中国版本图书馆 CIP 数据核字（2018）第 066671 号

本书版权登记号：图字 01-2018-0657

机器学习技术与实战：医学大数据深度应用

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：吴 怡 责任校对：李秋荣
印 刷：三河市宏图印务有限公司 版 次：2018 年 5 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：21.75
书 号：ISBN 978-7-111-59599-1 定 价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

献给我亲爱的 Eddie 和我的家人！

PREFACE

前　言

什么是机器学习？现在恐怕无需再做基本概念解释了。在本书中，我们谈机器学习的实用技术。我们知道，有了数据，就要做很多分析工作。其中很常见的、很基本的一个分析是，针对目标变量，我们需要从大量的候选变量（可能是几百个、几千个）中，探索、发现哪些变量与目标变量具有较强的广义相关性。我们可能应用很多不同的算法，一一对每个候选变量与目标变量进行相关性探索尝试，可有时还是没能找到一个有显著相关性的变量。不少人可能都觉得没办法了。但是，没有找到显著的独立相关变量，不意味着不存在任何相关变量了，数据中有可能存在着多个变量组合与目标变量具有较强的相关性（多变量相关组），或者说，与目标变量具有较强相关性的某个变量在数据中被“拆分”成了多个与目标变量不具有较强相关性的分变量。那么，在几百个甚至成千上万个候选变量中，如何有效地找到一个或多个多变量相关组呢？这是机器学习技术与工程实践中一个典型的深入课题。解决这个问题，就像下围棋一样，棋局太多、变化太多，着法也太多。机器学习中类似的分析课题有很多，这需要我们不断地探索、不断地实践、不断地创新、不断地积累，以便在千变万化的“棋局”中找到解决之道、制胜之道！

机器学习作为一种自动化、智能化的深度分析技术，从更高的层面上讲，其目的就是要从由数据代表的真实世界事物中探索和挖掘潜在规律和隐含机理，因此，机器学习除了是一门实用的应用技术外，它的发展前沿还是奥秘揭示、知识发现、科学探索！更高瞻远瞩一些，机器学习随着理论和实践的不断深入，已经不再是原先狭义的“数据利用”和“知识发现”了，正在越来越深入到数学发现、甚至哲学发现以及科学发现了。例如，机器学习通常从刻画客观事物的各类大数据中挖掘出内在的规律，并期望能得到可靠、精准的可预测性结果。但是，随着机器学习应用和研究的深入，我们发现了大量不可预测的现象与问题。通常，技术人员会想是数据出现了问题？还是算法出现了问题？因为人们的传统思维通常是建立在确定性理论基础之上的。但是，科学家们已经越来越多地意识到、甚至认识到了世界上大量不确定性现象的客观存在。

那么，数据中出现的这种不可预测性，很可能是由不确定性系统产生的。现实世界中，除了我们认识到的确定性系统之外，还存在着很多不确定性系统，这些系统中拥有大量的非线性的、无序的现象和事物。例如，量子力学中的不确定性原理、混沌学中确定性系统中的无序随机性，都属于不确定性，也就是说，至少是目前技术水平下，是不可预测的。但是，系统中存在着混沌性和无序性，并不意味着无规律性。实际上，很多系统中的非线性无序状态中蕴含着许多规律性，只不过现代的理论和技术比较有限，尚不能很好地认识和应用这些规律。例如，混沌学中洛伦茨奇异吸引子是一个美丽的无序状态，它是有规律的，数据的表现貌似随机，但却遵循着一定之规（数学模型）。

实际上，除了混沌学发现了大量的无序现象外，还有其他学科涉及不确定性系统的研究，例如，概率论也是研究无序（随机）问题的一个分支学科。无序（随机）与有序（确定）是相对的，而不同的无序（随机）之间是相对的。以上都体现了系统的不确定性，由数据表达的时候，就出现了不可预测性。这就需要机器学习或者数据挖掘的理论、技术与实践还要不断创新和发展。因此，我们说，机器学习在现在和未来，作为现实世界科学探索的一个工具和技术，将不断地探索和发现包括不确定性系统产生的大量客观规律，以便更好地服务于各行各业的应用实践！

我们在本书中尽可能将理论与实践相结合，既重于实践应用又深入理论原理。理论是灰色的，而实践则是最鲜活的。本书是机器学习应用方面的书籍，我们希望尽可能多讲些实践和案例，并多用图画、图表说明大部分的机器学习原理和应用，让读者更能贴近实际。

本书主要内容

第1章“机器学习基础”介绍机器学习应用的基础内容，希望能快速引领读者进入机器学习领域。该章包括机器学习中一些基本概念，如数据的“形状”、机器学习要素等；机器学习的应用概念，如事物与维度、分布与关系、描绘与预测、现象与知识、规律与因果；机器学习基础概念，如无限三维嵌套空间，分数维度空间，不确定论等。

第2章“数据探索”介绍机器学习应用活动的前期工作，即数据探索和数据准备工作，包括数据关系探索、数据特征探索、数据选择、数据处理。

第3章“机器学习技术”介绍机器学习的算法，一个好的、合适的算法在机器学习应用项目中起着至关重要的作用。本书从实际应用出发，介绍一些比较经典的算法，也包括一些我们为应用编写的新算法，以及一些算法流程，算法包括聚类分析、特性选择、特征抽取、关联规则、分类和预测、时间序列、深度学习等。

第4章“机器学习应用案例”介绍应用上一章中提到的一些算法开发商业应用的案例。

这些案例不仅体现了算法的实践应用，也展现了机器学习应用各个环节的工作内容。该章将主要介绍特性选择模型的应用、分类模型的应用等。

第5章“机器学习应用系统开发”介绍智能医学科研系统IMRS的设计思路与步骤，包括从应用需求的产生、解决思路、系统设计、应用实现、效果评价与总结等完整过程，具体剖析IMRS的几个重要模块的开发方法，包括异常侦测模型、特征抽取模型，以及算法开发。

第6章“机器学习系统应用（一）：结构数据挖掘”介绍如何使用机器学习应用系统IMRS。按照临床科研的普遍需求，我们将IMRS的功能划分为六个方向：分布探索、关系探索、特征探索、异常探索、推测探索和趋势探索，该章介绍前五个方向的应用。

第7章“机器学习系统应用（二）：非结构数据挖掘”继续介绍如何使用机器学习应用系统IMRS，包括文本挖掘技术、文本数据挖掘在医学上的应用、文本分词的实现、文本智能搜索、文本聚类与分类的应用、文本主题提取应用。

第8章“基于机器学习的人工智能应用”介绍人工智能在医学上的应用：智能医学诊断系统的设计思路与应用，还介绍了混沌人工智能的概念、应用及展望。

致谢

现在，大数据和机器学习是热门，长年从事这个领域工作的我及我的团队都很忙，能够出版这本书实属不易。需要感谢的是我公司的Sun Chen（孙辰），他是来自澳大利亚的资深数据分析师，悉尼大学统计学硕士毕业，在本书的编写和整理过程中做了不少的协助工作，在此表示由衷的感谢！当然，机械工业出版社的吴怡编辑给予了我一贯的支持，她严谨的学术态度和丰富的编辑专业经验，不仅是本书质量的保证，也给我留下了深刻的印象，再次向吴老师表示衷心的感谢！最后，还要特别感谢我的家人，他们是我事业的最有力支持者，本书要献给我亲爱的儿子Eddie和我所有的家人！

知识无止境，学习无止境！我和我的团队也还在不断地学习。书中的错误和不当之处可能难免，敬请广大读者指正，不胜感谢！

洪松林（Hong Song Lin）

2017年12月26日

CONTENTS

目　录

CHAPTER 1

第1章

机器学习基础

我们从本章开始一直到第4章结束，按照先后顺序，分别讲解机器学习应用的一些基础内容（第1章）、数据探索与准备阶段的工作和方法（第2章）、机器学习应用算法阶段的有关技术（第3章），以及基于前三个内容的机器学习应用的各种案例（第4章）。这四章构成了机器学习应用的一个较完整的讲解，可以算作机器学习应用的基础篇。本章又是这个基础篇的基础，希望能给读者一个机器学习应用的快速引领和轮廓印象。我们从第5章开始直到第7章结束，讲述了机器学习在行业的实践应用，即机器学习在医学领域的一个完整应用，较系统地阐述了医学领域机器学习应用系统的开发与实现（第5章），探讨了医学领域机器学习的实践应用（第6和7章）。最后一章（第8章）介绍了机器学习与人工智能的相关应用。

我们在本书中尽可能将理论与实践相结合，既重于实践应用又深入理论原理。理论是灰色的，而实践则是最鲜活的。对于一本机器学习应用方面的书籍，我们希望尽可能多讲些实践和案例，并多用图表说明大部分的机器学习原理和应用，让读者更能贴近实际。

1.1 认识机器学习

什么是机器学习？不同的人会给出不同的答案，很多人也会给出相似的答案，因为有很多经典的机器学习论著已经给出较为公认的定义。作为常年工作于机器学习应用项目的人来说，我们也有自己的一点认识，可能稍稍不同于理论书籍的概念。本节我们就从这个话题开始，介绍机器学习概念、机器学习与生活和机器学习与知识等内容。

1.1.1 机器学习概念

机器学习与数据挖掘差不多属于同义词。我们认为，数据挖掘应是一个更加广义的概念，甚至可以说不是一个传统意义上的定义，而是一类活动的集合，凡是有目的地探索数据中隐含的规律和知识的活动都可称为数据挖掘。在这里，我们重点强调的要素是：

- 有目的
- 探索性地获取
- 数据中隐含的规律和知识

我们稍后会详细讨论这三个要素。在这个定义中，我们并没有提及应用什么方法和手段获取数据中隐含的规律和知识，也就意味着不限任何方法和手段，无论是数学的还是非数学的，无论是复杂的还是简单的，只要能揭示数据中隐含的规律和知识，都可以被称为数据挖掘。机器学习则应用计算机算法从数据中自动学习新知识和新规律，并将之用于模拟或实现人类的智能行为的方法和技术。机器学习更侧重这样一个表述：像一台机器一样通过自主地、主动地从数据中进行学习而揭示数据中隐含的规律和知识的活动。因此，本书中数据挖掘和机器学习将“平滑”切换。

1.1.1.1 数据的“形状”

从字面而言，数据挖掘包含数据和挖掘两部分，二者同样重要，缺一不可。数据是机器学习的基础素材，我们首先谈一谈数据的形状。

数据的“形状”之一，大数据。经典意义上的机器学习，通常是指对海量数据进行分析。怎么样才算是海量数据？目前还没有明确的标准。而近几年，类似于海量数据，又产生了大数据的提法，其概念无论从内涵和外延上都有了扩展。但从本质上，我们认为，大数据和海量数据是相似的。在实践中，不单单是记录数多的就称为大数据，通常大数据是指数据量和数据维度均很大，数据形式很广泛，如数字、文本、图像、声音等。而大数据往往可能蕴含着丰富的规律和知识，所以在大数据之上应用机器学习就成了理所当然的活动了。

数据的“形状”之二，小数据。相对于大数据，在实践中还会存在不少特殊情况。例如在医学上有些疾病极为少见，只出现几百例，甚至几十例就几乎是该病的总体了，我们称之为小数据。业务中需要对这些小数据进行深入分析和探索，以便挖掘出罕见疾病的特征，并为相应的临床应对提出依据。对于这样规模的数据进行分析，如果按照记录数，依照传统机器学习观念、方法和技术，无法开展探索性的分析工作。我们认为，需求引领观念和技术，机器学习的一个发展分支应该是从规模较小的、有限的数据中探索其中的规律和知识，尽管目前的技术还很有限。

数据的“形状”之三，宽数据。还有一种情况是小数据高维度，小样本大信息，我们称之为宽数据。如某些基因组信息，就是数据量很少，通常只有几十例到几百例，但维度很高，通常有几百个到几千个。更极端情况的是个人大信息，即单个记录下的高维信息，如从宽带、移动支付、物联网、手机等媒介收集个人信息。在不远的将来会出现单独个体的高维数据，并需要关于解决此类机器学习的新理论和新算法。

数据的“形状”之四，深数据。我们还会遇到一种数据，涉及维度不是很宽，但是数据在某几个维度上跨度非常大，历史数据非常多，或者数据量的增长速度非常快，我们称之为深数据。如医学检查中 24 小时心电图监测、较长时段（如一小时以上）的脑电图监测，每小时会产生几十万至几百万条数据；再如，互联网服务商的 DNS 服务器对互联网访问事件的日

志记录，也是每小时会产生几十万至几百万条数据。这类数据，我们有时也称为流数据。对这些深数据的挖掘也是非常具有挑战性的，一方面由于它的数据量非常大，另一方面也由于对这类数据进行挖掘的实时性要求较高。

这些随着数据收集手段的进步而形成的各有特色的数据，正在逐步进入机器学习研究的视野。所以说，这门科学叫做机器学习，它应包括大数据机器学习、小数据机器学习、宽数据机器学习和深数据机器学习。我们需要做的是处理好各类数据来获取知识，研究解决各类型数据的挖掘的新理论和新算法，这些数据的分析算法不完全与经典大数据机器学习相同。例如医学上的个性化精确治疗，就离不开涉及个人的宽数据和深数据。

1.1.1.2　挖掘的思维

机器学习的目的是为了获得知识，很多书中也将机器学习称作 KDD。只要是为了获得知识，用什么工具并不重要，重要的是从数据里面获取知识，工具是从愿望到目的的桥梁，重要的是结果。此外，我们说在机器学习应用中，不是处理方法越复杂就越好，即使是非常简单的方法也可以睿智地理解数据。例如，世界大战中，统计学家沃德在被咨询飞机上什么部位的钢板需要加强时，他画了飞机的轮廓，标出返航战斗机上受敌军创伤的弹孔位置。统计积累一段时间后，机身各部位几乎都被标满了。最后，沃德建议，把剩下少数几个没有弹孔的位置加强，因为被击中这些位置的飞机都没有返航。最后实践验证了沃德对飞机改进的良好效果。

我们认为，不要被机器学习的传统概念限制思维。很多从业者或希望从事机器学习的人，把机器学习这个概念狭义化了。机器学习不是有限的几种工具或算法，例如聚类、分类和预测等，它是一个目的性导向的学科，目的是从数据中获取知识、规则，或其他可直接、间接用以产生效益的信息。广义上的机器学习是和概率统计、高等数学、数学分析、离散数学等数学分支无法清楚分割的，也是和数据库、网络、大数据等技术无法分割的，更是和各行各业的专业知识和业务需求无法分割的。

1.1.1.3　机器学习要素

下面说说我们提到的机器学习概念的要素，先说数据中隐含的规律或知识。我们知道数据是用以描述和反映人类社会中所发生的各种人文活动和事件及自然活动和事件的载体，而大量的人文事件和自然事件中通常蕴含着某些特点和规律。因此，我们利用各种形式的数据（包括数字形式或数据库形式，也包括书籍、图案、声音等形式）将这些活动和事件如实地描述和记录下来，然后应用各种技术手段来研究和挖掘这些数据中所隐含的东西，这些隐含的东西反映了人文或自然活动和事件的本质特征，这些本质特征通常又不是体现在人文或自然活动和事件的表面或较肤浅的层面。我们得到的这些本质特征可能表现为与某些事物相关的一些规律或知识，从而延伸了事物表面所展现的规律或知识，但又绝不类似于事物表面所展现的规律或知识。隐含知识与表象知识通常是完全两码事，隐含知识比表象知识具有更大的价值。所以说，机器学习要的不是事物的表面现象，而是事物所隐藏的东西。反过来说，展

现事物表象的知识不属于机器学习。

再说机器学习是一种探索性的活动。我们认为，由数据所表达的大量事物中通常可能蕴含了一些规律或知识，但谁也不敢保证一定有。另外，挖掘大量数据中所隐含的知识本身，无论从技术上还是从专业上都是一项极富挑战性的工作。因此，我们说机器学习是一种探索性质的活动。探索性质的活动意味着过程可能会很艰辛，结果可能不可预料。所以，如果机器学习的结果达不到我们的预期，一种可能是我们的技术、方法不行，一种可能是数据没有能够真实描绘、反映事物，还有一种可能是事物中没有蕴含着我们想要的东西。但是，由于通常隐含知识比表象知识具有更大的价值，需求引导我们不断地去追求，因此，我们会不停地探索。

最后，机器学习是有目的的活动。机器学习的方向是由业务需求所引领的，知识发现是一项目的性很强的工作。不同的机器学习目的涉及的技术、方法，甚至投入的人力、物力都大所不同，因此，选择恰当的目的使得机器学习工作可控、成本可控。在这里，我们着重讲的是机器学习的商业应用，而不是诸如有关机器学习基础研究、教学实验方面的工作。既然是商业应用，更一定要讲效益、投入产出比。因此，商业机器学习的目的性非常强，机器学习通常分为评估性初探、计划、评估、实施、再评估、部署、维护等过程。如果机器学习目的不明确、缺乏效果评估和风险评估，则项目的失败实在是在所难免。另一方面，即使是机器学习基础研究和教学实验，确定一个科学的工作目标也是必须的。我们说机器学习是有目的的，大家很容易理解，但机器学习本身作为具有探索性的一项工作，其工作目标的制订，尤其在商业项目中决定着项目的成败，因此，我们将其作为一个重点要素在此提出。

1.1.2 机器学习与生活

很多人一提到机器学习，觉得它一定很高深。的确，使用机器学习不是一件很容易的事，尤其经过机器学习应用项目的人感受更深，往往不得不跟一些复杂的算法、繁琐的数据处理打交道。但实际上，机器学习并不都是深奥和难懂的，很多原理就是生活的一部分。现实生活中，很多情况下，人们是在不知不觉中使用了一些机器学习的方法或思想。例如聚类，就是将具有相似特征的事物划分为一类，把特征不同的事物区分开来。实际上，聚类这个动作贯穿了我们生活的很多方面。例如，我们对事物分类的基础和分类的规则，大部分就是通过聚类活动实现的。例如，原始人面临猛兽和猎物，必须会区分这两者，知道自己可以吃掉谁，谁可以吃掉自己（这个是最早的分类之一）。可以说，不懂得这个分类的原始人基本上都灭绝了，如图 1-1 所示。而聚类提供的是什么呢？原始人经过大量的实践（数据的积累），在吃和被吃的两个分类下（那时候还没有家畜和宠物），如图 1-2 所示，对大量事例进行思考、总结、聚类（类内相似最小，类间相异最大），得到规则并实施，然后幸福地进化着。

再例如，关联规则的应用。所谓关联规则，就是在前件事件发生的条件下寻找高概率发生的后件事件。例如，原始人看见了前件——猎物，马上会关联后件——烤肉（原始人学会了使用火以后），于是他们实施的行动就是捕猎与钻木取火，如图 1-3 所示。原始人看见了前件——猛

兽，马上会关联后件——被吃，于是他们实施的行动就是狂奔与逃窜，如图 1-4 所示。

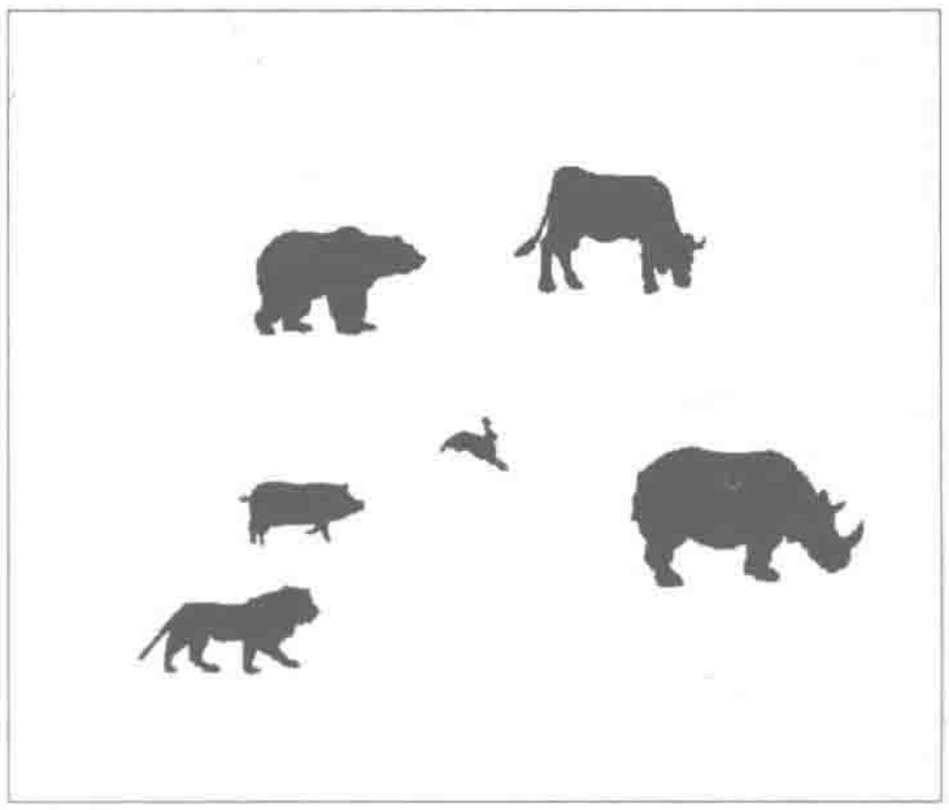


图　1-1

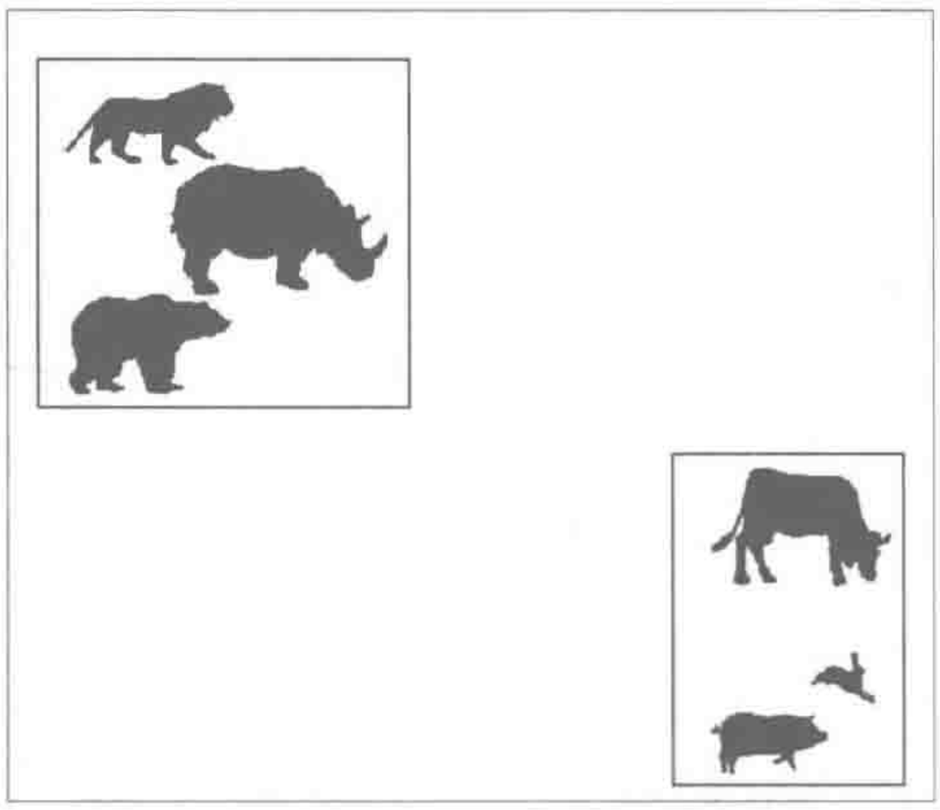


图　1-2

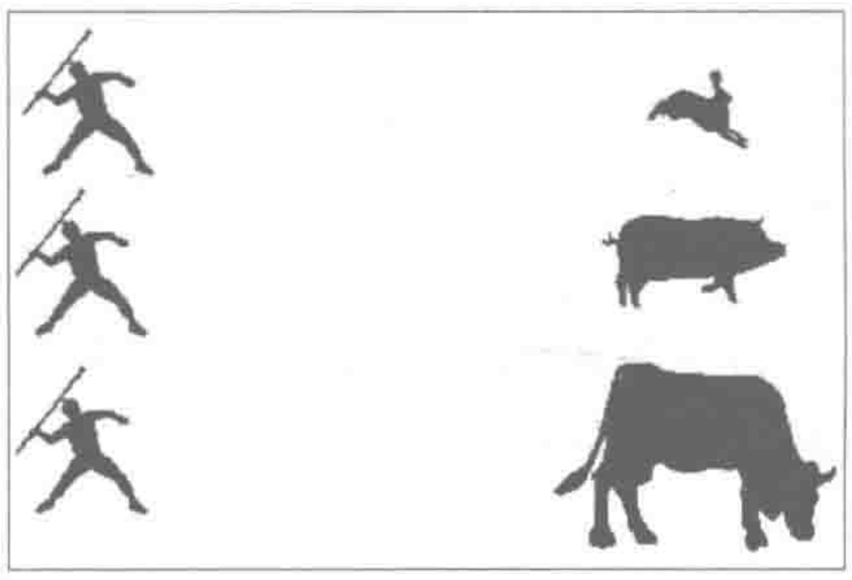


图　1-3

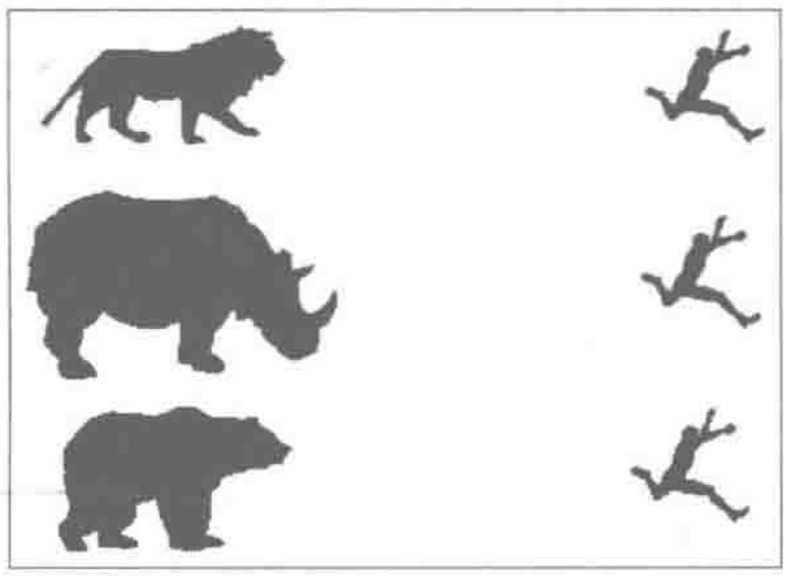


图　1-4

机器学习就是伴随着社会事件或自然活动的大量产生（数据的海量增长）而孕育出来的。所以说，机器学习不是孤立存在的，而是和社会生活及其衍生的需求有机结合的。

机器学习来源于生活，机器学习还要服务于生活。但很多时候，困扰人们的是机器学习复杂的理论、算法和公式。如果使用机器学习解决实践问题，我们希望就像医生不需要明白药品的合成制造原理和过程而能科学地使用药物、手机用户不必搞懂通信协议就能熟练地使用手机的各种功能一样，我们不必学习机器学习的算法理论与开发技术，也能做好我们生活中的机器学习。要实现机器学习的大众化，一个方法就需要将机器学习应用面向各个专业、面向生活的各个方面实现产品化和工具化，尽量使技术细节对使用者透明化，而使用户专注于机器学习技术在行业领域、生活领域的应用。

一般来说，机器学习的应用者需要知道机器学习各个算法的基本目的、基本原理。这里说的基本原理是高层面原理，而不是具体技术性的原理。例如聚类，我们只要知道它的原理是“物以类聚，人以群分”就足够了。机器学习技术的一个方向就是让每个人都可以方便地使用，从而享受新技术带来的好处。这样机器学习才可以真正融入生活的各个方面，成为一门成功的科学技术。

1.1.3 机器学习与知识

当然，机器学习现在还远没有走入人们的生活，大多情况下是应用于一些商务场合，也就是机器学习是以商业项目形式存在。根据经验，我们把机器学习项目分成几个不同的层面，如决策层面、设计层面、技术层面和应用层面，不同的层面需要不同的知识结构：

- 决策层面需要知道机器学习能干什么，能为本单位带来什么效益；
- 设计层面需要行业领域相关知识和机器学习技术相关的知识；
- 技术层面需要高等数学、概率学、统计学、数据库原理、分布计算、编程语言等知识，还需要掌握具体算法的原理；
- 应用层面需要知道机器学习怎样结合行业领域的需求，怎样应用机器学习的结果解决业务问题。

所以，机器学习的应用所需要的知识与人们在项目的职责与角色密切相关。使用机器学习需要从高等数学学起吗？还是懂得原理就可以了？决策者在行业目的的选择上需要懂得多深的机器学习知识？不同的目的，需要掌握的知识不一样。即使都是技术人员，不同的层面要求也不一样。机器学习应用技术人员需要掌握调节算法、算法的适用性和结果的合适表现形式；机器学习研究者需要探索新的理论，创新、改进新的算法的知识和能力。

要把机器学习应用讲得很明白，其实需要各个方面的很多知识。我们既要讲解机器学习原理，还要讨论机器学习具体的技术、行业知识、机器学习具体技术与行业知识的结合、机器学习结果在行业领域的含义和使用，等等。机器学习原理和理论并不是同一个事物，原理可以不借助各种公式而存在，并且可能相对简单。

机器学习应用有自身的特色，就是我们下面需要重点讲解的。而目前，对机器学习结果的分析还无法实现全自动化，现在更多的只是把现象（数据知识）而不是深入专业的知识（行业知识）展现给用户。这就需要我们应用者掌握深入的专业知识和技术知识，把机器学习自主发现的规律总结成专业知识，并应用于工作实践。

1.2 机器学习应用基础

机器学习是一种获得知识的技术，它的基础是数据，手段是各种算法，目的是获得数据中蕴含的知识。发现知识并非易事，人们总是受到各种各样的局限。在信息时代之前，数据的缺乏成为发现知识的限制。在天文学的早期，能有所发现不仅需要有敏锐的头脑和先进的观测设备，前人留下的大量观测数据也是必不可少的条件。实际上，就算是目前，数据的缺乏仍然是一个发现知识的主要瓶颈。随着数据采集和存储技术的发展，对大量数据的分析和使用成为了一个新的难题。机器学习是一门处理大数据的应用科学，它是随着对大数据分析处理的需要而诞生的新兴学科，出现至今只有短短二、三十年的时间，各方面还在不断发展和完善之中。对机器学习应用而言，知识的发现存在两个极限，一个是前面提到的数据极限，即要么数据非常庞大，要么数据的量小但维度非常大；另一个是算法极限，即我们针对很多