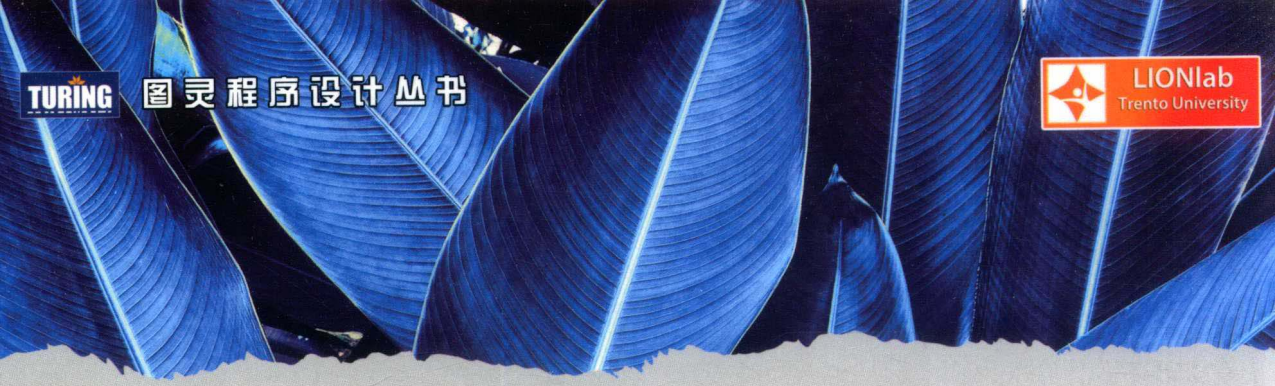


The LION Way

Machine Learning plus Intelligent Optimization

机器学习与优化

[意] 罗伯托·巴蒂蒂　毛罗·布鲁纳托　著
王彧弋　译

- 摒弃复杂的公式推导，从实践上手机器学习
- 人工智能领域先驱、IEEE会士巴蒂蒂教授领导的LION实验室多年机器学习经验总结

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

The LION Way

Machine Learning plus Intelligent Optimization

机器学习与优化

[意] 罗伯托·巴蒂蒂　毛罗·布鲁纳托　著
王彧弋　译

人民邮电出版社
北　京

图书在版编目(CIP)数据

机器学习与优化/（意）罗伯托·巴蒂蒂（Roberto Battiti），（意）毛罗·布鲁纳托（Mauro Brunato）著；王彧弋译. —北京：人民邮电出版社, 2018. 5
（图灵程序设计丛书）
ISBN 978-7-115-48029-3

Ⅰ. ①机… Ⅱ. ①罗… ②毛… ③王… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字（2018）第 044097 号

内 容 提 要

本书是机器学习实战领域的一本佳作，从机器学习的基本概念讲起，旨在将初学者引入机器学习的大门，并走上实践的道路。本书通过讲解机器学习中的监督学习和无监督学习，并结合特征选择和排序、聚类方法、文本和网页挖掘等热点问题，论证了“优化是力量之源”这一观点，为机器学习在企业中的应用提供了切实可行的操作建议。

本书适合从事机器学习领域工作的相关人员，以及任何对机器学习感兴趣的读者。

◆ 著 〔意〕罗伯托·巴蒂蒂 毛罗·布鲁纳托
译 王彧弋
责任编辑 朱 巍
执行编辑 温 雪 黄志斌
责任印制 周昇亮
◆ 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
三河市君旺印务有限公司印刷
◆ 开本：800×1000 1/16
印张：17.5 彩插：2
字数：420 千字 2018 年 5 月第 1 版
印数：1-3 500 册 2018 年 5 月河北第 1 次印刷
著作权合同登记号 图字：01-2014-4553 号

定价：89.00 元

读者服务热线：(010)51095186 转 600 印装质量热线：(010) 81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

版 权 声 明

目　录

第1章 引 言

人不应该过着野兽般的生活，而是要追寻美德与知识。

—— 但丁

1.1 学习与智能优化：燎原之火

优化是指为了找到更好的解决方案而进行的自动化搜寻过程。可以说，流程、方案、产品和服务之所以能持续改进，正是缘于优化为之**提供的强大动力**。优化不仅关乎方案的确定（从一些给定的可行方案中，选出最好的一个），**它还能主动创造出新的解决方案**。

优化催生了**自动化的创造和革新**。这看起来非常矛盾，因为自动化通常不会和创造与革新联系起来。因此，那些相信机器只能用来处理单调的重复性工作的人们在阅读本书时，会觉得书中的观点简直是胡言乱语，甚至会感受到如同被挑衅一般的愤怒。

自伽利略（1564—1642）之后，人们希望用科学改变世界，而这不仅需要哲学上的阐释，还需要**测量和实验**的支持。“测量那些可测量的，并使那些不可测量的变得可测量。”测量一开始看起来并不起眼，但它允许人们用务实的方式逐渐改变世界，只要人们还关心生产方式和生活质量。

几乎所有的商业问题都可以归结为**寻找一个最优决策值** x，这要通过**使某个收益函数** goodness(x) **最大化**来实现。为了能形象地理解，我们假设有一个集合变量 $\boldsymbol{x} = (x_1, \cdots, x_n)$，

它描述的可以是一个或多个待调节的旋钮，也可以是将要做出的选择，还可以是待确定的参数。在市场营销中，$\boldsymbol{x}$ 可以是一个向量，其数值表示为各类宣传活动（电视、报纸、各种网站、社交媒体）分配的预算，goodness($\boldsymbol{x}$) 则可以是由这些宣传活动而产生的新客户数量。在网站优化中，$\boldsymbol{x}$ 可以涉及图片、链接、话题和不同大小文本的使用，goodness($\boldsymbol{x}$) 则可以是该网站的普通访客成为客户的转化率。在工程学中，$\boldsymbol{x}$ 可以是一个汽车发动机的设计参数集，goodness($\boldsymbol{x}$) 则可以是该发动机每加仑汽油所能行驶的英里数。

将问题归结为“优化一个收益函数”也**激励着决策者，使用量化的目标，就可以用可衡量的方式来领会宗旨，也就可以专注于方针的制定而非执行的细枝末节**。当人们深陷于执行的泥潭中，以至于遗忘了目标时，企业就染上了“疫病”，此时如果外界环境发生了变化，这种“疫病”将会使企业无法做出及时的应对。

自动化是解决这个问题的关键：将一个问题形式化地表述后，我们把得到的收益模型输入计算机，计算机将自动创造出并找到一个或多个最佳的选项。另外，当条件和重点发生改变时，只需要修改一下收益函数的量化目标，再重启优化过程就可以了。当然，CPU 时间会是个问题，也并非每次都能保证找到全局最优解决方案。但可以肯定的是，使用计算机来搜寻，无论是速度还是范围，都远远领先于人力搜寻，并且这一领先优势会越来越明显。

然而，在大多数现实场景中，优化的惊人力量仍遭到很大程度的压制。优化在现实中没有被广泛采纳的主要原因是，**标准的数学优化理论假设存在一个需要最大化的收益函数**，也就是说，有一个明确定义的模型 goodness($\boldsymbol{x}$) 为每个输入配置 $\boldsymbol{x}$ 匹配一个结果。而目前，在现实的商业情境里，这个函数通常是不存在的。即使存在，靠人力找到这个函数也是极其困难、极其昂贵的。试想，问一个 CEO“请您告诉我，优化您业务的数学公式是什么”，显然不是咨询工作中开始对话的最佳方式。当然，一个经理对于目标应该会有一些想法和权衡，但是这些目标并没有以数学模型的方式给定，它们是动态的、模糊的，会随着时间改变，并且受限于估计误差和人们的学习进程。直觉被用来替代那些明确给定的、量化的和数据驱动的决策过程。

如果优化是燃料，那么点燃这些燃料的火柴就是**机器学习**。机器学习通过摒弃那种明确定义的目标 goodness($\boldsymbol{x}$) 来拯救优化：我们**可以通过丰富的数据来建立模型**。

机器学习与智能优化（learning and intelligent optimization，LION）结合了学习和优化，它从数据中学习，又将优化用于解决复杂的、动态的问题。LION 方法提高了自动化水平，并将数据与决策、行动直接联系起来。描述性分析和预测性分析之后，LION 的第三阶段（也是最终阶段）是**规范性分析**（prescriptive analysis）。**在自助服务的方式中，决策者手中直接握有更多的权力**，而不必求助于中间层的数据科学家。就像汽车的发动机一样，LION 包含一系列复杂的机制，但是用户（司机）并不需要知道发动机的内部工作原理，就可以享用它带来的巨大好处。在未来的几十年内，LION 方法带来的创新，将会像野火那样，以燎原之势延伸到大多数行业。那么企业就像野火频发的生态系统中的植物一样，只有适应并拥抱 LION 技术才能生存下来，并繁荣昌盛；否则，无论之前如何兴盛，在竞争逐渐加剧的挑战面前，都可能

土崩瓦解。

LION 范式关注的并不是数学上的收益模型，而是海量数据，以及如何针对多种具体选择（包括实际的成功案例）进行专家决策，或者如何交互地定义成功的标准。当然，这些都是建立在让人们感觉轻松愉快的基础之上的。例如，在市场营销中，相关数据可以描述之前的资金分配和宣传活动的成效；在工程学中，数据可以描述发动机设计的实验（真实的或模拟的）和相应的油耗测量方式。

1.2 寻找黄金和寻找伴侣

用于优化的机器学习需要数据。数据来源可以是**以往的优化过程**，也可以是**决策者的反馈**。

要了解这两种情境，先来看两个具体的例子。丹尼尔·克里金（Danie G. Krige，见图 1-1）是一名南非的采矿工程师，他曾遇到一个问题：如何在一张地图上找到挖掘金矿的最佳坐标 [74]。大约在 1951 年，他开创性地将统计学的思想应用于新金矿的估值，而这一方法仅需用到有限的几个矿坑。需要优化的函数是 $\mathrm{Gold}(\boldsymbol{x})$，即坐标 $\boldsymbol{x}$ 处的金矿的金量。当然，在一个新的地方 $\boldsymbol{x}$ 评估 $\mathrm{Gold}(\boldsymbol{x})$ 是非常昂贵的。你可以想象，挖一个新矿没那么快，也没那么简单。但是在一些试探性的挖掘之后，工程师们会积累一些把坐标 $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \cdots$ 和金量 $\mathrm{Gold}(\boldsymbol{x}_1), \mathrm{Gold}(\boldsymbol{x}_2), \mathrm{Gold}(\boldsymbol{x}_3)$ 关联起来的**实例知识**。克里金的直觉告诉他，用这些实例（**来自以往优化过程的数据**）可以建立起函数 $\mathrm{Gold}(\boldsymbol{x})$ 的模型。这个称为 $\mathrm{GoldModel}(\boldsymbol{x})$ 的模型归纳以往的实验结果，为地图上的每个位置 $\boldsymbol{x}$ 给出金量的估计值。通过优化，这个模型找到使预计黄金产量 $\mathrm{GoldModel}(\boldsymbol{x})$ 最大化的地点 $\boldsymbol{x}_{\mathrm{best}}$，于是这个 $\boldsymbol{x}_{\mathrm{best}}$ 成为下一个挖掘的地点。

图 1-1 丹尼尔·克里金，克里金法的发明者

可以用如图 1-2 所示的模型来形象地说明这个过程。先在地图上为每个矿坑插一根针，每根针的高度取决于在该处发现的金量。克里金的模型可以看作基于这些针的“训练”信息

在整个地图上方生成的一个曲面，使得给定位置的高度对应当地的预计黄金产量。因此，优化意味着在这个模型曲面上找到最高的那个点，并在对应的地点进行下一次挖掘。

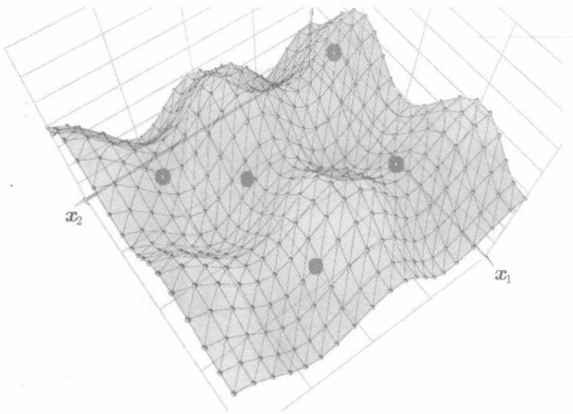


图 1-2 从样本中使用克里金法构造模型。一些样本在图中用点标示出来。表面的高度和颜色依赖于产金量（另见彩插）

这种技术现在被称为克里金法（Kriging），它背后的理念是未知点对应的值应该是其邻近已知点所对应的值的加权平均，权重与这些已知点到该未知点的距离相关。*高斯过程*、*贝叶斯推断*和*样条函数*（spline）都涉及了相关的建模方法。

第二个例子关于**决策者的反馈**。想象有这样一个约会服务：人们付费在数以百万计的候选人中匹配一个最佳的约会对象。在克里金法中，需要优化的函数是存在的，只是评估起来极为困难。对于这个案例，我们很难假设存在一个类似的函数 IdealMate($\boldsymbol{x}$)，它将个人特征 $\boldsymbol{x}$，例如美貌、智力等，与你的个人喜好联系起来。如果你不这么认为，且坚信存在这样一个函数，那么给你留一个作业，尝试用准确的数学术语来定义你心目中理想伴侣的 IdealMate 函数。即使你能准确地指出某些组成部分，例如 Beauty($\boldsymbol{x}$) 和 Intelligence($\boldsymbol{x}$)，但是在开始寻找最佳候选人之前，把这两个目标合并起来仍然是困难的。像“降低多少 IQ 值对应减少一点美貌”或者“美貌是否比智力重要，重要多少”这类问题是非常难回答的。假使你很痛苦地给出了一个初步答案，也肯定不会相信这个优化，在真正见到这个候选人之前，你不会为这个匹配服务付费，当然也不会对服务感到满意。你会想了解这个人的特征，而不仅仅是得到系统优化的肤浅的 IdealMate($\boldsymbol{x}$) 函数值。只有在考虑过不同的候选人并且对这个匹配服务进行反馈后，你才能希望找到最满意的另一半。

换句话说，在一开始，待优化函数中的某些信息是不全面的，只有决策者才能够调整优化的过程。**许多现实问题，即使不是大多数，都需要借助有学习参与的迭代过程来解决**。在了解了越来越多的案例后，用户会认识并调节自己的喜好，系统会从用户的反馈中建立起他的喜好模型。这一过程将持续下去，直到用户满意或者直到耗尽为这一决策分配的时间。

1.3 需要的只是数据

下面继续谈论商业用户的动机。如果你不关心这方面的内容，可以放心地跳过这部分，直接阅读 1.6 节。

商业领域里充斥着各种数字形式的数据。大数据指的是大量的半结构数据。顺便提一句，在 20 世纪七八十年代，数据对于当时的存储设备来说是庞大的，而如今的“大数据”更多是商业上的宣传概念：即便是最大的公司产生的所有数据，只需一台 PC 就足以处理了。

随着社交网络的爆发、电子商务的迅速扩张和物联网的兴起，网络正在掀起一场由结构化和非结构化数据引起的海啸。这场海啸驱使人们在信息技术领域花费多达数十亿美元。也有新的证据表明，标准的商业智能平台使用率正在下降，这是因为企业界已经不得不开始考虑一些非结构化的数据，而这些数据拥有无法估量的现实价值。例如，社交网络产生大量的数据，其中的大多数无法分类，也无法用传统数据的刚性层次结构来表示。试想，你该如何评估 Facebook 上一个“赞”的价值？况且非结构化数据需要用自适应方法来分析。再想想，随着时间的流逝，一个“赞”的价值会发生怎样的变化？由于这类问题的存在，我们需要在数据建模、自适应学习和优化等领域运用更加先进的技术。

为了让软件能够自我改进，并能快速适应新数据和调整后的业务目标，需要使用 LION 方法。这种方法的优势在于能够从过往的经验中学习、**在工作中学习**、应对不完全的信息，并快速适应新的情况，而这些能力通常只与人类的大脑联系起来。

LION 技术这种内在的灵活性是至关重要的，因为在求解过程开始之前，我们很可能无法确定哪些是对决策有影响的因素和重点。例如，我们要给一个市场营销的前景评分来估计其价值，应该考虑哪些因素？这些因素又对结果分别有多大程度的影响？如果使用 LION 方法的话，这些问题的答案就是：“这些都不是问题。”系统会开始自我训练，源源不断的数据加上终端用户的反馈将快速提升系统的性能。专家——这里指营销经理——可以通过表达他们自己的观点来改善系统的输出。

1.4 超越传统的商业智能

每一家企业都需要数据来满足 3 项基本需求：

(1) **了解**目前的业务流程，并评估以往的表现；

(2) **预测**商业决策的影响；

(3) 对业务的关键因素制定并执行明智且合理的决定，从而**提升**赢利能力。

传统的**描述型**商业智能（business intelligence，BI）擅于记录和可视化过往的表现。构建这样的记录意味着需要聘请顶级顾问，或雇用那些有统计、分析和数据库等领域知识的专业人员。专家必须要设计数据提取和操作的流程，然后交给程序员来实际执行。这是一个缓慢而繁琐的过程，毕竟大多数商业的境况都是瞬息万变的。

因此，那些严重依赖于 BI 的企业正在利用性能快照，尝试理解当前情况和未来趋势，并对此做出反应。这就如同开车的时候只盯着后视镜，很有可能会撞上什么东西。现在对于企业来说，就像是已经撞到了一堵僵化的墙，并且缺乏快速适应变化的能力。

预测分析确实在预见方案效果方面做得更出色，然而，**将数据驱动模型和优化进行整合**，自动创建完善的解决方案，才是 LION 真正的强大之处。**规范性分析**做到了引领我们直接从数据到最佳改进方案，**以及从数据到可执行的洞察力，再到行动本身**!

1.5 LION 方法的实施

对于处在不同业务状态的企业而言，全面采用 LION 方法作为商业实践的步骤会有所不同。更重要的是，相关数据的情况也会影响这一进程。显然，在数据收集完成的时候引进 LION 范式会相对容易，开销也更少。对某些企业来说，由于遗留系统的迁移和转换需要涉及大范围的整理，开销会非常大。这也正是那些老练的服务提供商能大显身手的地方。

除了整理和定义相关数据的结构之外，最重要的一点就是建立起数据分析团队和商业终端用户之间的合作。LION 方法通过自身的特性提供了一种合作方式，助其共同发现蕴藏在结构化或半结构化数据中的潜能。数据分析团队能够和商业终端用户高效地并肩合作，关键在于能够使业务目标的不断变化迅速反映到模型上。LION 方法的引入可以帮助数据分析团队在价值创造链中产生根本性的变化，它能揭示隐藏的商机，也能加快他们的商业伙伴对客户要求和市场变化的反应速度。

就业市场也将被打乱。从人类的实例中进行学习的软件将推导出我们在使用却又不明确了解的规则。这将消除进一步自动化的障碍，在许多需要适应性、常识和创造性的任务中，机器将会代替工人，也许会让中产阶级处在风险之中 [110]。

LION 方法可以说是一种极具颠覆性的**发现隐藏价值的智能方法，它能快速适应改变并改进业务**。通过恰当的规划和实施，LION 技术可以帮助企业在竞争中独领风骚，避免被燎原之火灼伤，同时也可以帮助个人在高技能人才的就业市场中保持竞争力。

1.6 “动手”的方法

因为这是一本关于从实例中进行（机器）学习的书，所以在学习这本书时也要遵从这一点。本书大多数的内容都是按照从实例中学习和**从实践中学习**的原则来安排的。当介绍不同的技术时，我们会讨论这些技术的基础理论，然后会总结出一些你“应该了解的梗概”。本书鼓励用现实中的情况来做实验，你可以在本书的网站上找到相关的例子和软件。这样做能让你体会到 LION 技术并不是只为专家准备的；它属于任何对快速且可测量的结果感兴趣的实践者。

第一次阅读本书时你可以跳过某些理论部分。但是某些理论知识是十分关键的，它们不仅能帮助开发新的、更加先进的 LION 技术，还能使你更加得心应手地使用这些技术。掌握一些基础的、未被稀释的理论，就像在陌生国度旅行时手中有地图指引。如果你是一艘不知要驶向何处的船，那么风往哪边吹都是无意义的。

我们会尽量兼顾开发人员和终端用户的感受。下面两个图标粗略地表示了不同章节的难度级别。当然，难易程度的真实感受跟读者的知识背景有关，因此可能与我们试验性的级别分类不同。

容易的话题

进阶的话题

本书作者以及读者群发布的数据、指导说明和教学短片都可以在本书的网站上找到：https://intelligent-optimization.org/LIONbook/。

我们感谢为这本书做出了贡献的人们。首先是照片和插画。Carlo Nicolini 提供了在 LION-4@VENICE 2010 会议期间拍摄的威尼斯照片。第 1 章的但丁像是 Domenico di Michelino 于 1465 年在佛罗伦萨完成的。George Chernilevsky 提供了第 2 章装着蘑菇的篮子的图片。第 9 章大脑图片是达·芬奇（1452—1519）的作品。聚类深度网络的图来自 Geoffrey Hinton。第 11 章的 Vapnik 教授的照片由 Yann LeCun 提供。超限学习机的图片来自 Guangbin Huang。储备池的结构图由 Herbert Jaeger 提供。第 13 章的威尼斯绘画由卡纳莱托在 1730 年完成。第 15 章的绘画是米开朗基罗于 1541 年完成的。我们也在维基百科中找到了一些解释性的图片。Hopfield 网络图来自 Gorayni，能级相图由 Mrazvan22 提供。本书作者和作者的儿子们都是维基百科条目积极的撰写者。第 14 章章首 Reschense 湖的照片来自 Markus Bernet。第 10 章的蟾蜍图片由 André Karwath 提供。

最后，我们感谢读者为提升这本书的品质所做的越来越多的贡献。他们包括 Patrizia Nardon、Fred Glover、Alberto Todeschini、Yaser Abu-Mostafa、Marco Dallariva、Enrico Sartori、Danilo Tomasoni、Nick Maravich、Drake Pruitt、Dinara Mukhlisullina、Rohit Jain、Jon Lehto、George Hart、Markus Dreyer、Yuyi Wang 和 Gianluca Bortoli。书中的漫画是 Marco

Dianti 赠予我们的礼物。我们十分乐意与读者沟通。如果你有评论、建议或者勘误[①]，请给我们发电子邮件，我们会把你的名字加在下一个版本中。你可以在 LIONlab 的网站上找到联系方式和电子邮件地址：https://intelligent-optimization.org/。

第 2 版补遗

现在你正在读的是本书的第 2 版：我们在此感谢许多读者发送的更正和改进建议。

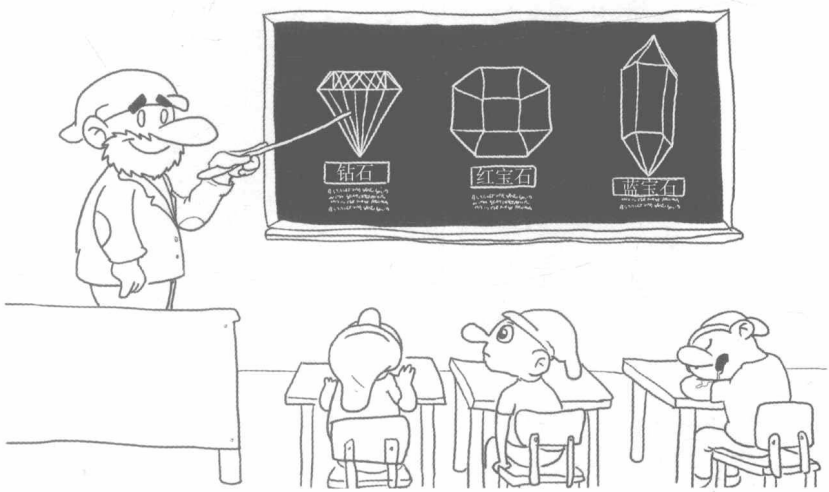
电子书

扫描如下二维码，即可购买本书电子版。

① 中文版勘误请读者到图灵社区的本书页面提交：http://www.ituring.com.cn/book/1413。—— 编者注

第 2 章　懒惰学习：最近邻方法

自然不允许跳跃。



如果你还记得小时候是如何识字的，那么你就可以理解什么是**从实例中学习**，尤其是监督学习。父母和老师给你展示一些带有英文字母（a、b、c，等等）的实例，然后告诉你：这是 a，这是 b，……

当然，他们并没有用数学公式或者精确的规律来描述这些字母的几何形状。他们只是展示了一些不同风格、不同形式、不同大小和不同颜色的**已标记的实例**。经过一些努力和失误之后，你的大脑就能够正确识别这些实例了。然而这不是关键，因为仅凭记忆你其实就能够做到这一点。重要的是，通过这些实例的训练，你的大脑还能从中提取出与认字真正相关的模式和规律，过滤掉不相关的“噪声”（比如颜色），从而进行**泛化**（generalize），以**识别**在训练阶段从未见过的**新实例**。这是很自然的结果，但确实是值得注意的成果。取得这一成果不需要什么先进的理论，也不需要博士学位。如果有一种方法也能如此自然而又轻松地解决商业问题，是不是很令人振奋呢？结合了从数据中学习和优化的 LION 范式就是这样的一种方法，我们将从这一熟悉的语境开始。

在**监督学习**中，由监督者（老师）给出一些**已标记的实例**，系统根据这些已标记的实例来完成训练。每一个实例是一个数列，它包括一个作为输入参数的向量 $\boldsymbol{x}$，称为**特征**（feature），和与之相对应的输出标记 y。

本书作者生活的地方有很多的山地和森林，因此采蘑菇是一项十分普及的消遣活动。虽然采蘑菇很受欢迎也很有趣，但是误食有毒的蘑菇将造成致命的危害（见图 2-1）。这里的小孩子在很小的时候就学会了如何区分可以食用的和有毒的蘑菇。到这里来的游客可以买到相关的书籍，书中有这两类蘑菇的图片和特征；他们也可以把采到的蘑菇带到当地的警察局，让专家帮他们免费检验这些蘑菇。

图 2-1　采蘑菇要区分可以食用的和有毒的

这里有一个被简化过的例子，如图 2-2 所示，假设我们用两个参数，比如高度和宽度，就能够区分这两种蘑菇。当然，一般来说，我们需要考虑更多的输入参数，像颜色、形状、气味等，甚至是更加令人困惑的正类（可以食用的）和负类实例的概率分布。

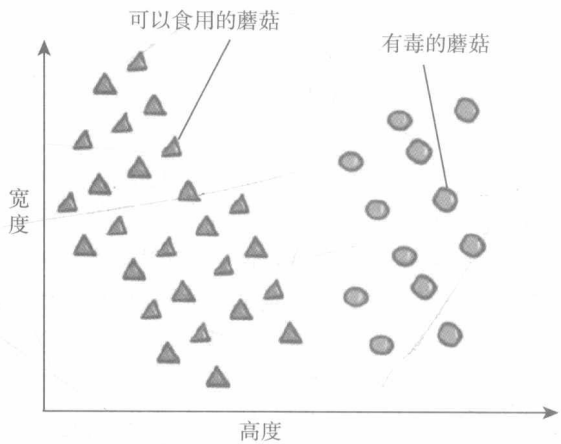


图 2-2　简化的例子：两个特征（宽度和高度）用以区分可以食用的和有毒的蘑菇

那些懒惰的初学者在采蘑菇的时候遵循简单的模式。通常他们在采摘蘑菇之前没有学习任何相关的知识，毕竟，他们到特伦蒂诺是来度假的，而不是来工作的。当发现一个蘑菇时，