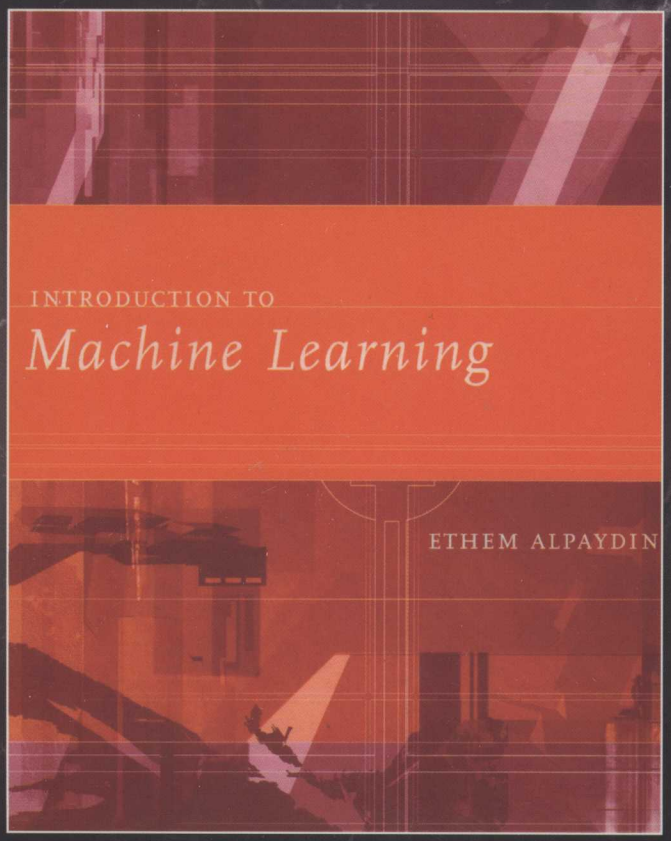
HZ BOOKS
华章教育

计 算 机 科 学 丛 书

# 机器学习导论

（土耳其）Ethem Alpaydın 著　范明 昝红英 牛常勇 译



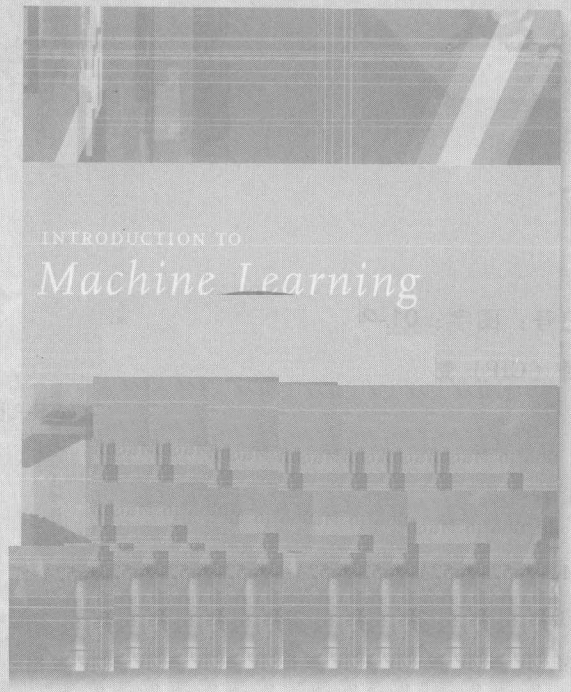

Introduction to Machine Learning

机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

# 机器学习导论

（土耳其）Ethem Alpaydın 著 范明 昝红英 牛常勇 译




**Introduction to Machine Learning**

机械工业出版社
China Machine Press

本书讨论了机器学习在统计学、模式识别、神经网络、人工智能、信号处理等不同领域的应用。其中涵盖的内容比较全面，且易于学习和掌握。主要内容包括：监督学习、贝叶斯决策理论、参数方法、多元方法、维度归约、非参数方法、决策树、线性判别式、多层感知器、隐马尔可夫模型、组合多学习器以及增强学习等。可作为高等院校计算机相关专业高年级本科生和研究生的教材，也可供研究机器学习方法的技术人员参考。

Ethem Alpaydın: *Introduction to Machine Learning* (ISBN 0-262-01211-1).



**本书版权登记号：图字：01-2006-5437**

**图书在版编目（CIP）数据**

机器学习导论/（土）阿培丁（Alpaydın, E.）著；范明等译. —北京：机械工业出版社，2009.6

（计算机科学丛书）

书名原文：Introduction to Machine Learning

ISBN　978-7-111-26524-5

Ⅰ.机…　Ⅱ.①阿…②范…　Ⅲ.机器学习　Ⅳ.TP181

中国版本图书馆 CIP 数据核字（2009）第 031371 号

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：李俊竹

北京慧美印刷有限公司印刷

2009 年 6 月第 1 版第 1 次印刷

184mm × 260mm · 18 印张

标准书号：ISBN 978-7-111-26524-5

定价：39.00 元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

# 出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章分社较早意识到“出版要为教育服务”。自1998年开始，华章分社就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章分社欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com
电子邮件：hzjsj@hzbook.com
联系电话：（010）88379604
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

华章教育

# 译者序

自从有计算机以来，人们就希望计算机能够学习。然而，机器学习真正取得实质性进展，能够成功地解决一些实际问题，并最终成为一个学科分支还是近20余年的事。

对于许多问题，我们的前人和先行者已经知道如何求解。例如，欧几里德告诉我们可以用辗转相除法求两个整数的最大公约数；Dijkstra告诉我们如何有效地求两点之间的最短路径；Hoare向我们展示了怎样将杂乱无章的对象快速排序……对于这些问题，我们清楚地知道求解步骤。因此，让计算机求解这些问题只需要设计算法和数据结构、进行编程，而不需要让计算机学习。

还有一些事情，人们可以轻而易举地做好，但是却无法解释清楚我们是如何做的。例如，尽管桌子千差万别、用途各异，但是我们一眼就能看出某个物体是否是桌子；尽管不同的人的手写阿拉伯数字大小不一、笔画粗细不同，但是我们还是可以轻易识别一个数字是不是8；尽管声音时大时小、有时可能还有点沙哑，但是我们还是可以不费力气地听出熟人的声音。诸如此类的例子不胜枚举。对于这些问题，我们不知道求解步骤。因此，让计算机来做这些事就需要让计算机学习。

我们知道桌子不是木材和各种材料的随机堆砌，手写数字不是像素的随机分布，熟人的声音也不是各种声波的随机混合。现实世界总是有规律的。机器学习正是从已知实例中自动发现规律，建立对未知实例的预测模型；根据经验不断提高，不断改进预测性能。

这是关于机器学习这一主题全面论述的教科书，适合作为高等院校计算机相关专业高年级本科生和研究生机器学习入门课程的教材。该书涵盖了监督学习、贝叶斯决策理论、参数方法、多元方法、维度归约、聚类、非参数方法、决策树、线性判别式、多层感知器、局部模型、隐马尔可夫模型、分类算法评估和比较、组合多学习器以及增强学习。作者对来自统计学、模式识别、神经网络、人工智能、信号处理、控制和数据挖掘等不同领域的机器学习问题和学习方法进行了统一论述。

现在，学习的本质还不十分清楚。然而，关于学习的理论认识已开始逐步形成，已经建立起来的一些机器学习方法已经成功地解决了许多实际问题。我们能够从这本书中学习机器学习，发现机器学习的新方法，不断提高对学习本质的认识。

全书共分16章和一个附录。昝红英翻译了第1～6章，牛常勇翻译了第13～16章，范明翻译了其余部分，并对全书译文进行了修改和最后定稿。

本书的翻译得到了原作者Ethem Alpaydın教授的支持。Ethem教授不仅为中文版写序，而且还耐心地解释了我们的一些疑问。在此，我们向Ethem教授表示感谢。

译文中的错误和不当之处，敬请读者朋友指正。意见和建议请发至mfan@zzu.edu.cn，我们不胜感激。

译　者

2009年春于郑州大学

# 致谢

获得好想法的途径是与有才干的人一起工作，与他们一起工作也是一种乐趣。Boğaziçi 大学计算机工程系是一个极好的工作场所，在我写这本书时，我的同事们为我提供了我所需要的所有支持。我也要感谢我过去和现在的学生，在他们身上，我实际检验了现在写进这本书中的内容。

在写本书时，我得到了土耳其科学院青年科学家奖励计划的资助（EATÜBA-GEBİP/2001-1-1）。

我特别感谢 Michael Jordan。对于他多年来的支持和最近对本书的支持，我深表感谢。他针对本书大体组织和第 1 章所给出的建议在内容和形式上都大大改进了本书。Taner Bilgiç、Vladimir Cherkassky、Tom Dietterich、Fikret Gürgen、Olcay Taner Yıldız 和 MIT 出版社的未留名审稿人也部分阅读了本书，并提供了非常宝贵的反馈。我希望他们在注意到我采纳了他们的建议但却没有特别致谢时，能够体会到我的感激之情。当然，书中的错误和不足应当由我个人负责。

我的父母信任我，我感谢他们永恒的爱和支持。无论我何时需要，Sema Oktuğ 总在身边，我将永远感激她的友谊。我还要感谢 Hakan Ünlü，在过去的几年中，我们无数次讨论了涉及生活、宇宙和万事万物的众多主题。

本书使用 Chris Manning 准备的 LATEX 宏排版，对此我很感谢他。我要感谢 MIT 出版社的编辑们，以及 Bob Prior、Valerie Geary、Kathleen Caruso、Sharon Deacon Warne、Erica Schultz 和 Emily Gutheinz，感谢他们在本书完成期间的不断支持和帮助。

# 符号表

| | |
|---|---|
| $x$ | 标量值 |
| $\boldsymbol{x}$ | 向量 |
| $\mathbf{X}$ | 矩阵 |
| $\boldsymbol{x}^T$ | 转置 |
| $\mathbf{X}^{-1}$ | 逆矩阵 |
| | |
| $X$ | 随机变量 |
| $P(X)$ | 概率质量函数，$X$ 是离散的 |
| $p(X)$ | 概率质量函数，$X$ 是连续的 |
| $P(X \mid Y)$ | 给定 $Y$，$X$ 的条件概率 |
| $E[X]$ | 随机变量 $X$ 的期望值 |
| $\text{Var}(X)$ | $X$ 的方差 |
| $\text{Cov}(X, Y)$ | $X$ 和 $Y$ 的协方差 |
| $\text{Corr}(X, Y)$ | $X$ 和 $Y$ 的相关性 |
| | |
| $\mu$ | 均值 |
| $\sigma^2$ | 方差 |
| $\Sigma$ | 协方差矩阵 |
| $m$ | 均值的估计 |
| $s^2$ | 方差的估计 |
| $\mathbf{S}$ | 协方差矩阵的估计 |
| $\mathcal{N}(\mu, \sigma^2)$ | 一元正态分布，均值为 $\mu$，方差为 $\sigma^2$ |
| $\mathcal{Z}$ | 单位正态分布：$\mathcal{N}(0, 1)$ |
| $\mathcal{N}_d(\mu, \Sigma)$ | $d$-变量正态分布，均值向量为 $\mu$，协方差矩阵为 $\Sigma$ |
| | |
| $x$ | 输入 |
| $d$ | 输入数：输入的维度 |
| $y$ | 输出 |
| $r$ | 要求的输出 |
| $K$ | 输出数(类) |
| $N$ | 训练实例数 |
| $z$ | 隐藏的值，内蕴维，潜在因子 |
| $k$ | 隐藏维数，潜在因子数 |

| | |
|---|---|
| $C_i$ | 类 $i$ |
| $\mathcal{X}$ | 训练样本 |
| $\{x^t\}_{t=1}^N$ | $x$ 的集合，上标 $t$ 遍取 1 到 $N$ |
| $\{x^t, r^t\}_t$ | 上标为 $t$ 的输入和期望输出的有序对的集合 |
| $g(x \mid \theta)$ | $x$ 的函数，其定义依赖于参数集 $\theta$ |
| $\arg\max_\theta g(x \mid \theta)$ | 参数 $\theta$，$g$ 关于它取最大值 |
| $\arg\min_\theta g(x \mid \theta)$ | 参数 $\theta$，$g$ 关于它取最小值 |
| $E(\theta \mid \mathcal{X})$ | 样本 $\mathcal{X}$ 上具有参数 $\theta$ 的误差函数 |
| $l(\theta \mid \mathcal{X})$ | 样本 $\mathcal{X}$ 上具有参数 $\theta$ 的似然函数 |
| $\mathcal{L}(\theta \mid \mathcal{X})$ | 样本 $\mathcal{X}$ 上具有参数 $\theta$ 的对数似然函数 |
| $1(c)$ | 如果 $c$ 为真，则值为 1，否则为 0 |
| $\#\{c\}$ | $c$ 为真的元素数目 |
| $\delta_{ij}$ | Kronecker $\delta$：如果 $i=j$，取 1，否则取 0 |

# 前言

机器学习使用实例数据或过去的经验训练计算机，以优化性能标准。当人们不能直接编写计算机程序解决给定的问题，而是需要借助于实例数据或经验时，就需要学习。一种需要学习的情况是人们没有专门技术，或者不能解释他们的专门技术。以语音识别，即将声学语音信号转换成 ASCII 文本为例。看上去我们可以毫无困难地做这件事，但是我们却不能解释我们是如何做的。由于年龄、性别或口音的差异，不同的人读相同的词发音却不同。在机器学习中，这个问题的解决方法是从不同的人那里收集大量发音样本，并学习将它们映射到词。

另一种需要学习的情况是要解决的问题随时间变化或依赖于特定的环境。我们希望有一个能够自动适应环境的通用系统，而不是为每个特定的环境编写一个不同的程序。以计算机网络上的包传递为例。最大化服务质量的、从源地到目的地的路径随网络流量的改变而改变。学习路由程序能够通过监视网络流量自动调整到最佳路径。另一个例子是智能用户界面，它能够自动适应用户的生物特征，即用户的口音、笔迹、工作习惯等。

机器学习在各个领域都有许多成功的应用：已经有了识别语音和笔迹的商用系统。零售商分析他们过去的销售数据，了解顾客行为，以便改善顾客关系管理。金融机构分析过去的交易，以便预测顾客的信用风险。机器人学习优化它们的行为，以便使用最少的资源来完成任务。在生物信息学方面，使用计算机不仅可以分析海量数据，而且还可以提取知识。这些只是我们（即你和我）将在本书讨论的应用的一部分。我们只能想象一下可使用机器学习实现的未来应用：可以在不同的路况、不同的天气条件下自己行驶的汽车，可以实时翻译外语的电话，可以在新环境（例如另一个星球的表面）航行的自动化机器人。机器学习的确是一个令人激动的研究领域！

本书讨论的许多方法都源于各种领域：统计学、模式识别、神经网络、人工智能、信号处理、控制和数据挖掘。过去，这些不同领域的研究遵循不同的途径，侧重点也不同。本书旨在把它们组合在一起，给出问题的统一处理并提供它们的解。

本书是一本入门教材，用于高年级本科生和研究生的机器学习课程，以及在业界工作、对这些方法的应用感兴趣的工程技术人员。预备知识是计算机程序设计、概率论、微积分和线性代数方面的课程。本书的目标是充分解释所有的学习算法，使得从本书给出的方程到计算机程序只是一小步。为了使这一任务更容易完成，对于某些情况，我们给出了算法的伪代码。

适当选取一些章节，本书可用作一学期的课程。再额外讨论一些研究论文的话，本书也可以作为两学期的课程，这时每章后的参考文献将很有用。

本书网页为 http://www.cmpe.boun.edu.tr/~ethem/i2ml/，我将在那里提供一些与本书有关的信息，如勘误表。我真诚地欢迎你将你的反馈意见发到我的邮箱：alpaydin@boun.edu.tr。

我非常喜欢写这本书；希望你能喜欢读它。

# 中文版序

机器学习领域在理论和应用两方面都发展迅速。无论是学术界还是产业界，人们都对能够通过实例学习的计算机程序表现出了极大的兴趣，并且所有国家都是如此。因此，看到本书的中文版出版我特别高兴，另外，我感谢范明教授为翻译本书所做出的努力，他在此之前翻译了几本统计学和数据挖掘的名著。我希望本书的读者能觉得它有益处，并且就像我乐于写它一样乐于阅读它。

Ethem Alpaydın
于伊斯坦布尔 Boğaziçi 大学
2008. 8

## Preface of the Chinese Edition

The field of machine learning is developing rapidly both in theory and applications. There is great interest in computer programs which can learn from examples, both in academia and industry, and this is true for all countries. It therefore gives me great pleasure to see the Chinese language edition of my book in print, and for the effort in dong the translation, I thank Professor Fan who previously have translated several well- known texts on statistics and data mining. I hope that the readers of my book will find it beneficial and enjoy reading it as much as Ienjoyed writing it.

Ethem Alpaydın
Boğaziçi University, Istanbul
August 2008

# 目　录

# 第1章 绪 论

## 1.1 什么是机器学习

随着计算机技术的发展，我们现在已经拥有存储和处理海量数据以及通过计算机网络从远程站点访问数据的能力。目前大多数的数据存取设备都是数字设备，记录的数据也很可靠。以一家连锁超市为例，它拥有遍布全国各地的数百家分店，并且在为数百万顾客提供数千种商品的零售服务。销售点的终端设备记录每笔交易的详细资料，包括日期、顾客识别码、购买商品和数量、消费总额等。这是典型的每日几个 G 字节的数据。只有分析这些数据，并且将它转换为可以利用的信息时，这些存储的数据才能变得有用，例如做预测。

我们不能确切地知道哪些人比较倾向于购买哪种特定的商品，也不知道应该向喜欢读海明威作品的人推荐哪位作者。如果我们知道，我们就不需要任何数据分析；我们只管供货并记录下编码就可以了。但是，正因为我们不知道，所以才只能收集数据，并期望从数据中提取这些问题或相似问题的答案。

我们确信存在某种过程，可以解释我们所观测到的数据。尽管我们不清楚数据产生过程（例如顾客行为）的细节，但是，我们知道数据产生不是完全随机的。人们并不是去超市随机购买商品。当人们买啤酒时，也会买薯片；夏天买冰淇淋，而冬天则为 Glühwein[⊖] 买香料。数据中存在确定的模式。

我们也许不能够完全识别该过程，但是我们相信，我们能够构造一个好的并且有用的近似（good and useful approximation）。尽管这样的近似还不可能解释一切，但其仍然可以解释数据的某些部分。我们相信，尽管识别全部过程也许是不可能的，但是我们仍然能够发现某些模式或规律。这正是机器学习的定位。这些模式可以帮助我们理解该过程，或者我们可以使用这些模式进行预测：假定将来，至少是不远的将来，情况不会与收集样本数据时有很大的不同，则未来的预测也将有望是正确的。

机器学习方法在大型数据库中的应用称为*数据挖掘*（data mining）。类似的情况如大量的金属氧化物以及原料从矿山中开采出来，处理后产生少量非常珍贵的物质。同样地，在数据挖掘中，需要处理大量的数据以构建简单有用的模型，例如具有高精度的预测模型。数据挖掘的应用领域非常广泛：除零售业以外，在金融业，银行分析他们的历史数据，构建用于信用分析、诈骗检测、股票市场等方面的应用模型；在制造业，学习模型可以用于优化、控制以及故障检测等；在医学领域，学习程序可以用于医疗诊断等；在电信领域，通话模式的分析可用于网络优化和提高服务质量；在科学研究领域，比如物理学、天文学以及生物学的大



⊖ Glühwein 是一种温热、有点甜味、加香料的葡萄酒。圣诞节期间，在欧洲很受欢迎。——译者注

量数据只有用计算机才可能得到足够快的分析。万维网(World Wide Web)是巨大的，并且在不断地增长，因此在万维网上检索相关信息不可能依靠人工完成。

然而，机器学习不仅仅是数据库方面的问题，它也是人工智能的组成部分。为了智能化，处于变化环境中的系统必须具备学习的能力。如果系统能够学习并且适应这些变化，那么系统的设计者就不必预见所有的情况，并为它们提供解决方案了。

机器学习还可以帮助我们解决视觉、语音识别以及机器人方面的许多问题。以人脸识别问题为例：我们做这件事毫不费力；即使姿势、光线、发型等不同，我们每天还是可以通过看真实的面孔或其照片来认出我们的家人和朋友。但是我们做这件事是下意识的，而且无法解释我们是如何做的。因为我们不能够解释我们所具备的这种技能，我们也就不可能编写相应的计算机程序。但是我们知道，脸部图像并非只是像素点的随机组合；人脸是有结构的、
2 对称的。脸上有眼睛、鼻子和嘴巴，并且它们都位于脸的特定部位。每个人的脸都有各自的眼睛、鼻子和嘴巴的特定组合模式。通过分析一个人脸部图像的多个样本，学习程序可以捕捉到那个人特有的模式，然后在所给的图像中检测这种模式，从而进行辨认。这就是*模式识别*(pattern recognition)的一个例子。

机器学习使用实例数据或过去的经验训练计算机，以优化某种性能标准。我们有依赖于某些参数的模型，而学习就是执行计算机程序，利用训练数据或以往经验来优化该模型的参数的。模型可以是*预测性的*(predictive)，用于未来的预测，或者是*描述性的*(descriptive)，用于从数据中获取知识，也可以二者兼备。

机器学习在构建数学模型时利用了统计学理论，因为其核心任务就是从样本中推理。计算机科学的角色是双重的：第一，在训练时，我们需要求解优化问题以及存储和处理通常所面对的海量数据的高效算法。第二，一旦学习得到了一个模型，它的表示和用于推理的算法解也必须是高效的。在特定的应用中，学习或推理算法的效率，即它的空间复杂度和时间复杂度，可能与其预测精确度同样重要。

现在，让我们更详细地讨论一些应用领域的例子，以进一步深入了解机器学习的类型和用途。

## 1.2 机器学习的应用实例

### 1.2.1 学习关联性

在零售业，例如超市连锁店，机器学习的一个应用是*购物篮分析*(basket analysis)。它的任务是发现顾客所购商品之间的关联性：如果人们在购买商品 $X$ 时也通常购买商品 $Y$，而有一名顾客购买了商品 $X$ 却没有购买商品 $Y$，则他(或她)即是商品 $Y$ 的潜在顾客。一旦我们发现这类顾客，我们就能针对他们实行打包销售策略。

为发现*关联规则*(association rule)，我们对学习形如 $P(Y \mid X)$ 的条件概率感兴趣，其中
3 $X$ 是我们知道的顾客已经购买的商品或商品集，$Y$ 表示在条件 $X$ 下可能购买的商品。

假定考察已有的数据，计算得到 $P(\text{chips} \mid \text{beer}) = 0.7$，那么我们就可以定义规则：购买啤酒(beer)的顾客中有70%的人也买了薯片(chips)。

我们也许想要区分不同的顾客。针对这个问题，我们需要估计 $P(Y \mid X, D)$，其中 $D$ 是

顾客的一组属性，如性别、年龄、婚姻状况等，这里假定我们已经得到了这些属性信息。如果是考虑书店而不是超市的销售问题，商品就可能是书或作者等。对于 Web 门户网站入口问题，项对应着到 Web 网页的链接，而我们可以估计用户可能点击的链接，并利用这些信息来预先下载这些网页，以取得更快的网页存取速度。

### 1.2.2 分类

信贷是金融机构(例如银行)借出的一笔钱，需要连本带息偿还，通常是分期偿还。对银行来说，重要的是能够提前预测贷款风险。这种风险是客户不履行义务和不全额还款的可能性。既要确保银行获利，又要确保不会因提供超出客户财力的贷款而给客户带来不便。

在信用评分(credit scoring)(Hand 1998)中，银行要计算在给定信贷额度和客户信息情况下的风险。客户信息包括我们已经获取的数据以及与计算客户财力相关的数据，即收入、存款、担保、职业、年龄、以往经济记录等。银行有以往贷款的记录，包括客户数据以及贷款是否偿还。通过这类特定的申请数据，我们可以推断出一般规则，表示客户属性及其风险性的关联性。也就是说，机器学习系统用一个模型来拟合过去的数据，以便能够对新的申请计算风险，从而决定接受或拒绝该项申请。

这是一个分类(classification)问题的例子，这里有两个类：低风险客户和高风险客户。客户信息作为分类器的输入(input)，分类器的任务是将输入指派到其中的一个类。

利用以往数据进行训练后，学习得到的规则可能具有如下形式

$$\text{IF income} > \theta_1 \text{ AND savings} > \theta_2 \text{ THEN low-risk ELSE high-risk}$$

其中 $\theta_1$ 和 $\theta_2$ 是合适的值(参见图 1-1)。这是判别式(discriminant)的一个例子，它是将不同类的样本分开的函数。

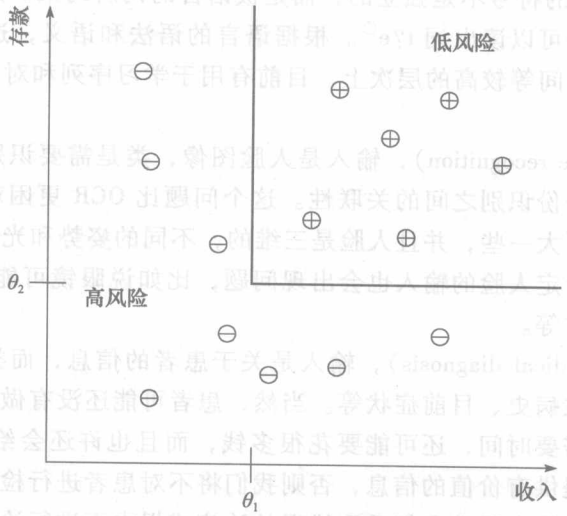



图 1-1 训练数据集示例，其中每个圆圈对应一个数据实例，输入值在对应的坐标上，符号则指示着类别。为简单起见，输入只包括客户的收入(income)和存款(savings)两种属性，两个类分别为低风险(“+”)和高风险(“-”)。图中还显示了分隔两类样本的判别式样例