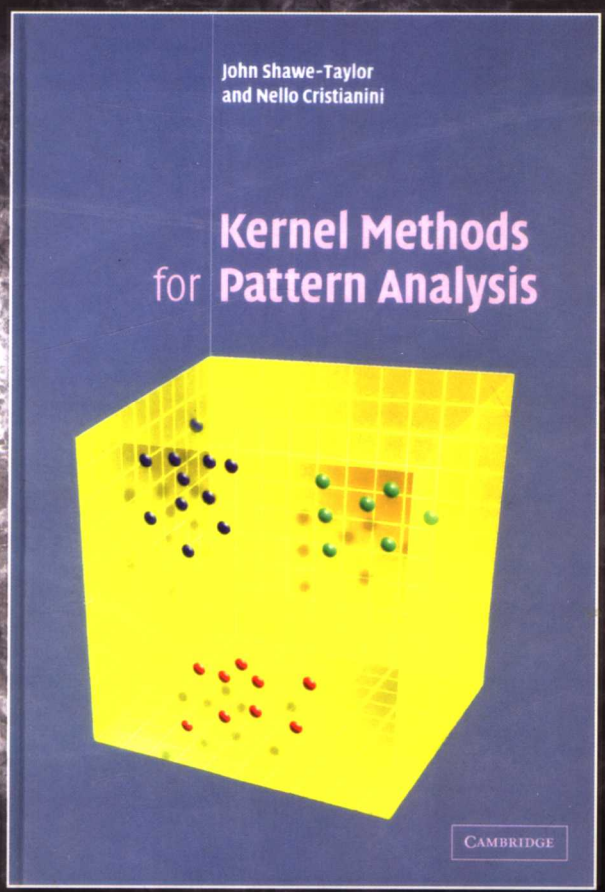
计 算 机 科 学 丛 书

模式分析的核方法

(英) John Shawe-Taylor (美) Nello Cristianini 著 赵玲玲 翁苏明 曾华军 等译



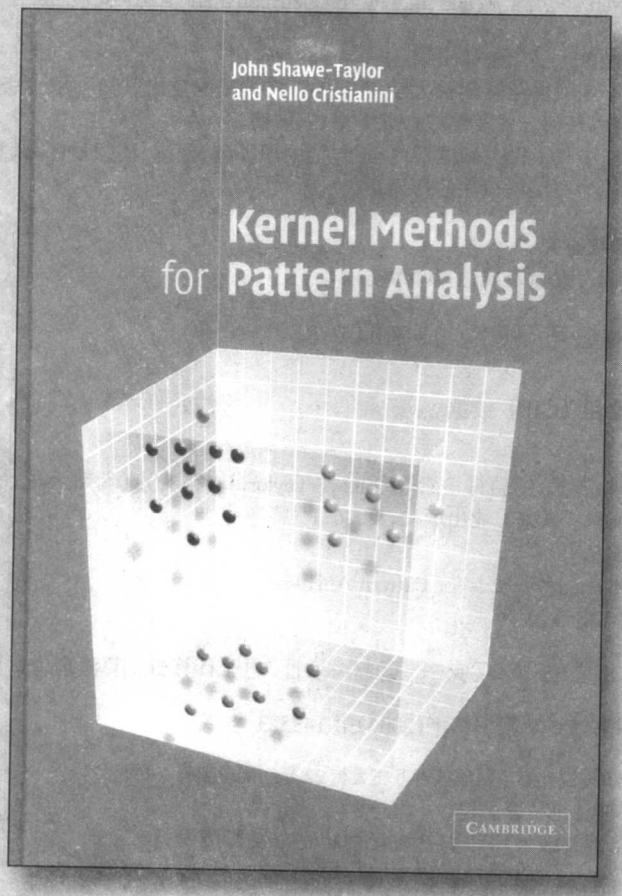
Kernel Methods for Pattern Analysis

机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

模式分析的核方法

(英) John Shawe-Taylor (美) Nello Cristianini 著 赵玲玲 翁苏明 曾华军 等译



Kernel Methods for Pattern Analysis

机械工业出版社
China Machine Press

本书详细介绍基于核的模式分析的基本概念及其应用，主要内容包括：主要理论基础，若干基于核的算法，从最简单的到较复杂的系统，例如核偏序最小二乘法、典型相关分析、支持向量机、主成分分析等。还描述了若干核函数，从基本的例子到高等递归核函数，从生成模型导出的核函数（如 HMM）到基于动态规划的串匹配核函数，以及用于处理文本文档的特殊核函数等。

本书适用于所有从事模式识别、机器学习、神经网络及其应用的学生、教师和研究人员。

本书版权登记号：图字：01-2004-6185

图书在版编目（CIP）数据

模式分析的核方法/（英）肖－泰勒（Shawe-Taylor, J.），（美）克瑞斯天尼（Cristianini, N.）著；赵玲玲等译. －北京：机械工业出版社，2006.1

（计算机科学丛书）

书名原文：Kernel Methods for Pattern Analysis

ISBN 7-111-17853-X

Ⅰ.模…　Ⅱ.①肖…　②克…　③赵…　Ⅲ.电子计算机－算法理论　Ⅳ.TP301.6

中国版本图书馆 CIP 数据核字（2005）第 143523 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：李红玉　李云静

北京瑞德印刷有限公司印刷 · 新华书店北京发行所发行

2006 年 1 月第 1 版第 1 次印刷

787mm×1092mm 1/16 · 20 印张

印数：0 001-4000 册

定价：48.00 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，还揭橥了学术的源变，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、从业人员较少的现状下，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图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray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、研究及庋藏。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，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步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步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum's Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业

的教学度身订造的。其中许多教材均已为M. I. T.，Stanford，U.C. Berkeley，C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

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，但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。教材的出版只是我们的后续服务的起点。华章公司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:

电子邮件: hzjsj@hzbook.com
联系电话: (010) 68995264
联系地址: 北京市西城区百万庄南街1号
邮政编码: 100037

专家指导委员会

（按姓氏笔画顺序）

尤晋元	王　珊	冯博琴	史忠植	史美林
石教英	吕　建	孙玉芳	吴世忠	吴时霖
张立昂	李伟琴	李师贤	李建中	杨冬青
邵维忠	陆丽娜	陆鑫达	陈向群	周伯生
周克定	周傲英	孟小峰	岳丽华	范　明
郑国梁	施伯乐	钟玉琢	唐世渭	袁崇义
高传善	梅　宏	程　旭	程时端	谢希仁
裘宗燕	戴　葵			

译 者 序

模式分析领域研究的是如何发现数据中潜在的关系。随着人们的注意力从线性关系转移到非线性关系，模式分析方法也发生了变化，从最初的统计模式识别，到后来的神经网络和决策树等方法，到本书所讨论的核方法，严格的理论分析推动着新技术的发展和更新。基于核的方法是从统计学习理论中发展出来的较新的研究方法，它有效解决了传统模式识别方法的局部极小化和不完全统计分析的缺点。目前基于核的模式分析方法已经应用于各种类型的数据（不管它们是向量、串或更复杂的对象），并且能够进行多种类型的数据分析，包括相关、回归、排列、聚类等等。

本书是一本综合介绍模式分析的核方法各项标准技术的著作，书中从核函数和基于核的算法的一般原理与性质开始，介绍核函数的特点和性质，接着展开讨论具体的算法，最后引出构造核的技术，其中特别列举了一些适合特定应用的核。本书的叙述循序渐进，内容深入浅出，既不失严谨又易于理解。此外，本书另一大特色是它的配套网站 www. kernel-methods. net 提供了大量在线参考文献的链接，读者可以很方便地查询到所需的内容。

本书作者 Nello Cristianini 是机器学习领域中的一个活跃的年轻学者，在这一领域的关键杂志和会议上都有数篇文章发表，另一位作者 John Shawe-Taylor 研究兴趣广泛，著作涉及学习系统理论分析、离散数学和计算机科学等领域。本书可以看做是对他们之前合著的一本书《An Introduction to Support Vector Machines》的综合和深入，它浓缩了一个研究团队在模式分析方面的 10 年研究成果，为读者进一步学习和掌握最新技术提供了一个理想的起点。

译者在翻译过程中力求忠实原著，专业术语尽量遵循各学科的标准。由于水平和时间有限，对原著的理解可能会有偏差，书中不妥之处在所难免，恳请读者批评和指正。

本书初稿主要由赵玲玲翻译，曾华军负责审阅，翁苏明对全书进行了修改和整理。另外，肖嵘、陈正、张本宇、林晨曦、薛荣贵、孙建涛、韩捷、韩定一也进行了部分书稿翻译和修改工作。最后，特别感谢李江红老师和林宙辰老师所给予的鼓励和支持。

前　言

对数据模式的研究与科学研究一样有非常漫长的历史。例如，考虑一下在天文学上取得重大突破的约翰尼斯·开普勒(Johannes Kepler)，他阐明了著名的三大行星运动定律，我们可以把这三个定律看做是开普勒从第谷·布拉赫(Tycho Brahe)编纂的大量的观测数据中发现的关系。

同样地，对于自动搜索模式的期望的历史至少与计算一样漫长。人们运用许多科学方法和工程方法，比如统计学、机器学习和数据挖掘等等，已在着手处理这个问题了。

模式分析(pattern analysis)处理的是(自动)检测和辨别数据中的关系这一问题。在模式分析领域，大多数统计方法和机器学习方法都假定，数据以向量形式存在，关系可以被表达成分类规则、回归函数或者聚类结构；人们通常把这些方法统称为“统计模式识别”。“句法模式识别”或者“结构模式识别”则代表了另外一种方法，其目的是从诸如串之类的数据中检测规则，这些规则往往按照语法或等价的抽象形式存在。

模式分析自动化算法的发展，经历了3次革命。20世纪60年代，引入了在向量集内检测线性关系的高效算法，并分析了这些算法的计算行为和统计行为。1957年引入的感知机(Perceptron)算法就是一个例子。如何检测非线性关系这一问题，是那个时候的主要研究目标。尽管如此，开发具有相同效率水平的算法，并且保证该算法得到统计理论的支持，已被证明是一个很困难的目标。

20世纪80年代，模式分析领域经历了一场“非线性革命”，几乎同时引入了后向传播多层神经网络算法和高效的决策树学习算法。尽管这些方法用到了启发式算法和不完全统计分析，它们第一次使得检测非线性模式成为可能。非线性革命的影响怎么强调都不过分：它激活了诸如数据挖掘和生物信息学的整个领域。然而，这些非线性算法，是建立在梯度下降法或贪心启发式法的基础上，因而受到局部极小化的限制。由于没有很好地理解它们在统计上的行为，人们利用这些算法时还经常遇到过度拟合的问题。

模式分析算法发展的第三个阶段发生在20世纪90年代中期，当时出现了新的被称为基于核的(kernel-based)学习方法的模式分析方法，该方法最终使得研究人员能够高效地分析非线性关系，而这种高效率原先只有线性算法才能够达到。该方法在统计分析方面进一步发展之后，在高维特征空间内也能够达到很高的效率，并且避免了过度拟合的危险。从各种角度，计算的、统计的和概念的角度来看，在这第三个阶段发展起来的非线性模式分析算法，和线性算法一样，高效而富有理论根据。神经网络和决策树中典型的局部极小化问题和过度拟合问题，也已得到解决。同时，这些方法在处理非向量型数据方面非常有效，这样就建立起了和模式分析的其他分支的联系。

基于核的学习方法，首先以支持向量机(Support Vector Machine，SVM)的形式出现，支持向量机是一种用来摆脱上面提到的计算和统计上的困难的分类算法。然而，很快就产生了基于核的算法，它能够解决分类以外的问题。人们越来越清楚地认识到，这种方法引起了模式分析领域的一场革命。这里，全部的新工具和新技术，都由严格的理论分析所推动，在计算效率的保证下制造出来或发展起来。

此外，这种方法能够消除不同的模式识别子学科之间存在的差距。它提供了一个统一的框架，来思考和操作各种类型的数据，不管它们是向量、串或更复杂的对象，同时也能够进行多种类型的模式分析，包括相关、排列、聚类等等。

本书概括地介绍了这种新方法。我们试图把一个年轻的、茁壮成长中的研究团队的10年深入研究，浓缩到本书的章节中。该团队的研究者们已经一起创造了一个模式分析方法类，该类已成为从业人员工具箱的一个重要部分。

本书介绍的算法能识别多种关系，从传统的分类和回归问题，到诸如排列和聚类等各种更专门化的问题，到包括主成分分析和典型相关分析的高级技术。而且，每一个模式分析问题，都可以和本书最后一部分论述的核函数库中的一类函数结合起来应用。这就意味着这种分析可以用于多种数据，从标准向量类型，到更复杂的诸如图像和文本文档等对象，到与生物序列、图和语法相关联的高级数据类型。

基于核的分析，对于数学家、科学家和工程师来说，是一个强大的新工具。它提供了非常丰富的方法，可以应用在模式分析、信号处理、句法模式识别和其他模式识别(从样条到神经网络)领域。简而言之，它提供了一个崭新的视角，我们仍然远没有了解它的全部潜力。

本书作者参与了基于核的学习算法的发展，对于这一方法的理论、实现、应用和普及，做出了许多贡献。他们的著作《An Introduction to Support Vector Machines》已经被许多大学当做教科书和研究参考书使用。作者也在一个由欧洲委员会(European Commission)资助的工作组的机构中，协助“神经和计算学习(NeuroCOLT)”研究，这个工作组在定义新研究日程和“图像和文本的核方法(KerMIT)”项目中起到了重要作用，而该项目已经应用于文档分析领域。

作者要感谢很多人，他们通过参加讨论、提出建议，或在许多情况下给予了非常详细和富于启发意义的反馈信息，对本书做出了贡献。特别感谢 Gert Lanckriet、Michinari Momma、Kristin Bennett、Tijl DeBie、Roman Rosipal、Christina Leslie、Craig Saunders、Bernhard Schölkopf、Nicolò Cesa-Bianchi、Peter Bartlett、Colin Campbell、William Noble、Prabir Burman、Jean-Philippe Vert、Michael Jordan、Manju Pai、Andrea Frome、Chris Watkins、Juho Rousu、Thore Graepel、Ralf Herbrich 和 David Hardoon。作者还要感谢欧洲委员会和英国基金理事会 EPSRC 对他们基于核的学习方法的研究的支持。

Nello Cristianini 是加州大学戴维斯分校(UC Davis)统计系的助理教授。Nello 要感谢加州大学伯克利分校(UC Berkeley)的计算机科学系和 Mike Jordan，感谢他们在 2001 年～2002 年 Nello 任访问讲师期间对他的款待。他也要感谢麻省理工学院的基于计算机的学习中心

(MIT CBLC)和 Tommy Poggio 2002 年夏天对他的款待,以及为他提供了理想的环境来写这本书的加州大学戴维斯分校(UC Davis)的统计系。本书的许多结构以 Nello 在加州大学伯克利分校、戴维斯分校讲授的课程和讲义为基础。

John Shawe-Taylor 是南安普顿大学(University of Southampton)的计算科学教授。John 要感谢伦敦大学皇家霍洛威学院(Royal Holloway)计算机科学系的同事们。在写作本书的大部分时间,他都在那里工作。

目　录

第一部分　基本概念

第1章　模式分析

模式分析解决的是如何自动检测数据中的模式这一问题，它在现代人工智能和计算机科学领域的许多问题中起着关键作用。我们根据模式理解某个数据源中内在的关系、规律性或者结构；通过检测提供的数据中的显著模式，系统能够对来自同一数据源的新数据做出预测。在这个意义上，系统通过“学习”关于生成数据的数据源的信息，获得了泛化能力(generalisation power)。许多重要的问题只有运用这种方法才能得到解决，这些问题涉及的领域包括生物信息学、文本分类、图像分析以及 Web 检索。近年来，模式分析已经成为一种标准的软件技术，出现在许多商业产品中。

一些早期的方法在寻找线性关系时很有效率，但是在处理非线性模式时，人们只能采用理论性较差的做法。本书描述的方法，把先前只限于线性系统的方法的理论性，和非线性方法的灵活性与适用性相结合，从而形成了一类非常强有力的、健壮的模式分析技术。

人们已经把统计模式识别和句法模式识别区分开来，前者主要处理分布满足某些统计假设的向量，而后者则处理结构化对象，例如序列或者形式语言，较少依赖于统计分析。本书介绍的方法兼顾这两个方向，因为它既能处理一般的数据类型，例如序列，同时又能处理统计模式分析中的典型问题，例如从有限的样本中学习。

1.1　数据中的模式

1.1.1　数据

本书涉及的是数据和从中识别有价值知识的方法。我们说的数据，意思是指任何观察、测量或者记录仪器的输出。因此它包括以数字格式保存的图像、描述物理系统状态的向量、DNA 序列、文本块、时间序列和商业交易的记录等等。我们说的知识，意思是指就数据之间的关系层面和数据内部的模式层面来说更为抽象的东西。这类知识能使我们对数据源做出预测，或者对数据中内在的关系做出推断。

一般来说，人工智能(Artificial Intelligence，AI)和计算机科学领域中许多非常有趣的问题都极为复杂，人们很难或者甚至都不可能指定一个明确的程序化解决方案。举个例子，考虑识别 DNA 序列中的基因这一问题，我们不知道如何指定一个程序从人类的 DNA 序列中挑选出代表基因的子序列。类似地，我们不能直接给计算机编程以识别照片中的脸孔。为了解决这些问题，学习系统提供了另外一种方法。通过充分利用从样本数据中提取的知识，学习系统通常能够自适应地推断这类任务的解决方案。我们把这种软件设计方法叫做*学习方法*

(learning methodology)。它也被称为*数据驱动*(data driven)方法或者*基于数据*(data based)的方法，该方法与*理论驱动*(theory driven)方法形成对比，后者能够精确定义所需算法。

近年来，学习方法适用的问题的范围扩大得非常快。例如文本分类、电子邮件过滤、基因检测、蛋白质同构体检测、Web 检索、图像分类、手写字符识别、贷款拖欠预测和分子性质确定等等。这些问题非常困难，有时甚至不可能采用标准方法来解决，但是人们已经证明它们可以用学习方法来解决。解决这些问题不只是为了研究人员的兴趣。例如，从分子的结构预测该分子的重要性质，可以为制药公司节省下数百万美元；而在通常情况下，制药公司要做昂贵的实验来测试临床候选药物。而若能够识别具有高度预测能力的生物标志(biomarker)蛋白质的结合体，并把这一结果用于早期的癌症诊断检验，也就可能挽救许多人的生命。

4

总的来说，模式分析这个领域研究的是运用学习方法发现*数据中的模式*(patterns in data)。它所寻找的这些模式包括许多不同的类型，例如分类、回归、聚类分析(有时统称为*统计模式识别*(statistical pattern recognition))、特征提取、语法推断和分析(有时统称为*句法模式识别*(syntactical pattern recognition))。在这本书里，我们会阐述所有这些领域中的概念，同时运用举例和案例研究的方法。这些例子和案例来源于上面提到的一些应用领域：包括生物信息学、机器视觉、信息检索和文本分类。

值得强调的是，虽然传统的统计学主要用*多元统计学*(multivariate statistics)的方法来处理向量数据，但上面提到的许多重要应用中的数据都是非向量形式的。我们还必须提及，计算机科学领域的模式分析主要集中在分类和回归，在这个意义上，模式分析和神经网络文献中的分类是同义的。引入*模式分析*(pattern analysis)这一术语，部分也是为了避免把上述比较狭窄的关注点和我们一般的定义相混淆。

1.1.2　模式

想像一个数据集，它包含关于太阳系行星位置的上千个观测值，例如九大行星每颗行星位置的日常记录。显然，某颗行星特定日子的位置依赖于同一行星前些日子的位置；基于这些知识，我们确实能够做出相当精确的预测。此数据集包含一定数量的冗余，也就是可以从数据的其他部分重新构造出来的信息，因此它们不是严格必需的。在这种情况下，我们说数据集是*冗余*(redundant)的：即能够从数据中提取简单的规律，并能用这些规律重新构造每颗行星每一天的位置。支配行星位置的这些规则，被称为开普勒定律。17 世纪时，约翰尼斯·开普勒(Johannes Kepler)通过分析第谷·布拉赫(Tycho Brahe)先前几十年记录下来的行星位置资料，发现了其三大定律。

我们可以把开普勒的发现看做模式分析(或者数据驱动分析)的一个先例。通过假定这些定律不变，人们可以预测将来的观测结果。这些定律符合行星数据表现出来的规律性，因此可以推断，它们也符合行星运动本身。它们说明，行星沿着以太阳为一焦点的椭圆形轨道运行；连接行星和太阳的那条轴线，在相等的时间段内扫过相等的面积；周期 P(绕太阳旋转一周的时间)和到太阳的平均距离 D 有关：都遵循 $P^3 = D^2$ 这一等式。

5

例 1.1　从表 1-1 我们可以观察到冗余数据集的两个潜在性质：一方面它们是*可压缩*(compressible)的，因为运用开普勒第三定律，仅从一列数据就能构造出这个表；另一方面它们又是*可预测*(predictable)的，例如，只要测量出周期，就能从定律推断出新发现的行星到太

阳的距离。这种预测能力是数据中存在可能的隐藏关系的直接结果。一旦发现这些关系，我们就能进行预测，从而更有效地处理新数据。 ■

表 1-1 数据中的模式的一个例子：对于所有的行星，量 D^2/P^3 保持不变。这意味着我们可以通过只给出其中的一列数据实现压缩，或者当发现以前未知的外部行星时，我们可以预测它的数值

	D	P	D^2	P^3
水星	0.24	0.39	0.058	0.059
金星	0.62	0.72	0.38	0.39
地球	1.00	1.00	1.00	1.00
火星	1.88	1.53	3.53	3.58
木星	11.90	5.31	142.00	141.00
土星	29.30	9.55	870.00	871.00

通常我们把要预测的特征作为其余特征的函数，例如，把距离作为周期的函数。为了做到这一点，这个关系必须是可逆的，所要预测的特征可以表示成其他值的函数。毫无疑问，我们将致力于随时寻找这类具有显式形式的关系。数据中也可能存在其他更为一般的关系，它们也能被检测到，并为人们所利用。例如，如果我们找到一个被表示成恒定函数 f 的一般关系，且这个函数满足

$$f(\mathbf{x})=0 \tag{1.1}$$

其中 $\mathbf{x}$ 是一个数据项，我们可以用 $\mathbf{x}$ 来识别异常的或错误的数据项，对于这些数据项这种关系不成立，即 $f(\mathbf{x})\neq 0$。然而，在这种情况下，由于它会要求我们根据方程(1.1)定义的流形(manifold)确定一个较低维的坐标系，实现压缩能力是很困难的。 6

开普勒定律是正确的，它适用于给定恒星系的所有行星。我们称这类关系为准确(exact)关系。上面给出的例子包括诸如贷款拖欠之类的问题，即基于贷款受理时提供的信息，预测哪个借款人不会偿还贷款。很明显，在这种情况下我们无法找到一个准确的预测，因为除了那些提供给系统的信息以外，还存在其他因素，而这些因素有可能是决定性的。例如，借款人有可能在获得贷款后不久就失业，从而无力履行偿还义务。在这些情况下，系统力所能及的就是找到在一定的概率下成立的关系。学习系统已成功地找到了这类关系。可压缩性和可预测性这两个性质也很明显。我们可以确定数据中的大部分都符合的关系，然后简单地追加一个例外情况列表。只要关系描述简明扼要，不存在太多的例外，就会减少数据集的大小。同样地，我们可以利用关系做出预测，例如，借款人是否会偿还贷款。由于这种关系在一定的概率下成立，很有可能实际结果会符合预测。我们把在一定的概率下成立的关系，叫做统计(statistical)关系。

人们在预测基于分子结构的物质的性质时，受到另一个问题的阻碍。在这种情况下，对于以实数形式存在的属性(例如沸点)来说，要寻找的关系必定只是近似的，因为我们无法得到准确的预测。通常可能希望预测中存在的期望误差很小，或者希望真值在很高的概率下存在于一定的预测区间之内，但是模式分析所寻求的关系必定是近似的。如果没有其他原因，

人们可以说，开普勒定律是近似的，因为它们没有考虑广义相对论。然而，当我们把注意力转到学习系统时，就会发现其中存在的近似的程度比开普勒定律更高。如果关系中的值包括一些不精确值，这些关系就叫做近似(approximate)关系。对于近似关系我们仍然谈论预测，虽然为此必须刻画估计值的精确性，并且很有可能还要刻画估计值适用的概率。为了再一次证明可压缩性，我们可以指定输出值和真实值之间的误差修正，如果误差修正很小，它们占的空间会比较少。

造成数据集冗余的关系，也就是我们通过挖掘数据集而提取的规律，在整本书中被称为
7 模式(pattern)。模式可以是类似于开普勒定律的确定性关系，也可以是如上文所述的近似关系，或者是只在一定概率下成立的关系。我们感兴趣的是不存在准确的规律的情形，特别是不存在可以简单地描述成像开普勒定律那样的规律的情形。出于这个原因，我们把模式(pattern)理解成数据中的任何关系，不管这种关系是准确的、近似的还是统计的。

例 1.2　考虑下面这个虚构的例子，该例子描述了在二维直角坐标系中的行星位置的一些观测值。注意它显然不是开普勒在第谷的数据中寻找出规律的那种情况。

x	y	x^2	y^2	xy
0.8415	0.5403	0.7081	0.2919	0.4546
0.9093	−0.4161	0.8268	0.1732	−0.3784
0.1411	−0.99	0.0199	0.9801	−0.1397
−0.7568	−0.6536	0.5728	0.4272	0.4947
−0.9589	0.2837	0.9195	0.0805	−0.272
−0.2794	0.9602	0.0781	0.9219	−0.2683
0.657	0.7539	0.4316	0.5684	0.4953
0.9894	−0.1455	0.9788	0.0212	−0.144
0.4121	−0.9111	0.1698	0.8302	−0.3755
−0.544	−0.8391	0.296	0.704	0.4565

图 1-1 左边的曲线显示(x,y)平面的数据。我们可以对隐含在这些位置下面的规律做许多假设。然而，如果我们考虑量 $c_1x^2+c_2y^2+c_3xy+c_4x+c_5y+c_6$，就会看到，选择某些参数时这个量是一个常数。正如图 1-1 左边所示的曲线那样，我们只用两个特征 x^2 和 y^2，就得到了一种线性关系。如果数据是随机的，或者如果轨道遵循的曲线不同于二次曲线，一般来说就不会是这种情况。实际上，数据中的这种不变性意味着行星沿着椭圆轨道运行。改变坐标系后，这种关系就变成线性的。■

在这个例子里，我们看到把坐标变换用于数据如何导致模式表示形式的改变。使用初始坐标系，模式被表示成一种二次形式，而使用单项式坐标系看起来则类似于线性函数。改变
8 描述数据的坐标系有可能改变模式的表示形式，这将是本书中反复出现的一个主题。

例子中的模式具有函数 f 的形式，该函数满足

$$f(\mathbf{x})=0$$