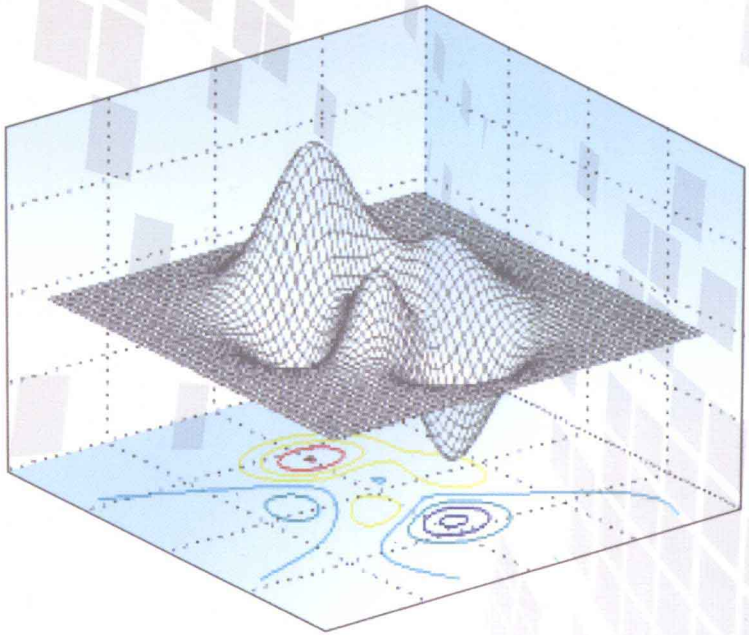
中国气象局培训中心培训系列教材

A Textbook on Support Vector Machines

支持向量机方法

应用教程

陈永义 熊秋芬 编著



气象出版社
China Meteorological Press

中国气象局培训中心培训系列教材

支持向量机方法应用教程

陈永义　熊秋芬　编著

气象出版社
China Meteorological Press

内容简介

支持向量机(Support Vector Machines,SVM)是令人瞩目的机器学习新方法,但其严密的理论结构又让应用者望而却步。本书是在作者多年进行的SVM应用方法研究、教学和推广的基础上完成的一本实用教程。

本书以SVM的应用为主线,用通俗、直观的语言,由浅入深地介绍了SVM分类和回归的基础原理、实现方法及特点,对学习机性能评估和统计学习理论也有所涉猎。书中提供了一个应用SVM方法学习建模和预报应用的软件平台,使读者能够快速"上手"为我所用地解决实际问题,并给出了天气预报中的应用实例。

本书对在广泛领域中,涉及分类、回归等问题的工程技术人员和研究工作者有参考价值,也可作为计算机、信息等相关专业本科生和研究生的SVM选修课教材。

图书在版编目(CIP)数据

支持向量机方法应用教程/陈永义,熊秋芬编著.—北京:气象出版社,2011.7

ISBN 978-7-5029-5236-5

Ⅰ.①支… Ⅱ.①陈… ②熊… Ⅲ.①向量计算机—算法理论—教材 Ⅳ.①TP301.6

中国版本图书馆CIP数据核字(2011)第114690号

支持向量机方法应用教程

Zhichixiangliangji Fangfa Yingyong Jiaocheng

出版发行:气象出版社

地　　址:北京市海淀区中关村南大街46号　　**邮政编码**:100081

总 编 室:010-68407112　　**发 行 部**:010-68409198

网　　址:http://www.cmp.cma.gov.cn　　**E-mail**:qxcbs@cma.gov.cn

责任编辑:何晓欢　张　斌　　**终　　审**:黄润恒

封面设计:博雅思企划　　**责任技编**:吴庭芳

印　　刷:北京京科印刷有限公司

开　　本:720 mm×960 mm　1/16　　**印　　张**:10.5

字　　数:215千字　　**印　　张**:1～1 500

版　　次:2011年7月第1版　　**印　　次**:2011年7月第1次印刷

定　　价:36.00元

前　言

“计算机学习”是当前国际上计算机理论和应用研究的最热门话题之一。尽管至今仍有学者质疑计算机能否像人一样“学习”，但是让计算机基于已有数据建立分类、回归等模型，从而应用于实际预报、预测、诊断、分析和决策等的研究，近年来已有长足进展。而支持向量机方法(Support Vector Machines，SVM)则是其中最受瞩目的一种新方法。

2000年，我初次接触SVM方法，立即被它的新颖思想所吸引——把样本空间“升维”，在无穷维空间中来解决问题，却又不增加计算的复杂性。然而，使这一切得以实现的关键竟是依据近百年前的一个纯数学定理。新的统计学习理论和SVM方法不愧是使机器学习研究腾飞的强劲双翅——前者是严密的理论结构，奠定了机器学习的理论基石，而后者是理论的巧妙实现，给出了有效的应用新方法。两翼双飞，二者结合得如此完美。我相信这一新方法能够在广阔的应用领域大显身手，自然也会在天气预报，特别是数值预报产品的释用上得到成功的应用。

2002年成都市气象台的冯汉中同志来中国气象局培训中心访问进修，他很快地掌握了SVM方法并用之成功实验了降水分类预报和单站气温预报，进而用VB语言开发出了基于SVM^{Light}的SVM学习建模、预报的应用软件平台CMSVM的雏形。本书的初稿就是在这一过程中形成的培训讲义(2003年)。

随后，中国气象局培训中心将SVM方法及其在气象中的应用作为重要的新技术之一，先后安排在省级首席预报员培训班、高级工程师培训班、数值预报释用班、气候系统监测及短期气候预测培训班、城市环境班、气象普及班、民航气象班、研究生班及SVM方法专题班等系列培训班进行讲授和推广，百余个气象台站和气象科研院所正式将SVM方法和CMSVM软件引入业务和科研中应用，国家气象中心的刘还珠老师还把SVM方法嵌入国家气象中心的“气象要素客观集成系统”加以推广应用。本书附录列出的论文索引从一个侧面反映了SVM方法近年来在气象业务中的推广应用情况。

为了便于推广应用，我们把相关培训教材和应用软件CMSVM 1.0

公开放在了网站上(http://streaml.cma.gov.cn/cmsvm/)。据我们收到的可靠反馈信息的不完全统计,它们还得到了计算机、遥感、水文、地质、地震、海洋、烟草、电力、农业、医学、经济和金融等领域的专业人员的关注和使用,并有南京信息工程大学、解放军理工大学、成都信息工程学院、云南大学、中山大学、中南林业科技大学、吉林大学、哈尔滨工业大学、华东师范大学、上海海事大学、西安电子科技大学、西安科技大学、北京科技大学、南京大学、中国科学院大气物理研究所等不同专业院校所的研究生应用我们的CMSVM做研究。

目前提供的这本书是我们多年的读书学习、应用研究和培训教学工作的一个粗略总结。SVM方法涉及较多的数学基础知识和理论,如泛函分析、统计分析、最优化理论、最优化算法等。作为面向气象及其他应用领域研究人员的一本应用教材,本书内容重点放在了讲清SVM的基本思想和具体应用方法上;而未能对相关的理论问题(如理论证明、算法实现、模型泛化能力界的估计依据等)加以全面介绍和深入讨论,是本书的缺憾之一。值得庆幸的是本书所引用的主要外文参考文献近年已相继有中译本出版,这些书是有志于从事核方法和机器学习理论研究的人员的好参考,有关内容都可以从中找到答案。

本书第1~6章力图对Vapnik V N四十年研究的成果,给出尽可能通俗、直观的介绍;第7章介绍了我们开发的一个应用SVM方法做学习建模、预报的软件平台CMSVM 2.0,该软件可在网上(http://streaml.cma.gov.cn/cmsvm)免费下载;第8章介绍了SVM方法在气象预报中的初步应用;附录给出了截至2009年年底公开发表的有关SVM方法在我国气象领域应用文章的目录索引。

本书虽然主要是面向气象的培训教程,但所介绍的理论和方法具有一定的通用性。天气预报常遇到的分类、回归等问题,在水文、海洋、地质、地震、医学、农业和人文等广泛的领域同样会遇到。“辞异而理同”,本书介绍的方法、实例和软件工具CMSVM 2.0,对多领域的工程技术人员和研究工作者均具有参考价值。

本书也可作为计算机、信息等相关专业本科高年级和研究生SVM选修课教材之用。如果重点选讲前6章,辅以CMSVM软件系统的应用实践,大约需要40学时。

本书第1~7章为陈永义撰写,第8章由熊秋芬执笔。

本书作者感谢国家自然科学基金(60072006)的资助，感谢中国气象局培训中心高学浩主任和俞小鼎教授的鼓励、关注和支持。向所有被引用的参考文献的作者一并表示诚挚谢意。

还要感谢王泳、李春辉在CMSVM软件平台实现上的辛勤劳动，由于他们的出色工作，使CMSVM 2.0软件得以顺利在气象系统推广应用。

限于作者们的水平，本书的内容难免会有疏漏，殷切地期望读者对本书提出修改意见。

陈永义
2010年7月于北京
yongyichen45@gmail.com

本书阅读建议

本书内容的层次结构安排，考虑到了不同读者的需求。

如果读者想对SVM方法及其应用做一粗略、宏观的了解并不实际应用，可以只阅读第1章。它对SVM方法的基本思想、特点及应用做了通俗概括的介绍。

如果读者熟悉并应用过其他的机器学习或统计分析方法（如判别分析、回归分析等），则可以在阅读第1章后直接进入第7章。第7章是我们开发的一个应用SVM方法的软件平台。熟悉CMSVM 2.0软件平台后，可直接用SVM方法解决实际应用问题并加以比较，然后根据感受再决定是否阅读其他章节。

第1～4章介绍了SVM方法的核心知识。如果读者想掌握该方法并应用其解决实际应用问题，就必须仔细阅读这4章，并要把它们基本读懂。在此基础上直接进入第7章，熟悉CMSVM软件平台后，就可以试用SVM来解决实际应用问题。

读者初步应用SVM方法解决分类或回归问题后，可能会产生一系列新问题。第5章的内容就是针对此情况撰写的，它帮助读者在几个方面对SVM方法作进一步的思考，逐步深入到计算机学习理论的核心问题。

第6章概略介绍了计算机学习的理论问题，是本书最难读的一章，对于只关心实际应用的读者可以跳过它。但对于关心统计学习理论问题的读者来说，这一章还算是较“通俗”的介绍。由于介绍得太过粗略，要彻底搞清楚这些理论问题，还必须补充阅读其他的书籍（如Vapnik的著作）。

第8章给出了几个天气预报中的应用实例和对不同方法的比较分析。虽然内容只限于气象领域，但触类旁通，对其他领域的应用也有参考价值。

当然，读者如果有兴趣并具备条件，最好的阅读建议是有重点地循序渐进地通读全书。这可以让你不但对SVM方法有较全面的了解，并能很好掌握一种软件工具，灵活方便地解决有关实际问题。

目　录

第1章 绪 论

1.1 计算机应用的历史回顾

在计算机[①]诞生以来的五十多年里，它给人类社会和人们生活带来了难以言尽的巨大变化。从应用角度来看，这半个世纪计算机技术的发展，大体上可以划分为三个阶段，或者说是历经了三个台阶——数值计算、数据处理和知识处理。

数值计算是人们发明计算机的初衷，希望借助它使人们能计算规模巨大的数值问题，并且算得更快、更准确。在这一方面人们已经取得了极大的成功，甚至可以说在数值计算的主要指标上，“电脑”已经超过了人脑。

如果计算机的应用只限于数值计算，那它不过是一个“大算盘”。人们很快就发现计算机还善于进行比数值计算更广义的数据处理。不但数值是数据，符号、文字、声音、图像等都是数据，计算机可以对这类广义的数据进行大容量、高速、高效、精确的管理。随着数据库理论的成熟和广泛应用，大量的数据库系统(Data Base System，DBS)、管理信息系统(Management Information System，MIS)、企业资源计划系统(Enterprise Resource Planning，ERP)、办公自动化与网络系统等几乎渗入各行各业，进入人类活动的各个领域，取得了难以估量的社会效益和经济效益。人们生产生活的各个方面都离不开计算机的数据处理。计算机在数据处理上取得了巨大的成功。

人类绝不满足于已有的成功。透过成功的眩目光环，人们发现几乎全部前面提到的成功应用，都仅局限于对数据的简单处理上，无非是对原始数据的录入、修改、删除、更新、查询、统计、打印等，没有任何智能处理和灵气。如果计算机的数据处理停留于此，那它无非是起到了一个“大账本”加上一个“大算盘”的功能。人类对计算机的新追求是希望计算机能够具有人类的智能，从而可以代替人脑干更多的事情，成为名副其实的电脑。这就是以知识处理为标志的计算机应用的新阶段。

现代计算机理论的鼻祖冯·诺伊曼(John von Neumann)在病逝前完成了他的最后一部著作《计算机和人脑》(冯·诺伊曼 2010)。在这本小册子中，冯·诺伊曼在人们对电脑的一片赞誉声中冷静地指出：无论是从逻辑原理还是从算法结构上看，当时的电子计算机都与人脑迥然不同，记忆、存储方式也不一样，计算机远不及人脑。从那时至今，虽然从硬件上看计算机的发展经历了四代(电子管、晶体管、集成电路、

① 计算机指能进行数学运算的机器。有用机械装置做成的，如手摇计算机；有用电子元器件组成的，如电子计算机。现特指电子计算机，或称电脑，下同。

大规模集成电路),但从原理上看它们都还属于冯·诺伊曼式计算机。我们在期待新一代非冯·诺伊曼式计算机早日诞生的同时,也要努力让目前的计算机具有更多的知识处理能力。

与数值计算、数据处理相比,知识处理具有更大的复杂性和困难性。近四十多年来,以人工智能为代表的研究工作不断深入,并取得了一系列令人瞩目的成果(超级电脑"深蓝"打败国际象棋特级大师卡斯帕罗夫,各种机器人代替人类从事多种工作等)。多种非经典逻辑(多值逻辑、模糊逻辑、格值逻辑、模态逻辑、概率逻辑、非单调逻辑、自动认识逻辑等)的提出,多种软计算方法(模糊数学、神经网络、遗传算法、免疫算法、粗糙集方法、数据挖掘算法等)的诞生,为知识处理提供了新手段。

智能是人类所独有的能力,如何让计算机具有智能是一个极具挑战性的课题。谈到智能,离不开知识。而人类知识在计算机中如何表示?计算机如何获取这些知识?如何运用这些知识?这些重要的基本问题至今都没有得到权威的解答,因而仍在探索之中。

如果说计算机应用的前两个台阶(数值计算和数据处理)均已成功跨越,则知识处理这第三个台阶到目前为止也不过是刚刚起步。虽然数值计算和数据处理的成功为知识处理提供了基础和条件,但知识处理的广度、深度、难度和复杂度等都远非数值计算和数据处理所能比拟的。也许这些问题需要等到全新一代的非冯·诺伊曼式计算机问世后才能得到彻底解决。

本书的讨论局限于知识处理方面的一个十分确定、具体的问题:计算机学习问题,而且只介绍计算机学习领域中近年兴起的一种新方法——支持向量机(Support Vector Machines,简称 SVM)方法。本书特别聚焦于 SVM 的分类、回归方法,及其在天气预报业务中的应用。

SVM 方法思路巧妙,理论基础坚实,特别适用于处理高度非线性分类、回归等实际问题。与已有的其他方法相比,SVM 方法具有明显的优势,近年来在很多领域得到了成功的应用,在天气预报领域的初步应用也显示了很好的前景。

1.2 计算机学习的基本问题

学习是人类最重要的一种思维活动。无知的孩童通过不断地观察、学习和实践,日积月累地形成自己的思想,走向成熟,甚至发展成为杰出人才。

用人类的思维和实践揭开人脑学习的秘密是一个历史性的难题。如果说在几十年前它还只是少数哲学大师和学术先觉探讨的话题,那么由于信息技术的迅猛发展和计算机的广泛应用,学习问题,特别是计算机学习问题,已成为广大研究人员、技术人员必须面对的实际课题。

从逻辑的角度看,学习的本质是归纳,而学习所得结果的应用才是演绎。有人说,只凭推理演绎是得不出创造性的东西的,只有归纳才有可能创新。经典逻辑提供

了多种推理演绎的手段，但对于归纳学习却很少提出好办法。

人类的学习是一个复杂的思维过程，目前提出的计算机学习方法只是对这一过程的粗略模拟和近似。类似于从认识个体中抽象出概念，机器学习是通过已有样本数据集建立数学模型。

图1.1和1.2分别给出了这两种学习过程的粗略图示。

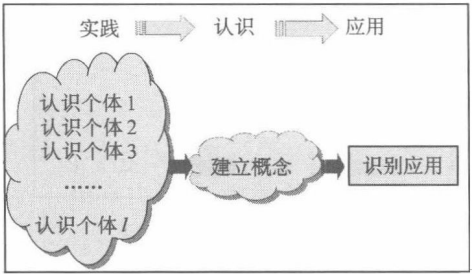


图1.1 人类学习过程图示

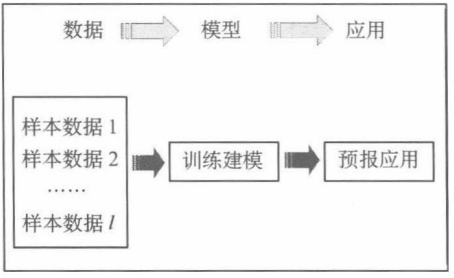


图1.2 机器学习过程图示

比较两者的异同可以看出：

(1) 人脑从观察、实践中学习，感触到的大量事实是学习的素材；计算机从数据中学习，给定的相关数据集是学习的依据，通常称这样的数据集为训练集。

(2) 人脑学习的过程是通过对大量事实的分析、比较、对比、归纳等高级思维活动(甚至是形象思维)从而获得有用的知识；计算机学习是采用人类指定的某种确定算法对训练集进行“训练”学习，建立用于分类、回归等的数学模型。

(3) 人脑获得的知识可以用来指导以后的实践；机器学习获得的数学模型可用来对未来同类相关数据做识别、判断、推理、预测和预报。

(4) 人脑通过实践可以修正、提升已有的知识；计算机可以扩充训练数据集重新训练学习，改进已得的模型。

(5) 人脑的学习过程至今尚无方法被准确地形式化，它涉及复杂的生理、物理、化学、心理等过程；机器学习问题则是被人们完全形式化了的一个数学问题，它是对人脑学习过程的粗化和约简。

通俗地说，学习问题就是从有限数量的已经发生的历史样本数据中，寻找隐藏于其中的某种依赖关系，从而建立起可应用于未来预测预报的模型。或者用一句话概括：机器学习就是基于数据用计算机建模。

机器学习本质上就是一种统计方法。很多计算机对实际的应用问题的处理本质上都可以归结为上述机器学习问题。如通常所说的模式识别、回归分析、密度函数估计、聚类分析、人工神经网络等，都可以看成是这里所叙述的机器学习的特例。

图1.3和1.4分别给出了机器学习的直观描述和数据描述。

历史样本数据集由若干个成员组成，通常称一个成员为一个样本。因此观测数

据集也称样本集。每一个样本数据通常包含因子量(用 n 维向量 $\boldsymbol{X}$ 表示)和实况量(用实数 y 表示),因此历史样本数据集的第 i 个样本数据可以表示成($\boldsymbol{X}_i$, y_i)如图 1.4。其中 $\boldsymbol{X}_i=(x_{i1}, x_{i2}, \cdots, x_{in})$为 n 维向量,它对应于选取的预报因子的值;y_i 为预报量的实况值。

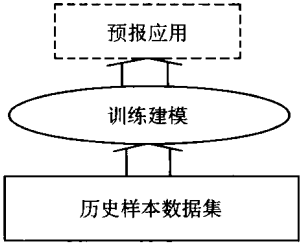


图 1.3 机器学习的直观描述

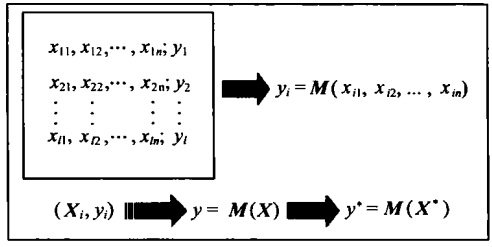


图 1.4 机器学习的数据描述

图 1.4 中所表示的 $y=\boldsymbol{M}(\boldsymbol{X})$是由样本数据集经过训练学习建立的模型。它是一个预报量与预报因子值的确定的函数关系,可用于预测预报。这一函数关系不仅应该同已有的历史样本数据充分符合(拟合),更重要的是要在对未来的预测预报进行应用时有良好的效果。

如果这一函数关系是线性函数,则对应的就是线性(分类或回归)模型,否则为非线性模型。本书称这样的一个函数关系为建立的一个学习机。

历史样本数据集是机器学习的基础,并不是随便凑成的一堆数据就可以称为样本数据集。必须指出,选取的预报因子和样本数据集的质量决定了建立的学习机是否有效。我们所说的样本数据集具有一定的客观性,它们是现实世界某一现象或过程的真实反映,但它们又具有一定的不确定性,往往是随机选取的。可以想象成是一只“上帝之手”把样本集的输入数据一一选取出来(我们相信数据集的生成是有规律的,但这个规律对我们来说是未知的)。而后它们通过一个完全确定过程的“黑箱”输出对应的数据,但该“黑箱”对于我们来说是未知的。机器学习的核心就是近似找出这一“黑箱”对应的输出输入关系。

用数学的语言来描述,图 1.3 可以细化为图 1.5:

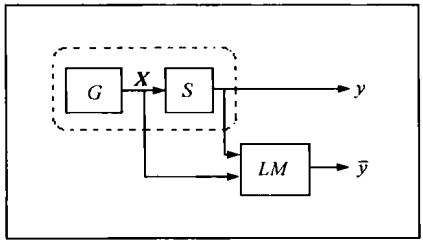


图 1.5 机器学习的图示

其中的框图 G 为输入随机向量(因子向量)产生器,所有输入数据看成是从满足某一固定但未知的概率分布函数 $F(\boldsymbol{X})$ 中独立生成的;框图 S 为产生器,它是一个黑箱系统,对每个输入向量产生一个确定的输出,产生输出的依据同样是客观存在且固定的,但对我们来说是未知的。一般认为这种依据是一个未知的条件概率分布函数 $F(y|\boldsymbol{X})$。产生器(G)和训练器(S)这两部分用虚线框起,它对应于实际问题的样本数据集。

我们称框图 LM 部分为一个学习机器,简称学习机。它通过对大量样本数据的训练学习,从指定的函数范围内确定一个数学模型(本书称建立一个学习机),作为对黑箱系统(训练器)的最好近似。所谓最好近似就是当把 G 产生的数据 $\boldsymbol{X}$ 同时输入到 S 和 LM 时,它们的输出值 y 和 $\bar{y}$ 最为接近。通常要预先指定待确定模型所属的函数类 $f(\boldsymbol{X},\boldsymbol{\alpha})$,从中选取一个最优函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$。这样,建立学习机的问题就变成确定最优函数所对应的一组参数 $\boldsymbol{\alpha}^*$ 的问题。

机器学习的过程就是由样本数据集建立学习机的过程。机器学习问题可以形式化为:

给定函数集 $f(\boldsymbol{X},\boldsymbol{\alpha})$,$\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$ 和 l 个独立同分布的样本数据训练集

$$(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_l,y_l)$$

其中 $\boldsymbol{X}_i\in R^n$,为 n 维向量,如何从给定的函数集 $f(\boldsymbol{X},\boldsymbol{\alpha})$ 中选择出能够最接近训练器 S 响应的函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$ 来呢?

这里产生了如下一些理论问题:

(1) 学习机的备选函数类 $f(\boldsymbol{X},\boldsymbol{\alpha})$ 如何确定?范围太大(比如不加限制)无法计算,范围太小(比如限定线性函数)难以保证精度。

(2) 依据什么选取原则确定最优函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$?通常采用误差最小化原则,是否还有其他原则?相应的误差如何估算?

(3) 样本数据集是随机产生的,一般来说,即使是对于同一个实际预报问题,对应于不同样本的数据集所建立的学习机模型是不会完全一样的。但这些学习机随样本数据的增加是否有一个稳定的趋势,即收敛性?若收敛,收敛速度如何?可否调控?

(4) 建立的学习机即使对训练数据的拟合率很好,但推广能力也未必好。通过历史样本数据建立的学习机的泛化能力(即推广能力)如何估计?学习机的性能好坏如何评价?

Vapnik V N 教授创立的统计学习理论针对上述问题提出了一套理论成果。SVM 方法是统计学习理论的一种实用化实现,对上述问题,特别是问题(2)—(4),从算法到程序实现,都给出了有效可行的解决方案。讨论这些困难的理论问题不是本书的重点,只在第 6 章中结合统计学习理论做了简要介绍。

早期的机器学习研究几乎完全局限在人工智能这一狭窄领域中；现在，机器学习已经开始走进不同学科的不同领域，成为一种具有通用性的支撑和服务技术。而目前，SVM 方法仍然在所有的机器学习方法中具有明显的优势。

很多实际应用问题本质上都可以归结为上述机器学习问题。如通常所说的聚类分析、模式识别、回归分析、密度函数估计、人工神经网络等，都可以看成是这里所说的机器学习的特例。比如线性回归分析，就是在线性函数类中采用最小二乘法选取与样本点偏差平方和最小的线性函数。

1.3 SVM 方法的基本思想

模式识别、回归分析、密度函数估计是机器学习的三个基本内容。学者们应用概率统计和函数逼近等方法取得了很多研究成果，特别是在处理线性问题上，有些甚至可以说是完美的。但对于本质上是非线性的问题还缺少较好的结果。

当我们面对数据处理而又缺乏理论模型时，统计分析方法是最先被采用的方法。然而传统的统计方法只有在样本数量趋于无穷大时才能有理论上的保证。而在实际应用中样本数目通常都是有限的，甚至是小样本的，基于大数定律的传统统计方法对此难以取得理想的效果。

Vapnik 等人提出的统计学习理论是一种专门的小样本理论，它避免了人工神经网等方法出现网络结构难于确定、过学习、欠学习，以及局部极小等问题，被认为是目前针对解决小样本的分类、回归等问题的最佳理论。这一方法数学推导严密，理论基础坚实。基于这一理论近年提出的 SVM 方法，为解决非线性问题提供了一条新思路。

早在 20 世纪 60 年代，Vapnik 就已奠定了统计学习的基本理论基础，如经验风险最小化原则下统计学习一致性的条件（收敛性、收敛的可控性、收敛与概率测度定义的无关性，号称机器学习理论的“三个里程碑”）、关于统计学习方法推广性的界的理论，以及在此基础上建立的小样本归纳推理原则等。

直到 20 世纪 90 年代中后期，能够实现统计学习理论和原则的实用化算法——SVM 方法才逐渐被完整地提出，并且在模式识别等人工智能领域得到成功应用，受到了广泛关注。

SVM 方法的基本思想是：定义最优线性超平面，并把寻找最优线性超平面的算法归结为求解一个最优化（凸规划）问题。进而基于 Mercer① 核展开定理（Mercer 1909），通过非线性映射 φ，把样本空间映射到一个高维乃至于无穷维的特征空间

① Mercer，著名数学家，这里指的是他在 1909 年提出的一个重要定理，详见第 3 章 3.5 节。

(Hilbert① 空间),使在特征空间中可以应用线性学习机的方法解决样本空间中的高度非线性分类和回归等问题。简单地说就是实现升维和线性化。

降维(即把样本空间向低维空间做投影)是人们处理复杂问题常用的简化方法之一,这样做可以降低计算的复杂性。而升维映射,即把样本向高维空间做映射,大大增加样本的维数。通常这样做必然会增加计算的复杂性,甚至会引起"维数灾",因而人们很少问津。但是对于分类、回归等问题来说,很可能在低维样本空间无法进行线性处理的样本集,在高维特征空间却可以通过一个线性超平面实现线性划分(或回归)。

图 1.6 给出一个在二维空间中无法线性划分但映射到三维空间却可以线性划分的例子。映射是这样的:二维空间的样本点(x_1, x_2)映射成三维空间的点$(x_1^2, \sqrt{2}x_1x_2, x_2^2)$。两类样本点的在二维空间是线性不可分的,但在三维空间中却可以用一个平面把它们完全划分开。很多在低维空间中看似无法实现的"特异"事情,在高一维的空间中却是平常事。

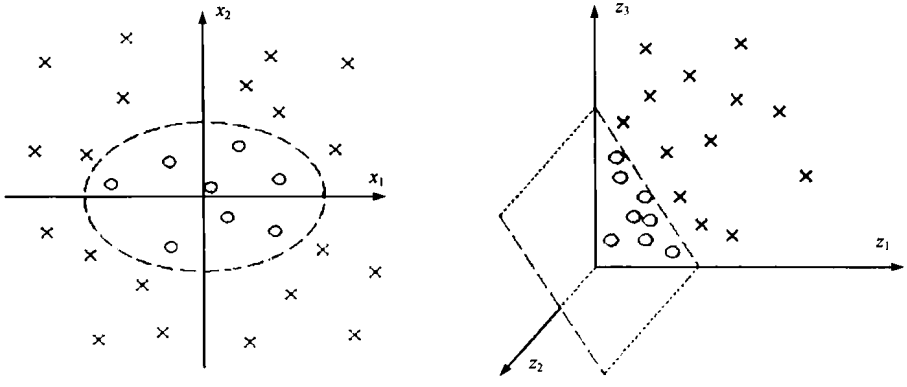


图 1.6　升维后实现线性可分的例子

SVM 的升维是通过一个非线性映射 φ,把样本空间中的点变换成一个称为特征空间的高维(甚至是无穷维)空间中的点,在特征空间中求解变换后的问题,这样做可以使原来难以解决的问题获得一个极为广阔的求解空间。在上面的例中,升维才只是增加一维:$\varphi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$,在 SVM 方法中,维数通常增加很多,有时增加到无穷。

线性化方法是人们解决复杂问题的一种常用手段。线性问题算法简单,结果确定,容易解决。很多线性问题都有很完满的解。所谓线性化方法,就是把非线性问题

① Hilbert 空间是应用最广泛的一类泛函空间,它是完备的内积空间,有关内积空间的定义详见第 3 章 3.1.3 节。

近似、转化为线性问题解决，或借助于已解决的线性问题从而解决非线性问题。

SVM 方法的线性化是先求得(在高维空间中)线性分类或回归的解决办法，然后把该方法应用在变换后的高维特征空间中。在高维特征空间中得到的是问题的线性解，但与之相对应的却是原来样本空间中问题所寻找的非线性解。这样就借助解线性问题的方法解决了原来的非线性问题。图 1.7 对此给出了一种直观图示。

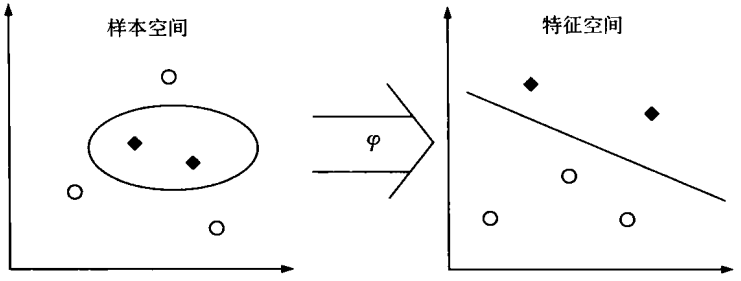


图 1.7 样本空间到特征空间的非线性映射示意图

一般的升维都会带来计算的复杂化。这里自然产生两个问题：一是如何找出非线性映射 φ 的数学表达式；二是如何简化升维带来的计算复杂性。

SVM 方法巧妙地解决了这两个难题：由于应用了核函数的展开定理，所以完全不需要知道非线性映射的显式表达式；由于是在高维特征空间中应用线性学习机的方法，所以与线性模型相比不但几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾”。这一切要归功于核的展开和计算理论。因此人们又称 SVM 方法为基于核的一种方法。对于核方法的研究是比 SVM 更为广泛和深刻的研究领域。

1.4 SVM 方法的特点和应用展望

SVM 是一种有坚实理论基础的新颖的小样本学习方法。它基本上不涉及概率测度的定义及大数定律等，因此不同于现有的统计方法。从本质上看，它避开了从归纳到演绎的传统过程，实现了高效的从训练样本到预报样本的“转导推理”(transductive inference)，大大简化了通常的分类和回归等问题。与常规的统计方法相比，SVM 方法具有如下特点：

(1) SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾”。如果说神经网络方法是对样本的所有因子加权的话，SVM 方法则对只占样本集少数的关键样本(支持向量)“加权”。当预报因子与预报对象间蕴涵的复杂非线性关系尚不清楚时，基于关键样本的方法可能优于基于因子的“加权”。

(2) 少数支持向量决定了最终结果，这不但可以帮助我们抓住关键样本、“剔除”