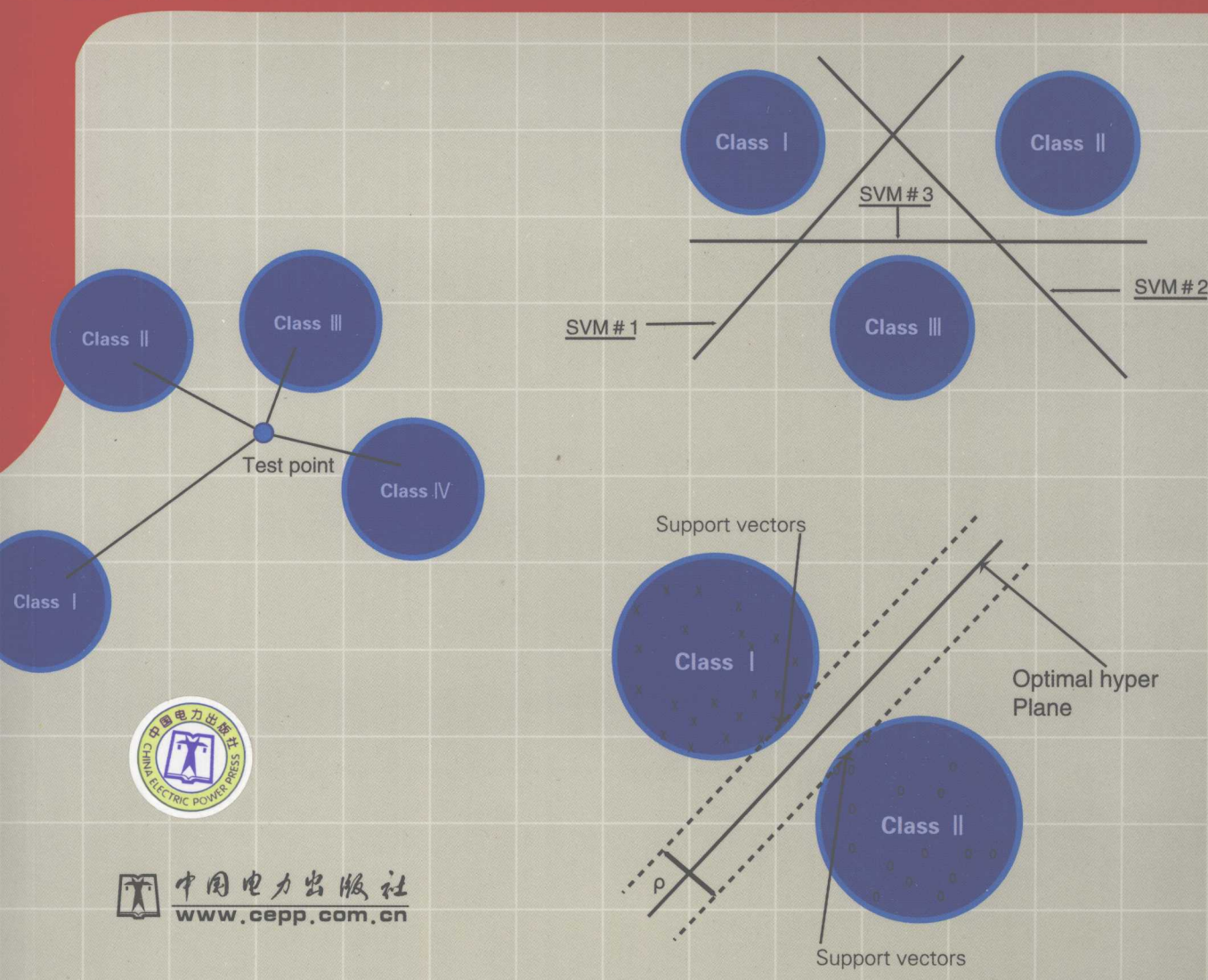
支持向量机理论及其应用分析

方瑞明 著

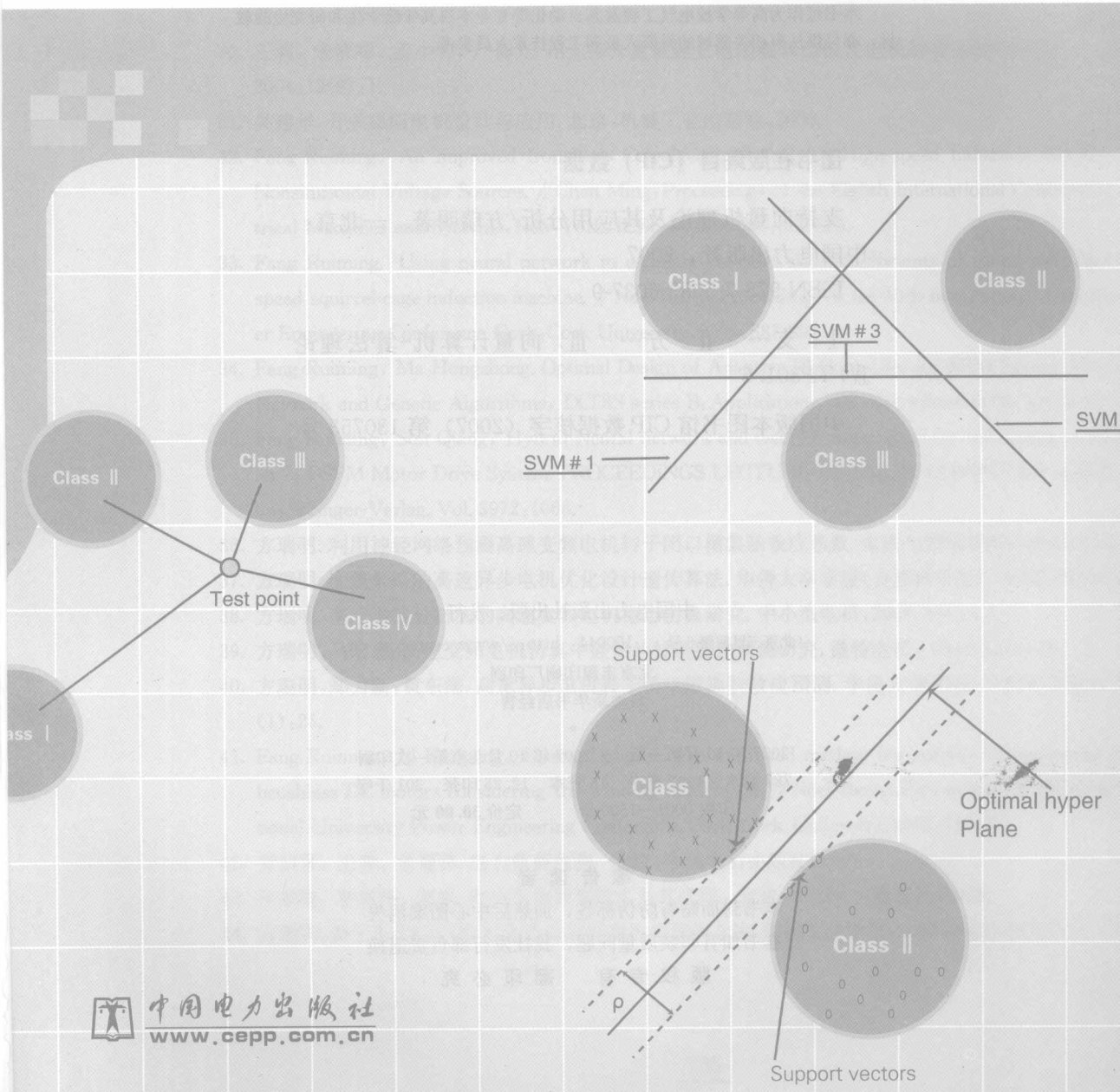


中国电力出版社
www.cepp.com.cn

华侨大学学术著作出版基金资助出版

支持向量机理论及其应用分析

方瑞明 著



中国电力出版社
www.cepp.com.cn

内 容 简 介

支持向量机是在统计学习理论基础上发展而来的一种通用学习机器。它建立在严密的统计学基础上，基于结构风险最小化准则取得实际风险，有效地提高了算法泛化能力，是处理有限样本学习的有效工具，在回归和模式识别领域具有良好的应用价值和发展前景。

本书介绍了支持向量机算法及其在电气工程领域中的应用。为了便于读者阅读和解决实际问题，书中首先对支持向量机的基本原理、训练算法、模型选择做了系统阐述。在此基础上，重点介绍了支持向量机在异步电机、电力变压器等电气设备的故障诊断以及交流电机非线性建模、电力系统短期负荷预测等领域中的应用，具有较强的实用性。

本书可作为高等学校电气工程及其自动化等专业本科高年级学生和研究生的教材，亦可供从事相关领域的科研人员和工程技术人员参考。

图书在版编目（CIP）数据

支持向量机理论及其应用分析/方瑞明著. —北京：中国电力出版社，2007

ISBN 978-7-5083-6037-9

Ⅰ. 支… Ⅱ. 方… Ⅲ. 向量计算机-算法理论 Ⅳ. TP301.6

中国版本图书馆 CIP 数据核字（2007）第 130753 号

中国电力出版社出版、发行

（北京三里河路 6 号 100044 http://www.cepp.com.cn）

北京丰源印刷厂印刷

各地新华书店经售

*

2007 年 10 月第一版 2007 年 10 月北京第一次印刷

787 毫米×1092 毫米 16 开本 12.25 印张 301 千字

印数 0001—1500 册 定价 **30.00** 元

序

Contents

电气工程作为一门独立的学科始于19世纪末期，发展至今已有100余年的历史。但是电气工程领域中仍存在许多实际问题，由于其内在的复杂性而难以用准确的数学模型进行描述。任何复杂电气系统的特性均能够由其观测数据表示，因此，长期以来，基于数据的机器学习方法在电气工程领域一直受到专家和研究者的重视。

支持向量机（Support Vector Machine，SVM）就是机器学习领域中涌现出来的一种新型算法，与机器学习中的经典方法－神经网络方法相比，它具有如下特点：

（1）支持向量机由统计学习理论发展而来，专门针对有限样本情况，其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值。

（2）支持向量机通过寻求结构化风险最小，能够实现经验风险和置信范围的最小化。

（3）支持向量机算法最终将转化成为一个二次型寻优问题，从理论上说，得到的将是全局最优点，解决了在神经网络方法中无法避免的局部极值问题。其拓扑结构由支持向量决定，而后者可以通过寻优过程得出，克服了神经网络拓扑结构选择难题。

（4）支持向量机算法将实际问题通过非线性变换转换到高维的特征空间，在高维空间中构造线性决策函数来实现原空间中的非线性决策函数，能保证其具有较好的推广能力，同时它巧妙地解决了维数问题，其算法复杂度与样本维数无关。

因此，自20世纪90年代贝尔实验室的Vapnik教授首次提出支持向量机的基本理论和概念以来，在国际范围内引起研究支持向量机理论和应用的热潮，各种杂志纷纷撰文介绍SVM的内容，许多学者纷纷将SVM理论应用于不同领域，如模式识别、回归分析、自动控制、函数逼近和信号处理等。我国对支持向量机的研究虽略晚于国外同行，但近年来关于支持向量机理论和应用的研究越来越受到研究者和技术开发界的关注。相信随着支持向量机基本原理、方法和应用技巧的深入研究和发展，其应用范围将越来越广泛。但是迄今为止，国内出版市场从工程应用的角度介绍支持向量机算法的相关著作仍显匮乏，因此有必要对相关文献中的支持向量机及其应用的研究报道进行系统整理，并出版相关专著。

方瑞明博士正是基于上述设想法，并结合自己长期从事支持向量机在电气工程中的应用研究专题所积累的资料与教学科研心得，撰写成本书，并取名为《支持向量机理论及其应用分析》，力求以电气工程领域中的实际应用为背景，阐明支持向量机的基础理论、应用方法和电气工程应用成果。我相信，本书的出版将有助于普及支持向量机及其在实际工程中的推广应用。在本书即将问世之际，愿为之简介并以为序。

湖南大学　王耀南

2007年1月于岳麓山下

前 言

Contents

人工智能对于解决电气工程领域的许多实际问题，如电气设备故障诊断、复杂设备与系统非线性建模、时间序列预测等，发挥了巨大的作用，并催生了电气工程领域许多新的技术，如智能化故障诊断技术、智能化建模技术、智能化预测技术等。然而，由于所面临对象日趋呈现复杂化的趋势，获取准确、完备、有效的领域知识越来越困难。已知的领域知识大都具有证据不充分或结论不完全的特点，领域知识的分散性、随机性和模糊性的特点使之表现出很强的不确定性。另一方面，复杂系统经常处在动态变化的过程中，其特点越来越不好把握。或受制于知识获取瓶颈(如专家系统)，或受制于有限样本(如神经网络)，上述智能技术在实际应用中的效果均受到影响。

支持向量机的诞生为这些问题的解决开辟了一条新的途径。支持向量机是由 Vapnik 等在 20 世纪 90 年代提出的一种新型机器学习算法，它建立在严密的统计学习理论基础之上，基于结构风险最小化准则取得实际风险，有效地提高了算法泛化能力，较好地解决了以往许多学习方法中小样本、非线性和高维数等实际难题，并克服了神经网络等学习方法中网络结构难以确定、收敛速度慢、局部极小点、过学习与欠学习以及训练时需要大量数据样本等不足。因此，支持向量机技术为电气工程领域智能化提供了一种新的解决方法。

目前已有一些研究者开始尝试将支持向量机引入电气工程领域，并取得了较好的效果，相信随着支持向量机基本原理、方法和应用技巧的深入研究和发展，其应用范围将越来越广泛。但是迄今为止，国内出版市场从工程应用的角度介绍支持向量机算法的相关著作都非常匮乏，更毋论电气工程领域。为促进支持向量机算法在电气工程领域中的进一步推广应用，作者以支持向量机及其在电气工程领域应用的研究成果为基础，并引用了国内外的一些最新研究成果，系统阐述支持向量机算法的基本原理、训练方法以及模型选择等，并全面介绍支持向量机算法在电气装置故障诊断、电气装置非线性建模以及电力负荷预测等领域中的应用方法。本书反映了支持向量机在电气工程领域应用的最新动态。全书由两部分组成，第一部分为支持向量机方法介绍，包括第 1 章支持向量机基本原理；第 2 章支持向量机训练算法；第 3 章支持向量机模型选择。第二部分主要介绍支持向量机在电气工程领域中的实际应用，包括第 4 章支持向量机在异步电机故障诊断中的应用；第 5 章支持向量机在电力变压器油中溶解气体诊断的应用；第 6 章支持向量机在交流电机非参数建模中的应用；以及第 7 章支持向量机在短期电力负荷预测中的应用。

本书注重系统性和可读性，力求准确无误、深入浅出。可供从事电气工程领域设计和应用的工程技术人员、研究人员参考使用，对于从事支持向量机工程应用的其他领域的工程技术人员、研究人员也具有一定的参考借鉴作用。亦可作为高等院校有关专业的教师和学生作为教科书和教学参考书。

本书写作过程中，得到了东南大学胡虔生教授、湖南大学王耀南教授、河海大学马宏忠教授的悉心指导和帮助。王耀南教授仔细审阅了书稿并热情为之作序，马宏忠教授在书稿的定稿过程中提出了许多有益的建议，日本神户大学大学院自然科学研究科阿部重夫教授与作者就书中的许多具体问题进行了坦诚而有效的讨论，并慷慨地与作者分享他所收集的大量文献资料，在此向他们表示深深的谢意。

此外，本书在写作过程中参考了大量的文献资料，已尽可能地列在书后的参考文献中，但其中难免有所遗漏，特别是一些资料经反复引用已难以查实原始出处。这里特向被漏列参考文献的作者表示歉意，并向所有作者表示诚挚的谢意。

限于作者的学识水平，书中的不妥之处在所难免，恳请读者批评指正！

作　者

2007 年 1 月于日本神户港岛

符 号 表

α_i：向量 $\boldsymbol{x}_i$ 对应的拉格朗日乘子；

ξ_i：向量 $\boldsymbol{x}_i$ 对应的松弛变量；

$\boldsymbol{A}^{-1}$：矩阵 $\boldsymbol{A}$ 的逆；

$\boldsymbol{A}^{\mathrm{T}}$：矩阵 $\boldsymbol{A}$ 的转置；

B：支持向量的边界集；

b_i：第 i 超平面的偏置项；

C：边际系数；

d：多项式核函数的阶数；

$g(x)$：将 x 映射到特征空间的映射函数；

γ：径向基函数参数；

$H(x, x')$：核函数；

l：特征空间的维数；

M：训练样本数；

m：输入变量数；

n：类别数；

S：支持向量集；

U：非边界支持向量集；

$\| x \|$：矢量 x 的欧式范数；

$\boldsymbol{W}_i$：第 i 超平面的系数矢量；

$\boldsymbol{X}_i$：训练样本中属于类 i 的样本集；

$| X_i |$：样本集 X_i 的样本数；

x_i：第 i 个 m 维训练样本；

y_i：对应输入 x_i 的输出(对于模式识别问题，y_i 为 1 或 -1；函数逼近问题，则为标量输出)。

目 录

Contents

1 支持向量机的基本原理

对样本数据进行训练并寻找规律，利用这些规律对未来数据或无法观测的数据进行预测是基于机器学习的基本思想。现有机器学习方法的重要理论基础之一是统计学。传统统计学研究的内容是样本无穷大时的渐进理论，即当样本数趋于无穷大时的极限特性。然而，现实应用当中，样本数目通常是有限的，因此研究在小样本数据量下的统计学习规律是一个非常有实用价值的课题。

诞生于20世纪70年代的统计学习理论系统地研究了机器学习问题，对有限样本情况下的统计学习问题提供了一个有效的解决途径，弥补了传统统计学的不足。与传统统计学相比，统计学习理论着重研究有限样本情况下的统计规律和学习方法，在这种体系下的统计推理不仅考虑了对渐进性能的要求，而且追求得到现有信息条件下的最优解。

支持向量机就是以统计学习理论为基础的一种新型机器学习算法，它具有严格的数学理论基础、直观的几何解释和良好的泛化能力，在处理小样本学习问题上具有独到的优越性。不仅如此，与机器学习的另一主流算法—人工神经网络相比，支持向量机避免了神经网络中的局部最优解问题和拓扑结构难以确定问题，并有效地克服了“维数灾难”，因此，支持向量机已经在许多工程实际问题中获得成功的应用。

1.1 统计学习理论基础

统计学习理论是一种研究小样本估计和预测的理论。它从理论上系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险的关系以及如何利用这些理论找到新的学习原则和方法。支持向量机算法就是在此基础上发展而来的一种通用学习方法。因此在介绍支持向量机的算法之前，有必要先介绍统计学习理论的重要概念和内容。

1.1.1 VC维

统计学习理论是关于小样本进行归纳学习的理论，其中一个重要的概念是VC维，它由俄罗斯数学家Vapnik和Chervonenkis在1960～1990年间提出并完善，故简称VC维。VC维是统计学习理论的核心概念，它是目前为止对函数集学习性能的最好描述指标。

VC维概念是建立在点集被“打散”的概念基础上。下面先介绍点集“打散”的概念。

定义1.1 设F是一个假设集，即由在$X \subset R^n$上取值为-1或$+1$的若干函数组成的集合。记$Z_m=\{x_1, \cdots, x_m\}$为X中的m个点组成的集合。考虑当f取遍F中的所有可能的假设时产生的m维向量$(f(x_1), \cdots, f(x_m))$。定义$N(F, Z_m)$为上述m维向量中不同的向量的个数。

定义1.2 设F是一个假设集，$Z_m=\{x_1, \cdots, x_m\}$为X中的m个点组成的集合，若$N(F, Z_m)=2^m$，称Z_m被F打散，或F打散Z_m。

定义1.3 (增长函数)增长函数$N(F, m)$定义为

$$N(F,m) = \max\{N(F,Z_m): Z_m \subset X\} \tag{1-1}$$

其中，$Z_m=\{x_1,\cdots,x_m\}$为 X 中的 m 个点组成的集合，$\max\{\cdot\}$是对整个 X 中点而言的。

定义 1.4 （VC 维）假设集 F 是一个由 X 上取值为+1 或−1 的函数值组成的集合。定义 F 的 VC 维为

$$VC\dim(F)=\max\{m:N(F,m)=2^m\} \tag{1-2}$$

当$\{m:\ N(F,\ m)=2^m\}$是一个无限集合时，定义 $VC\dim(F)=\infty$。

以二分类问题为例，m 是运用学习机器的函数集将点集以 2^m 种方法划分为两类的最大的点数目，即对于每个可能的划分，在此函数集中均存在一个函数 f_β，使得此函数对其中一个类取+1，另外一个类取−1。假设取在 2 维实平面的 3 个点，它们为“+”和“o”点的组合，如图 1-1 所示，若分别用“R”，“B”，“P” 3 个字符来表示，3 个点最多可以存在 2^3 种划分即(RP，B)(RB，P)(PB，R)(RPB，)(B，RP)(P，RB)(R，PB)(，RPB)；其中二元组的第一项指示的是+1 类，第二项指示的是−1 类。对于任意一个划分，均可以在函数集中找到一个有向线对应之。另外，该函数集合无法划分 2 维平面中任意 4 个点。因此，函数集合的 VC 维为 3。若对任意数目的样本都有函数能将它们打散，则函数集的 VC 维是无穷大。有界实函数的 VC 维可以通过用一定的阈值将它转化成指示函数来定义。

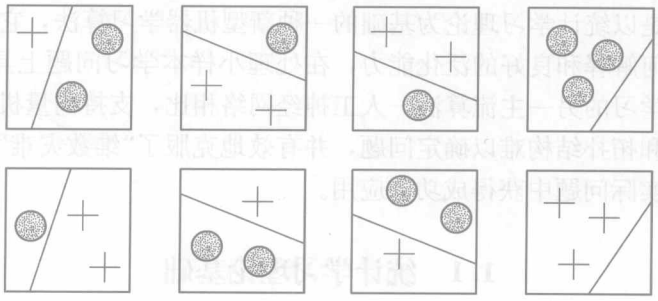


图 1-1　Z_3 被 F 打散示意图

因此，VC 维主要描述了组成学习模型的函数集合的容量，VC 维越大，函数集合越大，其相应的学习能力就越强。

1.1.2　经验风险最小化准则

一般地，学习问题可以表示为 y 与 x 之间存在的未知依赖关系，即遵循某一未知的联合概率 $F(x,\ y)$。机器学习问题就是根据 l 个独立的、相同分布的观测样本$(x_1,\ y_1)$，$(x_2,\ y_2)$，…，$(x_l,\ y_l)$，在一组函数$\{f(x,\ \omega)\}$中，求一个最优的函数 $f(x,\ \omega_0)$，对依赖关系进行估计，使期望风险

$$R(\omega)=\int L[y,f(x,\omega)]\mathrm{d}F(x,y) \tag{1-3}$$

最小。其中，$\{f(x,\ \omega)\}$称作预测函数集，ω 为函数的广义定义参数。$L[y,\ f(x,\ \omega)]$为用$\{f(x,\ \omega)\}$对 y 预测造成的损失。

然而，对于未知的概率分布 $F(x,\ y)$，要使得期望风险最小化，我们只有样本信息可以利用。这导致式(1-3)定义的期望风险是无法直接计算和最小化的。因此，传统的机器学习方法中，采用了经验风险最小化(Empirical Risk Minimization，ERM)准则，即采用样本定义经验风险

$$R_{emp}(\omega)=\frac{1}{l}\sum_{i=1}^{l}L[y_i,f(x,\omega)] \tag{1-4}$$

来逼近式(1-3)定义的期望风险。

从经验风险最小化到期望风险最小化，并没有可靠的理论依据，事实上，两者之间存在着一个学习的一致性问题。

定义 1.5 对于指示函数集 $L(y, x)$和概率分布函数 $F(y)$，如果下面两个序列概率地收敛到同一极限(如图 1-2 所示)，则称为经验风险最小一致性(ERM)。

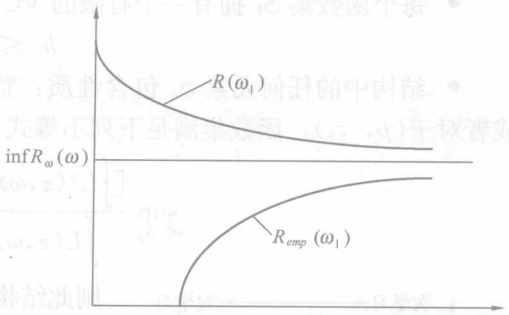


图 1-2 经验风险最小一致性

$$R(\omega_l) \xrightarrow[l\to\infty]{P} \inf_{\omega\in\Lambda} R(\omega) \tag{1-5}$$

$$R_{emp}(\omega_l) \xrightarrow[l\to\infty]{P} \inf_{\omega\in\Lambda} R(\omega) \tag{1-6}$$

式(1-5)判定期望风险 $R(\omega)$收敛到最好的可能值。式(1-6)使人们可以根据经验风险 $R_{emp}(\omega)$判定估计风险可能的最小值。

由定义可以得知，如果经验风险最小化方法是一致性，那么它必须提供一个函数序列 $L(y, x_i)$，$i=1, 2, \cdots, l$，使得期望风险和经验风险收敛到一个可能最小的风险值。

1.1.3 结构风险最小化准则

统计学习理论系统地研究了对于各种类型的函数集、经验风险和实际风险之间的关系，即泛化的界限。关于两类分类问题有如下结论：对指示函数集中所有函数(包括使经验风险最小的函数)，经验风险 $R_{emp}(\omega)$和实际风险 $R(\omega)$之间以至少$(1-\eta)$的概率满足如下的关系：

$$R(\omega) \leqslant R_{emp}(\omega) + \sqrt{\frac{h(\ln(2l/h)+1) - \ln(\eta/4)}{l}} \tag{1-7}$$

其中，h 是函数集的 VC 维，l 是样本数，η 是满足 $0\leqslant\eta\leqslant 1$ 的参数。

由式(1-7)可见，统计学习的实际风险 $R(\omega)$由两部分组成：一是经验风险 $R_{emp}(\omega)$，另一部分称为置信界限。置信界限反映了真实风险和经验风险差值的上界，即结构复杂所带来的风险，它和学习机器的 VC 维 h 及训练样本数 l 有关。

因此，传统机器学习方法中普遍采用的经验风险最小化原则在样本数据有限时是不合理的。如果要求学习风险最小，就需要不等式右边的两项相互权衡，共同趋于极小。另外，在获得的学习模型经验风险最小的同时，希望学习模型的推广能力尽可能大，这样就需要 h 值尽可能小，即，置信风险尽可能小。

根据风险估计公式(1-7)，如果固定训练样本数目 l 的大小，则，控制风险 $R(\omega)$的参量有两个：$R_{emp}(\omega)$与 VC 维 h。其中，经验风险依赖于学习机器所选定的函数 $f(x, \omega)$，这样，我们可以通过控制 ω 来控制经验风险。VC 维 h 依赖于学习机器所工作的函数集合。

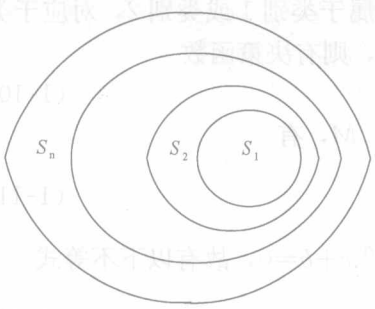


图 1-3 可同伦结构函数集

为了获得对 h 的控制，可以将函数集合结构化，建立 h 与各函数子结构之间的关系，通过控制对函数结构的选择来达到控制 VC 维 h 的目的。

定义 1.6 如图 1-3 所示，嵌套函数子集 $S_k=\{L(z, \omega), \omega\in\Lambda_k\}$，$S_1\subset S_2\subset\cdots\subset S_n$，若其中结构元素满足以

下性质：

- 每个函数集 S_k 拥有一个有限的 VC 维 h_k

$$h_1 \leqslant h_2 \cdots \leqslant h_n \tag{1-8}$$

- 结构中的任何元素 S_k 包含性质：整个界限函数集满足：$0 \leqslant L(z, \omega) \leqslant B_k$；$\omega \in \Lambda_k$ 或者对于(p, τ_k)，函数集满足下列不等式

$$\sup_{\omega \in \Lambda_k} \frac{\left[\int L^p(z,\omega)\mathrm{d}F(z)\right]^{\frac{1}{p}}}{\int L(z,\omega)\mathrm{d}F(z)} \leqslant \tau_k, p > 2 \tag{1-9}$$

则此结构称为可同伦结构。

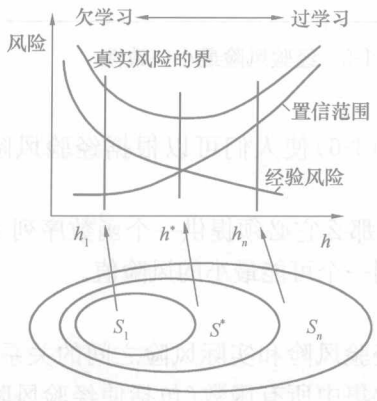


图 1-4　结构风险最小化原理图

定义 1.7(结构风险最小化原则)　所谓结构最小化准则，就是在可同伦结构的嵌套函数集 S_k 中，寻找一个中间子集 S^*，使得式(1-7)右端的结构风险达到最小。

如图 1-4 所示，学习风险由两个部分组成，即经验风险和置信范围。如果给定样本数目 l，那么，随着 VC 维数目 h 的增加，经验风险逐渐变小，而置信范围逐渐递增。真实风险的界是经验风险和置信范围之和，随着结构元素序号的增加，经验风险将减小，而置信范围将增加。最小的真实风险的上界是在结构的某个适当的元素上取得的。综合考虑经验风险与置信区的变化，可以求得最小的风险边界，它所对应的函数集的中间子集 S^* 可以作为具有最佳泛化能力的函数集合。

1.2　支持向量分类机

支持向量机是在统计学习理论的 VC 维理论和结构风险最小化原理的基础上发展起来的一种新的机器学习方法。它具有理论严密、适应性强、全局优化、训练效率高和泛化性能好等优点，它能非常成功地处理模式识别(分类、判别分析)和回归问题(时间序列分析)等诸多问题，并可进一步推广到预测和综合评价等领域，在理论和应用方面均有着光明的前景。本节首先介绍支持向量分类机。

1.2.1　硬间隔支持向量分类机

设 M 个 m 维训练样本输入数据 $x_i(i=1, 2, \cdots, M)$分属于类别 1 或类别 2，对应于类别 1，记 $y_i=1$；否则记 $y_i=-1$。若这些数据是线性可分的，则有决策函数

$$D(x) = \boldsymbol{W}^{\mathrm{T}}x + b \tag{1-10}$$

其中，$\boldsymbol{W}$ 是一个 m 维矢量，b 为偏置项。对于 $i=1, 2, \cdots, M$，有

$$\boldsymbol{W}^{\mathrm{T}}x_i + b \begin{cases} > 0, y_i = 1 \\ < 0, y_i = -1 \end{cases} \tag{1-11}$$

由于训练数据皆线性可分，不可能存在训练数据使得 $\boldsymbol{W}^{\mathrm{T}}x+b=0$，故有以下不等式

$$\boldsymbol{W}^{\mathrm{T}}x_i + b \begin{cases} \geqslant 1, & y_i = 1 \\ \leqslant -1, & y_i = -1 \end{cases} \tag{1-12}$$

式(1-12)中的1和−1也可以是一个常数$a(>0)$和$-a$。但不等式两端同除以常数a，仍可得到式(1-12)。式(1-12)可以等效为

$$y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)\geqslant 1,i=1,2,\cdots,M \tag{1-13}$$

超平面$D(x)=\boldsymbol{W}^{\mathrm{T}}x+b=c$，$-1<c<1$形成了一个对$x_i(i=1,2,\cdots,M)$的分类超平面。当$c=0$时，分类超平面位于$c=1$和$c=-1$两个超平面中间。分类超平面与最近的训练数据之间的距离称作间隔。

如果超平面$D(x)=1$和$D(x)=-1$都包含至少一个训练数据，则对于$-1<c<1$，$D(x)=0$具有最大间隔。区域$\{x\mid -1\leqslant D(x)\leqslant 1\}$称作决策函数的泛化区域。

图1-5中为满足式(1-13)的两个决策函数。事实上满足式(1-13)的决策函数可以有无数个，其泛化能力取决于分类超平面的位置。其中，具有最大间隔即为最优分类超平面，最优分类超平面的泛化能力最好。

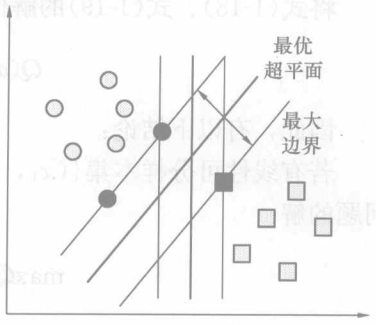


图1-5 二维空间的优化分类超平面

下面研究如何求取最优分类超平面。定义训练数据x距分类超平面的欧式距离为$|D(x)|/\|W\|$。则所有的训练数据均满足

$$\frac{y_kD(x_k)}{\|W\|}\geqslant\delta,k=1,2,\cdots,M \tag{1-14}$$

式中，δ为间隔。

如果(W,b)是决策函数的一个解，a是一个标量，则(aW,ab)也是解之一。因此，可以施加下列约束条件

$$\delta\|W\|=1 \tag{1-15}$$

由式(1-14)和式(1-15)可知，要求解最优超平面，需要找到满足式(1-13)的W的最小欧式范数。因此，优化超平面的求解转换为以下优化问题

$$\min Q(W)=\frac{1}{2}\|W\|^2$$

$$y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)\geqslant 1\qquad i=1,2,\cdots,M \tag{1-16}$$

上述优化问题是一个带不等式约束的二次规划目标函数问题。因此，能够保证收敛于全局最优解。这也是支持向量机优于神经网络算法的一个重要特点，后者很难克服局部极小点的问题。

由于删除严格满足式(1-16)中不等式约束的所有数据，仍可求得相同的最优解(最优超平面)，因此，定义满足式(1-16)中等式约束的训练数据为支持向量(图1-5中的实体点)。

式(1-16)中的优化变量为(W,b)，因此变量总数为输入变量数加1：$m+1$。当输入变量数较少时，可以直接通过二次规划方法求解式(1-16)。当输入变量的维数较高时，可以将式(1-16)转化为一个等效对偶问题，此时变量数等于训练样本数。

对这种不等式约束的最优问题，考虑在其Lagrange乘子空间求解，Lagrange乘子空间也常称为对偶空间。首先，需要找到Lagrange函数的鞍点，将式(1-16)中的约束问题转化为一个无约束问题

$$Q(W,b,\alpha)=\frac{1}{2}\boldsymbol{W}^{\mathrm{T}}W-\sum_{i=1}^{M}\alpha_i\{y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)-1\} \tag{1-17}$$

式中，$\alpha_i \geqslant 0$ 是非负拉格朗日乘子。

上式求极值时，对各变量的偏导为零，即式(1-17)当且仅当以下条件成立时

$$\frac{\partial Q(W,b,\alpha)}{\partial W} = W - \sum_{i=1}^{M} \alpha_i y_i x_i = 0 \tag{1-18}$$

$$\frac{\partial Q(W,b,\alpha)}{\partial b} = \sum_{i=1}^{M} \alpha_i y_i = 0 \tag{1-19}$$

同时，它还应满足 Karush-Kuhn-Tucker(KKT)互补条件

$$\alpha_i [y_i (\boldsymbol{W}^{\mathrm{T}} x_i + b) - 1] = 0, i = 1,2,\cdots,M \tag{1-20}$$

将式(1-18)、式(1-19)的解代入式(1-17)，得到

$$Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} x_j \tag{1-21}$$

因此，有以下结论：

若有线性可分样本集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_M, y_M)\}$，参数 α^* 是以下二次优化问题的解

$$\max Q(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} x_i$$

$$\text{约束条件为} \begin{cases} \sum_{i=1}^{M} y_i \alpha_i = 0 \\ \alpha_i \geqslant 0 \quad i = 1,2,\cdots,M \end{cases} \tag{1-22}$$

则权重向量 $W^* = \sum_{i=1}^{M} \alpha_i^* y_i x_i$ 决定了最优超平面。$\alpha_i > 0$ 所对应的训练数据即为类 1 或类 2 的支持向量。此时最佳分类决策函数为

$$D(x) = \sum_{i \in S} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} x + b \tag{1-23}$$

b 的值没有出现在对偶问题中，利用原始约束可以求得

$$b = y_i - \boldsymbol{W}^{\mathrm{T}} x_i \tag{1-24}$$

式中 x_i 为支持向量。从提高计算精度的角度出发，可以取平均值，此时

$$b = \frac{1}{|S|} \sum_{i \in S} (y_i - \boldsymbol{W}^{\mathrm{T}} x_i) \tag{1-25}$$

一旦求出最优分类决策函数，则未知数据 x 分类结果为

$$\begin{cases} \text{类 1} & D(x) > 0 \\ \text{类 2} & D(x) < 0 \end{cases} \tag{1-26}$$

若 $D(x) = 0$，则 x 位于边界上，不可分。若训练结果可以接受，此时区域$\{x \mid 1 > D(x) > -1\}$就是泛化区域。

1.2.2 软间隔支持向量分类机

硬间隔支持向量分类机是一个重要概念，它是分析和构造更加复杂的支持向量机的起点。硬间隔支持向量分类机的主要问题是它总是试图完美地产生一个没有训练误差的一致假设。然而，当训练数据有噪声时，特征空间一般不能线性分开。因此，我们需要将支持向量机扩展到线性不可分情况。

为了允许不可分性，在式(1-13)中引入非负松弛变量 ξ_i：

$$y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i, i=1,2,\cdots,M \tag{1-27}$$

由于引入了松弛变量 ξ_i，可行解总是存在的。如图 1-6 所示，对于训练数据 x_i，由于 $0<\xi_i<1$，虽然该数据不具有最大间隔，仍可正确分类；而于训练数据 x_j，由于 $\xi_j\geqslant 1$，该数据被最优超平面错误地分类了。

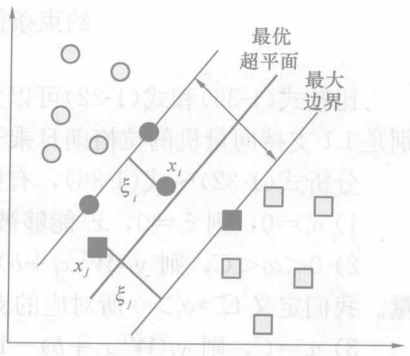


图 1-6 二维空间的不可分情况

为了得到没有最大间隔的训练数据数目最少的最优超平面，可以把下式最小化，即

$$Q(W)=\sum_{i=1}^{M}\theta(\xi_i),\theta(\xi_i)=\begin{cases}1,\xi_i>0\\0,\xi_i=0\end{cases}$$

然而，这是一个组合优化问题，很难求解。我们用式(1-28)代替

$$\min Q(W,b,\xi)=\frac{1}{2}\|W\|^2+C\sum_{i=1}^{M}\xi_i^P \tag{1-28}$$

约束条件为 $y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i \quad i=1,\cdots,M$

式中，$\xi=(\xi_1,\cdots,\xi_M)^{\mathrm{T}}$，$C$ 为边际系数，它用来平衡最大间隔和最小分类误差。事实上，参数 C 是一个变化范围很大的值，它的选择对支持向量机的性能具有很大的影响，后文将予以专门讨论。P 为范数，若选取 $P=1$，所得为 1 范数支持向量机(L1SVM)；$P=2$，所得为 2 范数支持向量机(L2SVM)。

下面以 1 范数情况为例继续讨论。

与线性可分情况类似，引入非负拉格朗日乘子 α_i 与 β_i，可得

$$\begin{aligned}Q(W,b,\xi,\alpha,\beta)=&\frac{1}{2}\|W\|^2+C\sum_{i=1}^{M}\xi_i\\&-\sum_{i=1}^{M}\alpha_i[y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)-1+\xi_i]-\sum_{i=1}^{M}\beta_i\xi_i\end{aligned} \tag{1-29}$$

式中，$\alpha=(\alpha_1,\cdots,\alpha_M)^{\mathrm{T}}$，$\beta=(\beta_1,\cdots,\beta_M)^{\mathrm{T}}$。

求解最优解，令上式对各变量的偏导为零，得

$$\frac{\partial Q(W,b,\xi,\alpha,\beta)}{\partial W}=W-\sum_{i=1}^{M}\alpha_i y_i x_i=0 \tag{1-30}$$

$$\frac{\partial Q(W,b,\xi,\alpha,\beta)}{\partial b}=\sum_{i=1}^{M}\alpha_i y_i=0 \tag{1-31}$$

$$\frac{\partial Q(W,b,\xi,\alpha,\beta)}{\partial \xi}=C-\alpha_i-\beta_i=0 \tag{1-32}$$

同时，它还应满足下列 KKT 互补条件

$$\alpha_i[y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)-1+\xi_i]=0, i=1,\cdots,M \tag{1-33}$$

$$\beta_i\xi_i=0, i=1,\cdots,M \tag{1-34}$$

$$\alpha_i\geqslant 0,\beta_i\geqslant 0,\xi_i\geqslant 0, i=1,\cdots,M \tag{1-35}$$

将式(1-30)、式(1-31)、式(1-32)代入式(1-29)，即可获得对应对偶问题。最大化：

$$\max Q(\alpha)=\sum_{i=1}^{M}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{M}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}x_j$$

$$\text{约束条件为}\begin{cases}\sum_{i=1}^{M} y_i\alpha_i = 0 \\ C \geqslant \alpha_i \geqslant 0 \quad i = 1,\cdots,M\end{cases} \tag{1-36}$$

比较式(1-36)和式(1-22)可以发现，硬间隔支持向量分类机和 L1 支持向量机的唯一区别是 L1 支持向量机的拉格朗日乘子不能超过边际系数 C。

分析式(1-32)～式(1-36)，有如下发现：

1) $\alpha_i=0$，则 $\xi_i=0$，x_i 能够被正确分类。

2) $0<\alpha_i<C$，则 $y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)-1+\xi_i=0$，$\xi_i=0$，此时 $y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)=1$，$x_i$ 为支持向量。我们定义 $C>\alpha_i>0$ 所对应的支持向量为非边界支持向量。

3) $\alpha_i=C$，则 $y_i(\boldsymbol{W}^{\mathrm{T}}x_i+b)-1+\xi_i=0$，$\xi_i\geqslant 0$，此时 x_i 为支持向量。我们定义 $\alpha_i=C$ 所对应的支持向量为边界支持向量。

若 $0\leqslant\xi_i\leqslant 1$，则 x_i 能够被正确分类；$\xi_i\geqslant 1$，x_i 被错分。

最佳分类决策函数与硬间隔支持向量机相同

$$D(x) = \sum_{i\in S}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}x + b \tag{1-37}$$

式中 S 为支持向量集。从提高计算精度的角度出发，b 取如下平均值

$$b = \frac{1}{|U|}\sum_{i\in U}(y_i - \boldsymbol{W}^{\mathrm{T}}x_i) \tag{1-38}$$

式中 U 为所有非边界支持向量集。

未知数据 x 分类结果为

$$\begin{cases}\text{类 1 if} D(x) > 0 \\ \text{类 2 if} D(x) < 0\end{cases} \tag{1-39}$$

若 $D(x)=0$，则 x 位于边界上，不可分。如果没有边界支持向量，分类区域为 $\{x \mid 1>D(x)>-1\}$就是泛化区域，与硬间隔支持向量分类机相同。

1.2.3 特征空间映射与核函数

支持向量机采用最优超平面决定最大泛化能力。但是，当训练数据不是线性可分时，尽管分类超平面已经优化，所获得的分类器的泛化性能也可能较低。因此，为加强线性可分性，可以将原来的输入空间映射到一个高维点积空间，即特征空间。

如果非线性矢量函数 $g(x)=[g_1(x),\cdots,g_l(x)]$将 m 维输入矢量 x 映射到 l 维特征空间，则特征空间的线性决策函数为

$$D(x)=\boldsymbol{W}^{\mathrm{T}}g(x)+b \tag{1-40}$$

1.2.3.1 Mercer 条件

下面研究映射函数的性质，首先介绍 Mercer 条件。

根据 Hilbert-Schmidt 定理，如果一个对称函数 $H(x,x')$满足

$$H(x,x')=\sum_{i,j=1}^{M} h_i h_j H(h_i,h_j)\geqslant 0 \tag{1-41}$$

式中，M 为自然数，h_i，h_j 为实数。

则存在一个映射函数 $g(x)$，能够将 x 映射到点积特征空间。该映射函数满足

$$H(x,x')=\boldsymbol{g}^{\mathrm{T}}(x)g(x') \tag{1-42}$$

如果式(1-41)成立，则