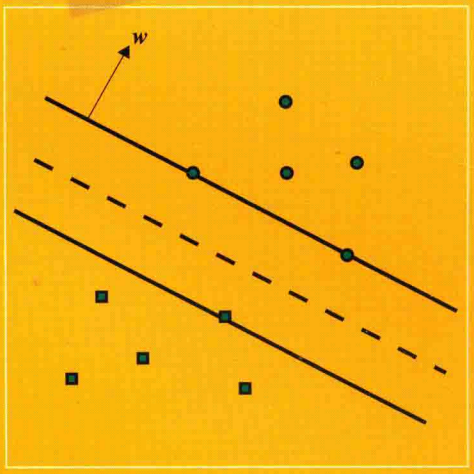
支持向量机数据扰动分析

蔡 春 著



清华大学出版社

支持向量机数据扰动分析

蔡 春 著

清华大学出版社
北 京

内 容 简 介

本书的内容包括支持向量机概述、支持向量分类机模型、加权支持向量分类机算法、线性支持向量分类机数据扰动分析、非线性支持向量分类机数据扰动分析、线性支持向量回归机的数据扰动分析。

本书内容丰富，深入浅出。为使数学基础不同的读者都能较好地对本门知识建立起概貌，结合自己的领域实际应用该门知识，本书特别重视的是：结合简单、典型的实例，讲清楚支持向量分类机数据扰动分析理论的产生背景，系统论述了支持向量分类机数据扰动分析体系。本书不仅可作为理工科人工智能方面研究生的扩充资料，也可供数学基础较强但对本方面知识有强烈学习愿望的其他各类读者自学之用，还可作为有关专业教师和科研人员的参考书。

图书在版编目(CIP)数据

支持向量机数据扰动分析/蔡春著. —北京：清华大学出版社，2019
ISBN 978-7-302-52598-1

Ⅰ. ①支…　Ⅱ. ①蔡…　Ⅲ. ①向量计算机－数据分析　Ⅳ. ①TP381.6

中国版本图书馆 CIP 数据核字(2019)第 044620 号

责任编辑：刘　颖
封面设计：傅瑞学
责任校对：刘玉霞
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市国英印务有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　　**印　张**：7.25　　**字　　数**：136 千字
版　　次：2019 年 4 月第 1 版　　**印　　次**：2019 年 4 月第1次印刷
定　　价：29.00 元

产品编号：063036-01

前　言

支持向量机(support vector machines,SVM)最初是20世纪90年代由万普尼克(Vapnik)提出。万普尼克等人在20世纪60年代开始研究有限样本情况下的机器学习问题，提出统计学习理论(statistical learning theory,SLT)，支持向量机就是在统计学习理论框架下发展起来的，其理论研究和应用方面都取得了突破性进展，开始成为数据挖掘的一种新技术，而且是一种很重要的新技术。

解决分类问题的支持向量机模型称为支持向量分类(support vector classification,SVC)或支持向量分类机，解决回归问题的支持向量机模型称为支持向量回归(support vector regression,SVR)或支持向量回归机。支持向量分类机在统计学习理论这一理论框架下产生，在应用中表现出令人满意的结果，它已初步表现出很多优于已有方法的性能，成为一种新的通用机器学习方法。利用支持向量分类机构造出的分类器可以自动寻找那些对分类有较好区分能力的支持向量、最大化两类样本点的间隔，因而支持向量分类机有较好的推广性能和较高的分类准确率，在解决小样本机器学习问题中表现出特有的优势，开始成为克服“维数灾难”和“过学习”等传统困难的有力手段。SVC正在成为继人工神经网络(artificial neural network,ANN)研究之后新的研究热点，并将有力地推动机器学习理论和技术的发展。

对于分类问题有两类：一类是线性可分问题，另一类是线性不可分问题。对于线性可分问题，支持向量分类机的基本思想就是最大化两类“间隔”，据此构造最优化模型，求解模型可以得到可分的线性平面；对于新的样本点的类别进行预测，就是把新样本点的数值代入所得到的线性平面，根据这个平面算出的值的正负性进行类别判断。对于线性不可分问题，理论上利用一个映射把原来的输入空间$\mathbb{R}^n$映射到希尔伯特

(Hilbert)空间(简记为 H 空间),引入超平面的思想;而这些想法就可以通过引入核函数来实现。核函数实质是卷积,求解原问题的沃尔夫(Wolfe)对偶问题而建立起决策函数,全部操作仍是在原来的输入空间$\mathbb{R}^n$上进行,而不管上述概念中的 H 具体是什么内积空间。

本书是关于支持向量分类机及回归机数据扰动分析的导论性专著,它着重于训练数据误差对分类平面的影响方面。本书简要概述了支持向量分类机的模型,支持向量分类机决策函数阈值,重点围绕线性支持向量分类机数据扰动分析,非线性支持向量分类机数据扰动分析理论体系进行论述。本书试图自我包容,只需要具备数学最优化理论的基础知识,所需的概念在每一章中均加以给出。

本书共分 6 章:概论、支持向量分类机算法及预备知识、加权支持向量分类机算法、加权线性支持向量分类机数据扰动分析、非线性支持向量分类机数据扰动分析、线性支持向量回归机的数据扰动分析。

本书的写作受到中国农业大学理学院教授邓乃扬、北京理工大学理学院教授刘宝光、中国农业大学理学院教授陈奎孚、加拿大曼尼托巴大学统计学院教授王熙逵的大力支持,在研究的具体开展中,我的同事吕书强老师也给我提出了很好的建议,在此表示感谢。另外也以此书献给我的家人、朋友,是他们给予我很多关心和厚爱,我才有精力完成此书。此外还要感谢清华大学出版社的刘颖老师,他深厚的数学功底,精心的编辑才保证此书顺利出版。

本书的出版得到北京联合大学学术出版的资助和北京市青年拔尖人才项目的资助(项目号 CIT&TCD201404080)。在此一并感谢!

蔡春

北京联合大学

2019 年 2 月

目　录

第1章

概　论

数据挖掘源于数据库技术的发展，现在数据库可以存储海量数据，数据的快速增加与数据分析方法滞后的矛盾越来越突出，人们希望对已有的海量数据进行科学分析，得到有价值的知识，这就促使了数据挖掘的产生。数据挖掘的方法很多，经典的是统计估计方法，比如回归分析、判别分析、聚类分析等。与经典统计方法相对的是新的学习方法即机器学习方法。目前机器学习方法的主流方法是支持向量机方法。

追溯支持向量机的知识背景，就要了解另一个比较新的概念——数据挖掘[1]，数据挖掘即从大量的数据中，抽取出潜在的、有价值的知识（模型或规则）的过程。数据挖掘的任务很多，“分类”是其中一项重要的任务，即在已知类别的样本集合上（训练集）建立分类模型，求解分类模型得到决策函数，利用决策函数对未知类别的样本（待测试样本）进行分类。SVM 最初是 20 世纪 90 年代由万普尼克（Vapnik）提出[2]，近年来在其理论研究和应用方面都取得了突破性进展，开始成为数据挖掘的一种新技术，而且是一种很重要的新技术。目前关于支持向量机已经出版了许多著作和会议论文集[3~6]。它在许多领域都获得了成功的应用，如：模式识别[7~9]，回归、函数拟合[10~19]等，现也被国内推广到经济预测[20,21]、文本分类[22~25]、人脸识别[26,27]、工程应用[28~37]、医学应用[38~40]等领域，逐渐成为国内外新的研究热点。

数据的获得有多种渠道，有用仪器测量的数据如医疗数据、建筑数据，有调查问卷获得的数据如消费数据，有各个单位报表的数据如企业数据，但无论如何，数据或多或少都有部分失真，对于部分失真的数据进行分析，我们就得考虑到数据的扰动对分析方法的影响。

本章首先介绍研究背景、提出问题，其次介绍支持向量分类机的基本思想，再次介绍支持向量分类机的发展历史、研究现状，最后对本书的研究内容、结构以及结论进行概述。

1.1 从机器学习到支持向量分类机

数据挖掘的方法很多，其中机器学习是数据挖掘的一种主流方法。基于数据的机器学习问题是人类智能研究的主要问题，它通过对已知事实的分析，总结规律，预测不能直接观测的规律。在机器学习过程中，统计学起着基础性的作用，但传统的统计学所研究的主要是渐近理论，即当样本趋向于无穷多时的统计性质。而在现实的问题中，我们所面对的样本数目通常是有限的，因此一些理论上很优秀的学习方法在实际中的表现却可能不尽如人意；虽然人们实际上一直知道这一点，但传统上仍以样本数目无穷多为假设来推导各种算法，希望这样得到的算法在样本较少时也能有较好的(至少是可接受的)表现。然而，相反的情况却经常出现，人们对于解决此类问题的努力一直在进行。

万普尼克等人在20世纪60年代开始研究有限样本情况下的机器学习问题[41]，提出统计学习理论。在统计学习理论建立过程中遇到了经验风险最小化与期望风险最小化不一致的情形，为了研究机器学习过程的一致性，万普尼克和切夫耐基(Chervonenkis)于1971年[42]提出了支持向量机的重要的基础理论——VC维(Vapnik-Chervonenkis dimension)理论。VC维是描述函数集复杂性的一个指标，VC维越大学习机器越复杂，学习机器越复杂推广能力就越难把握，为此直到20世纪90年代初期，VC维理论还没有得到很好的应用[43]。到20世纪90年代中期，随着其理论的不断发展和成熟，也由于神经网络(Neural Network，NN)等学习方法在理论上缺乏实质性进展，统计学习理论开始受到越来越广泛的重视。

万普尼克[44]进一步提出了具有划时代意义的原则——结构风险最小化(structural risk minimization，SRM)原则。在此基础上，20世纪90年代万普尼克和他的At&TBell实验室小组提出了支持向量分类机方法，该方法体现了结构风险最小化[45]原则的基本思想，进一步丰富和发展了统计学习理论，使抽象的学习理论转化为通用的实际算法。

1992年，博瑟(Boser)、吉翁(Guyon)和万普尼克在文献[45]中，提出了最优间隔分类器。1993年，科特斯(Cortes)和万普尼克在文献[46]中，进一步探讨了非线性软间隔的分类问题。1995年，万普尼克在文献[47]中，完整地提出了SVM分类方法。

SVM分类方法在统计学习理论这一理论框架下产生，在应用中表现出令人满意的结果，它已初步表现出很多优于已有方法的性能，成为一种新的通用机器学习方法[48]。利用SVM分类方法构造出的分类器可以自动寻找那些对分类有较好区分能力的支持向量、最大化两类样本点的间隔[49]，因而有较好的推广性能和较高

的分类准确率，在解决小样本机器学习问题中表现出特有的优势，开始成为克服“维数灾难”和“过学习”等传统困难的有力手段。SVM 正在成为继人工神经网络研究之后新的研究热点，并将有力地推动机器学习理论和技术的发展。

支持向量机是解决数据挖掘问题之一——分类问题的一种重要方法，其通过求解沃尔夫(Wolfe)对偶问题进而构造决策函数的求解方法，已经趋于成熟(特大型问题除外)，但是关于它的输入数据的误差，以及数据各种可能的变化对决策函数值的影响的分析，在理论上和实际计算实现方面都还是空白。输入数据一般都是某些特征的测定值，它只是真值的近似，使用这些近似值建立起支持向量分类机模型，求解模型，得解 w,b，以及决策函数。我们假定所得的 w,b 是所建模型的准确解，w,b 以及由其构造的决策函数与由数据真值构造的模型相对应的那个真的 w,b 以及决策函数有近似意义吗？其近似程度同输入数据的误差有怎样的定量关系？除了上述属于支持向量分类机模型的稳定性概念分析之外，下列各项考虑也是有意义的：比如说，输入数据特别是支持向量的某一个、某一维、某一个的某一维，甚至是全部数据在发生一定变化时，模型的解 w,b 乃至决策函数如何变化？能否无须重新求解模型而通过对原模型的解的某种修正得到其近似？更深入一层，能否从对这些变化结果的分析，透视出输入数据所代表的特征，以及不同特征对于决策函数的不同贡献？所有这些都可以纳入一个统一的框架之中，这就是数据的扰动分析。数据扰动分析是指针对数据的某种微小变化对扰动后造成的模型以及解的状况作定量的分析，比如给出构成解的各个要素对数据扰动变差的变化率。这样当数据扰动的范围将数据真值含于其中时，扰动分析的结果即可从定性、定量两个方面回答上述模型的稳定性问题，当扰动代表了我们所着意研究的数据变化时，扰动分析的结果就可以给出变化后的近似解，以及哪些变化对解影响巨大或者无足轻重，等等。

一般的非线性规划理论中的实用的可计算实现的灵敏度分析方法，为上述数据扰动分析问题提供了基础的理论工具。像引用非线性规划的沃尔夫对偶理论而形成的支持向量机系统的求解方法那样，我们也想把这种灵敏度分析方法引入到支持向量机这一特定的凸二次规划模型中，充分地具体化以求形成实用的数据扰动分析方法，作为支持向量分类机理论和方法的一个扩展。

1.2 支持向量分类机思想

1.2.1 分类问题的提出

数据挖掘的一个很重要的内容就是分类，分类问题不是什么新问题，但是随着

计算机的普遍应用以及数据挖掘的迅速发展赋予了它们新的意义，再次引起人们热切关注。国内在这方面发展也很快，下面通过一个例子引出分类问题[50]。

完全确诊某些疾病，可能需要进行创伤性探测或者昂贵的手段。因此利用一些有关的容易获得的临床指标进行推断，是一项有意义的工作。美国 Cleveland Hart Disease Database 提供的数据，就是这方面工作的一个实例。在那里对 297 个待诊病人进行了彻底的临床检测，确诊了他们是否有心脏病。同时，也记录了他们的年龄、胆固醇等 13 项有关指标。他们希望根据这些临床资料对待诊病人只检测这 13 项指标，来推断该病人是否有心脏病。这类问题称为分类(classification)问题，也称为模式识别问题[51]，在概率统计中则称为判别分析问题。我们采用“分类问题”这一术语。图 1.1 给出了 2 个指标(年龄、胆固醇水平)、10 个人(5 个患者用圆圈表示、5 个健康者用方块表示)的示意图。

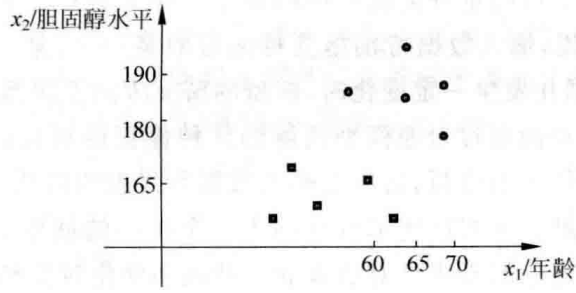


图 1.1 分类问题示意图

图 1.1 是 2 个指标 10 个样本点分两类的情形。一般地，可能有 n 个指标，即 $\boldsymbol{x}\in\mathbb{R}^n$($n$ 维列向量)，m 个样本点，记 m 个样本点的集合为 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}\in(X\times Y)^m$，其中 $\boldsymbol{x}_i\in X=\mathbb{R}^n$，$y_i\in Y=\{-1,1\}$，$\boldsymbol{x}_i\in\mathbb{R}^n$ 是输入指标向量或称为输入、或称为模式，空间 $\mathbb{R}^n$ 也因此称为输入空间，$y_i\in Y=\{-1,1\}$ 是输出指标，或称为输出，$i=1,2,\cdots,m$，这 m 个样本点组成的集合称为训练集，其中的样本点也称训练点。这时我们的问题是，给定一个新的模式 $\boldsymbol{x}$，根据训练集，寻找规则并按此规则推断它所对应的输出 y 是 1 还是 -1。分类问题用数学语言可以描述如下：

分类问题 根据给定的训练集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}\in(X\times Y)^m$，其中 $\boldsymbol{x}_i\in X=\mathbb{R}^n$，$y_i\in Y=\{-1,1\}$，$i=1,2,\cdots,m$，寻找 $\mathbb{R}^n$ 上的一个实值函数 $g(\boldsymbol{x})$，以便用决策函数 $f(\boldsymbol{x})=\operatorname{sgn} g(\boldsymbol{x})$ 推断任一模式 $\boldsymbol{x}$ 相对应的输出值。由此可见，求解分类问题实质上就是找到一个把 $\mathbb{R}^n$ 上的点分成两类的规则。

与分成两类的分类问题类似，还有分成多类的分类问题。它们的不同之处仅在于前者的输出只取两个值，而后者则可取多个值，我们这里只讨论分成两类的分

类问题。

参照机器学习领域中的术语，我们把解决上述分类问题的方法称为分类学习机。当 $g(\boldsymbol{x})$ 为线性函数 $g(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{x}+b$，由决策函数 $f(\boldsymbol{x})=\operatorname{sgn} g(\boldsymbol{x})$ 确定分类准则时，称为线性分类学习机；当 $g(\boldsymbol{x})$ 为非线性函数时，称为非线性分类学习机。分类的目的就是构造一个分类函数或分类模型（分类器），把未知类别的数据项映射到某一个给定类别。

1.2.2 分类问题的困难

分类问题的最终目标归结为构造分类函数，即在一组函数集合 $\{g(\boldsymbol{x},\boldsymbol{w})\}$ 中寻找一个最优的函数，使期望风险

$$R(\boldsymbol{w})=\int L(y,g(\boldsymbol{x},\boldsymbol{w}))\mathrm{d}F(\boldsymbol{x},y)$$

最小，其中 $\{g(\boldsymbol{x},\boldsymbol{w})\}$ 为预测分类函数集，$\boldsymbol{w}$ 为函数的参数；$L(y,g(\boldsymbol{x},\boldsymbol{w}))$ 表示用函数 $g(\boldsymbol{x},\boldsymbol{w})$ 对 y 进行预测带来的损失，我们称它为损失函数，它有不同的表达形式；$F(\boldsymbol{x},y)$ 为样本的分布。分类问题的困难在于：训练集是按照某个概率分布 $F(\boldsymbol{x},y)$ 选取的独立、同分布的样本点的集合，但概率分布函数 $F(\boldsymbol{x},y)$ 是未知的。常用的损失函数为 0-1 损失函数，其定义为

$$L(y,g(\boldsymbol{x},\boldsymbol{w}))=\begin{cases}0, & y=g(\boldsymbol{x},\boldsymbol{w}),\\ 1, & y\neq g(\boldsymbol{x},\boldsymbol{w})。\end{cases}$$

未知概率分布函数使得期望风险最小的分类学习目标无法计算，因此一种分类学习目标是使经验风险最小，即

$$R_{emp}(\boldsymbol{w})=\frac{1}{m}\sum_{i=1}^{m}L(y,g(\boldsymbol{x}_i,\boldsymbol{w}))。$$

用经验风险最小化只是体现了眼前的利益，对任意待输出模式的情况完全不知，也无法知晓，就经验风险最小化这一目标也很难实现。但这种思想却在多年的机器学习方法研究中占据了主要地位，直至发现神经网络的过学习问题，亦即追求经验风险最小导致了推广能力（对未来输出进行正确预测能力）的下降。这里可以举一个例子[52]：任意给定一组实数样本 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$y_i\in\{-1,1\}$，我们总可以选择适当的参数 α，使得函数 $\sin(\alpha x)$ 完全拟合给定的实数样本，如果样本点取得非常密集，我们知道函数 $\sin(\alpha x)$ 的图像起伏很大，这样对未来输出的正确预测能力就下降，即这个函数的推广能力下降。为此对有限样本学习提出了两个问题：一是经验风险最小并不一定意味着期望风险最小；二是函数的复杂性影响了学习机器的推广能力。为此提出了小样本情况下的机器学习理论——统计学习理论，主要研究内容分 4 个方面[2]：

(1) 基于经验风险最小化的学习过程一致性的条件；

(2) 学习过程收敛的速度；

(3) 学习过程的推广能力的界；

(4) 构造能够控制推广能力的算法。

1.2.3 支持向量分类机的基本思想

图 1.1 所示的训练集是一个二维的数据，可以用直线正确分开，这是一个二维线性可分问题。一般线性可分问题确切定义如下。

定义 1.2.1(线性可分) 考虑训练集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}\in(X\times Y)^m$，$\boldsymbol{x}_i\in X=\mathbb{R}^n$，$y_i\in Y=\{-1,1\}$，$i=1,2,\cdots,m$，若存在 $\boldsymbol{w}\in\mathbb{R}^n$，$b\in\mathbb{R}$ 和正数 ε，使得对所有 $y_i=1$ 的下标 i，有 $(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\geqslant\varepsilon$，而对所有 $y_i=-1$ 的下标 i，有 $(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\leqslant-\varepsilon$，则称训练集线性可分，同时也称相应的分类问题是线性可分问题。

给定训练集，其分类有两种情况：线性可分，线性不可分。

首先看线性可分的情形：假定训练集 $\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,m\}$ 可以被一个超平面 $(\boldsymbol{w}\cdot\boldsymbol{x})+b=0$ 正确地分开，其中 $\boldsymbol{x}_i\in\mathbb{R}^n$，标签 $y_i\in\{-1,+1\}$ 是点 $\boldsymbol{x}_i$ 的类别，支持向量分类机的基本思想就是最大化两类“间隔”，据此求出超平面的法方向 $\boldsymbol{w}$，然后利用法方向求出超平面的另一参数 b。图 1.2 给出线性可分的分类示意图。

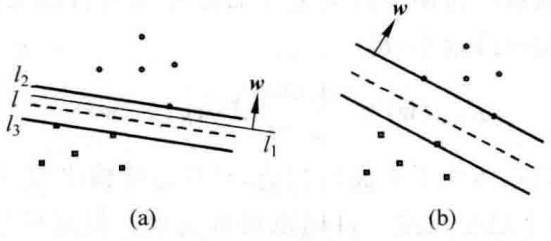


图 1.2 线性可分的分类图

从图 1.2(a)可以看出能把两类点正确分开的超平面很多，像图中的 l_1,l_2,l_3 超平面都可以将两类点分开，以及两个超平面的中间有无穷多个超平面也可以实现把两类点进行正确分类；但这两个超平面 l_2,l_3 间的距离并不是最大的，通过改变超平面的倾斜程度，可以看出图 1.2(b)给出的超平面间的距离最大。我们不难理解：数据点是不能动的，而这些超平面是可以动的，数据点离超平面的距离越大，数据模型就可以容忍数据的扰动，这样一来超平面对数据有更好的健壮性，即给出了最优超平面法方向 $\boldsymbol{w}\in\mathbb{R}^n$。进一步，对于选定的超平面的法方向 $\boldsymbol{w}\in\mathbb{R}^n$，平行移动超平面使之达到两类点的边界，选取参数 b，则中间的超平面为我们所求的

超平面。如果再调整 $\boldsymbol{w}$ 和 b 的尺度，可以把两条极端直线规范化为 $(\boldsymbol{w}\cdot\boldsymbol{x})+b=+1$ 和 $(\boldsymbol{w}\cdot\boldsymbol{x})+b=-1$。通过这种直线方程 $(\boldsymbol{w}\cdot\boldsymbol{x})+b=0$ 的规范化可知，此时这两条直线的距离即相应的“间隔”为 $\frac{2}{\|\boldsymbol{w}\|}$，可以得到最优超平面到边界超平面的距离为 $\frac{1}{\|\boldsymbol{w}\|}$（范数为 2 范数，后面不再声明），两类点所对应的边界超平面的间隔为 $\frac{2}{\|\boldsymbol{w}\|}$，为了求解方便，目标函数常常改为求最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2$。万普尼克给出了最大间隔超平面具有良好推广能力的定理[45]，根据此定理得到线性可分的支持向量分类机模型：

$$\begin{aligned}&\min_{\boldsymbol{w},b}\quad \frac{1}{2}\|\boldsymbol{w}\|^2\\&\text{s.t.}\quad y_i(\boldsymbol{x}_i\cdot\boldsymbol{w}+b)-1\geqslant 0,\quad i=1,2,\cdots,m。\end{aligned}\tag{1.1}$$

模型(1.1)是一个凸二次规划。模型的求解方法现在流行的是通过它的沃尔夫对偶问题来求解[49]，按沃尔夫对偶理论得到原问题(1.1)的沃尔夫对偶问题：

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\quad W(\boldsymbol{\alpha})=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\cdot\boldsymbol{x}_j\\&\text{s.t.}\quad \sum_{i=1}^{m}\alpha_i y_i=0,\\&\qquad \alpha_i\geqslant 0,\quad i=1,2,\cdots,m。\end{aligned}\tag{1.2}$$

照例称问题(1.1)为原问题，问题(1.2)为对偶问题。利用沃尔夫对偶问题(1.2)，不但使问题(1.1)更好处理，而且可以看出样本在对偶问题(1.2)的目标函数中仅仅以向量内积的形式出现，正是这一重要特点，使支持向量分类机方法能推广到非线性情况，这是沃尔夫对偶问题带来的一个最好副产品，现在对 SVM 的研究一般都从沃尔夫对偶问题开始，而不是直接求解原问题(1.1)。为此经常先求解沃尔夫对偶问题(1.2)，得到对偶问题的解后，利用二者之间的关系求出原问题的最优解，得到分类函数，具体二者之间的关系将在第 2 章中详细论述。

其次看线性不可分情形，有两种处理方式。

第一种处理方式如科特斯和万普尼克在 1993 年[46]引进的软间隔最优超平面概念，引入常数 C；引入非负松弛变量 ξ_i，ξ_i 是对经验误差的度量。优化问题转化为平衡经验误差及决策函数的推广能力，为此优化问题(1.1)变形为

$$\begin{aligned}&\min_{\boldsymbol{w},b,\boldsymbol{\xi}}\quad \frac{1}{2}\|\boldsymbol{w}\|^2+C\left(\sum_{i=1}^{m}\xi_i\right)\\&\text{s.t.}\quad y_i[(\boldsymbol{x}_i\cdot\boldsymbol{w})+b]+\xi_i\geqslant 1,\\&\qquad \xi_i\geqslant 0,\quad i=1,2,\cdots,m。\end{aligned}\tag{1.3}$$

这一方法适用于如排除个别样本点即成为线性可分情形，称为近似线性可分问题。最优化问题(1.3)的目标函数的第一项表示的是边界超平面的间隔，第二项表示经验误差；这里出现了惩罚参数 C，惩罚参数 C 的大小可以控制经验误差出现多少，从上述最优化问题的目标函数可以发现：惩罚参数 C 变大可以体现出我们重视经验误差，反之，C 变小可以体现出相对于经验误差来说，更重视决策函数的推广能力。问题(1.3)的直观表述可以参见图 1.3。

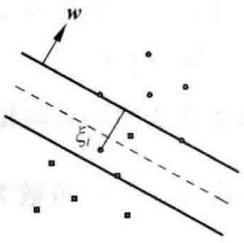


图 1.3 近似线性可分的分类图

真正求解问题(1.3)，我们同样是求解其对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,\cdots,m。 \end{aligned} \tag{1.4}$$

注 当参数 C 取正无穷，问题(1.3)退化为线性可分形式(1.1)，问题(1.4)退化为线性可分的情形(1.2)。

第二种处理方式是引入核函数[51]，对分类面是非线性函数的情况，可以将输入空间通过某种非线性变换 $\boldsymbol{\phi}: \mathbb{R}^n \to H$ 映射到一个特征空间 H。设空间 H 为内积空间，在空间 H 中存在线性的分类规则，可以构造线性的最优分类超平面，这时问题为求 H 的共轭空间的元素 $\bar{\boldsymbol{w}}, \bar{b}, \boldsymbol{\xi}$，使其满足下面的优化问题：

$$\begin{aligned} \min_{w,b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\bar{\boldsymbol{w}}\|^2 + C\left(\sum_{i=1}^{m}\xi_i\right) \\ \text{s.t.} \quad & y_i[(\bar{\boldsymbol{x}}_i \cdot \bar{\boldsymbol{w}}) + \bar{b}] + \xi_i \geqslant 1, \quad i = 1,2,\cdots,m, \\ & \xi_i \geqslant 0, \end{aligned} \tag{1.5}$$

其中 $\bar{\boldsymbol{x}}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$。

记函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$，则最优化问题(1.5)的沃尔夫对偶问题为

$$\begin{aligned} \max \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,\cdots,m。 \end{aligned} \tag{1.6}$$

如果 $\boldsymbol{\alpha}^*$ 是以上优化问题的解，决策函数为 $f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right)$。这

样通过引入函数 $K(\boldsymbol{x},\boldsymbol{y})$ 和求解沃尔夫对偶问题而建立起决策函数，全部操作仍是在原来的输入空间 $\mathbb{R}^n$ 上进行的，而不管上述概念中的 H 具体是什么内积空间[52,53]。更进一步地，只要有一个恰当的函数 $K(\boldsymbol{x},\boldsymbol{y})$ 就可以构造出问题(1.6)，进而得出决策函数 $f(\boldsymbol{x})=\operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b\right)$，而不必知道非线性变换 $\boldsymbol{\phi}(\boldsymbol{x})$ 和空间 H 是什么，此函数 K 就是所谓的核函数。

常用的核函数有多项式核函数、高斯径向基核函数、Sigmoid 核函数、B 样条核函数等[50]。

(1) 多项式核(Poly)函数

$$\begin{cases}K(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}\cdot\boldsymbol{y})^d,\\ K(\boldsymbol{x},\boldsymbol{y})=((\boldsymbol{x}\cdot\boldsymbol{y})+1)^d,\end{cases}\quad d=1,2,\cdots。\tag{1.7}$$

通常我们推荐使用第二个多项式核函数，可以避免黑塞(Hesse)矩阵为 $\mathbf{0}$ 的情况。

(2) 高斯径向基函数(radial basis function，RBF)，其表达式是

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})}{2\sigma^2}\right),\quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n。\tag{1.8}$$

写成向量的分量形式为

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{2\sigma_i^2}\right),\quad \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n。\tag{1.9}$$

RBF 核函数具有很强的生物背景和逼近任意非线性函数的能力，可以高速且以较高的精度完成预测工作，并且能以任意精度近似任何连续函数，在通常的情况下，我们认为 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2=\sigma^2$(这里定义参数为 σ^2，而不是 σ)，参数 σ^2 能控制函数的形状。

(3) Sigmoid 核函数

$$K(\boldsymbol{x},\boldsymbol{y})=\tanh(k(\boldsymbol{x}\cdot\boldsymbol{y})+v)，\text{其中 } k>0,v<0。\tag{1.10}$$

这个函数不是正定核，但它在某些实际应用中却非常有效。

(4) B 样条核函数，以 τ 为节点的一维 p 阶样条核函数：

$$K(x,x')=\sum_{j=1}^{m}(x-\tau_j)_+^p\,(x'-\tau_j)_+^p,\quad \forall x,x'\in\mathbb{R},\tag{1.11}$$

其中

$$x_+^p=\begin{cases}x^p, & x>0,\\ 0, & x\leqslant 0。\end{cases}\tag{1.12}$$

在实际应用中，我们常使用带有软间隔的支持向量分类机方法，它有两方面的好处：一方面当参数 C 变为正无穷时，$\phi(\boldsymbol{x}_i)=\boldsymbol{x}_i$，问题(1.5)则退化为问题(1.1)，

即退化为线性可分支持向量分类机模型，可以解决线性可分情况；另一方面，当样本点线性可分时，却因为存在少数样本点使得两类间隔过小，这时考虑用线性可分支持向量分类机所得到的决策函数并不见得是最好的，因为这些少数样本点的存在严重影响了最优分划超平面。我们常把问题(1.5)称为标准的支持向量分类机，也称为C-支持向量分类机(C-support vector classification，C-SVC)，它也被广泛地应用。

支持向量分类机的优点：(1)它是结构风险最小化的具体实现；(2)它具有良好的推广能力；(3)从线性分类出发，通过核函数实现非线性分类；(4)支持向量分类机模型是凸二次规划模型。支持向量分类机的核心思想是寻找一种算法在经验风险和推广能力二者之间去平衡。

1.3 支持向量分类机已有研究

支持向量机作为一种新的机器学习方法，它研究的目标就是从 m 个独立同分布观测样本

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)$$

出发，寻找一个函数对其依赖关系进行最优估计。根据学习内容，支持向量机分为支持向量分类机和支持向量回归机两类。

支持向量分类机算法出现不久就引起了国际上众多学者的关注，他们在支持向量分类机算法的理论研究和应用研究方面有很大进展。近些年，国内外有关学者在此方面的研究也很多，如文献[54～59]。

由于支持向量分类机在许多应用领域表现出较好的推广能力，自20世纪90年代提出以后，得到了广泛的研究。目前，对支持向量分类机的研究主要有：各种改进的支持向量分类机模型、支持向量分类机求解算法，统计学习理论基础，以及各种应用领域的推广等。

1.3.1 支持向量分类机模型研究现状

标准的支持向量分类机方法是我们前面介绍的算法，但人们通过增加函数项或修改变量系数等方法使标准的支持向量分类机中的最优化问题变形，用来解决某一类问题和适用某种优化算法。这里分别加以介绍。

1. C-支持向量分类机系列方法

由于万普尼克在1995年[2]最早提出的支持向量分类机方法含有常数 C，为此将这种方法称为标准的支持向量分类机方法，称为C-SVC方法。这一方法是基于前面介绍的支持向量分类机的基本思想而建立的。在文献[60]中，将原始最优化