華夏英才基金學術文庫

邓乃扬　田英杰　著

数据挖掘中的新方法
——支持向量机

科学出版社
www.sciencep.com

内 容 简 介

支持向量机是数据挖掘中的一个新方法．支持向量机能非常成功地处理回归问题(时间序列分析)和模式识别(分类问题、判别分析)等诸多问题，并可推广于预测和综合评价等领域，因此可应用于理科、工科和管理等多种学科．目前国际上支持向量机在理论研究和实际应用两方面都正处于飞速发展阶段．希望本书能促进它在我国的普及与提高．

本书对象既包括关心理论的研究工作者，也包括关心应用的实际工作者．对于有关领域的具有高等数学知识的实际工作者，略去书中的某些理论部分，仍能对支持向量机的本质有一个概括的理解，从而用它解决自己的问题．

本书适合高等院校高年级学生、研究生、教师和相关科研人员及相关领域的实际工作者使用．

图书在版编目(CIP)数据

数据挖掘中的新方法：支持向量机/邓乃扬，田英杰著．—北京：科学出版社，2004

ISBN 7-03-013281-5

Ⅰ．数…　Ⅱ．①邓…②田…　Ⅲ．支持向量机-优化方法　Ⅳ．TP13

中国版本图书馆 CIP 数据核字(2004)第 002705 号

责任编辑：林　鹏　陈玉琢　贾瑞娜/责任校对：朱光光

责任印制：钱玉芬/封面设计：黄华斌　陈　敬

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行　各地新华书店经销

*

2004年6月第一版　开本：B5 (720×1000)

2004年6月第一次印刷　印张：26 1/2

印数：1—3 000　字数：484 000

定价：53.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

序　言

数据挖掘源于数据库技术引发的海量数据和人们利用这些数据的愿望. 用数据管理系统存储数据, 用机器学习的方法分析数据、挖掘海量数据背后的知识, 便促成了数据挖掘 (data mining) 的产生. 概括地讲, 数据挖掘的任务是从大型数据库或数据仓库中提取人们感兴趣的、事先未知的、有用的或潜在有用的信息.

支持向量机 (support vector machine, SVM) 是数据挖掘中的一项新技术, 是借助于最优化方法解决机器学习问题的新工具. 它最初于 20 世纪 90 年代由 Vapnik 提出, 近年来在其理论研究和算法实现方面都取得了突破性进展, 开始成为克服 "维数灾难" 和 "过学习" 等传统困难的有力手段. 虽然它还处于飞速发展的阶段, 但是它的理论基础和实现途径的基本框架已经形成. 自 2000 年开始, 国外已陆续有几本专著出版. 据我们所知, 本书是国内第一本专门对它进行全面完整介绍和论述的书籍.

本书主要以分类问题 (模式识别, 判别分析) 和回归问题为背景, 系统阐述支持向量机和相应的最优化方法. 各章的主要内容如下: 第 1 章介绍最优化问题及其基本理论. 第 2 章对分类问题和回归问题直观地导出最基本的支持向量机. 第 3 章介绍核的理论, 这是推广基本的支持向量机的关键, 也是通过线性问题求解非线性问题的基础. 第 4 章介绍统计学习理论, 讨论支持向量机的统计学理论基础. 第 5 章和第 6 章分别详细研究支持向量分类机和支持向量回归机. 第 7 章介绍实现支持向量机的最优化算法. 第 8 章讨论支持向量机的应用, 包括解决实际问题时的一些处理方法和一些应用实例.

本书包括了我们自己的研究工作. 例如, 在做为支持向量机基础的原始问题和对偶问题解的关系上, 我们发现, 当前文献的论述存在着逻辑上的缺陷. 本书第一次在完整严密的逻辑基础上完善了各种支持向量机中的最优化问题的理论体系. 此外, 作为求解支持向量机中优化问题的方法, 本书介绍了我们自己的研究成果, 如处理大型问题的 Newton-PCG 型算法. 另外还应说明, 本书还包含了我们讨论班成员的若干研究工作.

本书所设定的读者包括关心理论与应用两方面的人士. 对于支持向量机的理论, 本书有系统而严谨的论述; 作为使用支持向量机的入门, 有直观的说明. 实际上我们特别强调该书的可读性, 强调直观对理解问题实质的重要作用. 我们通常总是首先用图像等直观手段引进各种概念、方法和结论, 并特别注意对它们的本质给予形象的解释和说明, 最后给出其严格证明. 仅仅关心实际应用的读者, 略去这些证明以及若干理论结论, 仍可以对所介绍的方法的本质有一个概括的理解. 本书对有关领域具有高等数学知识的实际工作者是一本实用读物. 我们希望本书的出

版，能普及和推广支持向量机在多种实际领域中的应用，也能促进我国对支持向量机的深入研究，特别是促进优化界朋友们的关心与参与.

本书得以出版，我们要感谢中国科学院科学出版基金和华夏英才基金的资助，同时也要感谢中国农业大学各级领导的支持和重点课程建设的资助. 本书已被选为中国农业大学研究生系列教材. 我们还要感谢国家自然科学基金多年来对我们研究工作的资助. 本书作者曾致力于最优化方法的研究多年，几年前开始组织和领导讨论班，学习与研究数据挖掘和支持向量机. 除本书两位作者外，讨论班的成员还有王来生教授、薛毅教授、钟萍副教授、经玲副教授、张春华、杨志民、刘广利、苏时光等多人. 在这里我们要特别感谢钟萍副教授和张春华. 此外，我们还要感谢刘宝光和张建中两位教授以及梁玉梅、张梅梅两位同学，他们都对本书提供了帮助.

由于作者水平所限，书中难免有不妥之处，欢迎读者批评指正.

符 号 表

$\mathbf{R}$	实数集合
$\mathbf{R}^n$	n维欧氏空间
(x_i, y_i)	训练点
$T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$	训练集
l	训练点个数
$\mathcal{X}$	输入空间
$\mathcal{Y}$	输出空间
$\mathcal{X} \times \mathcal{Y}$	训练点所在空间
$(\mathcal{X} \times \mathcal{Y})^l$	训练集所在空间
x	输入向量 (输入 / 模式)
$[x]_i$	向量x的第i个分量
x	Hilbert空间中的向量
$[\mathbf{x}]_i$	向量$\mathbf{x}$的第i个分量
y	输出指标 (输出)
$(x \cdot x')$	x与x'的内积
$\mathcal{H}$	内积空间， Hilbert 空间
$\mathcal{Z} = \{x_1, \cdots, x_l\}$	输入空间中的l个点组成的集合
$\mathcal{Z} = \{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$	Hilbert 空间中的 l 个点组成的集合
Φ	输入空间到 Hilbert 空间的映射
w	权向量
$[w]_i$	权向量w的第i个分量
$\mathbf{w}$	Hilbert 空间中的权向量
$[\mathbf{w}]_i$	权向量$\mathbf{w}$的第i个分量
b	阈值
$\mathrm{Conv}(\cdot)$	凸壳
$\mathrm{sgn}(\cdot)$	符号函数
$K(x, x')$	核函数
$\boldsymbol{K}$	核矩阵 (Gram 矩阵)
$\Vert \cdot \Vert_p$	p- 范数
$\Vert \cdot \Vert$	2- 范数
h	VC 维
C	惩罚参数
D	收缩凸壳的参数
$\ln$	自然对数
$\log_2$	底为 2 的对数

$\boldsymbol{\xi}$	松弛变量
ξ_i	松弛变量的第i个分量
γ	间隔
$\boldsymbol{\alpha}$	对偶变量， Lagrange 乘子
α_i	对偶变量的第i个分量
$P(\cdot)$	通常表示概率分布或概率

目　录

第 1 章　最优化问题及其基本理论

本章首先从具体实例出发，介绍最优化问题的数学形式及其几何解释，然后给出最优化问题解的最优性条件，最后讨论对偶理论.

1.1　最优化问题

1.1.1　最优化问题实例

什么是最优化？用数学语言概括地说，就是找出一个多变量函数的极小值点，我们通过两个实例予以说明.

例 1.1.1　曲线拟合问题

设有两个物理量 ξ,η. 根据某一物理定律知道它们满足下列函数关系

$$\eta = a + b\xi^c\,, \tag{1.1.1}$$

其中 a,b,c 是三个常数，在不同的情况下取不同的数值. 假定现在由实验测得了 (ξ,η) 的 l 组数据

$$(\xi_1,\eta_1),\cdots,(\xi_l,\eta_l)\,. \tag{1.1.2}$$

试选择 a,b,c 的值，使得曲线 $\eta = a + b\xi^c$ 尽可能靠近所有的实验点 $(\xi_i,\eta_i), i = 1,\cdots,l$(参看图 1.1.1).

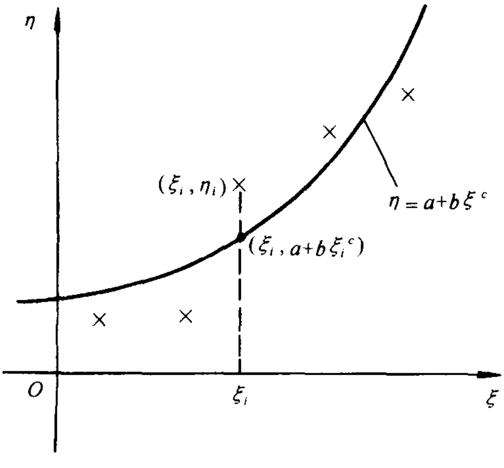

图 1.1.1

这个问题可用最小二乘原理求解，即选择 a,b,c 的值，使偏差的平方和

$$\delta(a,b,c)=\sum_{i=1}^{l}(a+b\xi_i{}^c-\eta_i)^2 \tag{1.1.3}$$

取最小值. 换句话说，欲求出三个变量的函数 $\delta(a,b,c)$ 的极小值点，以此作为问题的解.

例 1.1.2 生产安排问题

某工厂生产 D,E 两种产品，每种产品均经三道工序加工而成. 假定每生产 1 立方米 D 种产品需用 A 种机器加工 7 小时，用 B 种机器加工 3 小时，用 C 种机器加工 4 小时. 而每生产 1 立方米 E 种产品需用 A 种机器加工 2.8 小时，用 B 种机器加工 9 小时，用 C 种机器加工 4 小时. 又已知每生产 1 立方米 D 种产品可赢利 500 元，每生产 1 立方米 E 种产品可赢利 800 元. 现设一个月中 A 种机器工作时间不得超过 560 小时， B 种机器工作时间不得超过 460 小时， C 种机器工作时间不得超过 336 小时. 问每月 D,E 两种产品各生产多少可赢利最多？

设每月生产 D 种产品 $[x]_1$ 立方米，生产 E 种产品 $[x]_2$ 立方米. 这两种产品共赢利 $500[x]_1+800[x]_2$ 元. 我们要选择 $[x]_1,[x]_2$, 使赢利尽可能大. 然而受到机器加工时间的限制， $[x]_1,[x]_2$ 都不能太大. 事实上，这两种产品共需用 A 种机器加工 $7[x]_1+2.8[x]_2$ 小时，它不能超过 560 小时，需用 B 种机器加工 $3[x]_1+9[x]_2$ 小时，它不能超过 460 小时，需用 C 种机器加工 $4[x]_1+4[x]_2$ 小时，不能超过 336 小时. 总之，我们应该在约束条件

$$7[x]_1+2.8[x]_2\leqslant 560\,, \tag{1.1.4}$$

$$3[x]_1+9[x]_2\leqslant 460\,, \tag{1.1.5}$$

$$4[x]_1+4[x]_2\leqslant 336\,, \tag{1.1.6}$$

$$[x]_1\geqslant 0\,,\ [x]_2\geqslant 0 \tag{1.1.7}$$

下，寻找使函数 $500[x]_1+800[x]_2$ 取最大值的 $[x]_1^*$ 和 $[x]_2^*$.

1.1.2 最优化问题

1. 最优化问题的数学形式

前节考虑的两个例题中都包含着若干需要选择或调整的量，而调整的目的都是使这些量的某个函数达到其最小值或最大值，这类问题称为最优化问题. 由于一个函数 $f(x)$ 的最大值完全与函数 $-f(x)$ 的最小值相对应，所以能把最优化问题统

一归结为达到最小值的问题. 然而上述两个最优化问题之间还有不同之处: 在例 1.1.1 中, 要选择的量是不受限制的, 而在例 1.1.2 中, 要选择的量需在某个限定的范围内取值. 他们各自代表着一类最优化问题. 前者称为无约束问题, 后者称为约束问题. 下面分别给出它们的数学表达形式.

1) 无约束问题

考虑例 1.1.1 所代表的问题. 为了导出它们的标准形式, 我们把例 1.1.1 改写一下. 先分别记所选择的量 a,b,c 为 $[x]_1,[x]_2,[x]_3$, 并用一个 3 维欧氏空间中的向量 $x=([x]_1,[x]_2,[x]_3)^{\mathrm{T}}$ 表示它们, 再把欲求最小值的函数 $\delta(a,b,c)=\delta([x]_1,[x]_2,[x]_3)$ 记作 $f(x)=f([x]_1,[x]_2,[x]_3)$. 这样例 1.1.1 所述问题即可表示为

$$\min f(x),\quad x=([x]_1,[x]_2,[x]_3)^{\mathrm{T}}\in\mathbf{R}^3\ , \tag{1.1.8}$$

其中

$$f(x)=\sum_{i=1}^{l}([x]_1+[x]_2\xi_i{}^{[x]_3}-\eta_i)^2\ . \tag{1.1.9}$$

该问题的含义是求定义在 3 维欧氏空间 $\mathbf{R}^3$ 上的函数 $f(x)$ 的最小值点. 可以想像, 若有 n 个需要选择或调整的量时, 我们可以用 n 维欧氏空间上的向量 $x=([x]_1,\cdots,[x]_n)^{\mathrm{T}}$ 表示它们, 而把欲求最小点的函数记为 $f(x)$, 从而问题可表示为

$$\min f(x),\qquad x\in\mathbf{R}^n\ . \tag{1.1.10}$$

这就是对变量没有限制的最优化问题的一般形式, 即无约束问题的一般形式. 我们称其中的 $f(x)$ 为目标函数. 因此无约束问题就是寻求一个定义在 n 维欧氏空间 $\mathbf{R}^n$ 上的目标函数 $f(x)$ 的最小值点. 以下确切地给出问题的两种解的概念.

定义 1.1.3(无约束问题的整体解) 考虑无约束问题 (1.1.10). 称 x^* 为它的整体解或整体最优解, 或称 x^* 为目标函数 $f(x)$ 的整体 (全局) 极小点, 如果对任意的 $x\in\mathbf{R}^n$, 都有

$$f(x)\geqslant f(x^*)\ . \tag{1.1.11}$$

又称 x^* 是无约束问题 (1.1.10) 的严格整体解或严格整体最优解, 或称 x^* 是 $f(x)$ 的严格整体 (全局) 极小点, 如果对任意的 $x\in\mathbf{R}^n, x\neq x^*$, 都有

$$f(x)>f(x^*)\ . \tag{1.1.12}$$

定义 1.1.4(无约束问题的局部解) 考虑无约束问题 (1.1.10). 称 x^* 是它的局部解或局部最优解 (简称为问题的解或最优解), 或称 x^* 是目标函数 $f(x)$ 的局部极

小点 (简称为极小点), 如果对 $x^* \in \mathbf{R}^n$, 存在着 $\varepsilon > 0$, 使当 x 与 x^* 的距离小于 ε, 即当 $\|x - x^*\| < \varepsilon$ 时, 总有

$$f(x) \geqslant f(x^*) . \tag{1.1.13}$$

又称 x^* 为问题的严格局部解或严格局部最优解 (简称为严格解或严格最优解), 或称 x^* 是目标函数 $f(x)$ 的严格局部极小点 (简称为严格极小点), 如果当 $0 < \|x - x^*\| < \varepsilon$ 时, 总有

$$f(x) > f(x^*) . \tag{1.1.14}$$

2) 约束问题

考虑例 1.1.2. 若记 $x = ([x]_1, [x]_2)^{\mathrm{T}}$, 并令

$$f(x) = -(500[x]_1 + 800[x]_2) , \tag{1.1.15}$$

则使 $500[x]_1 + 800[x]_2$ 取最大值等价于使 $f(x)$ 取最小值. 因此, 若令

$$c_1(x) = 7[x]_1 + 2.8[x]_2 - 560 , \tag{1.1.16}$$

$$c_2(x) = 3[x]_1 + 9[x]_2 - 460 , \tag{1.1.17}$$

$$c_3(x) = 4[x]_1 + 4[x]_2 - 336 , \tag{1.1.18}$$

$$c_4(x) = -[x]_1 , \tag{1.1.19}$$

$$c_5(x) = -[x]_2 , \tag{1.1.20}$$

便可把问题简单地表示为

$$\min \quad f(x) , \ x = ([x]_1, [x]_2)^{\mathrm{T}} \in \mathbf{R}^2 , \tag{1.1.21}$$

$$\text{s.t.} \quad c_i(x) \leqslant 0 , \ i = 1, \cdots , 5 . \tag{1.1.22}$$

其中 $f(x)$ 和 $c_i(x)$ 由式 (1.1.15)~(1.1.20) 给出.

在某些问题中可能会有更多的需要选择的量和更多的约束条件. 当有 n 个需要选择的量, p 个不等式约束时, 问题可以表示为

$$\min \quad f(x) , \ x = ([x]_1, \cdots , [x]_n)^{\mathrm{T}} \in \mathbf{R}^n, \tag{1.1.23}$$

$$\text{s.t.} \quad c_i(x) \leqslant 0 , \ i = 1, \cdots , p. \tag{1.1.24}$$

这类问题称为不等式约束问题.

有些问题中的变量只受到等式约束的限制，这时约束问题的一般形式为

$$\min \quad f(x)\ ,\ x=([x]_1,\cdots,[x]_n)^{\mathrm{T}}\in\mathbf{R}^n, \tag{1.1.25}$$

$$\text{s.t.} \quad c_i(x)=0\ ,\ i=1,\cdots,q. \tag{1.1.26}$$

这类问题称为等式约束问题.

更加一般的约束问题可以既包含等式约束又包含不等式约束，其一般形式为

$$\min \quad f(x)\ ,\ x=([x]_1,\cdots,[x]_n)^{\mathrm{T}}\in\mathbf{R}^n, \tag{1.1.27}$$

$$\text{s.t.} \quad c_i(x)\leqslant 0\ ,\ i=1,\cdots,p, \tag{1.1.28}$$

$$c_i(x)=0\ ,\ i=p+1,\cdots,p+q. \tag{1.1.29}$$

这类问题称为一般约束问题. 在问题 (1.1.27)~(1.1.29) 中，称 $f(x)$ 为目标函数，称 $c_i(x)\leqslant 0, i=1,\cdots,p$ 和 $c_i(x)=0, i=p+1,\cdots,p+q$ 为约束条件，并称 $c_i(x)$ 为约束函数. 此外称满足约束条件的点为可行点，并称全体可行点组成的集合为可行域. 与定义 1.1.3 和定义 1.1.4 相对应，对问题 (1.1.27)~(1.1.29) 可给出如下定义:

定义 1.1.5(约束问题的整体解) 考虑一般约束问题 (1.1.27)~(1.1.29). 设 D 是它的可行域

$$D=\{x|c_i(x)\leqslant 0, i=1,\cdots,p\,;c_i(x)=0\,,i=p+1,\cdots,p+q\,;x\in\mathbf{R}^n\}\,. \tag{1.1.30}$$

称 x^* 是约束问题 (1.1.27)~(1.1.29) 的整体解或整体最优解，或称 x^* 是目标函数 $f(x)$ 在可行域 D 上的整体 (全局) 极小点，如果对任意的 $x\in D$, 都有

$$f(x)\geqslant f(x^*)\,. \tag{1.1.31}$$

称 x^* 是约束问题 (1.1.27)~(1.1.29) 的严格整体解或严格整体最优解，或称 x^* 是目标函数 $f(x)$ 在可行域 D 上的严格整体 (全局) 极小点，如果对任意 $x\in D, x\neq x^*$, 都有

$$f(x)>f(x^*)\,. \tag{1.1.32}$$

定义 1.1.6(约束问题的局部解) 考虑一般约束问题 (1.1.27)~(1.1.29). 设 D 是它的可行域

$$D=\{x|c_i(x)\leqslant 0, i=1,\cdots,p\,;c_i(x)=0\,,i=p+1,\cdots,p+q\,;x\in\mathbf{R}^n\}\,. \tag{1.1.33}$$