



北京大学

硕士研究生学位论文

题目： 放疗后非小细胞肺癌原发灶
退缩规律初步探讨

姓 名： 郑东勇

学 号： 10601050

院 系： 数学科学学院

专 业： 概率论与数理统计

研究方向： 数理统计理论及应用

导师姓名： 刘力平教授 姜明教授

二 0 0 九 年 六 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

现阶段我国肺癌的发病率持续增高，其中非小细胞肺癌占大多数。对于这类患者目前标准治疗为同时放化疗，然而放疗后大部分患者有部分肿瘤残存，这成为日后肿瘤复发和远地转移的根源。目前多数方法采取对肺内肿块的整体加大治疗剂量，然而整体加量方法可能会引起各类放疗并发症。如果能够结合放疗前后的临床数据建立肿瘤的退缩规律，在实际治疗的过程中对原始肿块整体适当加量的同时利用调强放疗技术给予预测的部位施加更高剂量，从而达到减少乃至消除肿瘤残存的目的。

本文利用回归分析方法处理得到的临床样本数据，得到非小细胞肺癌退缩规律的初步结果，主要工作有：

1. 简单介绍几种常用的回归方法，重点介绍支持向量机方法。
2. 结合已有的样本数据，比较各种方法的优劣，得到非小细胞肺癌退缩规律的初步结果。

关键词：非小细胞肺癌、退缩规律、回归分析、支持向量机、核函数

The Preliminary Study of Tumor Regression for Non-small Cell Lung Cancer after Radiotherapy

Dong Yong Zheng (Statistics)

Directed by Professor Li Ping Liu & Professor Ming Jiang

Lung cancer is one of the most common causes of cancer-related death. For a large number of patients surgery is inappropriate either because of locally advanced disease or because of comorbidity. For these patients the only potentially curative treatment modality is radical radiotherapy or concurrent radiochemotherapy. However, most patients after radiotherapy have tumor nidus, which has become the source of future recurrence and metastasis. The current popular method is increasing treating dose overall the tumor, but it may cause some complications.

If we can develop the shrinkage principle of the tumor with clinical data before and after treatment, then we can increase the treating dose overall tumor in treatment process, at the same time use Intensity Modulated Radiation Technology to give the forecast position a higher dose, this may results in the reduction or elimination of tumor nidus.

This thesis deals with clinical samples with regression analysis methods, and gets the preliminary results of shrinkage principle for non-small cell lung cancer.

Keywords: non-small cell lung cancer shrinkage principle regression analysis
support vector machine kernel function.

目录

第一章 序言	6
1.1 国内外研究热点.....	6
1.2 “瘤根靶向放疗”理论.....	7
1.3 技术路线和实验目的.....	8
第二章 传统回归分析方法	10
2.1 线性回归.....	10
2.1.1 最小二乘估计及性质.....	11
2.1.2 拟合和预测.....	12
2.1.3 异常值检验.....	13
2.1.4 最小一乘估计.....	13
2.2 非参数回归.....	15
2.2.1 核函数估计.....	15
2.2.2 非参数估计的优缺点.....	17
第三章 支持向量机	19
3.1 支持向量分类.....	19
3.1.1 线性可分情况.....	20
3.1.2 近似线性可分情况.....	23
3.1.3 一般情况.....	26
3.2 核函数.....	30
3.2.1 多项式空间和多项式核.....	30
3.2.2 Mercer 核和正定核.....	31
3.2.3 核函数的构造原则及常用的核函数.....	33
3.3 VC 维和结构风险最小化原则.....	35
3.3.1 损失函数和期望风险.....	35
3.3.2 经验风险和函数集的 VC 维.....	35
3.3.3 结构风险最小化原则.....	37
3.4 支持向量回归.....	41
3.4.1 常用的损失函数.....	42
3.4.2 支持向量回归.....	42
第四章 实验结果	52
4.1 线性回归.....	52
4.1.1 样本所含变量相关性分析.....	52
4.1.2 线性回归分析.....	54
4.2 最小一乘回归.....	61
4.3 非参数回归.....	61
4.4 支持向量回归.....	63
4.5 各种方法的比较.....	68
第五章 结论	69
5.1 结论.....	69
5.2 效果不理想可能存在的原因.....	69

第一章 序言

本文的工作来自于国家自然科学基金委员会《放疗后非小细胞肺癌原发灶退缩规律数学模型建立》课题组，课题编号：30870738，该课题是北京大学临床肿瘤学院和北京大学数学科学学院的合作项目，项目负责人为北京大学临床肿瘤医院的朱广迎教授。

我国肺癌的发病率持续增高，男性肺癌的发病率达到 50/10 万。非小细胞肺癌占肺癌总数的 80%，其中不能手术的局部晚期非小细胞肺癌约占 1/3。对于这类患者目前标准治疗为同时放化疗，然而放疗后肿瘤完全消退的患者仅占 2.6-12.1%，70%—80% 患者有部分肿瘤残存，这成为日后肿瘤复发和远地转移的根源 (Ataman OU, Barrett A, Filleron T, 2006)。如何在放疗的过程中采取适当的方法减少乃至消除放疗后可能出现的肿瘤残存，不仅是众多国内外专家研究的热点问题，同时也具有重大临床应用前景。

1.1 国内外研究热点

现代精确放射治疗是非小细胞肺癌的根治手段之一，该方法希望通过对肿瘤肿块进行精确放射治疗，达到使肿瘤缩小乃至消失的目的，然而在实际应用中肿瘤缩小的比例往往达不到临床的要求，这就需要提出新的治疗方法和治疗理论。

目前国内外专家的研究大多集中于对肺内肿块的整体加大治疗剂量，从而进一步缩小治疗以后的肿瘤残存，这可以在一定程度上提高治疗疗效，如图 1.1 所示。然而整体加量方法有其内在的缺陷：不符合肿瘤异质性的客观规律。由于该方法是对肿块的整体加量，这个量一般不能很大，所以很有可能对于肿瘤抗拒细胞来说放疗剂量显得不足，达不到大量杀死癌细胞的目的；同时对于瘤内放射敏感细胞来说治疗剂量又过高，这在临床上会引起各类严重的放疗并发症，如肺炎等。

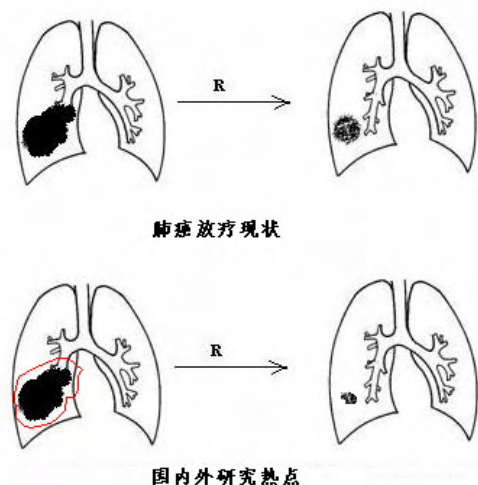


图 1.1

1.2 “瘤根靶向放疗”理论

整体加量的放疗方法由于固有的缺陷使得其治疗效果并不是很理想，反而可能会引起一些放疗并发症。为了改善整体加量方法，弥补其内在的缺陷，国外的一些研究者已着手利用 CT 研究肺癌肿块在放疗过程中体积的变化情况 (Woodford C, Yartsev S, Dar AR, 2007)，然后在放疗过程中及时缩小照射野以适应缩小的体积，进一步提高治疗疗效，并减少可能出现的并发症的概率。

然而这些研究很少关注放疗前后肿瘤的退缩规律，事实上如果能够结合放疗前后的一些临床数据建立起肿瘤的退缩规律，那么在实际治疗的过程中，可以通过分析放疗前病人的临床数据，然后结合已知的肿瘤退缩规律较为准确地预测放疗后残存灶可能的大小和部位。在对原始肿块整体适当加量的同时利用调强放疗技术给予预测的部位施加更高剂量，如图 2 所示，从而减少乃至消除肿瘤残存灶，这就是“瘤根靶向放疗”新理论。

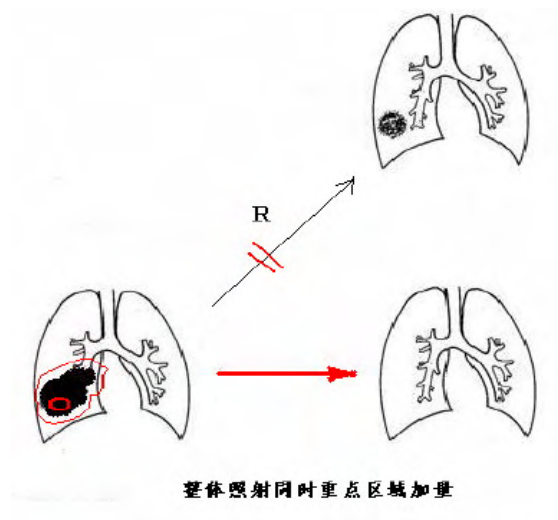


图 1.2

1.3 技术路线和实验目的

1. 技术路线

- i. 选择符合条件的患者，签署知情同意书。
- ii. 获得肺癌原发灶组织，免疫组化检测 Ki67、CD133 的表达。
- iii. 患者接受胸部定位 CT 扫描，以隆突下缘为原点，勾画肿瘤 GTV，确定肿瘤的体积和中心。
- iv. 原发灶接受放射治疗 ($> 60\text{Gy}$)，放疗 4 ~ 5 周。
- v. 在治疗中和治疗后患者再次进行胸部定位 CT 扫描，以隆突下缘为原点，勾画肿瘤 GTV，确定肿瘤的体积和中心。
- vi. 测定患者放疗后原发灶的退缩方向和体积退缩率
- vii. 分析影响放疗后原发灶的退缩方向和体积退缩率的临床因素，建立非小细胞肺癌放疗前后原发灶的退缩规律。

2. 实验目的

按照上面所述的技术路线收集数据，现阶段已有 31 个样本数据，其中临床因素包括原发病灶在治疗前后的体积大小、三维中心位置坐标以及患者性别。以后随着课题组项目的不断推进，样本容量和临床因素集合将会不断得到扩充。

本文的目的是在现有数据的基础上建立放疗前后原发病灶退缩规律的初步

模型，采用的方法分别是传统的回归分析方法（如线性回归和非参数回归）和支持向量机，文中重点介绍了支持向量机，作为最新涌现出来的分类和回归方法，其强大的功能使其在回归分析中得到越来越广泛的应用。

第二章 传统回归分析方法

回归分析是一种应用十分广泛的数理模型，它的任务是研究因变量 y 与自变量 x_1, x_2, \dots, x_p 之间的关系，并建立变量之间的经验公式，以便达到预测和控制的目的。它的研究内容主要有三个部分：

- 1) 是否有相关关系。即 x_1, x_2, \dots, x_p 的值确定以后， y 的值是否在某个确定的范围内。
- 2) 若有相关关系，如何找到适当的函数 $f(x_1, x_2, \dots, x_p)$ 使得

$$y = f(x_1, x_2, \dots, x_p) + e \quad (2.1)$$

其中 e 在某种意义下很小，叫做随机项或随机误差项，它是一个期望为零的随机变量（如果 $Ee \neq 0$ ，可以调整 f 使 $Ee = 0$ ）。

- 3) 如果得到的关系 (2.1) 中 e 的方差很小，则有

$$y \approx f(x_1, x_2, \dots, x_p) \quad (2.2)$$

此式称为回归关系式或经验公式，它可以用来预测和控制。

下面介绍两种常用的回归方法：线性回归和非参数回归。

2.1 线性回归

线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e \quad (2.3)$$

其中 β_i 是未知参数， e 为随机误差项， y 为响应变量， x_1, x_2, \dots, x_p 为自变量。

给出样本数据以后，记

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

回归方程可以简写成矩阵形式：

$$Y = X\beta + e \quad (2.4)$$

对于随机误差向量 e ，假设 $E(e) = 0$ ， $\text{Var}(e) = \sigma^2 I_n$ ，如果加上正态分布的假设，则 $e \sim N(0, \sigma^2 I_n)$ ，正态性假设常用于模型的假设检验和置信区间的估计。

2.1.1 最小二乘估计及性质

线性回归的任务是作一条直线 $Y = X\beta$ ，使得其能“最好地”代表样本点的趋势，或者说对样本点“拟合”的最好，什么叫拟合的最好？这要看给出的是怎样准则，最常用的准则是“最小二乘”准则。

最小二乘估计 $\hat{\beta}$ 是使拟合的残差平方和最小的 β 值，即

$$\hat{\beta} = \min_{\beta} Q(\beta) = \min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \min_{\beta} \|Y - X\beta\|^2 \quad (2.5)$$

对矩阵微分，得到下列结论：

$\hat{\beta}$ 是最小二乘估计的充要条件是 $\hat{\beta}$ 适合方程 $X'X\beta = X'Y$ 。

这说明最小二乘估计一定是存在的，因为

$$\text{rank}(X'X, X'Y) \geq \text{rank}(X'X) = \text{rank}(X) \geq \text{rank}(X'X, X'Y) \quad (2.6)$$

根据线性代数理论可知最小二乘估计 $\hat{\beta}$ 一定存在。

如果 $\text{rank}(X) = p + 1$ ， $X'X$ 一定可逆，则 $\hat{\beta} = (X'X)^{-1} X'Y$ 。在下面的论述中都假设 $\text{rank}(X) = p + 1$ 成立，如不成立，说明原自变量矩阵存在共线性，可以通过一定的变量选择使自变量的共线性消失。

最小二乘估计有下列性质:

- 1) $E\hat{\beta} = \beta$, 即最小二乘估计是无偏估计。
- 2) $Cov(\hat{\beta}, \hat{\beta}) = \sigma^2(X'X)^{-1}$
- 3) $EQ(\hat{\beta}) = (n-p-1)\sigma^2$, 从而得到随机项方差的无偏估计

$$\hat{\sigma}^2 = \frac{Q(\hat{\beta})}{n-p-1}。$$

2.1.2 拟合和预测

假定有了回归方程 $y = x'\hat{\beta}$, $\hat{\beta}$ 是根据已知样本得到的最小二乘估计。

1. 拟合

对于用于拟合的自变量 x 所对应的响应变量的均值估计为 $\hat{y} = c'\hat{\beta}$, 其中 $c = (1, x)'$, 方差为 $\text{var}(\hat{y}) = \sigma^2(c'(X'X)^{-1}c)$, 将式中的随机项方差 σ^2 用估计值 $\hat{\sigma}^2$ 代替, 从而得到拟合值的标准误差估计 $se(\hat{y} | x) = \hat{\sigma}\sqrt{c'(X'X)^{-1}c}$ 。

2. 预测

对于一个新的自变量值 x_0 , 可以得到其对应的响应变量 y 的预测值 $\hat{y} = (1, x_0)'\hat{\beta}$, 它是 $E(y)$ 的一个无偏估计。

$$\text{令 } c = (1, x_0)', \text{ 则 } E(y) = c'\beta, \text{ 可知 } \frac{c'(\hat{\beta} - \beta)}{\hat{\sigma}\sqrt{c'(X'X)^{-1}c}} \sim t(n-p-1),$$

从而可以得到 $E(y)$ 的 $1-\alpha$ 水平的置信区间 $[c'\hat{\beta} - u, c'\hat{\beta} + u]$, 其中 $u = \lambda\hat{\sigma}\sqrt{c'(X'X)^{-1}c}$, λ 满足 $P(|t| > \lambda) = \alpha$, $P(t)$ 为自由度为 $n-p-1$ 的 t 分布函数。

对于真值 y , 预测误差为 $\delta = y - \hat{y} = y - c'\beta + c'(\beta - \hat{\beta})$, 由于误差项 $e_0 = y - c'\beta$ 与前 n 个观测的随机项 e_1, e_2, \dots, e_n 独立而且有相同的正态分布, 所以 $\text{var}(\delta) = \sigma^2 + \sigma^2 c'(X'X)^{-1}c$, $\delta \sim N(0, \sigma^2 + \sigma^2 c'(X'X)^{-1}c)$, 于是

$t = \frac{y - c'\hat{\beta}}{\hat{\sigma}\sqrt{1+c'(X'X)^{-1}c}} \sim t(n-p-1)$, 从而可以得到真值 y 的 $1-\alpha$ 水平置信区间

$[c'\hat{\beta}-v, c'\hat{\beta}+v]$, 其中 $v = \lambda\hat{\sigma}\sqrt{1+c'(X'X)^{-1}c}$, λ 满足 $P(|t| > \lambda) = \alpha$, $P(t)$ 为自由度为 $n-p-1$ 的 t 分布函数。

要注意如果是通过数据分析得到的回归模型, 而选择的模型使 σ^2 的估计相对的小, 所以用 $\hat{\sigma}^2(1+c'(X'X)^{-1}c)$ 估计可能会低估预测方差。一个有效的准则是用回归模型去预测未参加拟合的样本, 视其预测的好坏给出预测误差的直接估计, 常用的是预测的均方误差的平方根。

2.1.3 异常值检验

在回归分析中的一个重要假定是, 使用的模型对所有的数据都是适合的。在实际的应用中, 可能存在一个或者多个观测值与模型不相符的情况。

假定第 i 个样本可能是个异常值, 我们通过如下的步骤对其进行检验

- i. 从数据中删除第 i 个样本, 余下的 $n-1$ 个样本用来拟合线性模型。
- ii. 使用删除后的数据估计 β 和 σ^2 , 分别记为 $\hat{\beta}_{(i)}$ 和 $\hat{\sigma}^2_{(i)}$ 。
- iii. 对于被删除的样本, 计算拟合值 $\tilde{y}_i = x_i'\hat{\beta}_{(i)}$ 。由于第 i 个样本没有被用于估计, 所以 $\text{var}(y_i - \tilde{y}_i) = \sigma^2(1+x_i'(X_{(i)}'X_{(i)})^{-1}x_i)$, 用 $\hat{\sigma}^2_{(i)}$ 代替 σ^2 得到方差的估计值。
- iv. 如果 y_i 不是一个异常值, 则 $E(y_i - \tilde{y}_i) = 0$, 检验假设 $E(y_i - \tilde{y}_i) = 0$ 的 t -

检验统计量为 $t_i = \frac{y_i - \tilde{y}_i}{\hat{\sigma}_{(i)}[1+x_i'(X_{(i)}'X_{(i)})^{-1}x_i]^{1/2}}$, 自由度为 $n-p-2$ 。

2.1.4 最小一乘估计

最小二乘回归因为便于计算且效果良好而得到了广泛的应用, 但其受到个别异常点的影响较大, 因为异常点往往有着比较大的偏差, 其平方值就更大, 为了降低平方和不得不更“将就”这些点, 使得模型偏离真实情况。而最小一乘回归

受的影响要少一点,也就是说最小一乘回归的稳健性比最小二乘的好。另一方面,在回归模型中一般假定随机误差服从正态分布,在这个假定下,最小二乘估计具有较好的性质,但是在一些实际应用中,误差项不能认为是正态的,而是服从一种尾部占更大比重的分布,在这些情况下,最小一乘估计的统计性质优于最小二乘估计。

最小一乘回归的计算

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}| = \min_{\beta} \|Y - X\beta\| \quad (2.7)$$

A. charnds (1955) 等三人在一篇文章中指出,上述问题可以转化为一个线性规划问题,在样本量不是很大的情况下,这个计算方法是很有用的,如果样本量很大,可以采用其他的迭代算法求解。具体转化方法如下:

i. 目标函数为 $Q(\beta) = \sum_{i=1}^n |y_i - x_i' \beta|$

ii. 引进两个 $p+1$ 维非负向量 d_1, d_2 , 令矩阵 $A = \begin{pmatrix} I_n & X & -X \\ I_n & -X & X \end{pmatrix}$,

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad r = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}, \quad l = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad (l \text{ 为 } n \text{ 维})。$$

iii. 引进线性规划问题:

$$\min(l' \quad 0 \quad 0) \begin{pmatrix} r \\ d_1 \\ d_2 \end{pmatrix} = \sum_{i=1}^n r_i \quad \text{s.t.} \quad B \begin{pmatrix} r \\ d_1 \\ d_2 \end{pmatrix} \geq \begin{pmatrix} y \\ -y \end{pmatrix}, \quad \begin{pmatrix} r \\ d_1 \\ d_2 \end{pmatrix} \geq 0 \quad (2.8)$$

通过求解这个线性规划问题得到解 r, d_1, d_2 。

iv. $\beta = d_1 - d_2$ 使得目标函数 (2.8) 达到最小,且目标函数的最小值就是 $\sum_{i=1}^n r_i$

2.2 非参数回归

回归分析的中心问题，就是寻找合适的 $f(x_1, x_2, \dots, x_p)$ 使得回归函数 $y = f(x_1, x_2, \dots, x_p) + e$ 成立，同时 $\text{var}(e)$ 应尽可能的小。为了估计函数 f ，大体上有两种方法。

一种是参数估计，即假定函数 f 的形式是已知的，它依赖于未知的参数 θ ，写成参数形式 $f = f(x, \theta)$ ，这样函数估计问题就转化为一个参数估计问题，线性回归就是一种典型的参数方法。参数估计有很多的优点，它的表达形式简单，分析起来比较方便；一旦得到参数 θ 的估计 $\hat{\theta}$ ，问题就得到了解决，同时也容易实现模型外推。但是现实问题可能是很复杂的，并非所有的关系都能用一个有限的数学式子来描述。在许多情况下，即使添加大量的自变量，模型的效果可能还是不理想。

另一种方法是非参数估计。考虑回归问题 $y_i = f(x_i) + e_i$ ，其中假设 $e_i \sim NID(0, \sigma^2)$ 。不同于参数方法对函数 f 进行参数化，非参数估计的目的是直接利用样本数据对 f 进行估计，下面给出一种常用的非参数方法：核函数估计。

2.2.1 核函数估计

先考虑简单的一维情况，样本数据为 (x_i, y_i) ，回归模型 $y_i = f(x_i) + e_i$ 。取定一个自变量的值 x_0 ，为了估计其对应的响应变量 y_0 的值，一种方法是在 x_0 附近建立一个 p 阶多项式 $y = a + b_1(x - x_0) + b_2(x - x_0)^2 + \dots + b_p(x - x_0)^p$ ， a, b_1, b_2, \dots, b_p 是待定的参数，记 $b = (b_1, b_2, \dots, b_p)'$ 。

考虑到样本数据 x_i 对 (x_0, y_0) 的影响不同（一般情况下离 x_0 越近，样本数据的影响也越大），为了刻画样本点的不同影响，引进核函数 $K(x)$ ，一般假定核函数

是对称的密度函数，常用的核函数有三角形密度函数、标准正态密度函数。核函数中一般带有一个带宽参数 h ，它刻画了一个 x_0 的局部邻域，一般认为只有在这个邻域内的样本对 x_0 有影响，可以通过调节 h 选择适当的邻域。

在最小二乘准则下，给定带宽参数 h ，估计问题转化为：

$$(\hat{a}, \hat{b}) = \min_{a,b} \sum_{i=1}^n (y_i - a - b_1(x_i - x_0) - \dots - b_p(x_i - x_0)^p)^2 K\left(\frac{x_i - x_0}{h}\right)$$

这是一个加权最小二乘估计问题，类似于线性回归最小二乘解法可以得到参数 (a,b) 的估计值 (\hat{a}, \hat{b}) ，其中 y_0 的估计值 $\hat{y}_0 = \hat{a}$ 。

图 2.1 是一个简单的例子 (John Fox, 2002) 说明，数据描述的是平均收入和社会威望的关系，样本容量为 102。

图(a)：样本 $x_{(80)}$ 附近的一个邻域，并调整 h 使其包含最近的 50 个样本点。

图(b)：选取的核函数 $K(x) = \begin{cases} (1 - |x|^3)^3 & |x| < 1 \\ 0 & |x| \geq 1 \end{cases}$ 。

图(c)： $x_{(80)}$ 附近的一个局部线性回归，加粗的点表示在 $x_{(80)}$ 处因变量的估计 \hat{y} 。

图(d)：对自变量变化区间上点进行逐个估计，并将估计值连接起来的得到的拟合曲线，可以通过调节 h 调整曲线的光滑性， h 越小拟合曲线越光滑。

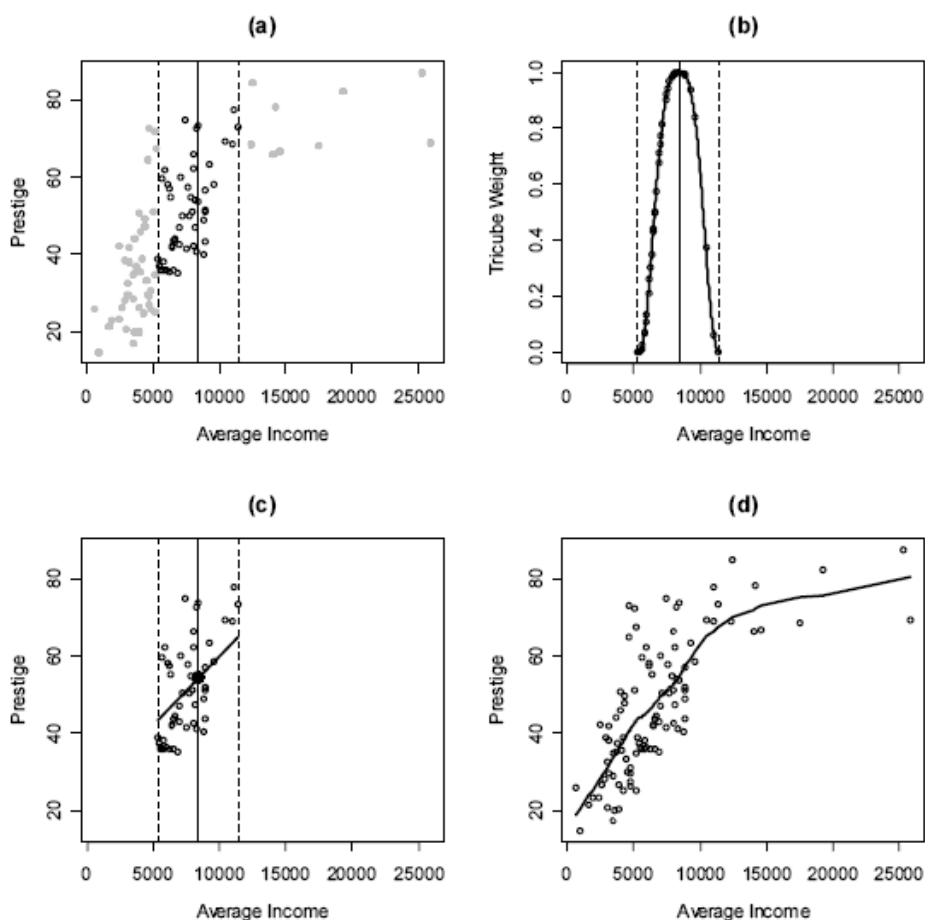


图 2.1

在最小一乘准则下的参数估计问题：

$$(\hat{a}, \hat{b}) = \min_{a, b} \sum_{i=1}^n |y_i - a - b_1(x_i - x_0) - \dots - b_p(x_i - x_0)^p| K\left(\frac{|x_i - x_0|}{h}\right)$$

多维的情况可以由一维推广得到，但要注意：在计算多维空间中的距离（一般是欧式距离）时，要注意不同自变量的量纲可能是不同的，所以在计算之前必须根据样本数据对自变量进行标准化处理。

2.2.2 非参数估计的优缺点

与参数回归估计相比，非参数回归有以下优点：

(1) 在利用样本数据对总体进行估计时，不依赖于总体所属的分布总体的分布形式，尤其是当对总体的分布不是很清楚时，因而非参数模型的适用性比较广，与

参数方法相比，具有较好的稳健性。

(2) 由于不必假定总体分布的具体形式，所以也无需多总体分布所具有的参数进行估计和检验。如果方法选择得当，非参数估计方法与参数估计的效果相差不多，尤其当参数估计的假设不满足时，非参数估计会比参数估计方法更为有效。

非参数回归也有其缺点：

(1) 如果对总体的了解足以确定它的分布类型，非参数估计就不如参数估计那样有更强的针对性。

(2) 它没有充分利用样本所携带的关于总体的信息，因而有时它的效率会低一些，或者在相同的精度下，非参数估计比参数估计需要更大的样本。

第三章 支持向量机

支持向量机 (Support Vector Machine, 简称 SVM) 最早是由 Vapnik (1992) 等人在研究统计学习理论上提出的分类设计准则. 刚开始提出时主要用于解决分类问题, 称为支持向量分类 (Support Vector Classification, 简称 SVC), 由于其强大的分类能力及在模式识别领域的成功运用, 使其逐步扩展到其他领域, 比如密度估计, 聚类分析和统计回归等, 特别是在回归领域的扩展, 为研究者提供了一个最小二乘、最大似然估计等传统方法以外的替代估计方法, 称之为支持向量回归 (Support Vector Regression, 简称 SVR)。下面从支持向量分类 (SVC) 开始介绍支持向量机的主要内容。

3.1 支持向量分类

分类问题是模式识别理论的中心问题, 在现实中有着广泛的应用。

考虑分类问题: 根据训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X = \mathbb{R}^n$, $y_i \in Y = \{1, -1\}$, $i = 1, 2, \dots, l$, 寻找一个 $X = \mathbb{R}^n$ 上的实值函数 $g(x)$, 以便用于决策函数 $f(x) = \text{sgn}(g(x))$, 其中 $\text{sgn}(x)$ 是符号函数,

$$\text{sgn}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x \leq 0 \end{cases}$$
, 从而对于任何模式 x 都能推断其对应的 y 值。求解分类问题, 实质上就是寻找一个把 \mathbb{R}^n 上的点分为两部分的规则。

上述分类问题是分成两类的问题, 相类似的还有分成多类的分类问题, 它们的不同之处在于后者的分类变量 y 可以取多个值。二分类问题是解决多分类问题的基础, 多分类问题往往转化为多个二分类问题, 不失一般性, 我们只讨论二分类问题。

3.1.1 线性可分情况

分类问题中最简单的是线性可分的情况，2 维情形下的线性可分问题图 3.1 所示。

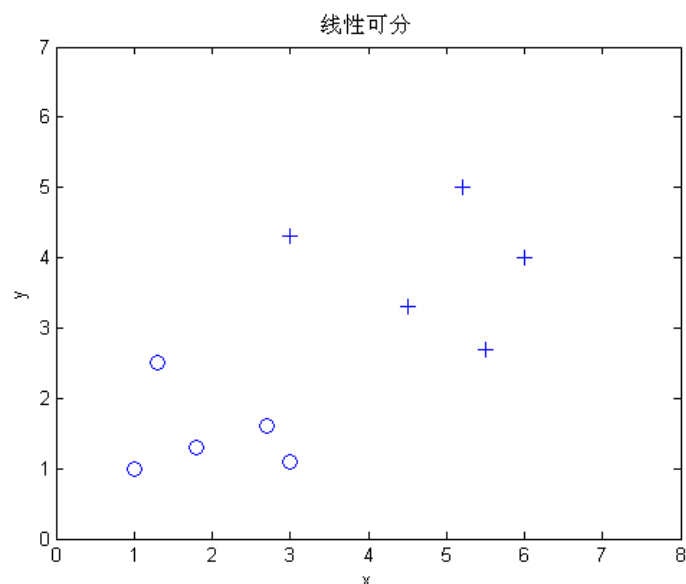


图 3.1

由于有很多直线都能将样本正确分类，所以有必要探讨一下哪条直线更好。

首先假定直线的法方向 w 已经确定，直线 l_1 就是以 w 为法方向且能正确分开两类点的直线，这样的直线不唯一，可以沿着法方向平行的移动直线 l_1 ，直到碰到某类训练样本为止，这样就得到了两条极端的直线 l_2 和 l_3 ，在 l_2 和 l_3 之间的平行直线都能正确的分类样本点，显然在这些直线中，以 l_2 和 l_3 “中间”的那条直线最好，以上分析了在已知法方向 w 的前提下如何得到最优分类直线的方法，这样就把问题归结为寻找法方向 w 的问题。对于给定的法方向 w ，类似于 l_2 、 l_3 能得到两条极端直线，称这两条直线之间的距离为该法方向对应的“间隔”，我们应该选取使“间隔”达到最大的法方向，如图 3.2 所示。

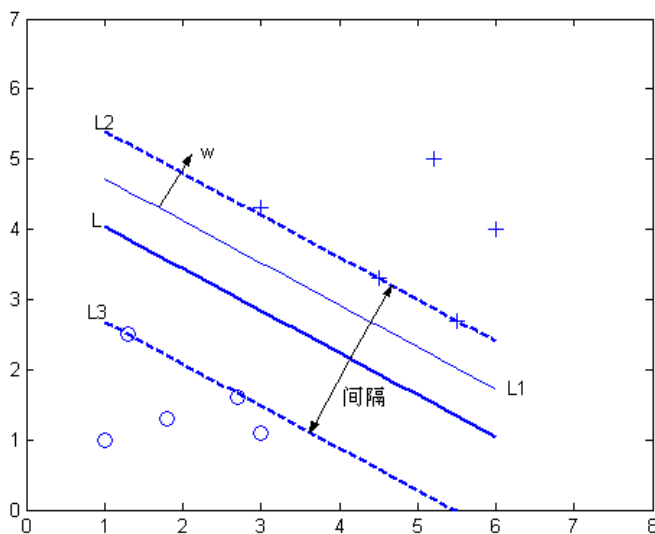


图 3.2

用数学语言描述这个过程。给定适当的法方向 \tilde{w} 后，得到的两条极端直线 l_2 和 l_3 分别可表示为 $(\tilde{w} \cdot x) + \tilde{b} = k_1$ 和 $(\tilde{w} \cdot x) + \tilde{b} = k_2$ ，调整 \tilde{b} 可以把这两条直线分别表示为 $(\tilde{w} \cdot x) + \tilde{b} = k$ 和 $(\tilde{w} \cdot x) + \tilde{b} = -k$ ，这时 l_2 和 l_3 中间的分类直线 l 可表示为 $(\tilde{w} \cdot x) + \tilde{b} = 0$ 。分别令 $w = \frac{\tilde{w}}{k}$ 、 $b = \frac{\tilde{b}}{k}$ ，则 l_2 和 l_3 及分类直线 l 可以分别表示成

$(w \cdot x) + b = 1$ ， $(w \cdot x) + b = -1$ 和 $(w \cdot x) + b = 0$ ，此时 l_2 和 l_3 之间的距离为 $\frac{2}{\|w\|}$ ，极

大化“间隔”导致了求解下列的最优化问题

$$\max_{w,b} \frac{2}{\|w\|} \tag{3.1.1}$$

$$s.t. \quad y_i((w \cdot x_i) + b) \geq 1, \quad i = 1, \dots, l \tag{3.1.2}$$

上述方法是对 2 维空间上的分类问题导出的，但对于一般 n 维空间的分类问题也同样适用，方法如下：

- i. 已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X = \mathbb{R}^n$ ，
 $y_i \in Y = \{1, -1\}$ ， $i = 1, 2, \dots, l$ ；

ii. 构造并求解最优化问题

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (3.1.3)$$

$$s.t. \quad y_i((w \cdot x_i) + b) \geq 1, \quad i=1, \dots, l \quad (3.1.4)$$

得到最优解 w^*, b^* ;

iii. 构造分类超平面 $(w^* \cdot x) + b^* = 0$, 得到决策函数 $f(x) = \text{sgn}((w^* \cdot x) + b^*)$ 。

原优化问题 (3.1.3)–(3.1.4) 的对偶问题是

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^l \alpha_i \quad (3.1.5)$$

s.t.

$$\sum_{i=1}^l y_i \alpha_i = 0 \quad (3.1.6)$$

$$\alpha_i \geq 0, \quad i=1, \dots, l. \quad (3.1.7)$$

设 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$ 为对偶问题 (3.1.5)–(3.1.7) 的任意解, 令

$$w^* = \sum_{i=1}^l \alpha_i^* y_i x_i, \quad b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (x_i \cdot x_j), \quad \text{其中下标 } j \in \{j \mid \alpha_j^* > 0\}, \quad (3.1.8)$$

由此得到的 (w^*, b^*) 是原优化问题 (3.1.3)–(3.1.4) 的解。

支持向量: 最优分类超平面是 $\sum_{i=1}^l \alpha_i^* y_i (x_i \cdot x) + b^* = 0$, 其中

$b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (x_i \cdot x_j)$, 下标 $j \in \{j \mid \alpha_j^* > 0\}$, 由此可以看出最优分类超平面只

与 $\alpha_i^* > 0$ 对应的样本点有关, 而与其它样本点无关。根据原优化问题的 KKT

条件, 得 $\alpha_i^* (y_i((w^* \cdot x_i) + b^*) - 1) = 0$, 所以 $\alpha_i^* > 0$ 对应的样本点正好都在间隔边

界上, 如图 3.3 所示。我们形象得称这样的样本点为支持向量, 它们支撑起最优

分类超平面。

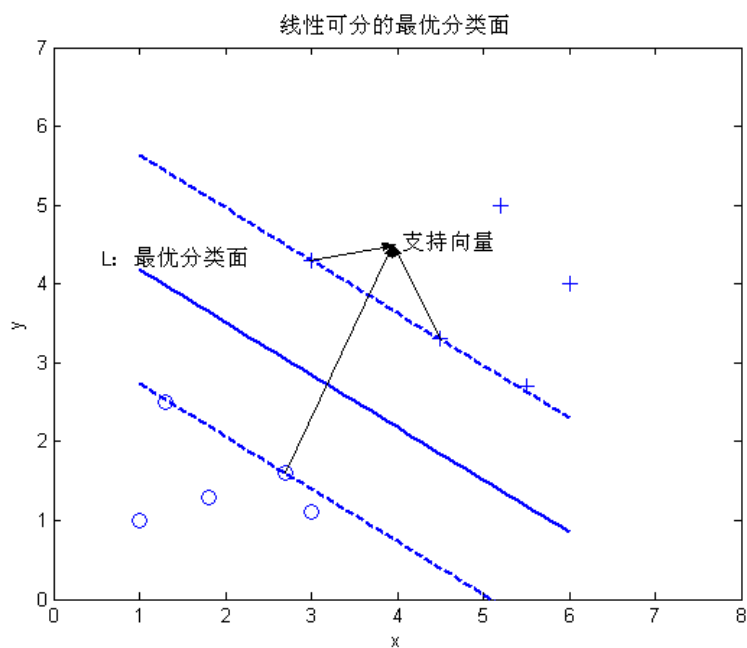


图 3.3

3.1.2 近似线性可分情况

在实际应用中，线性可分的条件并不一定满足，有可能存在少数样本点使得线性不可分，但是大部分样本点还是线性可分的，我们称之为近似线性可分，如图 3.4 所示。

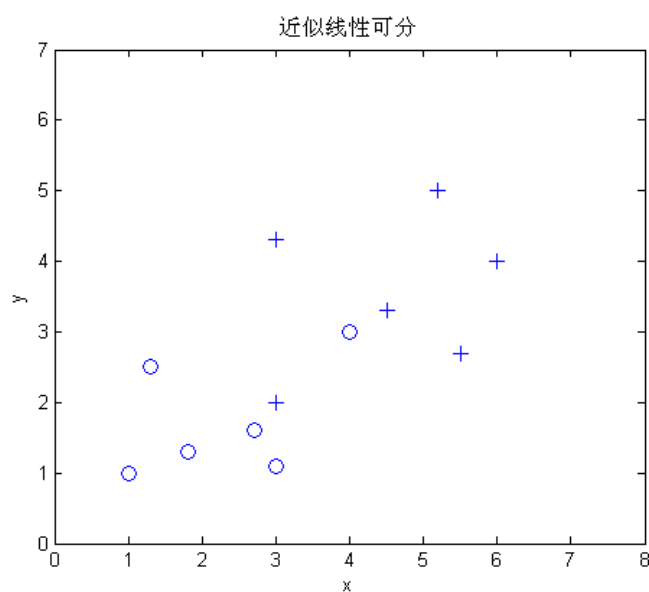


图 3.4

在近似线性可分情况下，不存在一个分类超平面能正确划分所有的样本点，即存在不满足约束条件 $y_i((w \cdot x_i) + b) \geq 1$ 的样本点，为了弱化这个约束条件，我们引入松弛变量 $\xi_i \geq 0$, $i = 1, 2, \dots, l$ ，得到了弱化的约束条件

$$y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, l. \quad (3.1.9)$$

显然，当 ξ_i 充分大时，样本点总可以满足约束条件(3.1.9)，此时得到的分类超平面与实际情况会相差很远，所以应该限制 ξ_i 使其不能太大。为此在目标函数中对大的 ξ_i 进行惩罚，比如在目标函数中添加 $\sum_i \xi_i$ ，这样线性可分情况下的原优化问题(3.1.3)-(3.1.4)转化为

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (3.1.10)$$

$$s.t. \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, l, \quad (3.1.11)$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, l, \quad (3.1.12)$$

其中惩罚参数 $C > 0$ 。

类似于线性可分情况，得到计算最优分类面的方法：

i. 已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X = \mathbb{R}^n$ ，

$$y_i \in Y = \{1, -1\}, \quad i = 1, 2, \dots, l;$$

ii. 构造并求解最优化问题

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (3.1.13)$$

$$s.t. \quad y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, l \quad (3.1.14)$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, l, \quad (3.1.15)$$

得到最优解 w^*, b^* ；

iii. 构造分类超平面 $(w^* \cdot x) + b^* = 0$ ，得到决策函数 $f(x) = \text{sgn}((w^* \cdot x) + b^*)$ 。

优化问题 (3.1.13) — (3.1.15) 的对偶问题是

$$\min_{\alpha} \quad \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^l \alpha_i \quad (3.1.16)$$

$$s.t. \quad \sum_{i=1}^l y_i \alpha_i = 0 \quad (3.1.17)$$

$$0 \leq \alpha_i \leq C, \quad i=1, \dots, l. \quad (3.1.18)$$

设 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$ 是对偶问题 (3.1.16) — (3.1.18) 的任意解, 如果存在 α^* 的分量 $\alpha_j^* \in (0, C)$, 则原始优化问题 (3.1.13) — (3.1.15) 对 (w, b) 的解 (w^*, b^*) 存在且唯一, 可以表示为 $w^* = \sum_{i=1}^l \alpha_i^* y_i x_i$, $b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (x_i \cdot x_j)$, 从而得到最优分类超平面 $w^* \cdot x + b^* = \sum_{i=1}^l \alpha_i^* y_i (x \cdot x_i) + b^* = 0$ 。

支持向量: 与线性可分情况相类似, 近似线性情况下的最优分类超平面只与 $\alpha_j^* > 0$ 对应的样本点有关, 稍有不同的是, 这里的 α_j^* 有一个上界 C , 我们分别称 $C > \alpha_j^* > 0$ 和 $\alpha_j^* = C$ 对应的样本点为界内支持向量和界上支持向量。对于正分类样本点来说, 划分区域 $(w^* \cdot x) + b^* \geq 1$ 、 $(w^* \cdot x) + b^* > 1$ 和 $(w^* \cdot x) + b^* = 1$ 分别称为带有间隔的正确分类区、间隔以外的正确分类区和间隔边界上的正确分类区; 对于负分类点来说也由相同的定义。

由原优化问题 (3.1.13) — (3.1.15) 的 KKT 条件

$$\alpha_i^* ((w^* \cdot x) + b^* - 1 + \xi_i^*) = 0, \quad \xi_i^* (C - \alpha_i^*) = 0, \quad i=1, \dots, l, \quad (3.1.19)$$

可以得到以下结论:

- 1) 不属于带有间隔的正确分类区的样本必是界上支持向量。
- 2) 属于间隔以外正确分类区的样本必不是支持向量。
- 3) 界内支持向量一定位于间隔边界上的正确分类区。

近似线性可分情况的最优分类面和支持向量如图 3.5 所示

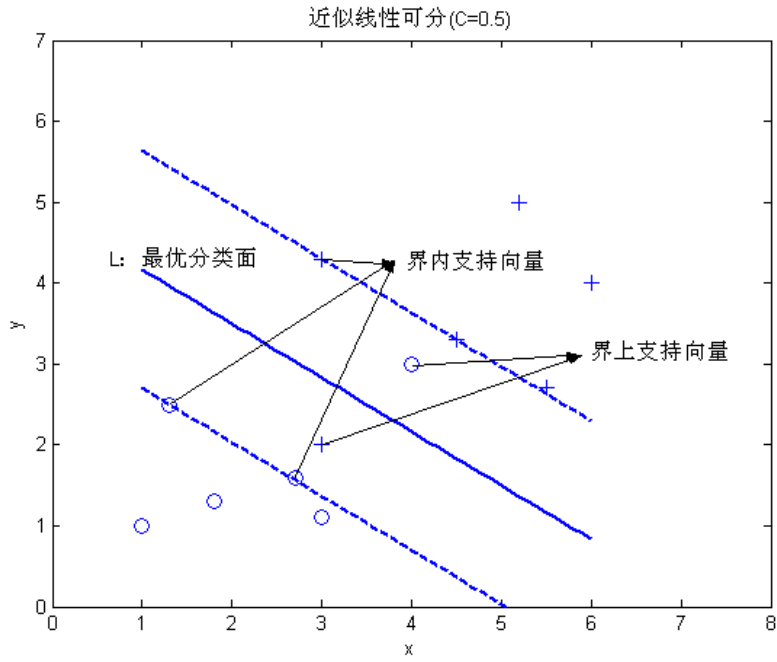


图 3.5

3.1.3 一般情况

在实际应用中存在一些分类问题，并不适合采用线性划分的方法进行分类，如图 3.6 所示的二维样本数据

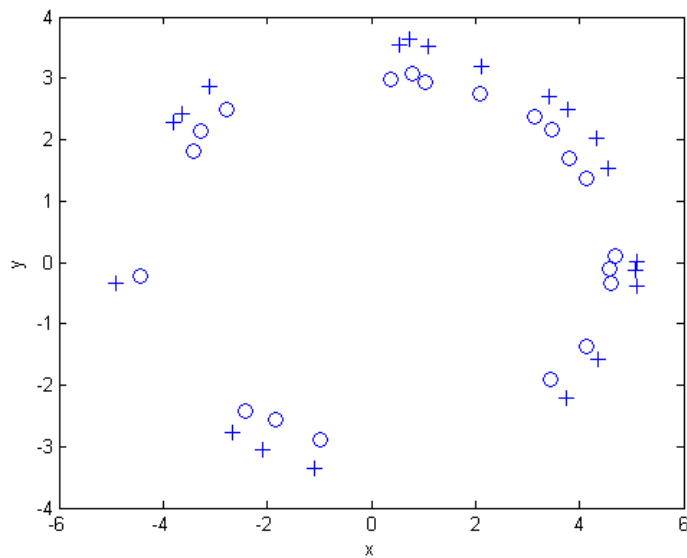


图 3.6

为了解决这一类分类问题，必须拓展前面所述的线性分类方法。观察图 3.6，

可以猜想合理的分类面可能是平面上的一个椭圆 $w_1x_1^2 + w_2x_2^2 + b = 0$ ，其中 w_1, w_2, b 都是参数。一个直观的方法就是在原样本数据上定义一个映射 $\phi: (x_1, x_2) \mapsto (x_1^2, x_2^2)$ ，然后在变换以后的数据基础上寻找最优的线性分类面。

下面考虑一般的情况，已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X = \mathbb{R}^n$ ， $y_i \in Y = \{1, -1\}$ ， $i = 1, 2, \dots, l$ ，以及空间 \mathbb{R}^n 上的映射 $\phi(x)$ ，这样得到变换以后的训练集 $T' = \{(\phi(x_1), y_1), \dots, (\phi(x_l), y_l)\} \in (X \times Y)^l$ ，在新得到的训练集上应用近似线性分类方法，得到下列原始优化问题

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (3.1.20)$$

$$s.t. \quad y_i((w \cdot \phi(x_i)) + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, l, \quad (3.1.21)$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, l, \quad (3.1.22)$$

其中惩罚参数 $C > 0$ 。

原始优化问题 (3.1.20) — (3.1.22) 的对偶问题是

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j (\phi(x_i) \cdot \phi(x_j)) - \sum_{i=1}^l \alpha_i \quad (3.1.23)$$

$$s.t. \quad \sum_{i=1}^l y_i \alpha_i = 0 \quad (3.1.24)$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, l. \quad (3.1.25)$$

设 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$ 是对偶问题 (3.1.23) — (3.1.25) 的任意解，如果存在 α^* 的分量 $\alpha_j^* \in (0, C)$ ，则原始优化问题 (3.1.20) — (3.1.22) 对 (w, b) 的解 (w^*, b^*) 存在且唯一，可以表示为 $w^* = \sum_{i=1}^l \alpha_i^* y_i \phi(x_i)$ ， $b^* = y_j - \sum_{i=1}^l y_i \alpha_i^* (\phi(x_i) \cdot \phi(x_j))$ ，从而得到最优分类超平面

$$w^* \cdot \phi(x) + b^* = \sum_{i=1}^l \alpha_i^* y_i (\phi(x) \cdot \phi(x_i)) + b^* = 0. \quad (3.1.26)$$

核函数：通过观察对偶问题 (3.1.23) — (3.1.25) 和最优分类面的表达式 (3.1.26)，发现在求解的过程中不涉及映射 $\phi(x)$ 的具体形式，而只涉及内积 $(\phi(x_i) \cdot \phi(x_j))$ ，不妨记这个内积函数为 $k(x_i, x_j)$ ，称之为核函数，有关核函数的具体内容将在 3.2 中介绍。在实际应用时，不需知道具体的映射公式 $\phi(x)$ ，而只需定义核函数 $k(x_i, x_j)$ ，这将大大简化计算的过程。

支持向量分类(SVC)

- i. 已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X = \mathbb{R}^n$ ， $y_i \in Y = \{1, -1\}$ ， $i = 1, 2, \dots, l$ ；

- ii. 选取适当的核函数 $K(x, x')$ 和参数 C ，构造并求解最优化问题

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{i=1}^l \alpha_i \quad (3.1.27)$$

$$s.t. \sum_{i=1}^l y_i \alpha_i = 0 \quad (3.1.28)$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \dots, l. \quad (3.1.29)$$

得到最优解 α^* ；

- iii. 选取 α^* 的一个正分量 $0 < \alpha_j^* < C$ ，计算 $b^* = y_i - \sum_{i=1}^l y_i \alpha_i^* K(x_i, x_j)$ 。

- iv. 构造分类超平面 $\sum_{i=1}^l \alpha_i^* y_i K(x, x_i) + b^* = 0$ ，得到决策函数

$$f(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i K(x, x_i) + b^*\right)。$$

对于本节开头的二维样本数据，选取核函数 $K(x, x') = (x \cdot x' + 1)^2$ ，惩罚参数

$C = 0.5$ ，得到的最优分类面如图 3.7 所示，可以看出分类的效果非常好。

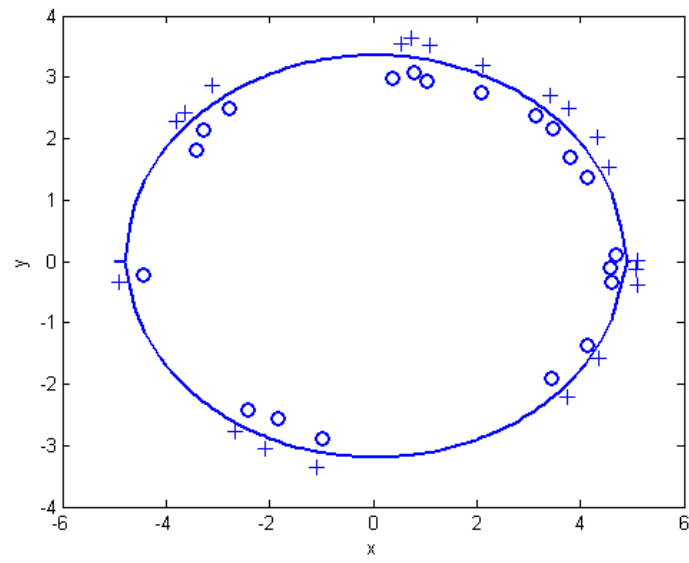


图 3.7

3.2 核函数

在介绍一般情况下的分类问题时，引入了核函数的概念，下面具体介绍核函数的有关内容。

在支持向量分类机中，需要选择核函数 $K(x, x')$ 或者是选择一个映射 $\phi(x)$ ，把 x 所在的输入空间 X 映射到另一个空间 H ， H 一般来说是一个 Hilbert 空间。在实际使用分类支持向量机时，核函数 $K(x, x')$ 起着直接作用，而不需要知道映射 $\phi(x)$ 具体是什么，只需直接选定核函数 $K(x, x')$ 。由于核函数强烈依赖于映射的选择，我们给出核函数的准确定义

核函数: 设 X 是 R^n 的一个子集，称定义在 $X \times X$ 上的函数 $K(x, x')$ 是核函数，如果存在从 X 到某一个 Hilbert 空间 H 的映射 $\phi(x)$ ，使得 $K(x, x') = (\phi(x) \cdot \phi(x'))$ 。

3.2.1 多项式空间和多项式核

我们在这节中先讨论一类特殊的 Hilbert 空间：多项式空间。

d 阶单项式: 设 $x = (x_1, x_2, \dots, x_n)^T \in R^n$ ，称乘积 $x_{j_1} x_{j_2} \dots x_{j_d}$ 为 x 的一个 d 阶单项式，其中 $j_1, j_2, \dots, j_d \in \{1, 2, \dots, n\}$ ， j_1, j_2, \dots, j_d 可以存在相同的数。

1. 齐次多项式空间

定义从 R^n 到齐次多项式空间 H 的映射

$$C_d: x = (x_1, x_2, \dots, x_n)^T \mapsto X = (x_{j_1} x_{j_2} \dots x_{j_d} \mid j_1, j_2, \dots, j_d \in \{1, 2, \dots, n\})^T \quad (3.2.1)$$

这样得到的空间 H 的维数 $n_H = n^d$ ，当 n 和 d 都不大时，空间 H 的维数 $n_H = n^d$ 可能会很大，如果直接在 H 上进行内积计算无疑是很困难的，考虑由 (3.2.1) 定义的从 R^n 到齐次多项式空间 H 的映射 C_d ，则空间 H 上的内积 $(C_d(x) \cdot C_d(x'))$ 可表

示为 $(C_d(x) \cdot C_d(x')) = K(x, x') = (x \cdot x')^d$ ，从而将空间 H 上的内积运算与原空间 R^n 上的内积联系起来，达到简化计算的目的。

2. 非齐次多项式空间

在 (3.2.1) 定义的 H 是由所有的 d 阶单项式组成的，把该空间的分量扩充到所有不超过 d 阶的多项式组成，得到从 R^n 到非齐次多项式空间 \tilde{H} 的映射

$$\tilde{C}_d: x = (x_1, x_2, \dots, x_n)^T \mapsto X = (x_{j_1} x_{j_2} \cdots x_{j_d}, \sqrt{d} x_{j_1} \cdots x_{j_{d-1}}, \dots, \sqrt{d} x_1, \dots, \sqrt{d} x_n, 1 \mid j_1, j_2, \dots, j_d \in \{1, 2, \dots, n\})^T \quad (3.2.2)$$

考虑由 (3.2.2) 定义的从 R^n 到非齐次多项式空间 \tilde{H} 的映射 \tilde{C}_d ，则空间 \tilde{H} 上的内积 $(\tilde{C}_d(x) \cdot \tilde{C}_d(x'))$ 可表示为 $(\tilde{C}_d(x) \cdot \tilde{C}_d(x')) = K(x, x') = (x \cdot x' + 1)^d$ 。

3.2.2 Mercer 核和正定核

1. Mercer 定理和 Mercer 核

设 X 是 R^n 的一个紧集，并设 $K(x, x')$ 为 $X \times X$ 上的连续对称函数

$$K(x, x') = K(x', x), \quad \forall x, x' \in X. \quad (3.2.3)$$

我们想知道当 $K(x, x')$ 满足什么条件时，它是一个核函数。

积分算子的半正定性： 给定满足上述条件的 $K(x, x')$ ，定义在 $L_2(X)$ 上的积分算子

$$T_K: T_K f(\cdot) = \int_X K(\cdot, x') f(x') dx', \quad \forall f \in L_2(X). \quad (3.2.4)$$

如果对于任意的 $f \in L_2(X)$ 有 $\int_{X \times X} K(x, x') f(x) f(x') dx dx' \geq 0$ 成立，则称积分算子

T_K 半正定。

Mercer 定理 (Mercer, 1909)： 设 X 是 R^n 的一个紧集，并设 $K(x, x')$ 为 $X \times X$ 上的连续对称函数，则 $K(x, x')$ 是一个核函数的充要条件是由 (3.2.4) 定义的积分算子

半正定, 即 $\int_{X \times X} K(x, x') f(x) f(x') dx dx' \geq 0, \forall f \in L_2(X)$, 并称 $K(x, x')$ 为 **Mercer 核**。如果 $K(x, x')$ 是一个 Mercer 核, 则存在一个映射 $\phi(x)$, 使得在相应的 Hilbert 空间 H 中有 $K(x, x') = (\phi(x), \phi(x'))$ 。

下面给出 Mercer 核满足的更易于验证的条件。

Gram 矩阵: 对于给定的函数 $K(x, x'): X \times X \mapsto R$ 和 $x_1, x_2, \dots, x_l \in X$, 称 $l \times l$ 矩阵 $K: K_{ij} = K(x_i, x_j)$ 为 $K(x, x')$ 关于 $x_1, x_2, \dots, x_l \in X$ 的 Gram 矩阵。

定理 3.2.1^[2] 设 X 是 R^n 的一个紧集, 并设 $K(x, x')$ 为 $X \times X$ 上的连续对称函数, 则 $K(x, x')$ 是一个 Mercer 核的充要条件是 $K(x, x')$ 关于任意的 $x_1, x_2, \dots, x_l \in X$ 的 Gram 矩阵半正定。

2. 正定核

设 X 是 R^n 的一个子集, 称定义在 $X \times X$ 上的函数 $K(x, x')$ 是核函数, 如果存在从 X 到某一个 Hilbert 空间 H 的映射 $\phi(x)$, 使得 $K(x, x') = (\phi(x), \phi(x'))$ 。

定理 3.2.2^[2] 设 X 是 R^n 的一个子集, 并设 $K(x, x')$ 为 $X \times X$ 上的连续对称函数, 则 $K(x, x')$ 是一个正定核的充要条件是 $K(x, x')$ 关于任意的 $x_1, x_2, \dots, x_l \in X$ 的 Gram 矩阵半正定。

如果函数 $K(x, x')$ 是一个 Mercer 核, 则对任何的 $c \in R^l$, 由 Mercer 定理知

$$\sum_{i,j} c_i c_j K(x_i, x_j) = \sum_{i,j} c_i c_j (\phi(x_i) \cdot \phi(x_j)) = \left\| \sum_i c_i \phi(x_i) \right\|^2 \geq 0, \text{ 因此 } K(x, x') \text{ 一定是正}$$

定核, 所以 Mercer 核一定是正定核。

3.2.3 核函数的构造原则及常用的核函数

核函数的构造原则：设 K_1, K_2 是 $X \times X$ 上的核函数， X 是 R^n 的一个子集，设常数 $a \geq 0$ ， $p(x)$ 为系数全为正数的多项式，则下列函数均为核：

$$1) \quad K(x, x') = K_1(x, x') + K_2(x, x') \quad (3.2.5)$$

$$2) \quad K(x, x') = aK_1(x, x') \quad (3.2.6)$$

$$3) \quad K(x, x') = K_1(x, x')K_2(x, x') \quad (3.2.7)$$

$$4) \quad K(x, x') = p(K_1(x, x')) \quad (3.2.8)$$

$$5) \quad K(x, x') = \exp(K_1(x, x')) \quad (3.2.9)$$

$$6) \quad K(x, x') = \exp(-\|x - x'\|^2 / \sigma^2) \quad (3.2.10)$$

常用的几种核函数

1. Gauss 径向基核

$$K(x, x') = \exp(-\|x - x'\|^2 / \sigma^2)$$

2. 多项式核

对于任给的正整数 d ，

$$\text{齐次多项式核：} \quad K(x, x') = (x \cdot x')^d$$

$$\text{非齐次多项式核：} \quad K(x, x') = (x \cdot x' + c)^d, \quad \text{其中 } c > 0$$

3. 傅立叶核

常用的傅立叶核有两种，它们都是一维傅立叶核的乘积，一维傅立叶核如下：

$$K_1(x, x') = \frac{1 - q^2}{2(1 - 2q \cos(x - x') + q^2)}, \quad \forall x, x' \in R, \quad \text{其中 } 0 < q < 1. \quad (3.2.11)$$

$$K_2(x, x') = \frac{\pi \cosh(\frac{\pi - |x - x'|}{\gamma})}{2\gamma \sinh(\frac{\pi}{\gamma})}, \quad \text{其中 } \gamma \text{ 为常数。} \quad (3.2.12)$$

4. Sigmoid 核

$$K(x, x') = \tanh(\kappa(x \cdot x') + \nu), \text{ 其中 } \kappa > 0, \nu < 0, \quad (3.2.13)$$

它不是正定核，但在某些实际应用中很有效。

3.3 VC 维和结构风险最小化原则

前面介绍的分类问题是根据已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X \subset \mathbb{R}^n$, $y_i \in Y = \{1, -1\}$, $i = 1, 2, \dots, l$, 求出决策函数 $f(x)$ 。对于得到的决策函数 $f(x)$, 我们希望能对其优劣性给出一个数量化的评价。

3.3.1 损失函数和期望风险

损失函数是评价预测准确程度的一种度量, 这里的预测是根据某个决策函数推断出的结果, 因此损失函数与决策函数 $f(x)$ 密切相关。

损失函数: $X \subset \mathbb{R}^n$, $Y = \{1, -1\}$, $f(x)$ 为决策函数, 定义映射 $L: X \times Y \times Y \mapsto [0, \infty)$ 使得对任意 $x \in X$, $y \in Y$, 都有 $L(x, y, y) = 0$, 称 L 为一个损失函数。常用于分

类问题的是 0-1 损失函数: $L(x, y, f(x)) = L(y - f(x)) = \begin{cases} 0 & y = f(x) \\ 1 & y \neq f(x) \end{cases}$ 。

期望风险: 设 $P(x, y)$ 为样本数据 T 在 $X \times Y$ 上的概率分布, L 为损失函数,

$f(x)$ 是定义 X 上的决策函数, 则决策函数 $f(x)$ 的期望风险

$$R(f) = E[L(x, y, f(x))] = \int_{X \times Y} L(x, y, f(x)) dP(x, y) \quad (3.3.1)$$

3.3.2 经验风险和函数集的 VC 维

(3.3.1) 式给出的期望风险 $R(f)$ 给出了刻画决策函数 $f(x)$ 优劣的准则, 但是由于 $P(x, y)$ 未知, 而仅仅知道训练集 T , 所以只能计算出 $f(x)$ 在这些样本点上的偏差。

1. 经验风险

已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X \subset \mathbb{R}^n$ ， $y_i \in Y = \{1, -1\}$ ， $i = 1, 2, \dots, l$ ，给定损失函数 L ，则决策函数 $f(x)$ 的经验风险

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^l L(x_i, y_i, f(x_i)) \quad (3.3.2)$$

经验风险最小化原则：任给训练集 T ，适当选定决策函数构成的集合 F ，经验风险最小化原则就是：在 F 中选择使经验风险 $R_{emp}(f)$ 达到最小的 $f(x)$ 。要注意不能只关心 $f(x)$ 而忽视对 F 的选择，考虑一种极端情况，假如 F 包含决策函数

$$\tilde{f}(x) = \begin{cases} y_i & x = x_i, i = 1, \dots, l \\ 1 & \text{其它} \end{cases}, \text{ 其中 } (x_i, y_i) \in T, \text{ 则这个决策函数的经验风险}$$

$R_{emp}(\tilde{f}) = 0$ ，但是很显然这个决策函数是不合理的，所以在使用经验风险最小化原则时要对假设 F 进行限制，至少不能包括 $\tilde{f}(x)$ 。

2. 函数集的VC维

在经验风险最小化原则中，需要选定适当的函数集 F ，其中关键的因素是 F 的“表达能力”或丰富程度。由 Vapnik 和 Chervonenkis (1968, 1971) 提出的 VC 维，是对 F 丰富程度的一个描述，在介绍 VC 维之前，先引入打散的概念。

打散：设 F 是由 $X \subset \mathbb{R}^n$ 上取值为 -1 和 1 的若干决策函数组成的集合， $Z_m = \{x_1, \dots, x_m\}$ 为 X 中 m 个点组成的集合，考虑当 f 取遍 F 中所有的函数产生的 m 维向量 $(f(x_1), \dots, f(x_m))$ ，定义 $N(F, Z_m)$ 为上述 m 维向量中不同向量的个数，如果 $N(F, Z_m) = 2^m$ ，则称 Z_m 被 F 打散，或 F 打散 Z_m 。二维空间的例子如图 3.8 所示， F 为二维线性函数的集合， F 将 Z_3 打散。

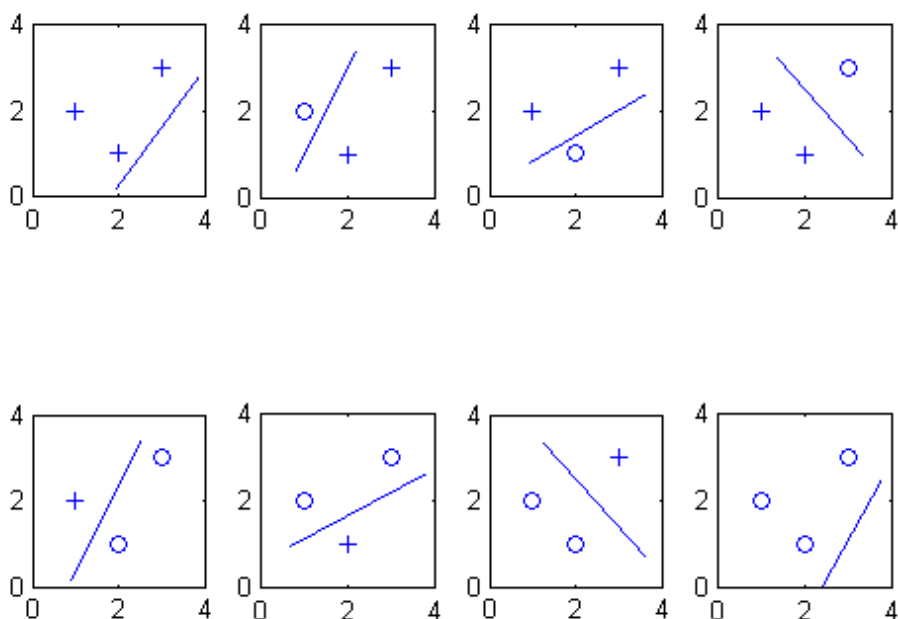


图 3.8

VC 维: 设 F 是由 $X \subset R^n$ 上取值为 -1 和 1 的若干决策函数组成的集合, $Z_m = \{x_1, \dots, x_m\}, x_i \in X$, 令 $N(F, m) = \max\{N(F, Z_m) : Z_m \subset X\}$, 定义 F 的 VC 维为

$$VC \dim(F) = \max\{m : N(F, m) = 2^m\} \tag{3.3.3}$$

设 F 是 n 维空间 R^n 上的线性指示函数的集合, 即

$$F = \{f(x, \alpha, c) = \text{sgn}((x' \alpha) + c)\}, \alpha \in R^n, c \in R, \tag{3.3.4}$$

则 F 的 VC 维为 $n+1$ 。

3.3.3 结构风险最小化原则

在本小节中先考虑经验风险和期望风险的关系, 然后再引出结构风险的概念。

经验风险最小化原则的一致性:

设训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X \subset \mathbb{R}^n$, $y_i \in Y = \{1, -1\}$, $i = 1, 2, \dots, l$, 为来自概率分布 $P(x, y)$ 的独立同分布样本数据, F 为一个函数集合, 记 $f^l (l = 1, 2, \dots)$ 是 F 中使经验风险 $R_{emp}(f) = \frac{1}{l} \sum_{i=1}^l L(x_i, y_i, f(x_i))$ 达到最小的函数。如对任意的 $\varepsilon > 0$, 有

$$\lim_{l \rightarrow \infty} P\{R(f^l) - \inf_f R(f) > \varepsilon\} = 0 \quad (3.3.5)$$

$$\lim_{l \rightarrow \infty} P\{R_{emp}(f^l) - \inf_f R(f) > \varepsilon\} = 0 \quad (3.3.6)$$

则称经验风险最小化原则对函数集 F 和概率分布 $P(x, y)$ 是一致的。

一致性概念的示意图如图 3.9 所示

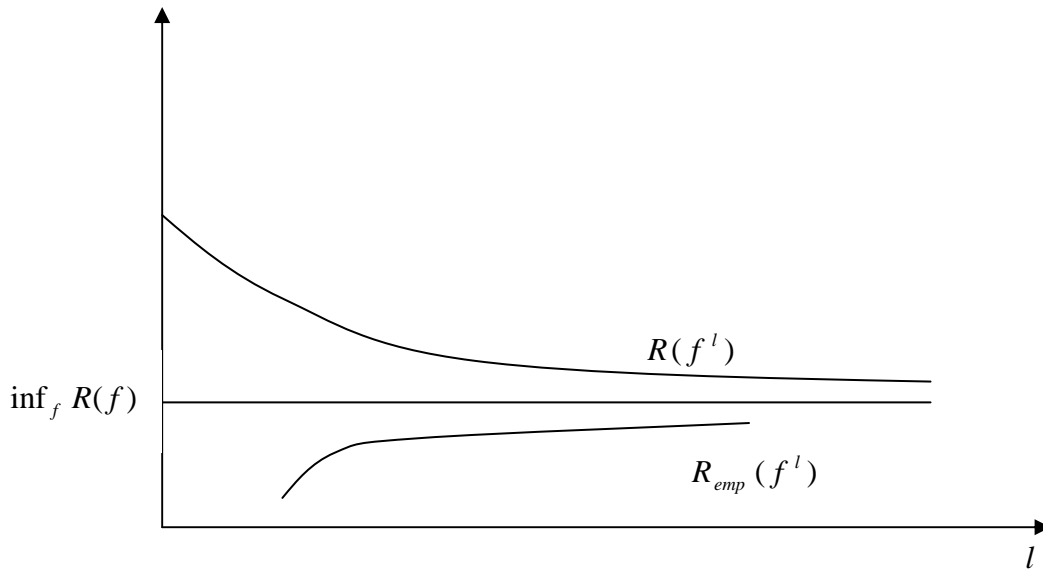


图 3.9

下列定理给出了一致性成立的充要条件 (Vapnik and Chervonenkis, 1989)

定理 3.3.2 设训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X \subset \mathbb{R}^n$, $y_i \in Y = \{1, -1\}$, $i = 1, 2, \dots, l$ 为来自概率分布 $P(x, y)$ 的独立同分布样本数据, F 为一个函数集合, 如对任意的 $\varepsilon > 0$, 都有

$$\lim_{l \rightarrow \infty} P\{\sup_{f \in F} (R(f) - R_{emp}(f)) > \varepsilon\} = 0 \quad (3.3.7)$$

则经验风险最小化原则对 F 和 $P(x, y)$ 是一致的。

定理 3.3.2 给出了经验风险最小化原则对 F 和 $P(x, y)$ 一致性的一个充分条件，下面的结论 (Vapnik 1979, 1998) 对式 (3.3.7) 中概率值大小给出一个估计，并由此引出结构风险的概念。

定理 3.3.3 $T, F, P(x, y)$ 的定义与定理 3.3.2 相同，记 h 为 F 的 VC 维，对任意的 $\varepsilon > 0$ ，如果 $l\varepsilon^2 \geq 2$ ， $l > h$ ，则有

$$P\{\sup_{f \in F} (R(f) - R_{emp}(f)) > \varepsilon\} \leq 4 \exp(h(\ln \frac{2l}{h} + 1) - \frac{l\varepsilon^2}{8}) \quad (3.3.8)$$

显然，由于 h 是一个定值，当 $l \rightarrow \infty$ 时，式 (3.3.8) 右端也趋向于 0。

令 $4 \exp(h(\ln \frac{2l}{h} + 1) - \frac{l\varepsilon^2}{8}) = \delta$ ，解得 $\varepsilon = \sqrt{\frac{8}{l}(h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta})}$ ，由 $l\varepsilon^2 \geq 2$ 限制条件得 $h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta} \geq \frac{1}{4}$ 。如 $l > h$ ， $h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta} \geq \frac{1}{4}$ ，则对于任何的概率分布 $P(x, y)$ 、任意的 $\delta \in (0, 1]$ 和 F 中的任意函数 f ，都能使下列不等式至少以 $1 - \delta$ 的概率成立

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{8}{l}(h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta})} \quad (3.3.9)$$

结构风险：称 $\sqrt{\frac{8}{l}(h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta})}$ 为置信区间， $R_{emp}(f) + \sqrt{\frac{8}{l}(h(\ln \frac{2l}{h} + 1) + \ln \frac{4}{\delta})}$ 为结构风险，它是经验风险和置信区间之和，是期望风险 $R(f)$ 的一个上界；经验风险 $R_{emp}(f)$ 依赖于 F ， F 越大 $R_{emp}(f)$ 可以越小，同时由于 F 的 VC 维 h 增大使得置信区间也增大；反之 F 越小，置信区间会减小，但同时 $R_{emp}(f)$ 可能较大，经验风险和置信区间有相互矛盾的倾向，为了兼顾两者，引入了结构风险最小化原则。

结构风险最小化原则: 结构风险最小化原则是寻找一个函数 f 使得其对应的结构风险最小, 为了寻找这样的 f , 适当选择一个嵌套的函数集 $\cdots F_{n-1} \subset F_n \subset F_{n+1} \cdots$, 在每个 F_n 中找到使得经验风险 $R_{emp}(f)$ 最小的函数 f_n , 得到一系列的函数 $\cdots f_{n-1}, f_n, f_{n+1} \cdots$, 观察与 f_n 与对应的结构风险随 n 的变化情况, 可以发现:

- 1) 置信区间随着 n 的增大而增大, 因为 F_n 的 VC 维 h_n 是递增的。
- 2) 经验风险随着 n 的增大而减小, 因为 F_n 是嵌套的。

最后结构风险最小化原则转化为寻找适当的 n^* , 使得置信区间和经验风险的和最小, 由此得到相应的 f_{n^*} 。

最后结构风险最小化原则的示意图如图 3.10

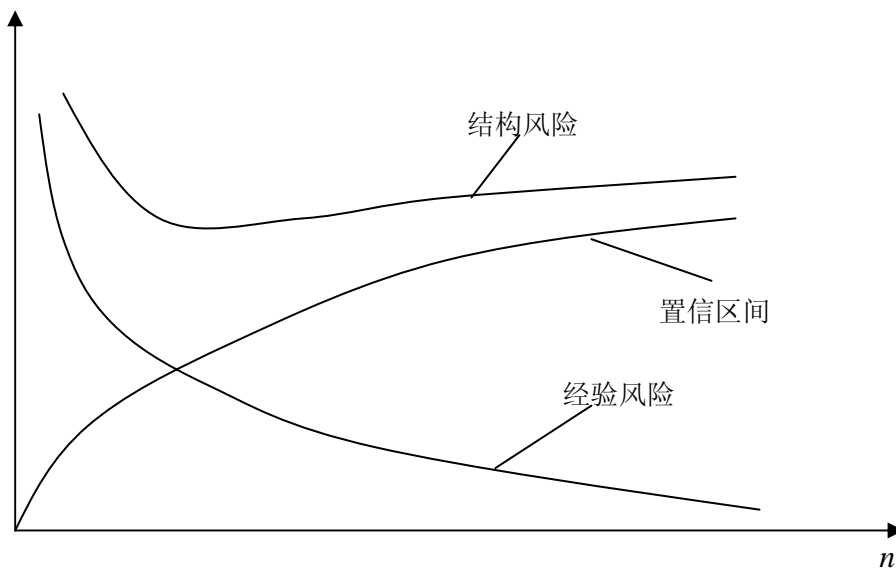


图 3.10

3.4 支持向量回归

与支持向量分类问题相似，在回归问题中给出训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$ ，其中 $x_i \in X \subset \mathbb{R}^n$ ， $y_i \in Y = \mathbb{R}$ ，这与分类问题稍有不同， $i = 1, 2, \dots, l$ 。

回归问题就是要在一个适当的函数集 F 中寻找一个合适的函数 $f(x)$ 使得 $f(x_i)$ 与 y_i 尽可能的接近。与分类问题相类似，我们定义损失函数来刻画 $f(x_i)$ 偏离 y_i 的程度，记损失函数为 $L(x, y, f(x))$ ，则常规回归问题转化为经验风险的最

优化问题：
$$\min_{f \in F} \sum_{i=1}^l L(x_i, y_i, f(x_i))$$
，我们可以定义不同的损失函数得到不同的拟合函数 $f(x)$ 。

与分类问题相似的，函数集 F 的选择对于回归问题的计算极为重要。 F 如果取的太小，虽然函数的形式比较简单，但拟合样本数据的精度不够； F 如果取的太大，拟合样本数据的精度可以很高，最极端的情况，定义

$$\tilde{f}(x) = \begin{cases} y_i & x = x_i, i = 1, \dots, l \\ 1 & \text{其它} \end{cases}, \text{ 其中 } (x_i, y_i) \in T, \text{ 则 } \tilde{f}(x) \text{ 可以完全精确的拟合样}$$

本数据，但是很显然 $\tilde{f}(x)$ 太复杂不具有实用性，它的过拟合现象特别严重，对于一个新的数据，将完全起不到预测的作用，这是因为函数集 F 的 VC 维过高，从而导致置信区间和结构风险过大的缘故。

在支持向量机中，引入了 $\frac{1}{2} \|w\|^2$ 来控制最优化函数的复杂度，再加上分类或者拟合的损失构成了总体的目标函数，这是对结构化风险的一个刻画，通过最小化总体目标函数使得整体的结构化风险得到有效的控制，所以通过支持向量机得到的决策函数具有较好的推广能力。

3.4.1 常用的损失函数

常用的损失函数有以下几种

$$1) \quad \varepsilon\text{-不敏感函数: } L(x) = \begin{cases} 0 & |x| \leq \varepsilon \\ |x| - \varepsilon & |x| > \varepsilon \end{cases} \quad (3.4.1)$$

$$2) \quad \text{拉普拉斯函数: } L(x) = |x| \quad (3.4.2)$$

$$3) \quad \text{高斯函数: } L(x) = \frac{1}{2}x^2 \quad (3.4.3)$$

$$4) \quad \text{鲁棒函数: } L(x) = \begin{cases} \frac{1}{2\sigma}x^2 & |x| \leq \sigma \\ |x| - \frac{\sigma}{2} & |x| > \sigma \end{cases} \quad (3.4.4)$$

函数图像如图 3.11 所示:

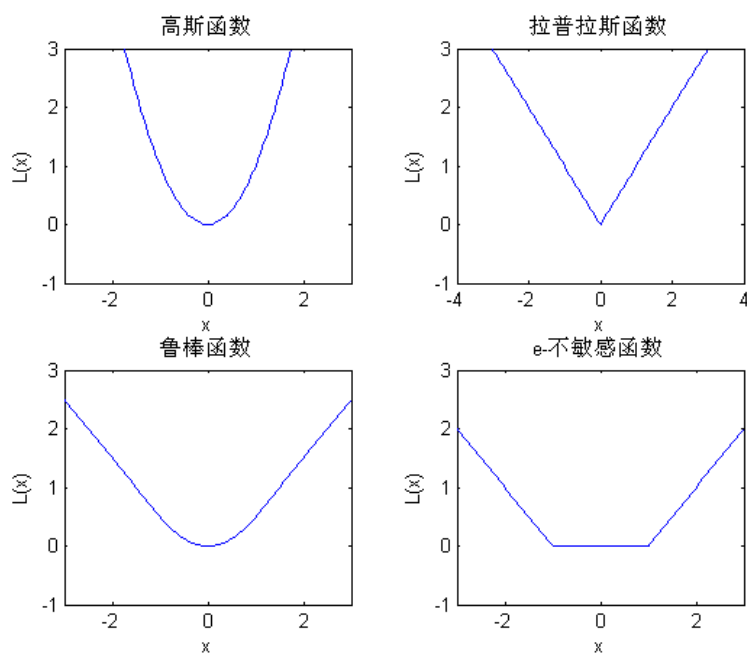


图 3.11

3.4.2 支持向量回归

支持向量回归来源于支持向量分类，在分类问题中，一般来说可以用较少的支持向量来表示决策函数，算法具有稀疏性。我们希望能把这一性质推广到回归问题中，在实际应用中，采用 ε -不敏感损失函数往往可以达到这一点，下面先

就 ε -不敏感损失函数展开讨论支持向量回归，最后给出对于一般损失函数的结论。

1. 线性 ε -带支持向量回归

考虑一类最简单的情况，样本点都在一条直线 $f(x) = (w \cdot x) + b$ 所构成的 ε -带 $f(x) \pm \varepsilon$ 以内，如图 3.12 所示，

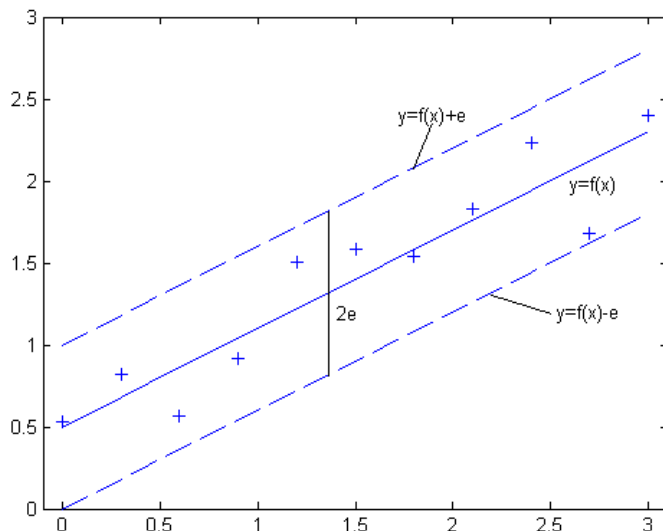


图 3.12

假设 ε_{\min} 是下列优化问题的解

$$\min_{w,b} \varepsilon \tag{3.4.5}$$

$$s.t. \quad -\varepsilon \leq y_i - (w \cdot x_i) - b \leq \varepsilon, \quad i = 1, 2, \dots, l. \tag{3.4.6}$$

选择合适的 ε 使得 $\varepsilon > \varepsilon_{\min}$ ，求解最优回归直线的问题转化为

$$\min_{w,b} \frac{1}{2} \|w\|^2 \tag{3.4.7}$$

$$s.t. \quad y_i - (w \cdot x_i) - b \leq \varepsilon, \quad i = 1, 2, \dots, l, \tag{3.4.8}$$

$$(w \cdot x_i) + b - y_i \leq \varepsilon, \quad i = 1, 2, \dots, l. \tag{3.4.9}$$

如果 $\varepsilon > \varepsilon_{\min}$ 满足，则优化问题(3.4.7)–(3.4.9)的解存在，且对于 w 是唯一的。

原优化问题(3.4.7)–(3.4.9)的对偶问题是

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,f=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) \quad (3.4.10)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.11)$$

$$\alpha_i^* \geq 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, l. \quad (3.4.12)$$

设 $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)$ 是对偶问题(3.4.10)—(3.4.12)的任意解, 令

$$\tilde{w} = \sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) x_i, \quad (3.4.13)$$

$$\text{并选择 } \tilde{\alpha} \text{ 的正分量 } \tilde{\alpha}_i > 0 \text{ 或者 } \tilde{\alpha}_i^* > 0, \text{ 令 } \tilde{b} = y_i - (\tilde{w} \cdot x_i) - \varepsilon, \quad (3.4.14)$$

则 (\tilde{w}, \tilde{b}) 是原优化问题(3.4.7)—(3.4.9)的解, 并 $\tilde{\alpha}_i > 0$ 或者 $\tilde{\alpha}_i^* > 0$ 对应的样本点 (x_i, y_i) 为原优化问题的支持向量。

2. “软化”线性 ε -带支持向量回归

与分类问题相类似, 通过引入松弛变量“软化”约束条件(3.4.8)—(3.4.9)得到“软化”的线性 ε -带支持向量回归, 选取合适的惩罚参数 C ,

其优化问题如下:

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (3.4.15)$$

$$s.t. \quad y_i - (w \cdot x_i) - b \leq \varepsilon + \xi_i^*, \quad i = 1, 2, \dots, l, \quad (3.4.16)$$

$$(w \cdot x_i) + b - y_i \leq \varepsilon + \xi_i, \quad i = 1, 2, \dots, l. \quad (3.4.17)$$

$$\xi_i^* \geq 0, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, l \quad (3.4.18)$$

优化问题(3.4.15)—(3.4.18)的解存在, 且对于 w 是唯一的, 对偶问题为

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,f=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) \quad (3.4.19)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.20)$$

$$C \geq \alpha_i^* \geq 0, \quad C \geq \alpha_i \geq 0, \quad i=1,2,\dots,l \quad (3.4.21)$$

设 $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)$ 是对偶问题 (3.4.19) — (3.4.21) 的任意解, 令

$$\tilde{w} = \sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) x_i, \quad (3.4.22)$$

并选择 $\tilde{\alpha}$ 的正分量 $C > \tilde{\alpha}_i > 0$ 或者 $C > \tilde{\alpha}_i^* > 0$, 令

$$\tilde{b} = y_i - (\tilde{w} \cdot x_i) - \varepsilon, \quad (3.4.23)$$

则 (\tilde{w}, \tilde{b}) 是原优化问题 (3.4.15) — (3.4.18) 的解, 并且 $\tilde{\alpha}_i > 0$ 或者 $\tilde{\alpha}_i^* > 0$ 对应的样本点 (x_i, y_i) 为原优化问题 (3.4.15) — (3.4.18) 的支持向量。

3. 非线性 ε -带支持向量回归

考虑将上述线性回归的方法推广到非线性的情况。对偶问题 (3.4.19) — (3.4.21) 只涉及内积运算 (x_i, x_j) , 类似于分类问题, 引入核函数, 从而把线性回归问题推广到处理非线性回归问题, 得到非线性 ε -带支持向量回归:

(1) 已知训练集 $T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l$, 其中 $x_i \in X = \mathbb{R}^n$, $y_i \in Y = \mathbb{R}$, $i=1,2,\dots,l$;

(2) 选取适当的核函数 $K(x, x')$ 和参数 C, ε , 构造并求解最优化问题

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,f=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i \cdot x_j) + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) \quad (3.4.24)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.25)$$

$$C \geq \alpha_i^* \geq 0, \quad C \geq \alpha_i \geq 0, \quad i=1,2,\dots,l \quad (3.4.26)$$

得到最优解 $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_1^*, \dots, \tilde{\alpha}_l, \tilde{\alpha}_l^*)$ 。

(3) 选取 $\tilde{\alpha}$ 的一个正分量 $0 < \tilde{\alpha}_j < C$ 或者 $0 < \tilde{\alpha}_j^* < C$, 计算

$$\tilde{b} = y_j - \sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_i, x_j) + \varepsilon. \quad (3.4.27)$$

$$(4) \text{ 构造决策函数 } f(x) = \sum_{i=1}^l (\tilde{\alpha}_i^* - \tilde{\alpha}_i) K(x_i, x) + \tilde{b}. \quad (3.4.28)$$

4. 一般损失函数对应的支持向量回归

现在考虑一般损失函数对应的支持向量回归，记损失函数为 $L(x)$ ，则原始优化问题为

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (L(\xi_i) + L(\xi_i^*)) \quad (3.4.29)$$

$$s.t. \quad y_i - (w \cdot x_i) - b \leq \xi_i^*, \quad i = 1, 2, \dots, l, \quad (3.4.30)$$

$$(w \cdot x_i) + b - y_i \leq \xi_i, \quad i = 1, 2, \dots, l. \quad (3.4.31)$$

$$\xi_i^* \geq 0, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, l \quad (3.4.32)$$

其对应的对偶问题为

$$\begin{aligned} \min_{\alpha, \alpha^*} \quad & \frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) \\ & - C \sum_{i=1}^l (T(\xi_i^*) + T(\xi_i)) \end{aligned} \quad (3.4.33)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.34)$$

$$\alpha_i^* \leq C \frac{dL(\xi_i^*)}{d\xi_i^*}, \quad \alpha_i \leq C \frac{dL(\xi_i)}{d\xi_i}, \quad i = 1, 2, \dots, l \quad (3.4.35)$$

$$\xi_i^* = \inf\{\xi_i^* \mid C \frac{dL(\xi_i^*)}{d\xi_i^*} \geq \alpha_i^*\}, \quad i = 1, 2, \dots, l \quad (3.4.36)$$

$$\xi_i = \inf\{\xi_i \mid C \frac{dL(\xi_i)}{d\xi_i} \geq \alpha_i\}, \quad i = 1, 2, \dots, l \quad (3.4.37)$$

$$\xi_i^* \geq 0, \quad \xi_i \geq 0, \quad \alpha_i^* \geq 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, l \quad (3.4.38)$$

考虑两种常见的损失函数：拉普拉斯函数和高斯函数，并引入核函数。

拉普拉斯函数

$L(x) = |x|$, $x \in \mathbb{R}$, 对偶问题(3.4.33) — (3.4.38)的具体形式为

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i \cdot x_j) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) \quad (3.4.39)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.40)$$

$$0 \leq \alpha_i^* \leq C, \quad 0 \leq \alpha_i \leq C, \quad i=1,2,\dots,l \quad (3.4.41)$$

高斯函数

$L(x) = x^2$, $x \in \mathbb{R}$, 对偶问题(3.4.33) — (3.4.38)的具体形式为

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i \cdot x_j) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i) + \frac{1}{2C} \sum_{i=1}^l (\alpha_i^2 + (\alpha_i^*)^2) \quad (3.4.42)$$

$$s.t. \quad \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \quad (3.4.43)$$

$$\alpha_i^* \geq 0, \quad \alpha_i \geq 0, \quad i=1,2,\dots,l \quad (3.4.44)$$

5. 一个简单的例子

考虑下列原始数据集(图 3.13)

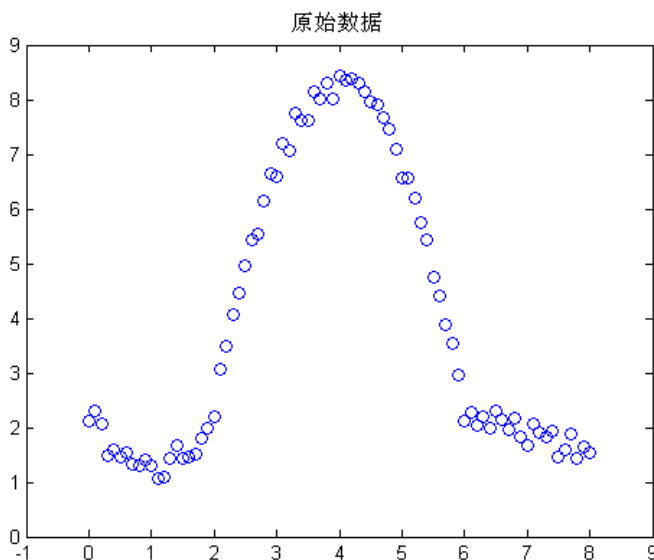


图 3.13

采用支持向量回归建立回归模型, 选取核函数为高斯函数, 参数为 p 分别等

于 1、2、0.5，损失函数为 ε -不敏感函数，参数为 $\varepsilon = 0.3$ ，惩罚参数为 $C = 1$ ，得到的结果如图 3.13-图 3.16 所示， p 值越大，得到的拟合函数越平滑，但拟合的效果不好；相反的 p 值越小，拟合的函数越不光滑，甚至会出现振荡的情况，但拟合的效果良好，可以看出选择合适的参数 p 对于建立正确的模型极为重要，在本例中， $p = 1$ 显得更为合理。

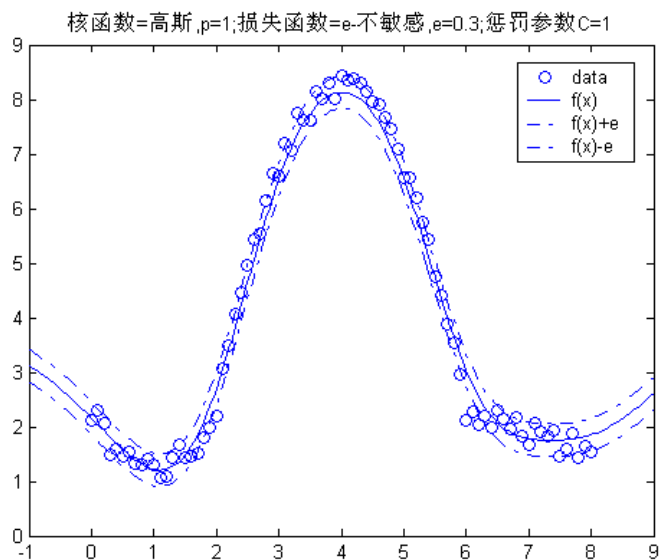


图 3.14

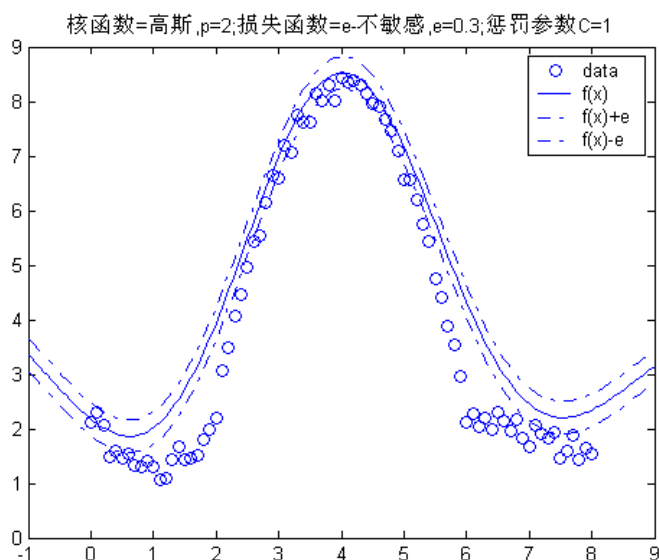


图 3.15

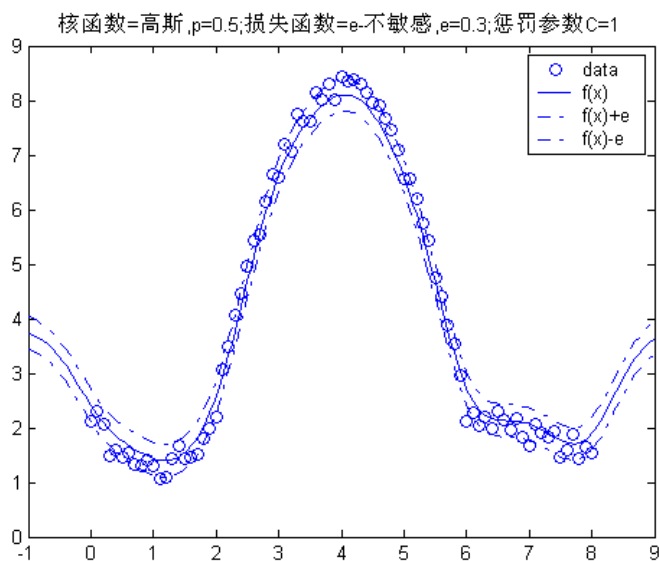


图 3.16

也可以选择 ε -不敏感函数以外的损失函数，如拉普拉斯函数，核函数为高斯函数，参数为 p 分别等于 1、2、0.5，惩罚参数为 $C = 1、10$ ，得到的结果如图 3.17-图 3.20 所示，与 ε -不敏感函数相类似，选择合适的参数 p 对于建立正确的模型极为重要， $p = 1$ 显得更为合理，其中惩罚参数 C 的选择在本例中对整体的回归函数影响不大。

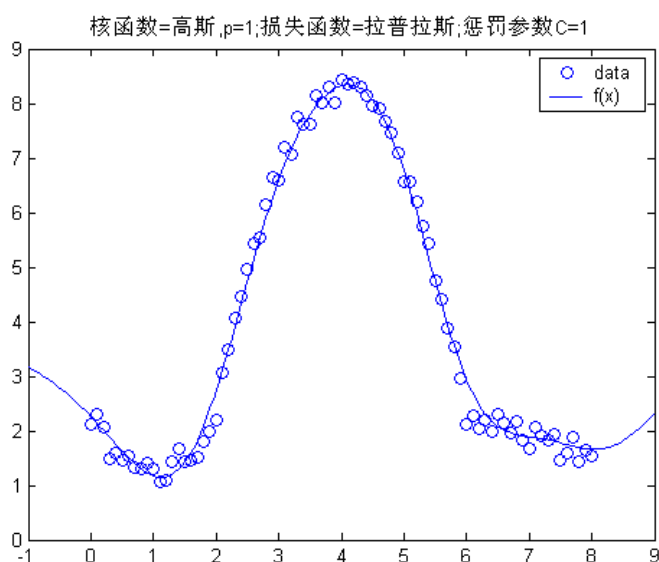


图 3.17

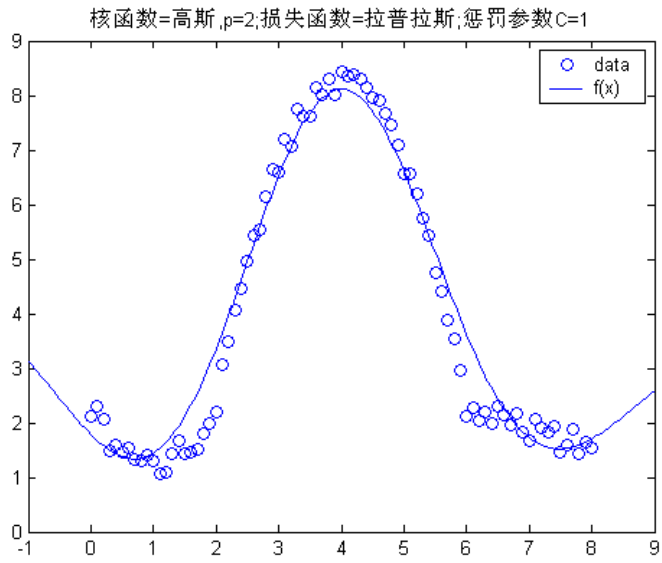


图 3.18

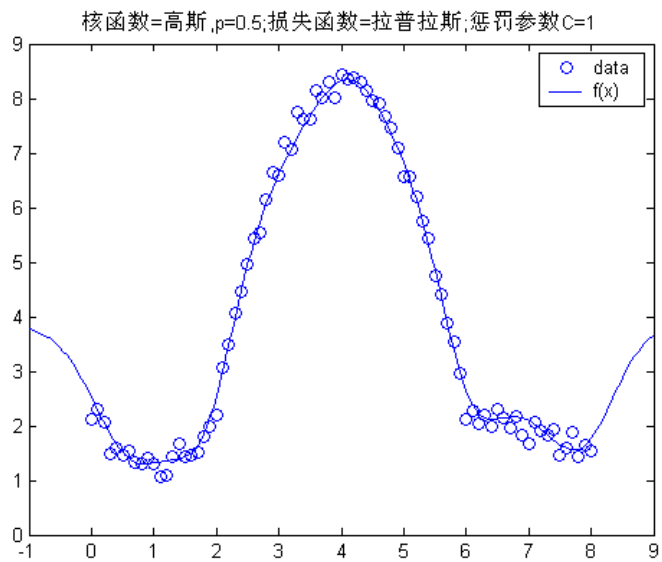


图 3.19

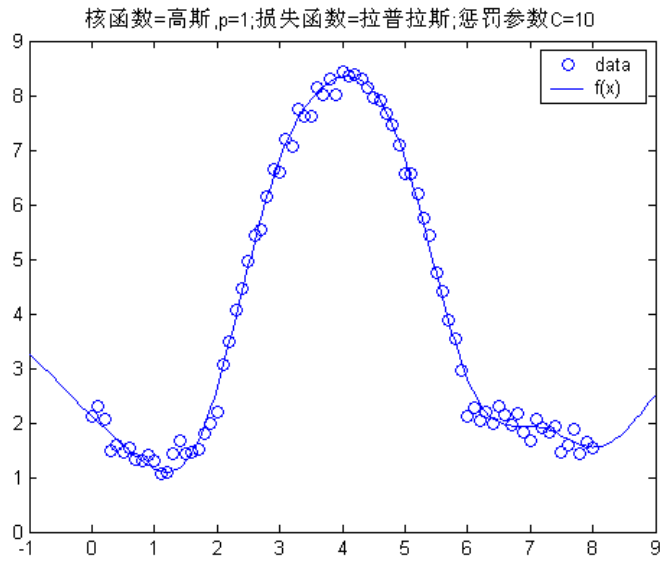


图 3.20

第四章 实验结果

实验数据：现阶段已有 31 个样本数据，其中临床因素包括原发病灶在治疗前后的体积大小、三维中心位置坐标以及患者性别。

4.1 线性回归

4.1.1 样本所含变量相关性分析

样本所含变量的相关性矩阵如表 4.1 所示（只显示上三角部分）：

治疗前				性别	治疗后				退缩比例
x	y	z	v		x'	y'	z'	v'	
1.00	-0.08	-0.16	0.27	-0.16	0.95	0.11	0.02	0.08	-0.12
	1.00	-0.18	-0.20	-0.50	-0.08	0.94	-0.17	-0.14	-0.09
		1.00	0.02	-0.06	-0.07	-0.24	0.82	0.20	0.31
			1.00	-0.03	0.21	-0.10	0.02	0.78	0.03
				1.00	-0.20	-0.49	-0.11	-0.03	0.08
					1.00	0.10	0.09	0.07	-0.10
						1.00	-0.25	-0.07	-0.02
							1.00	0.22	0.30
								1.00	0.56
									1.00

表 4.1

从表 4.1 中可以看出治疗后原发灶的中心坐标分量与治疗前原发灶对应的坐标分量具有很强的线性相关性，治疗后残留灶的体积也与治疗前的体积具有较强的相关性，这是可以解释的：从样本数据上看，原发灶的中心坐标移动量与本身的坐标相比一般都较小，治疗前后的坐标位置一般变化不大；对于体积来说，治

疗后的体积相比治疗前按一定的比例缩小，导致其相关系数较大，但由于缩小的比例差异性很大，所以体积变量表现出的相关性不如坐标变量。

如果直接对坐标因子 x' 、 y' 和 z' 进行回归，回归系数将分别主要由 x 、 y 和 z 决定，其他因子的作用可能会被忽略，所以我们改变因变量因子，将其改为个分量上的改变量 $x'-x$ 、 $y'-y$ 和 $z'-z$ ，仍将其记为 x' 、 y' 和 z' ，我们的目的是要回归在原有坐标基础上的改变量。体积因素采用的是体积退缩比例，改变因子以后的相关性分析如表 4.2 所示，可以看出目标因子与任一自变量因子都不存在很强的相关性，同时自变量因子之间也不存在很强的相关性。

治疗前				性别	治疗后				退缩比例
x	y	z	v		x'	y'	z'	v'	
1.00	-0.08	-0.16	0.27	-0.16	-0.23	0.52	0.31	-0.34	-0.12
	1.00	-0.18	-0.20	-0.50	0.01	0.09	0.04	0.19	-0.09
		1.00	0.02	-0.06	0.30	-0.21	-0.38	0.12	0.31
			1.00	-0.03	-0.23	0.21	0.01	-0.90	0.03
				1.00	-0.11	-0.10	-0.07	0.02	0.08
					1.00	-0.15	-0.12	0.30	0.07
						1.00	-0.10	-0.20	0.17
							1.00	-0.01	-0.05
								1.00	0.35
									1.00

表 4.2

4.1.2 线性回归分析

1. 坐标分量 X

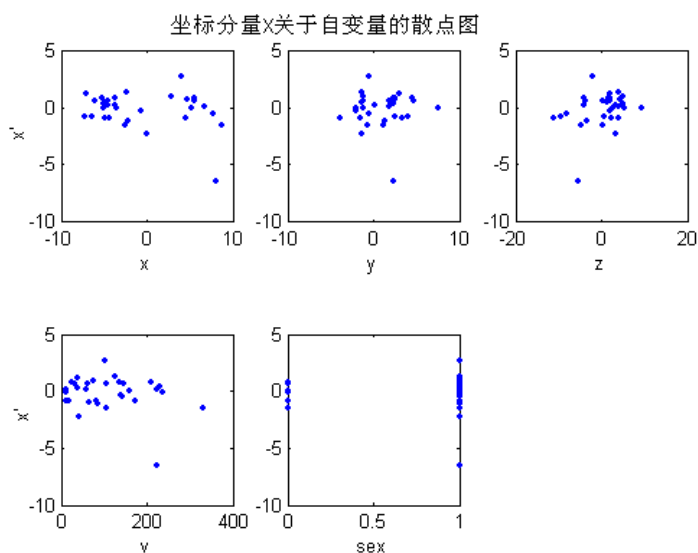


图 4.2

从散点图 4.2 上可以看出，坐标分量 x' 与自变量都不存在明显的线性关系，而且很有可能存在异常值点，回归结果如下：

系数	估计	标准差	t 值	p 值
截距	0.657185	0.859755	0.764	0.452
x	-0.050212	0.058933	-0.852	0.402
y	-0.049208	0.139554	-0.353	0.727
z	0.079935	0.062312	1.283	0.211
v	-0.003983	0.003655	-1.090	0.286
sex	-0.569444	0.772213	-0.737	0.468

标准误差：1.537，自由度 25

复相关系数的平方：0.1772

F-统计量：1.077，自由度为 (5, 25)，p 值：0.3971

从结果上看，线性回归的系数都不显著，复相关系数的平方很小，说明线性回归对于此样本数据不是很合适，数据本身可能存在未知的非线性关系。即使如此，

我们可以利用线性回归剔除样本中特别显著的异常值点,为进一步采用其他的分析方法打好基础。

线性回归残差对拟合值、QQ 图、标准残差对拟合值、Cook 距离图如图 4.3 所示:

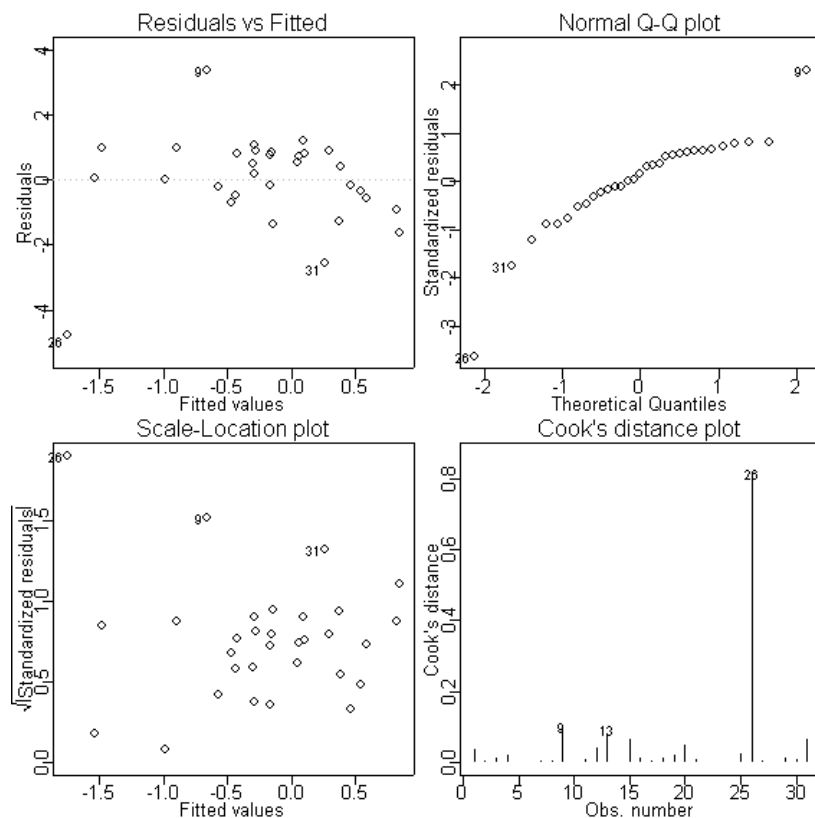


图 4.3

从残差对拟合值图上可以看出 9 号、26 号和 31 号样本很有可能是一个奇异点,对其采取异常值检验,其 p 值为 $p_1=0.0019563$, $p_2=1.56e-006$, $p_3=0.0079185$ 。26 号样本是最为显著的异常点,在以后的数据处理中将不会再包含这个样本。

2. 坐标分量 Y

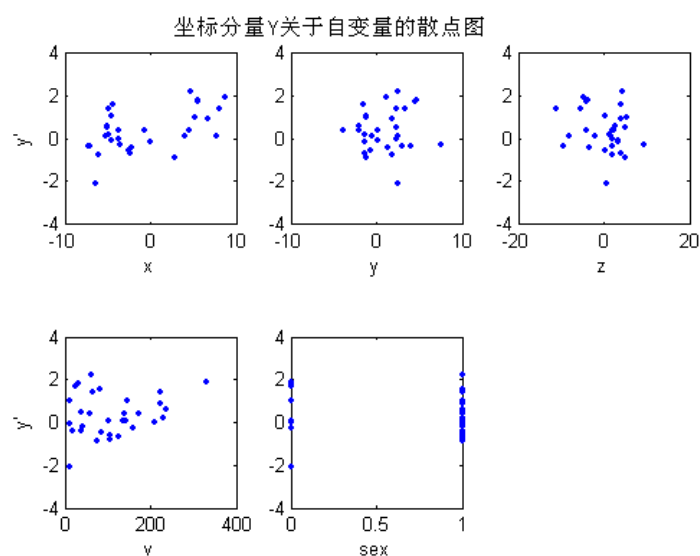


图 4. 4

从散点图 4. 4 上可以看出，坐标分量 y' 与自变量不存在明显的线性关系，也不存在明显的异常值点，回归结果如下：

系数	估计	标准差	t 值	p 值
截距	0.087757	0.525389	0.167	0.8687
x	0.094833	0.036403	2.605	0.0155
y	0.065337	0.085391	0.765	0.4516
z	-0.022207	0.036897	-0.602	0.5529
v	0.001486	0.002224	0.668	0.5103
sex	0.140539	0.473405	0.297	0.7691

标准误差：0.9041，自由度 24

复相关系数的平方：0.2847

F-统计量：1.91，自由度为 (5, 24)，p 值：0.1299

其残差对拟合值、QQ 图、标准残差对拟合值、Cook 距离图如图 4. 5 所示：

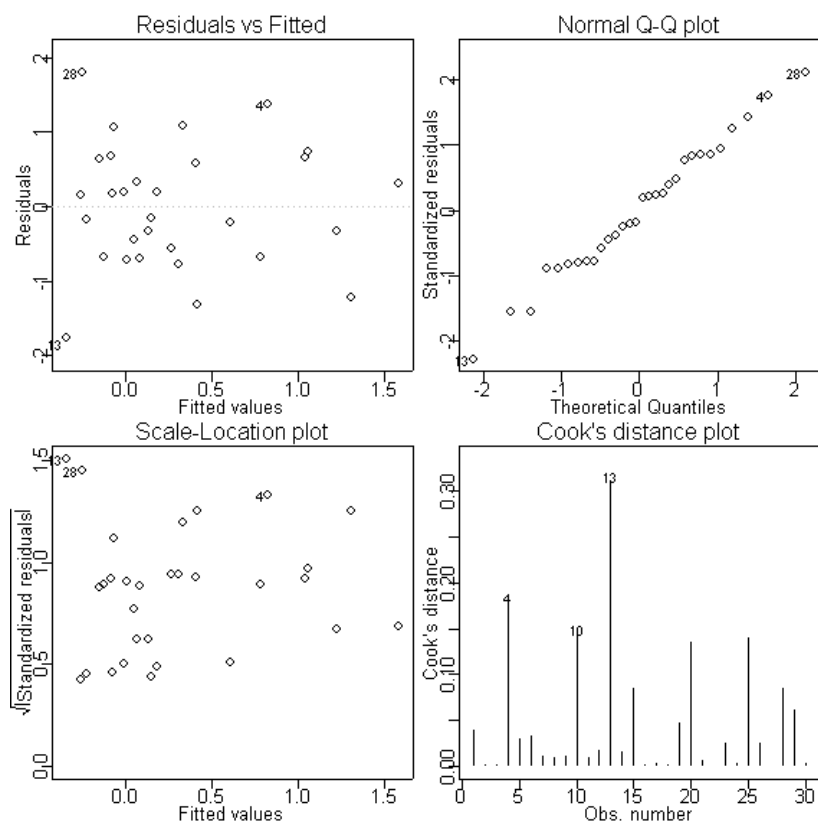


图 4.5

3. 坐标分量 Z

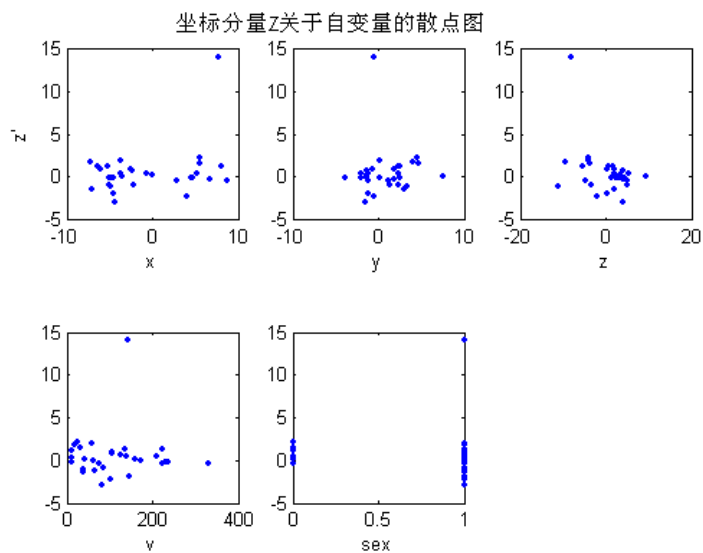


图 4.6

从散点图 4.6 上可以看出，坐标分量 z' 与自变量不存在明显的线性关系，可能存在一个异常值点，回归结果如下：

系数	估计	标准差	t 值	p 值
截距	1.119189	1.592349	0.703	0.4889
x	0.150853	0.110332	1.367	0.1842
y	-0.026994	0.258804	-0.104	0.9178
z	-0.214242	0.111827	-1.916	0.0674
v	-0.001327	0.006741	-0.197	0.8456
sex	-0.312211	1.434796	-0.218	0.8296

标准误差：2.74，自由度 24

复相关系数的平方：0.2231

F-统计量：1.378 ，自由度为 (5, 24)， p 值：0.2675

其残差对拟合值、QQ 图、标准残差对拟合值、Cook 距离图如图 4.7 所示：

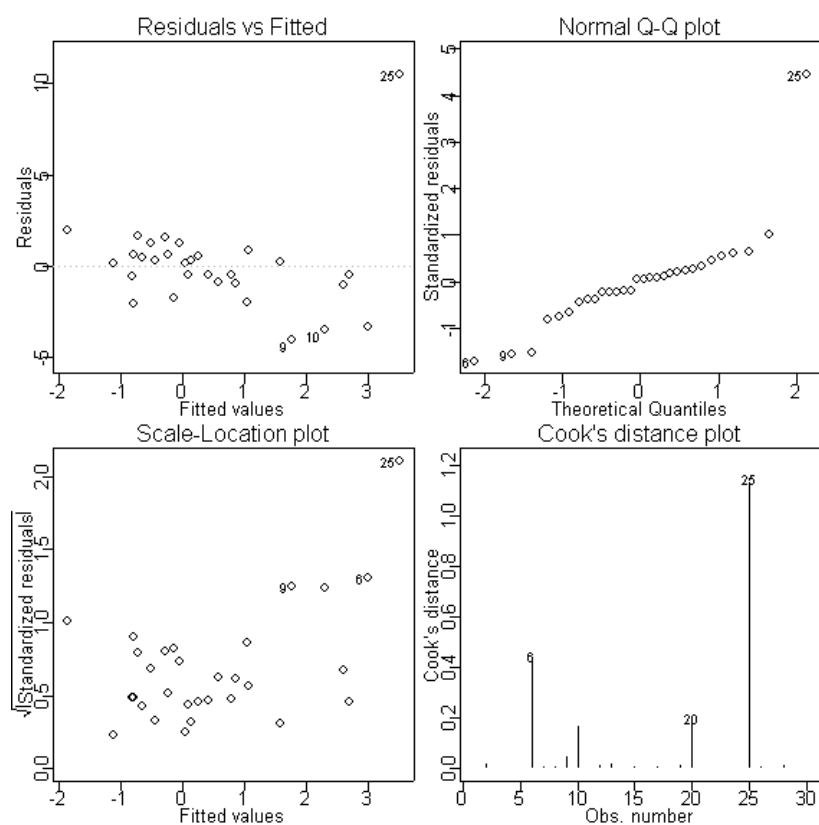


图 4.7

从残差对拟合值的图上看，25 号样本有可能是个异常值点，对其采取异常值检验，其 p 值 $p = 3.5168e-010$ ，说明 25 号样本是一个异常值点。

4. 体积退缩比例

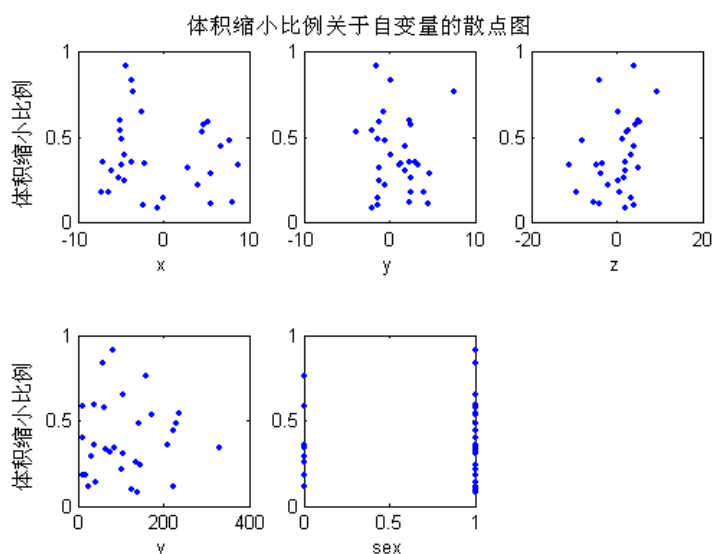


图 4.8

从散点图 4.8 上可以看出，体积缩小比例与自变量不存在明显的线性关系，也不存在明显的异常值点，回归结果如下：

系数	估计	标准差	t 值	p 值
截距	0.2889353	0.1320222	2.189	0.0390
x	-0.0022471	0.0096278	-0.233	0.8175
y	0.0091631	0.0214648	0.427	0.6734
z	0.0159829	0.0097906	1.632	0.1162
v	0.0002730	0.0005594	0.488	0.6301
sex	0.0727171	0.1200586	0.606	0.5507

标准误差：0.2272，自由度 23

复相关系数的平方：0.1266

F-统计量：0.6665，自由度为 (5, 23)，p 值：0.6526

其残差对拟合值、QQ 图、标准残差对拟合值、Cook 距离图如图 4.9 所示：

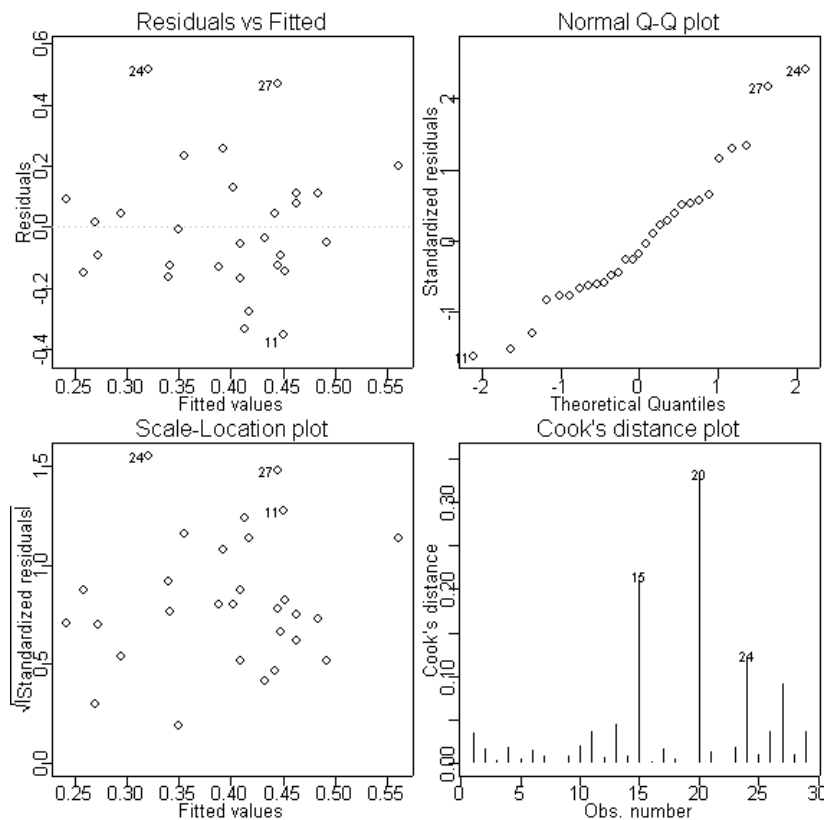


图 4.9

5. 通过对逐个目标分量进行线性回归，剔除了样本数据中两个异常值点，分别是 25 号和 26 号，对去除异常点以后的样本数据重新进行回归，得到的结果如下表 4.10 所示：

分量	x	y	z	体积缩小比例
标准误差	1.103cm	0.8756cm	1.183cm	0.2272
	空间距离的均方误差 1.4707cm			

表 4.10

4.2 最小一乘回归

在 4.1 中所做的线性回归中，一方面通过观察残差对拟合值和 Cook 距离图可以发现，对于线性模型影响较大的点往往都是残差较大的点，由于最小二乘回归模型对异常点比较敏感，所以可能最小一乘回归的效果会优于最小二乘回归；另一方面，对于空间坐标分量来说，我们直接关注的是空间距离的误差，所以在采用最小一乘回归的同时，将目标函数

$$\min F = \sum_{i=1}^n [(Y_{i1} - \alpha_0 - \sum_{j=1}^m \alpha_j X_{ij})^2 + (Y_{i2} - \beta_0 - \sum_{j=1}^m \beta_j X_{ij})^2 + (Y_{i3} - \gamma_0 - \sum_{j=1}^m \gamma_j X_{ij})^2]$$

改为

$$\min F' = \sum_{i=1}^n \sqrt{(Y_{i1} - \alpha_0 - \sum_{j=1}^m \alpha_j X_{ij})^2 + (Y_{i2} - \beta_0 - \sum_{j=1}^m \beta_j X_{ij})^2 + (Y_{i3} - \gamma_0 - \sum_{j=1}^m \gamma_j X_{ij})^2}。$$

得到的结果如下表 4.11 所示

分量	x	y	z	体积缩小比例
均方误差	1.4212cm			0.1500

表 4.11

与表 4.10 相比，最小一乘回归的效果优于最小二乘回归，其中空间距离的均方误差相对下降了 3.4%，体积缩小比例均方误差相对下降了 34%，体积缩小比例分量的改进比较明显。

4.3 非参数回归

在 4.1 中所做的线性回归中，复相关系数的平方值都比较小，说明原数据集中线性关系不是很明显，可能存在着其它的非线性关系，我们选用非参数的方法来初步的刻画这种非线性关系，得到的结果如下：

1. 对于空间坐标，目标函数采用

$$\min F = \sum_{i=1}^n \sqrt{\sum_{j=1}^3 (y_{ij} - a_j - b_j^T (x_i - x_0))^2 W\left(\frac{|x_i - x_0|}{h}\right)}$$

，其中 $W(x)$ 为加权函数， h 为

带宽参数，整体数据集作为训练样本，得到的均方误差如表 4.12 所示：

$W(x)$	$h=4.5$	$h=5$	$h=6$	$h=7$	$h=8$
线性函数	1.0865	1.1148	1.1807	1.2469	1.3030
二次函数	1.1344	1.1731	1.2691	1.3604	1.3842
Gauss 函数	1.1617	1.1937	1.2486	1.2922	1.3291

表 4.12

如果采取 Leave-one-out 方法做测试，即循环选取其中一个样本作为测试样本，余下的数据作为训练样本，得到预测均方误差，结果如表 4.13 所示：

$W(x)$	$h=4.5$	$h=5$	$h=6$	$h=7$	$h=8$
线性函数	1.8951	1.7392	1.7796	1.7906	1.7947
二次函数	1.8427	1.7428	1.7980	1.8042	1.8122
Gauss 函数	1.7766	1.7824	1.7892	1.7936	1.7974

表 4.13

综合考虑拟合均方误差和预测均方误差，发现一个比较合理的组合是在 $W(x)$ 为线性函数， $h=5$ 时取得的，此时拟合均方误差为 **1.1148cm**，优于最小二乘回归的 1.4707cm 和最小一乘回归的 1.4212cm，预测均方误差为 **1.7392cm**。

2. 对于体积缩小比例，目标函数为 $\min F = \sum_{i=1}^n \sqrt{(V_i - a - b^T(x_i - x_0))^2 W\left(\frac{|x_i - x_0|}{h}\right)}$ ，

整体数据集作为训练样本，得到的均方误差如表 4.14 所示

$W(x)$	$h=4.5$	$h=5$	$h=6$	$h=7$	$h=8$
线性函数	0.1221	0.1315	0.1356	0.1397	0.1396
二次函数	0.1319	0.1322	0.1417	0.1406	0.1436
Gauss 函数	0.1326	0.1346	0.1392	0.1393	0.1396

表 4.14

如果采取 Leave-one-out 方法做测试，即循环选取其中一个样本作为测试样本，余下的数据作为训练样本，得到预测均方误差，结果如表 4.15 所示：

$W(x)$	$h=4.5$	$h=5$	$h=6$	$h=7$	$h=8$	$h=1000$
线性函数	0.1839	0.1886	0.1810	0.1799	0.1772	0.1744
二次函数	0.1906	0.1900	0.1781	0.1760	0.1754	0.1744
Gauss 函数	0.1850	0.1814	0.1791	0.1773	0.1771	0.1744

表 4.15

综合考虑表 4.14 和表 4.15，可以发现拟合均方误差一般在 0.14 左右，预测均方误差一般在 0.18 左右，一个比较合理的组合是在 $W(x)$ 为线性函数， $h=6$ 时取得的，此时拟合均方误差为 **0.1356**，优于最小二乘回归的 0.2272 和最小一乘回归的 0.15，预测均方误差为 **0.1810**。

4.4 支持向量回归

1. 对于空间坐标分量，可以选择不同的核函数和损失函数，我们选取核函数分别为二次多项式核和高斯核，损失函数分别为拉普拉斯函数，用整体数据集作为训练样本估计拟合均方误差，采用 Leave-one-out 方法估计预测均方误差，得到的结果如下：

1) 二次多项式核 $K(x, x') = (x \cdot x' + 1)^2$ ，结果如表 4.16 所示，其中 C 表示惩罚系数， $R1$ 表示拟合均方误差， $R2$ 表示预测均方误差，可以看出随着惩罚系数 C 的增大， $R1$ 逐渐减小，趋于稳定，但 $R2$ 逐渐增大。与非参数回归相比，其最小拟合均方误差 1.0296cm 并不比非参数回归的 1.1148cm 改进多少；其预测均方误差则显得很不理想，所以二次多项式核支持向量回归的效果不如非参数回归。

	C=0.25	C=0.5	C=1	C=2	C=5	C=10	C=50	C=100
R1	1.2635	1.2264	1.1729	1.1326	1.0296	1.0624	1.1173	1.1173
R2	3.3337	3.6677	4.0859	4.757	6.0701	7.2123	9.798	12.347

表 4.16

2) 高斯核 $K(x_1, x_2) = \exp(-(x_1 - x_2)' \cdot (x_1 - x_2) / (2P^2))$, 结果如表 4.17 和表 4.18 所示, 其中 C 表示惩罚系数, P 表示高斯核函数参数, R1 表示拟合均方误差, R2 表示预测均方误差。

R1

	C=2	C=8	C=32	C=128	C=512	C=2048	C=8192
P=0.03125	0.2225	5.7e-09	5.7e-09	5.7e-09	5.7e-09	5.7e-09	5.7e-09
P=0.125	0.19871	5.95e-09	5.95e-09	5.95e-09	5.95e-09	5.95e-09	5.95e-09
P=0.5	0.25067	6.45e-09	6.45e-09	6.45e-09	6.45e-09	6.45e-09	6.45e-09
P=2	1.07	0.8680	0.6459	0.3100	0.01540	4.02e-07	4.02e-07
P=8	1.7241	1.6609	1.5994	1.5202	1.429	1.2448	1.1171
P=32	57.81	57.694	57.241	55.471	44.103	27.775	21.016
P=128	57.836	57.828	57.8	57.685	57.231	55.451	44.034

表 4.17

R2

	C=2	C=8	C=32	C=128	C=512	C=2048	C=8192
P=0.03125	1.7613	1.7627	1.7627	1.7627	1.7627	1.7627	1.7627
P=0.125	1.6835	1.6851	1.6851	1.6851	1.6851	1.6851	1.6851
P=0.5	1.6121	1.6139	1.6139	1.6139	1.6139	1.6139	1.6139
P=2	1.5501	1.732	2.0079	2.4416	2.8983	2.8839	2.8839
P=8	1.9146	2.0477	1.9804	2.0305	2.5352	3.347	4.2845
P=32	57.877	57.781	57.416	55.98	46.022	31.063	27.787
P=128	57.896	57.89	57.866	57.774	57.405	55.961	45.949

表 4.18

一个比较合理的点在 C=2, P=2 取得的, 此时 $R1=1.07$, $R2=1.55$, 在参数 (C=2,

P=2)重新画网格，得到结果如表 4.19 和表 4.20 所示

R1

	C=0.70711	C=1	C=1.4142	C=2	C=2.8284	C=4	C=5.6569
P=0.70711	0.7594	0.5690	0.35342	0.1514	0.00325	5.2e-009	5.2e-009
P=1	0.64752	0.4320	0.26272	0.11764	0.012703	0.006938	0.0004876
P=1.4142	0.958	0.8860	0.83238	0.7823	0.73935	0.69134	0.63865
P=2	1.1685	1.1349	1.1043	1.07	1.028	0.97524	0.92036
P=2.8284	1.603	1.568	1.5465	1.5336	1.5185	1.4963	1.4694
P=4	57.268	57.031	56.698	56.232	55.578	54.669	53.411
P=5.6569	57.488	57.344	57.142	56.856	56.455	55.892	55.106

表 4.19

R2

	C=0.70711	C=1	C=1.4142	C=2	C=2.8284	C=4	C=5.6569
P=0.70711	1.3924	1.3418	1.3077	1.2824	1.2651	1.2651	1.2651
P=1	1.2728	1.186	1.1311	1.0962	1.0838	1.084	1.0841
P=1.4142	1.3878	1.3515	1.3477	1.3787	1.4288	1.4801	1.5412
P=2	1.5391	1.5416	1.5394	1.5501	1.5731	1.618	1.6769
P=2.8284	1.9131	1.9199	1.9227	1.9404	1.9885	2.0361	2.1032
P=4	57.449	57.262	56.999	56.63	56.112	56.582	56.447
P=5.6569	57.617	57.502	57.339	57.11	56.789	56.338	55.707

表 4.20

从表 4.19 和表 4.20 可以看出，一方面支持向量回归的确非常强大，可以使得拟合的均方误差 R1 任意接近 0，但这是不合理的，即所谓的过度拟合；另一方面，预测的均方误差 R2 无法达到很小，都大于 1.00cm，这也说明单单使拟合均方误差达到很小是不合理的，实际上如果模型合理的话，R2 和 R1 之差不应该很大，我们选取 C=0.70711，P=1.4142 时的取值，R1=0.958cm，R2=1.3878cm。与非参数回归相比，支持向量回归的结果更好。

2. 对于体积缩小比例因子，采用类似的方法，选取核函数分别为二次多项式核

和高斯核，损失函数分别为拉普拉斯函数，用整体数据集作为训练样本估计拟合均方误差，采用 Leave-one-out 方法估计预测均方误差，得到如下结果：

1) 二次多项式核

其中 C 表示惩罚系数，R1 表示拟合均方误差，R2 表示预测均方误差，结果如表 4.21 所示

	C=0.25	C=0.5	C=1	C=2	C=5	C=10	C=50	C=100
R1	0.17757	0.17415	0.17583	0.17494	0.1863	0.18948	0.19164	0.19164
R2	0.24237	0.24741	0.25019	0.2768	0.3125	0.31247	0.31247	0.31247

表 4.21

从表 4.21 来看，与非参数方法相比，效果不如非参数回归

2) 高斯核

结果如表 4.22 和表 4.23 所示，其中 C 表示惩罚系数，P 表示高斯核函数参数，R1 表示拟合均方误差，R2 表示预测均方误差。

R1

	C=2	C=8	C=32	C=128	C=512	C=2048	C=8192
P=0.03125	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010
P=0.125	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010	6.6e-010
P=0.5	0.00488	1.38e-009	1.38e-009	1.38e-009	1.38e-009	1.38e-009	1.38e-009
P=2	0.1693	0.1570	0.1192	0.04578	1.42e-007	1.42e-007	1.42e-007
P=8	0.2064	0.2003	0.1898	0.1822	0.1698	0.1558	0.1596
P=32	0.2148	0.2084	0.2098	0.2115	0.2083	0.2041	0.1899
P=128	0.3352	0.2193	0.2148	0.2084	0.2100	0.2126	0.2119

表 4.22

R2

	C=2	C=8	C=32	C=128	C=512	C=2048	C=8192
P=0.03125	0.22422	0.22422	0.22422	0.22422	0.22422	0.22422	0.22422
P=0.125	0.22075	0.22075	0.22075	0.22075	0.22075	0.22075	0.22075
P=0.5	0.23385	0.23387	0.23387	0.23387	0.23387	0.23387	0.23387
P=2	0.25211	0.26128	0.31885	0.51921	0.59888	0.59888	0.59888
P=8	0.2418	0.26383	0.25569	0.27938	0.33555	0.46496	0.64611
P=32	0.22113	0.21999	0.24399	0.25884	0.27486	0.26684	0.25758
P=128	0.33707	0.21964	0.22109	0.21989	0.24416	0.25787	0.26958

表 4.23

一个比较合理的点在 C=2, P=2 取得的, 此时 $R1=0.1693$, $R2=0.2521$, 在参数 (C=2, P=2) 重新画网格, 得到结果如表 4.24 和表 4.25 所示

R1

	C=0.70711	C=1	C=1.4142	C=2	C=2.8284	C=4	C=5.6569
P=0.70711	0.0770	0.0493	0.0243	0.0186	0.0125	0.00388	2.70e-009
P=1	0.148	0.132	0.114	0.0922	0.0615	0.0232	0.0120
P=1.4142	0.174	0.170	0.166	0.159	0.147	0.138	0.125
P=2	0.174	0.173	0.171	0.170	0.169	0.171	0.168
P=2.8284	0.187	0.183	0.180	0.177	0.175	0.172	0.169
P=4	0.198	0.194	0.190	0.190	0.187	0.185	0.182
P=5.6569	0.205	0.204	0.202	0.199	0.195	0.192	0.191

表 4.24

R2

	C=0.70711	C=1	C=1.4142	C=2	C=2.8284	C=4	C=5.6569
P=0.70711	0.25191	0.25651	0.25947	0.25914	0.25917	0.25922	0.25924
P=1	0.2478	0.25019	0.25848	0.2768	0.295	0.31348	0.31249
P=1.4142	0.24479	0.24601	0.24664	0.24661	0.24845	0.25399	0.26726
P=2	0.24934	0.24593	0.24746	0.25211	0.25307	0.2554	0.25785
P=2.8284	0.24914	0.25449	0.26145	0.26068	0.25612	0.25162	0.25763
P=4	0.23706	0.24328	0.25073	0.25084	0.25477	0.26199	0.26759
P=5.6569	0.23256	0.23885	0.24531	0.25084	0.25585	0.25826	0.25336

表 4.25

从表 4.24 和表 4.25 可以看出，我们选取 C=2，P=1.4142 时的取值，**R1=0.159**，**R2=0.2466**。与非参数回归相比，支持向量回归的结果不如非参数回归。

4.5 各种方法的比较

方法	空间坐标 (cm)		体积缩小比例	
	R1	R2	R1	R2
最小二乘回归	1.4707cm		0.2272	
最小一乘回归	1.4212cm		0.1500	
非参数回归	1.1148cm	1.7392cm	0.1356	0.181
支持向量回归	0.958cm	1.3878cm	0.159	0.2466

表 4.26

从表 4.26 可以看出，在所用的四种方法中，对于空间坐标分量，支持向量回归的效果最佳；对于体积缩小比例来说，非参数回归效果更好。

第五章 结论

5.1 结论

1. 从表 4.26 上看, 对比空间坐标均方误差 0.5cm 和体积退缩比例均方误差 0.05 的临床要求, 可以说初步的结果不是很理想。
2. 从表 4.26 上看, 在所用的四种方法中, 对于空间坐标分量, 支持向量回归的效果最佳, 拟合均方误差和预测均方误差分别达到了 0.958cm 和 1.3878cm; 对于体积缩小比例来说, 非参数回归效果更好, 拟合均方误差和预测均方误差分别达到了 0.1356 和 0.181。

5.2 效果不理想可能存在的原因

1. 样本数据的总量不是很充足, 拟合的模型有可能对个别异常点比较敏感, 而使得模型偏离真实的情况。
2. 样本数据中缺乏足够的因子变量, 影响肿瘤退缩的临床因子有很多, 除了本身的位置、体积和患者的性别以外, 可能还存在对肿瘤退缩规律影响显著的其他因子未给出, 比如肿瘤治疗的剂量等。
3. 除了我们给出的几种方法以外, 可能还存在更适合拟合肿瘤退缩规律的其他方法。

参考文献

- [1] 陈家鼎, 孙山泽, 李东风, 刘力平. 2006. 数理统计学讲义. 高等教育出版社.
- [2] 邓乃扬, 田英杰. 2004. 数据挖掘中的新方法—支持向量机. 科学出版社.
- [3] Vladimir N. Vapnik 著, 张学工译. 2000. 统计学习理论的本质. 清华大学出版社.
- [4] Ataman OU, Barrett A, Filleron T, et al. Optimization of follow-up timing from study of patterns of first failure after primary treatment. An example from patients with NSCLC: a study of the REACT working group of ESTRO. *Radiother Oncol* 2006; 78(1):95-100.
- [5] Boser B E, Guyon I M, Vapnik V. A training algorithm for optimal margin classifier. *In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, ACM*, 1992: 144--52.
- [6] A. Charnes, W. W. Cooper, R. O. Ferguson. Optimal Estimation of Executive Compensation by Linear Programming. *MANAGEMENT SCIENCE* 1955 1(2), 138-151.
- [7] J. Fox. 2002. *An R and S-PLUS Companion to Applied Regression*. Sage Publications.
- [8] Jeffrey S. Simonoff. 2000. *Smoothing Methods in Statistics*. Springer-Verlag.
- [9] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, London, 1909, A209:415-446.
- [10] Vapnik V N, Chervoneenkis A Ja. 1968. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR*,181(4).
- [11] Vapnik V N, Chervoneenkis A Ja. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Apl.*, 16:264~280.
- [12] Vapnik V N. 1979. *Estimation of Dependencies based on empirical data*. Nauka, Moscow.
- [13] Vapnik V N and Chervonenkis A Ja. 1989. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting*, 2, Nauka, Moscow, 207~249.
- [14] Vapnik V N. 1998. *Statistical Learning Theory*. J.Wiley, New York.
- [15] Woodford C, Yartsev S, Dar AR, et al. Adaptive radiotherapy planning on decreasing gross tumor volumes as seen on megavoltage computed tomography images. *Int J Radiat Oncol Biol Phys* 2007; 69(4):1316-22.

致谢

衷心的感谢我的导师刘力平教授和姜明教授这几年来在学习上给予我的指导和帮助，他们严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从课题的选择到本文的最终完成，两位老师一直给予我支持、鼓励和帮助，并提出了众多宝贵的指导性建议，在此谨向两位老师致以诚挚的谢意和崇高的敬意！

此外我要感谢讨论班上的所有老师和同学，和他们一起几年的学习，大大丰富了我的视野，讨论班上活跃的学术气氛给我留下了深刻的印象，要特别感谢蒋欣同学和蔡啸天同学，他们对于我所讨论的问题也给予了许多中肯的建议，在此对大家给予我的支持和帮助表示深深的谢意！

最后要感谢北京大学临床肿瘤医院的朱广迎主任、石安辉大夫和李有泉博士，对我所讨论的问题，他们提供了丰富的材料和宝贵的临床数据，并对本文提出了很多建设性的意见，在此对于他们的支持和帮助表示衷心的感谢！

感谢北京大学和数学科学学院多年来对我的培养！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

(必须装订在提交学校图书馆的印刷本)

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校一年/两年/三年以后，在校园网上全文发布。

(保密论文在解密后遵守此规定)

论文作者签名： 导师签名：

日期： 年 月 日