数学建模算法与应用

第9章　支持向量机

支持向量机是数据挖掘中的一项新技术，是借助最优化方法来解决机器学习问题的新工具，最初由V.Vapnik 等人提出，近几年来在其理论研究和算法实现等方面都取得了很大的进展，开始成为克服“维数灾难”和“过学习”等困难的强有力手段，其理论基础和实现途径的基本框架都已形成。

支持向量机（Sport Vector Machine ,以下简称SVM）在模式识别等领域获得了广泛的应用。其主要思想是找到一个超平面，使得它能够尽可能多地将两类数据点正确分开，同时使分开的两类数据点距离分类面最远，如图 9.1（b)。

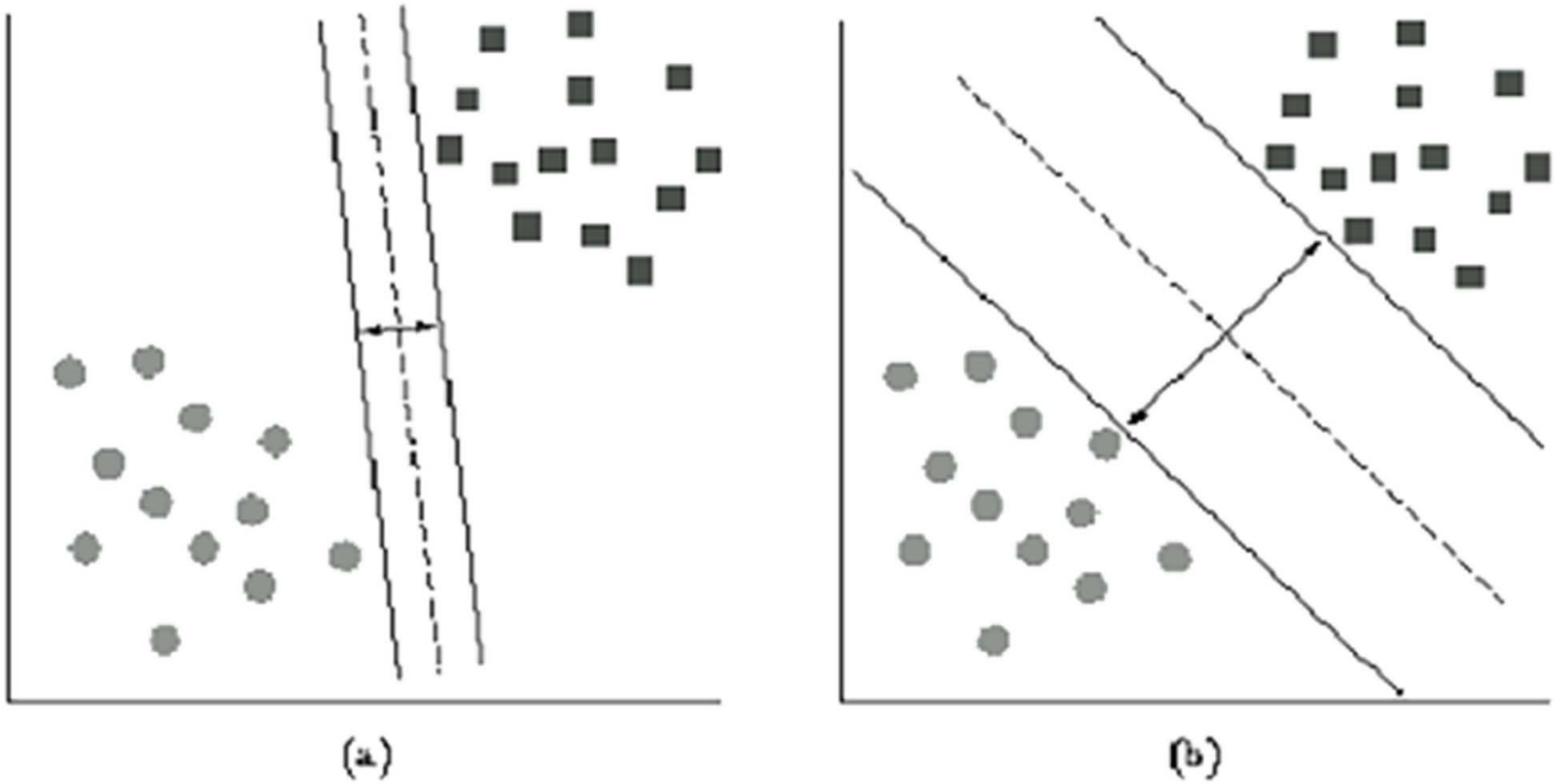

图 9.1　最佳超平面示意图

9.1 支持向量分类机的基本原理

根据给定的训练集

$$T=\{[a_1,y_1],[a_2,y_2],\cdots,[a_l,y_l]\}\in(\Omega\times Y)^l,$$

其中$a_i\in\Omega=R^n$，Ω称为输入空间，输入空间中的每一个点a_i由n个属性特征组成，$y_i\in Y=\{-1,1\},i=1,\cdots,l$。寻找$R^n$上的一个实值函数$g(x)$，以便用分类函数

$$f(x)=\mathrm{sgn}(g(x)),$$

推断任意一个模式x相对应的y值的问题为分类问题。

9.1.1 线性可分支持向量分类机

考虑训练集T，若$\exists \omega \in R^n$，$b \in R$和正数ε，使得对所有使$y_i = 1$的a_i有$(\omega \cdot a_i) + b \geq \varepsilon$（这里$(\omega \cdot a_i)$表示向量$\omega$和$a_i$的内积），而对所有使$y_i = -1$的$a_i$有$(\omega \cdot a_i) + b \leq -\varepsilon$，则称训练集$T$线性可分，称相应的分类问题是线性可分的。

记两类样本集分别为

$$M^+ = \{a_i \mid y_i = 1, [a_i, y_i] \in T\},$$

$$M^- = \{a_i \mid y_i = -1, [a_i, y_i] \in T\}.$$

定义M^+的凸包$\mathrm{conv}(M^+)$为

$$\mathrm{conv}(M^+) = \left\{ a = \sum_{j=1}^{N_+} \lambda_j a_j \mid \sum_{j=1}^{N_+} \lambda_j = 1, \quad \lambda_j \geq 0, j = 1, \cdots, N^+ \right.;$$

M^-的凸包$\mathrm{conv}(M^-)$为

$$\mathrm{conv}(M^-) = \left\{ a = \sum_{j=1}^{N_-} \lambda_j a_j \mid \sum_{j=1}^{N_-} \lambda_j = 1, \quad \lambda_j \geq 0, j = 1, \cdots, N^- \right.;$$

其中N_+表示+1类样本集M^+中样本点的个数，$N_$表示−1类样本集M^-中样本点的个数，定理 9.1 给出了训练集T线性可分与两类样本集凸包之间的关系。

定理 9.1　训练集T线性可分的充要条件是，T的两类样本集M^+和M^-的凸包相离，如图 9.2 所示。

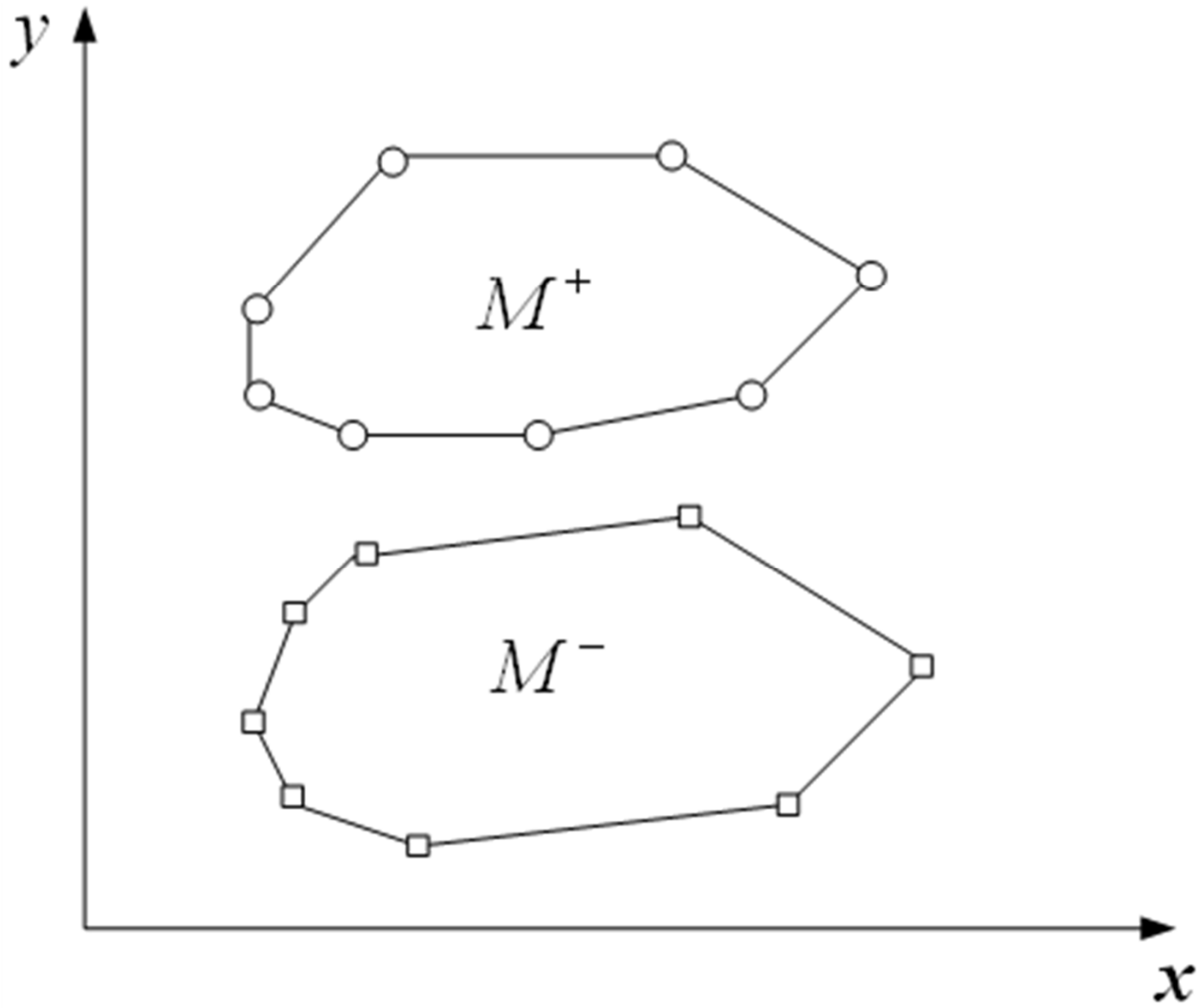

图 9.2　训练集$_T$线性可分时两类样本点集的凸包

证明：（1）必要性

若 T 是线性可分的，则存在超平面 $H=\{x\in R^n \mid (\omega\cdot x)+b=0\}$ 和 $\varepsilon>0$，使得

$$(\omega\cdot a_i)+b\geq\varepsilon,\ \forall a_i\in M^+ \text{且} (\omega\cdot a_j^{'})+b\leq-\varepsilon,$$

$$\forall a_j^{'}\in M^-.$$

而正类点集凸包中的任意一点x和负类点集凸包中的任意一点x'可分别表示为

$$x=\sum_{i=1}^{N_+}\alpha_i a_i \text{和} x'=\sum_{j=1}^{N_-}\beta_j a'_j,$$

其中$\alpha_i \geq 0, \beta_j \geq 0$且$\sum_{i=1}^{N_+}\alpha_i = 1, \sum_{j=1}^{N_-}\beta_j = 1$。

于是可以得到

$$(\omega\cdot x)+b=\left(\omega\cdot\sum_{i=1}^{N_+}\alpha_i a_i\right)+b$$

$$=\sum_{i=1}^{N_+}\alpha_i((\omega\cdot a_i)+b)\geq\varepsilon\sum_{i=1}^{N_+}\alpha_i=\varepsilon>0,$$

$$(\omega\cdot x')+b=\left(\omega\cdot\sum_{j=1}^{N_-}\beta_j a'_j\right)+b$$

$$=\sum_{j=1}^{N_-}\beta_j((\omega\cdot a'_j)+b)\leq-\varepsilon\sum_{j=1}^{N_-}\beta_j=-\varepsilon<0.$$

由此可见，正负两类点集的凸包位于超平面$(\omega \cdot x)+b=0$的两侧，故两个凸包相离。

（2）充分性

设两类点集M^+，M^-的凸包相离。因为两个凸包都是闭凸集，且有界，根据凸集强分离定理，可知存在一个超平面$H=\{x\in R^n\mid(\omega\cdot x)+b=0\}$强分离这两个凸包，即存在正数$\varepsilon>0$，使得对$M^+$，$M^-$凸包中的任意点$x$和$x'$分别有

$$(\omega\cdot x)+b\geq\varepsilon,$$

$$(\omega\cdot x')+b\leq-\varepsilon.$$

显然特别的，对于任意的$x \in M^{+}$，有$(\omega \cdot x)+b \geq \varepsilon$，对于任意的$x' \in M^{-}$，有$(\omega \cdot x')+b \leq -\varepsilon$，由训练集线性可分的定义可知$T$是线性可分的。

定义 9.1 空间 R^n 中超平面都可以写为 $(\omega \cdot x)+b=0$ 的形式，参数 (ω,b) 乘以任意一个非零常数后得到的是同一个超平面，定义满足条件

$$\begin{cases} y_i((\omega \cdot a_i)+b) \geq 0, \quad i=1,\cdots,l, \\ \min\limits_{i=1,\cdots,l} \left|(\omega \cdot a_i)+b\right| = 1. \end{cases}$$

的超平面为训练集 T 的规范超平面。

定理 9.2　当训练集T为线性可分时，存在唯一的规范超平面$(\omega \cdot x)+b=0$，使得

$$\begin{cases}(\omega \cdot a_i)+b \geq 1, & y_i=1,\\ (\omega \cdot a_i)+b \leq -1, & y_i=-1.\end{cases} \tag{9.1}$$

证明：规范超平面的存在性是显然的，下证其唯一性。

假设其规范超平面有两个，$(\omega' \cdot x) + b' = 0$ 和 $(\omega'' \cdot x) + b'' = 0$。由于规范超平面满足条件

$$\begin{cases} y_i((\omega \cdot a_i) + b) \geq 0, \quad i = 1, \cdots, l, \\ \min\limits_{i=1,\cdots,l} \left|(\omega \cdot a_i) + b\right| = 1. \end{cases}$$

由第二个条件可知

$$\omega' = \omega'', \quad b' = b'',$$

或者

$$\omega' = -\omega'', \quad b' = -b''.$$

第一个条件说明$\omega' = -\omega''$，$b' = -b''$不可能成立，故唯一性得证。

定义 9.2　式(9.1)中满足$(\omega \cdot a_i)+b=\pm 1$成立的$a_i$称为普通支持向量。

对于线性可分的情况来说，只有普通支持向量在建立分类超平面的时候起到了作用，它们通常只占样本集很小的一部分，故而也说明支持向量具有稀疏性。对于$y_i=1$类的样本点，其与规范超平面的间隔为

$$\min_{y_i=1}\frac{\left|(\omega\cdot a_i)+b\right|}{\|\omega\|}=\frac{1}{\|\omega\|},$$

对于$y_i=-1$类的样本点，其与规范超平面的间隔为

$$\min_{y_i=-1}\frac{\left|(\omega\cdot a_i)+b\right|}{\|\omega\|}=\frac{1}{\|\omega\|},$$

则普通支持向量间的间隔为$\frac{2}{\|\omega\|}$。

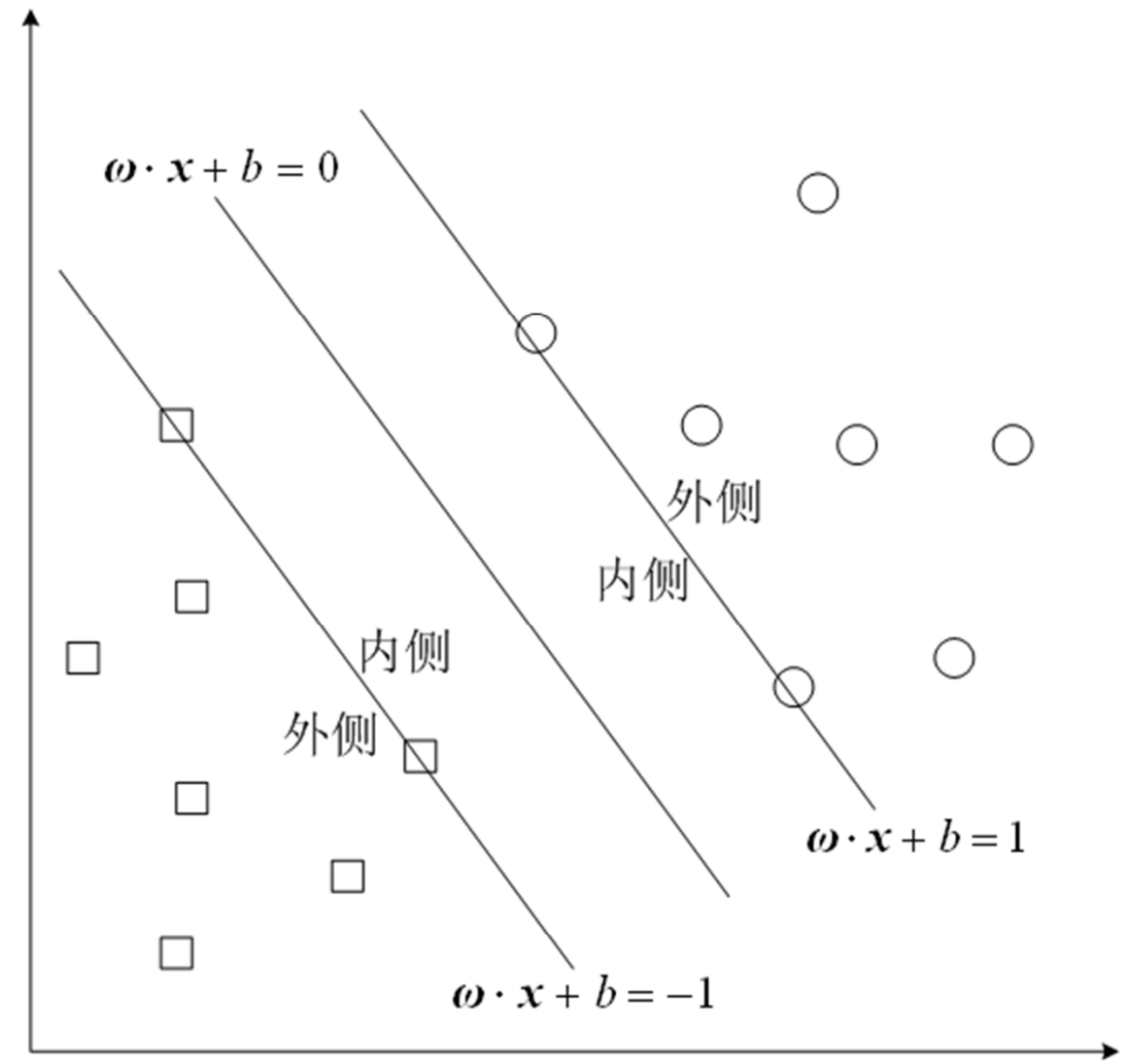

图 9.3　线性可分支持向量分类机

最优超平面即意味着最大化$\frac{2}{\|\omega\|}$，如图 9.3 所示，$(\omega \cdot x)+b=\pm 1$称为分类边界，于是寻找最优超平面的问题可以转化为如下的二次规划问题

$$\begin{aligned} &\min \quad \frac{1}{2}\|\omega\|^2, \\ &\text{s.t.} \quad y_i((\omega \cdot a_i)+b) \geq 1, \quad i=1,\cdots,l. \end{aligned} \tag{9.2}$$

该问题的特点是目标函数$\frac{1}{2}\|\omega\|^2$是ω的凸函数，并且约束条件都是线性的。

引入 Lagrange 函数

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{l}\alpha_i(1-y_i((\omega\cdot a_i)+b)),$$

其中$\alpha=[\alpha_1\cdots,\alpha_l]^T\in R^{l+}$为 Lagrange 乘子。根据对偶的定义，通过对原问题中各变量的偏导置零可得

$$\frac{\partial L}{\partial \omega}=0 \quad \Rightarrow \quad \omega=\sum_{i=1}^{l}\alpha_i y_i a_i\,,$$

$$\frac{\partial L}{\partial b}=0 \quad \Rightarrow \quad \sum_{i=1}^{l}\alpha_i y_i=0\,,$$

带入 Lagrange 函数化为原问题的 Lagrange 对偶问题

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (a_i \cdot a_j) + \sum_{i=1}^{l} \alpha_i$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ \alpha_i \geq 0, i = 1, \cdots, l. \end{cases} \tag{9.3}$$

求解上述最优化问题，得到最优解$\alpha^* = [\alpha_1^*, \cdots, \alpha_l^*]^T$，计算

$$\omega^* = \sum_{i=1}^{l} \alpha_i^* y_i a_i ,$$

由 KKT 互补条件知

$$\alpha_i^* (1 - y_i((\omega^* \cdot a_i) + b^*)) = 0,$$

可得只有当a_i为支持向量的时候，对应的α_i^*才为正，否则皆为零。选择α^*的一个正分量α_j^*，并以此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (a_i \cdot a_j),$$

于是构造分类超平面$(\omega^* \cdot x)+b^*=0$，并由此求得决策函数

$$g(x)=\sum_{i=1}^{l}\alpha_i^* y_i(a_i \cdot x)+b^*,$$

得到分类函数

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^* y_i(a_i \cdot x)+b^*),$$

(9.4)

从而对未知样本进行分类。

9.1.2　线性支持向量分类机

当训练集T的两类样本线性可分时，除了普通支持向量分布在两个分类边界$(\omega \cdot x)+b=\pm 1$上外，其余的所有样本点都分布在分类边界以外。此时构造的超平面是硬间隔超平面。当训练集T的两类样本近似线性可分时，即允许存在不满足约束条件

$$y_i((\omega \cdot a_i)+b) \geq 1$$

的样本点后，仍然能继续使用超平面进行划分。

只是这时要对间隔进行“软化”，构造软间隔超平面。简言之就是在两个分类边界$(\omega \cdot x)+b=\pm 1$之间允许出现样本点，这类样本点被称为边界支持向量。显然两类样本点集的凸包是相交的，只是相交的部分较小。线性支持向量分类机如图 9.4 所示。

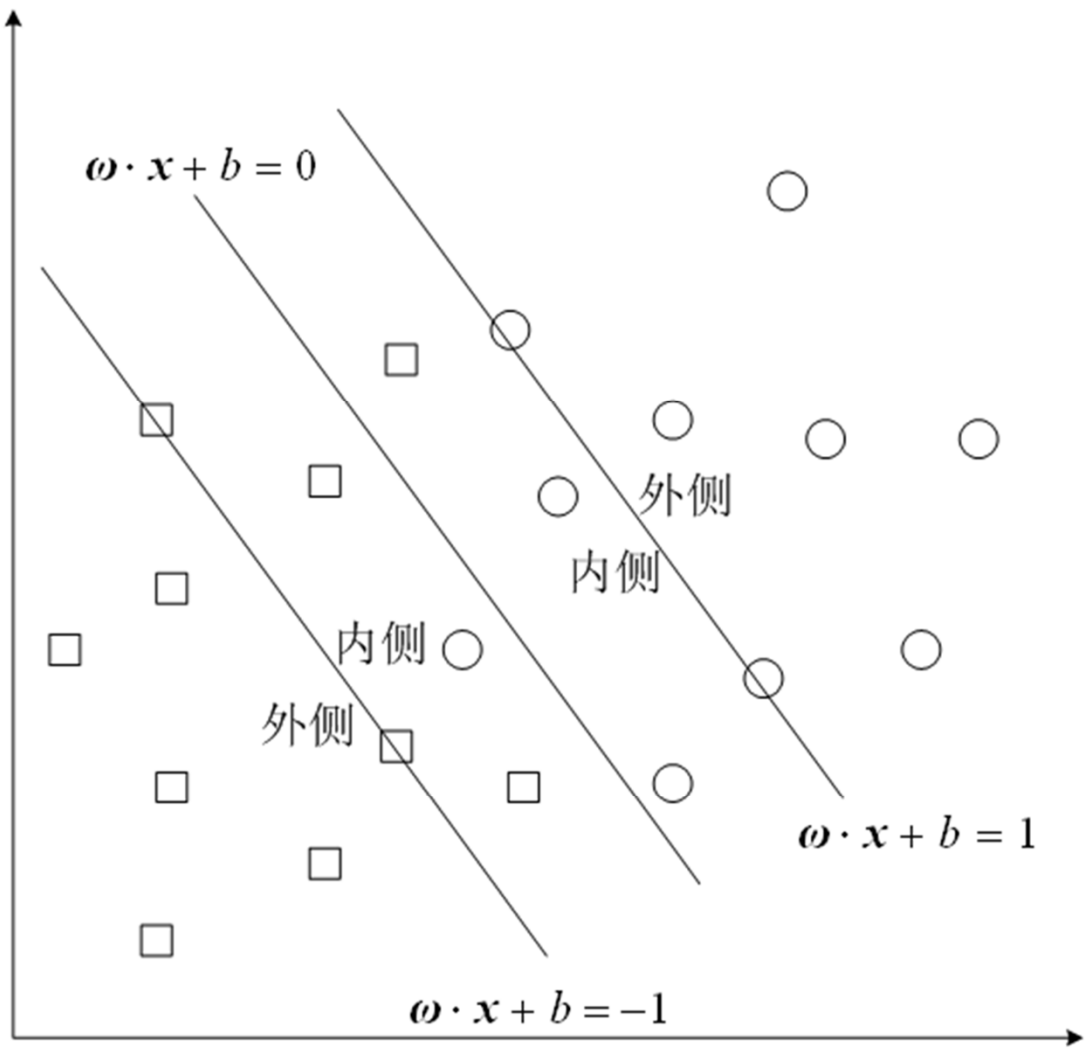

图 9.4　线性支持向量分类机

软化的方法是通过引入松弛变量

$$\xi_i \geq 0, \quad i = 1, \cdots, l,$$

来得到“软化”的约束条件

$$y_i((\omega \cdot a_i) + b) \geq 1 - \xi_i \quad i = 1, \cdots, l,$$

当ξ_i充分大时，样本点总是满足上述的约束条件，但是也要设法避免ξ_i取太大的值，为此要在目标函数中对它进行惩罚，得到如下的二次规划问题

$$\min \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i ,$$

$$\text{s.t.} \begin{cases} y_i((\omega \cdot a_i)+b) \geq 1-\xi_i, \\ \xi_i \geq 0, \quad i=1,\cdots,l. \end{cases} \tag{9.5}$$

其中$C>0$是一个惩罚参数。其 Lagrange 函数如下

$$L(\omega,b,\xi,\alpha,\gamma) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i$$

$$-\sum_{i=1}^{l}\alpha_i(y_i((\omega \cdot a_i)+b)-1+\xi_i)-\sum_{i=1}^{l}\gamma_i\xi_i,$$

其中$\gamma_i \geq 0$，$\xi_i \geq 0$。

原问题的对偶问题如下

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (a_i \cdot a_j) + \sum_{i=1}^{l} \alpha_i ,$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ 0 \le \alpha_i \le C, i = 1, \cdots, l. \end{cases} \tag{9.6}$$

求解上述最优化问题，得到最优解 $\alpha^* = [\alpha_1^*, \cdots \alpha_l^*]^T$，计算

$$\omega^* = \sum_{i=1}^{l} \alpha_i^* y_i a_i,$$

选择α^*的一个正分量$0 < \alpha_j^* < C$，并以此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (a_i \cdot a_j).$$

于是构造分类超平面$(\omega^* \cdot x) + b^* = 0$，

并由此求得分类函数

$$f(x)=\operatorname{sgn}(\sum_{i=1}^{l}\alpha_i^* y_i(a_i\cdot x)+b^*).$$

从而对未知样本进行分类，可见当$C=\infty$时，就等价于线性可分的情形。

9.1.3　可分支持向量分类机

当训练集T的两类样本点集重合的区域很大时，上述用来处理线性可分问题的线性支持向量分类机就不适用了，可分支持向量分类机给出了解决这种问题的一种有效途径。

通过引进从输入空间Ω到另一个高维的 Hilbert 空间H的变换$x \mapsto \phi(x)$将原输入空间Ω的训练集

$$T = \{[a_1, y_1], [a_2, y_2], \cdots, [a_l, y_l]\} \in (\Omega \times Y)^l,$$

转化为 Hilbert 空间H中新的训练集

$$\tilde{T} = \{[\tilde{a}_1, y_1], \cdots, [\tilde{a}_l, y_l]\} = \{[\phi(a_1), y_1], \cdots, [\phi(a_l), y_l]\},$$

使其在 Hilbert 空间H中线性可分，Hilbert 空间H也称为特征空间。然后在空间H中求得超平面$(\omega \cdot \phi(x)) + b = 0$，这个超平面可以硬性划分训练集$\tilde{T}$，

于是原问题转化为如下的二次规划问题

$$\min \quad \frac{1}{2}\|\omega\|^2,$$

$$\text{s.t.} \quad y_i((\omega \cdot \phi(a_i)) + b) \geq 1, \quad i = 1, \cdots, l.$$

采用核函数K满足

$$K(a_i, a_j) = (\phi(a_i) \cdot \phi(a_j))$$

将避免在高维特征空间进行复杂的运算。

不同的核函数形成不同的算法，主要的核函数有如下几类

线性内核函数 $K(a_i, a_j) = (a_i \cdot a_j)$;

多项式核函数 $K(a_i, a_j) = [(a_i \cdot a_j) + 1]^q$;

径向基核函数 $K(a_i, a_j) = \exp\left\{-\frac{\|a_i - a_j\|^2}{\sigma^2}\right\}$;

S 形内核函数 $K(a_i, a_j) = \tanh(v(a_i \cdot a_j) + c)$;

傅里叶核函数

$$K(a_i, a_j) = \sum_{k=1}^{n} \frac{1 - q^2}{2(1 - 2q\cos(a_{ik} - a_{jk}) + q^2)}.$$

同样可以得到其 Lagrange 对偶问题如下

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(a_i \cdot a_j) + \sum_{i=1}^{l} \alpha_i ,$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ \alpha_i \geq 0, i = 1, \cdots, l. \end{cases}$$

若K是正定核，则对偶问题是一个凸二次规划问题，必定有解。求解上述最优化问题，得到最优解$\alpha^* = [\alpha_1^*, \cdots, \alpha_l^*]^T$，选择$\alpha^*$的一个正分量$\alpha_j^*$，并以此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(a_i, a_j),$$

构造分类函数

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

从而对未知样本进行分类。

9.1.4 C-支持向量分类机

当映射到高维H空间的训练集不能被硬性分划时，需要对约束条件进行软化。结合 9.1.2 和 9.1.3 中所述，得到如下的模型

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(a_i, a_j) + \sum_{i=1}^{l} \alpha_i ,$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ 0 \le \alpha_i \le C, i = 1, \cdots, l. \end{cases} \tag{9.7}$$

得到最优解$\alpha^* = [\alpha_1^*, \cdots, \alpha_l^*]^T$，选择$\alpha^*$的一个正分量$0 < \alpha_j^* < C$，并以此计算

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(a_i, a_j),$$

构造决策函数

$$g(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(a_i, x) + b^*,$$

构造分类函数

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} y_i \alpha_i^* K(a_i, x) + b^*),$$

从而对未知样本进行分类。

当输入空间中两类样本点的分布区域严重重合时，选择合适的核函数及其参数，可以使映射到特征空间的每一类样本点的分布区域更为集中，降低两类样本点分布区域的混合程度，从而加强特征空间中两类样本集“线性可分”的程度，来达到提高分类的精度和泛化性能的目的。

但是就核函数及其参数的选取问题，目前尚无理论依据，同样的实验数据，采用不同的核函数，其精度往往相差很大，即便是对于相同的核函数，选取的参数不同，分类的精度也会有较大的差别。在实际应用过程中，往往针对具体的问题多次仿真试验，找到适合该问题的核函数，并决定其最佳参数。

下述定理给出了支持向量与 Lagrange 乘子之间的关系。

定 理 9.3 对偶问题(9.7)的最优解为 $\alpha^* = [\alpha_1^*, \cdots, \alpha_l^*]^T$，使得每个样本点 a_i 满足优化问题的KKT条件为

$$\alpha_i^* = 0 \Rightarrow y_i g(a_i) > 1,$$

$$0 < \alpha_i^* < C \Rightarrow y_i g(a_i) = 1,$$

$$\alpha_i^* = C \Rightarrow y_i g(a_i) < 1,$$

其中 $0 < \alpha_i^* < C$，所对应的 a_i 就是普通支持向量（记作NSV），位于分类间隔的边界 $g(x) = \pm 1$ 上，有 $|g(a_i)| = 1$；$\alpha_i^* = C$ 所对应的 a_i 就是边界支持向量（记作 BSV），代表了所有的错分样本点，位于分类间隔内部，有 $|g(a_i)| < 1$。$_{\text{BSV}} \cup _{\text{NSV}}$ 就是支持向量集。

9.2　支持向量机的 Matlab 命令及应用例子

Matlab 中支持向量机的命令有，训练支持向量机分类器的函数 svmtrain，使用支持向量机分类的函数 svmclassify，指定支持向量机函数使用的序列最小化参数函数 svmsmoset。下面我们通过一个例子说明有关函数的使用。

例 9.1　1991 年全国各省、区、市城镇居民月平均消费情况见表 9.1(表略),1-20 号省份为第一类，记为 G_1；21-27 号省份为第二类，记为 G_2。考察下列指标

x_1 人均粮食支出（元/人）；

x_2 人均副食支出（元/人）；

x_3 人均烟酒茶支出（元/人）；

x_4 人均文化娱乐支出（元/人）；

x_5 人均衣着商品支出（元/人）；

x_6 人均日用品支出（元/人）；

x_7 人均燃料支出（元/人）；

x_8 人均非商品支出（元/人）；

试判别西藏、上海、北京应归属哪一类。

解　用$i=1,\cdots,30$分别表示 30 个省市或自治区，第i个省（市或自治区）的第j个指标的取值为a_{ij}。$y_i=1$表示类，$y_i=-1$表示第二类。

计算得已知 27 个样本点的均值向量

$$\mu=[\mu_1,\cdots,\mu_8]$$
$$=[8.6115,36.7148,7.1993,10.0626,$$
$$16.3174,11.0833,1.8196,12.4274]$$

27 个样本点的标准差向量

$$\sigma=[\sigma_1,\cdots,\sigma_8]$$
$$=[1.5246,9.4167,1.7835,2.7343,$$
$$2.8225,2.3437,0.5102,1.9307]^{\cdot}$$

对所有样本点数据利用如下公式进行标准化处理

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{\sigma_j},\ \ i=1,\cdots,30, j=1,\cdots,8.$$

对应地，称

$$\tilde{x}_j=\frac{x_j-\mu_j}{s_j},\quad (j=1,2,\cdots,8)$$

为标准化指标变量。记$\tilde{x}=[\tilde{x}_1,\cdots,\tilde{x}_8]^T$。

记标准化后的 27 个已分类样本点数据行向量为 $b_i=[\tilde{a}_{i1},\cdots,\tilde{a}_{i8}]$（$i=1,\cdots,27$）。利用线性内核函数的支持向量机模型进行分类，求得支持向量为

$$b_{14},\ b_{15},\ b_{17},\ b_{19},\ b_{24},\ b_{25},\ b_{27},$$

线性分类函数为

$$\begin{aligned}c(\tilde{x})&=\sum_i \beta_i K(b_i,\tilde{x})+b\\&=0.4694K(b_{14},\tilde{x})+0.0476K(b_{15},\tilde{x})\\&\quad+0.6750K(b_{17},\tilde{x})+0.5979K(b_{19},\tilde{x})\\&\quad-0.4234K(b_{24},\tilde{x})-1.2908K(b_{25},\tilde{x})\\&\quad-0.0758K(b_{27},\tilde{x})+1.0269,\end{aligned}$$

其中$\beta_i = \alpha_i y_i$，$\tilde{x} = [\tilde{x}_1, \cdots, \tilde{x}_8]$，$K(b_i, \tilde{x}) = (b_i \cdot \tilde{x})$。

当$c(\tilde{x}) \geq 0$，$\tilde{x}$属于第 1 类；当$c(\tilde{x}) < 0$，$\tilde{x}$属于第 2 类。

用判别函数判别，得到西藏、上海、广东皆属于总体G_2，即属于高消费类型。

所有已知样本点回代分类函数皆正确，故误判率为 0。

9.3　乳腺癌的诊断

9.3.1　问题的提出

乳腺肿瘤通过穿刺采样进行分析可以确定其为良性或恶性。医学研究发现乳腺肿瘤病灶组织的细胞核显微图像的 10 个量化特征：细胞核直径、质地、周长、面积、光滑度、紧密度、凹陷度、凹陷点数、对称度、断裂度与该肿瘤的性质有密切的关系。现试图根据已获得的实验数据建立起一种诊断乳腺肿瘤是良性还是恶性的方法。

数据来自确诊的 500 个病例，每个病例的一组数据包括采样组织中各细胞核的这 10 个特征量的平均值、标准差和最坏值共 30 个数据，并将这种方法用于另外 69 名已做穿刺采样分析的患者。

这个问题实际上属于模式识别问题。什么是模式呢？广义地说，在自然界中可以观察的事物，如果能够区别它们是否相同或是否相似，都可以称之为模式。人们为了掌握客观事物，按事物相似的程度组成类别。模式识别的作用和目的就在于面对某一具体事物时将其正确地归入某一类别。

模式识别的方法很多，除了支持向量机外还有数理统计方法、聚类分析等方法。

9.3.2　支持向量机的分类模型

记$x_1,\cdots,x_{30}$分别表示 30 个指标变量，已知观测样本为$[a_i,y_i]$（$i=1,\cdots,n$，这里$n=500$），其中$a_i\in R^{30}$，$y_i=1$为良性肿瘤，$y_i=-1$为恶性肿瘤。

我们首先进行线性分类，即要找一个最优分类面 $(\omega \cdot x)+b=0$，其中 $x=[x_1,\cdots,x_{30}]$，$\omega \in R^{30}$，$b \in R$，w,b待定，满足如下条件

$$\begin{cases}(\omega \cdot a_i)+b \geq 1, & y_i=1 \text{ 时,} \\ (\omega \cdot a_i)+b \leq -1, & y_i=-1 \text{ 时,}\end{cases}$$

即有 $y_i((\omega \cdot a_i)-b) \geq 1$，$i=1,\cdots,n$，其中，满足方程 $(\omega \cdot a_i)+b=\pm 1$的样本为支持向量。

要使两类总体到分类面的距离最大，则有

$$\max \frac{2}{\|\omega\|} \Rightarrow \min \frac{1}{2}\|\omega\|^2,$$

于是建立 SVM 的如下数学模型。

模型 1

$$\min \quad \frac{1}{2}\|\omega\|^2,$$

$$\text{s.t.} \quad y_i((\omega \cdot a_i)+b) \geq 1, \quad i = 1,2,\cdots,n.$$

求得最优值对应的ω^*, b^*，可得分类函数

$$g(x) = \mathrm{sgn}((\omega^* \cdot x) + b^*).$$

模型 1 是一个二次规划模型，为了利用 Matlab 求解模型 1，下面把模型 1 化为其对偶问题。

定义广义拉格朗日函数

$$L(\omega,\alpha)=\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{n}\alpha_i[1-y_i((\omega\cdot a_i)+b)],$$

其中$\alpha=[\alpha_1,\cdots,\alpha_n]^T\in R^{n+}$。由 KKT 互补条件，通过对$\omega$和$b$求偏导可得

$$\frac{\partial L}{\partial \omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i a_i=0,\ \frac{\partial L}{\partial b}=\sum_{i=1}^{n}\alpha_i y_i=0,$$

得$\omega=\sum_{i=1}^{n}\alpha_i y_i a_i$，$\sum_{i=1}^{n}\alpha_i y_i=0$，代入原始拉格朗日函数得

$$L=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(a_i\cdot a_j).$$

于是模型 1 可以化为

模型 2

$$\max \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(a_i\cdot a_j),$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}\alpha_i y_i=0,\\ 0\le\alpha_i, i=1,2,\cdots,n.\end{cases}$$

解此二次规划得到最优解α^*，从而得权重向量

$\omega^*=\sum_{i=1}^{n}\alpha_i^* y_i a_i$。

由 KKT 互补条件知

$$\alpha_i^*[1-y_i((\omega^*\cdot a_i)+b^*)]=0,$$

这意味着仅仅是支持向量a_i，使得α_i^*为正，所有其它样本对应的α_i^*均为零。选择α^*的一个正分量α_j^*，并以此计算

$$b^*=y_j-\sum_{i=1}^{n}y_i\alpha_i^*(a_i\cdot a_j).$$

最终的分类函数表达式如下

$$g(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i^*y_i(a_i\cdot x)+b^*). \qquad (9.8)$$

实际上，模型 2 中的$(a_i \cdot a_j)$是核函数的线性形式。非线性核函数可以将原样本空间线性不可分的向量转化到高维特征空间中线性可分的向量。

将模型 2 换成一般的核函数$K(x, y)$，可得一般的模型。

模型 3

$$\max \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(a_i,a_j),$$

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0, \\ 0 \le \alpha_i, i = 1,2,\cdots,n. \end{cases}$$

分类函数表达式

$$g(x) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(a_i,x) + b^*\right). \qquad (9.9)$$

9.3.3　分类模型的求解

第i（$i=1,\cdots,569$）个样本点的第j（$j=1,\cdots,30$）个指标的取值记作a_{ij}。

对于给定的 500 个训练样本，首先计算它们的均值向量$\mu=[\mu_1,\cdots,\mu_{30}]$和标准差向量$\sigma=[\sigma_1,\cdots,\sigma_{30}]$，对所有样本点数据利用如下公式进行标准化处理

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{\sigma_j},\quad i=1,\cdots,569,j=1,\cdots,30.$$

对应地，称

$$\tilde{x}_j = \frac{x_j - \mu_j}{s_j}, \quad (j = 1,2,\cdots,30)$$

为标准化指标变量。记$\tilde{x} = [\tilde{x}_1,\cdots,\tilde{x}_{30}]^T$。

记标准化后的 500 个已分类样本点数据行向量为$b_i = [\tilde{a}_{i1},\cdots,\tilde{a}_{i,30}]$（$i = 1,\cdots,500$）。

利用二次核函数的支持向量机模型进行分类，求得支持向量总共为 73 个，记支持向量为b_i（$i \in I$）。

分类函数为

$$c(\tilde{x}) = \sum_{i \in I} \beta_i K(b_i, \tilde{x}) + b \qquad (9.10)$$

其中$\beta_i = \alpha_i y_i$，$\tilde{x} = [\tilde{x}_1, \cdots, \tilde{x}_{30}]$。

当$c(\tilde{x}) \geq 0$，$\tilde{x}$属于第 1 类,即良性肿瘤;当$c(\tilde{x}) < 0$，$\tilde{x}$属于第-1 类，即恶性肿瘤。

所有已知样本点回代分类函数皆正确，故误判率为 0。

把 69 个测试样本$b_j, j = 501,502,\cdots,569$，代入分类函数（9.10），按如下规则分类

$g(b_j) \geq 0$，第j个样本点为良性肿瘤，

$g(b_j) < 0$，第j个样本点为恶性肿瘤.

求解结果见表 9.2。

表 9.2　分类结果

良　　性								恶　　性							
1	3	6	7	8	9	11	12	2	4	5	10	13	15	17	18
14	16	20	21	23	24	25	26	19	22	27	29	34	36	37	53
28	30	31	32	33	35	38	39	54	55	56	63	64	65	66	67
40	41	42	43	44	45	46	47	68							
48	49	50	51	52	57	58	59								
60	61	62	69												

注：这里的数字代表病例序号。

注：该问题没有使用线性内核函数进行分类，由于线性内核函数的错判率为 1.2%。