第3章 监督学习方法

监督学习是指利用带有标记的样本数据来完成拟合或者分类任务。第 2 章介绍的统计学习方法需要已知先验概率 $p(w_i)$ 和类条件概率密度 $p(\boldsymbol{x}|w_i)$ 的形式，其中类条件概率需要通过参数化或者非参数化方法进行估计。然而，在特征维数较高、内在关系复杂或者样本数量较少时，很难精确估计类条件概率密度。若能够直接通过样本数据生成拟合函数或者判别函数，则能够在一定程度上避免统计学习方法的局限。

基于样本数据的拟合函数或者分类器设计需要确定三个要素：① 拟合函数或者分类器类型（参数化函数集合）；② 拟合或分类的目标准则，即最优化模型的目标函数；③ 拟合函数或者分类模型的参数值，即设计最优化算法。不同的函数类型、不同的目标准则以及不同的优化算法决定了不同的拟合或分类效果。

本章内容分成 5 节：

3.1 节介绍的最小二乘法主要用于解决线性回归问题，而线性概率回归方法主要用于解决逻辑回归（或分类）问题。

3.2 节首先介绍支持向量机，主要用于解决线性二分类问题；接着介绍软间隔支持向量机，用于处理少量样本无法线性分类的情形；最后介绍支持向量回归，用于解决鲁棒线性回归问题。

3.3 节介绍的核方法主要用于解决非线性拟合或分类问题。其主要思路是将低维空间的非线性回归或分类问题转换成高维空间的线性回归或分类问题，然后设计核函数并使用最小二乘法或者支持向量机来实现拟合或分类功能。

3.4 节介绍神经网络以及误差反向传播算法。由于神经网络的强大拟合能力，其能够实现复杂非线性拟合或分类功能，并广泛应用于语音和图像处理。

3.5 节介绍复合学习方法，通过构建多个学习器来实现稳定可靠的学习或分类任务。复合方法包括空间上的并行方式、时间上的序贯方式以及在结构上的嵌套方式。除了基本的集成学习算法外，本节还简要介绍了迁移学习、终身学习和元学习的基本原理。

3.1 最小二乘法

3.1.1 线性回归

统计回归是指在给定自变量 $x_1, \cdots, x_d$ 观测值的情况下对因变量 y 进行预测，在社会和科学领域有着很多应用。若因变量表示庄稼的收成，则自变量可以是雨量、光照、土质等因素。在动态系统中，因变量往往是系统输出，而自变量是系统的过去行为表现。

令 $\boldsymbol{x} = [x_1, \cdots, x_d]^{\mathrm{T}}$。从数学描述来看，统计回归是指寻找一个回归量 $g(\boldsymbol{x})$ 使其跟因变量的观测值 y 尽量接近，即 $|y - g(\boldsymbol{x})|$ 的值尽量小。$\hat{y} = g(\boldsymbol{x})$ 称为预测值。在没有 $\boldsymbol{x}$ 和 y 的先验信息时，线性回归通常有如下形式:

$$g(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}, \quad \boldsymbol{x}, \boldsymbol{\theta} \in \mathbb{R}^d \tag{3.1}$$

假如在 $t=1,2,\cdots,N$ 时刻能够获得自变量和因变量的观测值，分别记为 $\boldsymbol{x}_t$ 和 y_t，则线性回归的拟合函数可定义为

$$\begin{aligned} V_N(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{t=1}^{N}[y_t - \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{\theta}]^2 \\ &= \frac{1}{N}\|\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\theta}\|^2 \end{aligned} \tag{3.2}$$

式中

$$\boldsymbol{Y}_N = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{X}_N = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{bmatrix}$$

线性拟合的目的是去寻找 $\boldsymbol{\theta}$ 值使 $V_N(\boldsymbol{\theta})$ 达到最小，即

$$\begin{aligned} &\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} V_N(\boldsymbol{\theta}) \\ &\text{s.t.} \quad V_N(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\theta}\|^2 \end{aligned}$$

上述优化问题的最优解也称为最小二乘估计。

1. 最小二乘解

通过对函数 $V_N(\boldsymbol{\theta})$ 求导置零，得到如下正规方程：

$$\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{X}_N\boldsymbol{\theta} = \boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{Y}_N$$

当矩阵 $\boldsymbol{X}_N$ 列满秩时，上述正规方程有唯一解：

$$\hat{\boldsymbol{\theta}}_N = (\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{X}_N)^{-1}\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{Y}_N = \boldsymbol{X}_N^{\dagger}\boldsymbol{Y}_N$$

式中：“$\dagger$”为 Moore-Penrose 伪逆。

2. 最优解的几何意义

上述线性拟合的最小二乘解可以从几何角度进行解释。正规方程等式可以写成如下等价形式：

$$\boldsymbol{X}_N^{\mathrm{T}}(\boldsymbol{X}_N\hat{\boldsymbol{\theta}}_N - \boldsymbol{Y}_N) = 0$$

从上式可以看出，回归向量 $\boldsymbol{X}_N$ 和残差向量 $\boldsymbol{E}_N = \boldsymbol{Y}_N - \boldsymbol{X}_N\hat{\boldsymbol{\theta}}_N$ 正交。因此，可以通过建立上述正交方程来获得最小二乘解。

3.1.2 逻辑回归

针对二值分类问题，假设训练数据集为 $\{\boldsymbol{X}_N, C_N\} = \{\boldsymbol{x}_n, c_n\}_{n=1}^N$，其中 $c_n \in \{0,1\}$。令 $\boldsymbol{x}_n$ 属于类别 $c_n = 1$ 的概率为

$$p(c_n = 1|\boldsymbol{x}_n) = \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) = \frac{\mathrm{e}^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b}}{1 + \mathrm{e}^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b}} = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n - b}}$$

式中：$\sigma(\cdot) \in [0,1]$ 为 Logistic 或者 Sigmoid 函数。

根据上述概率定义可得

$$\log \frac{p(c_n = 1|\boldsymbol{x}_n)}{p(c_n = 0|\boldsymbol{x}_n)} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b$$

由于概率似然比的对数为线性函数，因此求解 $\boldsymbol{w}$ 和 b 的问题称为线性概率回归。

二项逻辑回归可以推广到多项逻辑回归。若离散型随机变量 c_n 的取值集合为 $\{1, 2, \cdots, K\}$，则多项逻辑回归模型可写成

$$\begin{aligned} p(c_n = k|\boldsymbol{x}_n) &= \frac{\mathrm{e}^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_n + b_k}}{1 + \sum_{k=1}^{K-1} \mathrm{e}^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_n + b_k}}, \quad k = 1, \cdots, K-1 \\ p(c_n = K|\boldsymbol{x}_n) &= \frac{1}{1 + \sum_{k=1}^{K-1} \mathrm{e}^{\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_n + b_k}} \end{aligned} \tag{3.3}$$

1. 逻辑回归的优化

对于二分类问题，假设各训练数据相互独立，似然函数可写成

$$\begin{aligned} p(C_N|\boldsymbol{X}_N, \boldsymbol{w}, b) &= \prod_{n=1}^N p(c_n|\boldsymbol{x}_n, \boldsymbol{w}, b) \\ &= \prod_{n=1}^N p^{c_n}(c_n = 1|\boldsymbol{x}_n, \boldsymbol{w}, b) p^{1-c_n}(c_n = 0|\boldsymbol{x}_n, \boldsymbol{w}, b) \end{aligned}$$

相应的对数似然函数可以写成

$$\begin{aligned} L(\boldsymbol{w}, b) &= \sum_{n=1}^N c_n \log \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) + (1 - c_n) \log(1 - \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)) \\ &= \sum_{n=1}^N \log(1 - \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)) + \sum_{n=1}^N c_n(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b) \end{aligned}$$

通过极大似然法可获得最优参数 $\boldsymbol{w}, b$：

$$\{\boldsymbol{w}^*, b^*\} = \arg\max_{\boldsymbol{w}, b} \quad L(\boldsymbol{w}, b)$$

由于上述最优化问题不能直接得到闭式解，因此用梯度上升法来迭代计算，其中涉及的梯度信息包括

$$\partial_{\boldsymbol{w}} L(\boldsymbol{w}, b) = \sum_{n=1}^{N} \left[c_n \boldsymbol{x}_n - \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n + b) \boldsymbol{x}_n \right]$$

$$\partial_b L(\boldsymbol{w}, b) = \sum_{n=1}^{N} \left[c_n - \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n + b) \right]$$

利用关系式 $\partial_x \sigma(x) = \sigma(x)[1 - \sigma(x)]$ 可简化上式的梯度计算。梯度上升法通过如下公式进行迭代更新：

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \eta \partial_{\boldsymbol{w}} L(\boldsymbol{w}^k, b^k)$$

$$b^{k+1} = b^k + \eta \partial_b L(\boldsymbol{w}^k, b^k)$$

式中：η 为迭代步长，k 为迭代次数。

2. 逻辑回归的凹函数特性

当 $b = 0$ 时，对数似然函数 $L(\boldsymbol{w})$ 是关于 $\boldsymbol{w}$ 的凹函数，因此梯度上升法能够收敛到全局最优解。接下来将证明 $L(\boldsymbol{w})$ 为凹函数。$L(\boldsymbol{w})$ 的 Hessian 矩阵为

$$H_{i,j} = \frac{\partial^2 L}{\partial w_i \partial w_j} = -\sum_{n=1}^{N} x_{n,i} x_{n,j} \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)[1 - \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)]$$

对于任意的向量 $\boldsymbol{z}$，如下不等式成立：

$$\begin{aligned} \sum_{i,j} z_i H_{i,j} z_j &= -\sum_{i,j,n} z_i x_{n,i} z_j x_{n,j} \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)[1 - \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)] \\ &\leqslant -\sum_{n} \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)[1 - \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n)] \left(\sum_i z_i x_{n,i} \right)^2 \leqslant 0 \end{aligned}$$

因此，$L(\boldsymbol{w})$ 是关于 $\boldsymbol{w}$ 的凹函数。

3.1.3 均方误差估计

对于线性拟合模型 $y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{w}$，其中 $\boldsymbol{x}$ 为随机输入向量，$\boldsymbol{w}$ 为权重向量，y 为期望输出。权重向量 $\boldsymbol{w}$ 通过最小化期望输出和实际输出之间的均方误差进行求解：

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} J(\boldsymbol{w})$$

$$\text{s.t.} \quad J(\boldsymbol{w}) = \mathcal{E}[\|y - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{w}\|^2]$$

通过对 $J(\boldsymbol{w})$ 关于 $\boldsymbol{w}$ 求导置零得到

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\mathcal{E}[\boldsymbol{x}(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w})] = 0 \tag{3.4}$$

由此得到

$$\hat{\boldsymbol{w}} = \mathcal{E}^{-1}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]\mathcal{E}[\boldsymbol{x}y] \tag{3.5}$$

式中：$\mathcal{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$ 和 $\mathcal{E}[\boldsymbol{x}y]$ 分别为自相关矩阵和互相关矩阵，通常需要知道概率分布函数来计算理论期望值或者通过对大量样本计算经验期望值。Robbins 和 Monro 利用随机逼近理论提供了一种解决方案。考虑如下形式方程：

$$\mathcal{E}[f(\boldsymbol{x}_k, \boldsymbol{w})] = 0$$

式中：$f(\cdot)$ 为非线性函数；$\boldsymbol{x}_k (k = 1, 2, \cdots)$ 为满足同分布的随机向量序列；$\boldsymbol{w}$ 为未知参数向量。

可以看出，线性方程 (3.4) 是上述方程一类特殊形式。随机逼近算法采用如下迭代策略：

$$\hat{\boldsymbol{w}}(k) = \hat{\boldsymbol{w}}(k-1) + \rho_k f(\boldsymbol{x}_k, \hat{\boldsymbol{w}}(k-1))$$

式中：ρ_k 为时变步长。

针对上述迭代方程，当序列 ρ_k 满足条件

$$\sum_{k=1}^{\infty} \rho_k \to \infty, \quad \sum_{k=1}^{\infty} \rho_k^2 < \infty$$

时，估计值 $\hat{\boldsymbol{w}}_k$ 将均方收敛到方程 $\mathcal{E}[f(\boldsymbol{x}_k, \boldsymbol{w})] = 0$ 的真实解 $\boldsymbol{w}^*$，即

$$\lim_{k \to \infty} \mathcal{E}[\|\hat{\boldsymbol{w}}(k) - \boldsymbol{w}^*\|^2] = 0$$

例 3.1（随机逼近）

考虑简单方程 $\mathcal{E}[\boldsymbol{x}_k - \boldsymbol{w}] = 0$。令 $\rho_k = 1/k$，则迭代策略为

$$\hat{\boldsymbol{w}}(k) = \hat{\boldsymbol{w}}(k-1) + \frac{1}{k}[\boldsymbol{x}_k - \hat{\boldsymbol{w}}(k-1)] = \frac{k-1}{k}\hat{\boldsymbol{w}}(k-1) + \frac{1}{k}\boldsymbol{x}_k$$

可以看出，当 k 值趋向于无穷时，有 $\hat{\boldsymbol{w}}(k) \to \sum_{i=1}^{k} \frac{\boldsymbol{x}_i}{k}$。

对于线性方程(3.4)，随机逼近的迭代方程可写成

$$\hat{\boldsymbol{w}}(k) = \hat{\boldsymbol{w}}(k-1) + \rho_k \boldsymbol{x}_k \left(y_k - \boldsymbol{x}_k^{\mathrm{T}}\hat{\boldsymbol{w}}(k-1)\right)$$

上述迭代算法将收敛到最小均方估计值。

3.2 支持向量机

支持向量机最初由学者 Vapnik 提出，其基本模型主要用于解决二分类问题。SVM 分类器不同于一般的分类器，具有更好的鲁棒性和泛化能力，因此被广泛应用于线性分类和非线性分类。

3.2.1 标准支持向量机

给定一个二分类数据集 $\mathcal{D}=\{\boldsymbol{x}_t, y_t\}_{t=1}^N$，其中 $y_t \in \{+1,-1\}$。若两类样本是线性可分的，则存在一个超平面

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b = 0$$

使得

$$y_t = \begin{cases} 1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b > 0 \\ -1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b < 0 \end{cases}$$

若两类样本线性可分，则意味存在 $\boldsymbol{w}$ 和 b 使得任意样本对 $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ 都满足 $y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b] > 0$。

1. 支持向量机模型

数据集 $\mathcal{D}$ 中每个样本 $\boldsymbol{x}_t$ 到分类超平面的距离为

$$\gamma_t = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b|}{\|\boldsymbol{w}\|} = \frac{y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b]}{\|\boldsymbol{w}\|}$$

定义间隔 $\boldsymbol{\gamma}$ 为整个数据集 $\mathcal{D}$ 中所有样本到超平面的最短距离：

$$\boldsymbol{\gamma} = \min_t \ |\boldsymbol{\gamma}_t|$$

若间隔 γ 越大，则分割超平面对两类数据的划分越稳定，不容易受到观测噪声等因素的影响（如图 3.1 所示）。支持向量机的设计目标是寻找超平面的参数 $(\boldsymbol{w}, b)$ 使得最小间隔 γ 最大化，即

$$\begin{aligned} &\max_{\boldsymbol{w},b} 2\gamma \\ &\text{s.t.} \quad \frac{y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b]}{\|\boldsymbol{w}\|} > \gamma, \quad \forall t = 1, \cdots, N \end{aligned}$$

由于满足 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b = 0$ 的 $(\boldsymbol{w}, b)$ 值存在尺度不确定性，为此可以限制 $\|\boldsymbol{w}\|\gamma = 1$，使得上述优化问题等价转换成

$$\begin{aligned} &\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \\ &\text{s.t.} \quad y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b] \geqslant 1, \quad \forall t = 1, \cdots, N \end{aligned}$$

或者

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{\|\boldsymbol{w}\|^2}{2} \\ &\text{s.t.} \quad y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b] \geqslant 1,\ \forall t = 1, \cdots, N \end{aligned} \tag{3.6}$$

不难看出，上述优化问题为凸优化问题。当两类样本线性可分时，该优化问题具有强对偶性。

数据集中所有满足 $y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b] = 1$ 的样本点称为支持向量。由于支持向量机是具有间隔最大的分类超平面，因此其通常具有唯一性。

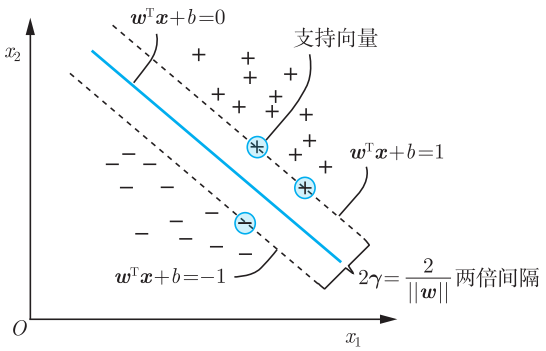

图 3.1　分类间隔与最优超平面

2. 支持向量机的优化

对支持向量机所对应的优化问题式 (3.6) 采用拉格朗日方法进行求解。其对应的拉格朗日函数可以写成

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{\|\boldsymbol{w}\|^2}{2} + \sum_{t=1}^{N} \lambda_t \left(1 - y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b]\right), \lambda_t \geqslant 0 \tag{3.7}$$

分别计算 L 关于 $\boldsymbol{w}$、b 的导数并置零，得到

$$\boldsymbol{w} = \sum_{t=1}^{N} \lambda_t y_t \boldsymbol{x}_t \tag{3.8}$$

$$0 = \sum_{t=1}^{N} \lambda_t y_t \tag{3.9}$$

将式 (3.8) 代入式 (3.7)，然后利用式 (3.9) 得到如下拉格朗日对偶函数：

$$J(\boldsymbol{\lambda}) = -\frac{1}{2} \sum_{t=1,k=1}^{N,N} \lambda_t \lambda_k y_t y_k \boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{x}_k + \sum_{t=1}^{N} \lambda_t$$

由此，支持向量机的对偶优化问题可以写成

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \ & -\frac{1}{2}\sum_{t=1,k=1}^{N,N}\lambda_t\lambda_k y_t y_k \boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{x}_k + \sum_{t=1}^{N}\lambda_t \\ \text{s.t.} \ & \sum_{t=1}^{N}\lambda_t y_t = 0 \\ & \lambda_t \geqslant 0 \end{aligned} \tag{3.10}$$

当原优化问题 (3.6) 具有强对偶性或者数据集线性可分时，可以通过最大化对偶优化问题来进行求解。不难看出对偶问题的目标函数是凹函数，而约束条件为线性约束。因此，最大化对偶优化问题能获得全局最优解。通常采用比较高效的序贯最小优化算法。每次迭代选择变量 λ_i 和 λ_j，在固定其他参数后，对偶优化问题式 (3.10) 仅需优化两个参数，其对应的约束可重写成

$$\lambda_i y_i + \lambda_j y_j = c \tag{3.11}$$

式中：$c = -\sum_{t\neq i,j}\lambda_t y_t$。

通过上述等式可以消去变量 λ_j，得到一个关于 λ_i 的单变量二次规划问题，仅存在的约束为 $\lambda_i \geqslant 0$。由此可以得到单变量二次规划问题的闭式解，从而提高优化效率。

根据优化理论中的 KKT 互补松弛条件，最优解 $(\lambda_t^*, \boldsymbol{w}^*, b^*)$ 需要满足

$$\lambda_t^*[1 - y_t(\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t + b^*)] = 0$$

若样本不在约束边界上，即 $1 - y_t(\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t + b^*) < 0$，则对应的 $\lambda_t = 0$；若样本在约束边界上，则 λ_t 可以不为零。

支持向量机的最优权重向量 $\boldsymbol{w}$ 需满足式 (3.8)，即 $\boldsymbol{w}$ 的最优值依赖 λ_t，而非零的 λ_t 值只发生在支持向量上。因此，支持向量机的分类超平面表达式只依赖支持向量，具有很好的稀疏特性。这就是该分类器称为“支持向量机”的原因。

通过求解对偶优化问题式 (3.10) 获得对偶变量 λ_t 的最优值后，根据式 (3.8) 得到最优权重 $\boldsymbol{w}^*$。到目前为止，如何求解偏置 b 的最优值还没有得到解决。通过互补松弛分析，非零 λ_t 所对应的样本 $\boldsymbol{x}_t$ 被确定为支持向量。若 $\boldsymbol{x}_t$ 为支持向量，则最优偏置 b 通过求解如下方程来获得：

$$y_t[\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t + b^*] = 1$$

或者

$$b^* = y_t - \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t \tag{3.12}$$

最终的支持向量分类器可以写成

$$y_t = \mathrm{sgn}[\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t + b^*]$$

3. 支持向量机算法描述及示例

综上所述，给定线性可分训练数据集合，支持向量机的分类超平面可以通过算法 3.1来获得。

算法 3.1（支持向量机学习算法）

输入：线性可分训练数据集合 $\{\boldsymbol{x}_t, y_t\}_{t=1}^N$。

（1）构造并求解对偶优化问题式 (3.10)，得到最优解 $\{\lambda_t^*\}_{t=1}^N$。

（2）通过式 (3.8) 计算权重向量 $\boldsymbol{w}^*$，确定非零 λ_t^* 对应的样本为支持向量，并通过式 (3.12) 得到最优偏置 b^*。

（3）求得分类超平面 $\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x} + b^* = 0$。

输出：分类决策函数 $y_t = \mathrm{sgn}[\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t + b^*]$。

接下来将通过一个简单例子来阐述支持向量机的应用。

例 3.2（支持向量机示例）

给定正样本点 $\boldsymbol{x}_1 = (3,3), \boldsymbol{x}_2 = (4,3)$ 和负样本点 $\boldsymbol{x}_3 = (1,1)$，试设计线性可分支持向量机。

解：该问题可以采用几何方法或代数方法来求解。

第一种方法为几何方法。根据支持向量机的基本原理，首先确定支持向量 $(3,3)$ 和 $(1,1)$，然后根据支持向量确定最优分类面为

$$x_1 + x_2 = 4$$

第二种方法为代数方法，即采用算法 3.1进行求解。首先，根据给定数据对偶优化问题可以写成

$$\begin{aligned}
&\min_{\boldsymbol{\lambda}} \frac{1}{2}\sum_i\sum_j \frac{1}{2}(18\lambda_1^2 + 25\lambda_2^2 + 2\lambda_3^2 + 42\lambda_1\lambda_2 - 12\lambda_1\lambda_3 - 14\lambda_2\lambda_3) - \lambda_1 - \lambda_2 - \lambda_3 \\
&\text{s.t.} \quad \lambda_1 + \lambda_2 - \lambda_3 = 0 \\
&\qquad \lambda_i \geqslant 0,\ i = 1,2,3
\end{aligned}$$

求解上述优化问题可得最优解 $\boldsymbol{\lambda} = (1/4, 0, 1/4)$。然后，根据式 (3.8) 和式 (3.12) 可得最优权重向量 $(1/2, 1/2)$ 和最优偏置值 -2。由此可以得到如下分类决策函数：

$$y = \mathrm{sgn}\left(\frac{x_1}{2} + \frac{x_2}{2} - 2\right)$$

3.2.2 软间隔与正则化

标准支持向量机主要是针对线性可分的数据集，即约束条件相对比较严格。当数据集线性不可分时，标准支持向量机无法找到最优解（如图 3.2 所示）。为解决该问题，引入松弛变量 ξ_t 来容忍不满足约束条件的样本，但是需要对其进行惩罚。根据该思想，改进支持向量机的优化问题可以写成

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ & \frac{\|\boldsymbol{w}\|^2}{2} + c\sum_{t=1}^{N}\xi_t \\ \text{s.t.} \ \ & y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b] \geqslant 1 - \xi_t \\ & \xi_t \geqslant 0, \quad \forall t = 1, \cdots, N \end{aligned} \tag{3.13}$$

其中，参数 $c > 0$ 用来控制间隔和松弛变量惩罚之间的平衡。引入松弛变量的间隔称为软间隔。

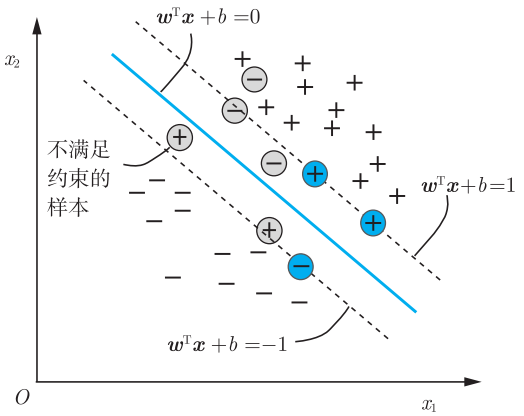

图 3.2 软间隔与正则化

类似于标准支持向量机的求解方式，构建如下拉格朗日函数：

$$\begin{aligned} L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\alpha}) = & \frac{\|\boldsymbol{w}\|^2}{2} + c\sum_{t=1}^{N}\xi_t \\ & + \sum_{t=1}^{N}\lambda_t[1 - \xi_t - y_t(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b)] - \sum_{t=1}^{N}\alpha_t\xi_t \end{aligned}$$

式中：$\lambda_t, \alpha_t \geqslant 0$ 为拉格朗日乘子。

对上述拉格朗日函数关于 $\boldsymbol{w}$、b、ξ_t 求偏导置零，得到

$$\begin{cases} \boldsymbol{w} = \sum\limits_{t=1}^{N} \lambda_t y_t \boldsymbol{x}_t \\ 0 = \sum\limits_{t=1}^{N} \lambda_t y_t \\ c = \lambda_t + \alpha_t \end{cases} \tag{3.14}$$

将上式代入拉格朗日函数后得到如下对偶优化问题：

$$\begin{aligned} \max_{\boldsymbol{\lambda}} & \sum_{t=1}^{N} \lambda_t - \frac{1}{2} \sum_{t=1,k=1}^{N,N} \lambda_t \lambda_k y_t y_k \boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{x}_k \\ \text{s.t.} & \sum_{t=1}^{N} \lambda_t y_t = 0 \\ & 0 \leqslant \lambda_t \leqslant c, \quad \forall t = 1, \cdots, N \end{aligned} \tag{3.15}$$

通过最大化上述凹优化问题可以获取 λ_t 的全局最优解，然后将其代入式 (3.14) 得到权重向量的 $\boldsymbol{w}$ 最优解。接下来将讨论如何求解最优偏置 b。

根据优化理论的 KKT 互补松弛条件，可以得到

$$\lambda_t \left(y_t [\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_t + b] + \xi_t - 1 \right) = 0$$

$$\alpha_t \xi_t = 0$$

$$c = \lambda_t + \alpha_t$$

通过 λ_t 的值可以判断样本 $\boldsymbol{x}_t$ 是否为支持向量。若 $\lambda_t > 0$，则样本 $\boldsymbol{x}_t$ 满足 $y_t[\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_t + b] + \xi_t = 1$。同时，根据约束 $\lambda_t + \alpha_t = c$，判断出当 $0 < \lambda_t < c$ 时，$\alpha_t > 0$ 并且 $\xi_t = 0$。因此，若 $0 < \lambda_t < c$，偏置 b 的最优解可以从其对应的样本和标记得到

$$b^* = \frac{1}{y_t} - \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x}_t = y_t - \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x}_t \tag{3.16}$$

最终，基于软间隔正则化的支持向量分类器可以写成

$$y_t = \operatorname{sgn}[\boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x}_t + b^*]$$

算法 3.2（软间隔支持向量机学习算法）

输入：训练数据集合 $\{\boldsymbol{x}_t, y_t\}_{t=1}^{N}$。

（1）选择惩罚参数 c，构造并求解对偶优化问题式 (3.15)，得到最优解 $\{\lambda_t^*\}_{t=1}^{N}$。

（2）通过式 (3.14) 计算权重向量 $\boldsymbol{w}^*$，确定 $0 < \lambda_t^* < c$ 对应的样本为支持向量，并通过

式 (3.16) 得到最优偏置 b^*。

(3) 求得分类超平面 $\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}+b^*=0$。

输出：分类决策函数 $y_t=\mathrm{sgn}[\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t+b^*]$。

3.2.3 支持向量回归

具有软间隔的支持向量机式 (3.13) 可以表示成经验风险 + 正则化的形式：

$$\min_{\boldsymbol{w},b}\sum_{t=1}^{N}\max(0,1-y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t+b])+\frac{\|\boldsymbol{w}\|^2}{2c}$$

式中：$\max(0,1-y_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t+b])$ 为损失函数；$\dfrac{\|\boldsymbol{w}\|^2}{2c}$ 为正则化项。

接下来将基于上述正则化形式设计支持向量回归模型。

不同于分类问题，我们希望获得一个回归模型 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t+b$，使其与 y_t 尽可能接近。支持向量回归与传统回归不同的地方在于容许 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t+b$ 和 y_t 之间存在最多 ϵ 的偏差，当 $|y_t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t-b|>\epsilon$ 时才计算损失 (如图 3.3 所示)。

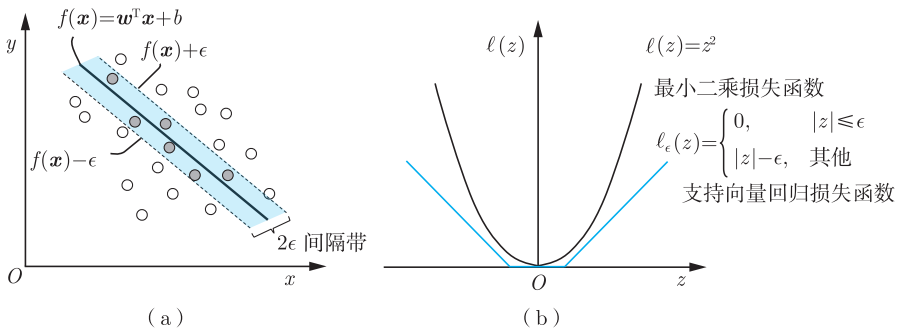

图 3.3 支持向量回归

支持向量回归的优化形式可以写成

$$\min_{\boldsymbol{w},b}\sum_{t=1}^{N}\max\{0,|y_t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t-b|-\epsilon\}+\frac{\|\boldsymbol{w}\|^2}{2c}$$

或者

$$\begin{aligned}\min_{\boldsymbol{w},b,\bar{\boldsymbol{\xi}},\underline{\boldsymbol{\xi}}}\ &\frac{\|\boldsymbol{w}\|^2}{2}+c\sum_{t=1}^{N}(\bar{\xi}_t+\underline{\xi}_t)\\ \text{s.t.}\quad &y_t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t-b\leqslant\epsilon+\bar{\xi}_t\\ &\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t+b-y_t\leqslant\epsilon+\underline{\xi}_t\end{aligned}$$

$$\bar{\xi}_t \geqslant 0, \underline{\xi}_t \geqslant 0, \quad \forall t = 1, \cdots, N$$

对应的拉格朗日函数可以写成

$$\begin{aligned} & L(\boldsymbol{w}, b, \bar{\boldsymbol{\xi}}, \underline{\boldsymbol{\xi}}, \bar{\boldsymbol{\lambda}}, \underline{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}, \underline{\boldsymbol{\alpha}}) \\ = & \frac{\|\boldsymbol{w}\|^2}{2} + c\sum_{t=1}^{N}(\bar{\xi}_t + \underline{\xi}_t) - \sum_{t=1}^{N}\bar{\alpha}_t\bar{\xi}_t - \sum_{t=1}^{N}\underline{\alpha}_t\underline{\xi}_t \\ & + \sum_{t=1}^{N}\bar{\lambda}_t[y_t - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t - b - \epsilon - \bar{\xi}_t] + \sum_{t=1}^{N}\underline{\lambda}_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b - y_t - \epsilon - \underline{\xi}_t] \end{aligned}$$

式中：$\bar{\boldsymbol{\lambda}}, \underline{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}, \underline{\boldsymbol{\alpha}} \geqslant 0$ 为拉格朗日乘子。

对上述拉格朗日函数关于 $\boldsymbol{w}$、$\boldsymbol{b}$、$\bar{\boldsymbol{\xi}}$、$\underline{\boldsymbol{\xi}}$ 求导置零，可得

$$\begin{aligned} \boldsymbol{w} &= \sum_{t=1}^{N}(\bar{\lambda}_t - \underline{\lambda}_t)\boldsymbol{x}_t \\ 0 &= \sum_{t=1}^{N}(\bar{\lambda}_t - \underline{\lambda}_t) \\ c &= \bar{\lambda}_t + \bar{\alpha}_t \\ c &= \underline{\lambda}_t + \underline{\alpha}_t \end{aligned} \tag{3.17}$$

将上式代入拉格朗日函数，得到如下对偶优化问题：

$$\begin{aligned} \max_{\bar{\boldsymbol{\lambda}}, \underline{\boldsymbol{\lambda}}} & \sum_{t=1}^{N} y_t[\bar{\lambda}_t - \underline{\lambda}_t] - \epsilon(\bar{\lambda}_t + \underline{\lambda}_t) \\ & - \frac{1}{2}\sum_{t=1,k=1}^{N,N}(\bar{\lambda}_t - \underline{\lambda}_t)(\bar{\lambda}_k - \underline{\lambda}_k)\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{x}_k \\ \text{s.t.} \quad & \sum_{t=1}^{N}(\bar{\lambda}_t - \underline{\lambda}_t) = 0 \\ & 0 \leqslant \bar{\lambda}_t, \underline{\lambda}_t \leqslant c, \quad \forall t = 1, \cdots, N \end{aligned} \tag{3.18}$$

通过求解上述对偶优化问题得到 $\bar{\lambda}_t$、$\underline{\lambda}_t$ 的最优值，代入式 (3.17) 进一步得到 $\boldsymbol{w}$ 的最优解。接下来讨论如何获偏置 b 的最优解。

根据优化理论的 KKT 松弛互补条件可以得到

$$\bar{\lambda}_t[y_t - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t - b - \epsilon - \bar{\xi}_t] = 0$$

$$\underline{\lambda}_t[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + b - y_t - \epsilon - \underline{\xi}_t] = 0$$

$$(c-\bar{\lambda}_t)\bar{\xi}_t=0,\quad (c-\underline{\lambda}_t)\underline{\xi}_t=0$$

$$c=\bar{\lambda}_t+\bar{\alpha}_t$$

$$c=\underline{\lambda}_t+\underline{\alpha}_t$$

若 $0<\bar{\lambda}_t<c$，则有 $\bar{\xi}_t=0$。因此，最优偏置可以从上述第一个等式计算得到

$$b=y_t+\epsilon-\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t \tag{3.19}$$

最终得到支持向量回归模型：$y_t=\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t+b^*$。

算法 3.3（支持向量回归学习算法）

输入：训练数据集合 $\{\boldsymbol{x}_t,y_t\}_{t=1}^N$，回归容许偏差 ϵ。

（1）选择惩罚参数 c，构造并求解对偶优化问题式 (3.18)，得到最优解 $\{\bar{\lambda}_t^*,\underline{\lambda}_t^*\}_{t=1}^N$。

（2）通过式 (3.17) 计算权重向量 $\boldsymbol{w}^*$，确定 $0<\bar{\lambda}_t^*<c$ 对应的样本为支持向量，并通过式 (3.19) 得到最优偏置 b^*。

输出：支持向量回归模型 $y_t=\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}_t+b^*$。

3.3 核方法与正则化

不同于线性回归，非线性映射通常具有比较复杂的几何特征，其通常记成

$$y=f(\boldsymbol{x})$$

为了能够便于处理，通常需要寻找 $f(\cdot)$ 函数的参数化形式，使得该参数化回归函数具有灵活表示形式，而且能够覆盖所有合理的非线性行为。接下来讨论如何采用基函数或者核函数来实现非线性函数的参数化表示。

3.3.1 广义线性模型

给定样本 $(\boldsymbol{x},y)$，广义线性估计可以表示成

$$\hat{y}=g(\boldsymbol{x},\boldsymbol{\alpha})=\sum_{k=1}^{K}\alpha_k\phi_k(\boldsymbol{x}) \tag{3.20}$$

式中：$\phi_k(\cdot)$ 为预先选择的基函数。

通常，基函数的选择有不同的形式，它们赋予了非线性函数研究的统一框架。

若采用泰勒展开来获得非线性函数的近似线性化表示，则其对应的基函数为

$$\phi_k(\boldsymbol{x})=\boldsymbol{x}^k:=\left\{\prod_{i=1}^{d}x_i^{\beta_i}\,|\,\sum_{i=1}^{d}\beta_i=k,\beta_i\geqslant 0\right\}$$

这类基函数也称为多项式基函数。多项式或者泰勒展开通常由 Weierstrass 定理来保证其拟合精度：任何定义在有界闭区间上的连续函数，总能够找到多项式展开使得其误差处处（一致）小于某一固定值。上述多项式展开有时也称为 Volterra 展开。

常见的基函数包括多尺度基函数 $\phi_k(x)$，具有如下特征：

（1）所有的基函数 $\phi_k(x)$ 均由一个母函数生成，该母函数表示为 $\kappa(\cdot):\mathbb{R}\to\mathbb{R}$；

（2）基函数 $\phi_k(x)$ 可以表示成

$$\phi_k(x)=\kappa[\beta_k(x-\gamma_k)]$$

式中：β_k 为尺度膨胀参数；γ_k 为平移参数。

常见的母函数包括傅里叶级数、高斯钟函数、分段定常函数和 Sigmoid 函数。

采用多尺度基函数，非线性映射函数 $f(x)$ 可以近似表示成

$$f(x)=\sum_{k=1}^{K}\alpha_k\kappa[\beta_k(x-\gamma_k)] \tag{3.21}$$

值得注意的是，单变量的基函数可以分为局部基函数和全局基函数。局部基函数是指函数值在局部发生剧烈变化，如高斯钟、分段定常和 Sigmoid 函数。全局基函数是指函数值对定义域上的所有值均会产生较大变化，如傅里叶变换和 Volterra 展开。

1. 有限维空间的线性可分容量

广义线性拟合能够解决复杂的非线性拟合问题，同时也能够解决低维空间不可分的分类问题。比如二维空间中的四点 $\{(0,0),(0,1),(1,0),(1,1)\}$，若映射函数为异或运算，则 $(0,0)$ 和 $(1,1)$ 为一类，而 $(0,1)$ 和 $(1,0)$ 为另一类。显然，这两类数据不能简单地由一条直线分割开来。低维空间中的不可分问题通常用以下方法来解决：先通过非线性方法将回归向量变换到高维空间，再进行线性分类。上述方法的合理性用模式可分性 Cover 定理来说明：“假设样本空间不是稠密分布的，将复杂的模式分类问题非线性地投影到高维空间，比投影到低维空间更容易线性可分。”

考虑 l 维空间中的 N 个点，若任意 $l+1$ 个点不在 $(l-1)$ 维超平面上时，则这些点具有良态分布。通常，具有良态分布的 3 个样本点不会出现在二维平面上的同一条直线上。

定理 3.1（Function Counting 定理） *若采用 $l-1$ 维的超平面对具有良态分布的 N 个 l 维样本进行二分类，其分类结果数 $\mathcal{P}(N,l)$ 可以表示成*

$$\mathcal{P}(N,l)=\begin{cases}2\displaystyle\sum_{i=0}^{l}\binom{N-1}{i}, & N>l+1\\ 2^N, & N\leqslant l+1\end{cases}$$

式中

$$\binom{N-1}{i}=\frac{(N-1)!}{(N-1-i)!i!}$$

考虑到 N 个样本所有可能的二分类组合数有 2^N 种，而且当 $N \leqslant l+1$ 时有 $\mathcal{P}(N,l) = 2^N$。因此，N 个 l 维样本可线性分类的概率为

$$P_N^l = \frac{\mathcal{P}(N,l)}{2^N} = \begin{cases} \dfrac{1}{2^{N-1}} \displaystyle\sum_{i=0}^{l} \begin{pmatrix} N-1 \\ i \end{pmatrix}, & N > l+1 \\ 1, & N \leqslant l+1 \end{cases}$$

通过对 Function Counting 定理的分析得到结论：在给定样本数量 N 的情况下，若式 (3.21) 中基函数的数量 K 越大，则线性可分的概率越大。因此，使用非线性函数将低维回归向量投影到高维空间行进线性分类的方法具有基本理论支撑。

2. 核函数

当 d 维回归向量通过 $\boldsymbol{\phi}(\boldsymbol{x})$ 转换到很高维新特征向量时，其对应的拟合计算量就会增加，从而导致维数灾难。基于新特征向量 $\boldsymbol{\phi}(\boldsymbol{x})$，接下来将考虑如何设计支持向量机。

类似于标准支持向量机的推导，新特征向量对应的最优权重 $\boldsymbol{w}$ 可以表示成

$$\boldsymbol{w}^* = \sum_{t=1}^{N} \lambda_t y_t \boldsymbol{\phi}(\boldsymbol{x}_t)$$

在新特征空间中，对新样本 $\boldsymbol{x}_{t+1}$ 的分类可以表示成

$$y_{t+1} = \operatorname{sgn}\left(\sum_{t=1}^{N} \lambda_t y_t \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_t) \boldsymbol{\phi}(\boldsymbol{x}_{t+1}) + b^* \right)$$

从上述分类器可以看出，虽然 $\phi(\boldsymbol{x})$ 可能维度很高而且难以表示，但是分类器中与 $\boldsymbol{\phi}(\boldsymbol{x})$ 相关的项 $\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}_t)\boldsymbol{\phi}(\boldsymbol{x}_{t+1})$ 是一个标量。因此，若知道 $\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{z})$ 的表达式，特征空间的分类器就能够快速计算。在这里，$\kappa(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{z})$ 称为内积核函数。

定理 3.2 对称函数 $\kappa(\boldsymbol{x},\boldsymbol{z}) = \kappa(\boldsymbol{z},\boldsymbol{x})$ 为正定核函数的充分必要条件为任意数据集 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ 所对应的如下核矩阵总是半正定的：

$$\boldsymbol{K} = \begin{bmatrix} \kappa(\boldsymbol{x}_1,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_m) \\ \kappa(\boldsymbol{x}_2,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_2,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_2,\boldsymbol{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\boldsymbol{x}_m,\boldsymbol{x}_1) & \kappa(\boldsymbol{x}_m,\boldsymbol{x}_2) & \cdots & \kappa(\boldsymbol{x}_m,\boldsymbol{x}_m) \end{bmatrix}$$

上述定理表明，若核函数所对应的核矩阵是半正定的，它就能作为核函数使用。反之，对于任意一个半正定核矩阵，总能找到一个与之对应的特征映射函数 $\boldsymbol{\phi}(\cdot)$。

采用核函数能够避免对高维特征向量的直接计算，然而核函数的选择对分类器的性能至关重要。在不知道特征映射函数的情况下，并不知道什么样的核函数是合适的。常用的核函数有如下几类：

高斯核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right)$$

式中：σ 为参数。

线性核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z})^r$$

式中：r 为参数。

多项式核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}+c)^r$$

式中：r 为参数。

拉普拉斯核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=\exp(-t\|\boldsymbol{x}-\boldsymbol{z}\|)$$

式中：t 为参数。

Spline 核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=B_{2p+1}(\|\boldsymbol{x}-\boldsymbol{z}\|^2)$$

式中：B_n 样条曲线由 $n+1$ 个单位区间函数 $[-1/2,1/2]$ 卷积得到。

Sigmoid 核：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=\tanh(\beta\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}+\theta),\quad \beta>0,\theta<0$$

基于上述几种基本核函数，通过如下一系列组合操作可以得到新的核函数：

（1）线性组合 $\alpha_1\kappa_1(\boldsymbol{x},\boldsymbol{z})+\alpha_2\kappa_2(\boldsymbol{x},\boldsymbol{z}),\forall\alpha_1,\alpha_2>0$ 是核函数；

（2）直积组合 $\kappa_1(\boldsymbol{x},\boldsymbol{z})\kappa_2(\boldsymbol{x},\boldsymbol{z})$ 是核函数；

（3）对于任意函数 $g(\boldsymbol{x}):\mathbb{R}^d\to\mathbb{R}$，$g(\boldsymbol{x})g(\boldsymbol{z})$ 和 $g(\boldsymbol{x})\kappa(\boldsymbol{x},\boldsymbol{z})g(\boldsymbol{z})$ 是核函数；

（4）指数和多项式组合：$\exp[\kappa(\boldsymbol{x},\boldsymbol{z})]$ 和 $[\kappa(\boldsymbol{x},\boldsymbol{z})]^r$ 是核函数。

3. 表示定理

表示定理在实际应用中具有重要作用：通过有限训练样本能够对经验损失函数进行快速优化，即使待估计函数具有很高维度。

定理 3.3（表示定理） 令 $\Omega:[0,+\infty)\to\mathbb{R}$ 为任意严格单调增函数，$L:\mathbb{R}^2\to\mathbb{R}$ 为任意非负损失函数，$\mathbb{H}$ 为再生核希尔伯特空间。下列最小正则化问题

$$\min_{f\in\mathbb{H}}\sum_{t=1}^{N}L[y_t,f(\boldsymbol{x}_t)]+\lambda\Omega(\|f\|^2)\tag{3.22}$$

的最优解具有如下表示形式：

$$f(\boldsymbol{x}) = \sum_{t=1}^{N} \theta_t \kappa(\boldsymbol{x}, \boldsymbol{x}_t)$$

上述表示定理对损失函数没有限制，对正则化项仅要求单调递增，甚至不要求其为凸函数。这意味着对于一般损失函数和正则化项，最优解都能表示为核函数的线性组合，这体现了核函数在实际应用中的重要性。

例 3.3（核表示的最小平方解）

令 $(\boldsymbol{x}_i, y_i)(i = 1, 2, \cdots, N)$ 为训练样本。试设计最小平方线性分类器，即

$$\min_{g \in \mathbb{H}} \sum_{i=1}^{N} (y_i - g(\boldsymbol{x}_i))^2$$

解：由表示定理能够得到

$$g(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \kappa(\boldsymbol{x}, \boldsymbol{x}_j)$$

因此，函数 g 的求解转换成参数 $\boldsymbol{a}$ 的求解。令

$$\begin{aligned} J(\boldsymbol{a}) &= \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{N} a_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{a}) \end{aligned}$$

则最优 $\boldsymbol{a}$ 可通过求解如下最小二乘问题获得：

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}} \ J(\boldsymbol{a})$$

通过对 $\boldsymbol{a}$ 求导置零得到

$$\boldsymbol{a}^* = \boldsymbol{K}^{-1} \boldsymbol{y}$$

从而最优函数 $g(\boldsymbol{x})$ 可以写成

$$g(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}} p(\boldsymbol{x}) = \boldsymbol{y}^{\mathrm{T}} \boldsymbol{K}^{-1} p(\boldsymbol{x})$$

式中

$$p(\boldsymbol{x}) = [\kappa(\boldsymbol{x}, \boldsymbol{x}_1), \cdots, \kappa(\boldsymbol{x}, \boldsymbol{x}_N)]^{\mathrm{T}}$$

3.3.2 核支持向量机

核方法的思想可以拓展标准支持向量机并用于解决非线性分类问题。类似于标准支持向量机的求解步骤，首先将对偶问题式 (3.10) 改写成

$$\begin{aligned}\max_{\lambda} \quad & -\frac{1}{2}\sum_{t=1,k=1}^{N,N}\lambda_t\lambda_k y_t y_k \kappa(\boldsymbol{x}_t,\boldsymbol{x}_k)+\sum_{t=1}^{N}\lambda_t \\ \text{s.t.} \quad & \sum_{t=1}^{N}\lambda_t y_t = 0 \\ & \lambda_t \geqslant 0\end{aligned}$$

式中：$\kappa(\boldsymbol{x}_t,\boldsymbol{x}_k)$ 为选择的核函数。

如式 (3.8) 所示，标准支持向量机的权重向量依赖特征 $\boldsymbol{x}_t$。若采用转换后的高维特征向量 $\boldsymbol{\phi}(\boldsymbol{x}_t)$，则权重向量具有很高的维度，很难进行计算或存储。为此，在核支持向量机处理过程中，往往不显式表示最优权重向量，而是给出如下形式的分类器：

$$y=\operatorname{sgn}\left(\sum_{t=1}^{N}\lambda_t y_t \kappa(\boldsymbol{x}_t,\boldsymbol{x})+b\right)$$

类似于式 (3.12)，上式中偏置 b 的值由支持向量来获得。若 $\lambda_{t_0}\neq 0$，则 $\boldsymbol{x}_{t_0}$ 或者 $\phi(\boldsymbol{x}_{t_0})$ 为支持向量。为了避免对高维特征向量 $\phi(\boldsymbol{x}_{t_0})$ 直接处理，最优偏置通过如下核函数计算得到：

$$b^*=y_{t_0}-\sum_{t=1}^{N}\lambda_t y_t \kappa(\boldsymbol{x}_t,\boldsymbol{x}_{t_0})$$

由于核函数的强大能力，可以构造如下核函数来解决异或逻辑函数线性不可分的难题：

$$\kappa(\boldsymbol{x},\boldsymbol{z})=(1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z})^2=\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x})\phi(\boldsymbol{z})$$

该核函数所对应的特征变换函数为

$$\boldsymbol{\phi}(\boldsymbol{x})=[1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2]^{\mathrm{T}}$$

3.3.3 正则化理论

式 (3.22) 给出一个正则化优化问题的广义表示形式，其中包含误差项和正则化项。给定样本集合 $\{\boldsymbol{x}_t,y_t\}_{t=1}^N$，岭回归优化问题可以写成

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{t=1}^{N}\|y_t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t\|^2+\frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{3.23}$$