

第3章

支持向量机

3.1 背景知识

支持向量机(SVM)是机器学习中的一种经典二分类算法,其核心目标是通过构建一个最优超平面(或称决策边界),在特征空间中将数据进行分类。支持向量机起源于统计学习理论,尤其是结构风险最小化原则,旨在通过最大化类别间的间隔来提升模型的泛化能力,从而提高分类的准确性。通过将数据样本映射到高维特征空间,支持向量机能够将低维空间中线性不可分的数据转化为高维特征空间中线性可分的形式,进而实现更高的分类精度。

支持向量机的理论基础可以追溯到 1963 年苏联数学家 Vladimir Vapnik 和 Alexey Chervonenkis 在统计学习理论方面的研究。他们首次提出了支持向量方法,作为解决模式识别问题的一种有效手段,为后来的支持向量机模型的发展奠定了基础。1964 年,Vapnik 和 Chervonenkis 提出了硬间隔支持向量机(Hard Margin SVM),该方法能够有效地解决线性可分的二分类问题。

然而,现实数据往往不是线性可分的,因此,1992 年,Bernhard Boser、Isabelle Guyon 和 Vladimir Vapnik 提出了一种创新方法,即通过引入核技巧将训练数据映射到更高维的特征空间,从而在这个高维特征空间中实现数据的线性可分。这一方法开创了非线性分类器的研究,将支持向量机推广到了线性不可分的情境。

1995 年,Corinna Cortes 和 Vapnik 进一步提出了软间隔支持向量机(Soft Margin SVM),该方法允许部分数据点在分类时违反间隔约束,从而增强了支持向量机对噪声和异常数据的鲁棒性,极大地提升了其在实际数据集上的适用性。

随着数据量的不断增长,传统的支持向量机在大规模数据集上的计算效率逐渐成为瓶颈。为了解决这一问题,1998 年,John Platt 提出了序列最小优化(SMO)算法,这一算法能够高效地求解支持向量机中的二次规划问题,大大提高了支持向量机的计算速度和可扩展性。

作为一种强有力的分类工具,支持向量机已经在多个领域中得到了广泛应用。例如,在构建一个年龄识别系统时,原始数据(如皱纹程度、皮肤松弛度、毛发特征、嘴唇状态等)往往呈现复杂的分布。支持向量机通过构建最优超平面,能有效地将“青年”与“老年”样本区分开来。在实际应用中,单纯依据单一特征往往难以准确分类,因为存在“少白头”“保养得当”

等非典型样本。支持向量机通过将这些特征映射到高维特征空间,能够捕捉特征之间的非线性组合关系,从而实现高精度的年龄分层判定。此外,支持向量机在处理小样本人脸数据时仍能保持较好的鲁棒性,这使其成为年龄识别、人脸属性分析等场景下的核心算法之一。

随着计算能力的不断提升以及大数据时代的到来,支持向量机的应用范围不断扩展,并在实际问题中展现出了强大的表现,特别是在高维数据分类中的优势,使得支持向量机成为处理复杂数据问题的有力工具。然而,尽管支持向量机在多个领域取得了显著成就,依然存在一些挑战亟待解决。首先,选择合适的核函数是支持向量机算法应用中的一个关键问题,不同核函数的选择会显著影响分类效果。其次,在数据量较大时支持向量机的训练时间可能会急剧增加,因此如何提升支持向量机的计算效率,尤其是在大规模数据集上的应用,仍然是一个等待解决的难题。最后,如何处理多分类问题也是支持向量机的一大挑战,尽管已有多种策略,如一对一、一对多等方法,但在一些特殊应用中,这些方法仍存在局限性。

本章将从硬间隔支持向量机的基本概念入手,逐步引入软间隔方法和核技巧,扩展到一般的非线性分类问题。最后,将通过具体的应用实例展示支持向量机在实际问题中的应用,包括医学、文本、图像等领域。通过本章的学习,读者将能够深入理解支持向量机的理论基础与实际应用,并掌握如何在不同场景下选择合适的支持向量机模型与核函数,从而提升模型的分类性能(图 3.1)。

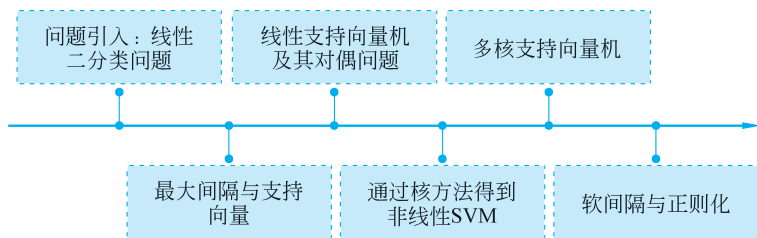


图 3.1 支持向量机章节框架图

3.1.1 问题引入

在许多实际应用中,如何从大量数据中发现潜在模式并进行有效分类是一个至关重要的问题。以基于人脸特征的年龄识别为例,从人脸图像中提取出能够区分不同年龄段的关键特征,如皱纹的深度、皮肤的松弛度以及毛发特征等,并通过合理的分类方法进行预测,往往是解决问题的关键。然而,传统的分类方法通常依赖于特定的假设,并且容易受到噪声和数据不平衡的影响,从而导致分类效果不理想。

针对这些问题,支持向量机提供了一套基于统计学习理论的解决方案。在正式推导支持向量机算法之前,首先回顾线性二分类问题的基本定义。

定义 3.1 假设有一个线性可分的训练数据集,包含 n 个样本,每个样本都是 d 维空间中的数据点,表示为 $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$, 其中, $\mathbf{x}_i$ 是样本特征, $y_i \in \{-1, 1\}$ 为样本类别标签。线性二分类问题的目标是找到一个分类超平面,将这两类数据点准确区分开。

线性决策面的形式如下:

$$\mathbf{w}^{(1)} \mathbf{x}_i^{(1)} + \mathbf{w}^{(2)} \mathbf{x}_i^{(2)} + \dots + \mathbf{w}^{(d)} \mathbf{x}_i^{(d)} + b = 0 \quad (3.1)$$

一般用向量形式表示为

$$\mathbf{w}^T \mathbf{x}_i + b = 0 \quad (3.2)$$

其中, $\mathbf{w} = (w^{(1)}; w^{(2)}; \dots; w^{(d)})$ 是决策面法向量; b 是偏置项。相应的分类决策函数可以表示为

$$f(\mathbf{x}_i) = \text{sgn}(\mathbf{w}^T \mathbf{x}_i + b) \quad (3.3)$$

对于线性可分的训练数据,通常存在无穷多个超平面能将两类数据**正确划分**,而支持向量机的目标是找出其中的最优超平面。该最优超平面能**最大化**不同类别样本间的距离,进而提高分类的准确性和泛化能力。

对于近似线性可分或线性不可分的场景,支持向量机的基本思想仍然一致。不同之处在于,支持向量机通过核函数实现非线性映射,将原始低维数据映射到高维特征空间,使得数据在该空间中线性可分,同时避免了直接计算高维特征的复杂度。此外,为了处理少量数据噪声,支持向量机引入了松弛变量,允许部分数据点在分类时偏离决策面,从而增强模型的鲁棒性和泛化能力。这使得支持向量机在面对噪声和数据不均衡时依然保持较好的分类性能。

3.1.2 间隔与支持向量

前文提到,支持向量机的目标是找到一个最优的线性决策面,其核心在于最大化分类间隔。为了理解这一点,我们需要明确判定分类超平面优劣的标准。

我们认为,一个好的分类超平面应该对局部扰动不敏感。具体来说,如果两个分类超平面都能正确分类样本,那么它们的优劣可以通过它们与最近样本点的距离来判断。距离越远,分类超平面对扰动的容忍度越高,表示分类超平面越“稳健”。以图 3.2 为例,图中圆圈半径代表了模型对样本位置扰动的容忍度。显然,右侧的分类超平面能够容忍更大的扰动。

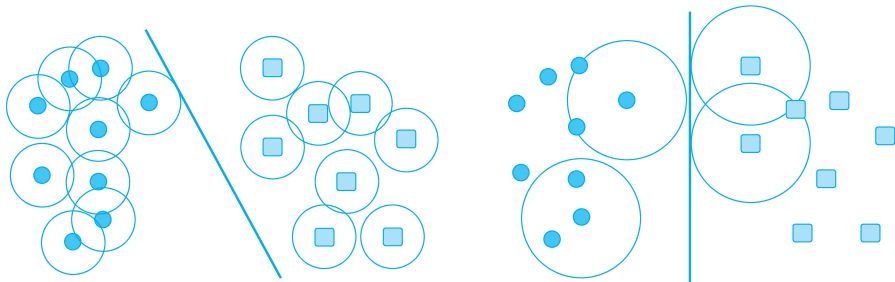


图 3.2 不同分类超平面对局部扰动的敏感程度示意图

在线性可分的情况下,随着圆圈半径增大,最后可以得到唯一的线性分类超平面,能容忍最大程度的扰动,这一线性分类超平面就是支持向量机算法寻找的最优分类超平面。

定义 3.2 如果一个超平面能够将训练样本没有错误地分开,并且两类训练样本中离超平面最近的样本与超平面之间的距离是最大的,则称该超平面为最优分类超平面。将两类训练样本中离最优分类超平面最近的样本称为支持向量,把两个异类支持向量到超平面的距离之和称为分类间隔,也称为硬间隔(图 3.3)。

支持向量机的名称正源于此。根据定义,最优分类超平面即具有最大分类间隔的决策面,而该间隔仅取决于支持向量。即使其他数据点移动,只要它们不转变为新的支持向量,

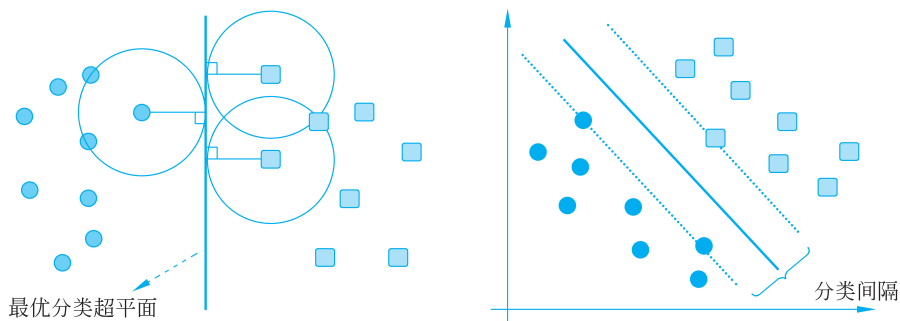


图 3.3 硬间隔示意图

最优超平面便保持不变。换言之,最优超平面具有**稀疏性**,仅由支持向量决定。注意到这与之前介绍的线性模型是完全不同的:线性模型平等考虑所有的样本,而支持向量机的分类边界仅由少数最难分类的样本(支持向量)决定,这正是其拥有良好泛化能力的关键所在。

3.2 支持向量机

3.2.1 理论推导

支持向量机以其强大的泛化能力和优雅的理论框架在机器学习领域占据了重要地位。前面的章节概览了支持向量机的核心思想,读者已经对支持向量机的基本理念及其在分类问题中的应用有了初步的了解。本节将在此基础上进一步展开,详细介绍支持向量机的理论推导过程和具体分类算法。支持向量机主要包含三种形式:硬间隔支持向量机、软间隔支持向量机以及核函数支持向量机,其复杂程度依次提升,分别适用于解决线性可分问题、近似线性可分问题以及线性不可分问题。下面将从最基础的硬间隔支持向量机入手,详细讲解其理论推导过程以及解决线性可分问题的具体步骤。通过这一过程,读者将深刻理解支持向量机的工作原理及其在模式识别领域中的价值。

硬间隔支持向量机专注于解决线性可分问题,其核心思想在于确定一个最优的决策边界,即分类超平面。该超平面不仅要正确地将不同类别的训练样本分开,还要保证样本与超平面之间的间隔达到最大。这种间隔最大化策略是支持向量机的一大特色,它为分类器提供了强大的泛化能力。从直观上理解,硬间隔支持向量机的目标仿佛是在寻找一条最宽阔的“街道”,其中央即为分类超平面,而“街道”的两侧则由支持向量所界定。这条“街道”的宽度反映了分类器的鲁棒性:宽度越大,分类器对样本局部扰动的容忍度越高。

从概念上理解,支持向量是那些恰好位于间隔边界上的样本点。它们对于确定最优分类超平面起到了决定性作用,因为它们是所有样本中距离超平面最近的点。移除这些支持向量将导致分类超平面的改变,因此,它们是模型构建中不可或缺的要素。最大间隔指的是分类超平面与其最近的支持向量之间的距离。在所有可能的分类超平面中,该距离最大,其目的在于增强分类器的泛化能力,使其在面对未知的测试样本时,能保持较低的泛化误差。最优分类超平面,即实现最大间隔的超平面,它不仅将两类样本完全分开,且与两侧支持向量保持等距。作为算法的求解目标,最优分类超平面能以极高的置信度对训练数据进行分类,并预期在处理新实例时保持良好的准确性。

综上所述,硬间隔支持向量机通过寻找最大化间隔的最优分类超平面,确保了模型卓越的泛化能力。在此过程中,支持向量的概念尤为关键。因此,如何在众多样本中找到关键的支持向量从而确定最优分类超平面,是本节的核心内容。接下来,将利用线性代数与凸优化理论,逐步推导出最优分类超平面的求解方法。

1. 基于间隔最大化的最优分类超平面

首先从一个线性可分的数据集开始,其中包含两类样本,分别用正标签和负标签表示:

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

其中, $\mathbf{x}_i \in \mathbb{R}^d$ 表示第 i 个样本的 d 维特征; $y_i \in \{+1, -1\}$ 表示第 i 个样本的标签, $y_i = +1$ 表示正标签样本, $y_i = -1$ 表示负标签样本。假设这些样本在特征空间中是线性可分的,意味着存在至少一个超平面,能够完全正确地将两类样本分开。超平面的一般形式可以表示为

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0 \quad (3.4)$$

为了找到最大化分类间隔的最优分类超平面,需要计算支持向量到最优分类超平面的距离。对于超平面 $\mathbf{w}^T \mathbf{x} + b = 0$, 首先求解其法向量。超平面的法向量满足: 对于超平面中任意向量都与该法向量垂直。任取超平面上的两个点 $\mathbf{x}_1, \mathbf{x}_2$, 分别满足

$$\mathbf{w}^T \mathbf{x}_1 + b = 0, \quad \mathbf{w}^T \mathbf{x}_2 + b = 0 \quad (3.5)$$

两式相减, 可得 $\mathbf{w}^T (\mathbf{x}_1 - \mathbf{x}_2) = 0$, 即超平面上任意一个向量都与 $\mathbf{w}$ 垂直, 故 $\mathbf{w}$ 即为超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的一个法向量。

对于超平面 $\mathbf{w}^T \mathbf{x} + b = 0$, 记超平面外一点为 $\mathbf{x}_0$, 其在超平面上的投影点为 $\mathbf{x}'_0$, 则有 $\mathbf{w}^T \mathbf{x}'_0 + b = 0$ 。且向量 $(\mathbf{x}_0 - \mathbf{x}'_0)$ 应与超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的法向量 $\mathbf{w}$ 平行, 即

$$\begin{aligned} |\mathbf{w}^T (\mathbf{x}_0 - \mathbf{x}'_0)| &= |\mathbf{w}^T \mathbf{x}_0 - \mathbf{w}^T \mathbf{x}'_0| = |\mathbf{w}^T \mathbf{x}_0 - (-b)| \\ &= |\mathbf{w}^T \mathbf{x}_0 + b| = |\mathbf{w}| * |(\mathbf{x}_0 - \mathbf{x}'_0)| \end{aligned} \quad (3.6)$$

则超平面外一点 $\mathbf{x}_0$ 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的距离为

$$d = |\mathbf{x}_0 - \mathbf{x}'_0| = \frac{|\mathbf{w}^T \mathbf{x}_0 + b|}{\|\mathbf{w}\|} \quad (3.7)$$

其中, $\|\mathbf{w}\|$ 表示向量 $\mathbf{w}$ 的模长。在欧氏空间中模长即为 L2 范数。

设 $\mathbf{x}_+$ 、 $\mathbf{x}_-$ 分别为正标签样本、负标签样本的支持向量, 则其到最优分类超平面的距离分别为

$$d_+ = \frac{|\mathbf{w}^T \mathbf{x}_+ + b|}{\|\mathbf{w}\|}, \quad d_- = \frac{|\mathbf{w}^T \mathbf{x}_- + b|}{\|\mathbf{w}\|} \quad (3.8)$$

考虑到最优超平面应满足正负标签样本的支持向量到超平面的距离相等, 则有 $d_+ = d_-$ 。因此, 正负标签样本的支持向量应分别满足

$$\mathbf{w}^T \mathbf{x}_+ + b = c, \quad \mathbf{w}^T \mathbf{x}_- + b = -c \quad (3.9)$$

且最优分类超面对应的最大分类间隔为

$$d = d_+ + d_- = \frac{|\mathbf{w}^T \mathbf{x}_+ + b|}{\|\mathbf{w}\|} + \frac{|\mathbf{w}^T \mathbf{x}_- + b|}{\|\mathbf{w}\|} = \frac{2c}{\|\mathbf{w}\|} \quad (3.10)$$

根据以上讨论, 对于线性可分数据, 可以找到一个使得分类间隔最大的最优超平面 $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0$, 且最大分类间隔为 $d = \frac{2c}{\|\mathbf{w}\|}$ 。然而, 尽管最优分类超平面和最大分类间隔唯一, 其所对应的参数 $(\mathbf{w}, b, c)$ 并不唯一。事实上, 如果将 $\mathbf{w}$ 、 b 、 c 等比例放缩为 $\alpha \mathbf{w}$ 、 αb 、

αc , 最优分类超平面本身并不会改变, 且上述最大分类间隔的计算也不会改变。最大分类间隔示意图见图 3.4。

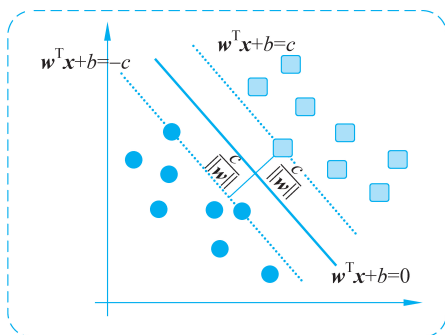


图 3.4 最大分类间隔示意图

推论 3.1 为了确定它们的尺度以保证参数的唯一性, 不妨固定 $c=1$, 则最优分类超面对应的分类间隔为

$$d = \frac{2}{\|w\|} \quad (3.11)$$

最终, 最大化分类间隔以找到最优分类超平面的目标可转换为求解以下问题: 找到 w 和 b , 使得超平面能够完全正确分类所有样本, 且对应的分类间隔 $\frac{2}{\|w\|}$ 最大。

2. 硬支持向量机的求解形式

为了将上述最大化分类间隔的问题转换为一个有约束的优化问题, 需要考虑以下两个条件。①分类正确性: 对于所有正标签样本, 有 $f(x) = w^T x + b \geq 1$; 对于所有负标签样本, 有 $f(x) = w^T x + b \leq -1$ 。②分类间隔最大化: 在保证所有样本分类正确的情况下, 需要最大化分类间隔 $\frac{2}{\|w\|}$, 这等价于最小化 $\frac{1}{2} \|w\|^2$ 。

定义 3.3 结合以上条件, 可以将支持向量机的原问题表述为一个有约束的优化问题:

$$\begin{aligned} \max_{w, b} \text{Margin}(w, b) &= \frac{2}{\|w\|} \\ \text{s.t. } y_i(w \cdot x_i + b) &\geq 1, \quad 1 \leq i \leq n \end{aligned} \quad (3.12)$$

该最大化问题可等价转换为一个最小化问题:

$$\begin{aligned} \min_{w, b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } y_i(w \cdot x_i + b) &\geq 1, \quad 1 \leq i \leq n \end{aligned} \quad (3.13)$$

其中, y_i 为样本 x_i 的标签, 取值为 +1 或 -1。

由于目标函数 $f(w, b) = \frac{1}{2} \|w\|^2$ 为凸函数, 则不等式约束 $g_i(w, b) = 1 - y_i(w^T x_i + b)$ 为凸函数, 该有约束优化问题可归纳为一般的凸优化问题求解:

$$\begin{aligned} \min_{w, b} f(w, b) \\ \text{s.t. } g_i(w, b) &\leq 0, \quad i = 1, \dots, k \\ h_i(w, b) &= 0, \quad i = 1, 2, \dots, l \end{aligned} \quad (3.14)$$

其中, $f(\mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2$; $g_i(\mathbf{w}, b) = 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)$; $h_i(\mathbf{w}, b)$ 不存在。

该凸优化问题的拉格朗日函数如下:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= f(\mathbf{w}, b) + \sum_{i=1}^k \alpha_i g_i(\mathbf{w}, b) + \sum_{j=1}^l \beta_j h_j(\mathbf{w}, b) \\ \text{s.t. } \alpha_i &\geq 0, \quad i = 1, 2, \dots, k \end{aligned} \quad (3.15)$$

其中, α_i 和 β_j 为拉格朗日乘子。

进一步考虑如下函数:

$$\theta(\mathbf{w}, b) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (3.16)$$

下面说明 $\theta(\mathbf{w}, b)$ 在满足约束的可行域内取值等于 $f(\mathbf{w}, b)$, 在可行域外取值为 $+\infty$ 。若 $(\mathbf{w}, b)$ 的某组取值不满足对 $g_i(\mathbf{w}, b)$ 或 $h_i(\mathbf{w}, b)$ 的约束条件, 即

$$\exists i, \text{ s.t. } g_i(\mathbf{w}, b) > 0 \quad \text{或} \quad h_i(\mathbf{w}, b) \neq 0$$

则可通过调整对应 α_i 或 β_i 的取值, 使得 $\theta(\mathbf{w}, b)$ 的取值满足

$$\theta(\mathbf{w}, b) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \left[f(\mathbf{w}, b) + \sum_{i=1}^k \alpha_i g_i(\mathbf{w}, b) + \sum_{j=1}^l \beta_j h_j(\mathbf{w}, b) \right] = +\infty \quad (3.17)$$

由此可知,

$$\theta(\mathbf{w}, b) = \begin{cases} f(\mathbf{w}, b), & (\mathbf{w}, b) \text{ 满足约束} \\ +\infty, & (\mathbf{w}, b) \text{ 不满足约束} \end{cases} \quad (3.18)$$

定义 3.4 综上, 仅需求解 $\min_{\mathbf{w}, b} \theta(\mathbf{w}, b) = \min_{\mathbf{w}, b} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 即可。因此支持向量机的原有约束优化问题可转化为如下极小极大问题求解:

$$\begin{aligned} \min_{\mathbf{w}, b} \theta(\mathbf{w}, b) &= \min_{\mathbf{w}, b} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \min_{\mathbf{w}, b} \max_{\boldsymbol{\alpha}; \alpha_i \geq 0} \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^k \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) \\ \text{s.t. } \alpha_i &\geq 0, i = 1, 2, \dots, k \end{aligned} \quad (3.19)$$

3.2.2 对偶问题

前面已经得到了支持向量机原问题的极小极大形式, 但该形式仍旧难以直接求解, 为了降低求解的难度、引入核函数实现非线性建模等目标, 进一步推导支持向量机原问题的对偶形式。

首先, 给出几个有关原问题和对偶问题的定义和定理。

定义 3.5 定义

$$\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (3.20)$$

再考虑极大化 $\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, 即

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (3.21)$$

则称上式为广义拉格朗日函数的极大极小问题, 亦即原问题的对偶问题。

定理 3.1 若原问题和对偶问题都有最优值, 则

$$d^* = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{w}} \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}; \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p^* \quad (3.22)$$

定理 3.2 考虑原问题(式(3.19))和对偶问题(式(3.21))。如果函数 $f(\mathbf{w})$ 和 $g_i(\mathbf{w})$ 是凸函数, $h_j(\mathbf{w})$ 是仿射函数; 并且不等式约束 $g_i(\mathbf{w})$ 是严格可行的, 即存在 $\mathbf{w}$, 对所有 i 有 $g_i(\mathbf{w}) < 0$, 则存在 $\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$, 使 $\mathbf{w}^*$ 是原问题的解, $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ 是对偶问题的解, 并且 $d^* = p^* = \mathcal{L}(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 。

定理 3.3 考虑原问题(式(3.19))和对偶问题(式(3.21))。如果函数 $f(\mathbf{w})$ 和 $g_i(\mathbf{w})$ 是凸函数, $h_j(\mathbf{w})$ 是仿射函数, 并且不等式约束 $g_i(\mathbf{w})$ 是严格可行的, 则 $\mathbf{w}^*$ 和 $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ 分别是原问题和对偶问题的解的充分必要条件是 $\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ 满足下面的 Karush-Kuhn-Tucker (KKT) 条件:

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = 0 \quad (3.23)$$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0, \quad i = 1, 2, \dots, k \quad (3.24)$$

$$g_i(\mathbf{w}^*) \leq 0, \quad i = 1, 2, \dots, k \quad (3.25)$$

$$\alpha_i^* \geq 0, \quad i = 1, 2, \dots, k \quad (3.26)$$

$$h_j(\mathbf{w}^*) = 0, \quad j = 1, 2, \dots, l \quad (3.27)$$

特别指出, 式(3.24)称为 KKT 的对偶互补条件。由此条件可知: 若 $\alpha_i^* > 0$, 则 $g_i(\mathbf{w}^*) = 0$ 。

根据定义 3.5, 可以写出支持向量机原问题的对偶问题:

$$\max_{\boldsymbol{\alpha}: \alpha_i \geq 0} \min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^n \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_{i=1}^n \alpha_i \quad (3.28)$$

不难验证, 对于线性可分的情况, 支持向量机的原问题和对偶问题满足定理 3.2 的条件, 即 $f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_2^2$ 和 $g_i(\mathbf{w}, b) = 1 - y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$ 为凸函数, 且 $g_i(\mathbf{w}, b) < 0$ 严格可行 (线性可分), 则存在 $\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*$, 使 $\mathbf{w}^*, b^*$ 是原问题的解, $\boldsymbol{\alpha}^*$ 是对偶问题的解, 并且 $d^* = p^* = \mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*)$, 因此可以通过求解对偶问题, 得到支持向量机原问题的解, 而对偶问题的求解往往更为简便。

为求解式(3.28)的对偶问题, 需要首先求解关于 $\mathbf{w}, b$ 的极小问题, 然后求解关于 $\boldsymbol{\alpha}$ 的极大问题, 步骤如下。

(1) 求解 $\min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$: 将拉格朗日函数 $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ 分别对 $\mathbf{w}, b$ 求偏导数, 并令所求导数等于零, 可以得到

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = 0$$

$$\nabla_b \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = - \sum_{i=1}^n \alpha_i y_i = 0$$

也即

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad (3.29)$$

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (3.30)$$

将式(3.29)、式(3.30)代入式(3.28)可得

$$\min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \quad (3.31)$$

(2) 求 $\max_{\boldsymbol{\alpha}: \alpha_i \geq 0} \min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$, 此时 $\min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ 可根据式(3.31)用 $\boldsymbol{\alpha}$ 表示:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} & -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \\ \text{s.t.} & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (3.32)$$

式(3.32)即为支持向量机原问题的对偶问题。

假设已经解得对偶问题的解 $\boldsymbol{\alpha}^*$, 如何据此得到原问题的解呢? 这时需要借助 KKT 条件(定理 3.3), 它描述了原问题和对偶问题解之间的关系。将支持向量机的原问题和对偶问题代入定理 3.3 可得

$$\begin{aligned} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*) &= \mathbf{w}^* - \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = 0 \\ \nabla_b \mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*) &= -\sum_{i=1}^n \alpha_i^* y_i = 0 \\ \alpha_i^* (1 - y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*)) &= 0, \quad i = 1, 2, \dots, n \\ 1 - y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) &\leq 0, \quad i = 1, 2, \dots, n \\ \alpha_i^* &\geq 0, i = 1, 2, \dots, n \end{aligned} \quad (3.33)$$

由式(3.33)可知, 支持向量机原问题的解 $\mathbf{w}^*$ 满足

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \quad (3.34)$$

其中, 至少有一个 $\alpha_i^* > 0$, 否则 $\mathbf{w}^* = \mathbf{0}$ 不可能满足不等式约束。记 $\mathcal{S} = \{i \mid \alpha_i > 0, i = 1, 2, \dots, n\}$ 表示大于 0 的 α_i 的索引的集合, 则对于任意的 $i \in \mathcal{S}$ 有 $1 - y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) = 0$, 因此 $b^* = y_i - \mathbf{w}^* \cdot \mathbf{x}_i$, 考虑 $\mathcal{S}$ 中的所有元素, 并代入式(3.34)可得

$$\begin{aligned} \mathbf{w}^* &= \sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i \\ b^* &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left(y_i - \sum_{j \in \mathcal{S}} \alpha_j^* y_j \mathbf{x}_j \cdot \mathbf{x}_i \right) \end{aligned} \quad (3.35)$$

至此, 从对偶问题的解 $\boldsymbol{\alpha}^*$ 求得了支持向量机原问题的解 $\mathbf{w}^*, b^*$, 可进一步得到分类超平面和分类决策函数:

$$\begin{aligned} \sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^* &= 0 \\ \hat{y} &= \text{sgn} \left(\sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^* \right) \end{aligned} \quad (3.36)$$

综上所述, 针对线性可分情况, 支持向量机的对偶解法可总结如下。

算法 3.1 线性可分支持向量机的对偶算法

输入: 线性可分的训练集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, 其中, $\mathbf{x}$ 为特征向量, y 为二值标签, 取值为 -1 或 1。

输出：支持向量机的分类超平面和分类决策函数。

(1) 求解支持向量机的对偶优化问题(式(3.32)),得到最优解 α^* 和集合 S ;

(2) 根据式(3.35)和式(3.36)即可计算出分类超平面和分类决策函数。

由式(3.35)和式(3.36)可知,支持向量机最终的参数和分类超平面仅取决于 $\alpha_i^* > 0$ 的样本点。由 KKT 条件中的对偶互补条件可知,这些样本点必然满足 $1 = y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*)$,即这些样本点都位于决策边界上,如图 3.5 所示。我们称 $\alpha_i^* > 0$ 的样本点为支持向量,支持向量机的分类超平面仅取决于支持向量。

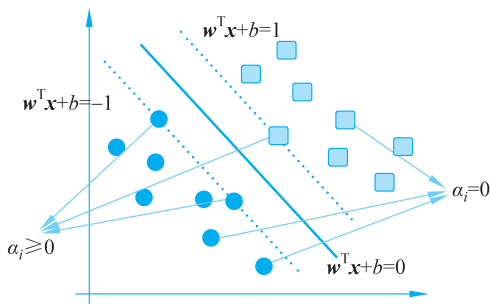


图 3.5 最大分类间隔示意图

例题 3.1 假设我们要构建一个简单的线性支持向量机模型来区分青年(正例)和老年(负例)。我们提取了人脸的两维核心特征(经标准化处理),第一维度代表皱纹特征指数(数值越大表示特征越显著),第二维度代表皮肤松弛度指数。给定青年样本(正例, $y=1$) $\mathbf{x}_1 = (1, 1)$, $\mathbf{x}_2 = (-1, 3)$, 老年样本(负例, $y=-1$) $\mathbf{x}_3 = (2, 3)$, $\mathbf{x}_4 = (2, 4)$, 计算支持向量、分类决策函数和分类超平面方程。

参考解法: 利用线性可分支持向量机的对偶算法。

(1) 首先写出对偶形式:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^4 \alpha_i - \frac{1}{2} \sum_{i=1}^4 \sum_{j=1}^4 y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{s.t.} \quad & \alpha_1 + \alpha_2 = \alpha_3 + \alpha_4 \\ & \alpha_i \geq 0, \quad i = 1, 2, 3, 4 \end{aligned}$$

(2) 将 $\alpha_4 = \alpha_1 + \alpha_2 - \alpha_3$ 代入优化目标函数得到

$$\begin{aligned} \max_{\alpha} \quad & 2\alpha_1 + 2\alpha_2 - \frac{1}{2}(10\alpha_1^2 + 10\alpha_2^2 + \alpha_3^2 + 12\alpha_1\alpha_2 - 6\alpha_1\alpha_3 - 2\alpha_2\alpha_3) \\ \text{s.t.} \quad & \alpha_1 \geq 0, \quad \alpha_2 \geq 0, \quad \alpha_3 \geq 0 \end{aligned}$$

(3) 将目标函数对 $\alpha_1, \alpha_2, \alpha_3$ 分别求偏导,令偏导数为 0,得到

$$\begin{aligned} 2 - 10\alpha_1 - 6\alpha_2 + 3\alpha_3 &= 0 \\ 2 - 10\alpha_2 - 6\alpha_1 + \alpha_3 &= 0 \\ -\alpha_3 + 3\alpha_1 + \alpha_2 &= 0 \end{aligned}$$

该方程组无解,说明最大值在边界处取得,即 $\alpha_i, i=1, 2, 3, 4$ 中存在 0。

当 $\alpha_1 = 0$ 时,解得 $\alpha_2 = \alpha_3 = \frac{2}{9}, \alpha_4 = 0$,代入可得 $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = -\frac{4}{9}$;

当 $\alpha_2 = 0$ 时,解得 $\alpha_4 = -4 < 0$,舍去;

当 $\alpha_3=0$ 时,解得 $\alpha_1=\alpha_2=\frac{1}{8}, \alpha_4=\frac{1}{4}$,代入可得 $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})=\frac{1}{4}$;

当 $\alpha_4=0$ 时,解得 $\alpha_1=\frac{1}{3}, \alpha_2=\frac{1}{9}, \alpha_3=\frac{4}{9}$,代入可得 $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})=\frac{4}{9}$;

由此可知,当 $\boldsymbol{\alpha}^*=\left[\frac{1}{3}, \frac{1}{9}, \frac{4}{9}, 0\right]$ 时 $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ 取得最大值。

(4) 将 $\boldsymbol{\alpha}^*$ 代入式(3.36)可得超平面方程和分类决策函数。

3.3 方法扩展

3.3.1 核函数

在前面的章节中,详细讨论了在训练数据样本线性可分的情况下硬间隔支持向量机的理论推导过程和求解方法。通过最大化分类间隔,构建了能够正确分类两类数据样本的最优超平面。然而,现实中的数据分布往往复杂多变,许多情况下样本在原始特征空间中呈现非线性分布,无法通过单一的线性超平面进行有效分割。例如,在基于人脸特征的年龄识别任务中,需要根据皱纹程度、皮肤松弛度、眼部老化程度、下颌线清晰度、毛发特征以及嘴角下垂等连续数值特征,将样本年龄层划分为“青年”或“老年”类别。由于面部衰老是一个复杂的生物过程,这些特征与年龄类别之间往往存在高阶的非线性关系。例如,单纯依靠“皱纹程度”的线性阈值可能无法简单区分,而某些脸部特征的组合(如“下颌线模糊”叠加“眼部重度老化”)才是判定老年的关键,这种复杂的特征交互导致两类样本在原始特征空间中相互交织,难以用一个平直的决策边界分开。此时,训练样本数据在原始特征空间中无法通过单一线性超平面进行有效分类,硬间隔支持向量机不再适用。针对此类问题,本节将介绍支持向量机的核函数方法,该方法通过隐式的高维特征空间映射,使得支持向量机能够处理复杂的非线性分类问题,同时有效避免了高维计算带来的“维数灾难”。

为了使表述更加严谨,对上述线性不可分问题作如下定义。

定义 3.6 假设训练样本数据的原始特征空间为 $\mathcal{X} \subseteq \mathbb{R}^d$, 给定训练集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, 其中 $y_i \in \{-1, 1\}$ 为样本标签。若不存在超平面 $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0$, 使得所有样本满足 $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$, 则称该问题为线性不可分。

此时,解决线性不可分问题的直观思路是将数据映射到更高维的特征空间,使得所有训练样本在该高维特征空间中线性可分。但是对于任意线性不可分问题,是否都存在这样的映射? Cover 定理为这一思想提供了理论支撑。Cover 定理指出,将非线性可分数据映射到足够高维的空间后,数据线性可分的概率显著增加。例如,二维平面上簇环形分布数据,在原始二维空间中无法用直线进行分割,但若将其映射到三维空间,通过引入 $x_1^2, x_2^2, x_1 x_2$ 等二次项特征,可能被一个平面完美分割(图 3.6)。这一理论确立了核函数方法的可行性。

具体而言,假设存在一个非线性映射函数 $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^D$ (其中 $D \gg d$), 将原始特征 $\mathbf{x}$ 转换为高维特征向量 $\phi(\mathbf{x})$, 使得训练样本在该高维特征空间中线性可分。那么只需在该高维特征空间中应用前述硬间隔支持向量机理论进行求解即可。在该高维特征空间中,分类超平面可表示为

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b = 0 \quad (3.37)$$

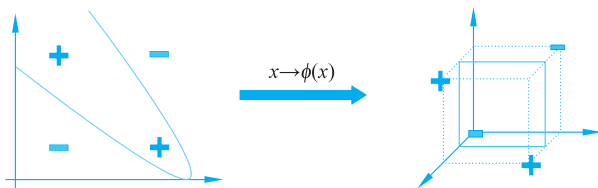


图 3.6 高维特征空间映射示意图

其中, $\mathbf{w}$ 为高维特征空间中的法向量。则支持向量机的原有约束优化问题在形式上与硬间隔支持向量机仍然保持一致:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1, \quad 1 \leq i \leq n \end{aligned} \quad (3.38)$$

其对应的对偶问题为

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) + \sum_{i=1}^n \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, n \end{aligned} \quad (3.39)$$

在上述过程中, 映射函数 $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^D$ 被称为基函数(图 3.7), 其将训练样本 $\mathbf{x}$ 从原始数据空间 $\mathbb{R}^d$ 映射到高维特征空间 $\mathbb{R}^D$, 从而可直接应用硬间隔支持向量机理论进行求解。然而, 在支持向量机的求解过程中将会涉及大量样本特征的内积运算, 而直接在原始高维特征空间中计算内积将会带来计算量的急剧增长。例如, 对于三维原始数据 $(x_1, x_2, x_3) \in \mathbb{R}^3$, 若采用二次多项式基函数映射 $(x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3, x_1^2, x_2^2, x_3^2)$, 特征维度将从 3 扩展至 9。对于三维原始数据, 在原始特征空间中计算内积时只涉及 3 次加乘运算; 而在二次多项式基函数空间中计算内积时需要 9 次加乘运算。实际上, 对于 d 维原始数据 $\mathbf{x} \in \mathbb{R}^d$, 其对应的 p 次多项式基函数的维度为 $\sum_{i=1}^p \binom{i+d-1}{i}$ (有

放回抽样的组合数问题)。当 d 或 p 较大时, 多项式基函数将引发维度灾难。例如, 当原始维度 $d=100$ 且多项式次数 $p=5$ 时, 映射后的维度将超过 10^7 , 在高维特征空间中显式计算样本内积的复杂度将变得完全不可接受。

然而在实际应用中, 数据分布往往复杂多变。例如在上述人脸年龄识别任务中, 为了将青年和老年样本分开, 可能需要引入皱纹程度与皮肤松弛度的高阶乘积项作为新特征, 这使得样本必须映射到超高维特征空间中才能实现线性可分, 从而导致基函数映射带来维度灾难问题。既然高维映射能够解决线性不可分问题, 但计算复杂度又如此之高, 那么是否存在一种方法能够绕过显式的高维计算? 支持向量机的对偶形式提供了一条关键线索。让我们重新回顾一下原始线性可分问题下硬间隔支持向量机的求解过程。线性可分问题下支持向量机的对偶问题形式为

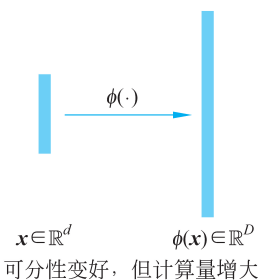


图 3.7 基函数示意图

$$\begin{aligned}
& \max_{\alpha} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^n \alpha_i \\
& \text{s.t. } \sum_i \alpha_i y_i = 0 \\
& \alpha_i \geq 0, \quad i = 1, 2, \dots, n
\end{aligned} \tag{3.40}$$

对其求解得到的分类超平面和分类判别函数为

$$\begin{aligned}
& \sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + b^* = 0 \\
& \hat{y} = \text{sgn} \left(\sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x} + b^* \right)
\end{aligned} \tag{3.41}$$

观察可以发现,支持向量机的求解仅依赖样本之间的内积运算 $\mathbf{x}_i^T \mathbf{x}_j$,而无须用到样本具体的特征值。这一观察提示我们,在使用基函数映射到高维特征空间中求解支持向量机时,其对应的优化问题也仅依赖映射到高维特征空间后的内积运算 $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ 。这一观察就催生了核函数(Kernel Function)的核心思想:若能直接在原始特征空间中计算样本在高维特征空间中的内积,即可绕过显式映射的复杂度,降低计算量。现在,给出核函数的定义。

定义 3.7 若存在一个映射 $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^D$, 对于原始特征空间中的任意样本 $\mathbf{x}_i, \mathbf{x}_j$, 其映射到高维特征空间后特征向量的内积为

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

则称 $K(\mathbf{x}_i, \mathbf{x}_j)$ 为核函数。

通过核函数,高维特征空间中的内积计算被转换为原始特征空间中的低维运算。例如,对于二维数据 $\mathbf{x} = (x_1, x_2)$, 若采用二次多项式映射:

$$\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

则其内积可简化为

$$\begin{aligned}
\phi(\mathbf{x})^T \phi(\mathbf{y}) &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T (y_1^2, \sqrt{2}y_1y_2, y_2^2) \\
&= x_1^2 y_1^2 + 2x_1x_2 y_1 y_2 + x_2^2 y_2^2 \\
&= (x_1 y_1 + x_2 y_2)^2 \\
&= ((x_1, x_2)^T (y_1, y_2))^2 \\
&= (\mathbf{x}^T \mathbf{y})^2 \\
&=: K(\mathbf{x}, \mathbf{y})
\end{aligned}$$

通过核函数 $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y})^2$, 能够在原始特征空间(2维)计算高维特征空间(3维)的内积(图 3.8)。现在比较一下采用核函数方法时带来的计算量优势。若采用显式计算,先将数据从原始特征空间映射到高维特征空间再进行内积计算,则需要 9 次乘法运算。而采用核函数方法,仅需 1 次原始内积和 1 次平方计算,计算复杂度从 $O(d^2)$ 降至 $O(d)$ 。因此,核函数方法能够有效解决基函数映射中高维特征空间内积运算导致的计算量急剧增长的瓶颈。

在将核函数引入支持向量机的求解过程中后,其对偶问题为

$$\max_{\alpha} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n \alpha_i$$

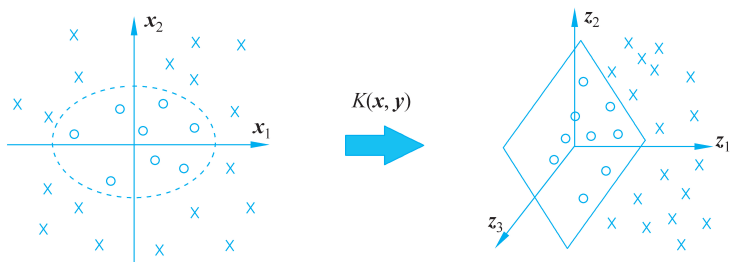


图 3.8 核函数示意图

$$\begin{aligned} \text{s.t. } & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, n \end{aligned} \quad (3.42)$$

求解该问题后可以得到,分类决策函数仅依赖于核函数的选择和支持向量:

$$\begin{aligned} \max_{\alpha} & -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n \alpha_i \\ \text{s.t. } & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, n \\ \hat{y} & = \text{sgn}(\mathbf{w}^T \phi(\mathbf{x}) + b) \\ & = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i (\phi(\mathbf{x}_i)^T \phi(\mathbf{x})) + b\right) \\ & = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \end{aligned} \quad (3.43)$$

其中,偏置项 b 可通过任意支持向量 $\mathbf{x}_k$ (满足 $\alpha_k > 0$) 计算:

$$b = y_k - \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_k) \quad (3.44)$$

可以发现,求解过程和判别函数均只涉及核函数的计算,而无须对基函数在高维特征空间中进行计算,从而验证了核函数方法能够有效解决基函数带来的维度灾难问题。值得一提的是,在应用核函数的支持向量机求解中,最优超平面仍然仅由支持向量(即分类边界上的样本)决定,其他样本对分类决策面没有影响。这与线性可分情况下的线性支持向量机的性质是一样的,唯一区别仅在于使用核函数将其扩展至非线性空间。

例题 3.2 现有一个简化的人脸年龄识别任务。我们提取了人脸的两项关键特征: $\mathbf{x}_1$ 表示“皱纹程度”(数值 1 表示皱纹程度高, 0 表示皱纹程度中等, -1 表示皱纹程度低), $\mathbf{x}_2$ 表示“下颌线模糊度”(数值 1 表示模糊, 0 表示一般, -1 表示清晰)。根据专家标注,我们将样本分为两类:“老年”为正例,“青年”为负例。现有 4 个训练样本,如图 3.9 所示,其特征数值可表示为:正例 $\mathbf{x}^1 = (1, 1)$, $\mathbf{x}^2 = (1, 0)$, 负例 $\mathbf{x}^3 = (0, 0)$, $\mathbf{x}^4 = (-1, 1)$ 。使用基函数 $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, 求解对应的支持向量、分类决策函数和分类超平面方程。

参考解法: 利用基于核函数的支持向量机的对偶算法。

样本编号	皱纹程度	下颌线清晰度	年龄层
1	高	模糊	老年
2	高	一般	老年
3	中	一般	青年
4	低	模糊	青年

图 3.9 核函数支持向量机例题

(1) 首先计算基函数 $\phi(\mathbf{x})$ 对应的核函数:

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = (\mathbf{x}^T \mathbf{y})^2$$

(2) 逐项计算 $K(\mathbf{x}_i, \mathbf{x}_j)$, 得到核矩阵:

$$\mathbf{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 4 \end{bmatrix}$$

(3) 写出对偶问题形式:

$$\max_{\alpha} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n \alpha_i$$

$$\text{s.t. } \sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \quad i = 1, 2, \dots, n$$

(4) 代入核函数值、样本标签以及 $\alpha_4 = \alpha_1 + \alpha_2 - \alpha_3$, 可得目标函数:

$$S(\alpha_1, \alpha_2, \alpha_3) = 2\alpha_1 + 2\alpha_2 - \frac{1}{2}(8\alpha_1^2 + 3\alpha_2^2 + 4\alpha_3^2 + 8\alpha_1\alpha_2 - 8\alpha_1\alpha_3 - 6\alpha_2\alpha_3)$$

$$\text{s.t. } \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0$$

(5) 将目标函数对 $\alpha_1, \alpha_2, \alpha_3$ 分别求偏导, 令偏导数为 0, 得到

$$2 - 8\alpha_1 - 4\alpha_2 + 4\alpha_3 = 0$$

$$2 - 3\alpha_2 - 4\alpha_1 + 3\alpha_3 = 0$$

$$4\alpha_1 + 3\alpha_2 - 4\alpha_3 = 0$$

该方程组无解, 说明最大值在边界处取得, 即 $\alpha_i, i=1, 2, 3, 4$ 中存在 0。

当 $\alpha_1 = 0$ 时, 解得 $\alpha_2 = \frac{8}{3}, \alpha_3 = 2, \alpha_4 = \frac{2}{3}$, 代入可得 $S = \frac{8}{3}$;

当 $\alpha_2 = 0$ 时, 解得 $\alpha_1 = \alpha_3 = \frac{1}{2}, \alpha_4 = 0$, 代入可得 $S = \frac{1}{2}$;

当 $\alpha_3 = 0$ 时, 解得 $\alpha_1 = -\frac{1}{4}$, 舍去;

当 $\alpha_4 = 0$ 时, 解得 $\alpha_1 = 0, \alpha_2 = \alpha_3 = 2$, 代入可得 $S = 2$;

由此可知, 当 $\alpha^* = \left[0, \frac{8}{3}, 2, \frac{2}{3}\right]$ 时目标函数取得最大值。

(6) 将 α^* 代入式(3.43)可得超平面方程和分类决策函数:

$$\mathbf{w}^* = \sum_{i=2}^4 \alpha_i y_i \phi(\mathbf{x}_i) = \left(2, \frac{2\sqrt{2}}{3}, -\frac{2}{3}\right)$$

$$b^* = -1$$

在支持向量机理论中,核函数方法是解决非线性可分问题的关键技术。该方法的独特之处在于,它不仅能处理复杂的非线性关系,还通过“核技巧”有效避免了高维特征空间带来的计算复杂性。在求解支持向量机时,我们仅需计算核函数,而无须显式定义将原始数据映射到高维特征空间的基函数形式,从而简化了求解过程。具体来说,核函数的基本思想是通过隐式映射,将原始特征空间中非线性可分的数据转换为高维特征空间中的线性可分数据。核心在于,这一映射是由核函数直接完成的。这意味着,我们无须知道基函数 $\phi(\mathbf{x})$ 的具体形式。核函数定义为 $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$, 即直接在原始特征空间中计算高维特征空间的内积,既保持了计算效率,又避免了显式的高维特征提取。这种隐式映射机制,使模型在避免“维数灾难”的同时,有效地捕捉了数据的潜在模式。

然而,不显式地定义基函数的具体形式也会带来问题:若没有定义明确的基函数形式而直接定义对应的核函数形式,那么该如何确定核函数的有效性呢?即如何确认该核函数所对应的基函数一定存在呢?该问题等价于:如何判断一个函数是否为某个特征空间下的内积运算? Mercer 定理为此提供了理论依据。Mercer 定理指出:

定理 3.4 一个对输入 $\mathbf{x}, \mathbf{y}$ 对称的函数 $K(\mathbf{x}, \mathbf{y})$ 能够表示成某个函数 $\phi(\mathbf{x})$ 的内积形式,当且仅当 $K(\mathbf{x}, \mathbf{y})$ 是半正定的:

$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0, \quad \forall g$$

或者也可以表述成等价形式:

$$\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}$$

对任意一组 $(x_1, x_2, \dots, x_n)$ 是半正定的。

Mercer 定理保证了只要一个对称函数满足半正定性,它就符合作为核函数的条件。这为核函数的构造提供了理论保证。在实际应用中,只要函数满足上述性质,就可以将其作为核函数,这极大地扩展了核函数的适用范围和设计灵活性。

然而,在选择核函数时,仅考虑有效性是不够的,还需要考虑其对应的特征空间能否实现样本的线性可分。使用不同的核函数,样本在高维特征空间的分布形式截然不同,这对于支持向量机的分类效果有显著影响。例如,若核函数选择过于简单,使得样本在高维特征空间仍无法线性可分,则支持向量机将无法有效分类;若核函数与数据分布不匹配,即使线性可分,最大分类间隔也可能显著缩小,从而降低模型的泛化性能。因此,在实际应用中,选择合适的核函数至关重要。常用的核函数主要包括以下几类。

首先,线性核函数是最为简单直接的形式,描述为

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y} \quad (3.45)$$

其常用于已经线性可分的数据集,或当特征数非常大而样本数较小时也非常有效。例如,如果使用“皱纹数量”“皮肤松弛程度”“眼部老化程度”等特征进行粗略划分,设定一个阈值就能将少量的青年和老年样本分开。此时线性核是有效的,不需要多余的转换步骤,并能够有效降低算法的计算复杂度。

相比之下,多项式核函数的形式更为灵活且复杂:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^d \quad (3.46)$$

这一核函数通过引入多项式特征来描述数据,可以捕获到数据间更加复杂的关系。这使得多项式核函数在数据分布呈现一定复杂性时,尤其是数据中包含显著多项式关系的任务时尤为有效。在人脸年龄识别中,特征之间往往存在“交互效应”。例如,“嘴角下垂”本身可能不代表老年,但如果“嘴角下垂”与“严重的皮肤松弛”同时出现(即特征的乘积项),则极大概率指向老年类别。多项式核函数通过将特征映射到高维特征空间,能够自动生成这些特征的乘积组合,从而捕捉到这种复杂的衰老模式。然而,选择适当的参数 c 和多项式的次数 d 至关重要,因为这两个参数会直接影响多项式核函数的灵活性和运算的复杂度。

径向基函数核(RBF 核),或者称为高斯核,是另一种广泛使用的核函数,其具体形式为

$$K(\mathbf{x}, \mathbf{y}) = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{\sigma^2}\right\} \quad (3.47)$$

RBF 核的优势在于其能够在特征空间中形成紧凑的局部区域,这种特性使得模型能够适应多样化的数据分布。在特征空间中,RBF 核通过高斯函数的形式,将距离近的数据点关联在一起,这增强了模型处理复杂模式的能力。对于人脸年龄识别而言,RBF 核非常直观:它基于样本间的“相似度”。“青年”样本通常在特征空间中聚集在一起(皮肤紧致、无皱纹、下颌清晰),形成一个紧密的簇。RBF 核能够很好地圈定这个“青年特征簇”的范围,将任何偏离该中心的样本(如出现皮肤松弛或有皱纹)划分为老年。特别是在面对高度非线性或只有复杂多项式关系才能描述的数据边界时,RBF 核表现优异。传统的线性核或者简单的多项式核在面对此类数据时往往难以奏效,因为它们无法有效地捕捉复杂的非线性结构。而 RBF 核具有的高维非线性映射能力,使其可以有效地在原始特征空间中进行复杂的划分,将原本难以识别的非线性边界清晰化。

然而,达到这种优异性能的关键在于如何设置核函数的 σ 参数,其决定了高斯函数的宽度,即数据点对分类超平面的影响范围。具体来说,较小的 σ 值意味着每个数据点的影响范围较小,模型仅对邻近的点有较强响应,这容易使决策边界变得过于复杂和破碎,从而在训练数据上表现很好但在测试数据上表现不佳,导致过拟合。另外,较大的 σ 值意味着数据点影响范围过大,决策边界趋于平滑甚至近似线性,可能无法充分捕捉数据的分布特征,从而导致欠拟合。因此,设定合适的 σ 值是 RBF 核在实践中的关键环节。在实际应用中,调优 σ 参数往往需要结合交叉验证等技术。通过在不同参数设置下评估模型的预测能力,选取在验证集上表现最优的参数值。此外,网格搜索、随机搜索以及贝叶斯优化等自动化算法也能有效帮助探索合适的参数空间,从而提升模型性能。

另一种值得介绍的是 Sigmoid 核(双曲正切核),其形式为

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\alpha \mathbf{x}^T \mathbf{y} + r) \quad (3.48)$$

Sigmoid 核是众多核函数中极具独特性的一种核函数,源自神经网络理论。其设计灵感来自神经元的激活过程,模拟了类似神经网络中 Sigmoid 激活函数的效果。正如激活函数在神经网络中负责引入非线性特性,Sigmoid 核在支持向量机中也旨在捕捉数据中的非线性关系。在人脸年龄识别场景下,可以将衰老看作一种“生理阈值”现象:当皱纹程度、眼部老化和皮肤松弛度累积超过某个临界点时,人脸状态会从“青年”激活突变为“老年”。Sigmoid 核能够模拟这种基于阈值的非线性转换机制。对于某些具有特定层次结构或非线性关联的数据集,它能够模拟复杂的数据分布并优化模型性能。在处理具有双态分布或高

非线性度的等问题时,Sigmoid 核赋予了支持向量机更高的灵活性。然而,Sigmoid 核并非通用的解决方案,其实际效果在很大程度上依赖于超参数 α 和 r 的设定。超参数 α 用于控制输入空间中数据点的尺度,其值的大小直接影响模型的灵敏度: α 值如果过大,模型可能变得过于敏感而导致过拟合,无法在未知数据上进行有效预测;如果过小,模型可能过于平滑而无法识别数据中的复杂细节特征。超参数 r 则用于控制 Sigmoid 函数的偏移量,直接影响决策边界的截距。设定正确的 r 值可以有效地校正模型输出,确保其能准确地反映输入数据的分布特点。由于 Sigmoid 核在这些参数方面的敏感性和不确定性,在实践中通常需要通过多次实验交叉验证来寻找参数的最佳配置。针对不同的数据集,采用网格搜索或者随机优化来调整 α 和 r ,从中选出模型在验证集上表现最佳的参数组合。除了参数调优带来的挑战外,Sigmoid 核在不同特征空间的稳定性和准确性可能也会受到其他因素的影响。尤其是在涉及复杂非线性边界的分类任务时,Sigmoid 核的表达力可能并不如 RBF 核那样强大。因此,Sigmoid 核常常作为一种试验性的选择,需要综合考虑数据的具体特性以及模型的泛化需求。尽管 Sigmoid 核在处理复杂非线性边界时可能不如 RBF 核稳健,但作为一种模拟神经网络机制的方法,它在处理特定模式的数据时依然具有重要的应用价值。

我们介绍了几类常用核函数,但选择合适的核函数不仅取决于数学形式,更需考虑计算效率、数据特性及调参需求等。首先是计算效率:线性核计算复杂度低,适合大规模稀疏数据;而 RBF 核和多项式核计算开销较大,需谨慎选择。其次是数据特性:对于线性关系明显的问题,线性核足矣;而对于局部性强或非线性复杂的问题,RBF 核通常效果更佳。此外,参数调优也是关键,利用交叉验证等数据驱动方法可辅助确定最优核函数。领域知识同样重要,例如在已知面部特征存在局部噪声(如光照影响导致的皮肤纹理误差)的年龄识别任务中,RBF 核往往是一个很好的起点。

尽管核函数方法为支持向量机提供了强大的非线性处理能力,但也存在局限性。显著的局限在于其“黑箱”特性:隐式特征映射使得模型的高维决策过程难以直观解读,降低了模型的可解释性。在人脸年龄识别的实际应用中,例如美容咨询或衰老分析,用户可能希望知道具体是“下颌线模糊”还是“眼角皱纹”导致了“老年”的判定。然而,核函数将这些特征映射到了无法可视化的超高维特征空间,使得我们很难给出一个直观的归因解释,这在对透明度要求极高的领域可能受到限制。此外,核矩阵的计算与存储对计算资源的需求较高,在大规模数据集上,训练时间和内存消耗仍是不可忽视的挑战。

总体来说,核函数方法赋予了支持向量机处理复杂非线性数据的能力。通过合理选择核函数及适度调优参数,支持向量机能够在模式识别任务中获得优异的表现。这种方法虽然有一定的局限性,但随着硬件不断发展和优化算法的进步,核函数方法无疑将在更多领域和应用中发挥更大的价值。

3.3.2 软间隔与正则化

在前面的讨论中,始终假设训练样本无论是在原始数据空间还是在经过核变换后的高维特征空间中都是线性可分的。然而,现实世界中的数据集往往不那么理想,导致这一假设在实际应用中往往难以满足。其中一个关键问题在于,很难确定合适的核函数以确保训练样本在特征空间中实现线性可分。通常情况下,不同的核函数会给样本分配不同的非线性变换,从而影响分类效果。然而,目前尚无通用的策略可以保证在所有情境中选择到最优的

核函数。此外,另一个常见且棘手的问题是离群值的处理。离群值是那些与大多数数据点明显不同的样本点,通常由测量误差或异常数据产生。在硬间隔支持向量机中,这些离群值会导致分类超平面出现过拟合现象。过拟合意味着模型过于贴合训练数据,从而失去对新数据的泛化能力。在这种情况下,模型的预测表现会变得不稳定,因为离群值可能严重扭曲最优分类超平面的位置(见图 3.10)。

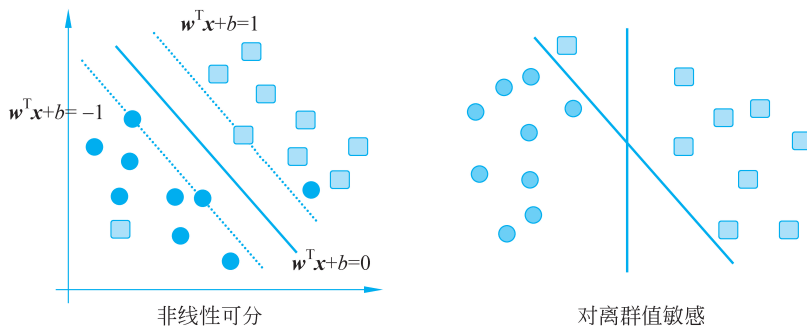


图 3.10 硬间隔支持向量机可能存在的问题

这些问题源于硬间隔支持向量机严格要求对所有样本必须分类正确。在硬间隔框架下,模型试图找到一个能够完美分类所有样本的决策边界。然而,这种严格的约束可能导致分类器在面对离群值或极端样本时,决策边界发生剧烈偏移。离群值的存在削弱了最优分类超平面的稳定性与泛化性,因为模型的决策边界会倾向于迁就这些噪声点,从而导致模型在处理非理想数据集时表现不佳。

为了解决上述难题,引入软间隔支持向量机的概念。在软间隔支持向量机中,允许模型对部分样本放宽分类要求,即允许函数间隔为 $1 - \xi_i$ ($\xi_i > 0$)。在这种情况下,优化目标不仅涉及在原始数据空间中最大化间隔,同时还增加了一项惩罚项 $C\xi_i$ 。通过这种方式,软间隔支持向量机在优化过程中考虑了两方面因素:一方面是建立一个大的间隔以增强模型的泛化能力,另一方面是惩罚违反间隔约束的样本,在最大化间隔与减少误分类之间寻求平衡。

超参数 C 在软间隔支持向量机中起到关键作用,它用于权衡模型的“大间隔”与“减小分类误差”之间的矛盾。当 C 取值较大时,模型倾向于惩罚错分样本,对误差的容忍度降低;而当 C 取值较小时,模型则更为宽松,允许更多的样本违反间隔约束以换取更宽的分类间隔。这种灵活性使得软间隔支持向量机能够更好地适应复杂的实际应用场景。

软间隔支持向量机的问题形式依然是一个有约束优化问题,然而它的求解相较于硬间隔支持向量机具有更多的挑战性。具体而言,软间隔支持向量机的原问题形式为

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ & \xi_i \geq 0, \quad 1 \leq i \leq n \end{aligned} \quad (3.49)$$

基于之前求解硬间隔支持向量机的经验,我们仍希望通过对偶形式来简化软间隔支持向量机的求解。但在应用对偶形式进行求解之前,首先需要验证软间隔支持向量机的原问题满足 KKT 条件成立的前提:

- (1) $f(\mathbf{w}, b, \xi) = \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i$ 为凸函数；
 (2) $g_i(\mathbf{w}, b, \xi) = 1 - \xi_i - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)$ 为凸函数；
 (3) 不存在等式约束条件 $h_i(\cdot)$ ；
 (4) 对任意样本 $\mathbf{x}_i, \mathbf{x}_j$, 显然存在 $\mathbf{w}, b$ 满足 $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1 - \xi_i$ 。
 其对应的 KKT 条件为

$$\begin{aligned}
 \alpha_i &\geq 0, \quad \mu_i \geq 0 \\
 y_i(\mathbf{w} \cdot \mathbf{x}_i + b) &\geq 1 - \xi_i \\
 \alpha_i(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i) &= 0 \\
 \xi_i &\geq 0, \quad \mu_i \xi_i = 0 \\
 \alpha_i + \mu_i &= C
 \end{aligned} \tag{3.50}$$

仿照硬间隔支持向量机的推导思路, 软间隔支持向量机的对偶形式为

$$\begin{aligned}
 &\max_{\substack{\alpha_i: \alpha_i \geq 0 \\ \mu_i: \mu_i \geq 0}} \min_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu) \\
 &= \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i (1 - \xi_i - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)) - \sum_{i=1}^n \mu_i \xi_i
 \end{aligned} \tag{3.51}$$

令 $\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu)$ 对 $\mathbf{w}, b, \xi$ 的偏导数为 0, 可将上式化简为

$$\begin{aligned}
 &\max_{\alpha} -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \\
 &\text{s.t.} \quad \sum_i \alpha_i y_i = 0 \\
 &0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n
 \end{aligned} \tag{3.52}$$

其判别函数为

$$\begin{aligned}
 \mathbf{w}^* &= \sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i \\
 b^* &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left(y_i - \sum_{j \in \mathcal{S}} \alpha_j^* y_j \mathbf{x}_j \cdot \mathbf{x}_i \right) \\
 \hat{y} &= \text{sgn} \left(\sum_{i \in \mathcal{S}} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^* \right)
 \end{aligned} \tag{3.53}$$

对比硬间隔和软间隔支持向量机的求解过程可以发现, 二者对偶问题的优化目标函数一致, 解法一致, 最终判别函数的表达式也一致。二者的不同之处在于, 硬间隔支持向量机要求 $\alpha_i \geq 0$, 而软间隔支持向量机要求 $0 \leq \alpha_i \leq C$ 。从 KKT 条件的角度来看, 当 $\alpha_i = 0$ 时, $\mu_i = C, \xi_i = 0, y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$, 该样本点为非支持向量。当 $0 < \alpha_i < C$ 时, $0 < \mu_i < C, \xi_i = 0, y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$, 该样本点为支持向量, 且在最大间隔边界上。当 $\alpha_i = C$ 时, $\mu_i = 0, \xi_i \geq 0, y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i \leq 1$, 该样本点为支持向量, 且越过了最大间隔边界。可通过图 3.11 所示理解软间隔支持向量机的求解过程。

综上所述, 软间隔支持向量机的提出旨在克服硬间隔支持向量机的局限性。硬间隔支持向量机严格要求所有样本位于决策边界的正确一侧, 这种理想化的假设在面对包含噪声和离群值的真实数据集时往往难以满足。相比之下, 软间隔支持向量机通过引入松弛变量