6.1 间隔与支持向量机

6.1.1 问题的提出

给定训练样本集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$, $y_i\in\{-1,+1\}$, 分类学习最基本的想法就是基于训练集 D 在样本空间中找到一个划分超平面, 将不同类别的样本分开. 能将训练样本分开的划分超平面有很多. 如下图.

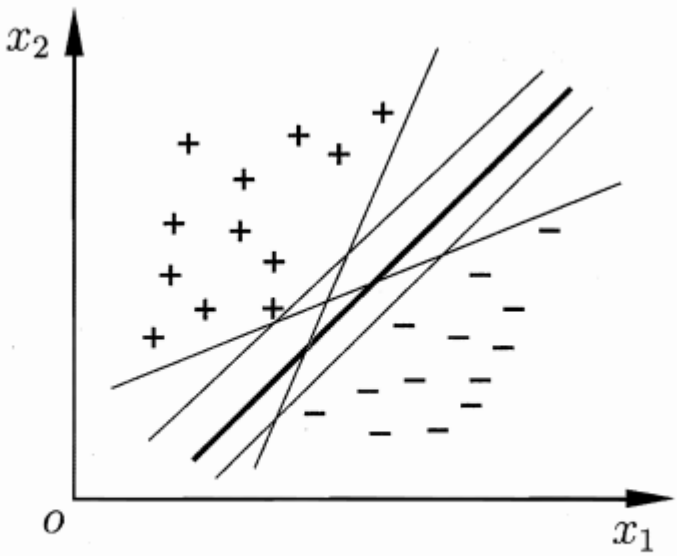

图 6.1 存在多个划分超平面将两类训练样本分开

可以看出, 粗线条的划分超平面"容忍性"最好. 换句话说, 这个划分超平面所产生的分类结果是最鲁棒的, 对未见示例的泛化能力最强.

6.1.2 数学表示

在样本空间中, 划分超平面可通过如下线性方程来描述:

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0 \tag{6.1}$$

其中, $\boldsymbol{w}=(w_1;w_2;\ldots;w_d)$ 是法向量, 决定了超平面的方向; b 为位移项, 决定了超平面与原点之间的距离, 划分超平面可被法向量 $\boldsymbol{w}$ 和位移 b 确定. 将法向量记为 $(\boldsymbol{w},b)$.

样本空间中任意点 $\boldsymbol{x}$ 到超平面 $(\boldsymbol{w},b)$ 的距离可以写成

$$r=\frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right|}{\|\boldsymbol{w}\|} \tag{6.2}$$

假设超平面 $(\boldsymbol{w},b)$ 能将训练样本正确分类, 即对于 $(\boldsymbol{x}_i,y_i)\in D$, 若 $y_i=+1$, 则有 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b>0$; 若 $y_i=-1$, 则有 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b<0$.

令

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geqslant +1, & y_i = +1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \leqslant -1, & y_i = -1 \end{cases} \tag{6.3}$$

如图 (6.2)

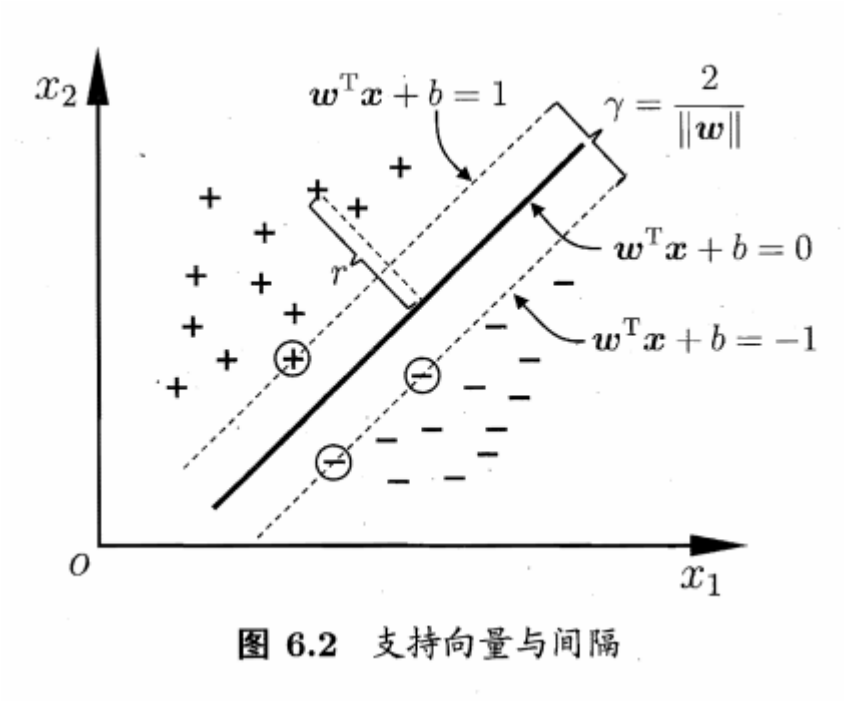

图 6.2　支持向量与间隔

如图 (6.2) , 距离超平面最近的这几个训练样本点使式 (6.3) 的等号成立, 它们被称为"支持向量机"(support vector) , 两个异类支持向量到超平面的距离之和为:

$$\gamma = \frac{2}{\|\boldsymbol{w}\|} \tag{6.4}$$

称为**"间隔"(margin)**

欲找到具有"最大间隔"(maximum margin)的划分超平面, 也就是找到能满足式(6.3)中约束的参数 $\boldsymbol{w}$ 和 b , 使得 γ 最大, 即

$$\begin{aligned} &\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geqslant 1, \quad i = 1, 2, \ldots, m \end{aligned} \tag{6.5}$$

最大化间隔, 即最大化 $\|\boldsymbol{w}\|^{-1}$, 也就是等价于最小化 $\|\boldsymbol{w}\|^2$. 则 (6.5) 可重写为

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geqslant 1, \quad i = 1, 2, \ldots, m \end{aligned} \tag{6.6}$$

这就是**支持向量机**(support vector machine, 简称SVM)的**基本型**.

6.2 对偶问题

6.2.1 对偶问题

我们的目的就是根据式 (6.6) 来求得最大间隔划分超平面所对应的模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \tag{6.7}$$

其中 $\boldsymbol{w}$ 和 b 是模型参数. (6.6)本身是一个凸二次规划问题, 可以直接求解, 但是有更高效的方法.

对式 (6.6) 使用拉格朗日乘子法可得到其"对偶问题"(dual poblem). 具体来说, 对式 (6.6) 的每条约束添加拉格朗日乘子 $\alpha_i \geqslant 0$, 则该问题的拉格朗日函数可以写成

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)\right)$$

> **注1:** 令 $\theta(w) = \max_{\alpha_i \geq 0} L(w, b, \alpha)$, 易验证,当某个约束条件不满足时，例如 $y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) < 1$ ，那么显然有 $\theta(w) = \infty$ (只要令 $\alpha_i = \infty$ 即可)。而当所有约束条件都满足时，则有 $\theta(w) = \frac{1}{2}\|\boldsymbol{w}\|^2$，亦即最初要最小化的量。
>
> 因此，在要求约束条件得到满足的情况下最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2$，实际上等价于直接最小化 $\theta(w)$（当然，这里也有约束条件，就是 $\alpha_i \geq 0, i = 1, \ldots, n$） , 因为如果约束条件没有得到满足， $\theta(w)$会等于无穷大，自然不会是我们所要求的最小值
>
> 具体写出来，目标函数变成了:
>
> $$\min_{\boldsymbol{w},b} \theta(\boldsymbol{w}) = \min_{\boldsymbol{w},b} \max_{\alpha_i \geq 0} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = p^*$$
>
> 把最小和最大的位置交换一下, 变成:
>
> $$\max_{\alpha_i \geq 0} \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = d^*$$

其中, $\boldsymbol{\alpha} = (\alpha_1; \alpha_2; \ldots; \alpha_m)$. 令 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 对 $\boldsymbol{w}$ 和 b 的偏导为零可得

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{6.9}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{6.10}$$

> **注2:** 对 $\boldsymbol{w}$ 求偏导有:
>
> $\frac{\partial L}{\partial w} = \frac{1}{2} \times 2 \times \boldsymbol{w} + 0 - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i - 0 = 0 \Longrightarrow \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$ (1. $\frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}{\partial \boldsymbol{w}} = \boldsymbol{x}_i$ 2. $\frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{w}$)
>
> $\frac{\partial L}{\partial b} = 0 + 0 - 0 - \sum_{i=1}^{m} \alpha_i y_i = 0 \Longrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$

将式 (6.9)带入式 (6.8) , 即可将 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 中的 $\boldsymbol{w}$ 和 b 消去, 再考虑式 (6.10) 的约束, 就得到式 **(6.6) 的对偶问题**

> **注3:**
>
> $$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{w}^T\boldsymbol{x}_i - \sum_{i=1}^{m}\alpha_i y_i b \\ &= \frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i - \boldsymbol{w}^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i - b\sum_{i=1}^{m}\alpha_i y_i \\ &= -\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i - b\sum_{i=1}^{m}\alpha_i y_i \end{aligned}$$
>
> 又 $0 = \sum_{i=1}^{m}\alpha_i y_i$, 进一步化简得到
>
> $$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= -\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \\ &= -\frac{1}{2}(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i)^T(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i) + \sum_{i=1}^{m}\alpha_i \\ &= -\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \end{aligned}$$
>
> 所以,$\max_{\alpha_i \geq 0}\min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$, 最后 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 的只有参数 α_i (α_j 一样). 因此
>
> $\max_{\alpha_i \geq 0}\min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$

(6.6) 的对偶问题

$$\begin{aligned} \max_{\boldsymbol{\alpha}} & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \\ s.t. & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & \alpha_i \geqslant 0, \quad i = 1, 2, \ldots, m \end{aligned} \tag{6.11}$$

6.2.2 KKT条件

解出 $\boldsymbol{\alpha}$ 后, 求出 $\boldsymbol{w}$ 和b 即可得到模型

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \\ &= \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} + b \end{aligned} \tag{6.12}$$

从对偶问题 (6.11)解出的 α_i 是式 (6.8) 中的拉格朗日乘子, 它恰对应着训练样本$(\boldsymbol{x}_i, y_i)$. 注意到式 (6.6) 中有不等式约束, 因此上述过程需满足 KKT(Karush-Kuhn-Tucker) 条件, 即要求

$$\begin{cases} \alpha_i \geqslant 0 \\ y_i f\left(\boldsymbol{x}_i\right)-1 \geqslant 0 \\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right)-1\right)=0 \end{cases}$$

> **注4:** 关于拉格朗日对偶性以及KKT条件, 详细知识参见统计学习方法附录C, 别问为什么是这样的, 再问就是"易得"

于是, 对任意训练样本 $\left(\boldsymbol{x}_i, y_i\right)$, 总有 $\alpha_i=0$ 或 $y_i f\left(\boldsymbol{x}_i\right)=1$.若 $\alpha_i=0$, 则该样本将不会在式 (6.12) 的求和中出现, 也就不会对 $f(\boldsymbol{x})$ 有任何影响; 若 $\alpha_i>0$, 则必有 $y_i f\left(\boldsymbol{x}_i\right)=1$, 所对应的样本点位于最大间隔边界上, 是一个支持向量. 这显示出支持向量机的一个重要性质:训练完成后, 大部分的训练样本都不需保留, 最终模型仅与支持向量有关 .

6.2.3 SMO算法

暂放

6.3 核函数

6.3.1 线性不可分的情况

前面的都是假设训练样本是线性可分的，即存在一个划分超平面能将训练样本正确分类.然而在现实任务中, 原始样本空间内也许并不存在一个能正确划分两类样本的超平面. 如图 6.3 中的" 异或 " 问题就不是线性可分的.

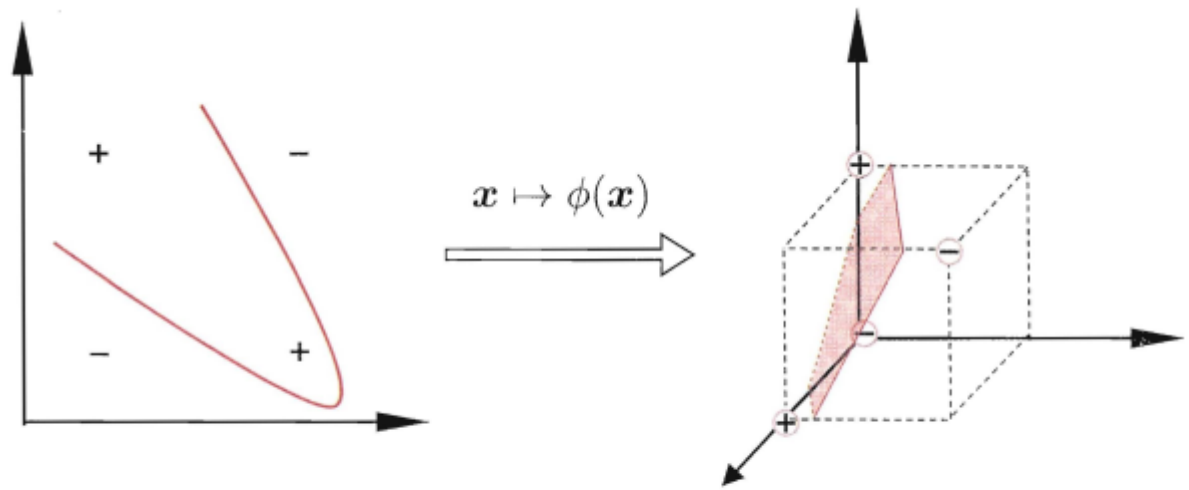

图 6.3　异或问题与非线性映射

对这样的问题, **可将样本从原始空间映射到一个更高维的特征空间, 使得样本在这个特征空间内线性可分.**

如在图 6.3 中, 若将原始的二维空间映射到一个合适的三维空间, 就能找到一个合适的划分超平面.

同时, **如果原始空间是有限维, 即属性数有限, 那么一定存在一个高维特征空间使样本可分**

6.3.2 数学表示

令 $\phi(\boldsymbol{x})$ 表示将 $\boldsymbol{x}$ 映射后的特征向量, 于是, 在特征空间中划分超平面所对应的模型可表示为:

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}} \phi(\boldsymbol{x})+b \tag{6.19}$$

其中, $\boldsymbol{w}$ 和 b 是模型参数. 类似于式 (6.6), 有

$$\begin{aligned} &\min_{\boldsymbol{w}, b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}} \phi\left(\boldsymbol{x}_i\right)+b\right) \geqslant 1, \quad i=1,2, \ldots, m \end{aligned} \tag{6.20}$$

其对偶问题是

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^m \alpha_i-\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \phi\left(\boldsymbol{x}_j\right) \tag{6.21}$$

$$\begin{aligned} s.t. & \sum_{i=1}^m \alpha_i y_i=0 \\ & \alpha_i \geqslant 0, \quad i=1,2, \ldots, m \end{aligned}$$

$\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \phi\left(\boldsymbol{x}_j\right)$ 表示样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 映射到特征空间之后的内积. 计算较为为难. 为避开, 设想这样一个函数:

$$\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\left\langle\phi\left(\boldsymbol{x}_i\right), \phi\left(\boldsymbol{x}_j\right)\right\rangle=\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \phi\left(\boldsymbol{x}_j\right) \tag{6.22}$$

即 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 在特征空间的内积等于它们在原始样本空间中通过函数 $\kappa(\cdot, \cdot)$ 计算的结果.

于是, 式 (6.21) 可以重写为:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^m \alpha_i-\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \tag{6.23}$$

$$\begin{aligned} s.t. & \sum_{i=1}^m \alpha_i y_i=0 \\ & \alpha_i \geqslant 0, \quad i=1,2, \ldots, m \end{aligned}$$

求解后即可得到

$$\begin{aligned} f(\boldsymbol{x}) & =\boldsymbol{w}^{\mathrm{T}} \phi(\boldsymbol{x})+b \\ & =\sum_{i=1}^m \alpha_i y_i \phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \phi(\boldsymbol{x})+b \\ & =\sum_{i=1}^m \alpha_i y_i \kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right)+b \end{aligned} \tag{6.24}$$

这里的函数 $\kappa(\cdot, \cdot)$ 就是**"核函数"**(kernel function)

式 (6.24) 显示出模型最优解可通过训练样本的函数展开, 这一展开式亦称 "支持向量展式"

6.3.3 有关核函数的定理

1 定理6.1 (核函数)

定理6.1 (核函数) 令 $\mathcal{X}$ 为输入空间, $\kappa(\cdot,\cdot)$ 是定义在 $\mathcal{X}\times\mathcal{X}$ 上的对称函数, 则 k 是核函数当且仅当对于任意数据 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$, "核矩阵" (kernel matrix) $\mathbf{K}$ 总是本正定的:

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

定理 6.1 表明, **只要一个对称函数所对应的核矩阵半正定, 它就能作为核函数使用.** 事实上, **对于一个半正定核矩阵, 总能找到一个与之对应的映射 ϕ** . 换言之, 任何一个核函数都隐式地定义了一个称为"再生核希尔伯特空间" (Reproducing Kernel Hilbert Space，简称 RKHS) 的特征空间.

2 常用的核函数及核函数的相关组合

表 6.1 列出了集中常用的核函数

表 6.1 常用核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^d$	$d\geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma>0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma>0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh(\beta\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+\theta)$	tanh 为双曲正切函数, $\beta>0,\theta<0$

此为, 还可以通过函数组合得到相关的核函数.

- 若 k_1 和 k_2 为核函数, 则对于任意正数 γ_1,γ_2 , 其线性组合

$$\gamma_1\kappa_1+\gamma_2\kappa_2 \tag{6.25}$$

  也是核函数

- 若 k_1 和 k_2 为核函数, 则核函数的直积

$$\kappa_1\otimes\kappa_2(\boldsymbol{x},\boldsymbol{z})=\kappa_1(\boldsymbol{x},\boldsymbol{z})\kappa_2(\boldsymbol{x},\boldsymbol{z}) \tag{6.26}$$

  也是核函数

- 若 k_1 为核函数, 则对于任意函数 $g(\boldsymbol{x})$,

$$\kappa(\boldsymbol{x},\boldsymbol{z})=g(\boldsymbol{x})\kappa_1(\boldsymbol{x},\boldsymbol{z})g(\boldsymbol{z})$$

也是函数

6.4 软间隔与正则化

6.4.1 软间隔的概念

前面我们假定训练样本在**样本空间**或**特征空间**中是**线性可分**的, 即存在一个超平面能将不同类的样本完全划分开. 然而, 在现实任务中往往很难确定合适的核函数使得训练样本在特征空间中线性可分; 即使恰好找到了 某个核函数使训练集在特征空间中线性可分, 也很难断定这个貌似线性可分的结果不是由于**过拟合**所造成的.

缓解该问题的一个办法是**允许支持向量机在一些样本上出错**.为此, 要引入"**软间隔**" (soft margin)的概念, 如图6.4所示.

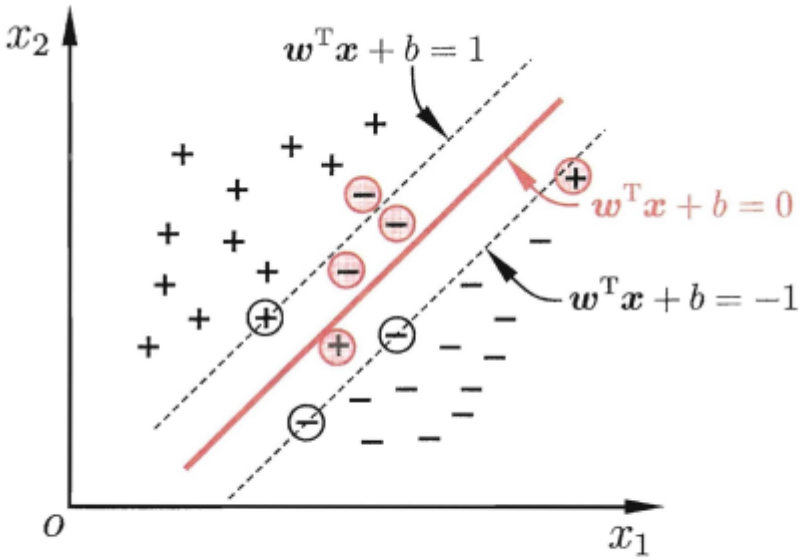

图 6.4 软间隔示意图. 红色圈出了一些不满足约束的样本.

6.4.2 数学表示

前面介绍的支持向量机形式是**要求所有样本均满足约束 (6.3)**, 即**所有样本都必须划分正确**, 这称为" **硬间隔**" (hard margin), 而**软间隔则是允许某些样本不满足约束**

$$y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1 \tag{6.28}$$

在**最大化间隔的同时, 不满足约束的样本应尽可能的少**. 则优化目标可以写为

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}\ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)-1\right) \tag{6.29}$$

其中 $C>0$ 是一个常数, $\ell_{0/1}$ 是 "0/1损失函数"

$$\ell_{0/1}(z) = \begin{cases} 1, & \text{if } z < 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.30}$$

显然, 当 C 为无穷大时, 式 (6.29) 迫使所有样本均满足约束 (6.28) , 于是式 (6.29) 等价于 式 (6.6) ; 当 C 取有限值时, 式 (6.29) 允许一些样本不满足约束.

然而, $\ell_{0/1}$ 非凸、非连续, 数学性质不太好, 使得式 (6.29) 不易直接求解. 人们通常用其他一些函数来代替 $\ell_{0/1}$, 称为"**替代损失**" (surrogate loss). 替代损失函数一般具有较好的数学性质, 如它们通常是凸的连续函数且是 $\ell_{0/1}$ 的上界. 图 6.5 给出了三种常用的替代损失函数:

- hinge 损失: $\ell_{\text{hinge}}(z) = \max(0, 1-z)$ (6.31)
- 指数损失(exponential loss): $\ell_{exp}(z) = \exp(-z)$ (6.32)
- 对率损失(logistic loss): $\ell_{\text{exp}}(z) = \exp(-z)$ (6.33)

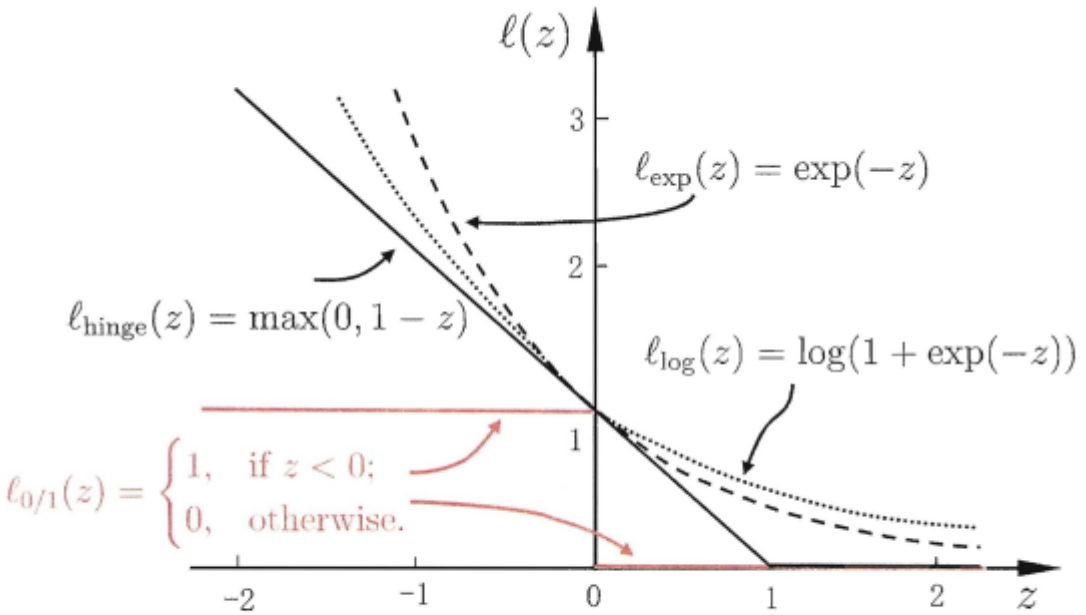

图 6.5 三种常见的替代损失函数: hinge损失、指数损失、对率损失

若采用hinge损失, 则式 (6.29) 变成

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)\right) \tag{6.34}$$

> 注5: 同6.29一样的思路来理解

接着, 引入"**松弛变量**"(slack variable) $\xi_i \geqslant 0$, 可将式 (6.34) 重写为

$$\min_{\boldsymbol{w},b,\xi_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i \tag{6.35}$$

$$s.t. \quad y_i\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i+b\right) \geqslant 1-\xi_i$$

$\xi_i \geqslant 0, i=1,2, \ldots, m$

这就是常用的**"软间隔支持向量机"**

> 注6: 式(6.35)是式(6.34)的上限，最小化式(6.35)的同时也会最化小式(6.34)，这是因为:
>
> 由式(6.35)的约束条件 $y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i+b\right) \geqslant 1-\xi_i$ 可得 $\xi_i \geqslant 1-y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i+b\right)$，再加上约束条件 $\xi_i \geqslant 0$，即 $\xi_i \geqslant \max\left(0,1-y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i+b\right)\right)$，因此式(6.35)是式(6.34)的上限

显然, 式 (6.35) 中每个意昂本都有一个对应的松弛变量, 用以表征该样本不满足约束 (6.28) 的程度.

与式 (6.6) 相似, 这还是一个二次规划问题. 于是, 类似于式 (6.8) , 通过拉格朗日乘子法可得到式 (6.35) 的拉格朗日函数

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = & \frac{1}{2}\|\boldsymbol{w}\|^2+C \sum_{i=1}^m \xi_i \\ & +\sum_{i=1}^m \alpha_i\left(1-\xi_i-y_i\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i+b\right)\right)-\sum_{i=1}^m \mu_i \xi_i \end{aligned} \tag{6.36}$$

其中, $\alpha_i \geqslant 0, \mu_i \geqslant 0$ 是拉格朗日乘子.

> 注7: **拉格朗日基本形式**为
>
> $$\min_{x \in \mathbf{R}^n} f(x)$$
>
> $$s.t. \quad c_i(x) \leqslant 0, \quad i=1,2, \cdots, k$$
>
> $$h_j(x)=0, \quad j=1,2, \cdots, l$$
>
> $$L(x, \alpha, \beta)=f(x)+\sum_{i=1}^k \alpha_i c_i(x)+\sum_{j=1}^l \beta_j h_j(x)$$
>
> 引入拉格朗日函数有
>
> $$L(x, \alpha, \beta)=f(x)+\sum_{i=1}^k \alpha_i c_i(x)+\sum_{j=1}^l \beta_j h_j(x)$$

令 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$ 对 $\boldsymbol{w}, b, \xi_i$ 的偏导为零可得

$$\boldsymbol{w}=\sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i \tag{6.37}$$

$$0=\sum_{i=1}^{m}\alpha_i y_i \tag{6.38}$$

$$C=\alpha_i+\mu_i \tag{6.39}$$

将式 (6.37)-(6.39) 代入式 (6.36) 即可得到式 (6.35) 的**对偶问题**

$$\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \tag{6.40}$$

$$s.t.\sum_{i=1}^{m}\alpha_i y_i=0$$

$0\leqslant\alpha_i\leqslant C,\quad i=1,2,\ldots,m$

软间隔的对偶问题式 (6.40) 与硬间隔下的对偶问题 (6.11) 对比可看出, 两者唯一的差别就再与对偶变量的约束不同: 前者是 $0\leqslant\alpha_i\leqslant C$, 后者是 $0\leqslant\alpha_i$. 于是，可采用 6.2节中同样的算法求解式 (6.40); 在引入**核函数**后能得到与式 (6.24) 同样的**支持向量展式**.

类似于式 (6.13), 对软间隔支持向量机, KKT条件要求

$$\begin{cases}\alpha_i\geqslant 0,\quad \mu_i\geqslant 0\\ y_i f\left(\boldsymbol{x}_i\right)-1+\xi_i\geqslant 0\\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right)-1+\xi_i\right)=0\\ \xi_i\geqslant 0,\quad \mu_i\xi_i=0\end{cases} \tag{6.41}$$

对任意训练样本 $\left(\boldsymbol{x}_i, y_i\right)$, 总有 $\alpha_i=0$ 或 $y_i f\left(\boldsymbol{x}_i\right)=1-\xi_i$. 若 $\alpha_i=0$, 则该样本不会对 $f(\boldsymbol{x})$ 有任何影响; 若 $\alpha_i>0$, 则必有 $y_i f\left(\boldsymbol{x}_i\right)=1-\xi_i$, 即该样本是支持向量: 由式 (6.39)可知, 若 $\alpha_i<C$, 则 $\mu_i>0$, 进而有 $\xi_i=0$, 即该样本恰在最大间隔边界上; 若 $\alpha_i=C$, 则有 $\mu_i=0$, 此时若 $\xi_i\leqslant 1$, 则该样本落在最大间隔内部, 若 $\xi_i>1$, 则该样本被错误分类.

由此可看出, 软间隔支持向量机的最终模型**仅与支持向量有关**, 即通过采用hinge损失函数仍**保持了稀疏性**.