7 Support Vector Machine 支持向量机

间隔最大化，支持向量确定决策边界

没有什么比精巧理论更实用的了。

Nothing is more practical than a good theory.

—— 弗拉基米尔·万普尼克 (Vladimir Vapnik) | 俄罗斯统计学家、数学家 | 1936 ~

- `numpy.hstack()` 水平方向将数组堆叠起来
- `numpy.vstack()` 竖直方向将数组堆叠起来
- `sklearn.svm.SVC` 支持向量机算法函数

7.1 支持向量机

弗拉基米尔 · 万普尼克 (Vladimir Vapnik) 和他的同事们发明并且完善了**支持向量**机 (Support Vector Machine, SVM)。弗拉基米尔·万普尼克为机器学习发展奠定了大量理论基础，有兴趣的大家可以翻看他的作品——*The Nature of Statistical Learning Theory*。

弗拉基米尔·万普尼克 (Vladimir Vapnik) | 俄罗斯统计学家、数学家 | 1936 ~
支持向量机发明者之一。关键词：• 支持向量机 • 核技巧

原理

图 1 所示为支持向量机核心思路。如图 1 所示，一片湖面左右散布着蓝色 • 红色 • 礁石，游戏规则是，皮划艇以直线路径穿越水道，保证船身恰好紧贴礁石。寻找一条路径，让该路径通过的皮划艇宽度最大。很明显，图 1 (b) 中规划的路径好于图 1 (a)。

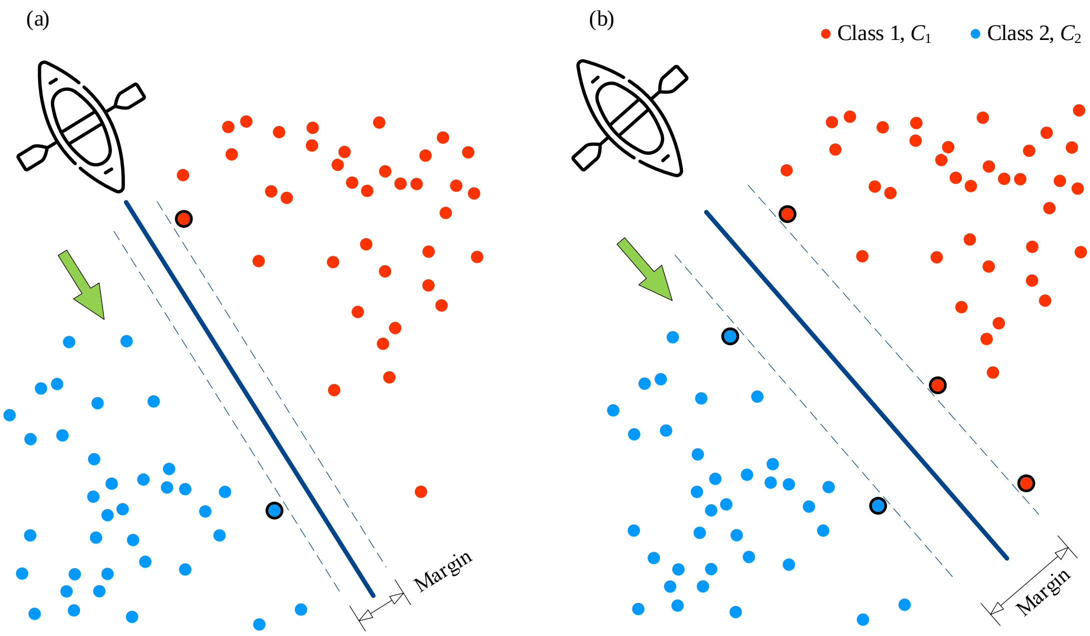

图 1. 支持向量机原理

图 1 (b) 中加黑圈 ○ 的五个点，就是所谓的**支持向量** (support vector)。

图 1 中深蓝色线，便是**决策边界**，也称**分离超平面** (separating hyperplane)。本书为了统一称呼，下文都使用决策边界。特别提醒大家注意一点，加黑圈 ○ 支持向量确定决策边界位置；其他数据并没有起到任何作用。因此，SVM 对于数据特征数量远高于数据样本量的情况也有效。

图 1 中两条虚线之间宽度叫做**间隔** (margin)。正如，本章副标题所言，支持向量机的优化目标为——间隔最大化。

线性可分、线性不可分

从数据角度，图 1 两类数据用一条直线便可以分割开来，这种数据叫做**线性可分** (linearly separable)。线性可分问题采用**硬间隔** (hard margin)；白话说，硬间隔指的是，间隔内没有数据点。

实践中，并不是所有数据都是线性可分。多数时候，数据**线性不可分** (non-linearly separable)。如图 2 所示，不能找到一条直线将蓝色 ● 红色 ● 数据分离。

对于线性不可分问题，就要引入两种方法——**软间隔** (soft margin) 和**核技巧** (kernel trick)。

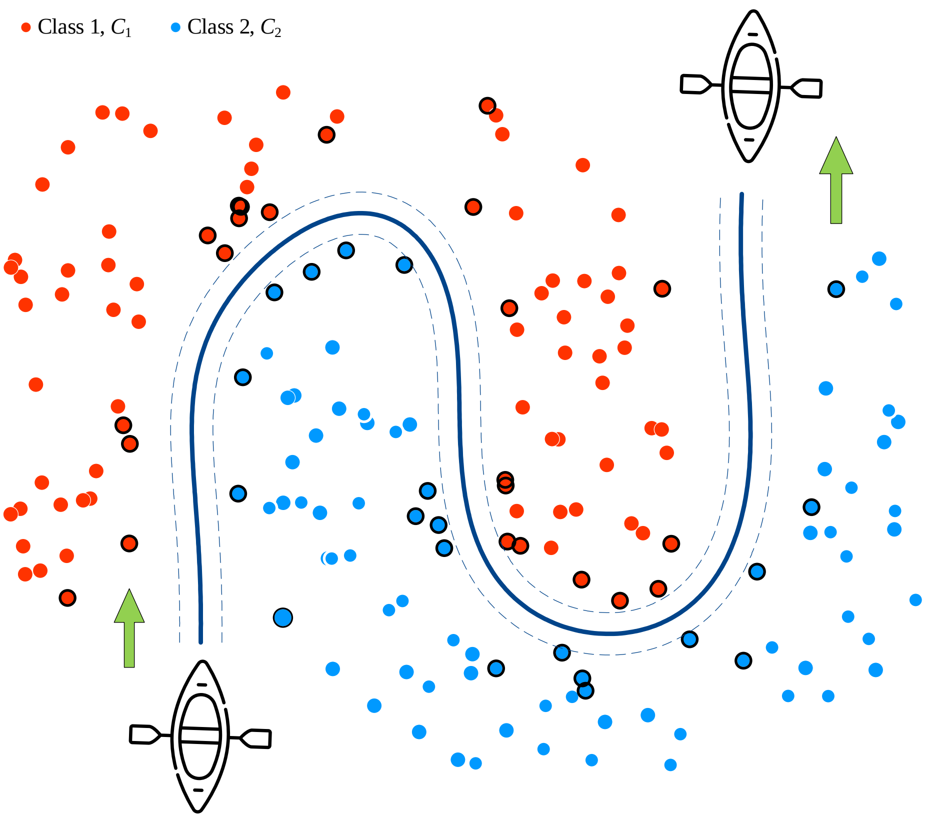

图 2. 线性不可分数据

软间隔

白话说，如图 3 所示，软间隔相当于一个**缓冲区** (buffer zone)。软间隔存在时，用决策边界分离数据时，有数据点侵入间隔，甚至超越间隔带。

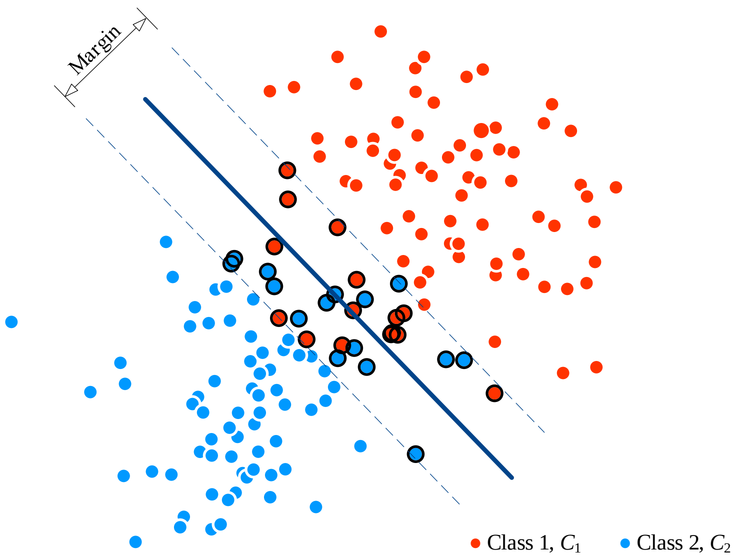

图 3. 软间隔

核技巧

核技巧将数据映射到高维特征空间，是一种数据升维。如图 4 所示，样本数据有两个特征，用平面可视化数据点位置。很明显图 4 给出的原始数据线性不可分。

采用核技巧，将图 4 二维数据，投射到三维核曲面上；很明显，在这个高维特征空间，容易找到某个水平面，将蓝色 • 红色 • 数据分离。利用核技巧，分离线性不可分数据变得更容易。

通常，采用支持向量机解决线性不可分问题，需要并用软间隔和核技巧。如图 5 所示，SVM 分类环形数据中，核技巧配合软间隔。

另外，支持向量机也可以用来处理回归问题，对应的方法为**支持向量回归** (Support Vector Regression, SVR)。本章将主要介绍硬间隔、支持向量和软间隔；下一章，将介绍核技巧。本章和下一章有一定比例的公式推导，这对理解支持向量机原理有帮助，希望大家耐心阅读。

《矩阵力量》第 19 章为本章提供大量数学工具，建议大家回顾。

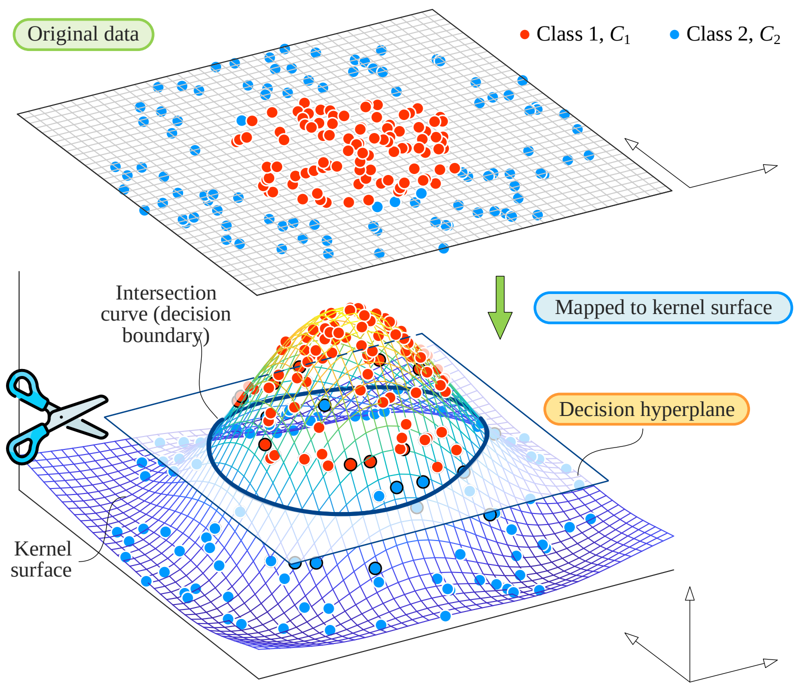

图 4. 核技巧原理

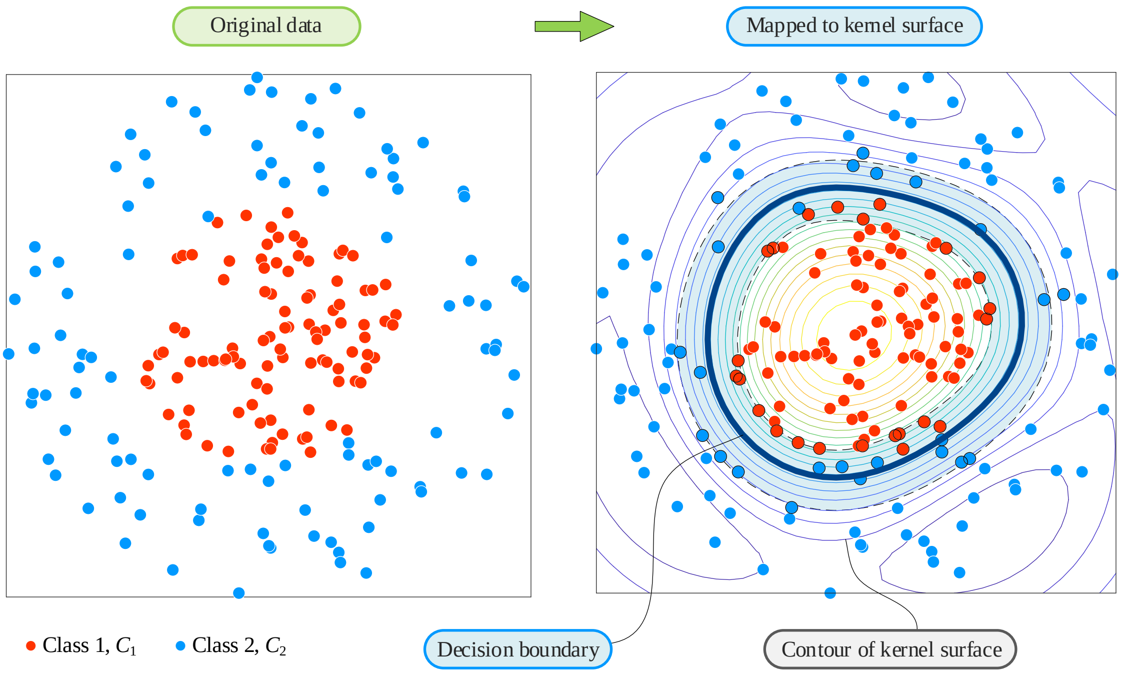

图 5. 核技巧配合软间隔

7.2 硬间隔：处理线性可分

支持向量机中硬间隔方法用来处理线性可分数据。利用《矩阵力量》一册讲解的向量几何知识，这一节将构造 SVM 中支持向量、决策边界、分类标签和间隔等元素之间的数学关系。

决策边界

如图 6 所示，决策边界定义如下：

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{1}$$

其中，$\boldsymbol{w}$ 和 b 为模型参数；$\boldsymbol{w}$ 为 $f(\boldsymbol{w})$ 的梯度向量，形式为列向量。(1) 中，列向量 $\boldsymbol{w}$ 和 $\boldsymbol{x}$ 行数均为特征数 D。

很明显 (1) 为**超平面** (hyperplane)。注意，图 6 所示间隔宽度为 $2h$ ($h > 0$)。

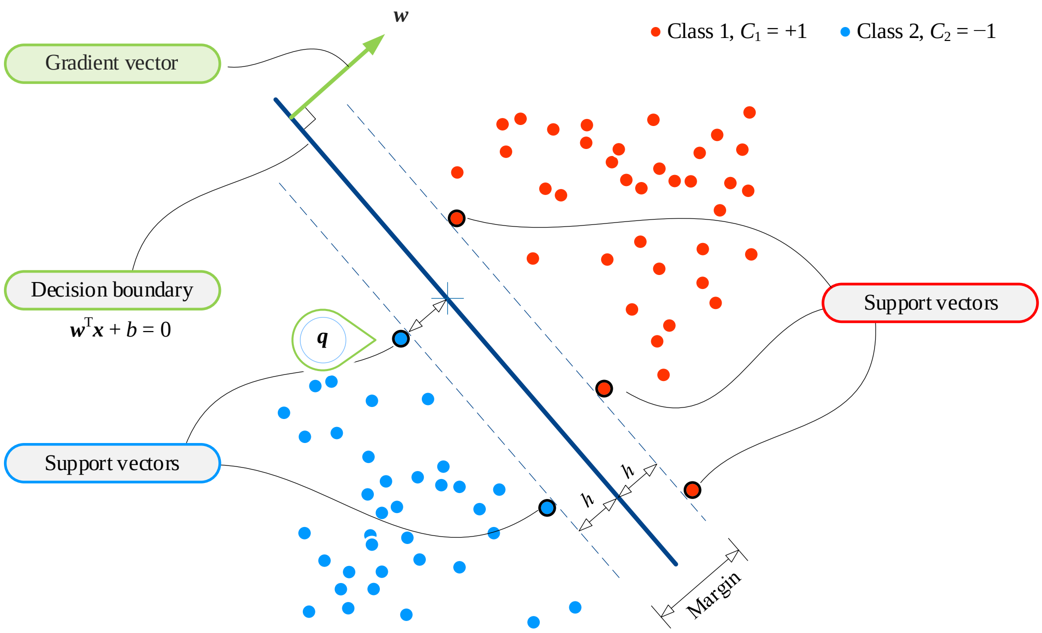

图 6. 硬间隔 SVM 处理二分类问题

(1) 可以展开为：

$$f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \ldots + w_Dx_D + b = 0 \tag{2}$$

特别地，对于 $D = 2$ 时，决策边界形式为：

$$w_1x_1 + w_2x_2 + b = 0 \tag{3}$$

分类

对于二分类 ($K = 2$) 问题，决策边界“上方”的数据点满足：

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b > 0 \tag{4}$$

展开 (4) 得到：

$$w_1 x_1 + w_2 x_2 + \ldots + w_D x_D + b > 0 \tag{5}$$

决策边界“下方”的数据点满足：

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b < 0 \tag{6}$$

展开 (6) 得到：

$$w_1 x_1 + w_2 x_2 + \ldots + w_D x_D + b < 0 \tag{7}$$

准确地说，以 (1) 中 $f(\boldsymbol{x}) = 0$ 为基准，“上方”对应 $f(\boldsymbol{x}) > 0$；“下方”对应 $f(\boldsymbol{x}) < 0$。

决策函数

对任意查询点 $\boldsymbol{q}$，二分类决策函数 $p(\boldsymbol{q})$ 则可以表达为：

$$p(\boldsymbol{q}) = \operatorname{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q} + b\right) \tag{8}$$

其中，sign() 为**符号函数** (sign function)。

如图 6 所示，对于二分类 ($K = 2$) 问题，决策边界“上方”的数据点，预测分类为+1；决策边界“下方”的数据点，预测分类为−1。

支持向量到决策边界距离

图 6 中，某一支持向量坐标位置用列向量 $\boldsymbol{q}$ 表达。支持向量 $\boldsymbol{q}$ 到 (1) 对应的决策边界的距离为：

$$d = \frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q} + b\right|}{\|\boldsymbol{w}\|} = \frac{\left|\boldsymbol{w}\cdot\boldsymbol{q} + b\right|}{\|\boldsymbol{w}\|} \tag{9}$$

对于上式陌生的读者，请回顾《矩阵力量》第 19 章第 6 节。

一般情况点线距离不考虑正负。但是，对于分类问题，考虑距离正负便于判断点和超平面关系。

(9) 分子去掉绝对值符号得到：

$$d = \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q} + b}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}\cdot\boldsymbol{q} + b}{\|\boldsymbol{w}\|} \tag{10}$$

d 大于 0 时，点在超平面上方；d 小于 0 时，点在超平面下方。如图 7 所示，$\boldsymbol{q}_1$ 位于直线上方；而 $\boldsymbol{q}_2$ 位于直线下方。

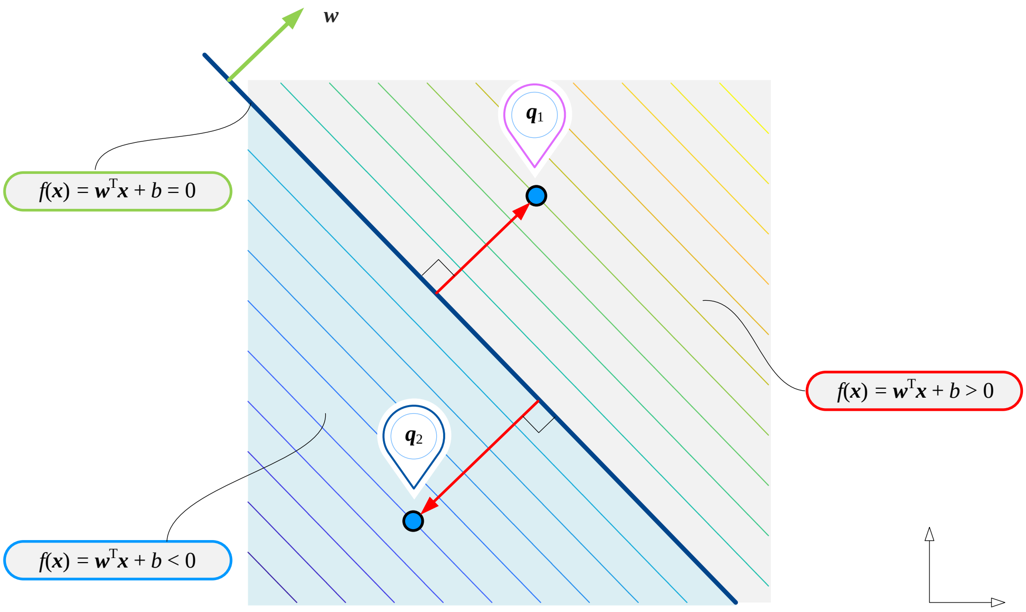

图 7. 直线外一点到直线距离，和平面外一点到平面距离

支持向量到硬间隔距离

如图 8 所示，硬间隔“下边界”为 l_1，l_1 到决策边界距离为$-h$。而支持向量 A、B 在 l_1 上，因此满足：

$$\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}{\|\boldsymbol{w}\|}=-h \tag{11}$$

硬间隔“上边界”为 l_2，l_2 到决策边界为$+h$。支持向量 C 在 l_2 上，满足：

$$\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}{\|\boldsymbol{w}\|}=+h \tag{12}$$

如图 8 所示，决策边界 (深蓝色线) 成功分离样本数据。距离决策边界大于等于 h 的样本点，标记为 $y = +1$；距离决策边界小于等于$–h$ 的样本点，标记为 $y = –1$，即：

$$\begin{cases}\dfrac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}{\|\boldsymbol{w}\|}\geq +h, & y=+1\\ \dfrac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}{\|\boldsymbol{w}\|}\leq -h, & y=-1\end{cases} \tag{13}$$

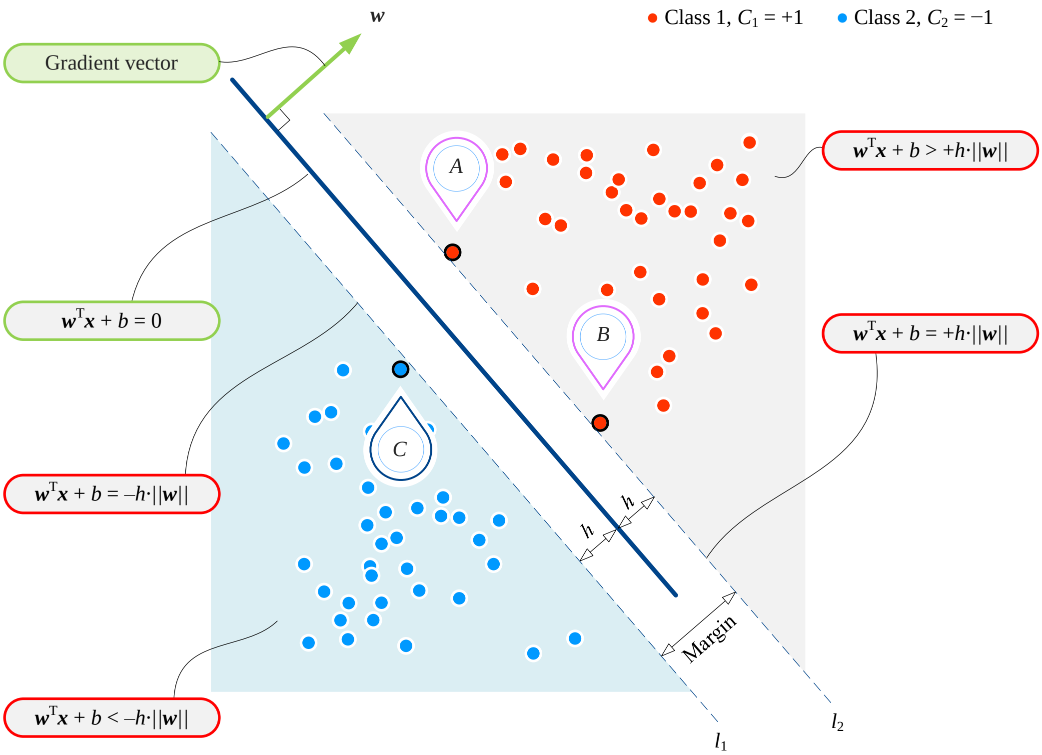

图 8. 硬间隔、决策边界和支持向量之间关系

整理 (13)，得到：

$$\begin{cases} \dfrac{\mathbf{w}^{\mathrm{T}}\mathbf{x}+b}{\|\mathbf{w}\|h} \geq +1, & y=+1 \\ \dfrac{\mathbf{w}^{\mathrm{T}}\mathbf{x}+b}{\|\mathbf{w}\|h} \leq -1, & y=-1 \end{cases} \tag{14}$$

合并 (14) 两式可以得到：

$$\frac{\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}+b\right)y}{\|\mathbf{w}\|h} \geq 1 \tag{15}$$

特别地，图 8 中三个支持向量点 A、B、C 满足下式：

$$\frac{\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}+b\right)y}{\|\mathbf{w}\|h} = 1 \tag{16}$$

进一步简化运算

令：

$$\|\mathbf{w}\|h = 1 \tag{17}$$

(16) 可以简化为：

$$\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right)y \geq 1 \tag{18}$$

利用内积来表达 (18)：

$$\left(\boldsymbol{w}\cdot\boldsymbol{x}+b\right)y \geq 1 \tag{19}$$

将 (17) 代入 (11) 和 (12)，可以得到间隔上下边界的解析式：

$$\begin{cases}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=+1\\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=-1\end{cases} \tag{20}$$

根据 (18)，间隔宽度 $2h$ 可以用 $\boldsymbol{w}$ 表达：

$$2h=\frac{2}{\|\boldsymbol{w}\|} \tag{21}$$

7.3 构造优化问题

支持向量机的核心思想——最大化间隔。本节利用**拉格朗日乘子法** (method of Lagrange multipliers) 构造并求解支持向量机优化问题。本节内容相对来说“很不友好”，但是极其重要，建议大家耐心读完。

对拉格朗日乘子法感到陌生的话，请回顾《矩阵力量》第 18 章。

最大化间隔宽度

以 $\boldsymbol{w}$ 和 b 为优化变量，最大化 (21) 给出的间隔宽度：

$$\begin{aligned}&\underset{\boldsymbol{w},b}{\arg\max}\ \frac{2}{\|\boldsymbol{w}\|}\\ &\text{subject to}\ \left(\boldsymbol{x}^{(i)}\boldsymbol{w}+b\right)y^{(i)}\geq 1,\quad i=1,2,3,\ldots,n\end{aligned} \tag{22}$$

其中，i 为样本数据点序数，$i = 1,2, \ldots, n$。n 为样本数据数量。

最小化问题

(22) 等价于如下最小化问题：

$$\begin{aligned}&\underset{\boldsymbol{w},b}{\arg\min}\ \frac{\|\boldsymbol{w}\|^2}{2}=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}{2}=\frac{\boldsymbol{w}\cdot\boldsymbol{w}}{2}\\ &\text{subject to}\ \left(\boldsymbol{x}^{(i)}\boldsymbol{w}+b\right)y^{(i)}\geq 1,\quad i=1,2,3,\ldots,n\end{aligned} \tag{23}$$

拉格朗日函数

构造**拉格朗日函数** (Lagrangian function) $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$:

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{\boldsymbol{w} \cdot \boldsymbol{w}}{2} + \sum_{i=1}^{n} \lambda_i \left(1 - y^{(i)}\left(\boldsymbol{x}^{(i)}\boldsymbol{w} + b\right)\right) \tag{24}$$

其中，$\boldsymbol{\lambda}$ 为拉格朗日乘子构造的列向量：

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_n \end{bmatrix}^{\mathrm{T}} \tag{25}$$

这样含不等式约束优化问题，转化为一个无约束优化问题。

偏导

$L(\boldsymbol{w}, b, \boldsymbol{\lambda})$ 对 $\boldsymbol{w}$ 和 b 偏导为 0，得到如下一系列等式：

$$\begin{cases} \dfrac{\partial L(\boldsymbol{w}, b, \boldsymbol{\lambda})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)\mathrm{T}} = \boldsymbol{0} \\ \dfrac{\partial L(\boldsymbol{w}, b, \boldsymbol{\lambda})}{\partial b} = \sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \end{cases} \tag{26}$$

这部分内容用到了《矩阵力量》第 17 章介绍的多元微分相关数学工具。

整理 (26) 可以得到：

$$\begin{cases} \boldsymbol{w} = \sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)\mathrm{T}} \\ \sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \end{cases} \tag{27}$$

注意，$\boldsymbol{w}$ 为列向量，而 $\boldsymbol{x}^{(i)}$ 为行向量。

简化拉格朗日函数

将上 (27) 带入 (24)，消去式中 $\boldsymbol{w}$ 和 b：

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\lambda}) &= \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}{2} + \sum_{i=1}^{n} \lambda_i \left(1 - y^{(i)}\left(\boldsymbol{x}^{(i)}\boldsymbol{w} + b\right)\right) \\ &= \frac{\left(\sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)}\right)^{\mathrm{T}} \left(\sum_{j=1}^{n} \lambda_j y^{(j)} \boldsymbol{x}^{(j)}\right)}{2} + \sum_{i=1}^{n} \lambda_i \left(1 - y^{(i)} \left(\sum_{j=1}^{n} \lambda_j y^{(j)} \boldsymbol{x}^{(j)}\right) \cdot \boldsymbol{x}^{(i)} - y^{(i)} b\right) \\ &= \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}\right)}{2} - \sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}\right) + \sum_{i=1}^{n} \lambda_i - b \sum_{i=1}^{n} \lambda_i y^{(i)} \\ &= \sum_{i=1}^{n} \lambda_i - \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}\right)}{2} \end{aligned} \tag{28}$$

拉格朗日函数 $L(\boldsymbol{w}, b, \boldsymbol{\lambda})$ 简化为 $L(\boldsymbol{\lambda})$：

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}\right)}{2} \tag{29}$$

对偶问题

利用拉格朗日乘子法，这样便将 (23) 优化问题转化成一个以 $\boldsymbol{\lambda}$ 为变量的优化问题：

$$\begin{aligned} &\underset{\boldsymbol{\lambda}}{\arg\min} \quad \sum_{i=1}^{n} \lambda_i - \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \lambda_j y^{(i)} y^{(j)} \left(\boldsymbol{x}^{(i)} \cdot \boldsymbol{x}^{(j)}\right)}{2} \\ &\text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \lambda_i y^{(i)} = 0 \\ \lambda_i \geq 0, \quad i, j = 1, 2, 3, \dots, n \end{cases} \end{aligned} \tag{30}$$

这个优化问题常被称作**拉格朗日对偶问题** (Lagrange duality)，也称**对偶问题** (duality)。

发现二次型、格拉姆矩阵

大家是否发现 (29) 中的二次型？

对二次型陌生的读者，请回顾《矩阵力量》第 5 章。

举个例子，当 $n = 2$，即两个样本数据，(29) 可以展开为：

$$L(\boldsymbol{\lambda}) = (\lambda_1 + \lambda_2) - \frac{1}{2}\left(\lambda_1\lambda_1 y^{(1)} y^{(1)} \left(\boldsymbol{x}^{(1)} \cdot \boldsymbol{x}^{(1)}\right) + 2\lambda_1\lambda_2 y^{(1)} y^{(2)} \left(\boldsymbol{x}^{(1)} \cdot \boldsymbol{x}^{(2)}\right) + \lambda_2\lambda_2 y^{(2)} y^{(2)} \left(\boldsymbol{x}^{(2)} \cdot \boldsymbol{x}^{(2)}\right)\right) \tag{31}$$

(31) 整理为如下二次型：

$$L(\boldsymbol{\lambda}) = (\lambda_1 + \lambda_2) - \frac{1}{2}\begin{bmatrix} \lambda_1 y^{(1)} & \lambda_2 y^{(2)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{(1)} \cdot \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(1)} \cdot \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(2)} \cdot \boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} \cdot \boldsymbol{x}^{(2)} \end{bmatrix} \begin{bmatrix} \lambda_1 y^{(1)} \\ \lambda_2 y^{(2)} \end{bmatrix} \tag{32}$$

类似地，(29) 可以整理为：

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \begin{bmatrix} \lambda_1 y^{(1)} \\ \lambda_2 y^{(2)} \\ \vdots \\ \lambda_n y^{(n)} \end{bmatrix}^{\mathrm{T}} \underbrace{\begin{bmatrix} \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)} \rangle & \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle & \cdots & \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(n)} \rangle \\ \langle \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)} \rangle & \langle \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)} \rangle & \cdots & \langle \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(n)} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(1)} \rangle & \langle \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(2)} \rangle & \cdots & \langle \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n)} \rangle \end{bmatrix}}_{\text{Gram matrix}} \begin{bmatrix} \lambda_1 y^{(1)} \\ \lambda_2 y^{(2)} \\ \vdots \\ \lambda_n y^{(n)} \end{bmatrix} \tag{33}$$

相信大家已经在上式中看到了久违的**格拉姆矩阵** (Gram matrix)！

决策边界

利用 (27)，决策边界可以整理为：

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = \underbrace{\left(\sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)}\right)}_{\text{Coefficients}} \boldsymbol{x} + b = 0 \tag{34}$$

需要大家注意区分，行向量 $\boldsymbol{x}^{(i)}$ 为第 i 个数据点，$\boldsymbol{x}$ 为未知量构成的列向量。也就是说，$\sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)}$ 求和结果为行向量。

分类决策函数 $p(\boldsymbol{x})$ 则为：

$$p(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right) = \operatorname{sign}\left(\underbrace{\left(\sum_{i=1}^{n} \lambda_i y^{(i)} \boldsymbol{x}^{(i)}\right)}_{\text{Coefficients}} \boldsymbol{x} + b\right) \tag{35}$$

7.4 支持向量机处理二分类问题

本节利用具体实例介绍如何实现硬间隔支持向量机算法。

实例

图 9 所示为 20 个样本数据，容易发现样本数据线性可分，下面利用支持向量机进行预测分类。

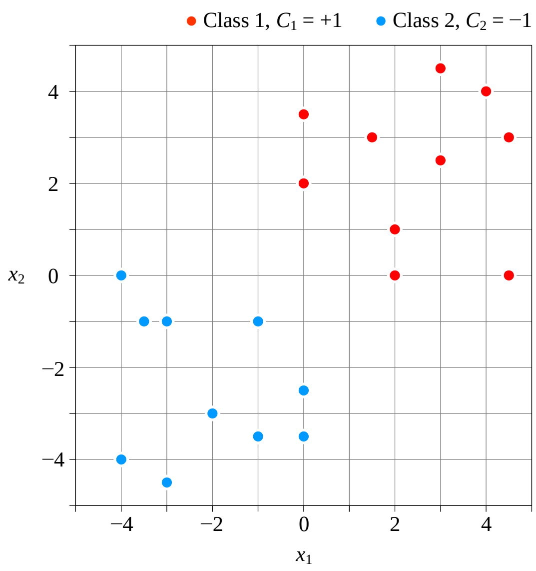

图 9. 20 个样本数据点平面位置

决策边界

对于 $D = 2$ 的情况，将 (1) 展开：

$$w_1 x_1 + w_2 x_2 + b = 0 \tag{36}$$

w_2 不等于 0 时，将 (36) 写成大家熟悉的一次函数形式：

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2} \tag{37}$$

硬间隔

根据 (20)，硬间隔“上边界”l_1 对应的函数为：

$$w_1 x_1 + w_2 x_2 + b = 1 \quad \Rightarrow \quad x_2 = -\frac{w_1}{w_2} x_1 - \frac{b-1}{w_2} \tag{38}$$

间隔“下边界” l_2 对应的函数为：

$$w_1 x_1 + w_2 x_2 + b = -1 \quad \Rightarrow \quad x_2 = -\frac{w_1}{w_2} x_1 - \frac{b+1}{w_2} \tag{39}$$

再次注意，因为 (37) 中 w_2 不能为 0，因此 (37) 存在局限性。这种表达方式仅为方便大家理解。

分类结果

图 10 为分类结果。容易发现，一共存在三个支持向量——A (0, 2)、B (2, 0) 和 C (−1, −1)。剩余 17 个样本数据对决策边界没有丝毫影响。

图 10 中深蓝色直线为决策边界，对应解析式：

$$\frac{x_1}{2} + \frac{x_2}{2} = 0 \quad \Rightarrow \quad x_1 + x_2 = 0 \quad \Rightarrow \quad x_2 = -x_1 \tag{40}$$

分类决策函数 $p(\boldsymbol{x})$ 为：

$$p(x_1, x_2) = \operatorname{sign}(x_1 + x_2) \tag{41}$$

间隔“上”边界 l_1 对应的函数为：

$$\frac{x_1}{2} + \frac{x_2}{2} = 1 \quad \Rightarrow \quad x_2 = -x_1 + 2 \tag{42}$$

间隔“下”边界 l_2 对应的函数为：

$$\frac{x_1}{2}+\frac{x_2}{2}=-1 \quad \Rightarrow \quad x_2=-x_1-2 \tag{43}$$

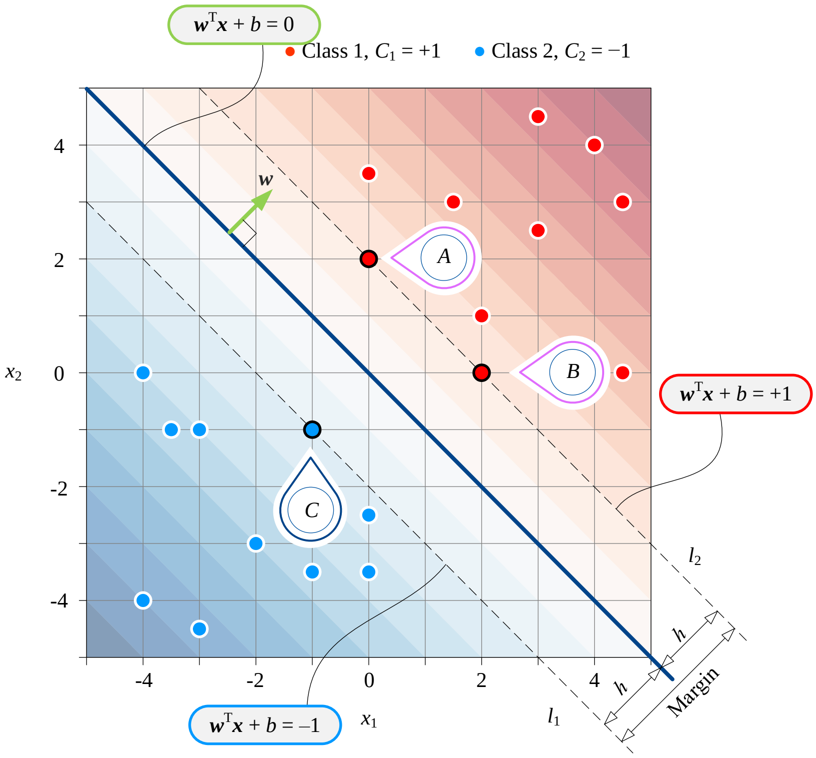

图 10. 硬间隔分类结果

预测分类

将 (4, 4) 代入 (41)，可以判断 (4, 4) 的预测分类为+1：

$$p(4,4)=\operatorname{sign}(4+4)=+1 \tag{44}$$

将 (−2, −3) 代入 (41)，可以判断 (−2, −3) 的预测分类为−1：

$$p(-2,-3)=\operatorname{sign}(-2-3)=-1 \tag{45}$$

将 (3, −3) 代入 (41)，结果为 0，可以判断 (3, −3) 位于决策边界上：

$$p(3,-3)=\operatorname{sign}(3-3)=0 \tag{46}$$

支持向量影响决策边界

图 11 所示为删除点 A 后，支持向量变化，以及决策边界和间隔位置。再次强调，支持向量算法中，除支持向量之外的样本数据对决策边界没有影响。

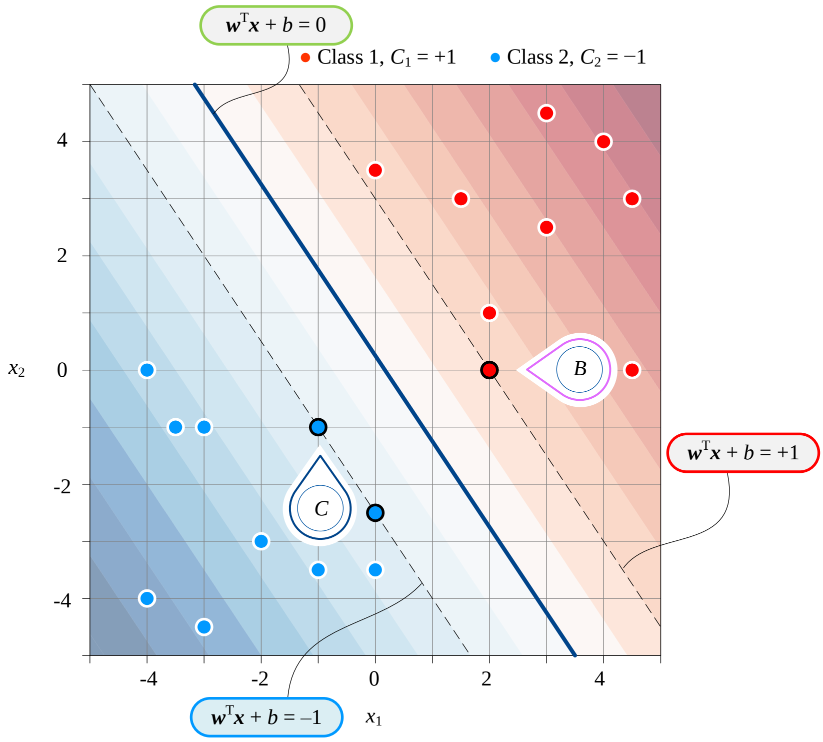

图 11. 删除点 A 后硬间隔 SVM 分类结果

7.5 软间隔：处理线性不可分

本章第一节提到，支持向量机可以采用**软间隔** (soft margin) 处理**线性不可分** (non-linearly separable data)。白话说，**硬间隔** (hard margin) 处理“泾渭分明”的分类数据，一条直线将样本数据彻底分离，如图 12 (a) 所示。而软间隔处理的数据呈现“你中有我，我中有你”，如图 12 (b) 所示。

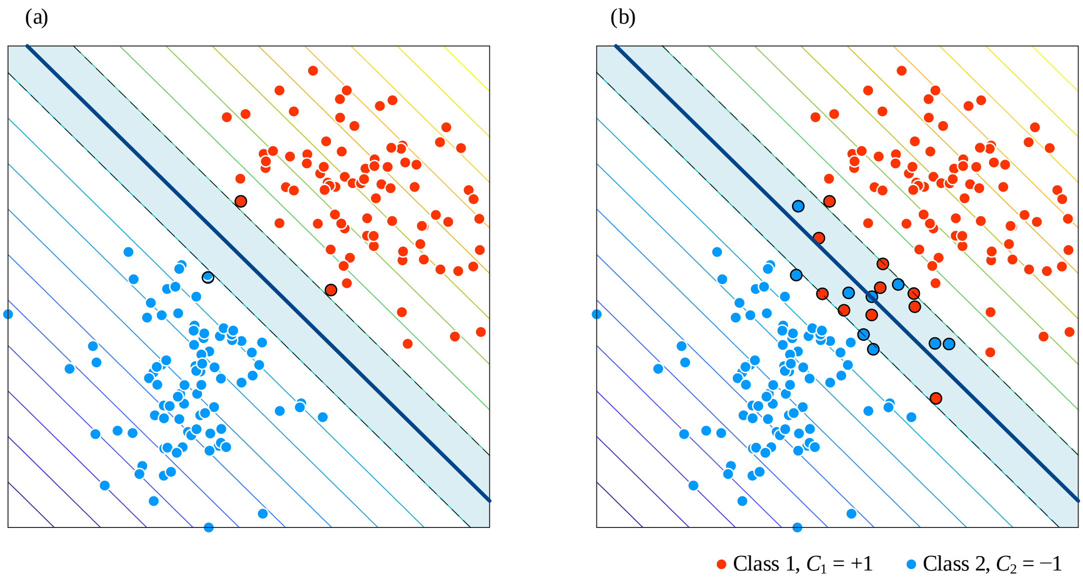

图 12. 比较硬间隔和软间隔

软间隔 SVM 方法的核心思想是牺牲部分数据点分类准确性，来换取更宽的间隔。

软间隔有两个重要参数：

- **松弛变量** (slack variable) ξ，一般读作 /ksaɪ/
- **惩罚因子** (penalty parameter) C

松弛变量

松弛变量用来模糊间隔边界，图 13 所示为原理图。

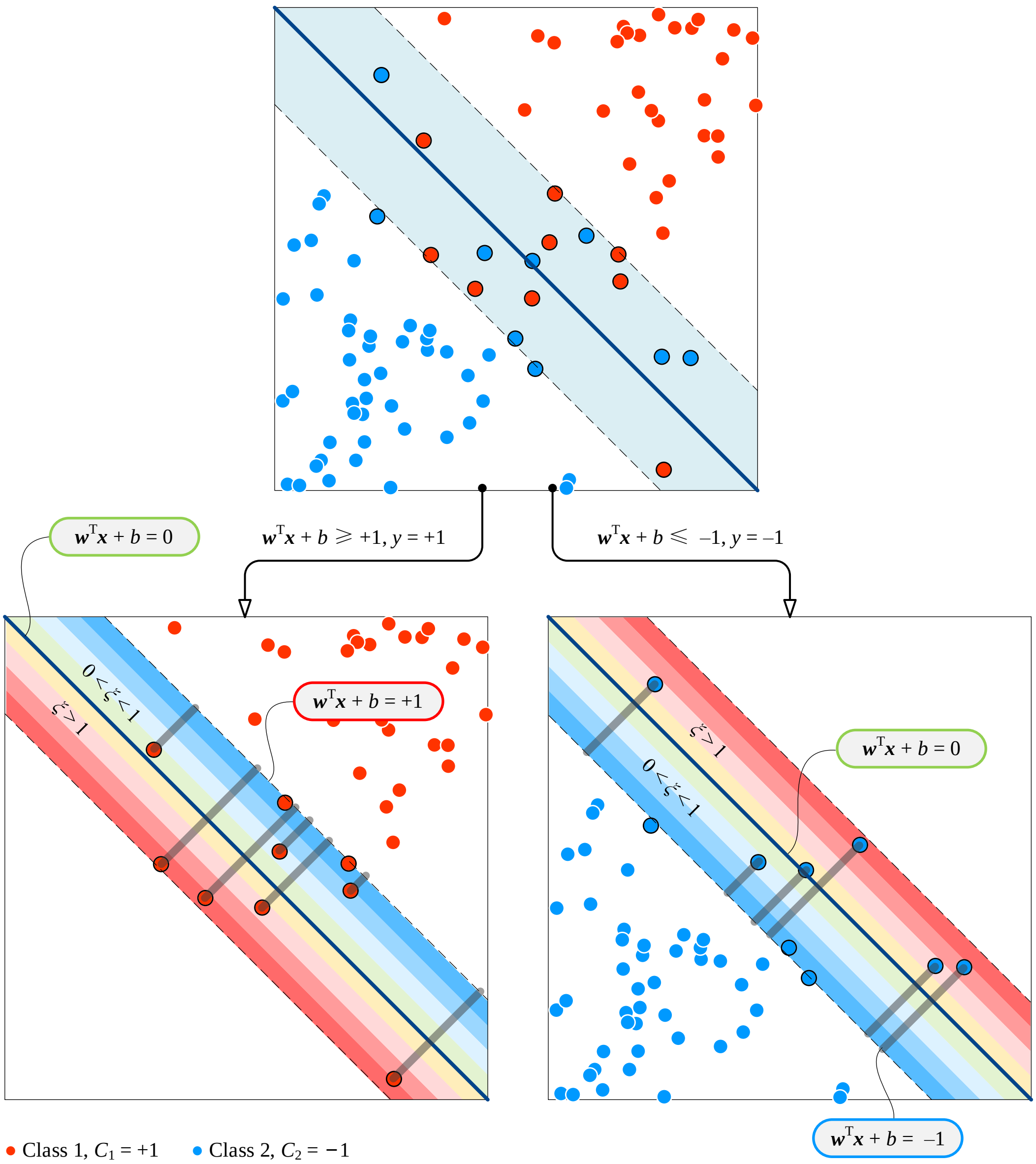

图 13. 软间隔中松弛变量作用

引入松弛变量 ξ，(19) 被改造为：

$$\left(\boldsymbol{w}\cdot\boldsymbol{x}+b\right)y \geq 1-\xi \tag{47}$$

当 $y = +1$，

$$\left(\boldsymbol{w}\cdot\boldsymbol{x}+b\right) \geq 1-\xi \tag{48}$$

当 $y = -1$

$$\left(\boldsymbol{w}\cdot\boldsymbol{x}+b\right) \leq -1+\xi \tag{49}$$

如图 13 所示，当 $\xi = 0$，样本数据位于正确分类区域内或正确间隔边界上；当 $\xi > 0$，样本数据位于软间隔范围之内，甚至在错误的分类区域内。图 13 中，红色带对应松弛变量 ξ 较大区域，蓝色带对应松弛变量 ξ 较小区域。

图 13 中，软间隔内任一数据点 $\boldsymbol{x}^{(i)}$ 距离各自边界距离为：

$$d_i = \frac{\xi_i}{\|\boldsymbol{w}\|} \tag{50}$$

优化问题

下面，在 (23) 基础上引入惩罚因子 C，构造软间隔 SVM 优化问题：

$$\begin{aligned} &\underset{\boldsymbol{w},b,\xi}{\arg\min}\ \frac{\boldsymbol{w}\cdot\boldsymbol{w}}{2}+C\sum_{i=1}^{n}\xi_i \\ &\text{subject to}\ \begin{cases} y^{(i)}\left(\boldsymbol{x}^{(i)}\boldsymbol{w}+b\right) \geq 1-\xi_i, & i=1,2,3,\dots,n \\ \xi_i \geq 0 \end{cases} \end{aligned} \tag{51}$$

惩罚因子 C 为用户设定参数，它调整松弛变量惩罚项的影响力。C 较大时，优化问题更在意分类准确性，牺牲间隔宽度；间隔可以窄一些，分类错误少犯一些。C 取值较小时，间隔更宽一些，间隔内的样本数据较多，分类错误可以多一点。

也可以采用 L^2 范数来构造松弛变量惩罚项，此时 (51) 被改造成：

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\arg\min}\ \frac{\boldsymbol{w}\cdot\boldsymbol{w}}{2}+C\sum_{i=1}^{n}\xi_i^2 \\ &\text{subject to}\ \begin{cases} y^{(i)}\left(\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}+b\right) \geq 1-\xi_i, & i=1,2,3,\dots,n \\ \xi_i \geq 0 \end{cases} \end{aligned} \tag{52}$$

惩罚因子影响分类结果

图 14 所示为惩罚因子 C 取不同值时，支持变量、决策边界和间隔宽度变化。

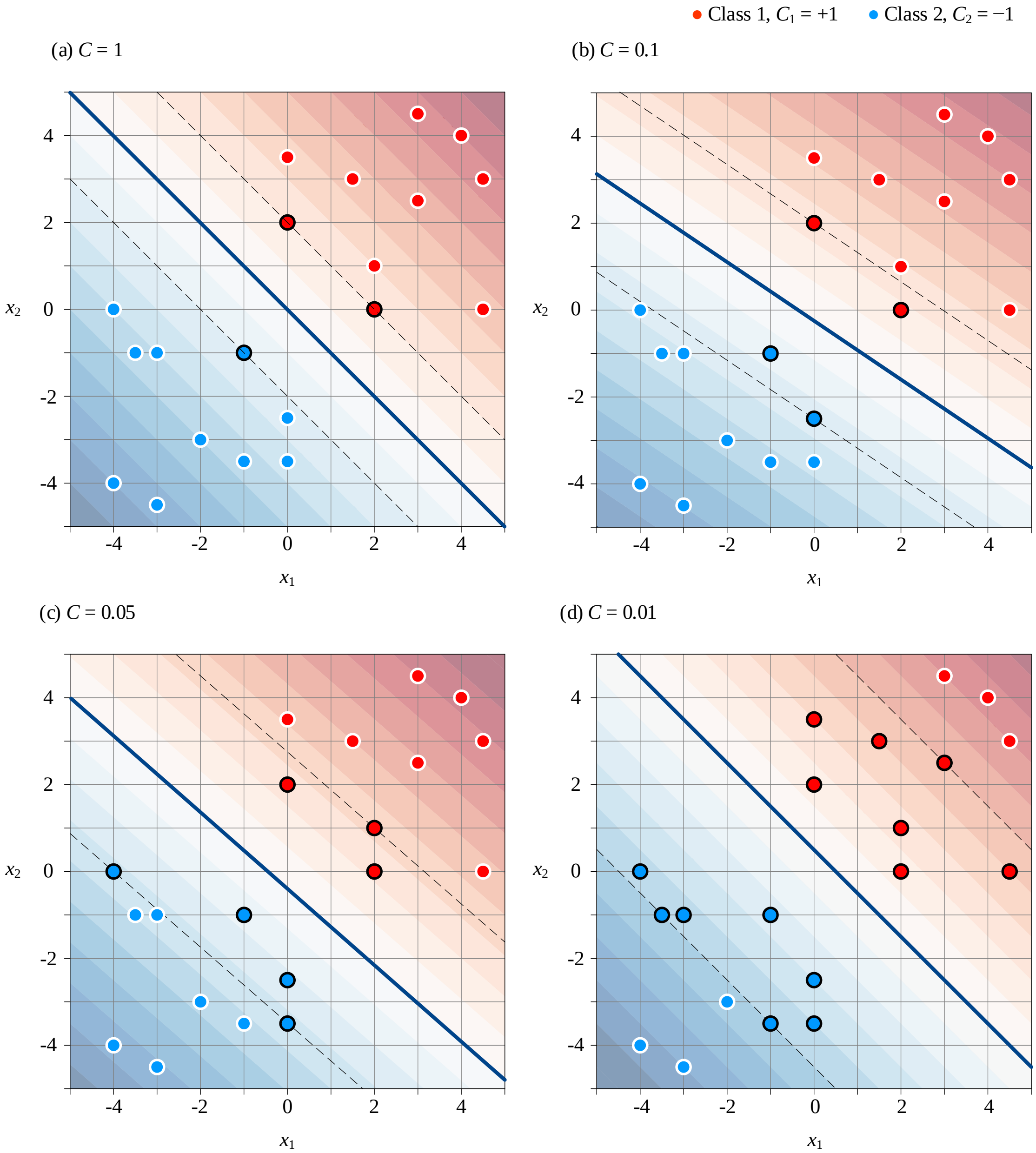

图 14. 惩罚因子对软间隔宽度和决策边界影响

代码 Bk7_Ch07_01.py 利用 SVM 实现分类，并绘制图 10、图 11 和图 14。