华中科技大学
电子信息与通信学院
School of Electronic Information and Communications

机器学习导论

– Introduction to Machine Learning

第 05 章：支持向量机

(Chapter 05: Support Vector Machine)

华中科技大学电信学院

王邦　博士, 教授博导

wangbang@hust.edu.cn

第 05 章：支持向量机

问题引入 – 重新审视 logistic 回归 (1)

logistic 回归将特征的线性组合作为自变量，通过 Sigmoid 函数，将自变量映射到 (0,1) 区间上，作为分类为正例的概率值。

- 假设函数

$$h(\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b)}}. \tag{1}$$

$$y = \begin{cases} 1 & h(\boldsymbol{x}) \geq 0.5 \\ 0 & h(\boldsymbol{x}) < 0.5 \end{cases} \tag{2}$$

  y 为预测标签值。

- 可见，$h(x)$ 值或者说分类结果只和 $\boldsymbol{w}^T\boldsymbol{x} + b$ 有关，由此得到决策边界的概念：

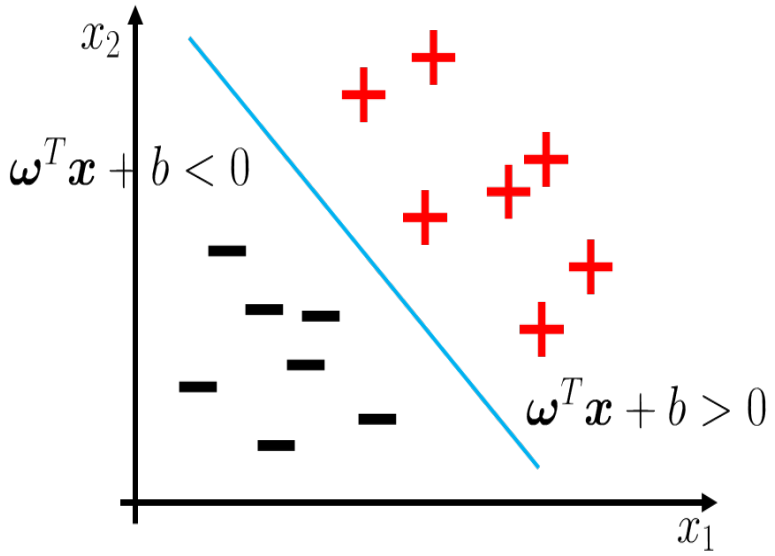
$$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0. \tag{3}$$

问题引入 -重新审视 logistic 回归 (2)

- 从决策边界的角度来说，我们希望分类器到达的效果：

$$\text{标签 } y = 1 \text{ 时，} \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b > 0$$
$$\text{标签 } y = 0 \text{ 时，} \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b < 0$$

- 几何直观上, 样本点离决策边界的间隔越大越好

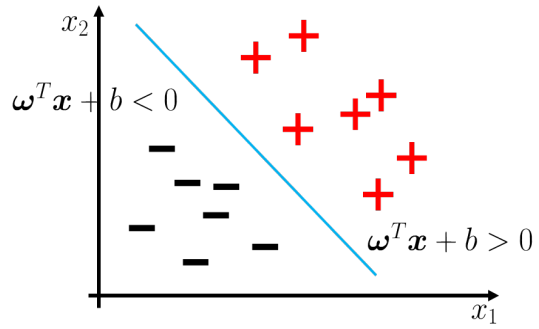


本章从间隔的角度重新去考虑分类问题。

基本概念 (1)

支持向量

从几何的观点来看，决策边界是一个**划分超平面**（二维空间中表现为直线，以下讨论限于线性可分的情况）。



在样本空间中，**划分超平面**定义为：$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$

- $\boldsymbol{w}$ 为 **法向量，决定超平面的方向。**
- b 为**位移项，决定超平面与原点之间的距离。**

在 SVM 中，定义：

$$\begin{cases} \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \geq +1 & y = +1 \\ \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \leq -1 & y = -1 \end{cases} \tag{4}$$

说明

+1 代表正例，-1 代表反例

$\geq +1 (\leq -1)$ 代表更高的分类确信度 (不仅仅满足于只大于 (小于)0)，即有更高的概率是正例 (反例))。

使上式等号成立的样本点称作**支持向量**（support vector）

基本概念 (2)

间隔

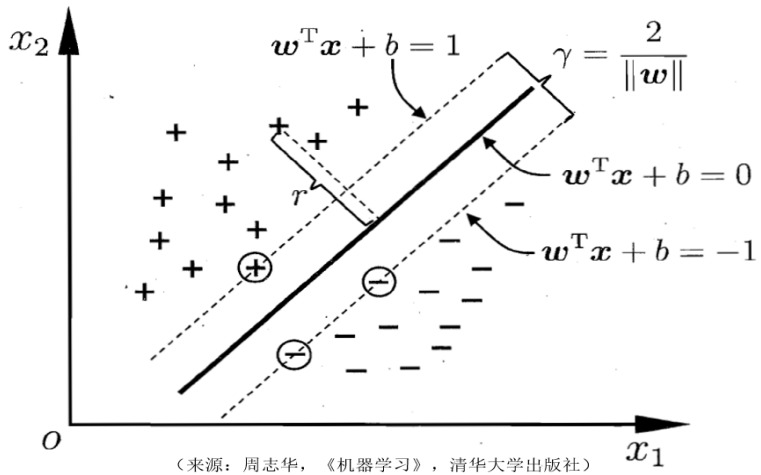
- 任一样本点到划分超平面的距离： $r = \frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{||\boldsymbol{w}||}$
- 由支持向量使式 (4) 等号成立可知，对于支持向量：

$$|\boldsymbol{w}^T\boldsymbol{x}+b| = 1$$

则两个**异类**支持向量到超平面的距离之和为：

$$\gamma = \frac{2}{||\boldsymbol{w}||} \tag{5}$$

称为**间隔** (margin)



（来源：周志华，《机器学习》，清华大学出版社）

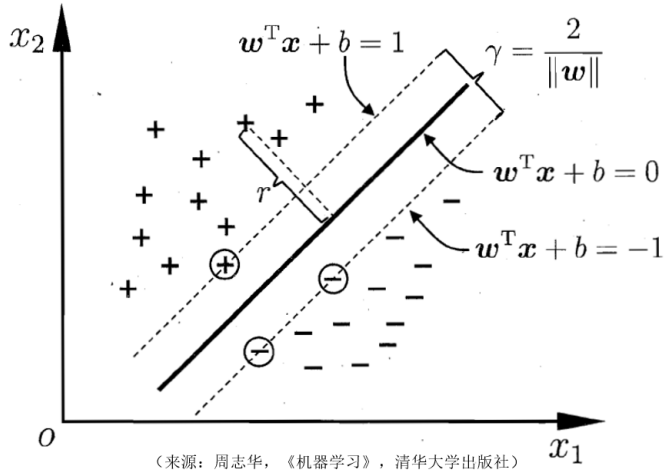
模型优化与方法 (1)

优化目标

我们希望分类器具有以下效果：

标签 $y = +1$ 时，$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \geq +1$

标签 $y = -1$ 时，$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \leq -1$



（来源：周志华，《机器学习》，清华大学出版社）

优化目标： 从间隔的角度考虑，希望获得**具有最大间隔的划分超平面**。寻找满足式 (4) 中约束的参数 $\boldsymbol{w}, b$，使得间隔 γ (5)最大：

$$\begin{aligned} &\max_{\boldsymbol{w},b} \quad \frac{2}{||\boldsymbol{w}||} \\ &s.t. \quad y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) \geq 1, i = 1, 2, ..., m \end{aligned} \tag{6}$$

模型优化与方法 (2)

SVM 模型

- 优化目标：$\max\limits_{\boldsymbol{w},b} \frac{2}{||\boldsymbol{w}||}$ s.t. $y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) \geq 1, i = 1, 2, ..., m$
- 显然，为了最大化间隔，仅需最大化 $||\boldsymbol{w}||^{-1}$，等价于最小化 $||\boldsymbol{w}||^2$，优化目标可重写为：

$$\begin{aligned} &\min_{\boldsymbol{w},b} \quad \frac{1}{2}||\boldsymbol{w}||^2 \\ &s.t. \quad y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) \geq 1, i = 1, 2, ..., m \end{aligned} \tag{7}$$

- 式 (7) 就是**支持向量机（SVM）的基本型**。
- 我们希望求解式 (7) 来得到大间隔划分超平面所对应的模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b \tag{8}$$

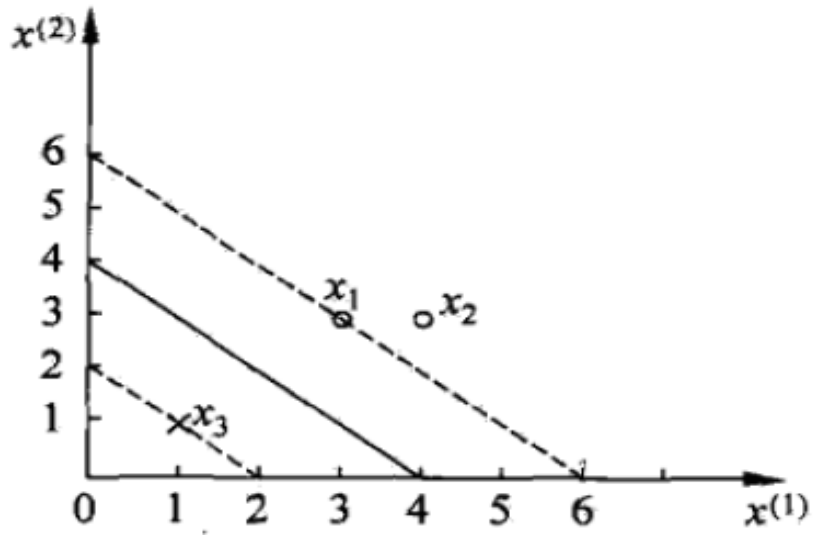
这是一个**凸二次规划问题**，能直接用现成的优化计算包求解。

模型优化与方法 (3)

例子：已知一个数据集的样本为 $(\boldsymbol{x}_i, y_i) = ((x_1^{(1)}, x_2^{(2)})^{\mathsf{T}}, \pm 1)$。其中**正样本**是 $\boldsymbol{x}_1 = (3,3)^{\mathsf{T}}$，$\boldsymbol{x}_2 = (4,3)^{\mathsf{T}}$，**负样本** $\boldsymbol{x}_3 = (1,1)^{T}$。求最大间隔分离超平面 $\boldsymbol{w}\boldsymbol{x}^{\mathsf{T}} + b$，其中 $\boldsymbol{w} = (w_1, w_2)$。

解： 根据 SVM 的基本型 (7)，构造约束最优化问题：

$$\begin{aligned} \min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 &= \min_{\boldsymbol{w},b} \frac{1}{2}(w_1^2 + w_2^2) \\ s.t.\ \ 3w_1 + 3w_2 + b &\geq 1 \\ 4w_1 + 3w_2 + b &\geq 1 \\ -w_1 - w_2 - b &\geq 1 \end{aligned}$$



求得此最优化问题的解 $w_1 = w_2 = \frac{1}{2}$，$b = -2$，所以最大间隔分离超平面为

$$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$$

其中，$\boldsymbol{x}_1 = (3,3)^{\mathsf{T}}$ 与 $\boldsymbol{x}_3 = (1,1)^{\mathsf{T}}$ 为支持向量，如图所示。

拉格朗日对偶问题 (1)

拉格朗日对偶问题：

- 使用拉格朗日乘子法，对式 (7) 每条约束添加**拉格朗日乘子** $\alpha_i \geq 0$，可得其“**对偶问题**” (dual problem)。SVM 基本型的拉格朗日方程可写为

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)), \tag{9}$$

  其中 $\boldsymbol{\alpha} = (\alpha_1; \alpha_2; ...; \alpha_m)$。

- 令 L 对 $\boldsymbol{w}$ 和 b 的偏导为 0 可得:

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{10}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{11}$$

拉格朗日对偶问题 (2)

拉格朗日对偶问题： 代入可得到式 (7) 的对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j \\ s.t. & \sum_{i=1}^{m} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, i = 1, 2, ..., m. \end{aligned} \tag{12}$$

- 原来的 SVM 基本型最小化问题 (7) 变换成拉格朗日对偶问题 (12) 的最大化问题。
- 拉格朗日对偶问题 (12) 仅涉及拉格朗日乘子和训练数据；而原问题 (7) 除涉及拉格朗日乘子外还涉及决策边界的参数。尽管如此，这两个优化问题是等价的。

拉格朗日对偶问题 (3)

拉格朗日对偶问题： 解出 $\boldsymbol{\alpha}$ 后，划分超平面方程可写为

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x} + b. \tag{13}$$

因此，**通过对偶问题解出拉格朗日乘子向量 $\boldsymbol{\alpha}$，即可求得超平面**。由于式 (7) 中有不等式约束，因此上述过程需满足 KKT（Karush-Kuhn-Tucker）条件，即：

$$\alpha_i \geq 0; \tag{14}$$

$$y_i f(\boldsymbol{x}_i) - 1 \geq 0; \tag{15}$$

$$\alpha_i (y_i f(\boldsymbol{x}_i) - 1) \geq 0 \tag{16}$$

以上约束条件说明，对于任意训练样本 $(\boldsymbol{x}_i, y_i)$，总有 $\alpha_i = 0$ 或 $y_i f(\boldsymbol{x}_i) = 1$.

- 若 $\alpha_i = 0$，则说明该样本不会在划分超平面的求和式中出现（即不会对 $f(\boldsymbol{x})$ 产生影响）；
- 若 $\alpha_i > 0$，则 $y_i f(\boldsymbol{x}_i) = 1$，对应的样本点为支持向量.

所以，支持向量机的模型训练完成后，大部分训练样本都不需保存，最终模型仅与支持向量有关。

拉格朗日对偶问题 (4)

拉格朗日对偶问题： 拉格朗日对偶问题式 (12) 解出 $\boldsymbol{\alpha}^*$ 后，按照式 (10) 求取 $\boldsymbol{w}^*$，即

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \boldsymbol{x}_i \tag{17}$$

b^* 可以通过求解支持向量公式 (16) 得到，即：

$$\alpha_i^* [y_i(\boldsymbol{w}^* \cdot \boldsymbol{x}_i + b^*) - 1] = 0, \tag{18}$$

上式中 $\boldsymbol{x}_i$ 为支持向量，α_i^* 为其对应的最优拉格朗日乘子。

由于可能存在多个支持向量，且 α_i^* 是通过数值计算得到的，因此可能存在数值误差，计算出的 b^* 值可能不唯一，取决于 (16) 中使用的支持向量。实践中，通常使用求得的平均值 b^* 作为决策边界的参数。

最后求得的划分超平面方程可写为

$$f(\boldsymbol{x}) = {\boldsymbol{w}^*}^{\mathsf{T}} \boldsymbol{x} + b^* \tag{19}$$

拉格朗日对偶问题 (5)

例子： 问题同前。已知一个数据集的样本为 $(\boldsymbol{x}_i, y_i) = ((x_1^{(1)}, x_2^{(2)})^{\mathsf{T}}, \pm 1)$。其中**正样本**是 $\boldsymbol{x}_1 = (3,3)^{\mathsf{T}}$，$\boldsymbol{x}_2 = (4,3)^{\mathsf{T}}$，**负样本** $\boldsymbol{x}_3 = (1,1)^{T}$。求最大间隔分离超平面 $\boldsymbol{w}\boldsymbol{x}^{\mathsf{T}} + b$，其中 $\boldsymbol{w} = (w_1, w_2)$。

解： 根据式 (12)，对偶问题是

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{3} \alpha_i - \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j \\
&= \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3) \\
&s.t. \quad \alpha_1 + \alpha_2 - \alpha_3 = 0, \\
&\qquad \alpha_i \geq 0, i = 1,2,3.
\end{aligned}$$

解这一最优化问题。将 $\alpha_3 = \alpha_1 + \alpha_2$ 代入上式并记为

$$s(\alpha_1, \alpha_2) = -4\alpha_1^2 - \frac{13}{2}\alpha_2^2 - 10\alpha_1\alpha_2 + 2\alpha_1 + 2\alpha 2$$

拉格朗日对偶问题 (6)

对 α_1，α_2 求偏导数并令其为 0，易知 $s(\alpha_1, \alpha_2)$ 在点 $(\frac{3}{2}, -1)^T$ 取极值，但该点不满足约束条件 $\alpha_2 \geq 0$，所以最大值应在边界上达到。

- 当 $\alpha_1 = 0$ 时，最大值 $s(0, \frac{2}{13}) = \frac{2}{13}$；
- 当 $\alpha_2 = 0$ 时，最大值 $s(\frac{1}{4}, 0) = \frac{1}{4}$。

于是 $s(\alpha_1, \alpha_2)$ 在 $\alpha_1 = \frac{1}{4}, \alpha_2 = 0$ 达到最大，此时 $\alpha_3 = \alpha_1 + \alpha_2 = \frac{1}{4}$.
这样，$\alpha_1 = \alpha_3 = \frac{1}{4}$ 对应的实例点 $\boldsymbol{x}_1$，$\boldsymbol{x}_3$ 是支持向量。接下来计算 $\boldsymbol{w}$ 和 b

$$\boldsymbol{w} = \sum_{i=1}^{3} \alpha_i y_i \boldsymbol{x}_i = \frac{1}{4} \times 1 \times (3,3)^T + 0 + \frac{1}{4} \times (-1) \times (1,1) = (0.5, 0.5)$$

$$b_1 = 1 - \boldsymbol{w} \cdot \boldsymbol{x}_1 = 1 - (0.5, 0.5) \cdot \begin{pmatrix} 3 \\ 3 \end{pmatrix} = -2$$

$$b_3 = -1 - \boldsymbol{w} \cdot \boldsymbol{x}_3 = -1 - (0.5, 0.5) \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = -2$$

$$b = (b_1 + b_3)/2 = -2$$

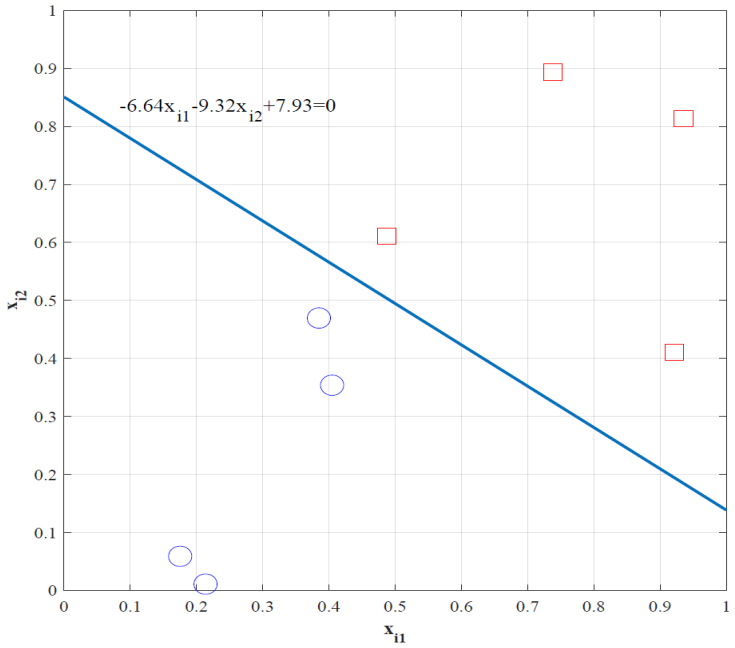
最后求得划分超平面为：

$$\frac{1}{2} x^{(1)} + \frac{1}{2} x^{(2)} - 2 = 0$$

拉格朗日对偶问题 (7)

例子： 考虑下图所给的二维数据集，它包含 8 个训练实例。使用二次规划的方法，可以求解公式 (12) 给出的优化问题，得到每一个训练实例的拉格朗日乘子 α_i，如下表的最后一列所示。

x_{i1}	x_{i2}	y_i	拉格朗日乘子
0.3858	0.4687	1	65.5261
0.4871	0.6110	−1	65.5261
0.9218	0.4103	−1	0
0.7382	0.8936	−1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	−1	0
0.2146	0.0099	1	0



仅前两个实例具有非零的拉格朗日乘子，这两个实例对应于该数据集的支撑向量。

拉格朗日对偶问题 (8)

- 令 $\boldsymbol{w} = (w_1, w_2)$，$b$ 为决策边界的参数，则

$$w_1 = \sum_{i=1}^{8} \alpha_i y_i x_{i1} = 65.5261 \times 1 \times 0.3858 + 65.5261 \times (-1) \times 0.4871 = -6.64$$

$$w_2 = \sum_{i=1}^{8} \alpha_i y_i x_{i2} = 65.5261 \times 1 \times 0.4687 + 65.5261 \times (-1) \times 0.6110 = -9.32$$

- 位移项 b 可以用每个支持向量进行计算：

$$b_1 = 1 - \boldsymbol{w} \cdot \boldsymbol{x}_1 = 1 - (-6.64)(0.3858) - (-9.32)(0.4687) = 7.9300$$

$$b_2 = -1 - \boldsymbol{w} \cdot \boldsymbol{x}_2 = -1 - (-6.64)(0.4871) - (-9.32)(0.611) = 7.9289$$

  对这些值取平均，得到 $b = 7.93$。

- 所以对应的划分超平面 $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$ 为

$$-6.64x_{i1} - 9.32x_{i2} + 7.93 = 0$$

SMO 算法

式 (12) 的拉格朗日对偶问题式是一个二次规划问题，其问题规模正比于训练样本数，在实际任务中开销过多。

SMO (Sequential Minimal Optimization) 是减少计算开销的高效算法。

SMO 算法基本思想：

- 选取一对变量 α_i 和 α_j 。
- 固定其他变量，求解上述优化目标得到更新后的 α_i 和 α_j。
- 重复上述两步骤直至收敛。

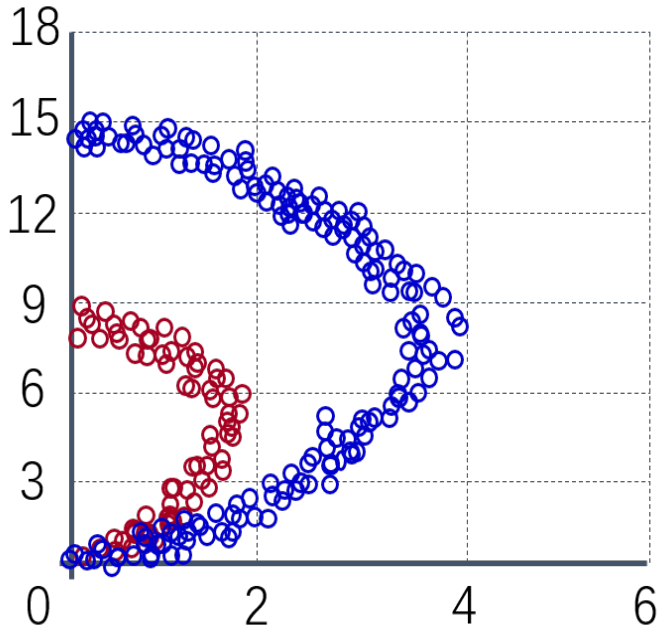
SMO 采用启发式方法选择 α_i 和 α_j。

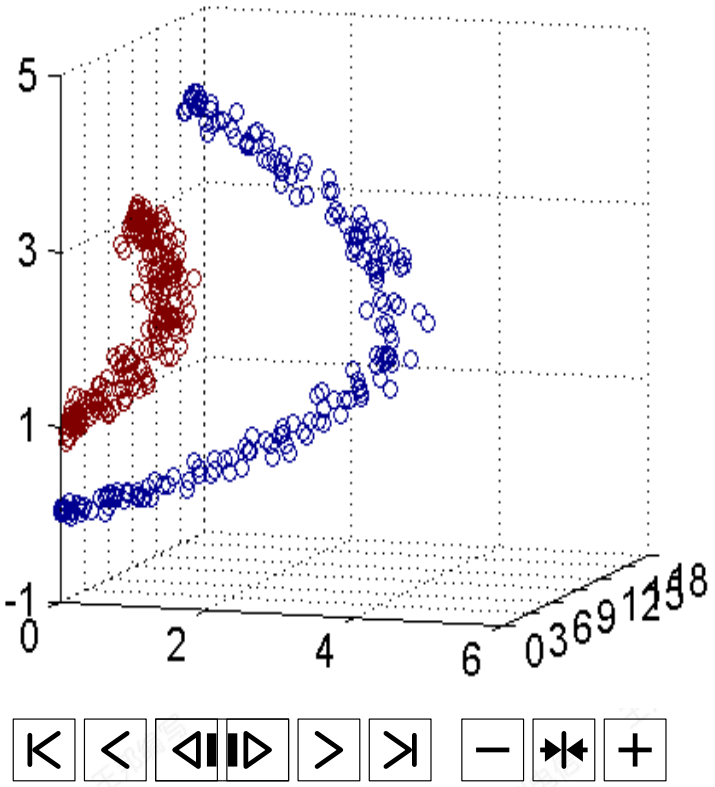
- 第一个参数 α_1 的选择：先从**所有样本**中选择一个违反 KKT 条件的参数，再选择第二个参数，一轮迭代更新后，再从**非边界样本点**中选择一个违反 KKT 条件的参数，再选择第二个参数。
- 第二个参数 α_2 的选择：从剩下的 (除已被选择的第一个参数外) 的参数中选择 $|E_1, E_2|$ 最大的作为第二个参数。

核函数 (1)

线性可分： 在前面的讨论中，我们假设样本点线性可分。实际问题中，这一假设并不总是成立。

样本线性不可分： 利用**核函数** (Kernel Function) 将原始样本点**映射到一个高维空间**，从而使其线性可分。





如果原始空间是有限维，则一定存在一个高维特征空间使样本线性可分。

核函数 (2)

- **什么是核函数？**

  令 $\phi(\boldsymbol{x})$ 表示样本 $\boldsymbol{x}$ 映射后的特征向量，则核函数满足下式

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \phi(\boldsymbol{x}_i)^{\mathsf{T}} \phi(\boldsymbol{x}_j) \tag{20}$$

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

- **为什么是核函数（而不是其它映射函数）？**

  映射后的特征空间**维数可能很高**，直接计算 $\phi(\boldsymbol{x}_i)^{\mathsf{T}} \phi(\boldsymbol{x}_j)$，通常都是困难的，为了避开这个障碍，所以选用核函数。（这称为“核技巧”（kernel trick））

核函数 (3)

核函数情况下的 **SVM** 模型：

优化目标：

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \\ &\alpha_i \geq 0, i = 1, 2, ..., m. \end{aligned} \tag{21}$$

解出 $\boldsymbol{\alpha}$ 后及相应的支持向量后，得到划分超平面的核函数展开式，亦称“支持向量展开式”。

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{\omega}^{\mathsf{T}} \phi(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{m} \alpha_i y_i \phi(\boldsymbol{x}_i)^{\mathsf{T}} \phi(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{m} \alpha_i y_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b. \end{aligned} \tag{22}$$

核函数 (4)

定理 (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 κ 是核函数当且仅当对于任意数据 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_m\}$，“核矩阵” (kernel matrix) $\boldsymbol{K}$ 总是半正定的：

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

上述定理表明，只要一个对称函数所对应的核矩阵半正定，就能作为和函数使用。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。

核函数 (5)

常用核函数

- 在不知道特征映射的形式时，我们并不知道什么样的核函数是合适的，而核函数也仅是隐式地定义了这个特征空间。**“核函数选择”成为支持向量机的最大变数。**
- 几种常用的核函数：

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

- 选用核函数的一些基本经验：文本数据常用线性核，未知分布的数据可先用高斯核。

核函数 (6)

核函数组合

- 核函数也可由函数组合得到，设 $\kappa_1\kappa_2$ 均为核函数，则下列组合也是核函数：

1. 对于任意的正数 γ_1 和 γ_2

$$\gamma_1\kappa_1 + \gamma_2\kappa_2$$

2. 核函数的直积

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

3. 对于任意函数 $g(\boldsymbol{x})$

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

核函数 (7)

例子： 现有 5 个一维数据 (x, y)，如图所示。

$(x_1 = 1; y_1 = +1)$，$(x_2 = 2; y_2 = +1)$，

$(x_3 = 4; y_3 = -1)$，$(x_4 = 5; y_4 = -1)$，

$(x_5 = 6; y_5 = +1)$。

解：选择如下的二次项核函数：$\kappa(x_i, x_j) = (x_i x_j + 1)^2$，设置 $C = 100$。从下式中求解出 $\alpha_i, i = 1, 2, ..., 5$。

$$\max \quad \sum_{i=1}^{5} \alpha_i - \frac{1}{2} \sum_{i=1}^{5} \sum_{j=1}^{5} \alpha_i \alpha_j y_i y_j (x_i x_j + 1)^2$$

$$s.t. \quad \sum_{i=1}^{5} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq 100.$$

通过二次规划求解，得到：$\alpha_1 = 0,\ \alpha_2 = 2.5,\ \alpha_3 = 0,\ \alpha_4 = 7.333,\ \alpha_5 = 4.833$。

支持向量为：$\{x_2 = 4, x_4 = 5, x_5 = 6\}$

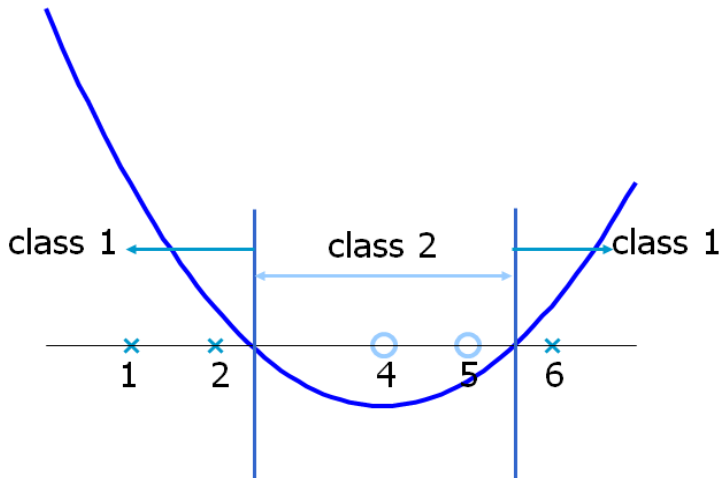
核函数 (8)

根据支持向量展开式 (22)，得到：

$$
\begin{aligned}
f(x) &= \sum_{i=1}^{5} \alpha_i y_i \kappa(x, x_i) + b = \sum_{i=1}^{5} \alpha_i y_i (x x_i + 1)^2 + b \\
&= 2.5 y_2 (x_2 x + 1)^2 + 7.333 y_4 (x_4 x + 1)^2 + 4.833 y_5 (x_5 x + 1)^2 + b \\
&= 0.6667 x^2 - 5.333 x + b.
\end{aligned}
$$

b 满足 $f(x_2) = 1, f(x_4) = -1, f(x_5) = 1$，得到 $b = 9$，故最后得到

$$
f(x) = 0.6667 x^2 - 5.333 x + 9
$$



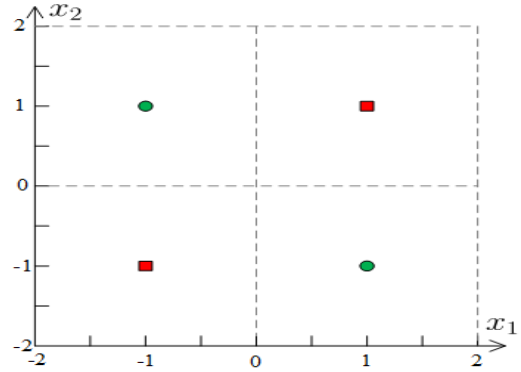
SVM 求解 XOR 问题 (1)

XOR 问题： 用 SVM 处理 XOR(异或) 问题。

4 个样本和期望的响应如下所示：

输入向量 $\boldsymbol{x}$	期望的响应 y
$(-1,-1)$	-1
$(-1,1)$	$+1$
$(1,-1)$	$+1$
$(1,1)$	-1

样本**线性不可分**！



需要往高维特征空间进行映射，使其线性可分！将一个样本 $\boldsymbol{x}_i$ 表示为 $\boldsymbol{x}_i = (x_{i1}, x_{i2})$，取以下**核函数**进行映射：

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j)^2 \tag{23}$$

展开为：$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 + x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}$.

例如：若 $\boldsymbol{x}_i = (1,1)$，$\boldsymbol{x}_j = (-1,1)$，则

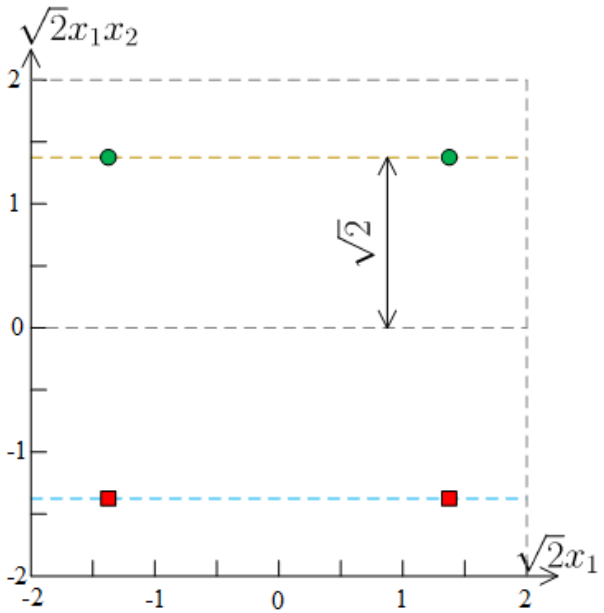
$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 + 1^2 \times (-1)^2 + 2 \times 1 \times 1 \times (-1) \times 1 + 1^2 \times 1^2 + 2 \times 1 \times (-1) + 2 \times 1 \times 1 = 1.$

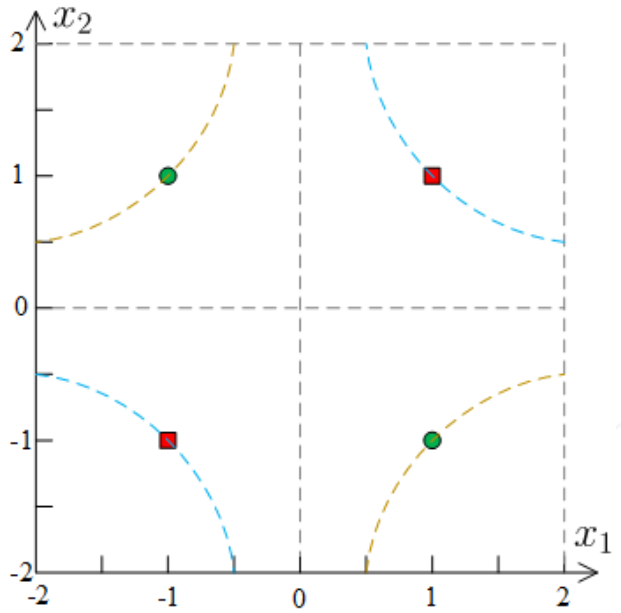
SVM 求解 XOR 问题 (2)

- 根据核函数展开式，可以得到基函数，也就是输入向量在高维空间的映射

$$\phi(\boldsymbol{x}_i) = (1, x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2, \sqrt{2}x_{i1}, \sqrt{2}x_{i2})^{\mathsf{T}}$$

在此例中，输入为**二维空间**向量，通过基函数映射到了**六维空间**. 下面左图为六维空间在平面的投影，可以看出数据变得线性可分。下图右边为六维空间的支撑向量所在决策平面对应于原空间为双曲线 $\boldsymbol{x}_1\boldsymbol{x}_2 = \pm 1$





SVM 求解 XOR 问题 (3)

- 将输入样本代入式 (23) 中，可得到 4×4 的 κ 矩阵中各元素的值为

$$\kappa = \begin{vmatrix} \kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_3, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_4, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_4) \end{vmatrix} = \begin{vmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{vmatrix}$$

- 根据式 (21)，可知接下来需要**寻找拉格朗日乘子 $\{\alpha_i\}$ 使得以下目标函数 L_p 最大**

$$L_p = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

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

$$\alpha_i \geq 0, i = 1, 2, ..., m.$$

SVM 求解 XOR 问题 (4)

- 将数据代入, 即使下式最大化

$$L_p = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 9\alpha_2^2 + 2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2)$$

约束条件为:

$$-\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 = 0$$

$$\alpha_i, i = 1, 2, 3, 4.$$

- L_p 对 α_i 求偏导并令导数等于 0 得:

$$9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1$$

$$-\alpha_1 - 9\alpha_2 + \alpha_3 - \alpha_4 = 1$$

$$-\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 = 1$$

$$\alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1$$

SVM 求解 XOR 问题 (5)

- 解得拉格朗日系数的最优值为：

$$\alpha_1 = \frac{1}{8}, \alpha_2 = \frac{1}{8}, \alpha_3 = \frac{1}{8}, \alpha_4 = \frac{1}{8}$$

- 观察可知，没有拉格朗日系数等于 0，所以四个样本都是支撑向量。此时对应的目标函数最优值为 $L_p^{op} = \frac{1}{4}$
- 接下来求解对应六维空间的划分超平面对应的法向量 $\boldsymbol{w}$

$$\boldsymbol{w} = \sum_{i=1}^{4} \alpha_i y_i \phi(\boldsymbol{x}_i) = \frac{1}{8}[-\phi(\boldsymbol{x}_1) + \phi(\boldsymbol{x}_2) + \phi(\boldsymbol{x}_3) - \phi(\boldsymbol{x}_4)]$$

$$= \frac{1}{8}\left[- \begin{pmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ -\sqrt{2} \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ \sqrt{2} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ -1/\sqrt{2} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

SVM 求解 XOR 问题 (6)

- 由式 (12) 所示的 KKT 条件可对每个支持向量求位移项 b_i

$$b_1 = -1 - \boldsymbol{w} \cdot \phi(\boldsymbol{x}_1) = -1 - 0 - 0 - (\sqrt{2})(-\frac{1}{\sqrt{2}}) - 0 - 0 - 0 = 0$$

$$b_2 = 1 - \boldsymbol{w} \cdot \phi(\boldsymbol{x}_2) = 1 - 0 - 0 - (-\sqrt{2})(-\frac{1}{\sqrt{2}}) - 0 - 0 - 0 = 0$$

$$b_3 = 1 - \boldsymbol{w} \cdot \phi(\boldsymbol{x}_3) = 1 - 0 - 0 - (-\sqrt{2})(-\frac{1}{\sqrt{2}}) - 0 - 0 - 0 = 0$$

$$b_4 = -1 - \boldsymbol{w} \cdot \phi(\boldsymbol{x}_4) = -1 - 0 - 0 - (\sqrt{2})(-\frac{1}{\sqrt{2}}) - 0 - 0 - 0 = 0$$

- 对这些值取平均，得到：

$$b = 0.$$

SVM 求解 XOR 问题 (7)

- 根据划分超平面的定义可得，XOR 问题对应的模型为

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}) + b \quad = (0\ 0\ \frac{-1}{\sqrt{2}}\ 0\ 0\ 0) \begin{pmatrix} 1 \\ \boldsymbol{x}_{i1}^2 \\ \sqrt{2}\boldsymbol{x}_{i1}\boldsymbol{x}_{i2} \\ \boldsymbol{x}_{i2}^2 \\ \sqrt{2}\boldsymbol{x}_{i1} \\ \sqrt{2}\boldsymbol{x}_{i2} \end{pmatrix} + 0 = -\boldsymbol{x}_{i1}\boldsymbol{x}_{i2}$$

- 对应的划分超平面 $\boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}) + b = 0$ 为:

$$-\boldsymbol{x}_{i1}\boldsymbol{x}_{i2} = 0$$

软间隔 SVM (1)

- **硬间隔：** 前面的讨论中，样本点或样本点映射后都是完全线性可分的。
- 现实中存在的问题

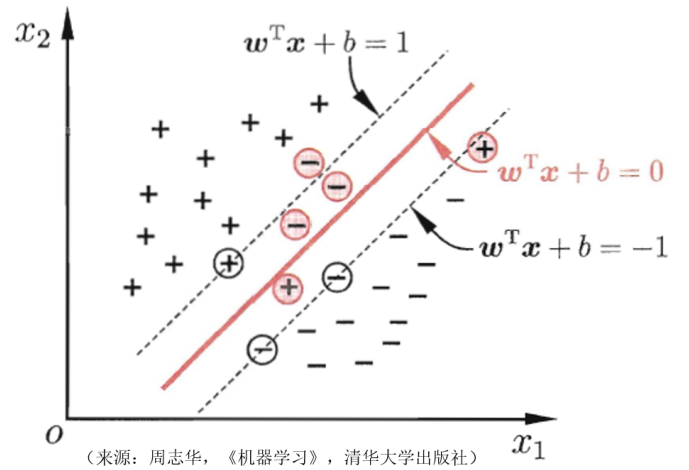
  1. 现实任务中往往很难确定合适的核函数使得训练样本在特征空间中线性可分。

  2. 即使找到了某个核函数使得训练数据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。

- 由此，引入软间隔，**允许 SVM 在一些样本上出错**，即允许某些样本不满足约束：

$$y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1$$

右图中红色样本即表示不满足约束条件的样本



（来源：周志华，《机器学习》，清华大学出版社）

软间隔 SVM (2)

- 对于软间隔，在**最大化间隔**的同时，还希望不满足约束的样本应该**尽可能的少**，所以优化目标式 (7) 改写为

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) - 1\right) \tag{24}$$

其中，$C > 0$ 是常数，$\ell_{0/1}$ 是 “0/1 损失函数”：

$$\ell_{0/1} = \begin{cases} 1 & z < 0 \\ 0 & otherwise \end{cases} \tag{25}$$

- C 趋于无穷大时，上式等价于硬间隔的约束条件。
- C 取有限值时，允许某些样本不满足约束条件。

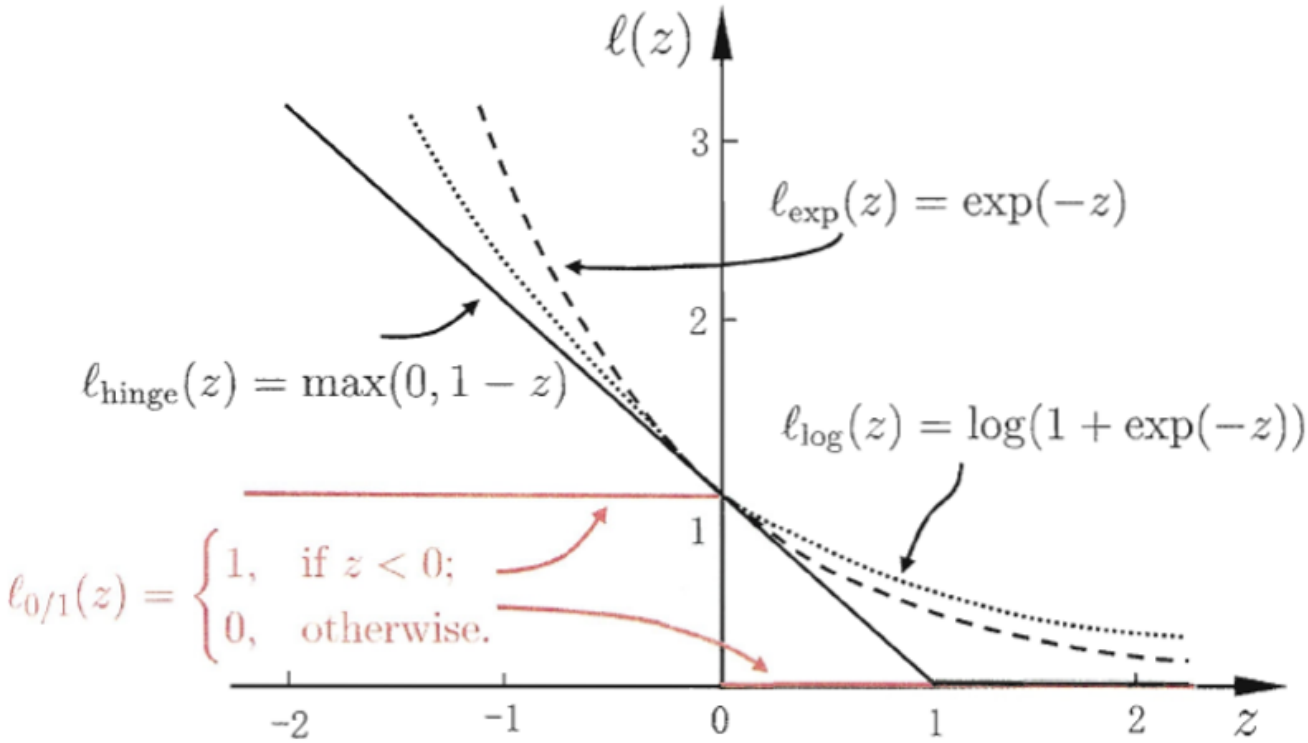
软间隔 SVM (3)

- $\ell_{0/1}$ 非凸、非连续，数学性质不好, 常用以下损失函数代替

  hinge 损失: $\ell_{hinge}(z) = \max(0, 1 - z)$;

  指数损失 (exponential loss): $\ell_{exp}(z) = \exp(-z)$;

  对数损失 (logistic loss): $\ell_{log}(z) = \log(1 + \exp(-z))$.



（来源：周志华，《机器学习》，清华大学出版社）

软间隔 SVM (4)

- 若果采用 hinge 损失替换 $\ell_{0/1}$，则优化目标式 (24) 变为

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, 2, ..., m. \end{aligned} \tag{26}$$

其中，ξ_i 称为**松弛变量** (slack variable)，且 $\xi_i \geq 0$

$$\xi_i = max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

这就是常用的“软间隔支持向量机”，解出 $\boldsymbol{\alpha}$ 即可得到划分超平面。

软间隔 SVM (5)

- 式 (26) 优化目标的对偶问题变为

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_j \\ &s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C, i = 1, 2, ..., m. \end{aligned} \tag{27}$$

  将式 (27)与硬间隔下的对偶问题式 (12)对比可以发现，两者唯一不同是对偶变量的约束条件：

  式 (27)中：$0 \leq \alpha_i \leq C$

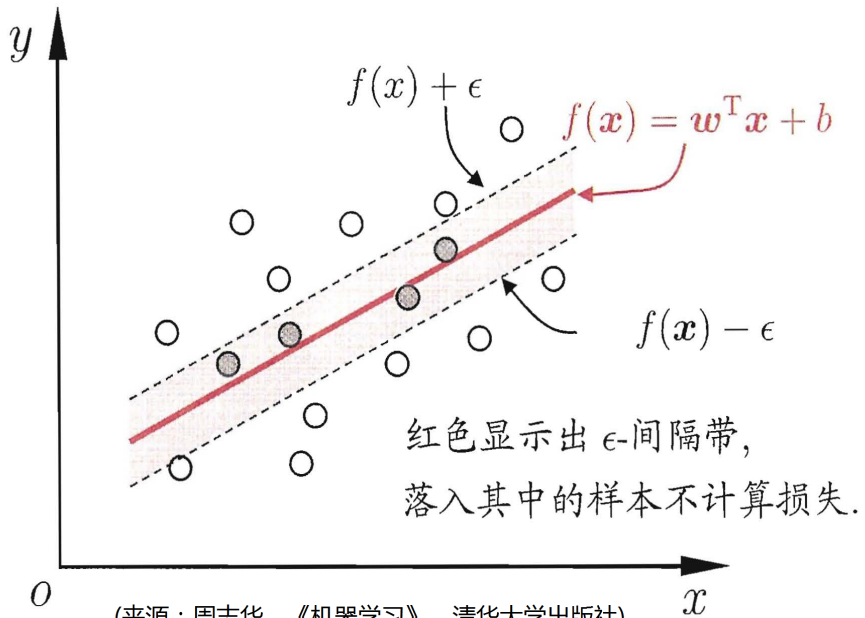
  式 (12)中：$0 \leq \alpha_i$

- 类似的，KKT 条件也会做出相应的变化。但是，软间隔支持向量机和硬间隔下一样，**最终模型仅与支持向量有关**.

支持向量回归 (1)

回顾：回归问题 给定训练样本 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\}$，$y \in \mathbb{R}$，希望学得一个回归模型：$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b$，其中 $\boldsymbol{w}$ 和 b 是待学习的模型参数。学习目标是最小化模型输出 $f(\boldsymbol{x})$ 与真实输出 y 之间的误差损失；而盖损失当且仅当 $f(\boldsymbol{x})$ 与 y 完全相同时，损失才为 0。

支持向量回归 (Support Vector Regression, SVR)：假设能容忍 $f(x)$ 与 y 最多有 ϵ 的偏差，则相当于以 $f(x)$ 为中心，构建了一个宽度为 2ϵ 的间隔带；若样本落入此范围中，则认为预测正确。



(来源：周志华，《机器学习》，清华大学出版社)

支持向量回归 (2)

- **SVR 优化目标：**

$$\min_{\boldsymbol{\omega},b} \quad \frac{1}{2}||\boldsymbol{\omega}||^2 + C\sum_{i=1}^{m}\ell_\epsilon(f(\boldsymbol{x}_i) - y_i) \tag{28}$$

其中 C 为正则化常数，ℓ_ϵ 为 ϵ-不敏感损失函数：

$$\ell_\epsilon(z) = \begin{cases} 0 & if\ |z| \leq \epsilon \\ |z| - \epsilon & otherwise \end{cases}$$

- 引入松弛变量 ξ_i 和 $\hat{\xi}_i$，可将上式重写为：

$$\begin{aligned} \min_{\boldsymbol{\omega},b,\xi_i\hat{\xi}_i} \quad & \frac{1}{2}||\boldsymbol{\omega}||^2 + C\sum_{i=1}^{m}(\xi_i + \hat{\xi}_i) \\ s.t. \quad & f(\boldsymbol{x}_i) - y_i \leq \epsilon + \xi_i, \\ & y_i - f(\boldsymbol{x}_i) \leq \epsilon + \hat{\xi}_i, \\ & \xi_i \geq 0, \hat{\xi}_i \geq 0, i = 1, 2, ..., m. \end{aligned} \tag{29}$$

支持向量回归 (3)

- **SVR 的对偶问题：**

$$\begin{aligned} \max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} \quad & \sum_{i=1}^{m} y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j \\ s.t. \quad & \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i) = 0, \\ & 0 \le \alpha_i, \hat{\alpha}_i \le C. \end{aligned} \tag{30}$$

- 上述过程同样需要满足 KKT 条件。解出上式中的 α_i 和 $\hat{\alpha}_i$ 后，可得到：

$$f(\boldsymbol{x}) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x} + b.$$

其中

$$b = y_i + \epsilon - \sum_{j=1}^{m}(\hat{\alpha}_j - \alpha_j)\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_i.$$

实践中常选取多个或所有满足 $0 < \alpha_i < C$ 的样本求解 b 后取平均值。

小结

SVM 方法的特点：

- SVM 的目标是对特征空间划分的最优超平面，核心思想是最大化分类间隔。
- SVM 利用内积核函数向高维空间进行特征映射，使得原线性不可分特征变得可能。
- 支持向量是 SVM 的训练结果，在 SVM 分类决策中起决定作用的是支持向量。
- SVM 得到的决策函数只由少数的支持向量所确定，计算复杂性取决于支持向量的数目，而不是样本空间的维数，能够避免了“维数灾难”。
- SVM 方法已经在图像识别、信号处理和基因图谱识别等方面得到了成功的应用。
- 支持向量方法也为样本分析、因子筛选、信息压缩、知识挖掘和数据修复等提供了新工具。
- SVM 本质上是二分类器，可通过前面讲的多分类学习方法扩展到多分类问题。