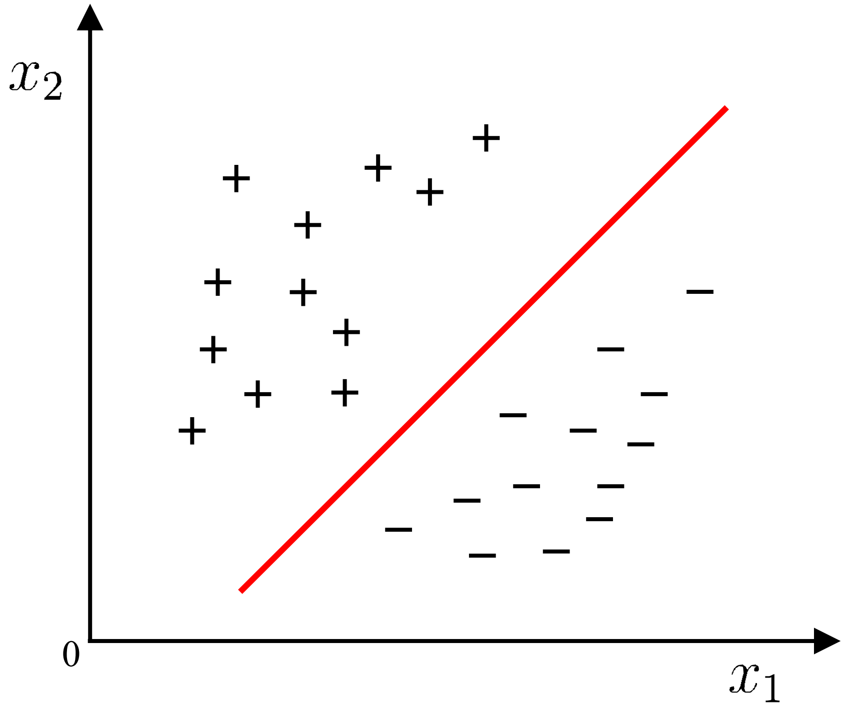
第六章：支持向量机

主讲：连德富 特任教授 | 博士生导师
邮箱：liandefu@ustc.edu.cn
手机：13739227137
主页：http://staff.ustc.edu.cn/~liandefu

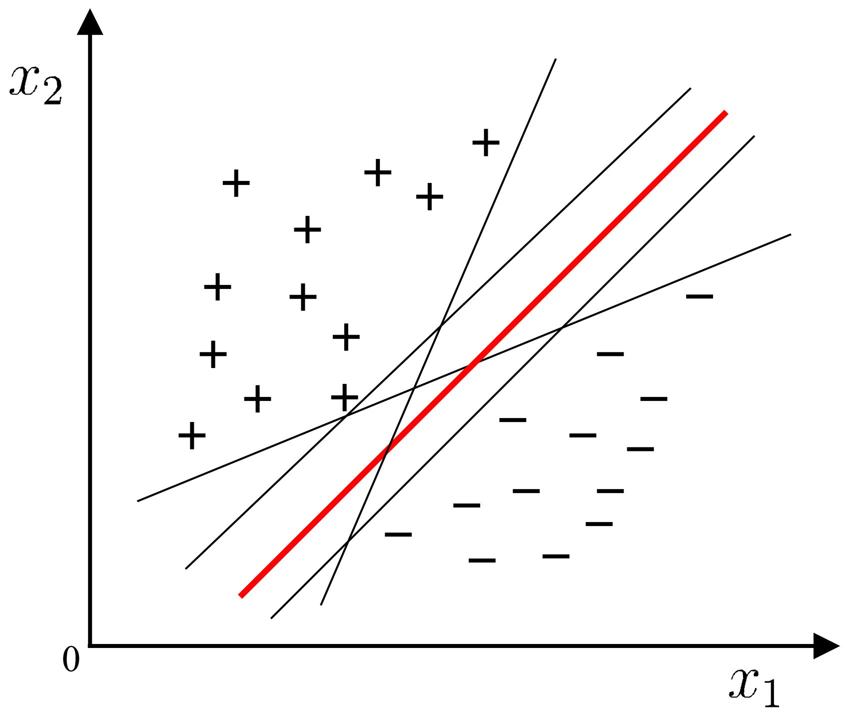
线性模型

在样本空间中寻找一个超平面，将不同类别的样本分开



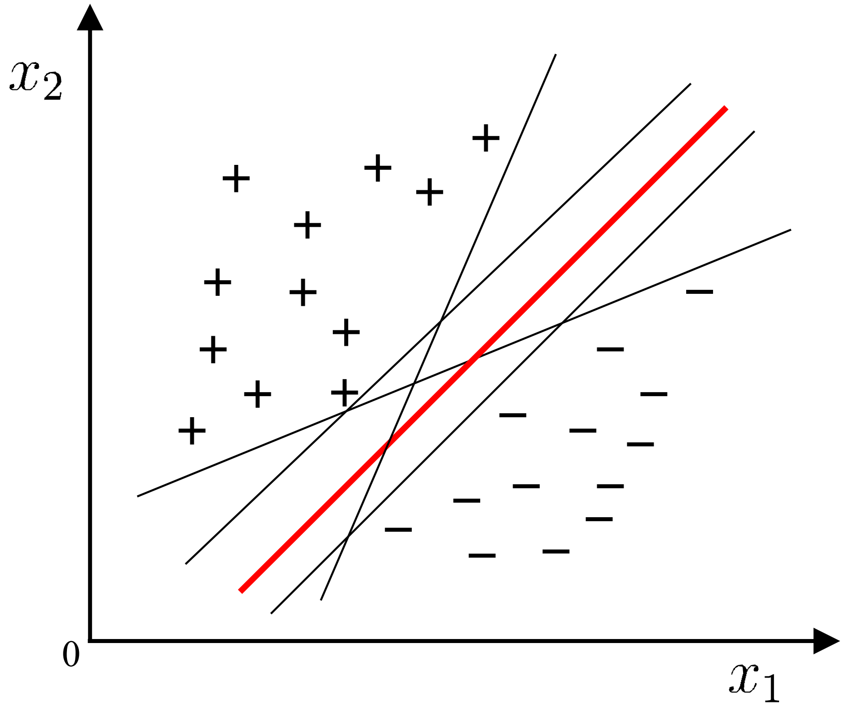
线性模型

- 将训练样本分开的超平面可能有很多, 哪一个好呢?



线性模型

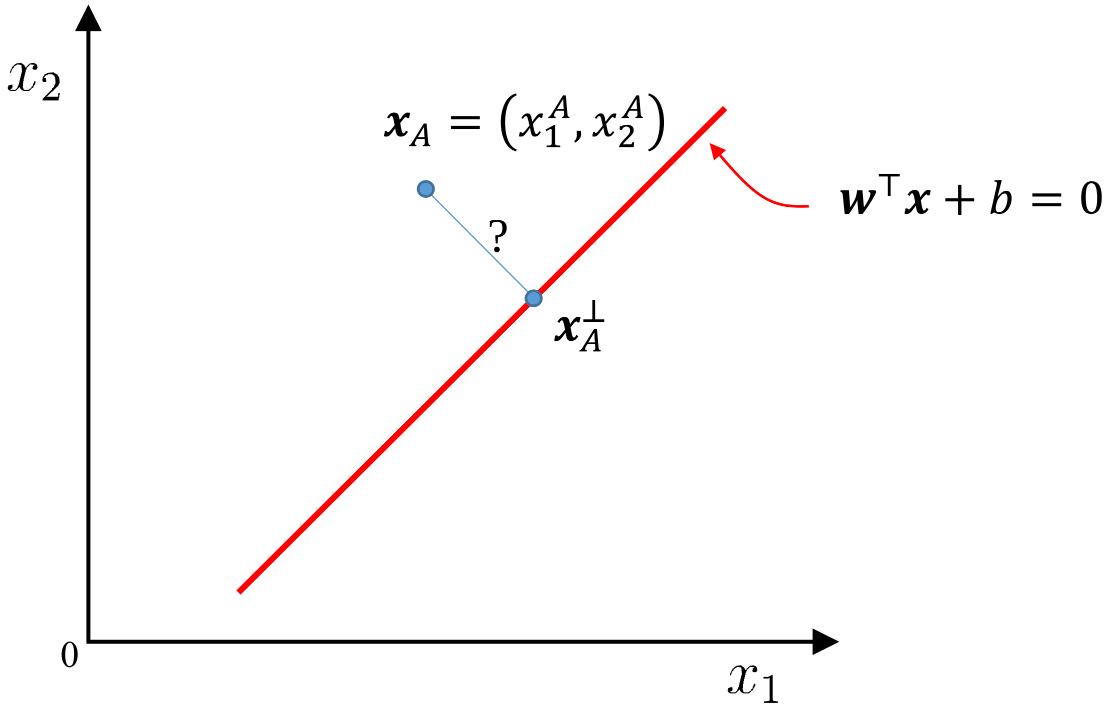
- 将训练样本分开的超平面可能有很多, 哪一个好呢?



应选择"正中间", 容忍性好, 鲁棒性高, 泛化能力最强

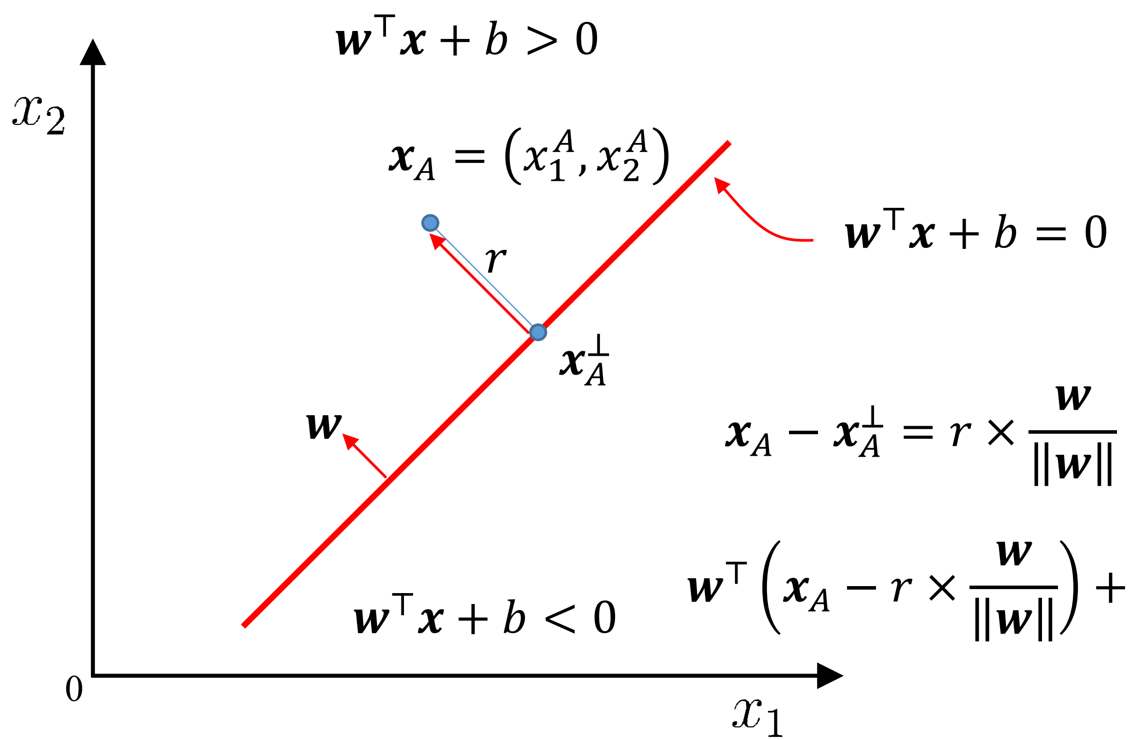
线性模型

- 超平面方程: $\boldsymbol{w}^{\top}\boldsymbol{x} + b = 0$



线性模型

- 超平面方程: $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$



$$\boldsymbol{x}_A - \boldsymbol{x}_A^\perp = r \times \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$\boldsymbol{w}^\top \left(\boldsymbol{x}_A - r \times \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + b = 0$$

$$r = \frac{|\boldsymbol{w}^\top \boldsymbol{x}_A + b|}{\|\boldsymbol{w}\|}$$

线性模型（线性可分）

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

$$\Rightarrow \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0$$

- 分类平面对于向量$\boldsymbol{w}' = [\boldsymbol{w}, b]$的长度具有不变性

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \quad \xrightarrow[\alpha > 0]{\boldsymbol{w}' \mapsto \alpha \boldsymbol{w}'} \quad y_i(\alpha \boldsymbol{w}^\top \boldsymbol{x}_i + \alpha b) > 0$$

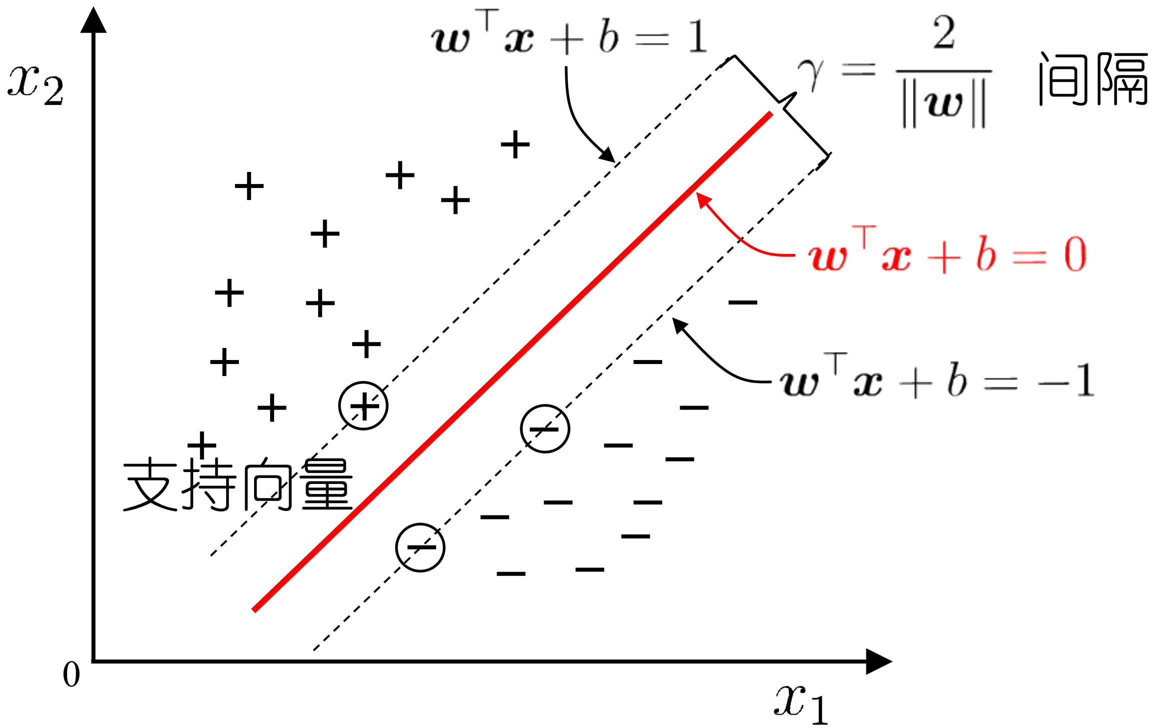
- 假设$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq \epsilon$，则$y_i\left(\frac{\boldsymbol{w}^\top}{\epsilon} \boldsymbol{x}_i + \frac{b}{\epsilon}\right) \geq 1$

$$y_i\left(\frac{\boldsymbol{w}^\top}{\epsilon} \boldsymbol{x}_i + \frac{b}{\epsilon}\right) \geq 1 \quad \xLeftrightarrow{\frac{1}{\epsilon}(\boldsymbol{w}, b) \mapsto (\boldsymbol{w}', b')} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$

线性模型（线性可分）

- $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$

若$y_i = +1$，则有$\boldsymbol{w}^\top \boldsymbol{x}_i + b \geq +1$
若$y_i = -1$，则有$\boldsymbol{w}^\top \boldsymbol{x}_i + b \leq -1$



支持向量机基本型

- 最大间隔:寻找参数$\boldsymbol{w}$和b, 使得间隔r最大

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t.}\ y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 \qquad i = 1, \cdots, m$$

$$\Longleftrightarrow$$

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.}\ y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 \qquad i = 1, \cdots, m$$

对偶问题

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \qquad i = 1, \cdots, m$$

凸函数

凸函数

凸优化问题

$$\min_{x} f(\boldsymbol{x})$$

$$s.t. \quad g_i(\boldsymbol{x}) \leq 0, i = 1,2, \ldots, m$$

$$\boldsymbol{a}_i^\top \boldsymbol{x} = b, j = 1,2, \cdots, n$$

其中$f(\boldsymbol{x}), g_i(\boldsymbol{x})$是凸函数

对偶问题

原问题 凸优化

$$\min_{x} f(\boldsymbol{x})$$
$$s.t.\quad g_i(\boldsymbol{x}) \leq 0, i = 1,2,\ldots,m$$
$$h_j(\boldsymbol{x}) = 0, j = 1,2,\cdots,n$$

广义拉格朗日函数

$$L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = f(\boldsymbol{x}) + \sum_i u_i\, g_i(\boldsymbol{x}) + \sum_j v_j\, h_j(\boldsymbol{x})$$

其中$u_i \geq 0$

对偶问题 凸优化

$$\max_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u},\boldsymbol{v})\ s.t.\ \ \boldsymbol{u} \succcurlyeq 0$$

其中$g(\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v})$

对偶问题

原问题 凸优化

$$\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|^2/2$$

$$s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,m$$

广义拉格朗日函数

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

其中$\alpha_i \geq 0$

对偶问题 凸优化

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})\ s.t.\ \ \boldsymbol{\alpha} \succcurlyeq 0$$

其中$g(\boldsymbol{\alpha}) = \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$

对偶问题

- $g(\boldsymbol{\alpha}) = \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$

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

$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$是关于$\boldsymbol{w}$和$b$的凸函数，　在其梯度等于0时取得最优值

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{b}} = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

对偶问题

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

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \qquad \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$g(\boldsymbol{\alpha}) = \frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i\right\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\left(\sum_{j=1}^{m} \alpha_j y_j \boldsymbol{x}_j \right)^\top \boldsymbol{x}_i \right) \right) + \sum_{i}^{m} \alpha_i$$

$$= \sum_{i}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

对偶问题

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

$$s.t. \;\; \boldsymbol{\alpha} \geqslant 0 \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0$$

对偶问题

原问题 凸优化

$$\min_{x} f(\boldsymbol{x})$$
$$s.t.\quad g_i(\boldsymbol{x}) \leq 0, i = 1,2, \dots, m$$
$$h_j(\boldsymbol{x}) = 0, j = 1,2, \cdots, n$$

对偶问题 凸优化

$$\max_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u}, \boldsymbol{v})\ s.t.\ \ \boldsymbol{u} \geqslant 0$$

其中$g(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$

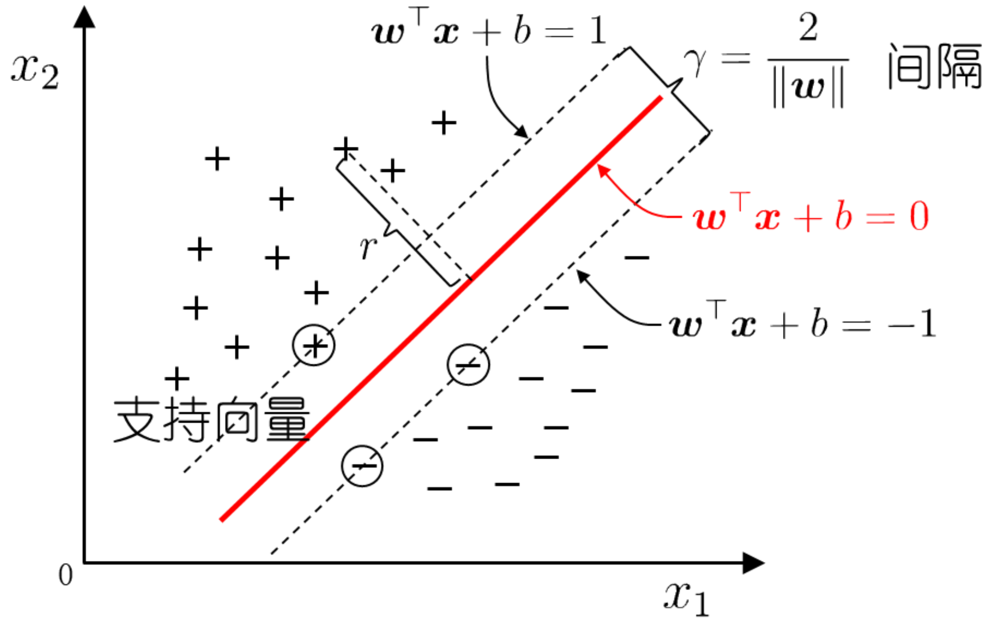
强对偶性

$$f^\star = g^\star = \max_{\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{u}, \boldsymbol{v})$$

Slater条件

原问题为凸优化问题，且可行域中至少有一个点使得不等式约束严格成立

对偶问题



- 对于线性可分问题，一定存在 $(\boldsymbol{w}, b)$ 使得 $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$，
  $i = 1,2, \dots, m$
- 设$\epsilon < 1$，则$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > \epsilon$ ⇨ $y_i\left(\frac{\boldsymbol{w}^\top}{\epsilon} \boldsymbol{x}_i + \frac{b}{\epsilon}\right) > 1$

则$\left(\frac{\boldsymbol{w}}{\epsilon}, \frac{b}{\epsilon}\right)$严格满足不等式。因此，该优化问题满足强对偶性条件

对偶问题

对偶问题

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \underbrace{\boldsymbol{x}_i^\top \boldsymbol{x}_j}_{K_{ij}}$$

$$s.t.\ \ \boldsymbol{\alpha} \geqslant 0 \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0$$

通过序列最小优化（SMO）求解最优的$\boldsymbol{\alpha}$

基本思路：不断执行如下两个步骤直至收敛

第一步：选取一对需更新的变量α_i和α_j

第二步：固定α_i和α_j以外的参数, 求解对偶问题更新α_i和α_j

$$g(\alpha_i,\alpha_j) = \alpha_i + \alpha_j - \frac{1}{2}\left(\alpha_i^2 K_{ii} + \alpha_j^2 K_{jj} + 2\alpha_i\alpha_j y_i y_j K_{ij} + \alpha_i y_i \sum_{k\neq i,j} a_k y_k K_{ik} + \alpha_j y_j \sum_{k\neq i,j} a_k y_k K_{jk}\right)$$

仅考虑α_i和α_j时, 对偶问题的约束变为

$$\alpha_i y_i + \alpha_j y_j = -\sum_{k\neq i,j}^{m} \alpha_i y_i = \zeta \quad \Rightarrow \quad \alpha_i = (\zeta - \alpha_j y_j) y_i$$

对偶问题

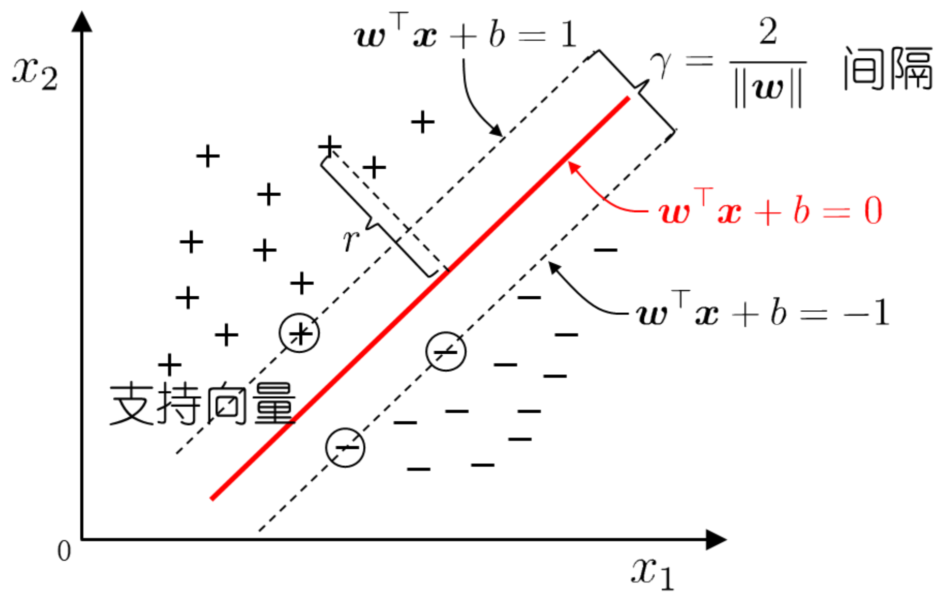
- 终止条件：KKT条件

$$
\begin{cases}
\alpha_i \geq 0 \\
y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \\
\alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1) = 0
\end{cases}
$$

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$$

$$\alpha_i > 0 \Rightarrow y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$$

支持向量机解的稀疏性: 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



对偶问题

- 如何选择两个变量?

$$\begin{cases} \alpha_i \geq 0 \\ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \\ \alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1) = 0 \end{cases}$$

若α_i和α_j有一个违反了KKT条件，目标函数就会在迭代后增大 Osuna et al. 1997

- KKT条件违背的程度越大，则变量更新后可能导致的目标函数值增幅越大

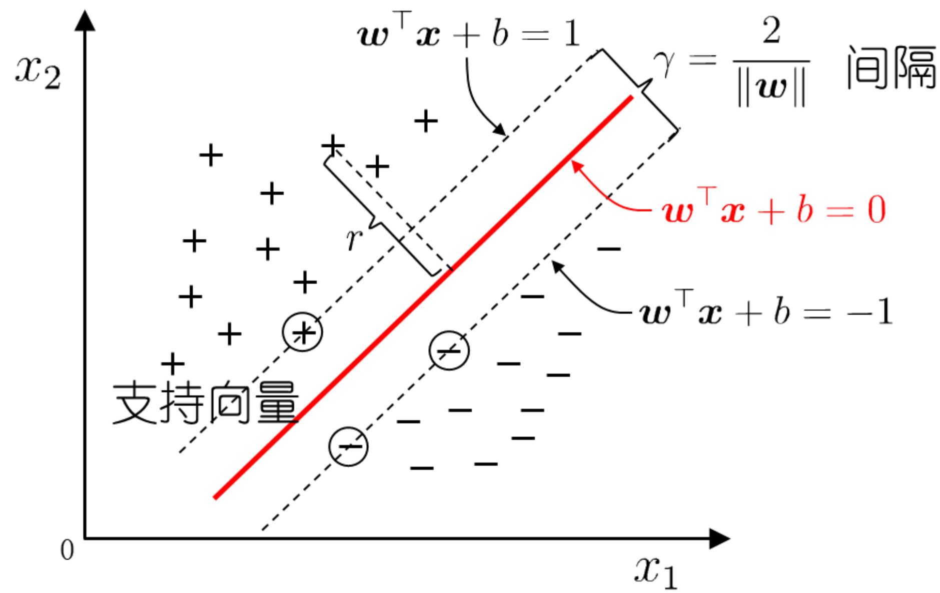
第一个变量：选取违背KKT条件程度最大的变量
第二个变量：与第一个变量的间隔最大的变量

对偶问题

给定最优$\boldsymbol{\alpha}$，求解$\boldsymbol{w}$和b

$$\boldsymbol{\alpha}^\star = \max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})$$

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



对于任意的支持向量 $\boldsymbol{x}_i \in S$，均满足$\boldsymbol{w}^\top \boldsymbol{x}_i + b = y_i$，则

$$b^\star = \frac{1}{|S|} \sum_{i \in S} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}^\star\right)$$

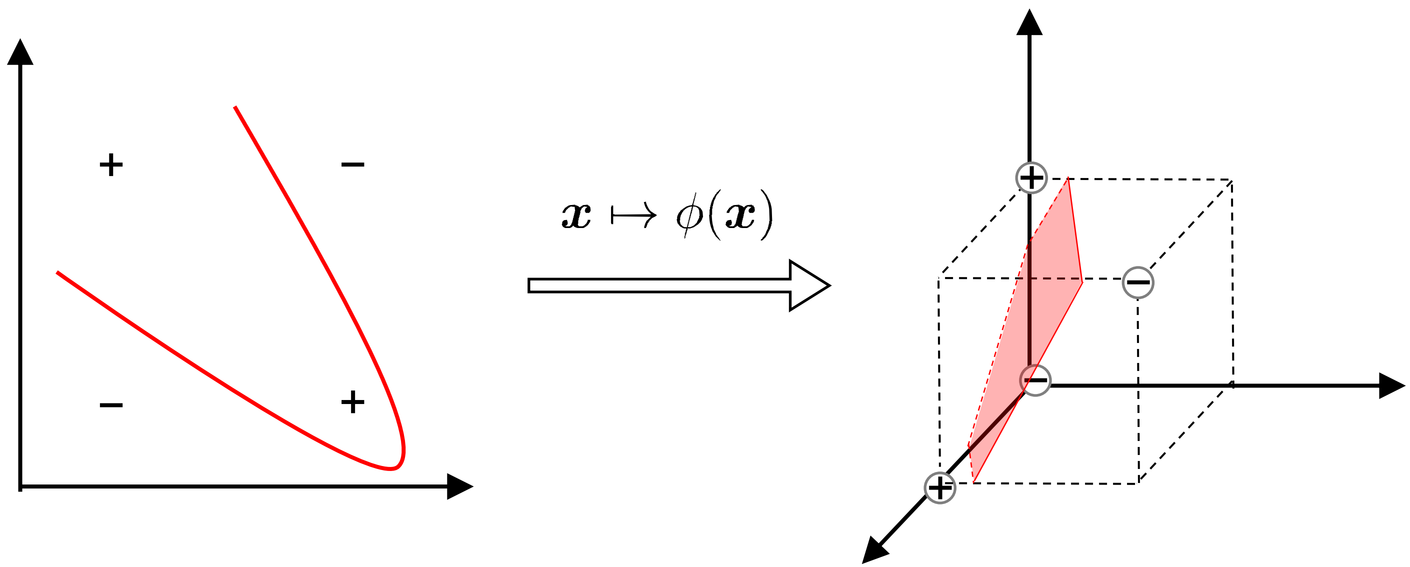
支持向量机解的稀疏性: 训练完成后, 大部分的训练样本都不需保留, 最终模型仅与支持向量有关

$$y = \boldsymbol{x}^\top \boldsymbol{w} + b = \sum_{i}^{m} \alpha_i^\star y_i \boldsymbol{x}^\top \boldsymbol{x}_i + b^\star$$

线性模型（线性不可分）

- 若不存在一个能正确划分两类样本的超平面, 怎么办

> 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



核支持向量机

- 设样本$\boldsymbol{x}$映射后的向量为$\phi(\boldsymbol{x})$, 划分超平面为$\boldsymbol{w}^\top\phi(\boldsymbol{x})+b=0$

原问题

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2/2$$
$$s.t.\ \ y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\geq 1, i=1,2,\ldots,m$$

对偶问题

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})=\sum_{i=1}^m \alpha_i-\frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j \boxed{\phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_j)}$$
$$s.t.\ \ \boldsymbol{\alpha}\geqslant 0 \text{ and } \sum_{i=1}^m \alpha_i y_i=0$$

只以内积的形式出现

预测函数

$$y=\boldsymbol{x}^\top\boldsymbol{w}+b=\sum_{i}^{m}\alpha_i^\star y_i \boxed{\phi(\boldsymbol{x})^\top\phi(\boldsymbol{x}_i)}+b^\star$$

核支持向量机

- 由于特征空间维数可能很高，甚至是无穷维，因此直接计算 $\phi(\boldsymbol{x})^\top\phi(\boldsymbol{x}_i)$ 通常是困难的。
- 可以设计函数 $\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_i)$

核函数

对偶问题

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})=\sum_{i=1}^m \alpha_i-\frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j \kappa\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$$
$$s.t.\ \ \boldsymbol{\alpha}\geqslant 0 \text{ and } \sum_{i=1}^m \alpha_i y_i=0$$

预测函数

$$y=\boldsymbol{x}^\top\boldsymbol{w}+b=\sum_{i}^{m}\alpha_i^\star y_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b^\star$$

核支持向量机

- 若已知$\phi(\cdot)$，则可写出核函数$\kappa(\cdot,\cdot)$；但现实任务中通常不知道$\phi(\cdot)$的形式
  - 核函数是否存在?
  - 什么样的函数可以作为核函数?

令$\mathcal{X}$为输入空间，$\mathcal{H}$为特征空间，如果存在$\phi(\cdot)$: $\mathcal{X} \mapsto \mathcal{H}$，使得对所有$\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$，函数$\kappa(\cdot,\cdot)$满足条件

$$\kappa(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{z})$$

则称$\kappa(\cdot,\cdot)$为核函数

Mercer定理：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$\kappa$是核函数当且仅当对于任意数据集$D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$，核矩阵K总是半正定的

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

核支持向量机

- 常用核函数:

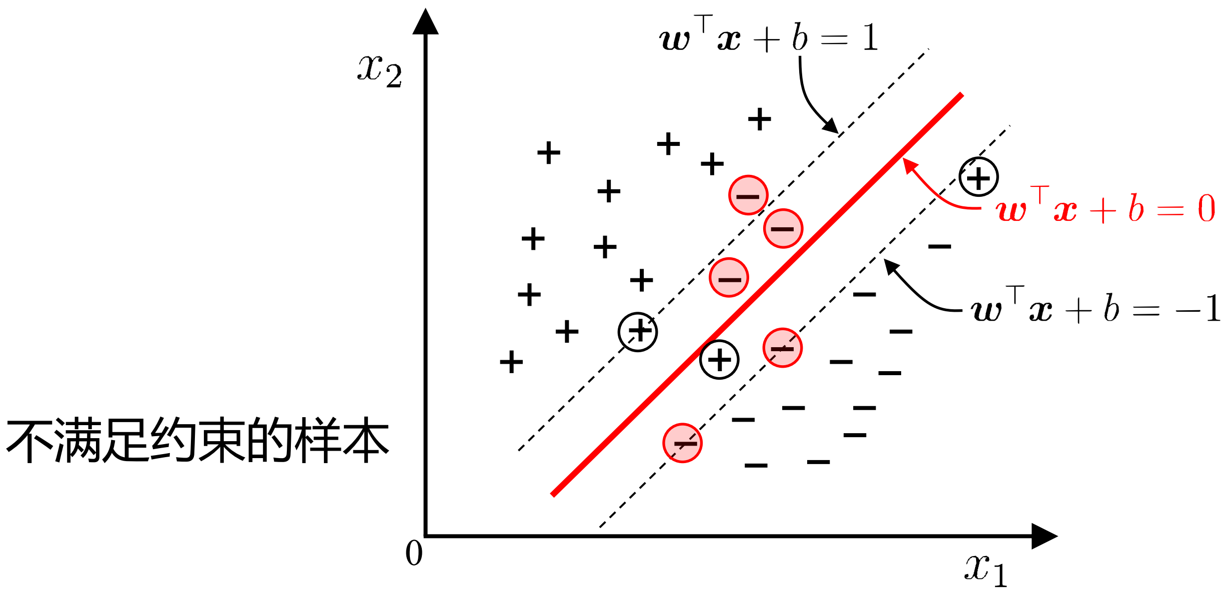
名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j)^d$	$d \geq 1$为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\\|\boldsymbol{x}_i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

- 函数组合得到
  - 核函数线性组合$\gamma_1\kappa_1 + \gamma_2\kappa_2$
  - 核函数直积 $\kappa_1 \otimes \kappa_2(\boldsymbol{x}, \boldsymbol{z}) = \kappa_1(\boldsymbol{x}, \boldsymbol{z}) \otimes \kappa_1(\boldsymbol{x}, \boldsymbol{z})$
  - $g(\boldsymbol{x})\kappa(\boldsymbol{x}, \boldsymbol{z})g(\boldsymbol{z})$

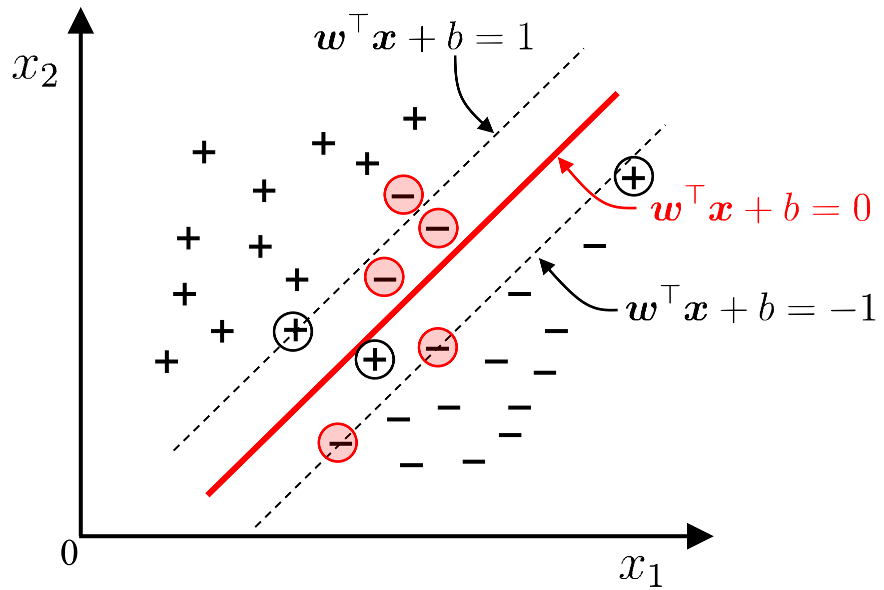
线性模型（线性不可分）

- 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
- 一个线性可分的结果也很难断定是否是有过拟合造成的

引入软间隔的概念, 允许支持向量机在一些样本上不满足约束



软间隔支持向量机



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

正则化

$$s.t.\ \ y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, i = 1,2,\ldots,m$$

为每个样本$(\boldsymbol{x}_i, y_i)$引入松弛变量ξ_i

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

软间隔支持向量机

原问题 凸优化

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

$$s.t.\ \ y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b) \geq 1-\xi_i, i=1,2,\ldots,m$$

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

广义拉格朗日函数

$$L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1+\xi_i) - \sum_{i}^{m}\mu_i\xi_i$$

$$\alpha_i, \mu_i \geq 0$$

对偶问题

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)$$

$$s.t.\ C \succcurlyeq \boldsymbol{\alpha} \succcurlyeq 0 \text{ and } \sum_{i=1}^{m}\alpha_i y_i = 0$$

软间隔支持向量机

对偶问题

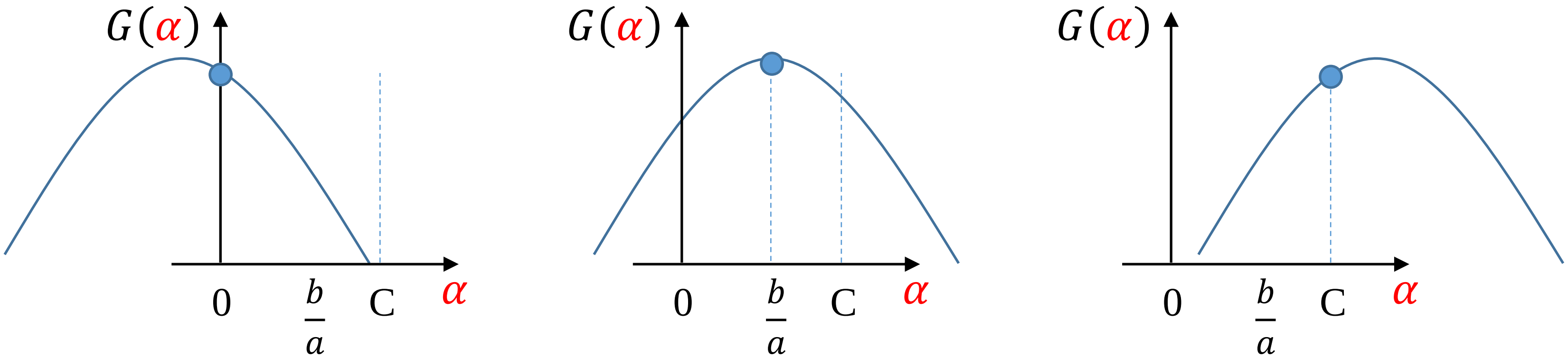
$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

$$s.t.\ C \geqslant \boldsymbol{\alpha} \geqslant 0 \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0$$

SMO

$$\max_{\alpha} G(\alpha) = -a\alpha^2 + 2b\alpha + c$$

$$\text{s.t. } 0 \le \alpha \le C$$



软间隔支持向量机

- 终止条件：KKT条件

$$\begin{cases} \alpha_i \geq 0, \mu_i \geq 0 \\ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \\ \alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i) = 0 \\ \xi_i \geq 0, \mu_i \xi_i = 0 \end{cases}$$

对于样本$(\boldsymbol{x}_i, y_i)$，若$\alpha_i > 0$，则$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1 - \xi_i$.　该样本是支持向量

若$\alpha_i < C$，由$C = \alpha_i + \mu_i$得出$\mu_i > 0$，从而得出$\xi_i = 0$.　该样本在最大间隔边界上

若$\alpha_i = C$，由$C = \alpha_i + \mu_i$得出$\mu_i = 0$

若$\xi_i \leq 1$，　该样本在最大间隔内

若$\xi_i > 1$，　该样本被错误分类

软间隔支持向量机

- 目标函数视角

基本想法：最大化间隔的同时，让不满足约束的样本应尽可能少.

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}(y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i)+b)-1)$$

其中 $\ell_{0/1}$是0/1损失函数

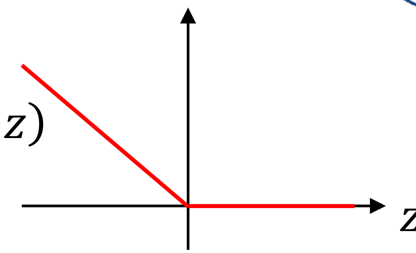
$$\ell_{0/1}(z) = \begin{cases} 1, z < 0 \\ 0, otherwise \end{cases}$$

存在的问题：0/1损失函数非凸、非连续，不易优化！

软间隔支持向量机

$$\ell_{0/1}(z) = \begin{cases} 1, z < 0 \\ 0, otherwise \end{cases}$$

$\max(0, -z)$



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

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

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

$$\text{s.t.} \quad 1 - y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \leq \xi_i$$

$$y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i$$

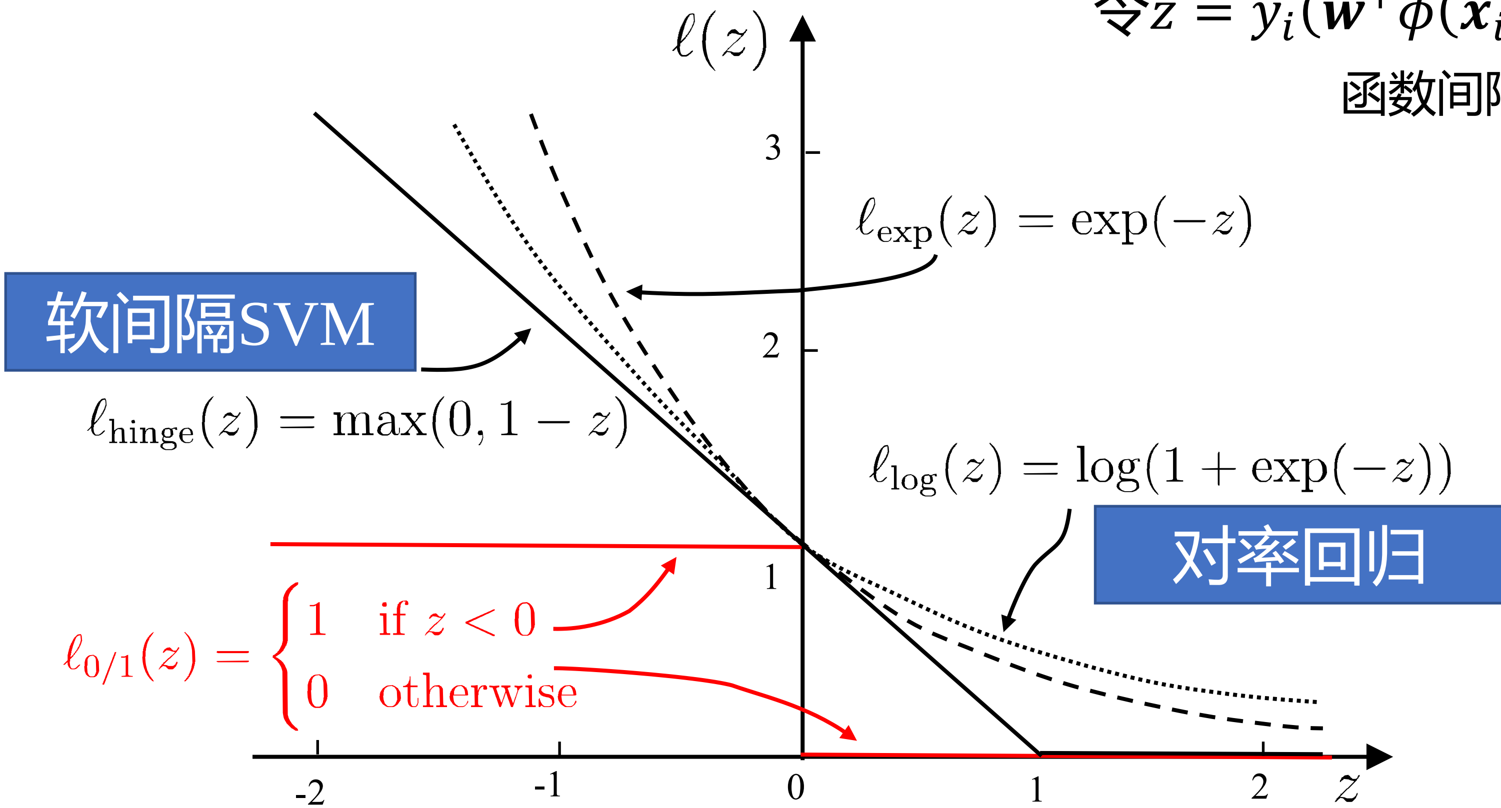
$$0 \leq \xi_i$$

替代损失函数

$y_i \in \{\pm 1\}$

令$z = y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b)$

函数间隔



替代损失函数数学性质较好，一般是0/1损失函数的上界

正则化

- 支持向量机学习模型的更一般形式

$$\min_{\mathrm{f}} \Omega(f) + C\sum_{i}^{m} \ell(f(x_i), y_i)$$

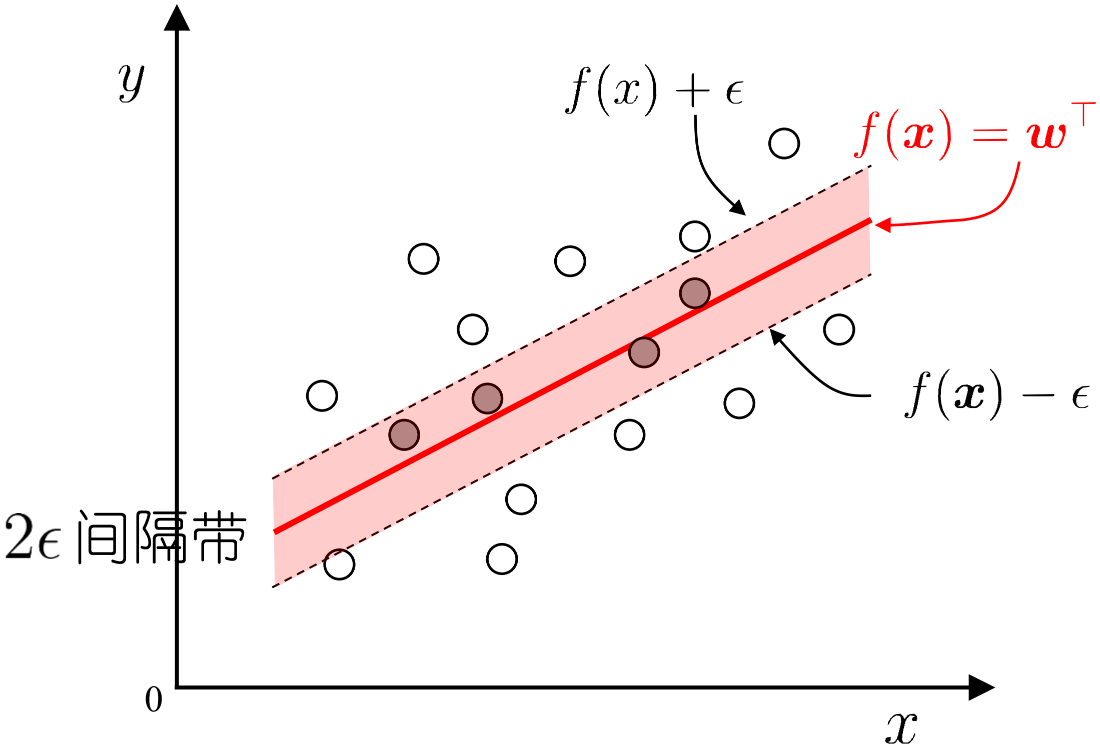
结构风险，描述模型的某些性质

经验风险，描述模型与训练数据的契合程度

- 通过替换上面两个部分，可以得到许多其他学习模型
  - 对数几率回归(Logistic Regression)
  - 最小绝对收缩选择算子(LASSO)
  - ……

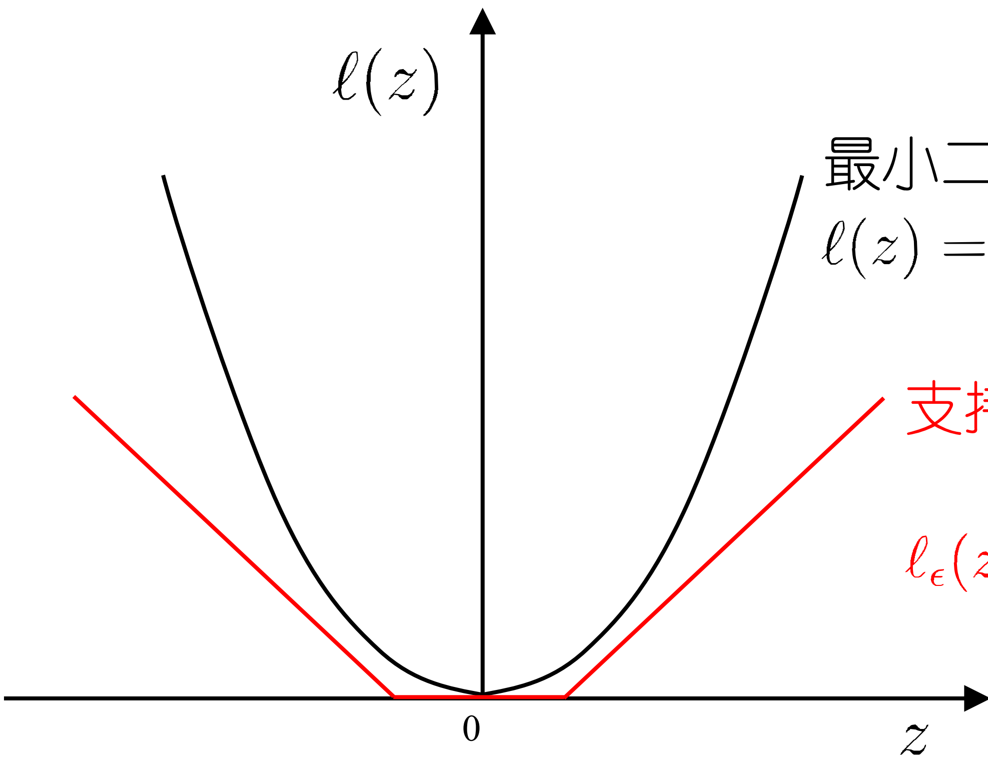
支持向量回归

- 允许模型输出和实际输出间存在 2ϵ的偏差.



损失函数

- 落入中间2ϵ 间隔带的样本不计算损失, 从而使得模型获得稀疏性



最小二乘损失函数

$$\ell(z) = z^2$$

支持向量回归损失函数

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

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^\top \boldsymbol{x}_i + b$$

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

支持向量回归

原问题

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)$$

$$\begin{aligned} \text{s.t.} \quad & f(\boldsymbol{x}_i) - y_i \leq \epsilon + \xi_i, \\ & y_i - f(\boldsymbol{x}_i) \leq \epsilon + \hat{\xi}_i, \\ & \xi_i \geq 0 \quad \hat{\xi}_i \geq 0, i = 1,2,\ldots,m \end{aligned}$$

对偶问题

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

$$s.t.\ C \geqslant \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}} \geqslant 0 \text{ and } \sum_{i=1}^{m}(\alpha_i - \hat{\alpha}_i) = 0$$

预测函数

$$y = \boldsymbol{w}^{\top}\phi(\boldsymbol{x}) + b = \sum_{i}^{m}(\hat{\alpha}_i^{\star} - \alpha_i^{\star})y_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}) + b^{\star}$$

核方法

- 支持向量机 $y = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_i^m \alpha_i^\star y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b^\star$
- 支持向量回归 $y = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_i^m (\hat{\alpha}_i^\star - \alpha_i^\star) y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b^\star$

无论是支持向量机还是支持向量回归, 学得的模型总可以表示成
核函数的线性组合

更一般的结论（表示定理）：令$\mathbb{H}$为核函数κ对应的再生核希尔伯特空间, $\|h\|_{\mathbb{H}}$表示在$\mathbb{H}$空间中关于h的范数，对于任意单调增函数$\Omega: [0, \infty) \mapsto \mathbb{R}$和任意非负损失函数$\ell: \mathbb{R}^m \mapsto [0, \infty)$, 优化问题

$$\min_{\mathrm{h} \in H} F(h) = \Omega(\|h\|_{\mathbb{H}}) + \ell\big(h(\boldsymbol{x}_1), \cdots, h(\boldsymbol{x}_m)\big)$$

的解总可以写成$h^\star = \sum_i^m \alpha_i \kappa(\cdot, \boldsymbol{x}_i)$

核线性判别分析

- 通过表示定理可以得到很多线性模型的”核化”版本
  - 核SVM
  - 核LDA
  - 核PCA
  - ……
- 核LDA: 先将样本映射到高维特征空间, 然后在此特征空间中做线性判别分析

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \mathbf{S}_b^\phi \boldsymbol{w}}{\boldsymbol{w}^\top \mathbf{S}_w^\phi \boldsymbol{w}}$$

$$\mathbf{S}_b^\phi = \left(\boldsymbol{\mu}_1^\phi - \boldsymbol{\mu}_0^\phi\right)\left(\boldsymbol{\mu}_1^\phi - \boldsymbol{\mu}_0^\phi\right)^\top$$

$$\mathbf{S}_w^\phi = \sum_{i=0}^{1} \sum_{x \in X_i} \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^\phi\right)\left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^\phi\right)^\top$$

$$h(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{i}^{m} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x})$$

$$\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{N} \boldsymbol{\alpha}}$$

$$\mathbf{M} = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^\top$$

$$\mathbf{N} = \mathbf{K}\mathbf{K}^\top - \sum_{i=0}^{1} m_i \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^\top$$

核线性判别分析

由表示定理可得 $\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i)$ $\quad \boldsymbol{\mu}_i^{\phi} = \frac{1}{m_i} \sum_{\boldsymbol{x} \in X_i} \phi(\boldsymbol{x})$

$$\mathbf{S}_b^{\phi} = \left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)\left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)^{\top}$$

$$\boldsymbol{w}^{\top}\phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^{\top} K(:, \boldsymbol{x})$$

$$\boldsymbol{w}^{\top}\boldsymbol{\mu}_i^{\phi} = \frac{1}{m_i} \sum_{\boldsymbol{x} \in X_i} \boldsymbol{w}^{\top}\phi(\boldsymbol{x}) = \frac{1}{m_i} \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\alpha}^{\top} K(:, \boldsymbol{x}) = \boldsymbol{\alpha}^{\top} \underbrace{\frac{1}{m_i}\mathbf{K}\mathbf{1}_i}_{\hat{\boldsymbol{\mu}}_i}$$

$$\boldsymbol{w}^{\top}\mathbf{S}_b^{\phi}\boldsymbol{w} = \boldsymbol{w}^{\top}\left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)\left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)^{\top}\boldsymbol{w} = \boldsymbol{\alpha}^{\top}\underbrace{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^{\top}}_{M}\boldsymbol{\alpha}$$

核线性判别分析

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i) \qquad \boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^\top K(:, \boldsymbol{x})$$

$$\begin{aligned}
\boldsymbol{w}^\top \boldsymbol{S}_w^\phi \boldsymbol{w} &= \sum_{i=0}^{1} \sum_{x \in X_i} \boldsymbol{w}^\top \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^\phi\right)\left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^\phi\right)^\top \boldsymbol{w} \\
&= \sum_{i=1}^{m} \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i)^\top \boldsymbol{w} - \sum_{i=0}^{1} m_i \boldsymbol{w}^\top \boldsymbol{\mu}_i^\phi \left(\boldsymbol{\mu}_i^\phi\right)^\top \boldsymbol{w} \\
&= \boldsymbol{\alpha}^\top \left(\boldsymbol{K}\boldsymbol{K}^\top - \sum_{i=0}^{1} m_i \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^\top \right) \boldsymbol{\alpha}
\end{aligned}$$

$$\sum_{i=1}^{m} \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i)^\top \boldsymbol{w} = \sum_{i=1}^{m} \boldsymbol{\alpha}^\top K(:, \boldsymbol{x}_i) K(:, \boldsymbol{x}_i)^\top \boldsymbol{\alpha} = \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{K}^\top \boldsymbol{\alpha}$$

$$\boldsymbol{w}^\top \boldsymbol{\mu}_i^\phi = \frac{1}{m_i} \sum_{x \in X_i} \boldsymbol{w}^\top \phi(\boldsymbol{x}) = \frac{1}{m_i} \sum_{x \in X_i} \boldsymbol{\alpha}^\top K(:, \boldsymbol{x}) = \boldsymbol{\alpha}^\top \underbrace{\frac{1}{m_i} \boldsymbol{K} \boldsymbol{1}_i}_{\hat{\boldsymbol{\mu}}_i}$$

$$\sum_{i=0}^{1} m_i \boldsymbol{w}^\top \boldsymbol{\mu}_i^\phi \left(\boldsymbol{\mu}_i^\phi\right)^\top \boldsymbol{w} = \sum_{i=0}^{1} m_i \boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^\top \boldsymbol{\alpha} = \boldsymbol{\alpha}^\top \left(\sum_{i=0}^{1} m_i \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^\top \right) \boldsymbol{\alpha}$$

作业

- 6.4
- 6.6
- 6.9

支持向量回归的对偶问题如下，

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

$$s.t.\ C \geqslant \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}} \geqslant 0 \text{ and } \sum_{i=1}^{m}(\alpha_i-\hat{\alpha}_i)=0$$

请将该问题转化为类似于如下标准型的形式（$\boldsymbol{u},\boldsymbol{v},\boldsymbol{K}$均已知），

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top\boldsymbol{v}-\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}$$

$$s.t.\ C \geqslant \boldsymbol{\alpha} \geqslant 0 \text{ and } \boldsymbol{\alpha}^\top\boldsymbol{u}=0$$

例如在软间隔SVM中$\boldsymbol{v}=\mathbf{1},\boldsymbol{u}=\boldsymbol{y},\boldsymbol{K}[i,j]=y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.