线性SVM

软间隔最大化

软间隔

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

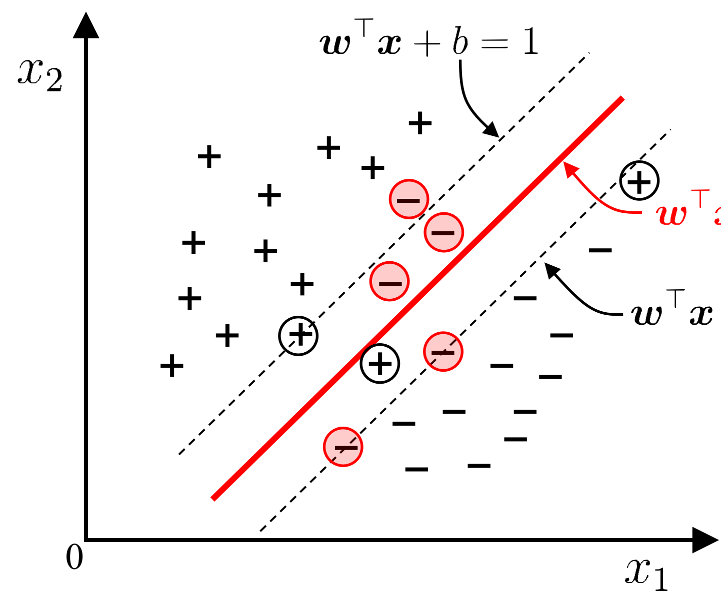
-Q:对于线性可分问题，硬间隔最大化非常好。但如果样本中的噪声和特异点使训练数据线性不可分呢？

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

$$\text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1, 2, \cdots, N$$

-A:引入"软间隔"的概念，允许支持向量机在一些样本上不满足间隔≥1的约束，即对每个样本点引入一个松弛变量 $\xi_i \geqslant 0$。因此，约束条件变为

$$y_i(w \cdot x_i + b) \geqslant 1 - \xi_i$$



不满足约束的样本

软间隔最大化

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

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

$$\text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1, 2, \cdots, N$$

- 线性不可分情况下的线性SVM的学习问题变为

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

$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \cdots, N$$

$$\xi_i \geqslant 0, \quad i = 1, 2, \cdots, N$$

- 优化目标的两层含义：

  1. 间隔尽量大　　2. 不满足约束样本尽可能少

- C是惩罚参数，超参数，起调和作用

拉格朗日对偶

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

- 原始优化问题的拉格朗日函数是

$$L(w,b,\xi,\alpha,\mu) \equiv \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(w\cdot x_i+b)-1+\xi_i) - \sum_{i=1}^{N}\mu_i\xi_i$$

$$\alpha_i \geqslant 0, \mu_i \geqslant 0$$

代入

- 对偶问题是拉格朗日函数的极大极小问题

1. 求 $L(w,b,\xi,\alpha,\mu)$ 对w,b,ξ的极小

$$\nabla_w L(w,b,\xi,\alpha,\mu) = w - \sum_{i=1}^{N}\alpha_i y_i x_i = 0$$

$$\nabla_b L(w,b,\xi,\alpha,\mu) = -\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\nabla_{\xi_i} L(w,b,\xi,\alpha,\mu) = C - \alpha_i - \mu_i = 0$$

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$C - \alpha_i - \mu_i = 0$$

拉格朗日对偶

- 得到 $$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

- 再对 $\min\limits_{w,b,\xi} L(w,b,\xi,\alpha,\mu)$ 求α的极大，得到对偶问题：

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

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

$$\left.\begin{aligned} & C - \alpha_i - \mu_i = 0 \\ & \alpha_i \geqslant 0 \\ & \mu_i \geqslant 0, \quad i = 1,2,\cdots,N \end{aligned}\right\} \quad 0 \leqslant \alpha_i \leqslant C$$

$$\boxed{\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & 0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,\cdots,N \end{aligned}}$$

- 设 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)^{\mathrm{T}}$ 是对偶问题的一个解，若存在α*的一个分量 α_j^*，$0 < \alpha_j^* < C$，则原始问题的解w*,b*可按下式求得

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i \qquad b^* = y_j - \sum_{i=1}^{N} y_i \alpha_i^* (x_i \cdot x_j)$$

KKT条件

$$\nabla_w L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = w^* - \sum_{i=1}^{N}\alpha_i^* y_i x_i = 0 \longrightarrow w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$$

$$\nabla_b L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = -\sum_{i=1}^{N}\alpha_i^* y_i = 0$$

$$\nabla_\xi L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = C - \alpha^* - \mu^* = 0$$

$$\alpha_i^*(y_i(w^*\cdot x_i + b^*) - 1 + \xi_i^*) = 0$$

$$\mu_i^*\xi_i^* = 0$$

$$y_i(w^*\cdot x_i + b^*) - 1 + \xi_i^* \geqslant 0$$

$$\xi_i^* \geqslant 0$$

$$\alpha_i^* \geqslant 0$$

$$\mu_i^* \geqslant 0, \quad i = 1,2,\cdots,N$$

若存在 α_j^*，$0 < \alpha_j^* < C$，

则 $y_j(w^*\cdot x_j + b^*) - 1 = 0$

那么 $b^* = y_j - \sum_{i=1}^{N} y_i\alpha_i^*(x_i\cdot x_j)$

分离超平面：

$$\sum_{i=1}^{N}\alpha_i^* y_i(x\cdot x_i) + b^* = 0$$

分类决策函数：

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i(x\cdot x_i) + b^*\right)$$

线性SVM算法

- 输入：线性不可分训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

  $x_i \in \mathcal{X}=\mathbf{R}^n \qquad y_i \in \mathcal{Y}=\{-1,+1\}, \quad i=1,2,\cdots,N$

- 输出：分离超平面和分类决策函数

1、构造并求解约束最优化问题

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

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

$$\boxed{0 \leqslant \alpha_i \leqslant C, \quad i=1,2,\cdots,N}$$

求得最优解：$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$

线性SVM算法

2、计算 $w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$

并选择α^*的一个分量α_j^*，适合条件 $0 < \alpha_j^* < C$ ，计算

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

3、求得分离超平面

$$w^* \cdot x + b^* = 0$$

分类决策函数

$$f(x) = \mathrm{sign}(w^* \cdot x + b^*)$$

对任一适合条件 $0 < \alpha_j^* < C$ 的α_j^*，都可以求出b^*，理论上线性不可分情况的b^*不唯一。

软间隔的支持向量

$$\nabla_w L(w^*,b^*,\xi^*,\alpha^*,\mu^*)=w^*-\sum_{i=1}^{N}\alpha_i^* y_i x_i=0$$

$$\nabla_b L(w^*,b^*,\xi^*,\alpha^*,\mu^*)=-\sum_{i=1}^{N}\alpha_i^* y_i=0$$

$$\nabla_\xi L(w^*,b^*,\xi^*,\alpha^*,\mu^*)=C-\alpha^*-\mu^*=0$$

$$\alpha_i^*(y_i(w^*\cdot x_i+b^*)-1+\xi_i^*)=0$$

$$\mu_i^*\xi_i^*=0$$

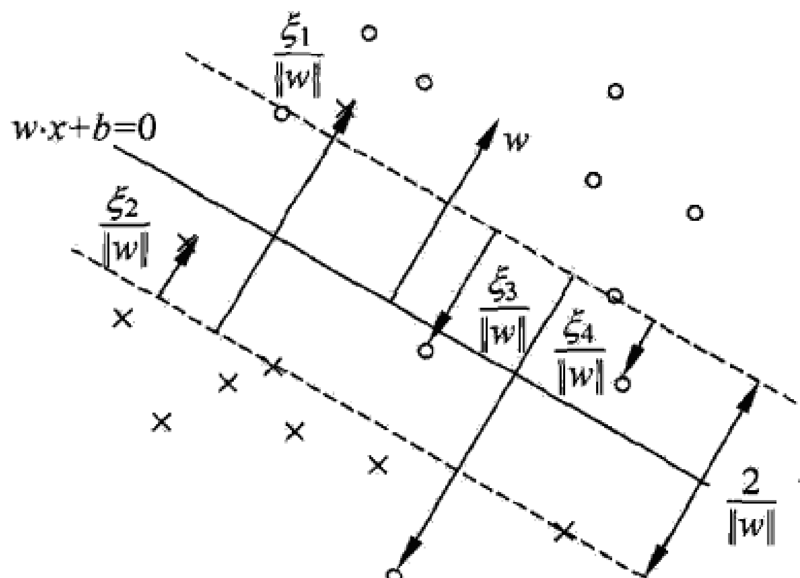
$$y_i(w^*\cdot x_i+b^*)-1+\xi_i^*\geqslant 0$$

$$\xi_i^*\geqslant 0$$

$$\alpha_i^*\geqslant 0$$

$$\mu_i^*\geqslant 0,\quad i=1,2,\cdots,N$$

- 对任一训练样本，总有$\alpha_i^*=0$或 $y_i f(\boldsymbol{x}_i)=1-\xi_i$
- $\alpha_i^*=0$，该样本对分类面无影响
- $\alpha_i^*>0$，支持向量



软间隔的支持向量

$$\nabla_w L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = w^* - \sum_{i=1}^{N} \alpha_i^* y_i x_i = 0$$

$$\nabla_b L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = -\sum_{i=1}^{N} \alpha_i^* y_i = 0$$

$$\nabla_\xi L(w^*,b^*,\xi^*,\alpha^*,\mu^*) = C - \alpha^* - \mu^* = 0$$

$$\alpha_i^*(y_i(w^* \cdot x_i + b^*) - 1 + \xi_i^*) = 0$$

$$\mu_i^* \xi_i^* = 0$$

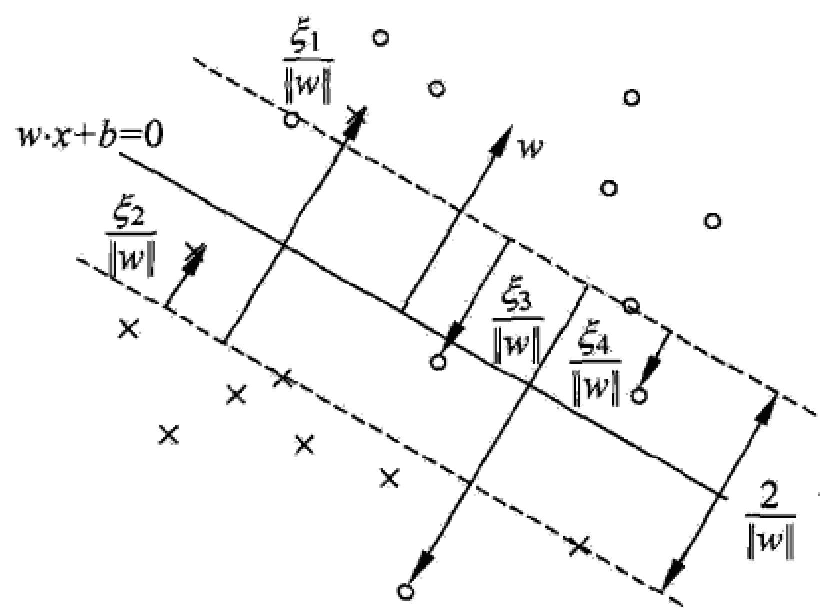
$$y_i(w^* \cdot x_i + b^*) - 1 + \xi_i^* \geqslant 0$$

$$\xi_i^* \geqslant 0$$

$$\alpha_i^* \geqslant 0$$

$$\mu_i^* \geqslant 0, \quad i = 1,2,\cdots,N$$

- 若$\alpha_i^* < C$，则 $\xi_i = 0$，间隔边界上
- 若$\alpha_i^* = C$， $0 < \xi_i < 1$，分类正确，在间隔边界与超平面之间
- 若$\alpha_i^* = C$， $\xi_i = 1$，在超平面上
- 若$\alpha_i^* = C$， $\xi_i > 1$，误分类



另一种解释

- 最小化以下目标函数：最大化间隔的同时，让不满足约束的样本应尽可能少

$$\sum_{i=1}^{N}[1-y_i(w\cdot x_i+b)]_+ +\lambda\|w\|^2$$

- 合页损失函数（hinge loss）

$$[z]_+=\begin{cases} z, & z>0 \\ 0, & z\leqslant 0 \end{cases}$$

- 0/1损失不易优化求解
- 合页损失为“替代损失”

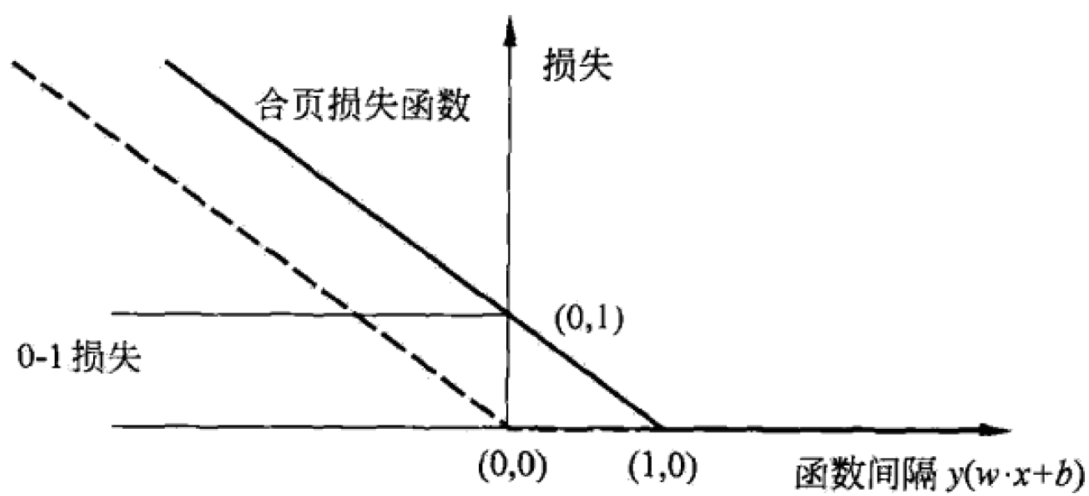


图 7.6　合页损失函数

思考

- 线性SVM原始最优化问题

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

$$\text{s.t.} \quad y_i(w\cdot x_i + b) \geqslant 1-\xi_i\,, \quad i=1,2,\cdots,N$$

$$\xi_i \geqslant 0\,, \quad i=1,2,\cdots,N$$

- 等价于最优化问题

$$\min_{w,b} \quad \sum_{i=1}^{N}[1-y_i(w\cdot x_i + b)]_+ + \lambda\|w\|^2$$

正则化

$$\min_{w,b} \lambda\|w\|^2+\sum_{i=1}^{N}\left[1-y_i\left(w\cdot x_i+b\right)\right]_+$$

描述间隔大小

描述训练集上的误差

- 支持向量机学习模型的更一般形式

$$\min_{f} \quad \Omega(f)+C\sum_{i=1}^{m} l(f(\boldsymbol{x}_i),y_i)$$

结构风险，描述模型的某些性质，正则化项

经验风险，描述模型与训练数据的契合程度

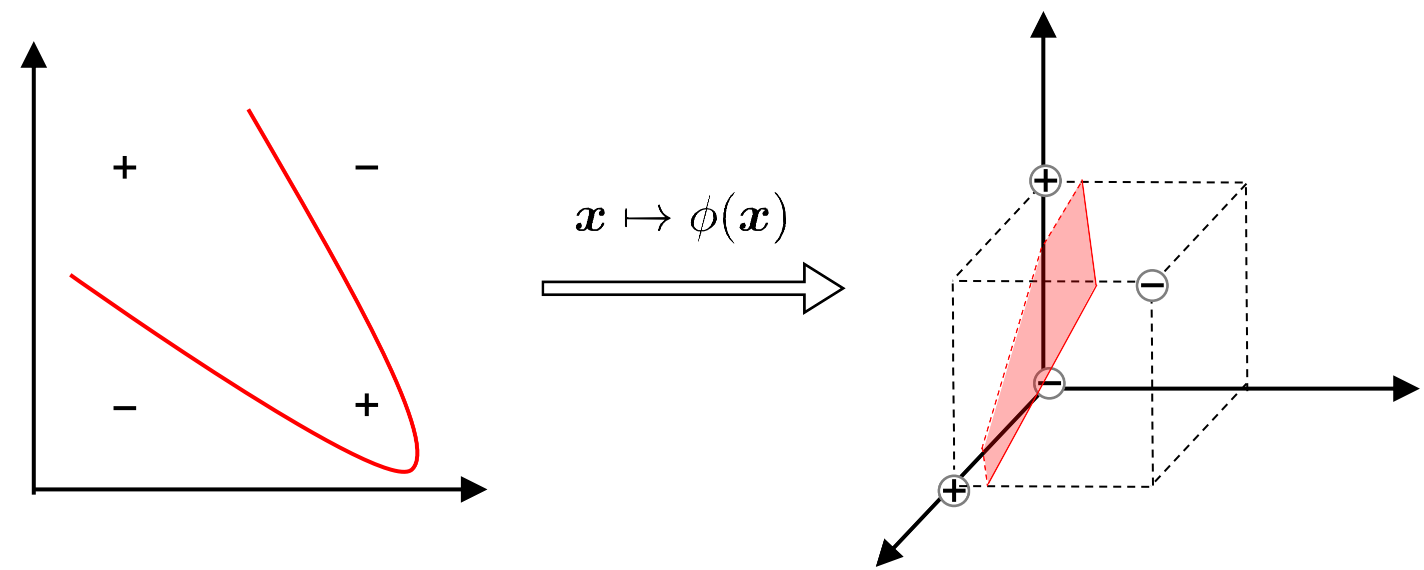
- 通过替换上面两个部分，可以得到许多其他学习模型
  - 对数几率回归(Logistic Regression)
  - 最小绝对收缩选择算子(LASSO)
  - ……

核函数与核方法

线性不可分

-Q:若不存在一个能正确划分两类样本的线性超平面，怎么办?

-A: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$f(\boldsymbol{x})=\boldsymbol{w}^{\top}\phi(\boldsymbol{x})+b$.

原始问题

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

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

对偶问题

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

$$\text{s.t.}\ \ \sum_{i=1}^{m}\alpha_i y_i=0,\ \ \alpha_i\geq 0,\ \ i=1,2,\ldots,m.$$

只以内积的形式出现

预测

$$f(\boldsymbol{x})=\boldsymbol{w}^{\top}\phi(\boldsymbol{x})+b=\sum_{i=1}^{m}\alpha_i y_i \boxed{\phi(\boldsymbol{x}_i)^{\top}\phi(\boldsymbol{x})}+b$$

核函数

- $\phi(x)$一般是无限维，可以不知道$\phi(x)$的显式表达，只要知道一个如下所示的核函数，则优化式依然可解。

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

- Mercer定理(充分非必要)：只要一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用.

- 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，$\beta > 0$, $\theta < 0$

核函数

- 特征空间的选择对SVM的性能至关重要
- $\phi(x)$未知，难以确定什么样的核函数是合适的
- 文本数据通常采用线性核，情况不明时可先尝试高斯核

- 若κ_1和κ_2为核函数，
  - 对于任意正数γ_1、γ_2，$\gamma_1\kappa_1 + \gamma_2\kappa_2$也是核函数
  - 直积$\kappa_1 \otimes \kappa_2 = \kappa_1(x,z)\kappa_2(x,z)$也是核函数
  - 对于任意函数$g(x)$，$\kappa(x,z) = g(x)\kappa_1(x,z)g(z)$也是核函数

表示定理

SVM决策函数 $$f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

结论：不考虑偏移项b，SVM学得的模型可以表示成核函数的线性组合.

- 更一般的结论(表示定理)：对于任意单调增函数 Ω 和任意非负损失函数 l，优化问题

$$\min_{h \in \mathbb{H}} \quad F(h) = \Omega(\boxed{\|h\|_{\mathbb{H}}}) + l(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_m))$$

$\mathbb{H}$空间中关于h的范数

的解总可以写为 $h^* = \sum_{i=1}^{m} \alpha_i \kappa(\cdot, \boldsymbol{x}_i)$.

- 适用于一般的损失函数和正则项，显式出核函数的巨大威力。

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

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

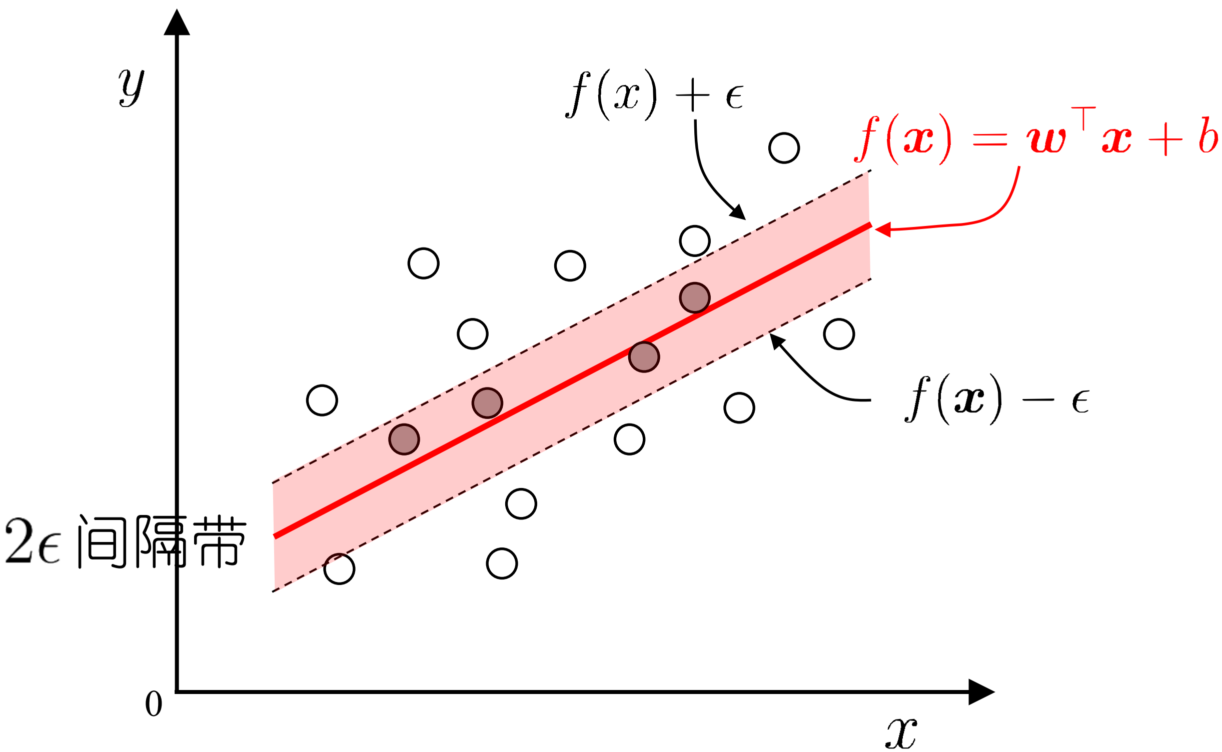
$$h(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x})$$

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

支持向量回归

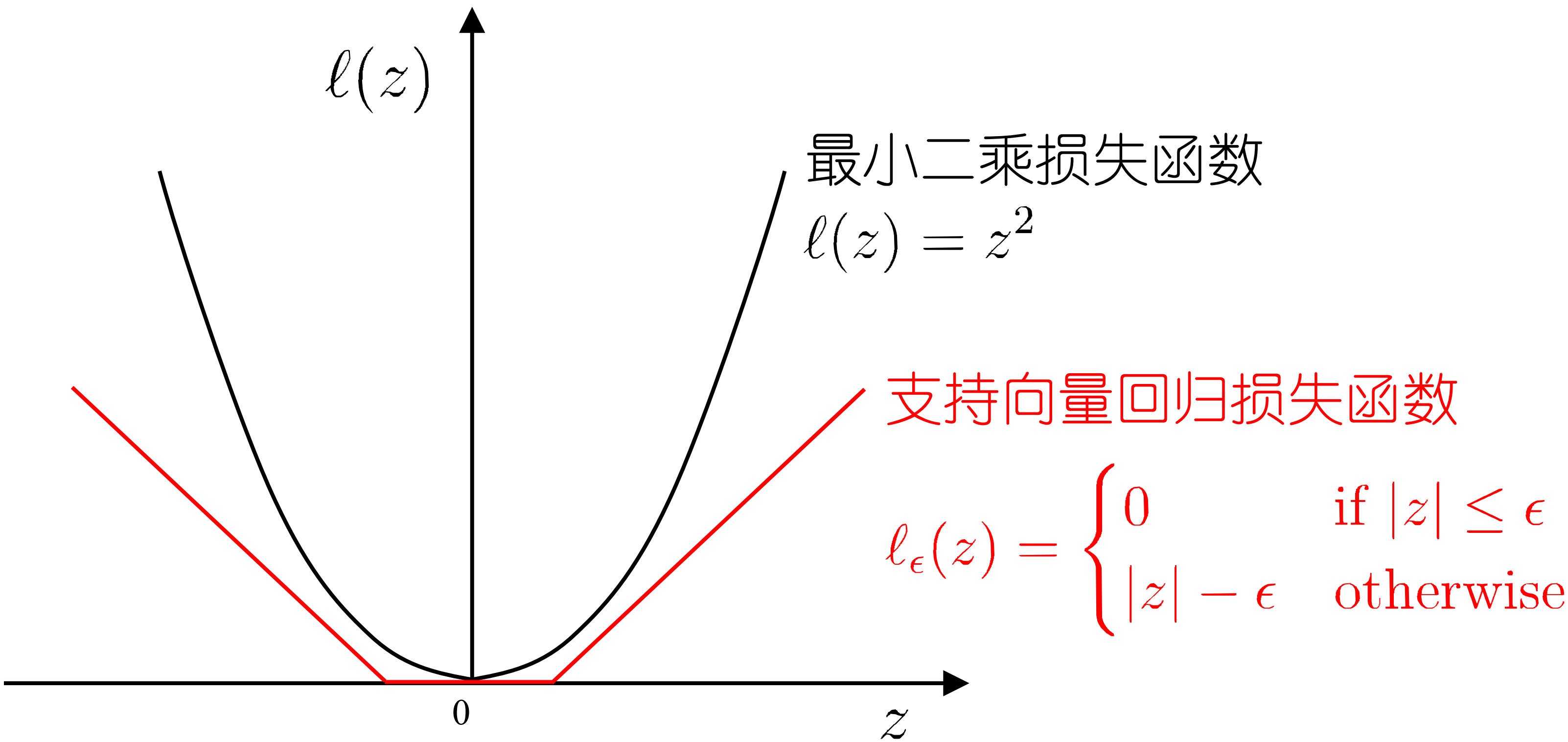
支持向量回归

特点：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.



形式化

原始问题

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

对偶问题

$$\begin{aligned}\min_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{m}(\alpha_i(\epsilon - y_i) + \hat{\alpha}_i(\epsilon + y_i)) \\ \text{s.t.} \quad & \sum_{i=1}^{m}(\alpha_i - \hat{\alpha}_i) = 0, \\ & 0 \leq \alpha_i \leq C,\ 0 \leq \hat{\alpha}_i \leq C.\end{aligned}$$

预测

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top\phi(\boldsymbol{x}) + b = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)y_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

Take Home Message

- 支持向量机的"最大间隔"思想
- 对偶问题及其解的稀疏性
- 引入"软间隔"缓解特征空间中线性不可分的问题
- 通过向高维空间映射解决线性不可分的问题
- 将核方法推广到其他学习模型
- 将支持向量的思想应用到回归问题上得到支持向量回归

成熟的SVM软件包

- LIBSVM
  http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- LIBLINEAR
  http://www.csie.ntu.edu.tw/~cjlin/liblinear/
- SVM^{light}、SVM^{perf}、SVM^{struct}
  http://svmlight.joachims.org/svm_struct.html
- Pegasos
  http://www.cs.huji.ac.il/~shais/code/index.html