

Machine Learning

# 支持向量机 (Support Vector Machines)

梁毅雄

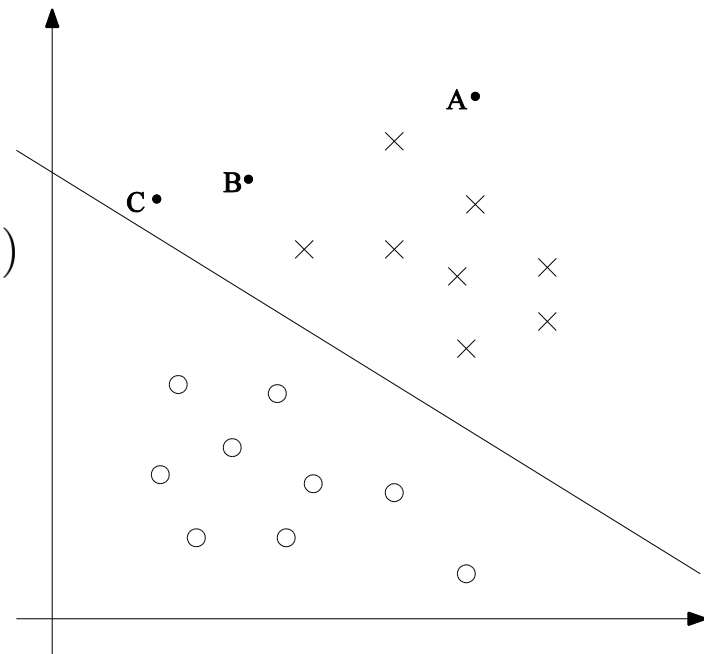
[yxliang@csu.edu.cn](mailto:yxliang@csu.edu.cn)

Some materials from Andrew Ng, Barnabás Póczos and others

# Margins: Intuition

Logistic regression:  $p(y = 1|x; \theta) = h_{\theta}(x) = g(\theta^T x)$

在测试新样本时，当  $\theta^T x \gg 0$  或者  $\theta^T x \ll 0$   
我们可以“very confident”给出预测结果



While in training, we'd have found a good fit to the training data if we can find  $\theta$  so that  $\theta^T x^{(i)} \gg 0$  whenever  $y^{(i)} = 1$ , and  $\theta^T x^{(i)} \ll 0$  whenever  $y^{(i)} = 0$

# Margins: Intuition

重新定义符号如下：

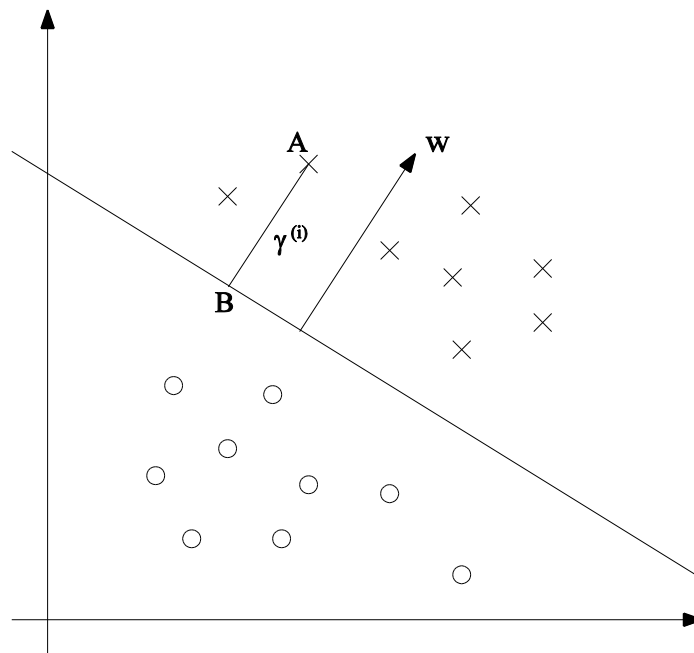
$$y = \{0, 1\} \Rightarrow y = \{-1, +1\}$$

$$\theta_0 \Rightarrow b$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \Rightarrow w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

“Confidence”:  $\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$

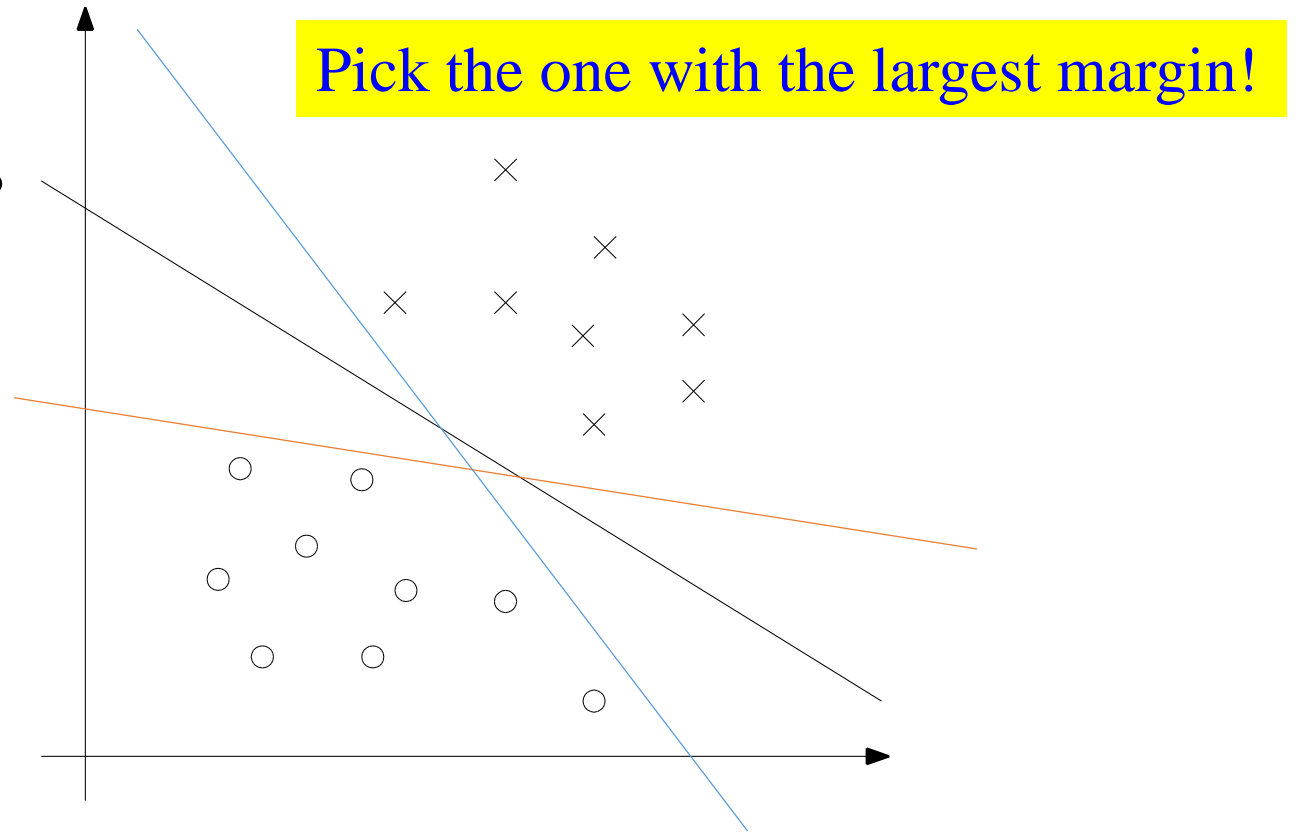
$$y = \text{sign}(w^T x + b) = \begin{cases} +1, & w^T x + b > 0 \\ -1, & w^T x + b < 0 \end{cases}$$



正确分类的条件:  $y^{(i)}(w^T x^{(i)} + b) > 0$

# Margins: Intuition

Which line is better?



# 最大间隔分类器(Max Margin Classifier)

Pick the one with the largest margin!

如何计算margin?

$$w^T x^+ + b = +1 \quad w^T x^- + b = -1$$

$$x^+ = x^- + \lambda w$$

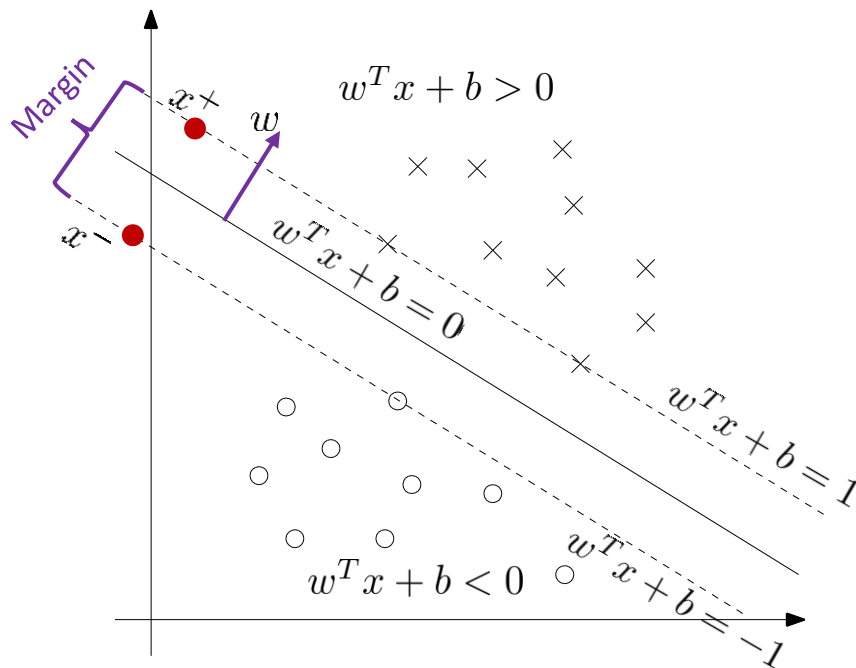
$$\text{margin} = \|x^+ - x^-\| = ?$$

$$w^T (x^+ - x^-) = 2$$

$$\lambda = \frac{2}{w^T w}$$

$$\text{margin} = \|x^+ - x^-\| = \|\lambda w\| = \frac{2}{\|w\|}$$

$$\text{最大化margin: } \max_w \frac{2}{\|w\|} \quad \text{等价于} \quad \min_w \frac{1}{2} \|w\|^2$$



# The Primal Hard SVM

假设数据线性可分，即  $y^{(i)}(w^T x^{(i)} + b) \geq 1$

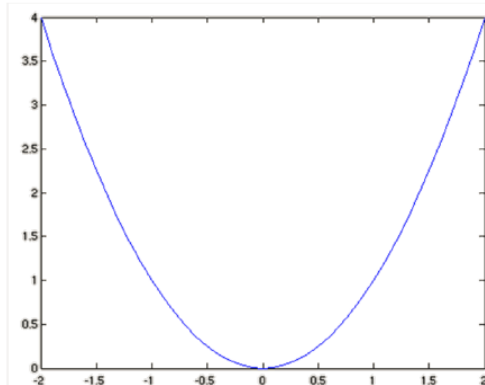
$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m \end{aligned}$$

- 属于带约束的优化问题(线性约束条件 + 二次目标函数)
- 典型的二次规划 ( Quadratic Programming, QP ) 问题
- Efficient algorithms and commercial code exist for QP

# Constrained Optimization

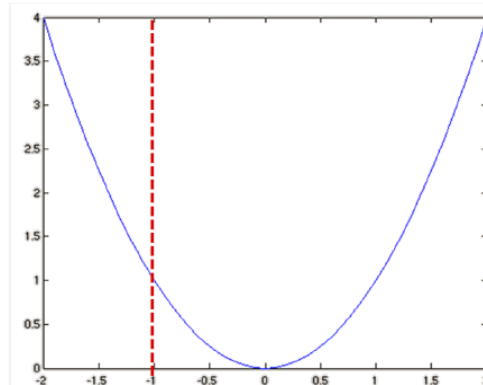
$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & x \geq b \end{aligned}$$

$$\min_x x^2$$



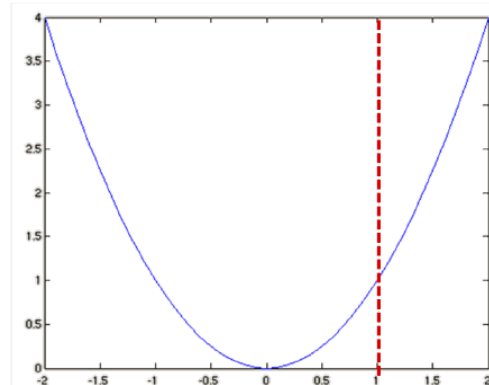
$$x^* = 0$$

$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & x \geq -1 \end{aligned}$$



$$x^* = 0$$

$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & x \geq 1 \end{aligned}$$



$$x^* = 1$$

# Lagrange Multiplier

$$\begin{array}{ll}\min_w & f(w) \\ \text{s.t.} & h_i(w) = 0, \quad i = 1, \dots, l.\end{array}$$

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^l \beta_i h_i(w)$$

$\beta_i$ : **Lagrange multipliers.**

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$



# Lagrange Multiplier

$$\begin{array}{ll} \min_w & f(w) \\ \text{s.t.} & g_i(w) \leq 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l. \end{array} \quad \mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w).$$

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta). \quad \alpha_i, \beta_i: \text{Lagrange multipliers.}$$

Let some  $w$  be given. If  $w$  violates any of the primal constraints (i.e., if either  $g_i(w) > 0$  or  $h_i(w) \neq 0$  for some  $i$ ), then you should be able to verify that

$$\begin{aligned} \theta_{\mathcal{P}}(w) &= \max_{\alpha, \beta: \alpha_i \geq 0} f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w) \\ &= \infty. \end{aligned}$$

# Lagrange Multiplier

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) \quad \mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

等价于原来的优化问题 (Primal problem)

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta)$$

$$\max_{\alpha, \beta: \alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

对偶优化问题 (Dual problem)

$$d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^* \quad \text{弱对偶性}$$

# Lagrange Multiplier

$$d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*$$

Under certain conditions, we will have  $d^* = p^*$  强对偶性

- $f$  and the  $g_i$ 's are convex (its Hessian is positive semi-definite)
- $h_i$ 's are affine, i.e., there exists  $a_i, b_i$ , so that  $h_i(w) = a_i^T w + b_i$

Under our above assumptions, there must exist  $w^*, \alpha^*, \beta^*$  so that  $w^*$  is the solution to the primal problem,  $\alpha^*, \beta^*$  are the solution to the dual problem, and moreover  $p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$ .

# Lagrange Multiplier

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l. \end{aligned}$$

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w)$$

Moreover,  $w^*, \alpha^*$  and  $\beta^*$  satisfy the **Karush-Kuhn-Tucker (KKT) conditions**, which are as follows:

$$\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, n$$

$$\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, l$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \dots, k$$

$$g_i(w^*) \leq 0, \quad i = 1, \dots, k$$

$$\alpha^* \geq 0, \quad i = 1, \dots,$$

The KKT **dual complementarity** condition:  
If  $\alpha_i^* > 0$ , then  $g_i(w^*) = 0$

Moreover, if some  $w^*, \alpha^*, \beta^*$  satisfy the KKT conditions, then it is also a solution to the primal and dual problems.

# Lagrange Multiplier

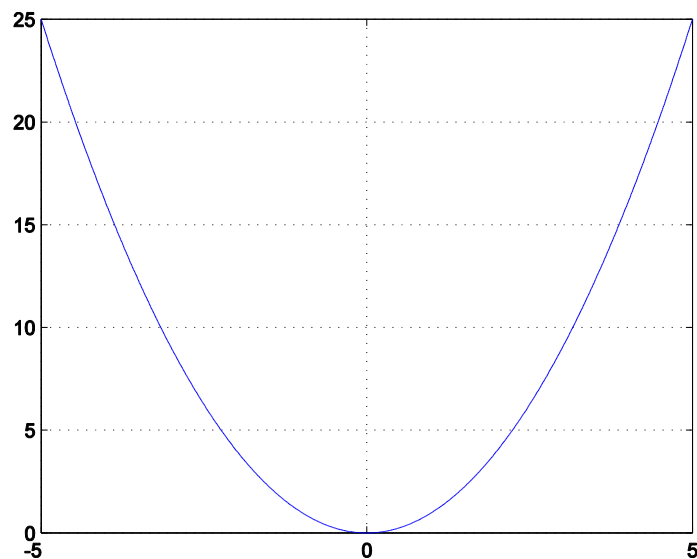
$$\min_w f(w)$$

$$\text{s.t. } g_i(w) \leq 0, \quad i = 1, \dots, k$$

$$h_i(w) = 0, \quad i = 1, \dots, l.$$

$$\max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w)$$



$$\min_x f(x) = x^2$$

$$\text{s.t. } x \geq b \quad \text{满足强对偶条件}$$

$$\mathcal{L}(x, \alpha) = x^2 - \alpha(x - b)$$

$$\min_x \max_{\alpha} \mathcal{L}(x, \alpha)$$

$$\text{s.t. } \alpha \geq 0$$

# Lagrange Multiplier

$$\min_x \quad f(x) = x^2$$

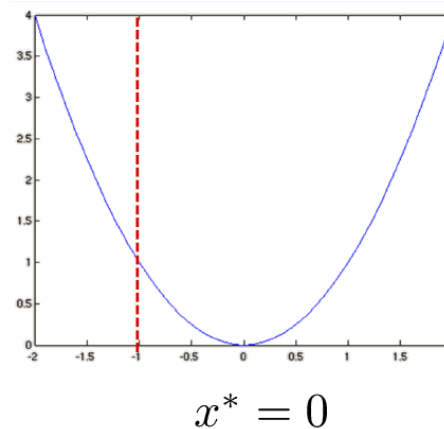
$$\text{s.t.} \quad x \geq b$$

$$\begin{aligned} \min_x \max_{\alpha} \quad & \mathcal{L}(x, \alpha) = x^2 - \alpha(x - b) \\ \text{s.t.} \quad & \alpha \geq 0 \end{aligned}$$

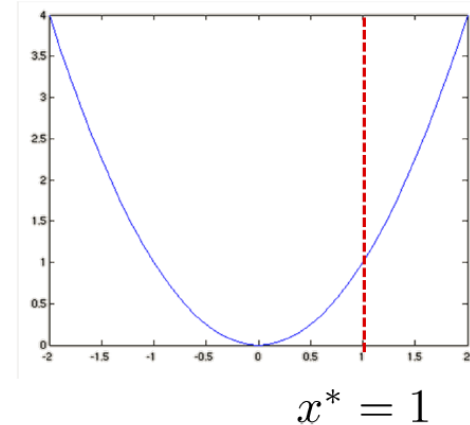
$$\frac{\partial \mathcal{L}}{\partial x} = 0 \Rightarrow x^* = \frac{\alpha}{2}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = 0 \Rightarrow \alpha^* = \max(2b, 0)$$

$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & x \geq -1 \end{aligned}$$



$$\begin{aligned} \min_x \quad & x^2 \\ \text{s.t.} \quad & x \geq 1 \end{aligned}$$



# SVM: From Primal to Dual

Primal problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m \end{aligned}$$

Lagrange function:

满足强对偶条件

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i \left[ y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0 \Rightarrow w^* = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial}{\partial b} \mathcal{L}(w, b, \alpha) = \sum_{i=1}^m \alpha_i^* y^{(i)} = 0$$

# SVM: From Primal to Dual

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i \left[ y^{(i)} (w^T x^{(i)} + b) - 1 \right]$$



$$w^* = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}$$

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b \sum_{i=1}^m \alpha_i y^{(i)}$$



$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$



# Solving the Dual: The SMO Algorithm

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

- 这仍然是一个二次规划问题，可采用通用的QP算法求解，但其规模正比于训练样本数
- Sequential Minimal Optimization (SMO) algorithm是一种高效算法，基本思想是采用坐标下降法，在更新 $\alpha_i$ 时固定其他 $\alpha_j$ ，由于 $\sum_{i=1}^m \alpha_i y^{(i)} = 0$ ， $\alpha_i$ 的值可由其他的 $\alpha_j$  ( $j \neq i$ )表示

# The Dual Hard SVM

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

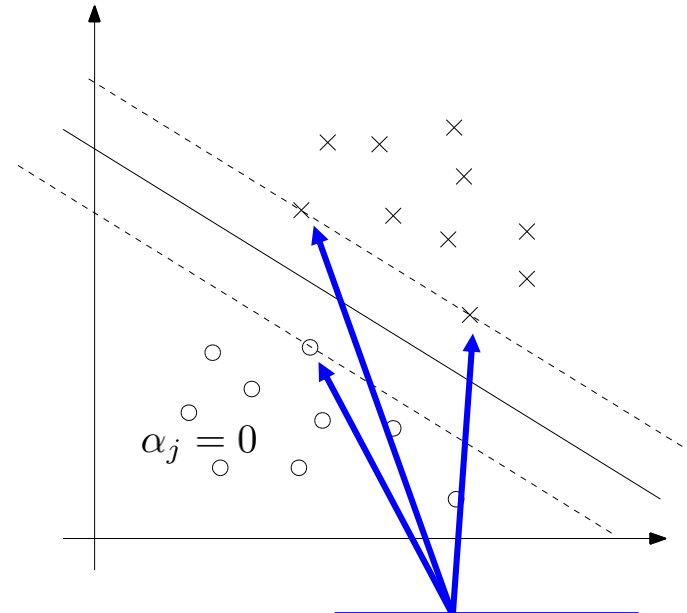
得到上式的最优解  $\alpha^*$  后, 可代入  $w^* = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)}$  得到最优的  $w^*$   
如何求  $b$  ?

$$b^* = - \frac{\max_{i: y^{(i)} = -1} w^{*T} x^{(i)} + \min_{i: y^{(i)} = 1} w^{*T} x^{(i)}}{2}$$

$$\text{分类:} \quad y = \text{sign}(w^T x + b) = \begin{cases} +1, & w^T x + b > 0 \\ -1, & w^T x + b < 0 \end{cases}$$

# The Dual Hard SVM

$$\begin{aligned}w^T x + b &= \left( \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\&= \sum_{i=1}^m \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b\end{aligned}$$



- The KKT **dual complementarity** condition:

$$\alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] = 0, \forall i \Rightarrow \begin{cases} \alpha_i > 0, & y^{(i)}(w^T x^{(i)} + b) - 1 = 0 \\ \alpha_i = 0, & y^{(i)}(w^T x^{(i)} + b) - 1 > 0 \end{cases}$$

$\alpha_i > 0$  支持向量

- $y = \text{sign} \left( \sum_{i=1}^m \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b \right)$

- 一般情况下只有少数训练样本对应的Lagrange Multiplier大于零(支持向量), 分类面则是由这些支持向量决定
- 决策时只需计算新样本与所有支持向量的内积

# From Hard SVM to Soft SVM

$$(w_{\text{hard}}^*, b_{\text{hard}}^*) = \arg \min_{w, b} \frac{1}{2} ||w||^2$$

s.t.  $y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m$

$$(w_{\text{hard}}^*, b_{\text{hard}}^*) = \arg \min_{w, b} \sum_{i=1}^m \ell_{0-\infty} \left( y^{(i)}(w^T x^{(i)} + b) \geq 1 \right) + \frac{1}{2} ||w||^2$$

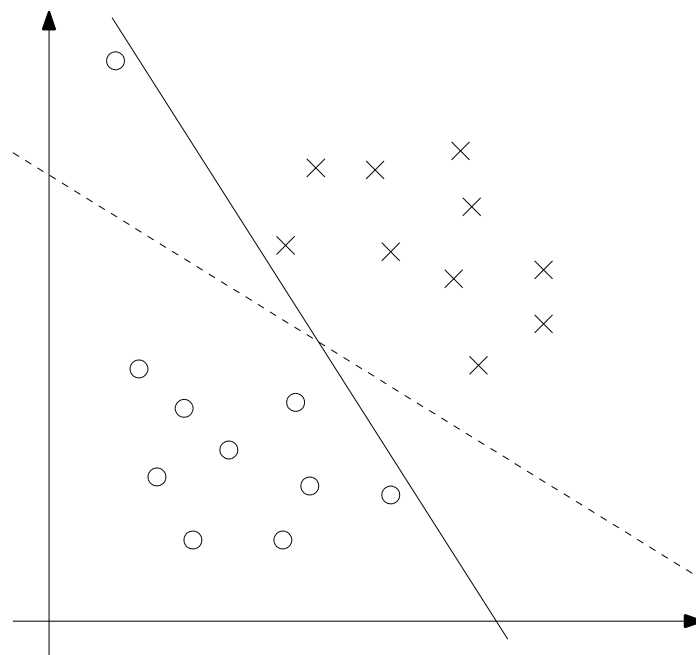
$$\ell_{0-\infty} \left( y^{(i)}(w^T x^{(i)} + b) \geq 1 \right) = \begin{cases} 0, & \text{if } y^{(i)}(w^T x^{(i)} + b) \geq 1 \\ \infty, & \text{otherwise} \end{cases}$$

$$J(\theta) = L(\theta) + \lambda R(\theta)$$


# From Hard SVM to Soft SVM

$$(w_{\text{hard}}^*, b_{\text{hard}}^*) = \arg \min_{w, b} \sum_{i=1}^m \ell_{0-\infty} \left( y^{(i)} (w^T x^{(i)} + b) \geq 1 \right) + \frac{1}{2} \|w\|^2$$

- 仅能处理线性可分问题
- 实际情况下训练数据中可能存在“特异点”(outlier)，把这些点去掉后数据是线性可分的
- 或者是去掉这些数据后，能得到更大的margin

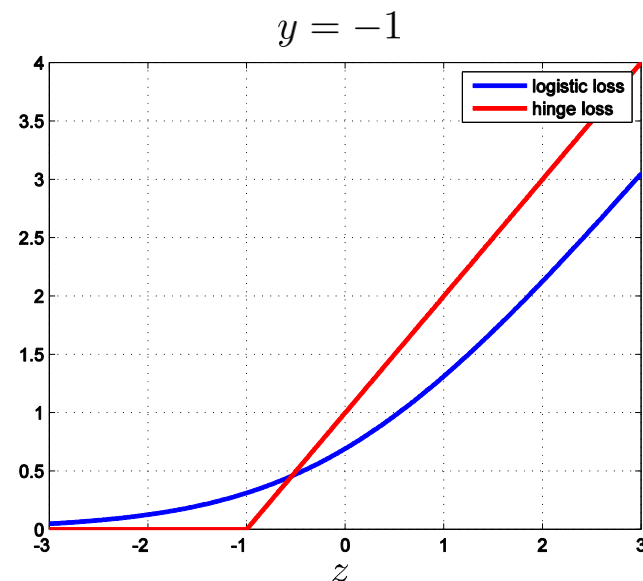
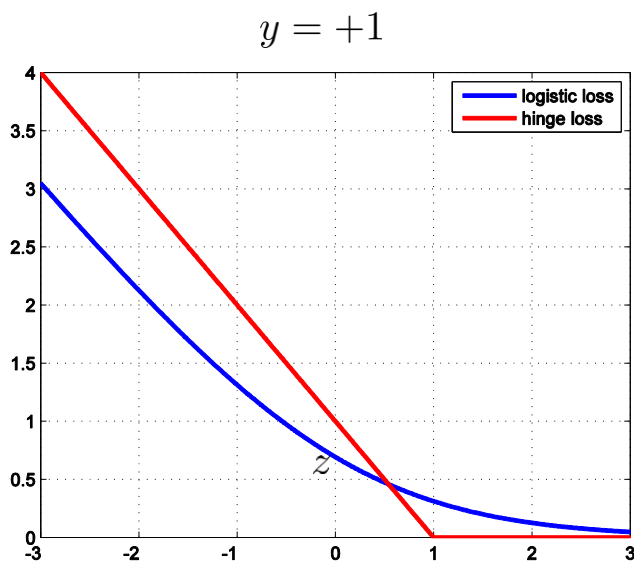


# From Logistic Loss to Hinge Loss

$$z = \theta^T x = w^T x + b$$

$$\text{Logistic Loss: } \ell = \begin{cases} -\log\left(\frac{1}{1+e^{-z}}\right) = \log(1 + e^{-z}), & y = +1 \\ -\log\left(1 - \frac{1}{1+e^{-z}}\right) = \log(1 + e^z), & y = -1 \end{cases} \Rightarrow \ell = \log(1 + e^{-yz})$$

$$\text{Hinge Loss: } \ell = \max(1 - yz, 0)$$



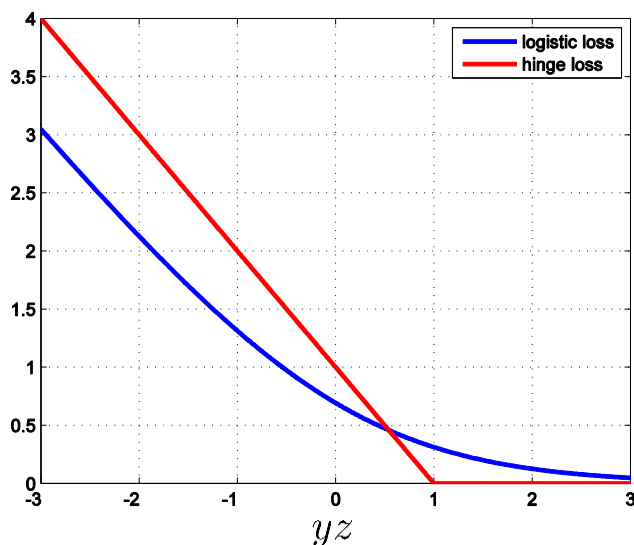
# From Logistic Loss to Hinge Loss

$$z = \theta^T x = w^T x + b$$

$$\text{Logistic Loss: } \ell = \begin{cases} -\log\left(\frac{1}{1+e^{-z}}\right) = \log(1 + e^{-z}), & y = +1 \\ -\log\left(1 - \frac{1}{1+e^{-z}}\right) = \log(1 + e^z), & y = -1 \end{cases}$$

$$\Rightarrow \ell = \log(1 + e^{-yz})$$

$$\text{Hinge Loss: } \ell = \max(1 - yz, 0)$$



# The Primal Soft SVM problem

$$(w_{\text{hard}}^*, b_{\text{hard}}^*) = \arg \min_{w, b} \sum_{i=1}^m \ell_{0-\infty} \left( y^{(i)} (w^T x^{(i)} + b) \geq 1 \right) + \frac{1}{2} \|w\|^2$$


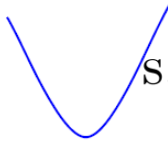
↓ 换成hinge loss

$$(w_{\text{soft}}^*, b_{\text{soft}}^*) = \arg \min_{w, b} C \sum_{i=1}^m \max \left( 1 - y^{(i)} (w^T x^{(i)} + b), 0 \right) + \frac{1}{2} \|w\|^2$$

$\xi_i = \max \left( 1 - y^{(i)} (w^T x^{(i)} + b), 0 \right)$  松弛因子

↓ 等价于

$$(w_{\text{soft}}^*, b_{\text{soft}}^*) = \arg \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \quad C: \text{惩罚因子}$$

 +  s.t.

$$y^{(i)} (w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, m$$

$$\xi_i \geq 0, \quad i = 1, \dots, m.$$



# The Dual Soft SVM

$$\begin{aligned}(w_{\text{soft}}^*, b_{\text{soft}}^*) &= \arg \min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad &y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, m \\ &\xi_i \geq 0, \quad i = 1, \dots, m.\end{aligned}$$

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i \left[ y^{(i)}(x^T w + b) - 1 + \xi_i \right] - \sum_{i=1}^m r_i \xi_i.$$

另 $\mathcal{L}(w, b, \xi, \alpha, r)$ 对 $w, b, \xi$ 的偏导数为零, 有

$$\begin{aligned}w^* &= \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\ \sum_{i=1}^m \alpha_i^* y^{(i)} &= 0 \\ C &= \alpha_i + r_i\end{aligned}$$

# The Dual Soft SVM

$$(w_{\text{soft}}^*, b_{\text{soft}}^*) = \arg \min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

思考：是否可以用梯度下降法求解？

Primal problem

s.t.  $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, m$

$\xi_i \geq 0, \quad i = 1, \dots, m.$

Dual problem

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

s.t.  $0 \leq \alpha_i \leq C, \quad i = 1, \dots, m$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0,$$

# The Dual Soft SVM

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

The Dual Hard SVM

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0, \end{aligned}$$

The Dual Soft SVM

思考: 惩罚因子  $C$  趋近无穷大会发生什么情况?

# The Dual Soft SVM

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0, \end{aligned}$$

KKT conditions:

$$\begin{cases} \alpha_i \geq 0, \quad r_i \geq 0 \\ y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i \geq 0 \\ \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i] = 0 \\ \xi_i \geq 0, \quad r_i \xi_i = 0 \end{cases}$$

$$\begin{aligned} \alpha_i = 0 & \Rightarrow y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i \geq 0 \\ \alpha_i > 0 & \Rightarrow y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i = 0 \\ 0 < \alpha_i < C & \Rightarrow r_i > 0, \xi_i = 0 \\ \alpha_i = C & \Rightarrow r_i = 0 \end{aligned}$$

$$\begin{aligned} w^* &= \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\ \sum_{i=1}^m \alpha_i^* y^{(i)} &= 0 \\ C &= \alpha_i + r_i \end{aligned}$$

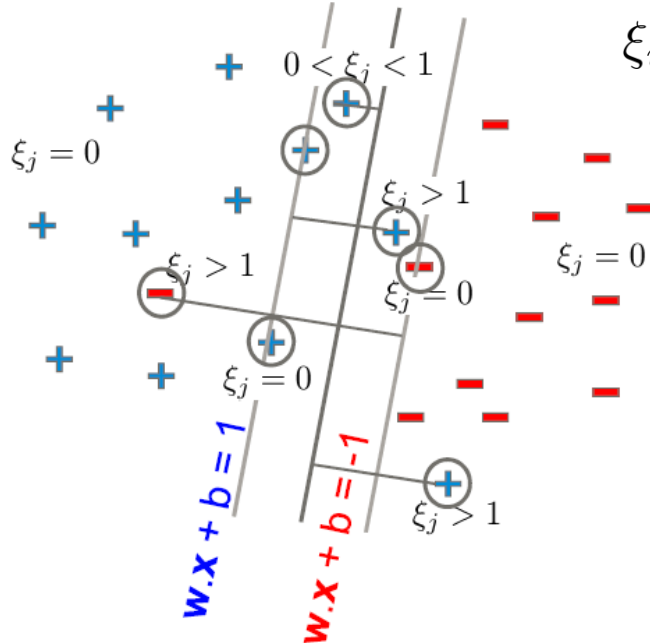
# Support vectors in Soft SVM

$$(w_{\text{soft}}^*, b_{\text{soft}}^*) = \arg \min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

s.t.

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, m$$

$$\xi_i \geq 0, \quad i = 1, \dots, m.$$



- Margin support vectors:  
 $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$
- Nonmargin support vectors:  
 $\xi_i > 0$

# The Dual Soft SVM

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0, \end{aligned}$$

- 这仍然是一个二次规划问题，可采用通用的QP算法求解，但其规模正比于训练样本数
- Sequential Minimal Optimization (SMO) algorithm是一种高效算法，基本思想是采用坐标下降法，在更新 $\alpha_i$ 时固定其他 $\alpha_j$ ，由于 $\sum_{i=1}^m \alpha_i y^{(i)} = 0$ ， $\alpha_i$ 的值可由其他的 $\alpha_j$  ( $j \neq i$ )表示

# Solving the Dual: The SMO Algorithm

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0, \end{aligned}$$

具体而言，SMO算法不断执行下面两个基本步骤直到收敛：

- 选取一对需要更新的变量(假设为 $\alpha_1$ 和 $\alpha_2$ ),满足 $\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^m \alpha_i y^{(i)}$ ;
- 固定 $\alpha_3, \dots, \alpha_m$ , 则 $-\sum_{i=3}^m \alpha_i y^{(i)} = \zeta$ 为一常量, 求 $W(\alpha)$ 的最优解等价于求一个带约束的二次函数优化问题

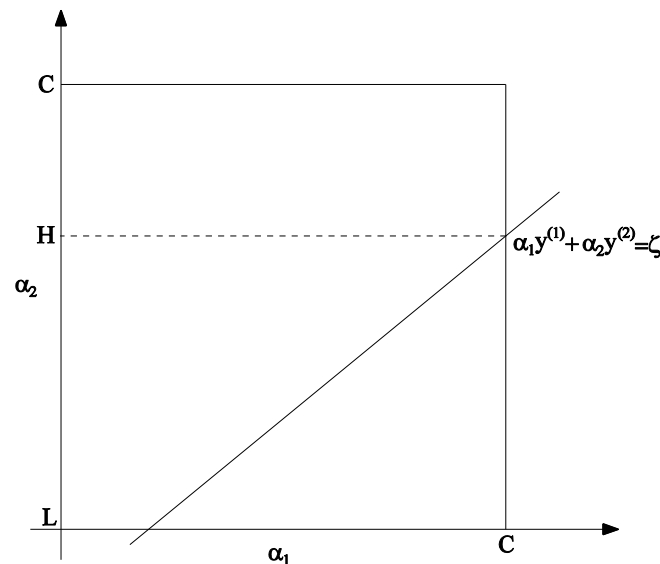
$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = \zeta \Rightarrow \alpha_1 = (\zeta - \alpha_2 y^{(2)}) y^{(1)}$$

代入到目标函数中:

$$W(\alpha_1, \alpha_2, \dots, \alpha_m) = W((\zeta - \alpha_2 y^{(2)}) y^{(1)}, \alpha_2, \dots, \alpha_m).$$

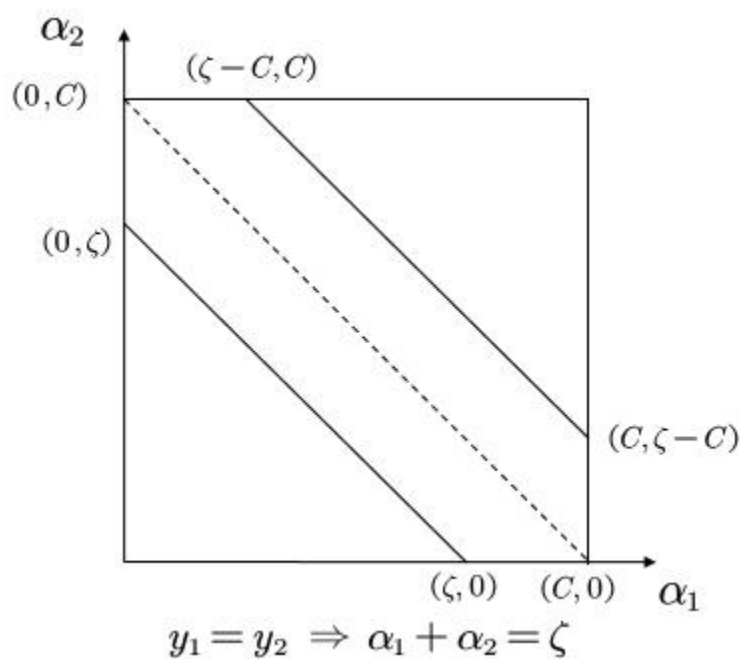
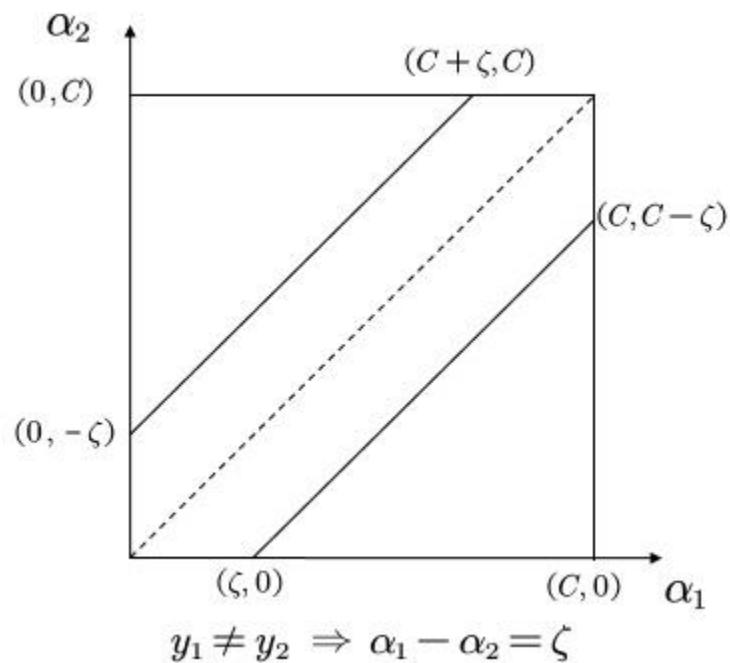
这是个关于 $\alpha_2$ 的二次函数:  $a\alpha_2^2 + b\alpha_2 + c$

约束条件为  $L \leq \alpha_2 \leq H$ , 这里 $L = 0$ .



# Solving the Dual: The SMO Algorithm

$$\alpha_1 y_1 + \alpha_2 y_2 = \zeta \Rightarrow \alpha_1 = (\zeta - \alpha_2 y_2) y_1$$





## The SMO Algorithm Implementation (optional)

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \end{aligned}$$

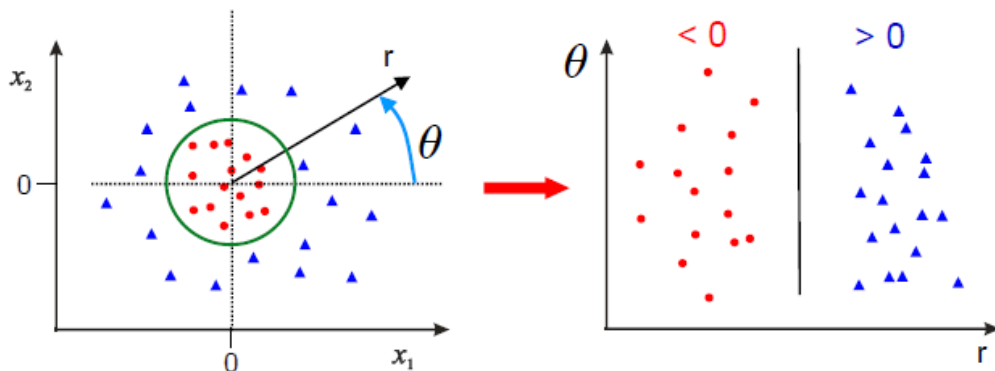
每次如何选择需要更新的 $\alpha_i$ 和 $\alpha_j$ ?

选取最不符合要求的两个参数:

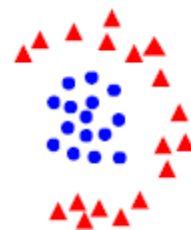
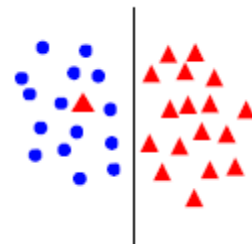
- 选出违反 KKT 条件最严重的样本点、以其对应的参数作为第一个参数
- 选出与第一个参数对应的样本点间隔最大的样本对应的参数

# 核 (Kernel)

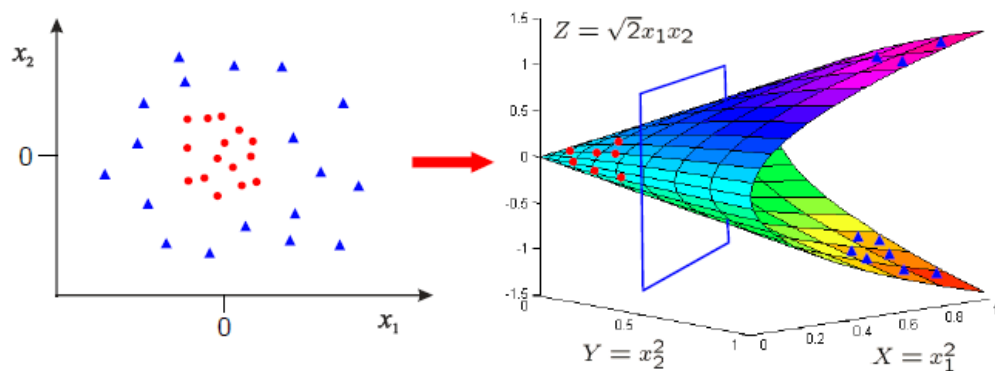
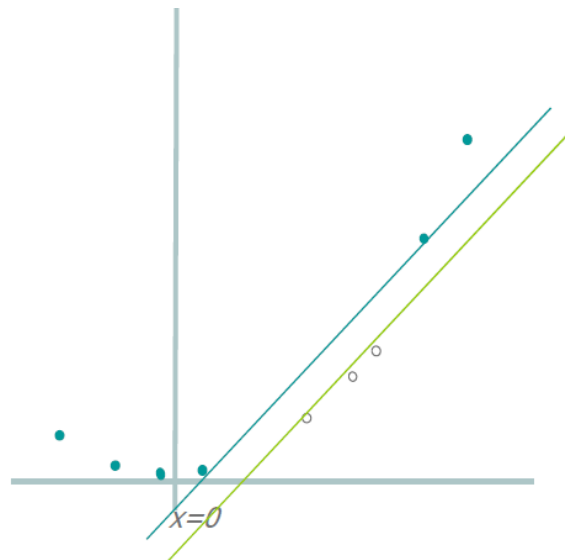
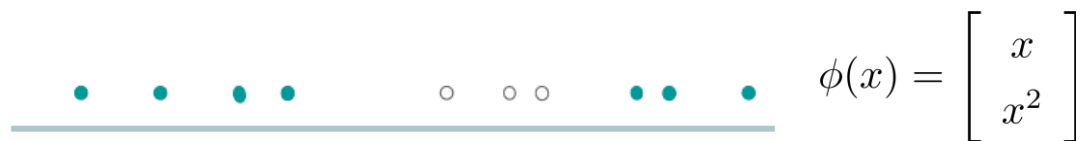
- 通过引入松弛因子，Soft SVM能处理部分“特异点” outlier导致的线性不可分问题
- 但如果数据本身是线性不可分的，如何处理？
  - 显式将数据变换到新的空间(如采用极坐标、多项式升维)，使其线性可分，如



$$\phi : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} r \\ \theta \end{bmatrix}$$



# 核 (Kernel)



$$\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$$

# 核 (Kernel)

映射为:  $x \rightarrow \phi(x)$

对应的SVM分类准则为:  $y = \text{sign} \left( \sum_{i=1}^m \alpha_i y^{(i)} \langle \phi(x^{(i)}), \phi(x) \rangle + b \right)$

- 只有少数的 $\alpha > 0$
- 只需要知道测试样本 $x$ 与支持所有支持向量的内积, 无需明确知道对应的映射

可以直接定义核函数来计算内积:  $k(x, z) = \langle \phi(x), \phi(z) \rangle = \phi(x)^T \phi(z)$

SVM分类准则:  $y = \text{sign} \left( \sum_{i=1}^m \alpha_i y^{(i)} k(x^{(i)}, x) + b \right)$

# 核 (Kernel)

如:  $k(x, z) = (x^T z)^2$ , 其中  $x, z \in \mathbb{R}^n$

$$\begin{aligned} k(x, z) &= \left( \sum_{i=1}^n x_i z_i \right) \left( \sum_{j=1}^n x_j z_j \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i x_j z_i z_j \\ &= \sum_{i,j=1}^n (x_i x_j) (z_i z_j) \end{aligned}$$

若  $n = 3$ ,  $\phi(x) =$

$$O(n^2) \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}$$

# 核 (Kernel)

$$\begin{aligned}k(x, z) &= (x^T z + c)^2 \\&= \sum_{i,j=1}^n (x_i x_j)(z_i z_j) + \sum_{i=1}^n (\sqrt{2c} x_i)(\sqrt{2c} z_i) + c^2\end{aligned}$$

如  $n = 3$ , 对应的映射为  $\phi(x) =$

多项式核:  $k(x, z) = (x^T z + c)^d$ , 对应  $\phi(x)$  的维度  $\approx n^d$

高斯核(RBF Kernel):  $k(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$ ,  
对应  $\phi(x)$  的维度  $\infty$

$$\begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ \sqrt{2c} x_3 \\ c \end{bmatrix}$$

# RBF Kernel

高斯核(RBF Kernel):  $k(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$ , 对应 $\phi(x)$ 的维度 $\infty$ , 因为(这里省略 $\sigma$ )

$$\begin{aligned}k(x, z) &= \exp\left(-\frac{1}{2}\|x - z\|^2\right) = \phi(x)^T \phi(z)? \\&= \exp\left(-\frac{1}{2}(\|x\|^2 + \|z\|^2 - 2x^T z)\right) \\&= \exp\left(-\frac{1}{2}\|x\|^2\right) \exp\left(-\frac{1}{2}\|z\|^2\right) \exp(x^T z) \\&= C_x C_z \sum_{i=0}^{\infty} \frac{(x^T z)^i}{i!}\end{aligned}$$

# Kernel Matrix and Mercer Kernel

假定存在某个核函数 $K(\cdot, \cdot)$ 对应某个隐式映射 $\phi(\cdot)$ , 给定训练集 $\{x^{(1)}, \dots, x^{(m)}\}$ , 对应的核矩阵 $K \in \mathbb{R}^{m \times m}$ 为对称矩阵, 其元素 $K_{ij} = K(x^{(i)}, x^{(j)})$

容易证明该矩阵是半正定的:

$$\begin{aligned} z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\ &= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\ &= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\ &= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\ &= \sum_k \left( \sum_i z_i \phi_k(x^{(i)}) \right)^2 \\ &\geq 0. \end{aligned}$$

**Mercer定理.** 函数 $K(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ 为核函数的充要条件为: 给定训练集 $\{x^{(1)}, \dots, x^{(m)}\}$ , 对应的核矩阵 $K \in \mathbb{R}^{m \times m}$ 为对称半正定矩阵.



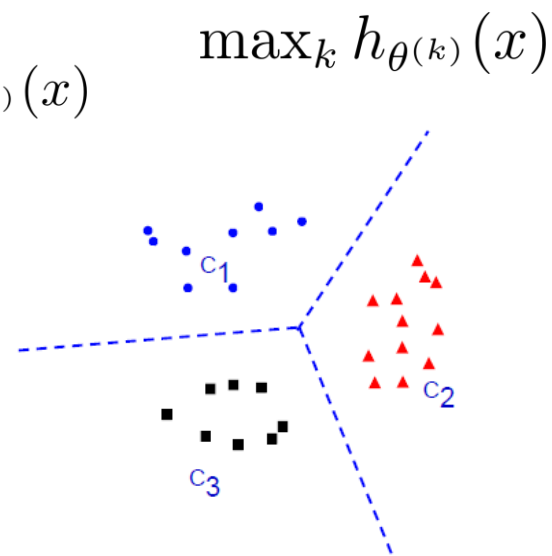
# Multi-Class SVM

- 将二分类器拓展处理多分类问题的基本思路:

- 训练: 采用one vs. rest策略训练 $K$ 个分类器 $h_{\theta^{(k)}}(x)$
- 测试: 选择分类器输出最大的值

- 是否可以采用该策略拓展SVM?

$$y = \arg \max_k \left( w^{(k)T} x + b^{(k)} \right)$$

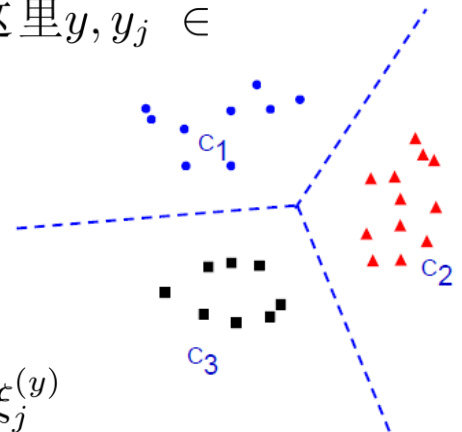


But  $(w^{(k)}, b^{(k)})$  may not be based on the same scale

# Multi-Class SVM

- 采用类似Softmax的思想, 同时学习 $K$ 个参数 $(w^{(k)}, b^{(k)})$ , 满足:  $\max_k h_{\theta^{(k)}}(x)$   
 $w^{(y_j)^T} x_j + b^{(y_j)} \geq w^{(y)^T} x_j + b^{(y)} + 1, \forall y \neq y_j, \forall j.$  这里 $y, y_j \in \{1, 2, \dots, K\}$

- Margin: gap between correct class and nearest other class



$$(w_{\text{multiclass}}^*, b_{\text{multiclass}}^*) = \arg \min_{w, b, \xi} \frac{1}{2} \sum_{k=1}^K \|w^{(k)}\|^2 + C \sum_{j=1}^m \sum_{y \neq y_j} \xi_j^{(y)}$$

s.t.

$$w^{(y_j)^T} x_j + b^{(y_j)} \geq w^{(y)^T} x_j + b^{(y)} + 1, \forall y \neq y_j, \forall j$$
$$\xi_j^{(y)} \geq 0, \forall y \neq y_j, \forall j.$$

采用了joint optimization, 保证各类的参数矢量  $(w^{(k)}, b^{(k)})$  have the same scale

# SVM vs. Logistic Regression

	SVM	Logistic Regression
Loss function	Hinge Loss	Logistic Loss (Cross-Entropy Loss)
High dimensional features with kernels	Yes!	No (but there is kernel logistic regression too)
Solution sparse	Often yes!	Almost always no!
Semantics of output	“Margin”	“Real probabilities”

# 支持向量回归

- Soft SVM Classifier:

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

- 可以写成更一般的形式

$$\min_{w,b} C \sum_{i=1}^m \ell \left( y^{(i)}, h_{w,b}(x^{(i)}) \right) + \frac{1}{2} \|w\|^2$$

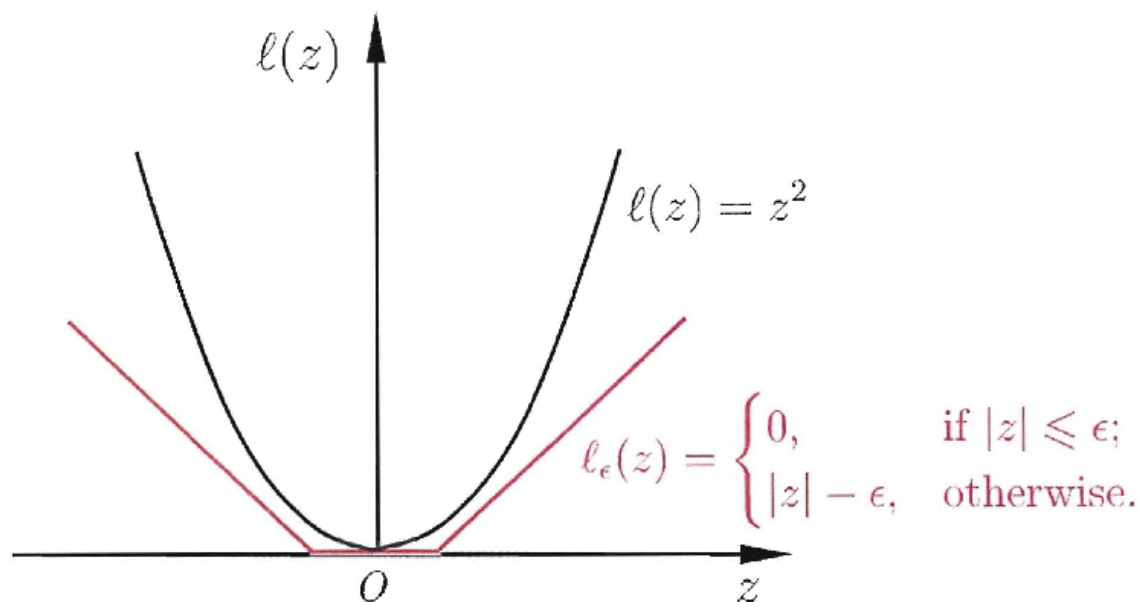
- 若 $\ell(\cdot, \cdot)$ 为平方损失, 则演变成Ridge regression, 当且仅当 $h(x) = y$ 时, 损失才为零. SVR的基本思想是可以容忍一定的错误, 即当预测的值 $h(x)$ 和实际的值差别不大于 $\epsilon$ 时, 损失仍为零, 对应的损失函数为

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

# 支持向量回归

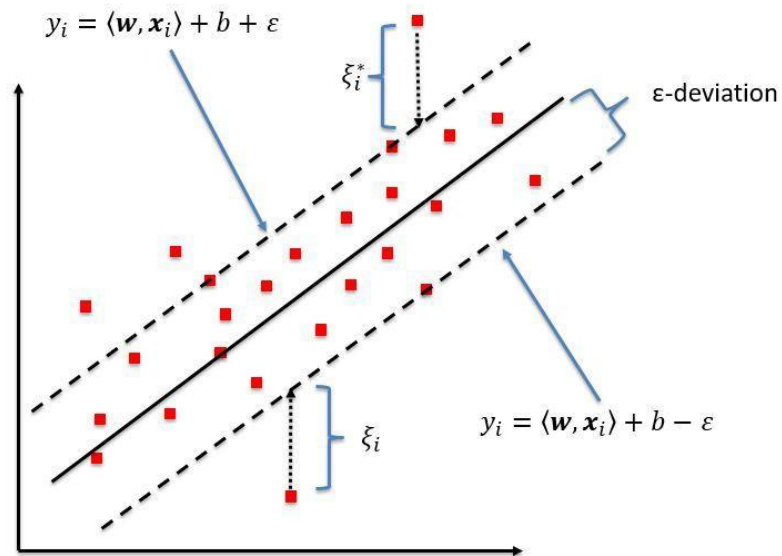
$\epsilon$ -不敏感损失函数  $\ell_\epsilon(z)$

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



# 支持向量回归： 线性回归

- 引入松弛因子 $\xi_i, \xi_i^*$

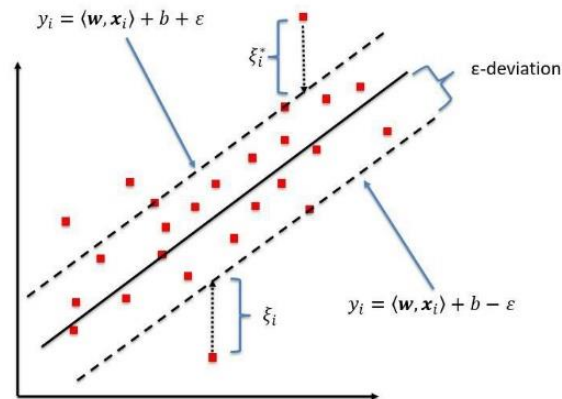


$$\begin{aligned} \min_{w, b, \xi_i, \xi_i^*} \quad & C \sum_{i=1}^m (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & h_{w,b}(x^{(i)}) - y^{(i)} \leq \epsilon + \xi_i, \\ & y^{(i)} - h_{w,b}(x^{(i)}) \leq \epsilon + \xi_i^*, \\ & \xi_i \geq 0, \xi_i^* \geq 0, i = 1, 2, \dots, m \end{aligned}$$

这里 $h_{w,b}(x) = w^T x + b = \langle w, x \rangle + b$

# 支持向量回归： 线性回归

$$\begin{aligned}
 \min_{w, b, \xi_i, \xi_i^*} \quad & C \sum_{i=1}^m (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2 \\
 \text{s.t.} \quad & h_{w,b}(x^{(i)}) - y^{(i)} \leq \epsilon + \xi_i, \\
 & y^{(i)} - h_{w,b}(x^{(i)}) \leq \epsilon + \xi_i^*, \\
 & \xi_i \geq 0, \xi_i^* \geq 0, i = 1, 2, \dots, m
 \end{aligned}$$



- 再次引入拉格朗日乘子  $r_i \geq 0, r_i^* \geq 0, \alpha_i \geq 0, \alpha_i^* \geq 0$ , 对应的拉格朗日函数为

$$\begin{aligned}
 L(w, b, \alpha, \alpha^*, \xi, \xi^*, r_i, r_i^*) = & C \sum_{i=1}^m (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2 - \sum_{i=1}^m r_i \xi_i - \sum_{i=1}^m r_i^* \xi_i^* \\
 & + \sum_{i=1}^m \alpha_i (h_{w,b}(x^{(i)}) - y^{(i)} - \epsilon - \xi_i) + \sum_{i=1}^m \alpha_i^* (y^{(i)} - h_{w,b}(x^{(i)}) - \epsilon - \xi_i^*)
 \end{aligned}$$

- 令拉格朗日函数  $L(w, b, \alpha, \alpha^*, \xi, \xi^*, r_i, r_i^*)$  对  $w, b, \xi, \xi^*$  的偏导数为零有

$$w = \sum_{i=1}^m (\alpha_i^* - \alpha_i) x^{(i)}, \quad 0 = \sum_{i=1}^m (\alpha_i^* - \alpha_i), \quad C = \alpha_i + r_i, \quad C = \alpha_i^* + r_i^*$$

# 支持向量回归： 线性回归

- 将

$$w = \sum_{i=1}^m (\alpha_i^* - \alpha_i) x^{(i)}, \quad 0 = \sum_{i=1}^m (\alpha_i^* - \alpha_i), \quad C = \alpha_i + r_i, \quad C = \alpha_i^* + r_i^*$$

代入到拉格朗日函数中化简得到只关于 $\alpha_i, \alpha_i^*$ 的函数，最大化该函数可得SVR的对偶问题

$$\begin{aligned} \max_{\alpha, \alpha^*} \quad & \sum_{i=1}^m y^{(i)} (\alpha_i^* - \alpha_i) - \epsilon (\alpha_i^* + \alpha_i) - \frac{1}{2} \sum_{i,j=1}^m (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, m \\ & \sum_{i=1}^m (\alpha_i^* - \alpha_i) = 0. \end{aligned}$$

仍然属于典型的二次规划问题.



# 支持向量回归：线性回归

- KKT条件为

$$\begin{cases} \alpha_i(h_{w,b}(x^{(i)}) - y^{(i)} - \epsilon - \xi_i) = 0, \\ \alpha_i^*(y^{(i)} - h_{w,b}(x^{(i)}) - \epsilon - \xi_i^*) = 0, \\ (C - \alpha_i)\xi_i = 0, (C - \alpha_i^*)\xi_i^* = 0 \end{cases}$$

- 将 $w = \sum_{i=1}^m (\alpha_i^* - \alpha_i)x^{(i)}$ 代入 $h_{w,b}(x) = w^T x + b$  有

$$h_{w,b}(x) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle x^{(i)}, x \rangle + b$$

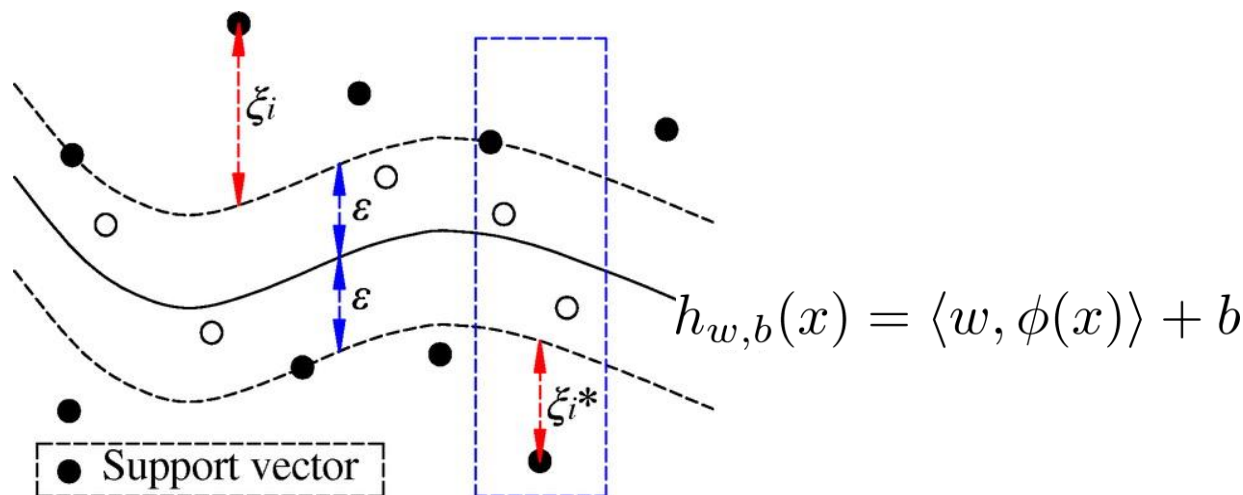
对应 $(\alpha_i^* - \alpha_i) \neq 0$ 对应的样本为SVR的支持向量，必位于 $\epsilon$ -间隔带之外。

- 如何求 $b$ ? 根据KKT条件，对任意训练样本 $(x^{(i)}, y^{(i)})$ ，有 $\alpha_i(h_{w,b}(x^{(i)}) - y^{(i)} - \epsilon - \xi_i) = 0$ 和 $(C - \alpha_i)\xi_i = 0$ ，若满足 $0 < \alpha_i < C$ ，则必有 $\xi_i = 0$ ，从而有

$$b = y^{(i)} + \epsilon - \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle x^{(i)}, x \rangle$$

理论上可以去任意满足 $0 < \alpha_i < C$ 的样本进行计算，实际应用中取多个(或所有)满足条件的样本计算后取平均值。

# 支持向量回归：非线性回归



$$h_{w,b}(x) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle x^{(i)}, x \rangle + b$$

变为

$$h_{w,b}(x) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle \phi(x^{(i)}), \phi(x) \rangle + b = \sum_{i=1}^m (\alpha_i^* - \alpha_i) k(x^{(i)}, x) + b$$

# Thanks!

Any questions?