

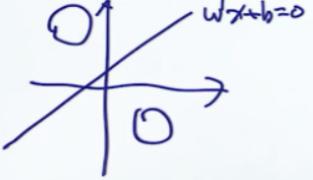
线性划分回顾，将乘到 $wx+b$ 左边将1、2等式归一

Linear Classifier

$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad y_i \in \{-1, 1\} \quad i=1, 2, \dots, n$

$\Theta = \{w, b\} \quad \underline{w^T x + b = 0} \leftarrow \text{Decision Boundary}$

① $w^T x_i + b \geq 0$ 时 $y_i = 1$ ② $w^T x_i + b < 0$ 时 $y_i = -1$ $\Rightarrow (w^T x_i + b) \cdot y_i \geq 0$

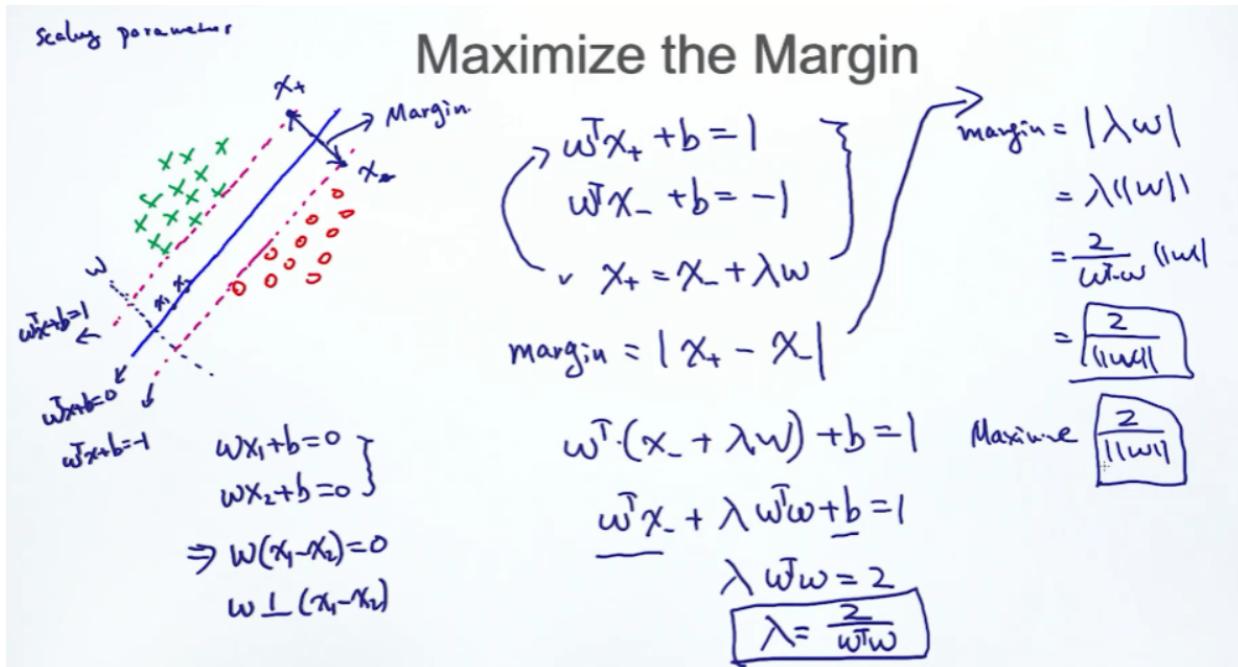


四 题

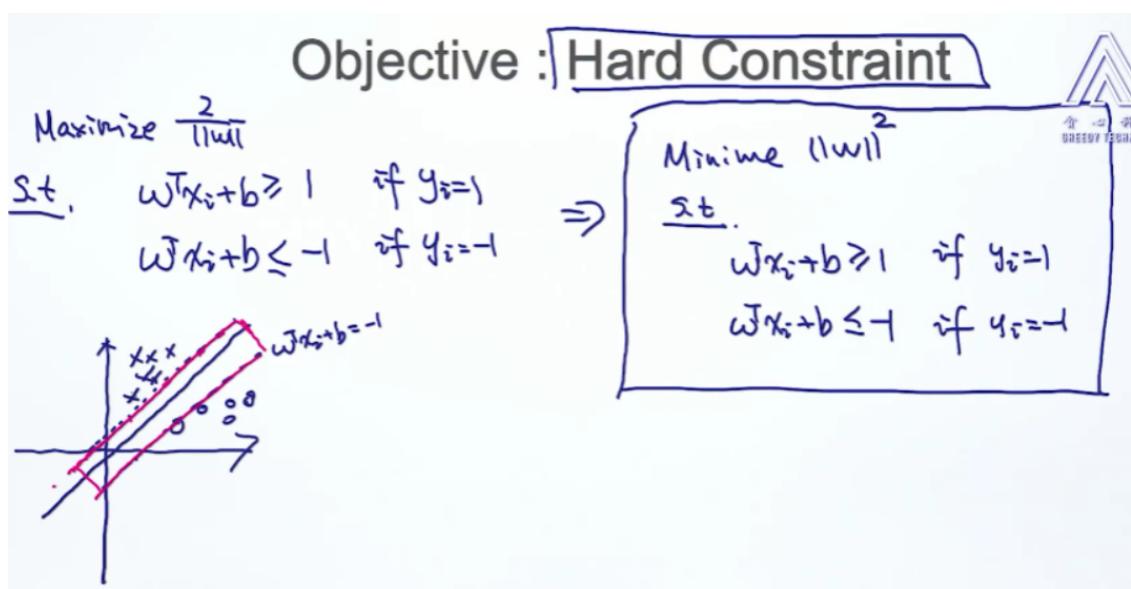
123三条线中2号线的性质最好，有最大的安全区域，最鲁棒



推导目标函数的过程，左边证明为什么 w 向量垂直于直线，中间推导两条直线公式，右边选出margin的最大值

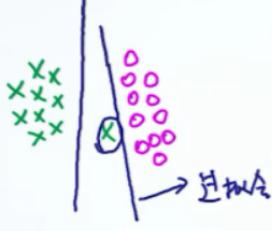


hard constraint



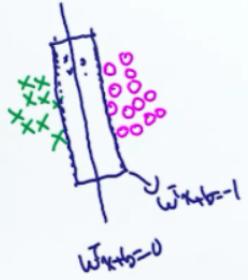
soft constraint. hard constraint会导致过拟合， soft constraint允许样本犯一定范围的错误

Objective : Soft Constraint



$$\begin{aligned} \min & \|w\|^2 \\ \text{s.t.} & w^T x_i + b \geq 1, \text{ if } y_i = 1 \\ & w^T x_i + b \leq -1 \text{ if } y_i = -1 \end{aligned}$$

Hard-constraint.



$$\begin{aligned} \min & \|w\|^2 \\ \text{s.t.} & (w^T x_i + b) y_i \geq 1 \end{aligned}$$

$\downarrow \lambda = \infty$

$$\begin{aligned} \min & \|w\|^2 + \lambda \sum_{i=1}^n \varepsilon_i \\ \text{s.t.} & (w^T x_i + b) y_i \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0 \end{aligned}$$

Degree of mistake
犯了多大的错误

1式使得c越小越好，根据2式推导处c大于a. 所以c是一个定值，等于 $1 - (wx + b)y$ 设其为a. 当a小于0时不起任何作用,c取0.当a大于0时c起作用，从而的好hinge loss。

Convert to Hinge Loss

$$\begin{aligned} \min & \|w\|^2 + \lambda \sum_{i=1}^n \varepsilon_i \\ \text{s.t.} & (w^T x_i + b) y_i \geq 1 - \varepsilon_i \end{aligned}$$

$\varepsilon_i \rightarrow \text{slack variable}$

$$\varepsilon_i \geq 1 - (w^T x_i + b) y_i$$

$$[\varepsilon_i = a]$$

$$\varepsilon_i = 1 - (w^T x_i + b) y_i$$

$$\varepsilon_i \geq 0 \Rightarrow 1 - (w^T x_i + b) y_i \geq 0$$

then $1 - (w^T x_i + b) y_i \leq 0$
 $(w^T x_i + b) y_i \geq 1 \Rightarrow$ 不适用 hinge cost.

$$\min \|w\|^2 + \lambda \sum_{i=1}^n \max(0, 1 - (w^T x_i + b) y_i)$$

$\max(0, x) :$
 3 if $x < 0 \Rightarrow 0$
 if $x > 0 \Rightarrow x$

Linear SVM

Hinge Loss

使用梯度下降法求w和b if判断a等式是不是大于0

Stochastic Gradient Descent for Hinge Loss Objective

$$\min \|\mathbf{w}\|^2 + \lambda \sum_{i=1}^n \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

Regularization
Loss

$$\max(0, x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$$

\mathbf{w}^*, b^* (参数)

$i=1 \dots n$

if $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0$

$$\mathbf{w}^* = \mathbf{w} - \eta_t \cdot 2\mathbf{w}$$

else $\mathbf{w}^* = \mathbf{w} - \eta_t \left(2\mathbf{w} + \lambda \frac{\partial(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))}{\partial \mathbf{w}} \right)$

$$b^* = b - \eta_t \left(\lambda \frac{\partial(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))}{\partial b} \right)$$

SGD
迭代

对偶

Linear SVM : Primal Form

Linear SVM

primal form

Dual Form

kernel trick

$$\min_x f(x)$$

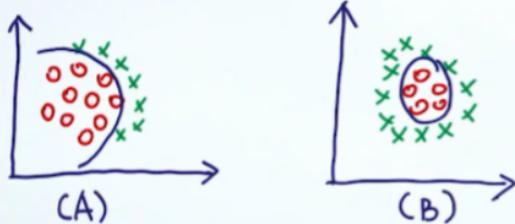
$$\text{s.t. } g_i(x) \leq 0$$

$$h_j(x) = 0$$

x^*

(sub-optimal)

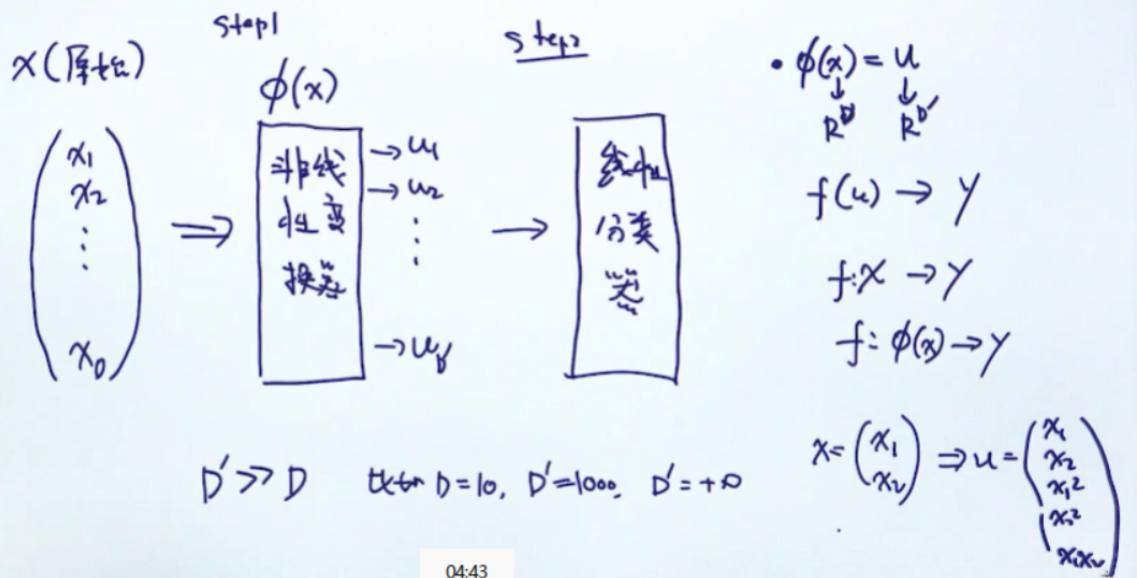
Disadvantages of Linear SVM



- ① 利用非线性核
- Neural Network
- ② 把数据映射到高维空间
↓
学习一个线性模型



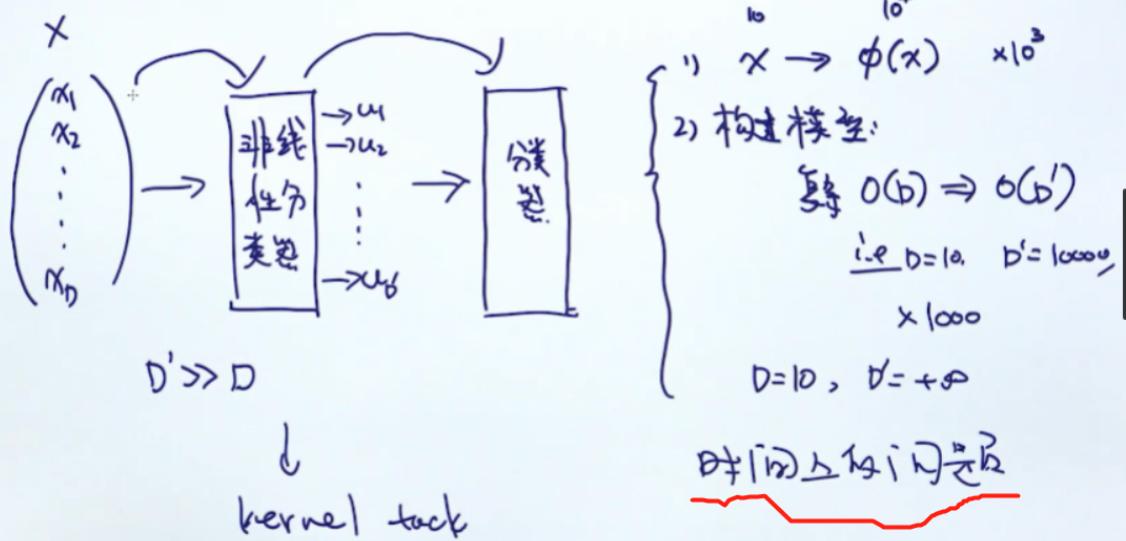
Mapping Feature to High Dimensional Space



04:43

映射到高维空间会有时间复杂度带来的问题

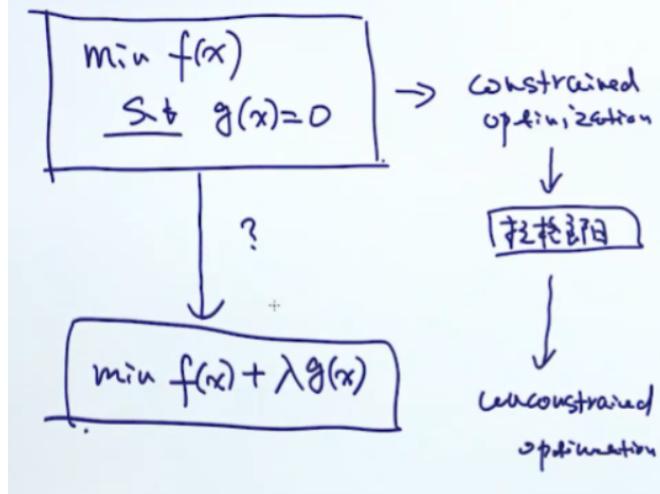
The Disadvantages of Mapping to High Dimension



拉格朗日乘子

×625 (51%)

Lagrangian: Equality Constraint



例：
 $\begin{aligned} &\min x_1^2 + x_2^2 \\ &\text{s.t. } x_2 - x_1 = -1 \end{aligned}$

↓

$\min x_1^2 + x_2^2 + \lambda (x_2 - x_1 + 1)$

$L(x, \lambda)$

$\begin{pmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \frac{\partial L}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} 2x_1 - \lambda \\ 2x_2 - \lambda \\ x_2 - x_1 + 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

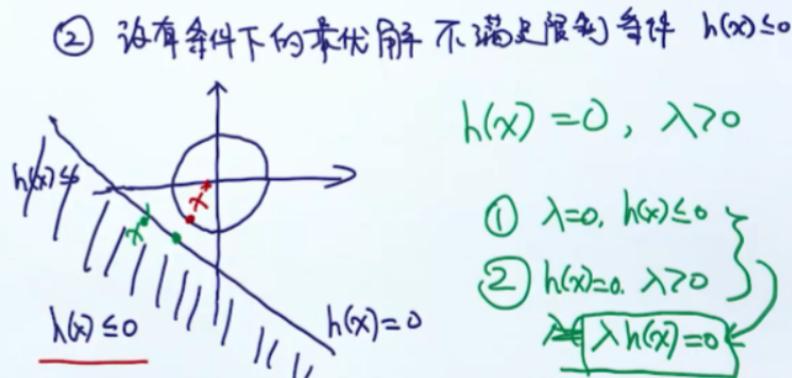
$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}$

当 $x(x)$ 小于 0 时 拉格朗日函数形式

Lagrangian: Intuition of Inequality Constraint

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & h(x) \leq 0 \end{array}$$

$$\begin{array}{ll} \min & f(x) + \lambda h(x) \\ \text{s.t.} & \lambda h(x) = 0 \\ & h(x) \leq 0 \end{array}$$



kkt条件

KKT Conditions

$$\begin{array}{l} \min f(x) \\ \text{s.t.} : g_i(x) = 0, i=1, 2, \dots, R \\ h_j(x) \leq 0, j=1, 2, \dots, R' \end{array} \Rightarrow \begin{array}{l} \min f(x) + \sum_{i=1}^R \lambda_i g_i(x) + \sum_{j=1}^{R'} \mu_j h_j(x) \\ \text{s.t.} \quad \lambda_i, \mu_j \geq 0, \forall i, j \\ \mu_j h_j(x) = 0, \forall j \\ h_j(x) \leq 0, \forall j \end{array}$$

KKT conditions

svm 的kkt条件

KKT Condition of SVM



SVM - Hard Constraint

$$\begin{aligned} \text{minimize}_{\omega} & \frac{1}{2} \|\omega\|_2^2 \\ \text{s.t.} & y_i(\omega^T x_i + b) - 1 \geq 0 \quad i=1, \dots, n \end{aligned}$$

对偶问题
↓

$$\begin{aligned} \text{minimize}_{\omega, b, \lambda} & \frac{1}{2} \|\omega\|_2^2 + \sum_{i=1}^n \lambda_i [1 - y_i(\omega^T x_i + b)] \\ \text{s.t.} & \begin{cases} \lambda_i \geq 0 \quad \forall i \\ \lambda_i [1 - y_i(\omega^T x_i + b)] = 0 \quad \forall i \\ 1 - y_i(\omega^T x_i + b) \leq 0 \end{cases} \end{aligned}$$

SVM FA
KKT conditions

primal problem
 $\theta = (\omega, b, \lambda)$

为什么svm用对偶

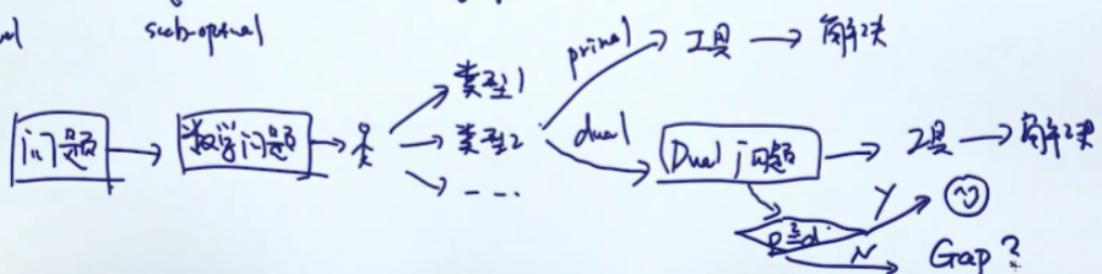
Primal-Dual Problem



- ① primal 问题有可能比较难解决
- ② 在 dual 问题上发现一些有趣的东西 ←

$$\begin{array}{ccc} \text{Primal} & \stackrel{?}{=} & \text{Dual} \\ \downarrow & & \downarrow \\ \text{optimal} & & \text{sub-optimal} \end{array}$$

Gap $\leq \langle \text{primal}, \text{dual} \rangle$
理想: Gap = 0



Dual Derivation of SVM

$$\boxed{\begin{aligned} \lambda &= \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^n \lambda_i [1 - y_i(w^T x_i + b)] \\ \text{s.t. } \lambda_i &\geq 0, \quad \forall i \\ \lambda_i [1 - y_i(w^T x_i + b)] &= 0, \quad \forall i \\ 1 - y_i(w^T x_i + b) &\leq 0, \quad \forall i \end{aligned}}$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w + \sum_{i=1}^n \lambda_i (-y_i x_i) \Rightarrow \boxed{w = \sum_{i=1}^n \lambda_i y_i x_i}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \boxed{\sum_{i=1}^n \lambda_i y_i = 0}$$

Dual

$$\begin{aligned} \frac{1}{2} \|w\|_2^2 &= \frac{1}{2} w^T w = \frac{1}{2} \left(\sum_{i=1}^n \lambda_i y_i x_i \right)^T \left(\sum_{j=1}^n \lambda_j y_j x_j \right) \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i y_i x_i^T \lambda_j y_j x_j^T \\ \sum_{i=1}^n \lambda_i [1 - y_i(w^T x_i + b)] &= \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \lambda_i y_i (\sum_{j=1}^n \lambda_j y_j x_j^T) \\ &= \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \lambda_i y_i (\sum_{j=1}^n \lambda_j y_j x_j^T) \\ &= \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j \\ \therefore L &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^n \lambda_i \end{aligned}$$

Recall: Mapping to High Dimension

$$\begin{array}{c} x \\ \left(\begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right) \end{array} \rightarrow \begin{array}{c} \phi(x) = u \\ \left[\begin{array}{c} \text{非线性} \\ \text{特征} \\ \text{映射} \end{array} \right] \end{array} \rightarrow \begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_D \end{array} \quad \boxed{\text{分类器}}$$

$$D' \gg D$$

$$O(D) \rightarrow O(D')$$

$$\begin{aligned} x &= (x_1, x_2), \quad \text{非线性} \quad z = (z_1, z_2) \\ x^T \cdot z &= (x_1 z_1 + x_2 z_2) \\ \phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1 x_2) \quad \phi(z) = (z_1^2, z_2^2, \sqrt{2}z_1 z_2) \\ \phi(x)^T \phi(z) &= (x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 x_2 z_1 z_2) \\ &= (x_1 z_1 + x_2 z_2)^2 = \boxed{(x^T z)^2} \end{aligned}$$

604 (48%)

Dual Derivation of SVM

Dual Formation of SVM

$$\text{minimize} \quad -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j + \frac{1}{2} \sum_{i=1}^n \lambda_i$$

$$\begin{aligned} \text{s.t. } \lambda_i &\geq 0, \quad \forall i \\ \sum_{i=1}^n \lambda_i y_i &= 0, \quad w = \sum_{i=1}^n \lambda_i y_i x_i \end{aligned}$$

$$\lambda_i [1 - y_i(w^T x_i + b)] = 0$$

$$\begin{aligned} \phi(\cdot) &=? \\ O(\phi(x)^T \phi(y)) &\approx O(x^T y) \end{aligned}$$

未知参数
\$\boxed{\lambda}\$

Kernel Trick

$$f(x) \cdots \boxed{x^T \Phi(\cdot)}$$

Kernel trick = 不直接对特征向量 SVM

$$(x_i, x_j) \rightarrow (x_i^T x_j, \gamma_i \gamma_j)$$

+
 Mercer's Theorem

Linear kernel :

$$k(x, y) = x^T y \Rightarrow \text{Linear SVM}$$

Polynomial kernel :

$$k(x, y) = (1 + x^T y)^d$$

Gaussian kernel :

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

$$\sum_{i=1}^n \xi_i = 1$$

① Linear SVM

- Hard constraint
- Soft - constraint (slack variable)

② 优化 (KKT 条件)

- 等号条件
- 不等式条件

③ Dual Formulation (kernel trick)

- primal $\stackrel{?}{=} \text{dual}$