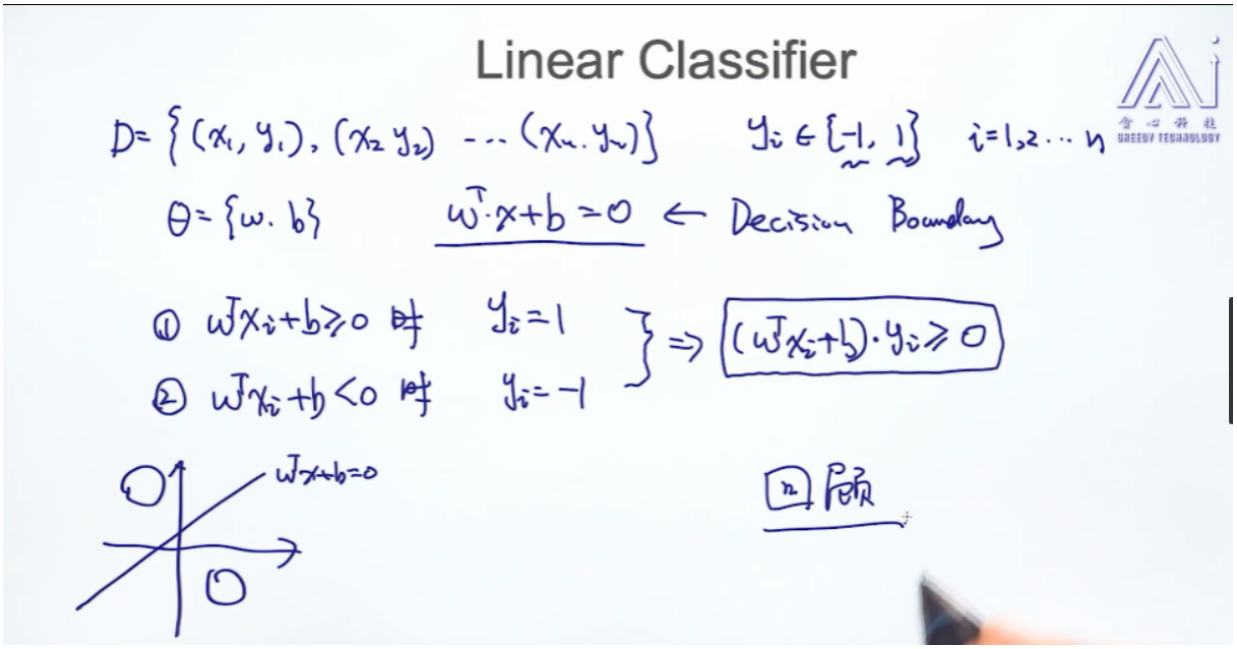
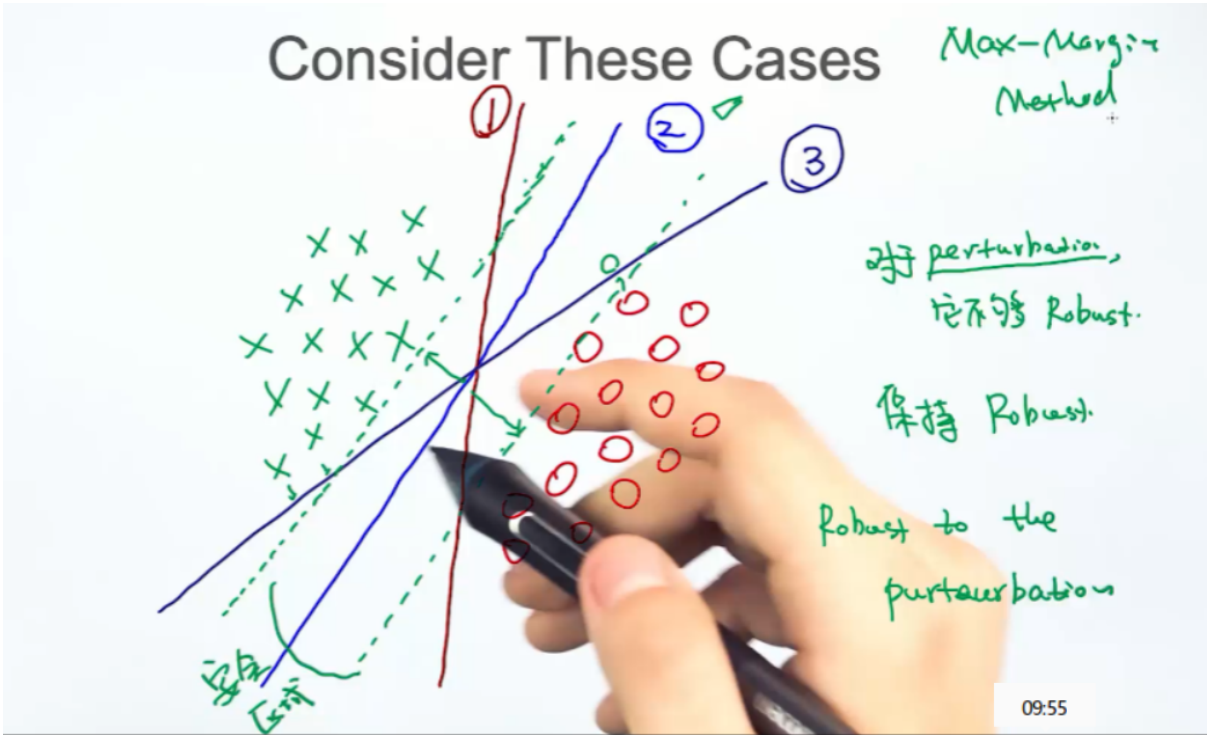
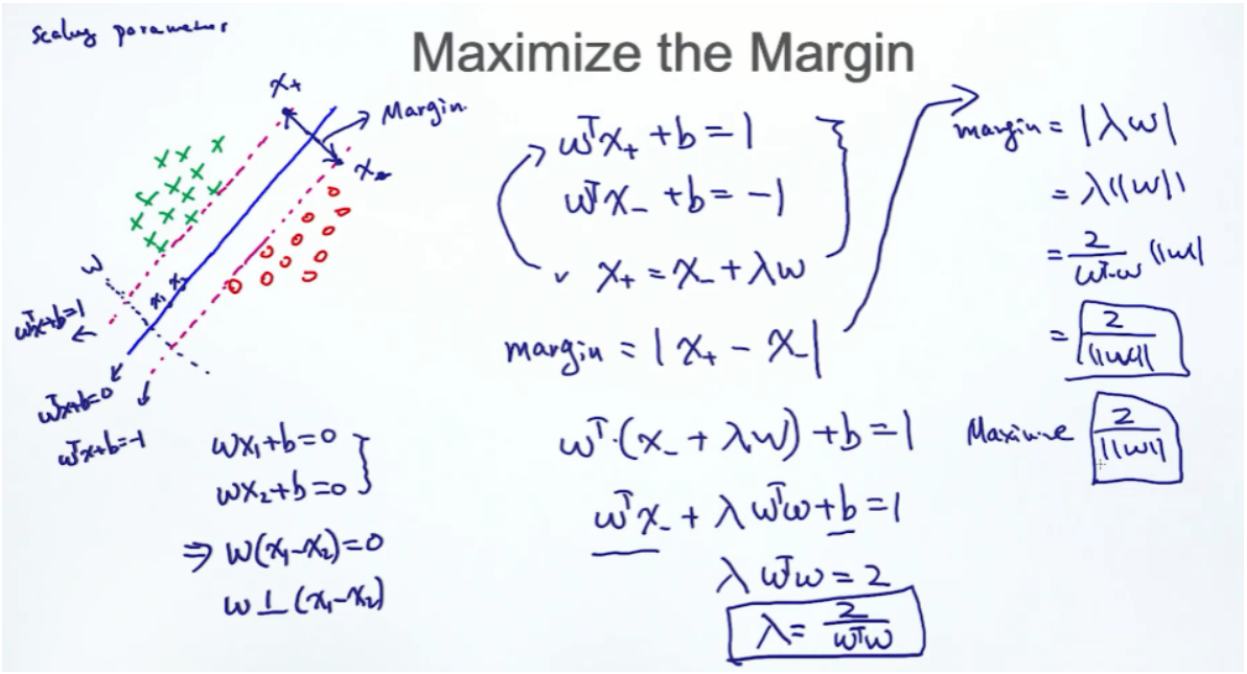
线性划分回顾，将乘到wx+b左边将1、2等式归一



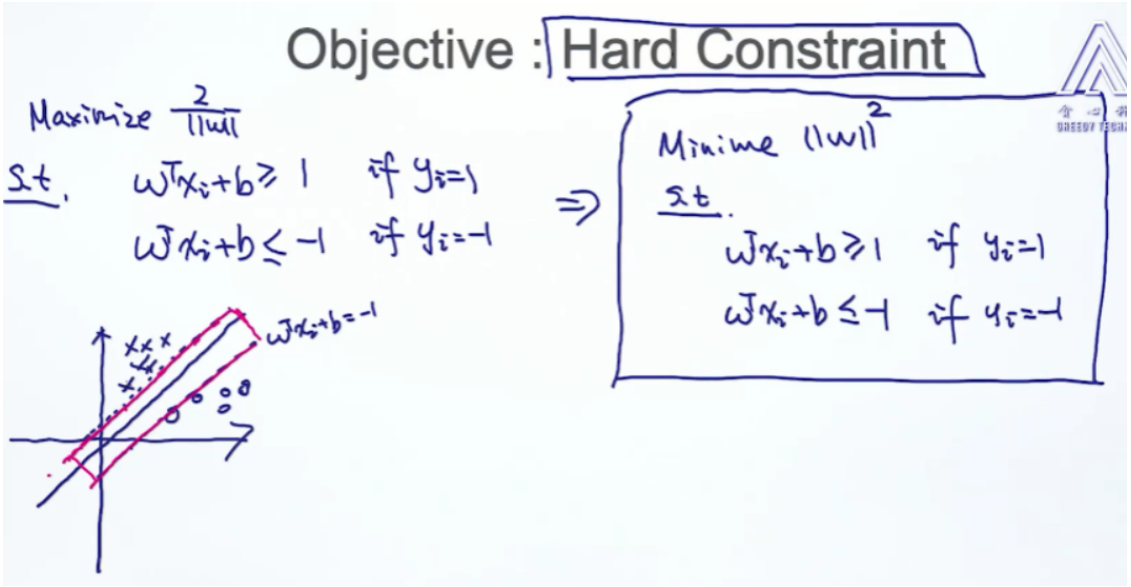
123三条线中2号线的性质最好，有最大的安全区域，最鲁棒



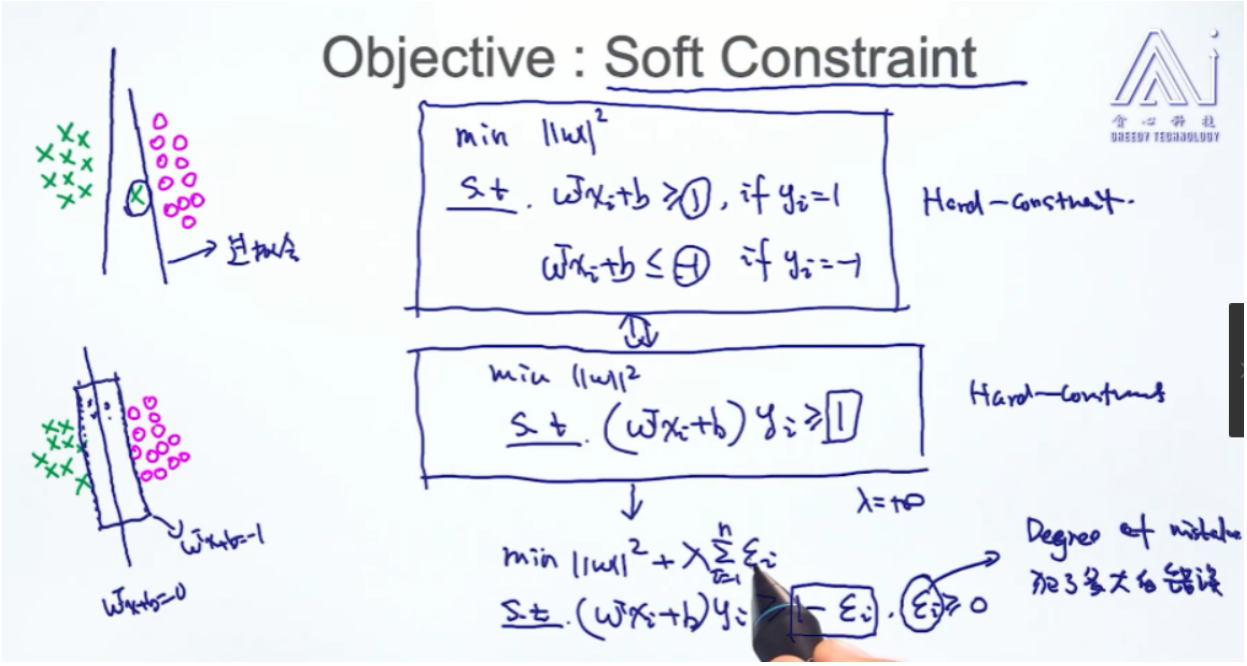
推导目标函数的过程，左边证明为什么w向量垂直于直线，中间推导两条直线公式，右边选出margin的最大值



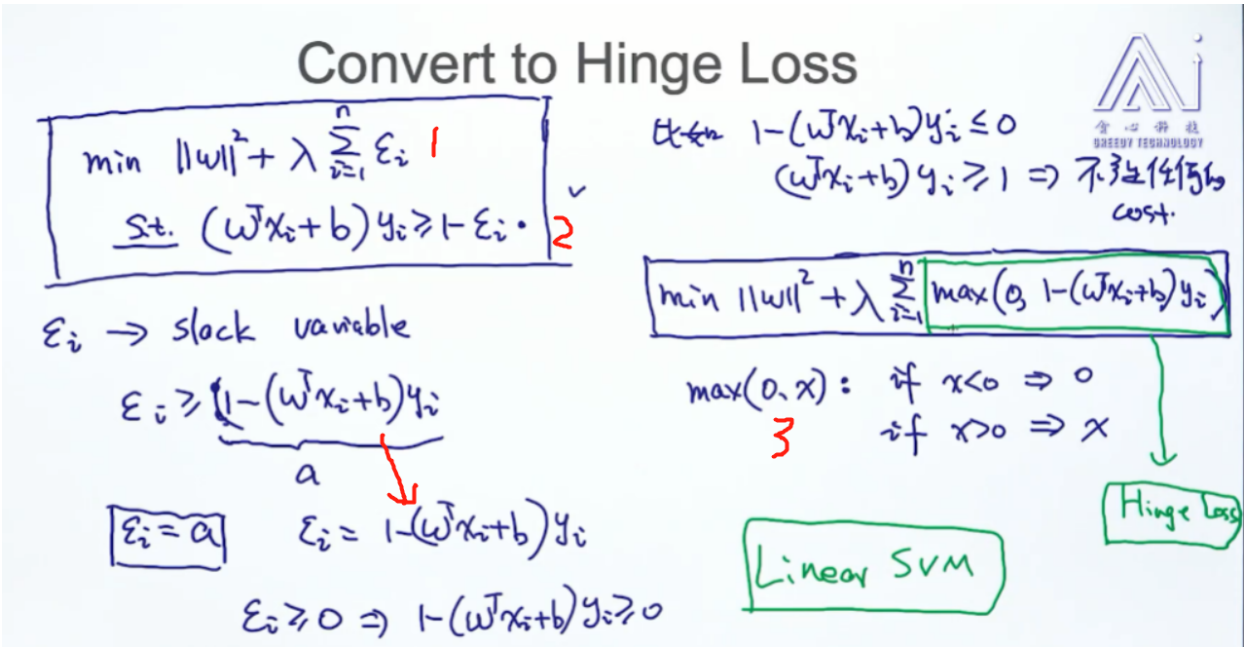
hard constraint



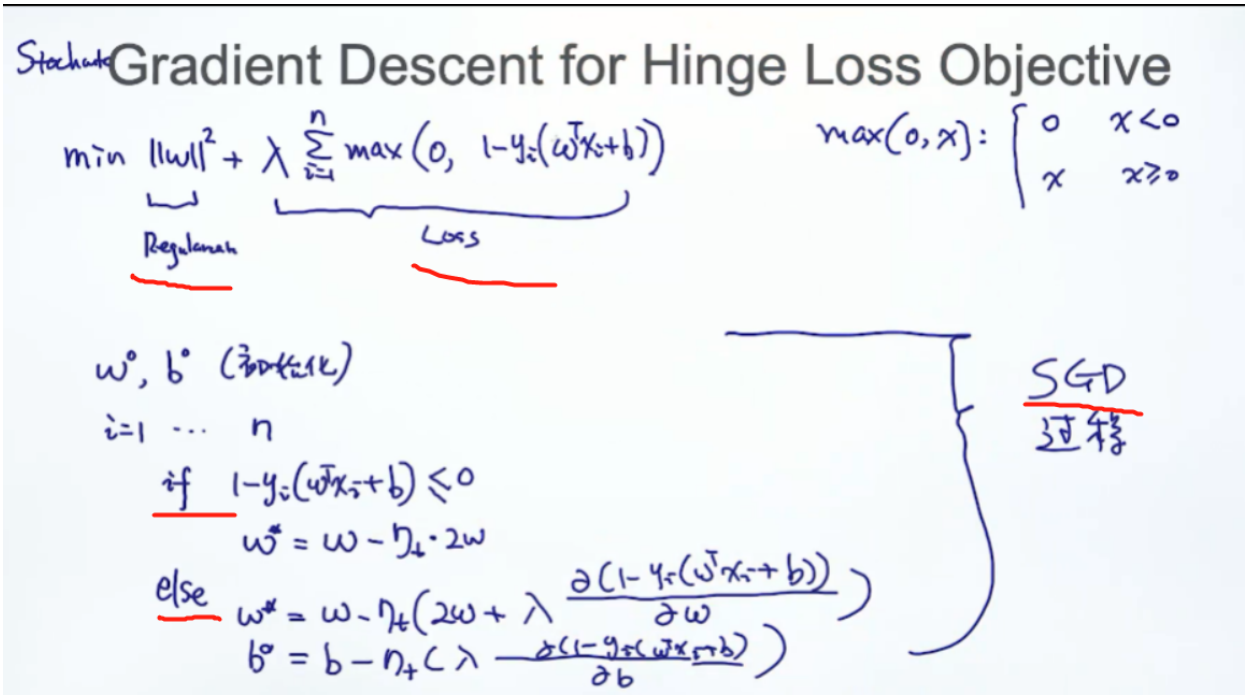
soft constraint. hart constraint会导致过拟合， soft constraint允许样本犯一定范围的错误



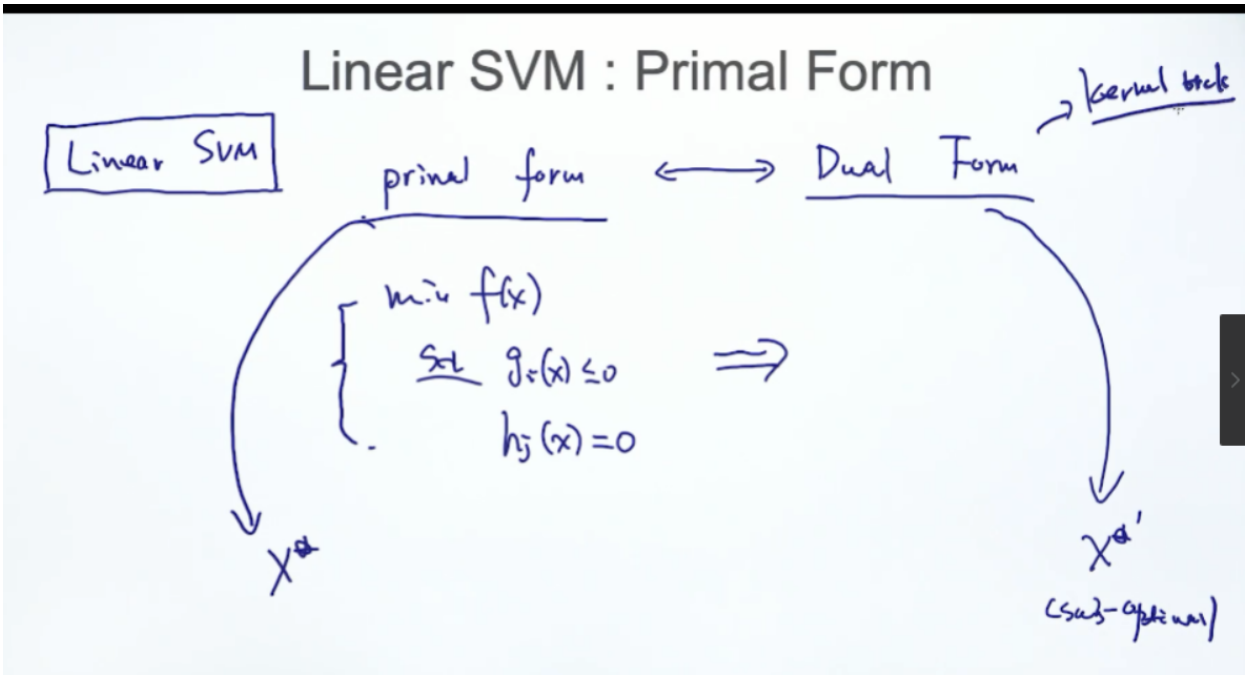
1式使得c越小越好，根据2式推导处c大于a. 所以c是一个定值，等于1-(wx+b)y 设其为a.当a小于0时不起任何作用,c取0.当a大于0时c起作用，从而的好hinge loss。

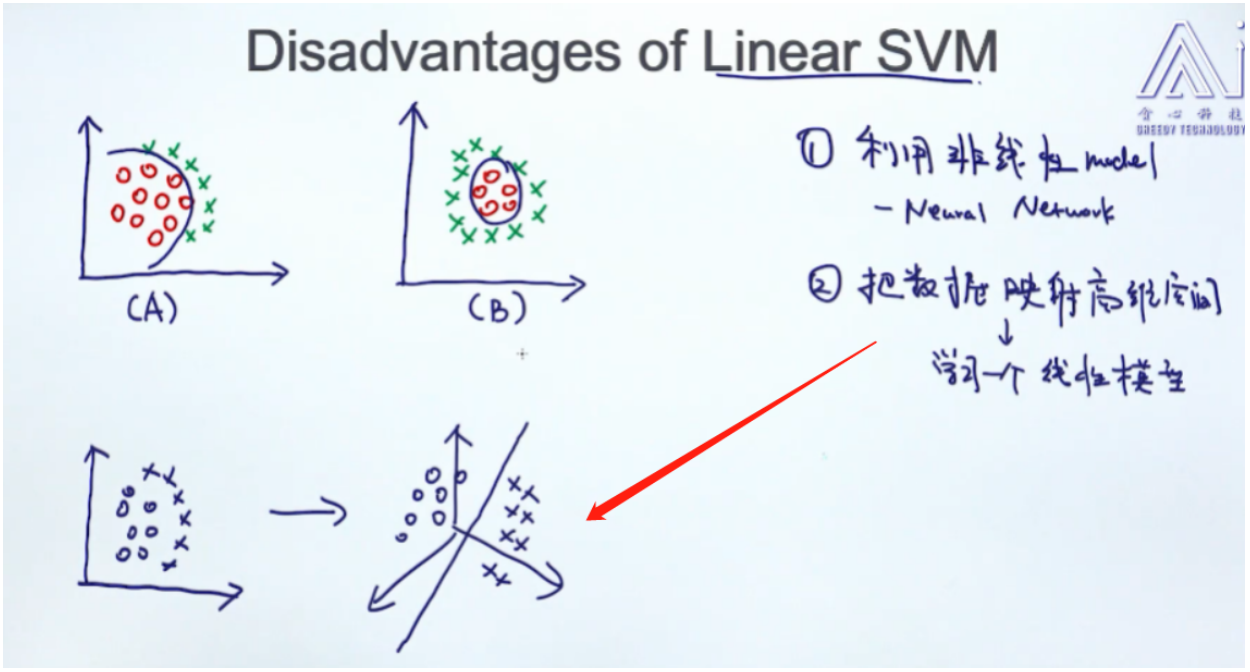


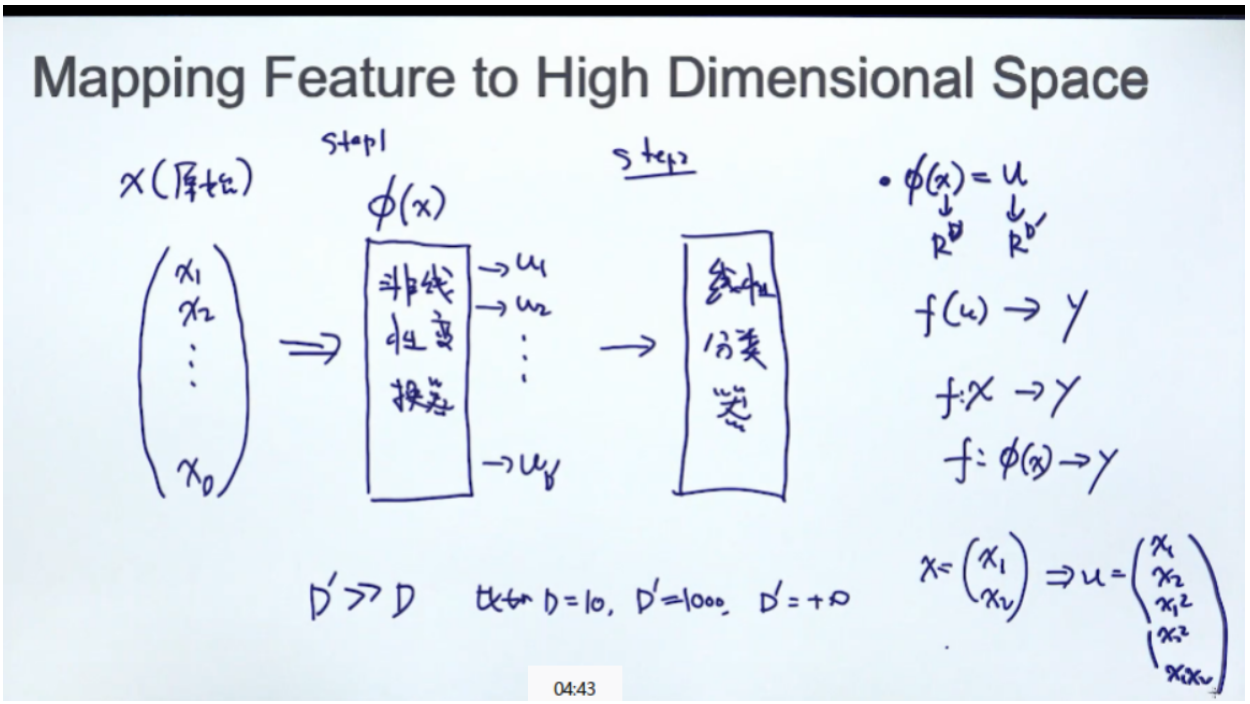
使用梯度下降法求w和b if判断a等式是不是大于0


Stochastic Gradient Descent for Hinge Loss Objective
min ||w||^2 + λ Σ_{i=1}^{n} max(0, 1 - y_i(w^T x_i + b))
Regularization
Loss
max(0, x): 0 x<0; x x≥0
w°, b° (random)
i=1 ... n
if 1 - y_i(w^T x_i + b) ≤ 0
w* = w - η_t · 2w
else
w* = w - η_t(2w + λ ∂(1 - y_i(w^T x_i + b))/∂w)
b° = b - η_t(λ ∂(1 - y_i(w^T x_i + b))/∂b)
SGD

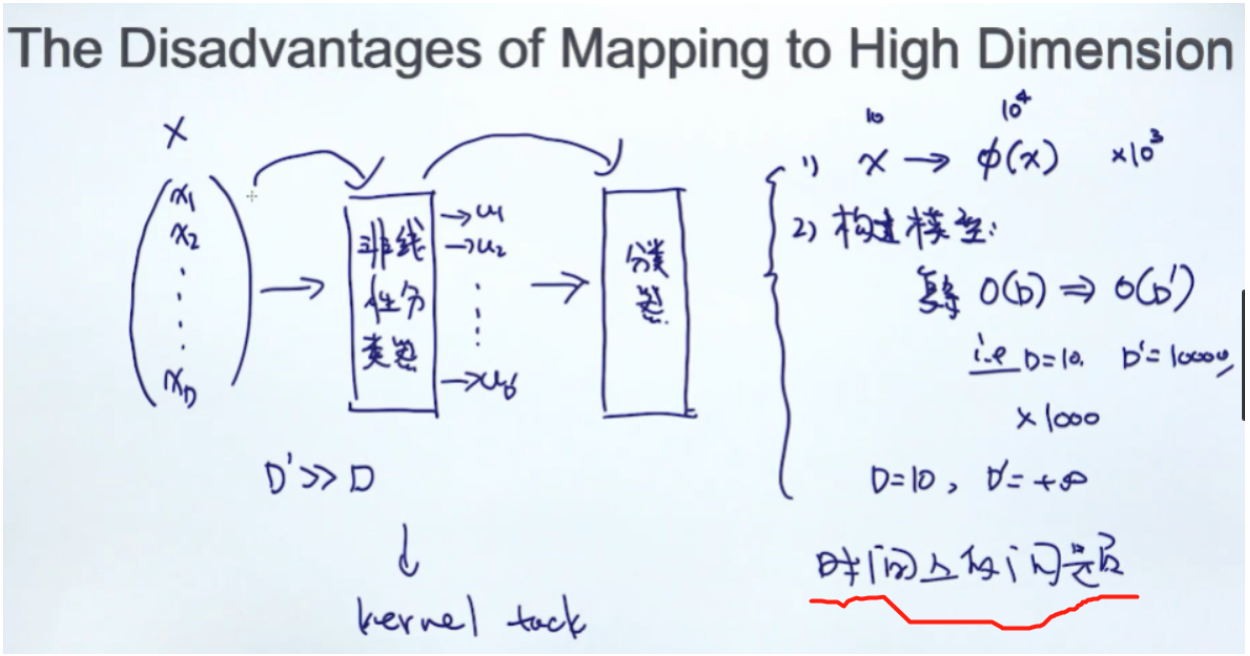
对偶


Linear SVM : Primal Form
Linear SVM
primal form
Dual Form
kernel trick
min f(x)
s.t. g_i(x) ≤ 0
h_j(x) = 0
X*
X*'
(sub)-optimal

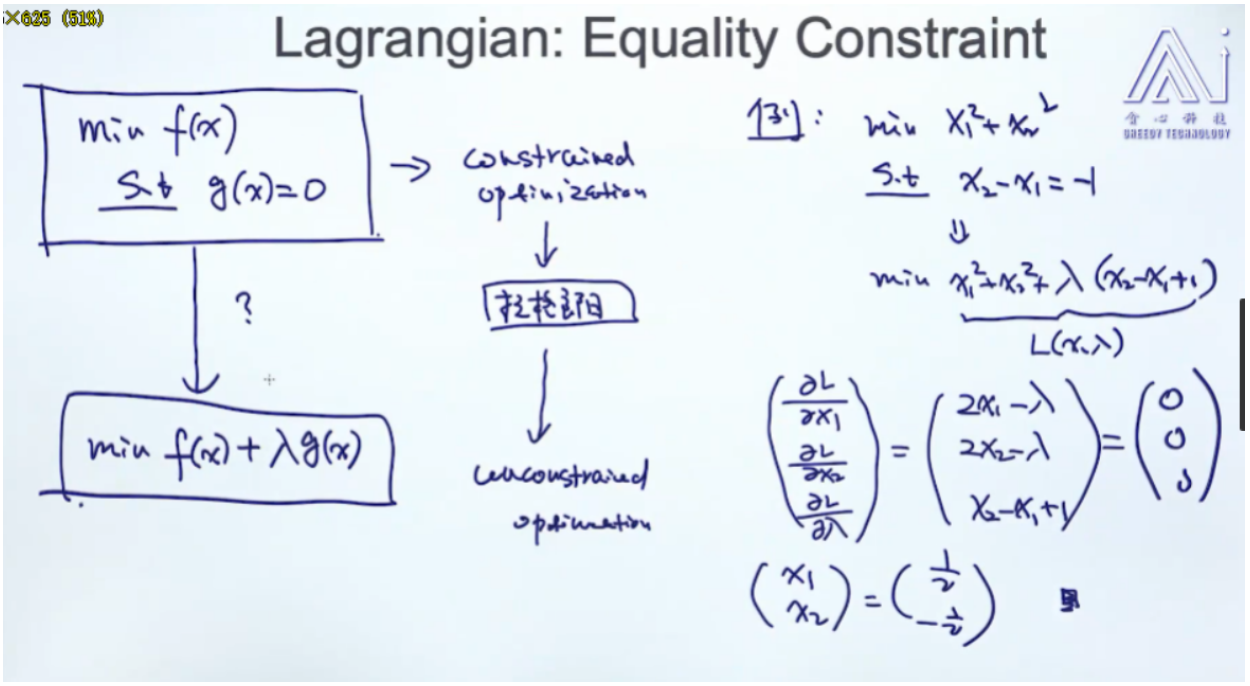




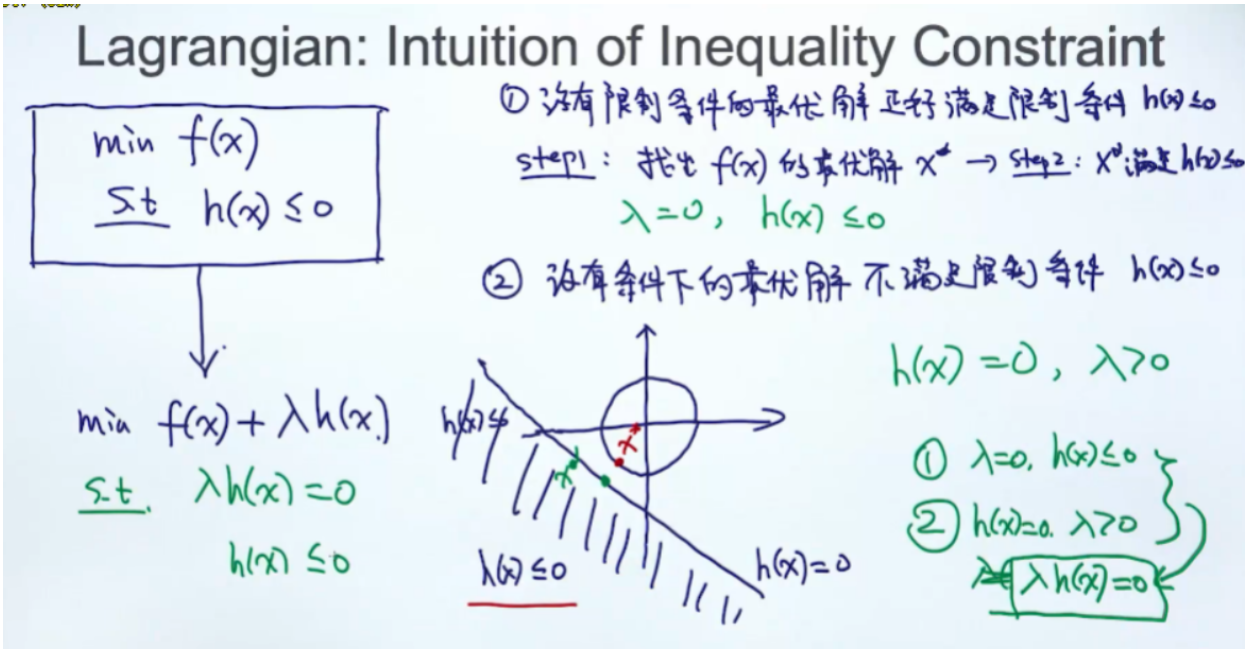
映射到高维空间会有时间复杂度带来的问题



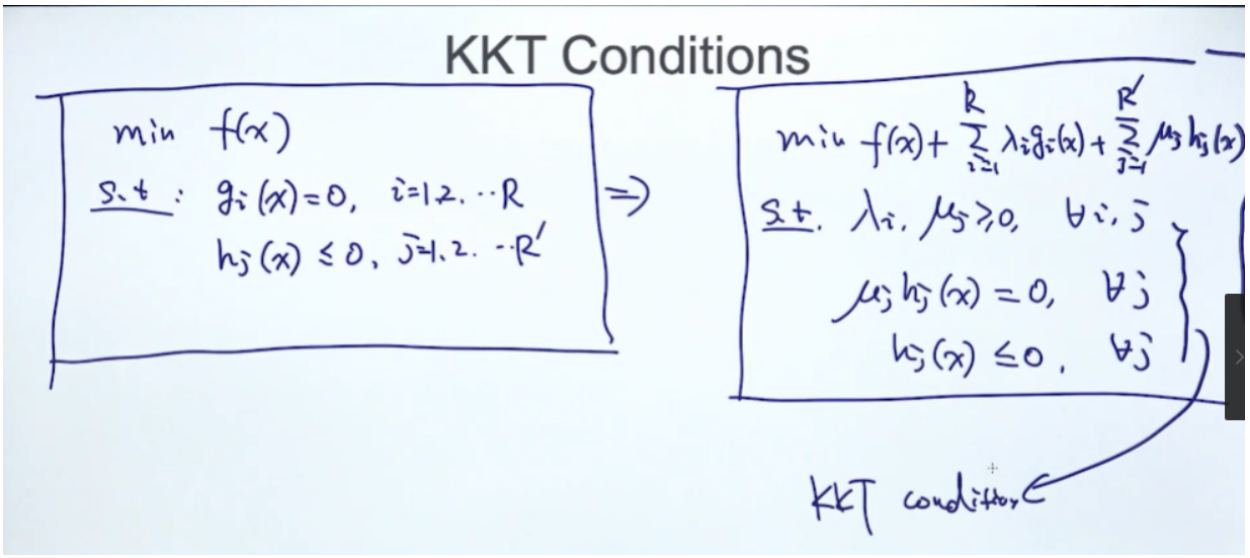
拉格朗日乘子



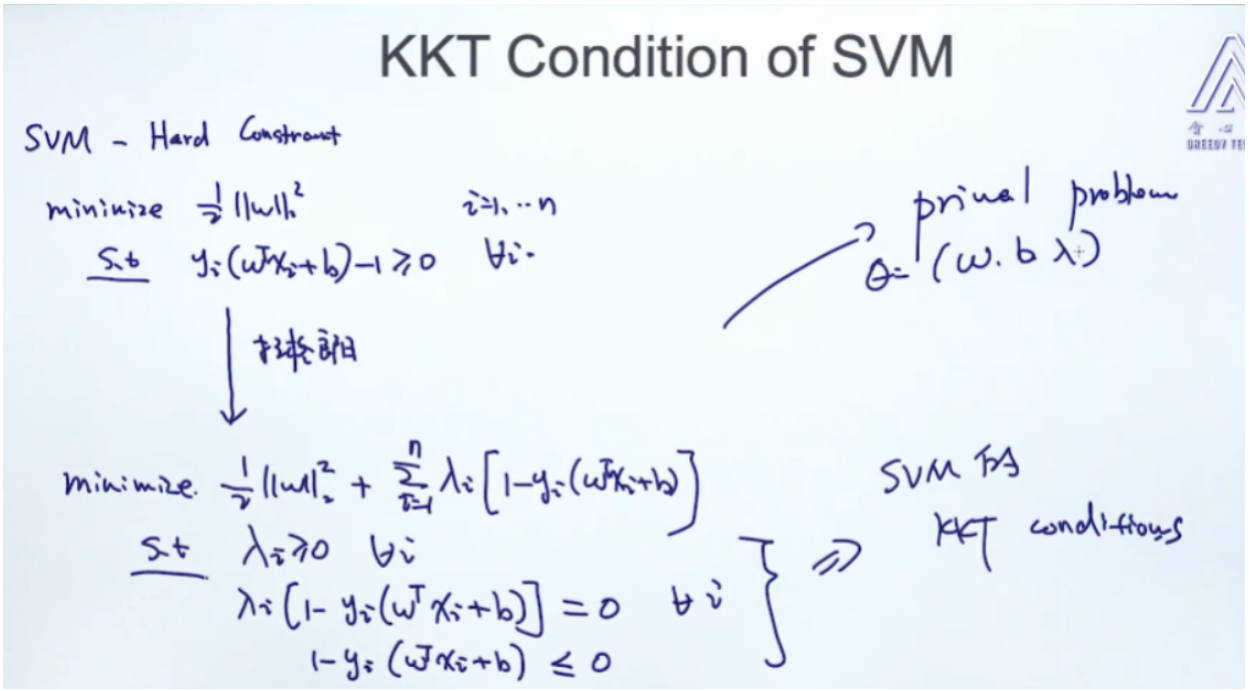
hx(x)小于0时 拉格朗日函数形式



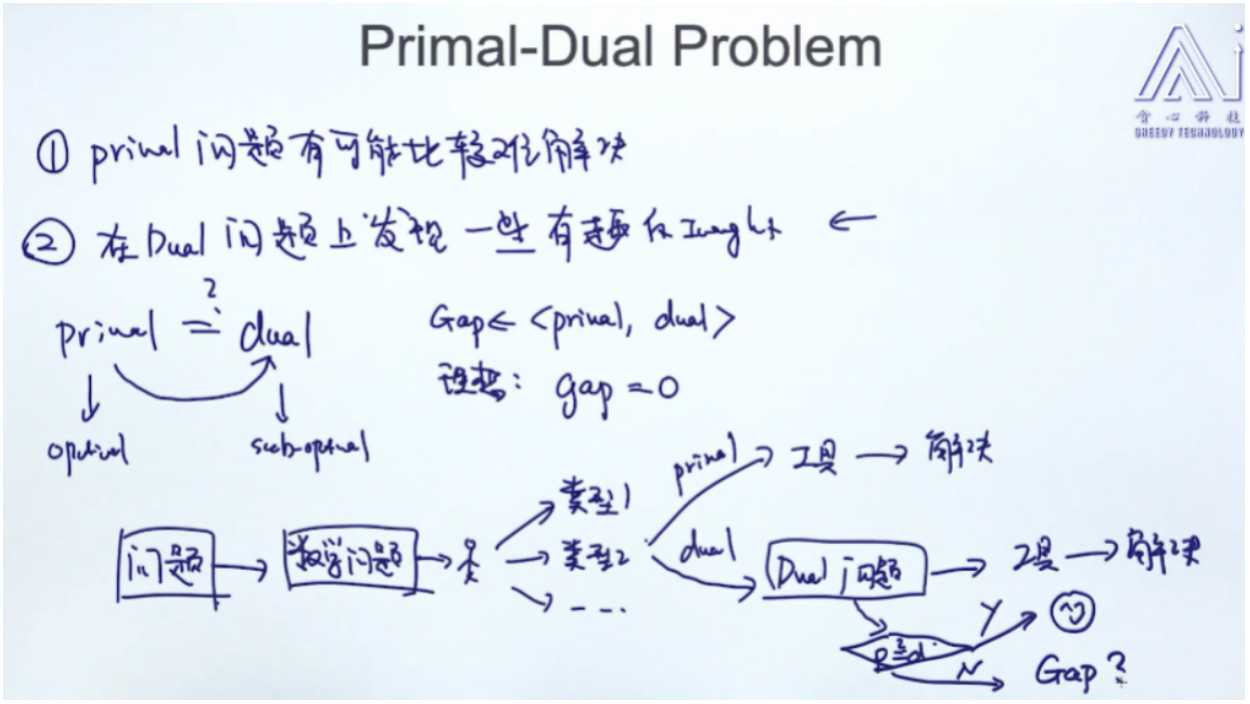
kkt条件



svm 的kkt条件


KKT Condition of SVM
SVM - Hard Constraint
minimize $\frac{1}{2}\|w\|_2^2$ $i=1,\dots n$
s.t $y_i(w^Tx_i+b)-1\geq 0$ $\forall i$
拉格朗日
primal problem
$\theta(w,b,\lambda)$
minimize $\frac{1}{2}\|w\|_2^2+\sum_{i=1}^{n}\lambda_i[1-y_i(w^Tx_i+b)]$
s.t $\lambda_i\geq 0$ $\forall i$
$\lambda_i[1-y_i(w^Tx_i+b)]=0$ $\forall i$
$1-y_i(w^Tx_i+b)\leq 0$
SVM 的 KKT conditions

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 insight
primal ≟ dual
optimal
sub-optimal
Gap ≤ <primal, dual>
理想: gap = 0
问题
数学问题
类型1
类型2
...
primal
dual
工具
解决
Dual 问题
Gap?

Dual Derivation of SVM

$$L = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \lambda_i \left[1 - y_i(w^T x_i + b)\right]$$

s.t. $\lambda_i \geq 0, \quad \forall i$

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

$1 - y_i(w^T x_i + b) \leq 0 \quad \forall i$

$$\frac{1}{2}\|w\|_2^2 = \frac{1}{2} w^T \cdot w = \frac{1}{2}\left(\sum_{i=1}^{n} \lambda_i y_i x_i\right)^T \cdot \left(\sum_{j=1}^{n} \lambda_j y_j x_j\right)$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T \cdot x_j$$

$$\sum_{i=1}^{n} \lambda_i \left[1 - y_i(w^T x_i + b)\right] = \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i \cdot y_i (w^T x_i + b)$$

$$= \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i y_i w^T x_i - \sum_{i=1}^{n} \lambda_i y_i \cdot b$$

$$= \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \lambda_i y_i \left(\sum_{j=1}^{n} \lambda_j y_j x_j\right)^T \cdot x_i$$

$$= \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T \cdot x_j$$

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

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

$$\therefore \boxed{L = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T \cdot x_j + \sum_{i=1}^{n} \lambda_i}$$

Dual

Recall: Mapping to High Dimension

$x \xrightarrow{\phi(x) = u}$ 非线性映射 $\xrightarrow{\phi(\phi(x)) = y}$ 分类器

$x = (x_1, x_2), \quad z = (z_1, z_2)$

$$x^T \cdot z = (x_1 z_1 + x_2 z_2)$$

$\phi(x) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2) \quad \phi(z) = (z_1^2, z_2^2, \sqrt{2} z_1 z_2)$

$$\phi(x)^T \cdot \phi(z) = (x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2)$$

$$= (x_1 z_1 + x_2 z_2)^2 = (x^T z)^2$$

$D' \gg D$

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

$(x^T \cdot z)^2$

Dual Derivation of SVM

Dual Formation of SVM

$\phi(\cdot) = ?$

$O(\phi(x)^T \phi(y)) \approx O(x^T y)$

minimize $-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \underbrace{x_i^T \cdot x_j}_{\phi(x_i)^T \cdot \phi(x_j)} + \frac{1}{2} \sum_{i=1}^{n} \lambda_i$

s.t. $\lambda_i \geq 0 \quad \forall i$

$\sum_{i=1}^{n} \lambda_i y_i = 0, \quad w = \sum_{i=1}^{n} \lambda_i y_i x_i$

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

$\Rightarrow$ 未知参数 λ

Kernel Trick

$f(x) \cdots x_i^T x_j$ $\phi(\cdot)$

kernel trick：不仅仅是针对 SVM

$(x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2)$

Mercer's Theorem

Linear kernel：

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

Polynomial kernel：

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

Gaussian kernel：

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

总结

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 $\stackrel{?}{=}$ dual