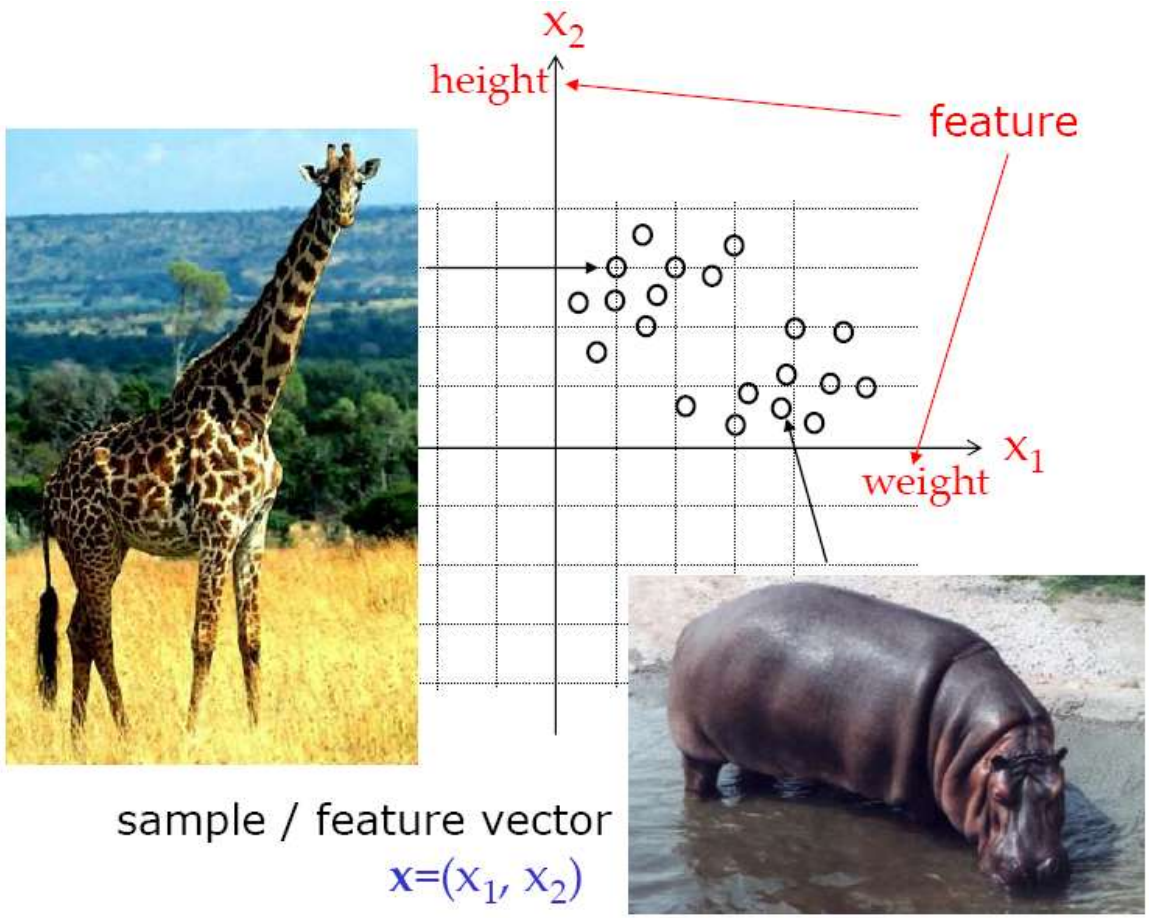
Support Vector Machine
Part 1

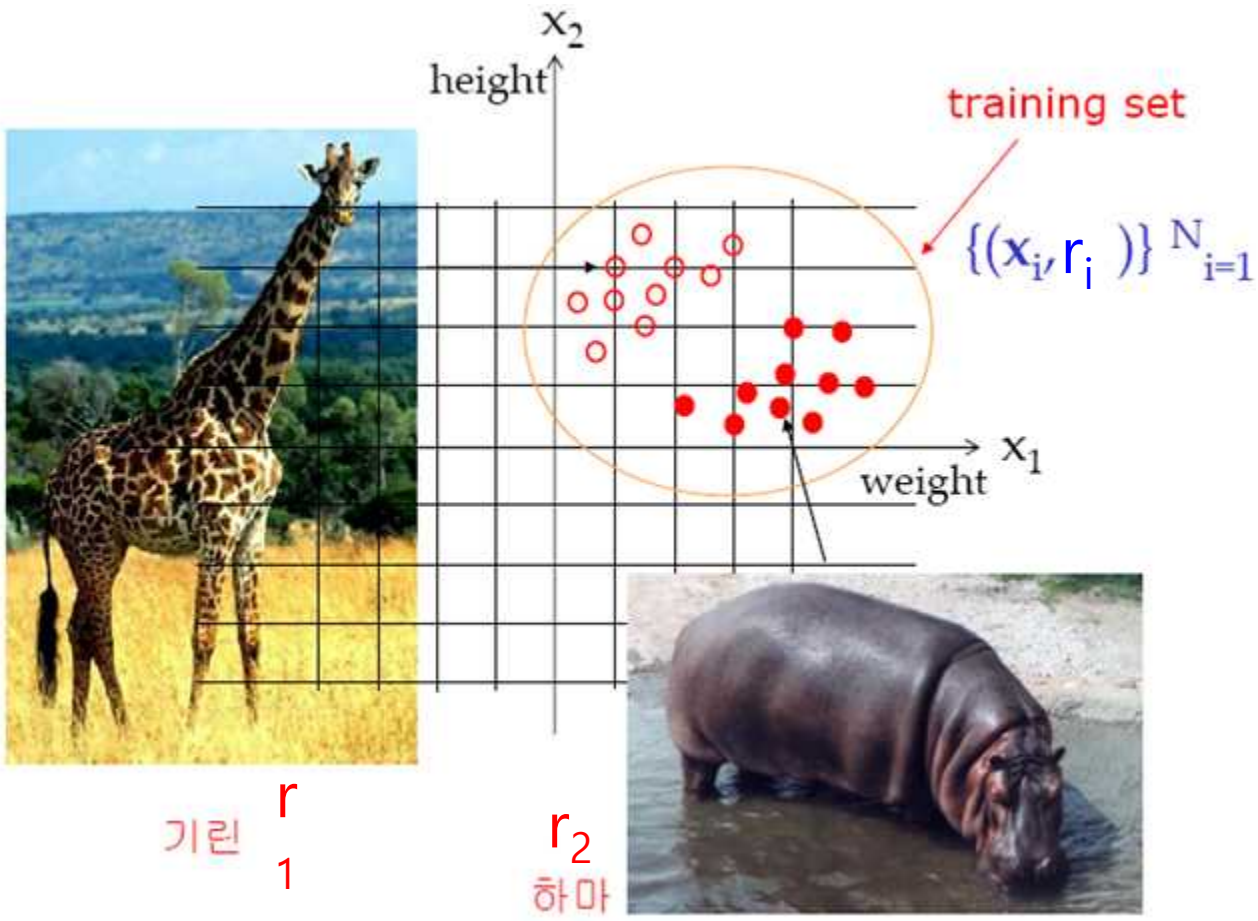
Feature Space

- Sample



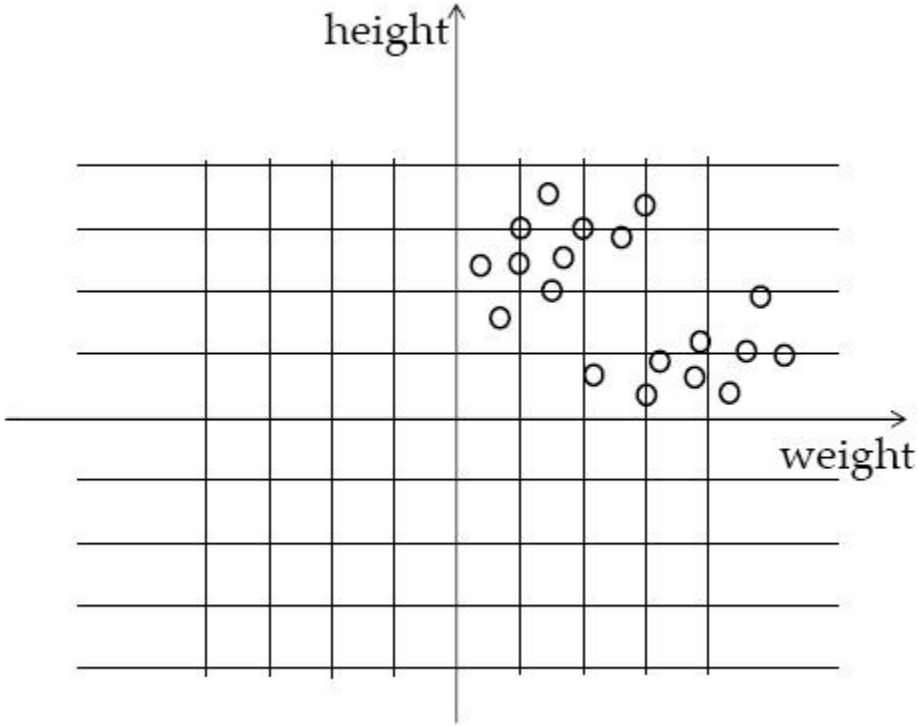
Feature Space

- Training Set



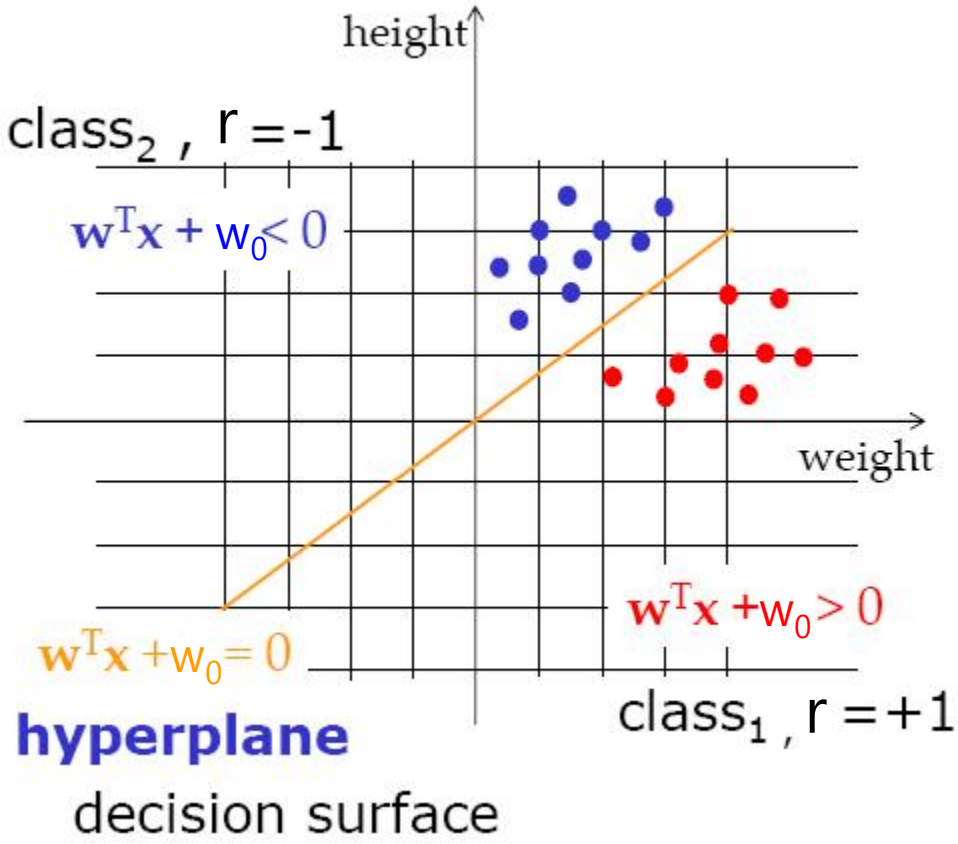
Feature Space

- How to classify them using computer ?



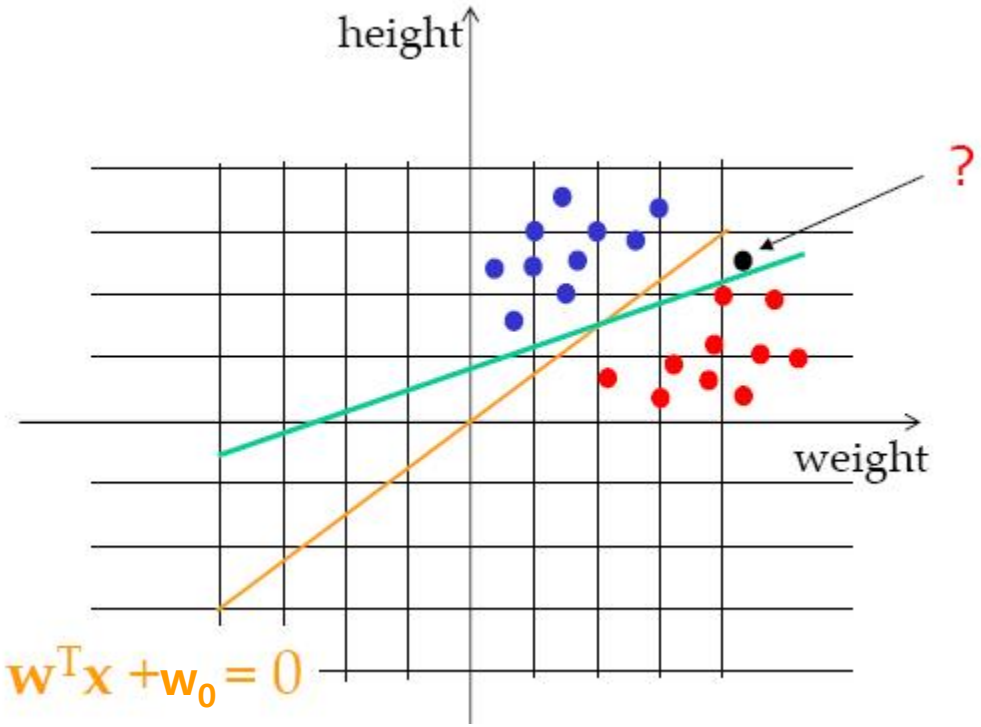
Optimal Hyperplane

- Linear classification



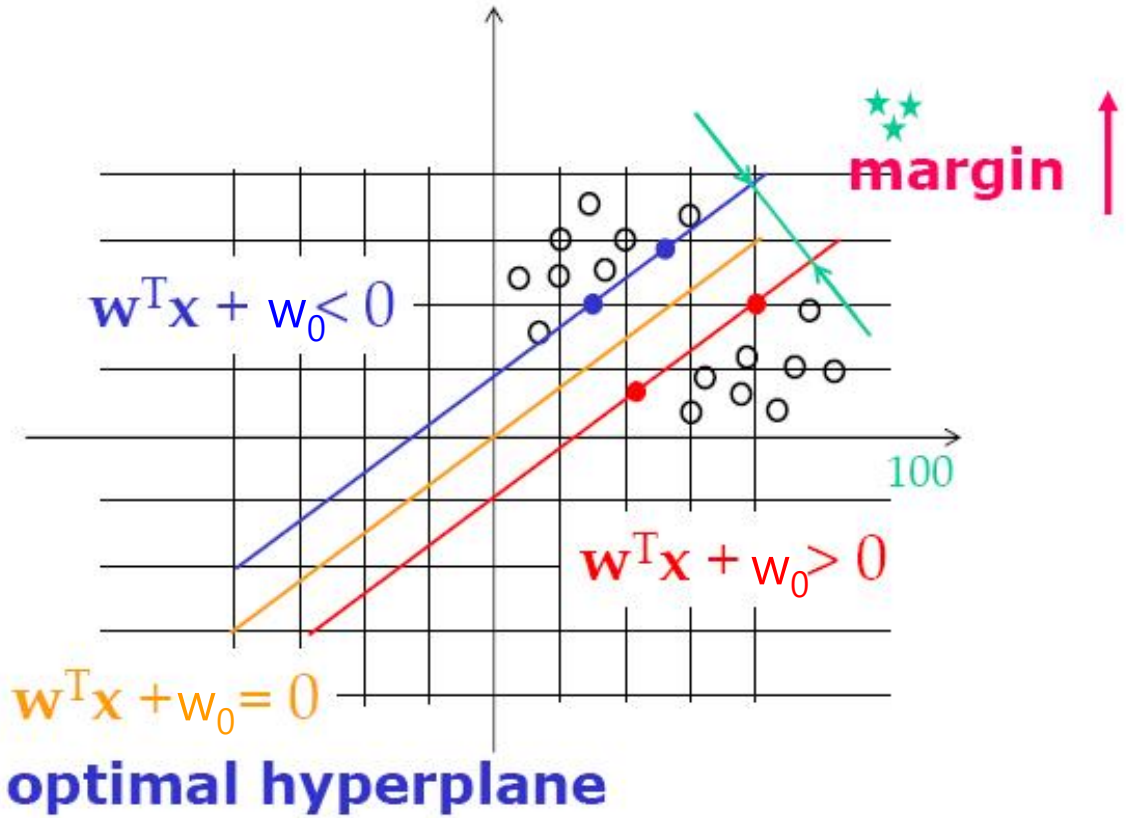
Optimal Hyperplane

- Linear classification



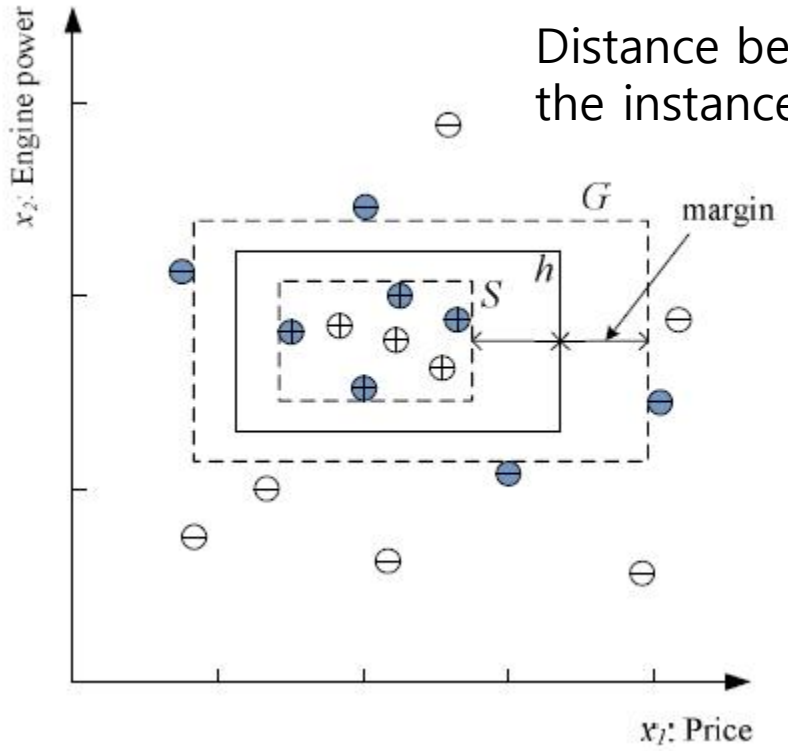
Optimal Hyperplane

- Margin



Optimal Hyperplane

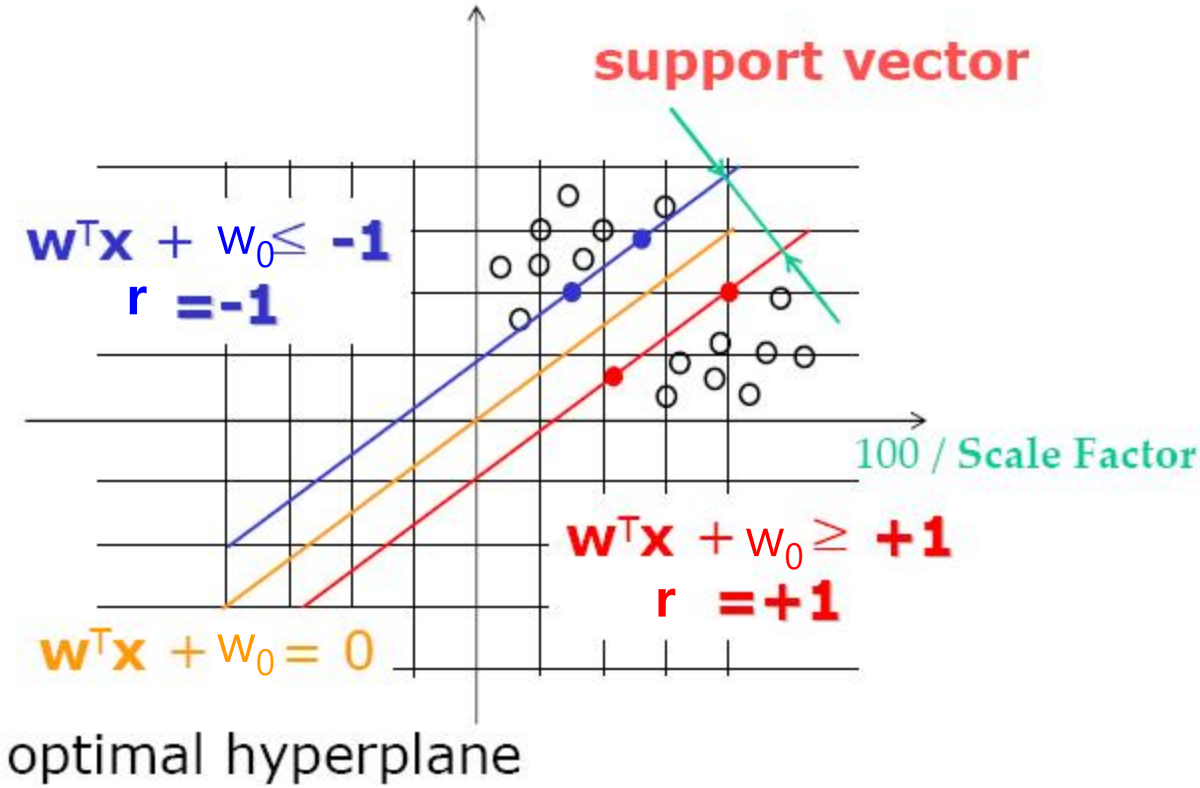
- Choose *h* with largest margin
  - Minimize the error function



Distance between the boundary and the instances closest to it

Optimal Hyperplane

- Support vector



Optimization Problem

$$\mathcal{X} = \left\{ \boldsymbol{x}^t, r^t \right\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \boldsymbol{x}^t \in C_1 \\ -1 & \text{if } \boldsymbol{x}^t \in C_2 \end{cases}$$

find $\boldsymbol{w}$ and w_0 such that

$$\boldsymbol{w}^T \boldsymbol{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$\boldsymbol{w}^T \boldsymbol{x}^t + w_0 \leq -1 \text{ for } r^t = -1$$

which can be rewritten as

$$r^t \left(\boldsymbol{w}^T \boldsymbol{x}^t + w_0 \right) \geq +1$$

(Cortes and Vapnik, 1995; Vapnik, 1995)

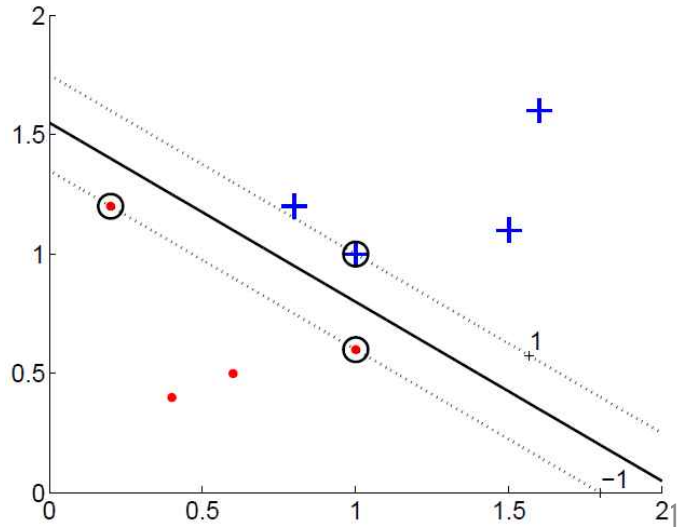
Optimization Problem

- Margin
  - Distance from the discriminant to the closest instances on either side
    - Distance of x^t to the hyperplane is
    - Margin for support vectors

$$\frac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$$

  - Maximize t $\left|\frac{1}{\|\mathbf{w}\|} - \frac{-1}{\|\mathbf{w}\|}\right| = \frac{2}{\|\mathbf{w}\|}$

    ⇒ Minimize the ||w||



Optimization Problem

– Constrained optimization problem

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2$$

: Convex cost function

$$\text{s.t. } r^t\left(\boldsymbol{w}^T\boldsymbol{x}^t + w_0\right) \geq +1, \forall t$$

: Linear constraints

- Standard quadratic optimization problem
- Convex optimization problem

  : Local solution ⇒ Global optimal solution

Optimization Problem

- Lagrange Multipliers Method

cost function: $f(\mathbf{x})$

constraint function: $g(\mathbf{x})=0$

$$\mathbf{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda\, g(\mathbf{x})$$

Lagrange function Lagrange multiplier

A Brief Overview of Optimization Theory

(2001년 추계 CVPR 튜토리얼)

- **Theorem**: $f \in C^1$ has a min. at $x^* \Rightarrow \frac{\partial f}{\partial x}(x^*) = 0$.

  This condition, together with convexity of f, is also a sufficient condition.

- Example 1: $\min . f(x) = \frac{1}{2}(x_1^2 + x_2^2)$

  Solution:

$$\frac{\partial f}{\partial x} = 0 \Rightarrow \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2}\right] = [x_1 \ x_2] = 0 \quad \therefore \ x^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- In a constrained min. problem,

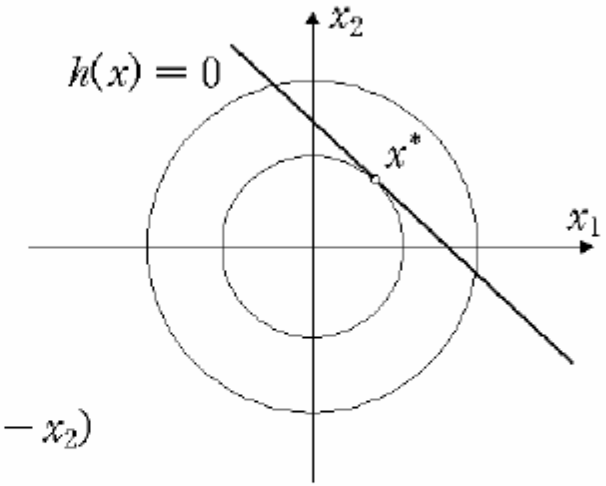
$$f \in C^1 \text{ has a min. at } x^* \not\Rightarrow \frac{\partial f}{\partial x}(x^*) = 0$$

A Brief Overview of Optimization Theory

- Example 2:

$$\min . f(x) = \frac{1}{2}(x_1^2 + x_2^2)$$

$$\text{s.t. } h(x) = 1 - x_1 - x_2 = 0$$



Solution: Define the Lagrange function

$$L(x, \lambda) = f(x) + \lambda h(x)$$
$$= \frac{1}{2}(x_1^2 + x_2^2) + \lambda(1 - x_1 - x_2)$$

$$\frac{\partial L}{\partial x} = 0 \Rightarrow [x_1 - \lambda \;\; x_2 - \lambda] = 0 . \;\therefore\; x_1 = x_2 = \lambda .$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow 1 - x_1 - x_2 = 0 . \;\therefore\; 1 - 2\lambda = 0 . \;\therefore\; \lambda = \frac{1}{2} .$$

$$\therefore\; x^* = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

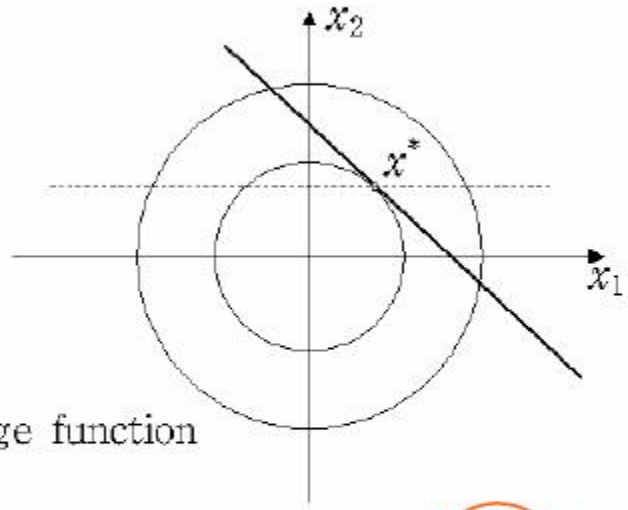
A Brief Overview of Optimization Theory

- Example 3:

$$\min.\ f(x) = \frac{1}{2}(x_1^2 + x_2^2)$$

$$\text{s.t.}\ \ h(x) = 1 - x_1 - x_2 = 0,$$

$$g(x) = \frac{3}{4} - x_2 \le 0$$



Solution: Define the generalized Lagrange function

$$L(x, \lambda, \alpha) \triangleq f + \lambda h + \alpha g$$

$$= \frac{1}{2}(x_1^2 + x_2^2) + \lambda(1 - x_1 - x_2) + \alpha\left(\frac{3}{4} - x_2\right), \quad \alpha \ge 0$$

$$\frac{\partial L}{\partial x} = 0 \Rightarrow [x_1 - \lambda,\ x_2 - \lambda - \alpha] = 0. \ \therefore\ x_1 = \lambda,\ x_2 = \lambda + \alpha$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow 1 - x_1 - x_2 = 0. \ \therefore\ 2\lambda + \alpha = 1$$

Also, $\alpha \geq 0$ and $\frac{3}{4} - x_2 \leq 0$.

One more condition is needed to solve the problem.

Karush

→ The Kuhn-Tucker complementarity condition

$\alpha(\frac{3}{4} - x_2) = 0$ i.e. $\alpha = 0$ or $x_2 = \frac{3}{4}$

① If $\alpha = 0$, then $\lambda = \frac{1}{2}$; thus $x_1 = x_2 = \frac{1}{2}$ ✗ $(\because x_2 \geq \frac{3}{4})$

② If $x_2 = \frac{3}{4}$, then $\begin{cases} \lambda + \alpha = \frac{3}{4} \\ 2\lambda + \alpha = 1 \end{cases}$ $\therefore \begin{cases} \lambda = \frac{1}{4}, \alpha = \frac{1}{2} \\ x_1 = \frac{1}{4}, x_2 = \frac{3}{4} \end{cases}$

$\therefore x^* = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}$

A Brief Overview of Optimization Theory

Theorem (Kuhn-Tucker Theorem)

Given an opt. prob. with convex domain $\Omega \subseteq R^n$

$$\left.\begin{array}{ll} \min & f(x),\ x \in \Omega \ (x \text{ is primal variable}) \\ \text{s.t.} & g_i(x) \le 0,\ i = 1, \cdots, k \\ & h_j(x) = 0,\ j = 1, \cdots, m \end{array}\right\}$$

primal opt. prob.

with $f \in C^1$ convex, and g_i, h_j affine, the following are necessary and sufficient conditions for a point $x^* \in \Omega$ to be an opt.:

For $L(x, \alpha, \lambda) \triangleq f(x) + \sum_{i=1}^{k} \alpha_i g_i(x) + \sum_{j=1}^{m} \lambda_j h_j(x) = f + \alpha^T g + \lambda^T h$,

$\exists\ \alpha^*$ and λ^* s.t. $\dfrac{\partial L}{\partial x}(x^*, \alpha^*, \lambda^*) = 0$, $\dfrac{\partial L}{\partial \lambda}(x^*, \alpha^*, \lambda^*) = 0$

$g_i(x^*) \le 0$, $\alpha_i^* \ge 0$ for $i = 1, \cdots, k$,

and $\alpha_i^* g_i(x^*) = 0,\ i = 1, \cdots, k$

Optimization Problem

- Problem

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t } r^t\left(\boldsymbol{w}^T\boldsymbol{x}^t + w_0\right) \geq +1, \forall t$$

- Lagrange function (Primal)

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{t=1}^{N}\alpha^t\left[r^t\left(\boldsymbol{w}^T\boldsymbol{x}^t + w_0\right) - 1\right]$$
$$= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{t=1}^{N}\alpha^t r^t\left(\boldsymbol{w}^T\boldsymbol{x}^t + w_0\right) + \sum_{t=1}^{N}\alpha^t$$

- Solution

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{t=1}^{N}\alpha^t r^t \boldsymbol{x}^t$$
$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N}\alpha^t r^t = 0$$

$$\text{KKT condition} \Rightarrow \alpha^t\left(1 - r^t\left(w^T x^t + w_0\right)\right) = 0$$

Support Vector Machine

- Support Vector Machine (SVM) (Vapnik 1995)

$$\text{KKT condition}: \alpha^t \left(1 - r^t \left(w^T x^t + w_0\right)\right) = 0$$

$$\begin{aligned} &\text{If } \alpha^t \neq 0 \text{ then } r^t \left(w^T x^t + w_0\right) = 1 \\ &\quad \Rightarrow \quad x^t \text{ is support vector} \\ &\text{If } r^t \left(w^T x^t + w_0\right) \neq 1 \text{ then } \alpha^t = 0 \\ &\quad \Rightarrow \quad x^t \text{ is not support vector} \end{aligned}$$

$$\boldsymbol{w} = \sum_{t=1}^{N} \alpha^t r^t x^t$$

: $\boldsymbol{w}$ is only related to support vectors x^t

Most α^t are 0 and only a small number have $\alpha^t > 0$; they are the support vectors

Support Vector Machine

- Dual Problem

$$L_d(\alpha^t) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{t=1}^{N} \alpha^t \left[r^t \left(\boldsymbol{w}^T \boldsymbol{x}^t + w_0 \right) - 1 \right]$$

$$= \frac{1}{2}\left(\boldsymbol{w}^T \boldsymbol{w} \right) - \boldsymbol{w}^T \sum_t \alpha^t r^t \boldsymbol{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}\left(\boldsymbol{w}^T \boldsymbol{w} \right) + \sum_t \alpha^t$$

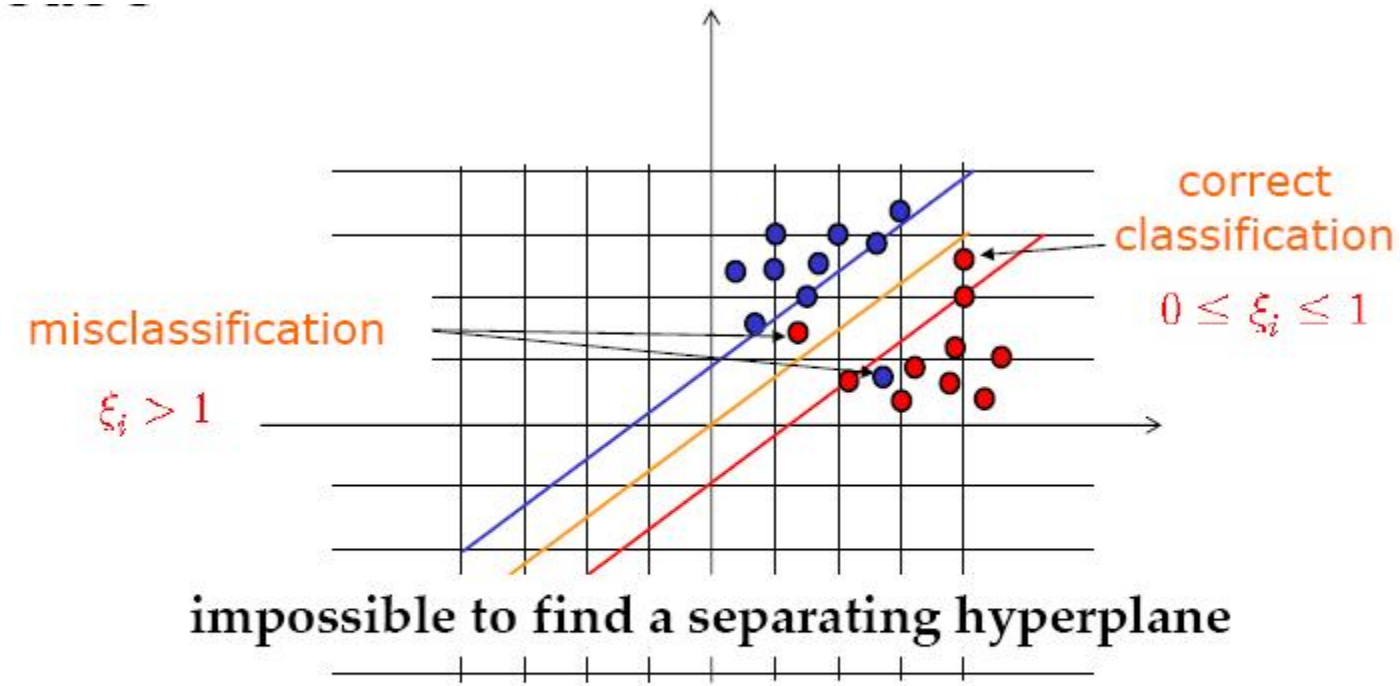
$$= -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s \left(\boldsymbol{x}^t \right)^T \boldsymbol{x}^s + \sum_t \alpha^t$$

subject to $\sum_t \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$

$$\alpha^{t*} = \arg\min_{\alpha^t} L_d$$

Soft Margin Hyperplane

- Non-separable case



impossible to find a separating hyperplane

give them up as errors while minimizing the probabilities of classification error averaged over the training set

Soft Margin Hyperplane

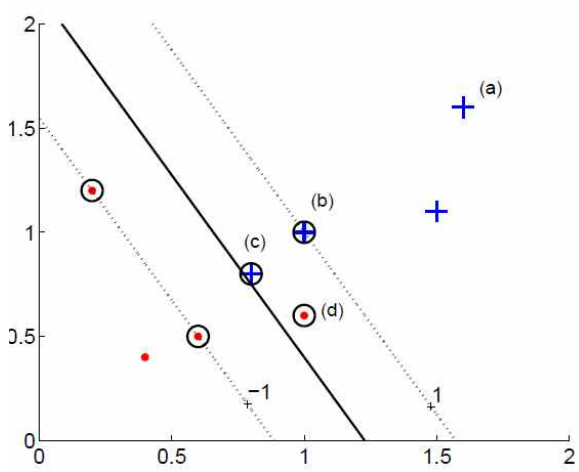
- Soft Error
  - Problem: can't satisfy $r^t\left(\boldsymbol{w}^T\,\boldsymbol{x}^t + w_0\right) \geq 1 \quad ,\forall t$
  - Relaxing the equation using slack variables $\xi^t \geq 0$

$$r^t\left(\boldsymbol{w}^T x^t + w_0\right) \geq 1 - \xi^t$$

$\xi^t = 0$: no problem

$0 < \xi^t < 1$: correct classification

$\xi^t \geq 1$: misclassification



  - Soft error

$$\sum_t \xi^t$$

$\#\left\{\xi^t > 1\right\}$: number of misclassification points

Soft Margin Hyperplane

- Cost Function

$$\text{mininize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t$$

- Lagrange function

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\nolimits_t \xi^t - \sum\nolimits_t \alpha^t \left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1 + \xi^t\right] - \sum\nolimits_t \mu^t \xi^t$$

- Solution

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{t=1}^{N} \alpha^t r^t \boldsymbol{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

$$\frac{\partial L_p}{\partial \xi^t} = 0 \Rightarrow C - \alpha^t - \mu^t = 0$$

$$0 \le \alpha^t \le C \quad \left(\because \mu^t \ge 0\right)$$

The optimal value of C is determined experimentally

ν (nu) -SVM

- Replacing C with the parameter ν (Scholkopf, 2000)

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{N}\sum_t \xi^t$$

subject to

$$r^t\left(\boldsymbol{w}^T\boldsymbol{x}^t + w_0\right) \geq \rho - \xi^t, \xi^t \geq 0, \rho \geq 0$$

  - ρ : optimization variable (scales the margin),
    margin size = $\rho / \| \boldsymbol{w} \|$

  - $\nu \in [0,1]$: parameter
    lower bound on the fraction of support vectors
    upper bound on the fraction of instances having margin errors

$$\sum_t \alpha^t r^t = 0$$

$$0 \leq \alpha^t \leq \frac{1}{N}$$

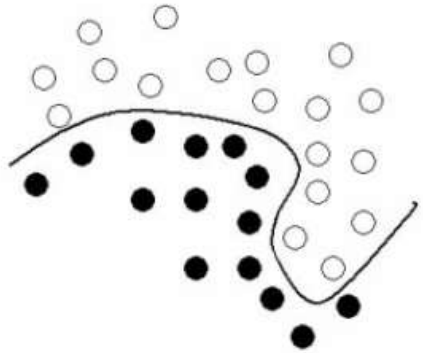
$$\sum_t \alpha^t \leq \nu$$

Nonlinear SVM

- Nonlinear SVM
  - Basis function

$$\boldsymbol{z}=\boldsymbol{\varphi}(\boldsymbol{x}) \quad , z_j=\phi_j(\boldsymbol{x}), j=1\cdots,k$$

$$g(\boldsymbol{x})=\boldsymbol{w}^T\,\boldsymbol{\varphi}(\boldsymbol{x})=\sum_{j=1}^{k} w_j\phi_j(\boldsymbol{x})$$



  - Solution

$$\boldsymbol{w}=\sum_t \alpha^t r^t \boldsymbol{\varphi}\left(\boldsymbol{x}^t\right)$$

$$g(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x})=\sum_t \alpha^t r^t \boldsymbol{\varphi}\left(\boldsymbol{x}^t\right)^T\boldsymbol{\varphi}(\boldsymbol{x})$$

$$=\sum_t \alpha^t r^t K\left(\boldsymbol{x}^t,\boldsymbol{x}\right)$$

Nonlinear SVM

Kernel

Kernel: a function k that takes 2 variables and computes a scalar value (a kind of **similarity**)

$$k(\mathrm{x}, \mathrm{y}) = (\Phi(\mathrm{x}) \cdot \Phi(\mathrm{y}))$$

Kernel Matrix: m ×m matrix K with elements $K_{ij} = k(x_i, x_j)$.

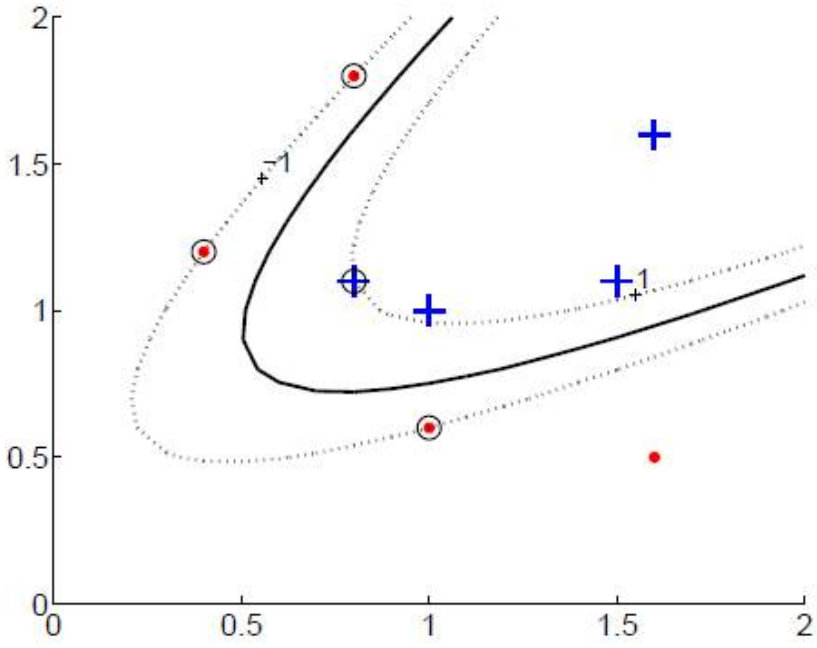
Standard Kernels

Polynomial Kernel	$k(x, x_i) = (x^T x_i + 1)^d$
Radial Basis Function Kernel	$k(x, x_i) = \exp(- \parallel x - x_i \parallel^2 / 2\sigma^2)$
Sigmoid Kernel	$k(x, x_i) = \tanh(\beta_0 x^T x_i + \beta_1)$

Nonlinear SVM

- Polynomials of degree q:

$$K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T \mathbf{x}^t + 1)^q$$



$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^2$$

$$= (x_1 y_1 + x_2 y_2 + 1)^2$$

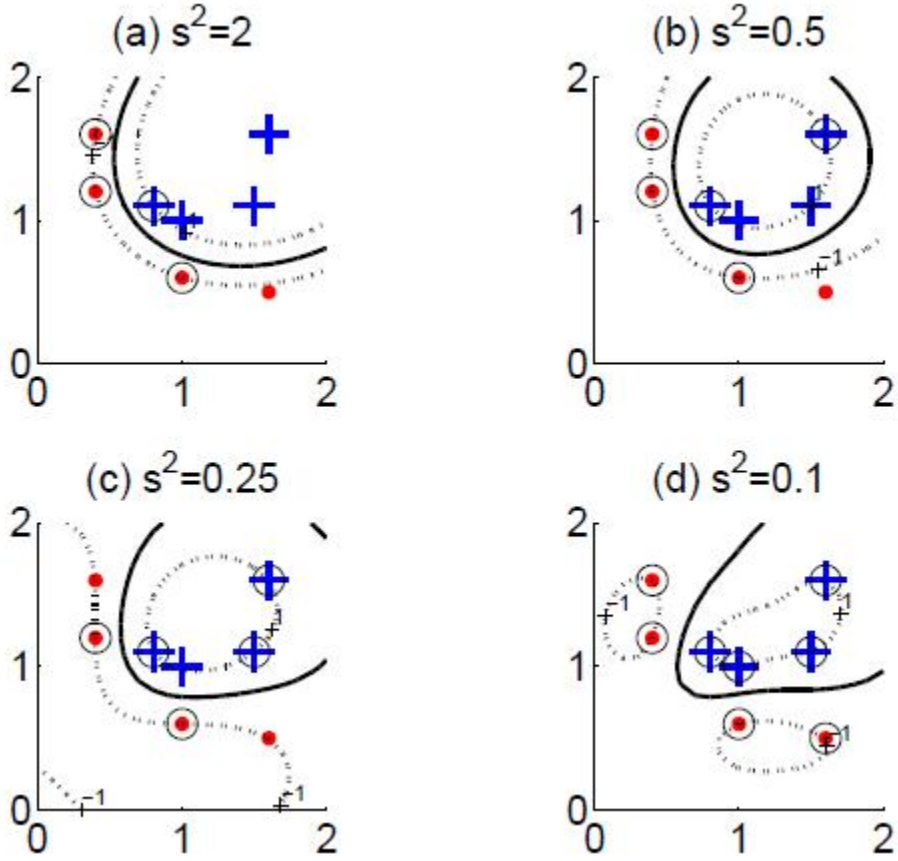
$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$

$$\phi(\mathbf{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

Nonlinear SVM

- Radial-basis functions:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \exp\left[-\frac{\left\|\mathbf{x}^t - \mathbf{x}\right\|^2}{2s^2}\right]$$



Nonlinear SVM

Example: XOR Problem
(small scale QP problem)

- Training set: $\{(-1\ -1; -1), (-1\ +1; +1), (+1\ -1; +1), (+1\ +1; -1)\}$
  $\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$
- Let kernel $k(\mathbf{x}, \mathbf{x}_i)=(1+\mathbf{x}^T\mathbf{x}_i)^2$, $\mathbf{x}=(x_1, x_2)^T$, $\mathbf{x}_i=(x_{i1}, x_{i2})^T$
- Then $k(\mathbf{x}, \mathbf{x}_i) = (1+\mathbf{x}^T\mathbf{x}_i)^2 = (1+x_1x_{i1}+x_2x_{i2})^2$
  $= 1 + x^2_1x^2_{i1} + 2x_1x_2x_{i1}x_{i2} + x^2_2x^2_{i2} + 2x_1x_{i1} + 2_{x2}x_{i2}$
- A Mapping: $\varphi(\mathbf{x})=(1,\ x^2_1, x^2_2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2)^T$
- Kernel Matrix:

$$\Phi = \begin{bmatrix} 1 & 1 & 1 & -\sqrt{2} & -\sqrt{2} & \sqrt{2} \\ 1 & 1 & 1 & -\sqrt{2} & \sqrt{2} & -\sqrt{2} \\ 1 & 1 & 0 & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 1 & 1 & 1 & \sqrt{2} & \sqrt{2} & \sqrt{2} \end{bmatrix}$$

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}_{4 \times 4}$$

Nonlinear SVM

Example: XOR Problem

- Dual Problem:

$$L_D(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - (9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 9\alpha_2^2 + 2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2)/2$$

- Optimizing L_D: $9\alpha_1 - 2\alpha_2 - \alpha_3 + \alpha_4 = 1$, $-\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 = 1$
  $-\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 = 1$, $\alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1$

- Optimal Lagrange multipliers: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/8 > 0$

  All the samples are support vectors

- Optimal $\mathbf{w}$:

$$\mathbf{w} = \sum \alpha_i y_i \varphi(\mathbf{x}_i) = (1/8)(-1)\,\varphi(\mathbf{x}_1) + (1/8)(+1)\varphi(\mathbf{x}_2) + (1/8)(+1)\,\varphi(\mathbf{x}_3) + (1/8)(-1)\,\varphi(\mathbf{x}_4) = (0\ 0\ 0\ 0\ 0\ -1\sqrt{2})^T$$

- Optimal Hyperplane:

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^T\varphi(\mathbf{x})) = \mathrm{sgn}(-x_1x_2)$$

Multiclass SVM

- Multiclass Classification
  - SVM : binary classification
  - Multiclass SVM: multiple binary classification
- One-to-rest vs. one-to-one

  1) One-to-rest
     - I 번째 class 와 I 클래스를 제외한 나머지 M-1 클래스로 이진 분류
     - M 개의 이진 분류 SVM 을 학습하여 사용
  2) One-to-one
     - M 개의 클래스 중 2개를 선택하여 이진 분류
     - M(M-1)/2 개의 이진 분류 SVM 을 학습하여 사용

Open Software

Software	Developer	Language	Environment	Algorithms	URL
SVMFu	R. Rifkin M. Nadermann (MIT)	C++	Unix-like system	Osuna *et al.*, SMO(Platt)	http://www.ai. mit.edu
LIBSVM	C.C. Chang, C.H. Lin (National Taiwan Univ.)	C++, Java	Python, R, Matlab, Perl	SMO(Platt), SVMLight(Joachims)	http://www.csie. ntu.edu.tw/ libsvm
SVMLight	T. Joachims, (Univ. of Dortmund)	C	Solaris, Linux, IRIX, Windows NT	T. Joachims	http://www.svmlight. joachims.org
SVMTorch	R. Collobert, (IDIAP, Switzerland)	C, C++	Windows	R. Collobert	http://www.idap.ch /learning/SVMTorch. html

SVM Demo

https://jgreitemann.github.io/svm-demo

SVM Summary

- Optimal separating hyperplane : max. margin
- Global solution (convexity)
- Analytic solution
- Model selection problem: kernel selection
- Classification & Regression 문제에 모두 적용