CS570 Artificial Intelligence & Machine Learning

# Support Vector Machines

Kee-Eung Kim
Department of Computer Science
KAIST

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

# Support Vector Machines (classification) → assume binary

☐ Key idea: find the optimal separating hyperplane

- $\mathcal{X} = \{\mathbf{x}^t, r^t\}_t$ where $r^t = \begin{cases} +1 \text{ if } \mathbf{x}^t \in C_1 \\ -1 \text{ if } \mathbf{x}^t \in C_2 \end{cases}$
- Find $\mathbf{w}$ and $w_0$ such that

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

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

- Equivalently,

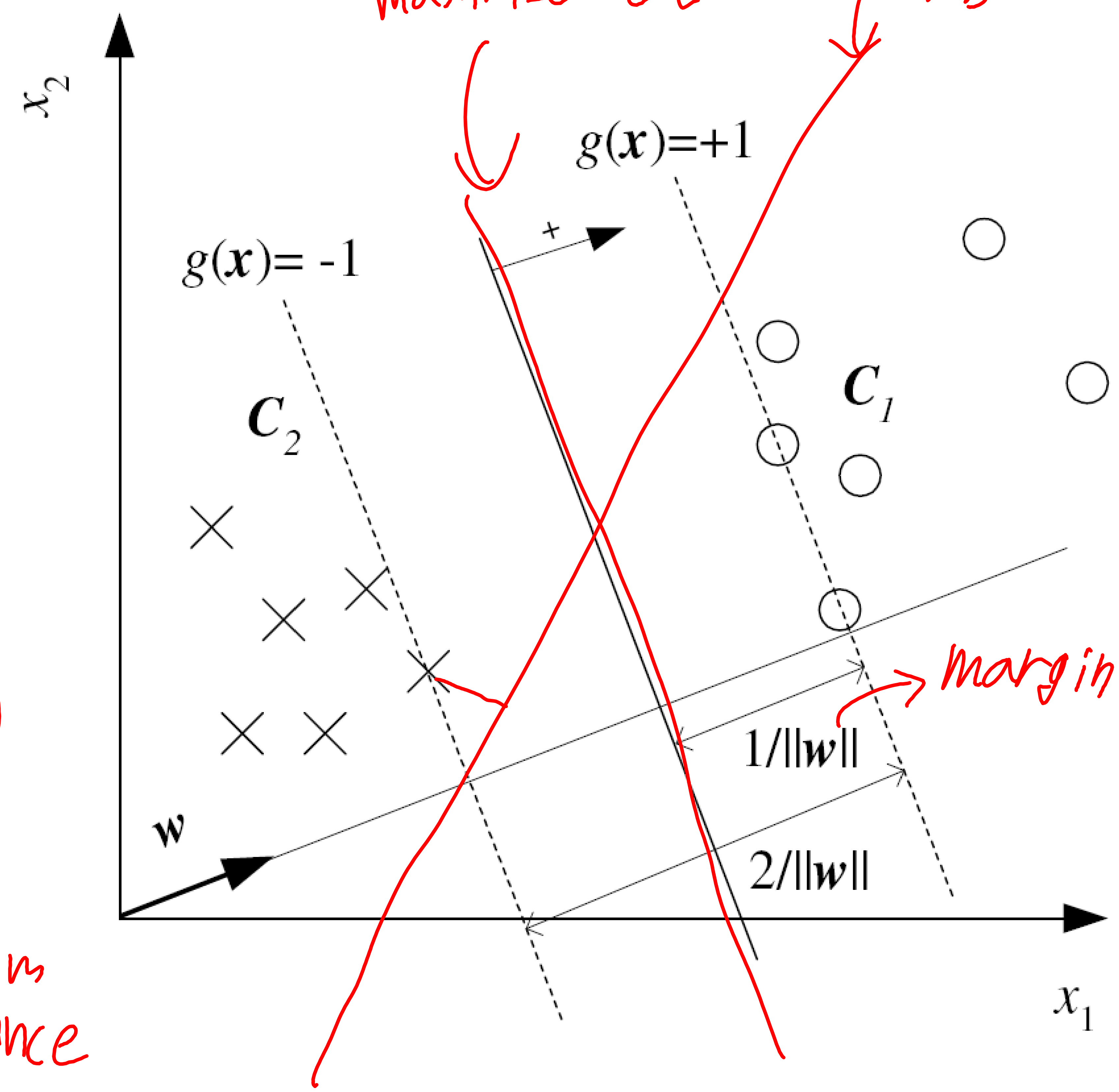
  항상 만족

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

# Margins

거리의 최소값을 maximize 하는 선

노이즈에 취약함




- ☐ Distance from the discriminant to the closest instance on either side

- ☐ Distance of x to the hyperplane:

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

- ☐ Want: $\dfrac{r^t(\mathbf{w}^T\mathbf{x}^t + w_0)}{||\mathbf{w}||} \geq \rho, \forall t$

$r^t$ → 부호를 유지시키기 위해

$\rho$ → margin, → minimum distance

↳ 이걸 만족시키는 w는 사실 많음

- ☐ For a unique solution, fix $\rho||\mathbf{w}|| = 1$ and thus to maximize margin, minimize $||\mathbf{w}||$ : $\min \frac{1}{2}||\mathbf{w}||^2$ subject to $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t$

→ ||w||를 최소화
→ ρ(margin) 최대화

↳ ρ가 커짐

  - Quadratic programming problem! → 매트랩 quad prog (?)

# Maximizing Margins

SVM 훈련시키기
→ margin 최대화

- $\min \frac{1}{2}\|\mathbf{w}\|^2$ subject to $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t$

D+1 variables
N constraints

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

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t(\mathbf{w}^T\mathbf{x}^t + w_0) + \sum_{t=1}^{N} \alpha^t$$

↳ 데이터셋이 클 경우, 최적화값 구하는데 시간이 많이 필요
↳ 따라서 변수가 많되 constraint가 적은게 나음

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

- $L_d = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \underbrace{\mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t}_{w} - \underbrace{w_0 \sum_t \alpha^t r^t}_{0} + \sum_t \alpha^t$

$$= -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_t \alpha^t$$

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

· t, s : index to the dataset

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

1개

w = Σ α^t r^t x^t
→ α^t ≠ 0 일때만 x^t가 contribution
→ 그래서 support VM

- Most $\alpha^t = 0$ and only small number have $\alpha^t > 0$; $x^t$ with $\alpha^t > 0$ are the support vectors

→ optimizer가 찾아낸 대다수의 α는 0일 것임

hyperplane을 결정짓는 가까운 점들

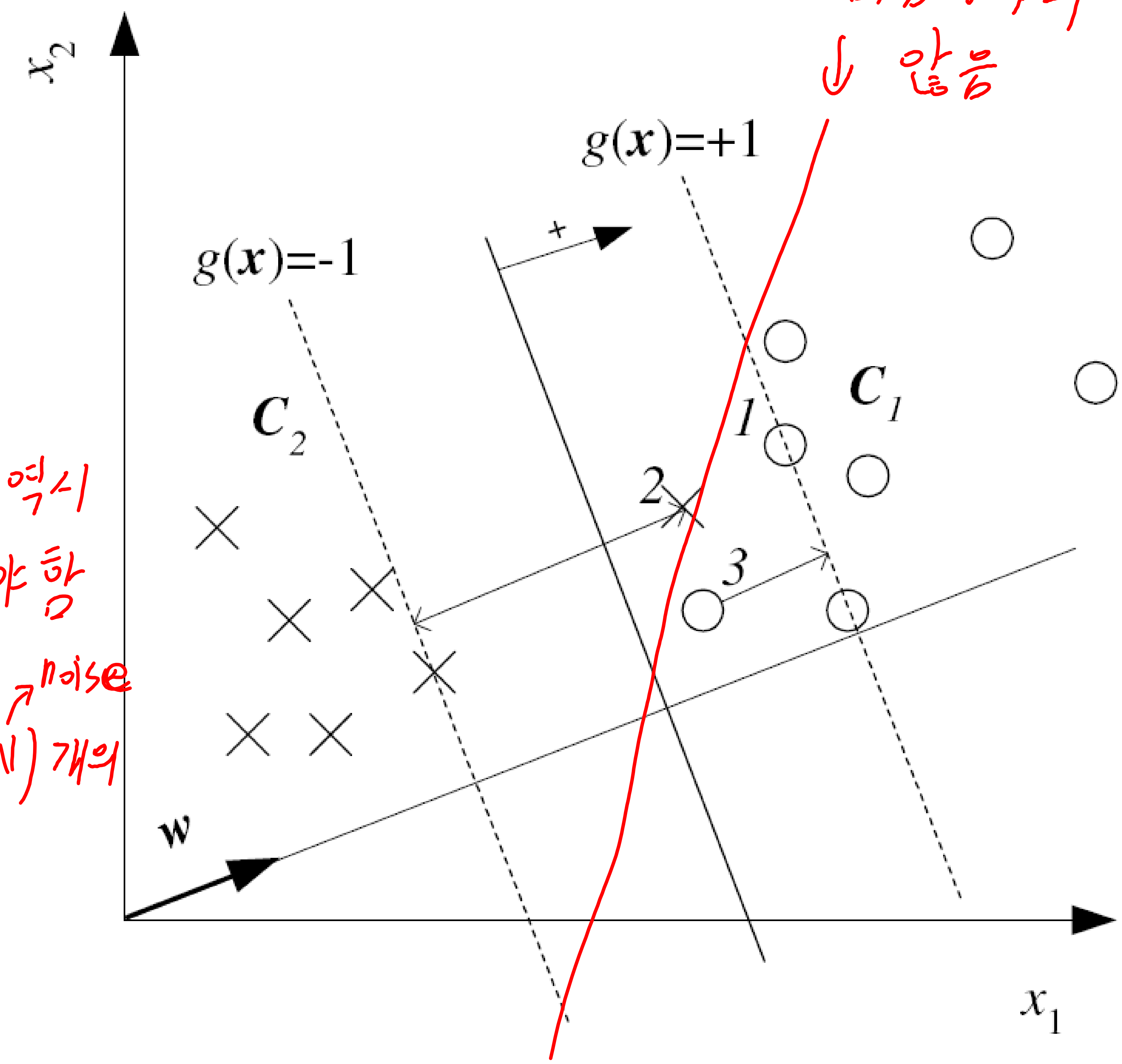
(확률 통계에서의 support → 확률값이 0이 아님

# Soft Margins

- ☐ If not linearly separable

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t$$ error

바람직하지 않음




- ☐ Soft error $\sum_t \xi^t$

- ☐ New objective function:

$$\min \left[\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t\right]$$

이젠 에러 역시 최소화해야 함

tunable

subject to

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t, \forall t$$

$$\xi^t \geq 0, \forall t$$

noise

☆ D+1+(N) 개의 변수

- ☐ New primal is

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

Korea Advanced Institute of Science and Technology
한국과학기술원

# Kernel Machines

- Preprocess input x by basis functions
  - Suppose $\mathbf{z} = \varphi(\mathbf{x})$
  - Prepare transformed training set $\mathcal{Z} = \{\varphi(\mathbf{x}^t), r^t\}$
  - Linear model in space Z is nonlinear model in space X

    $g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} \quad g(\mathbf{x}) = \mathbf{w}^T\varphi(\mathbf{x})$ ↳ key idea
- SVM on the transformed space Z
  - $\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \varphi(\mathbf{x}^t)$
  - $g(\mathbf{x}) = \mathbf{w}^T\varphi(\mathbf{x}) = \sum_t \alpha^t r^t \varphi(\mathbf{x}^t)^T\varphi(\mathbf{x})$ → $K(x^t, x)$ (커널)
  - $g(\mathbf{x}) = \sum_t \alpha^t r^t K(\mathbf{x}^t, \mathbf{x})$
    - linear → $x^T x$
    - non-linear
- Kernel functions K
  - Polynomials of degree q: $K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T\mathbf{x}^t + 1)^q$
    - $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^2 = (x_1y_1 + x_2y_2 + 1)^2$

      $= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1x_2y_1y_2 + x_1^2y_1^2 + x_2^2y_2^2$

      $\varphi(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$
  - Radial-basis functions: $K(\mathbf{x}^t, \mathbf{x}) = \exp[-\|\mathbf{x}^t - \mathbf{x}\|^2/\sigma^2]$
  - Sigmoid functions: $K(\mathbf{x}^t, \mathbf{x}) = \tanh(2\mathbf{x}^T\mathbf{x}^t + 1)$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

# SVM for Regression ※ $r^t$가 연속적

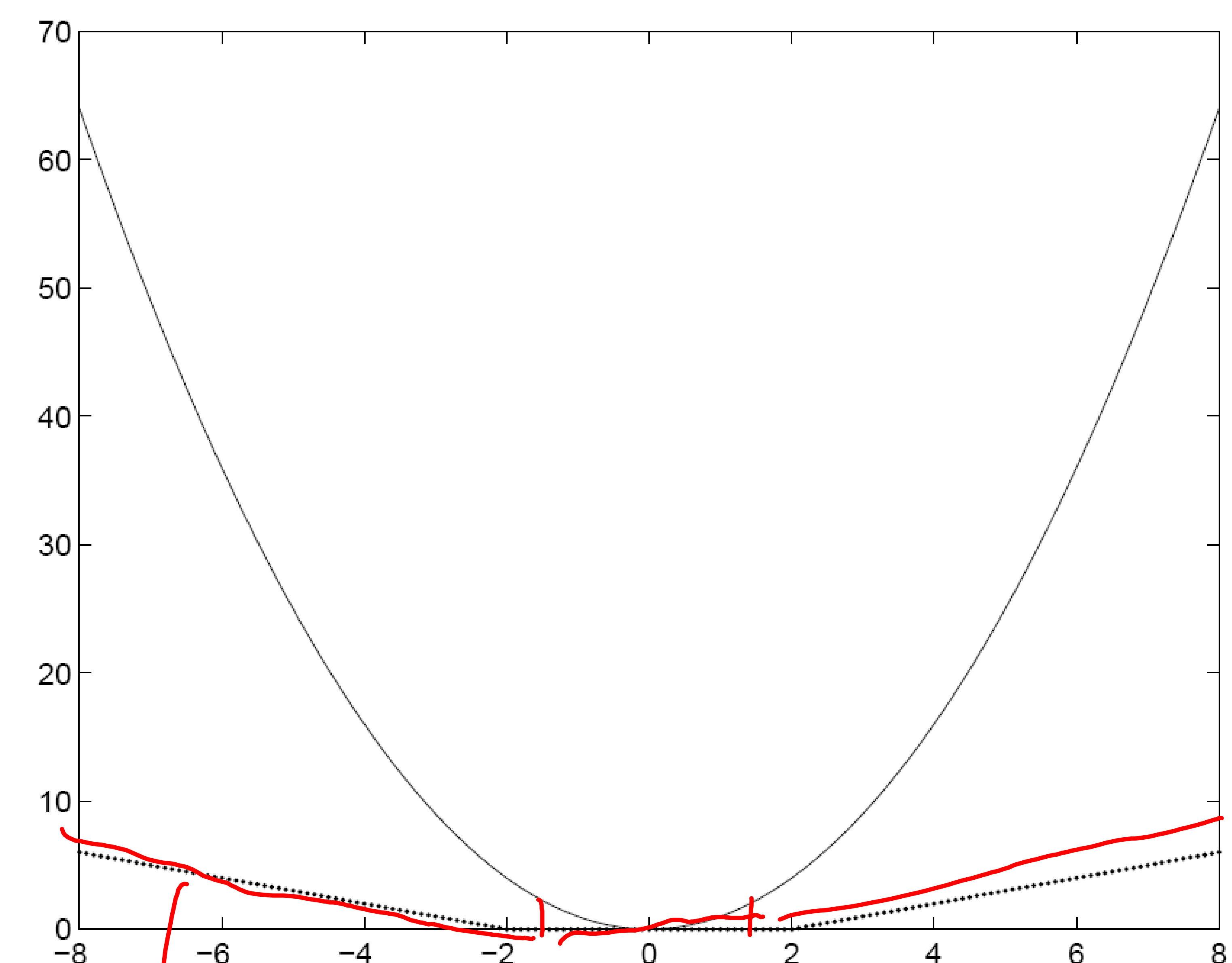
- Assume a linear model (possibly kernelized)
  - $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$

그리고 예전 에러 함수를 쓰면

이걸 안쓰면 formulation이 어려움

- Use ε-sensitive error function (instead of squared error function)

$$Err(r^t, f(\mathbf{x}^t)) = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| < \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$




Err가 선형증가

Err=0

- Problem formulation:

$$\min \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t (\xi_+^t + \xi_-^t) \right\}$$

subject to

$$r^t - (\mathbf{w}^T\mathbf{x} + w_0) \le \epsilon + \xi_+^t$$
$$(\mathbf{w}^T\mathbf{x} + w_0) - r^t \le \epsilon + \xi_-^t$$
$$\xi_+^t, \xi_-^t \ge 0$$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원