Statistical Machine Learning

1주차

담당: 15기 염윤석

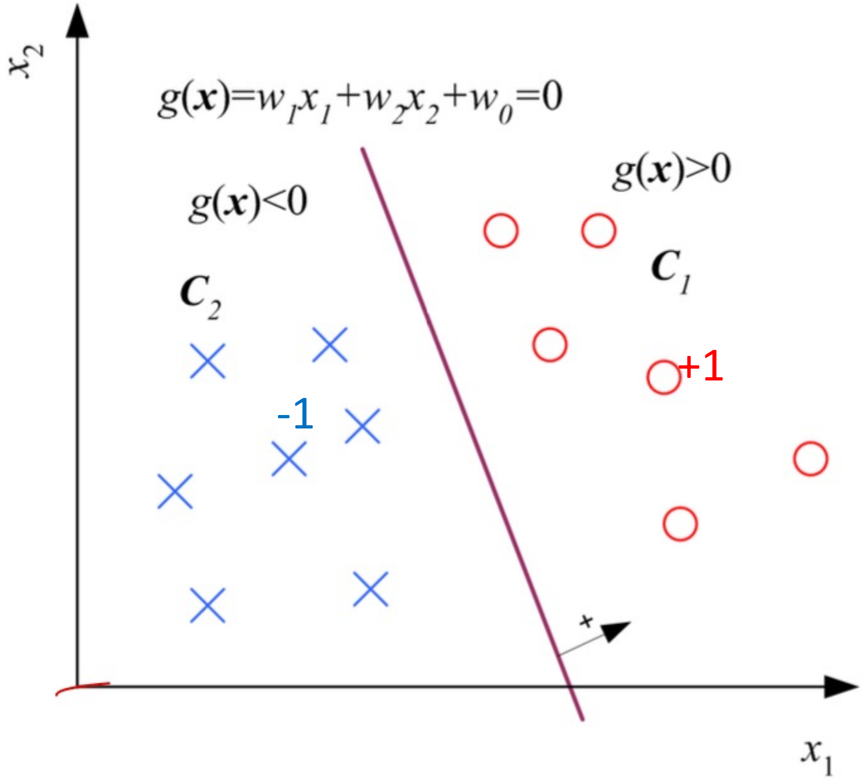
1. Linear SVM

2. Kernel SVM

3. SVM-Regression

1. Linear SVM - Classification

Linear Discriminant



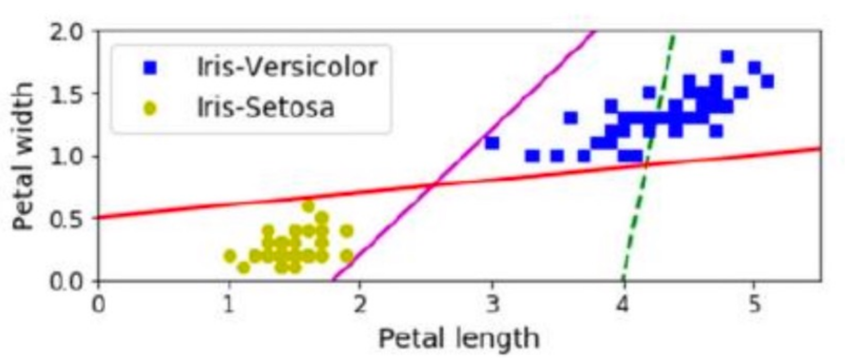
Decision Boundary or separating hyperplane

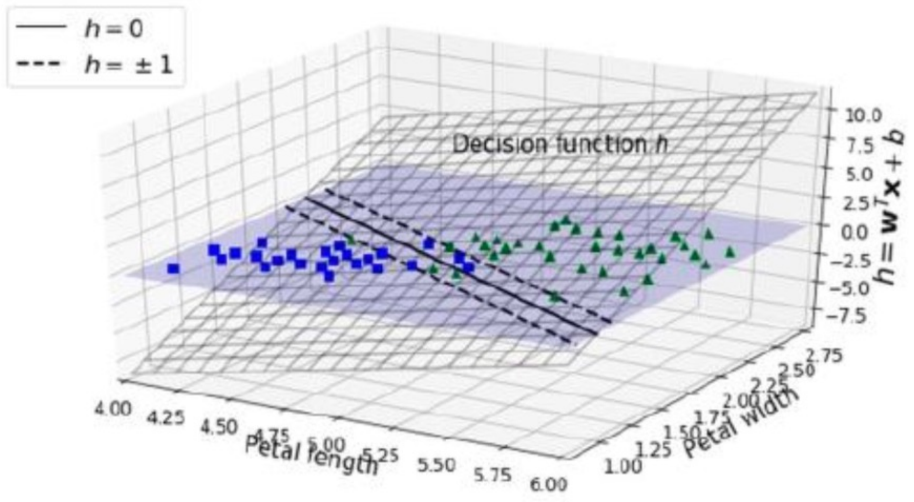
Decision Boundary : $g(x) = w^T x + w_0 = 0$

$X = \{x^t, r^t\} \mid r^t = \begin{cases} +1 \\ -1 \end{cases}$

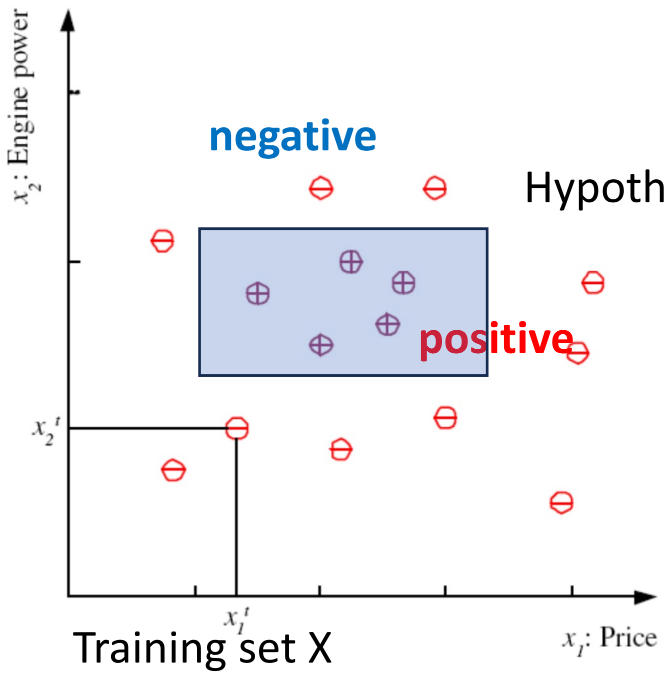
$$w^T x + w_0 \geq +1, for\ r^t = +1$$
$$w^T x + w_0 \leq -1\ , for\ \ r^t = -1$$

Hyperplane


Iris-Versicolor
Iris-Setosa
Petal width
Petal length


$h = 0$
$h = \pm 1$
Decision function h
$h = \mathbf{w}^T\mathbf{x} + b$
Petal length
Petal width

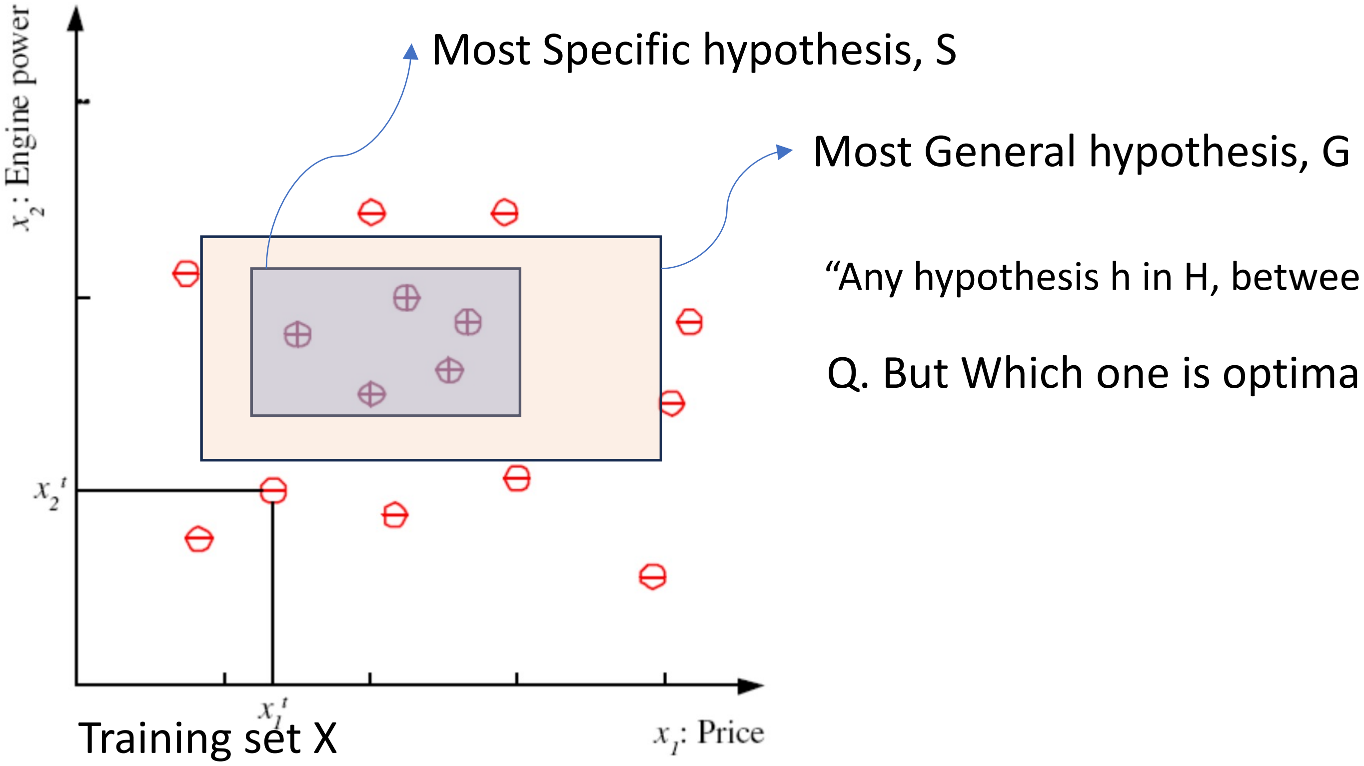
S, G and the Version Space



Hypothesis H = Assumption = Model

Training set X

S, G and the Version Space

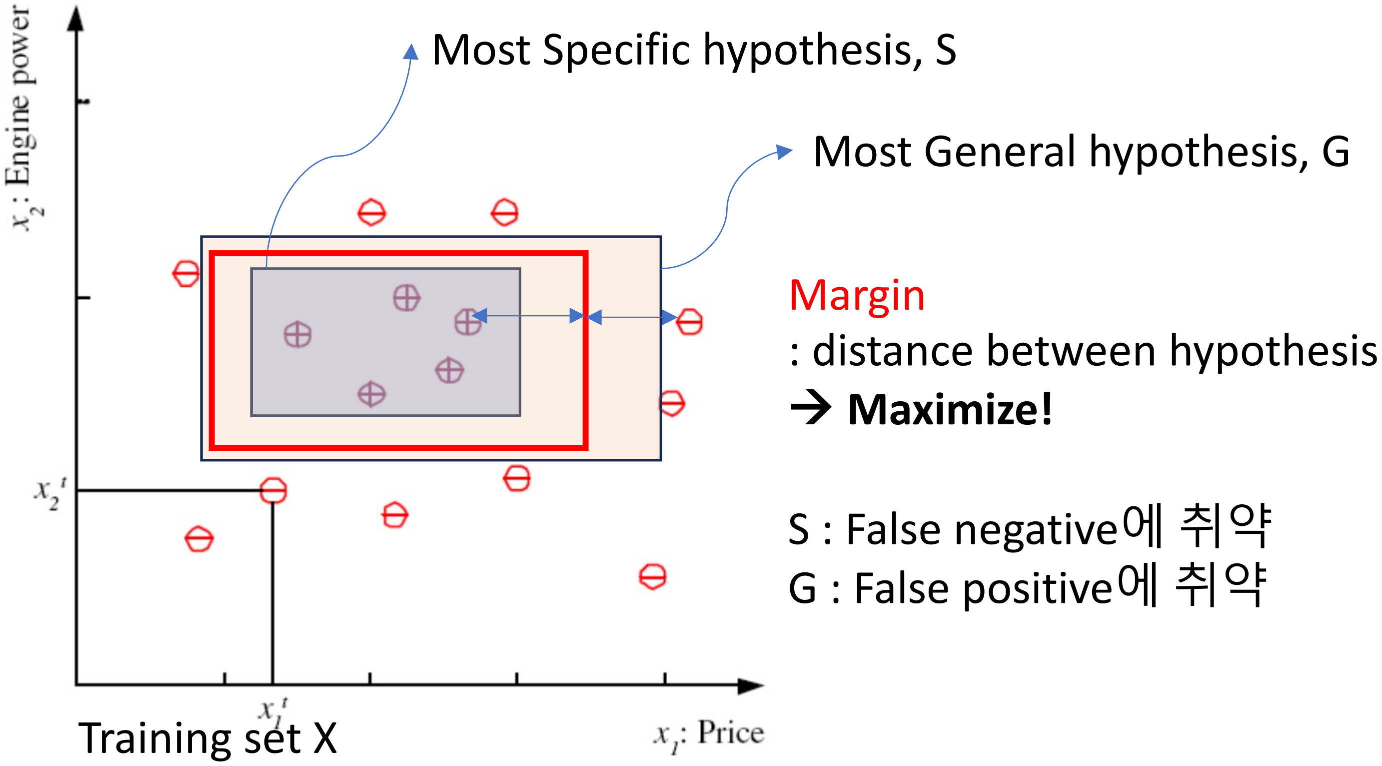


Training set X

"Any hypothesis h in H, between S & G is consistent and make up the Version space"

Q. But Which one is optimal?

Margin



Training set X

Margin
: distance between hypothesis and the closest positive and negative instances
→ **Maximize!**

S : False negative에 취약
G : False positive에 취약

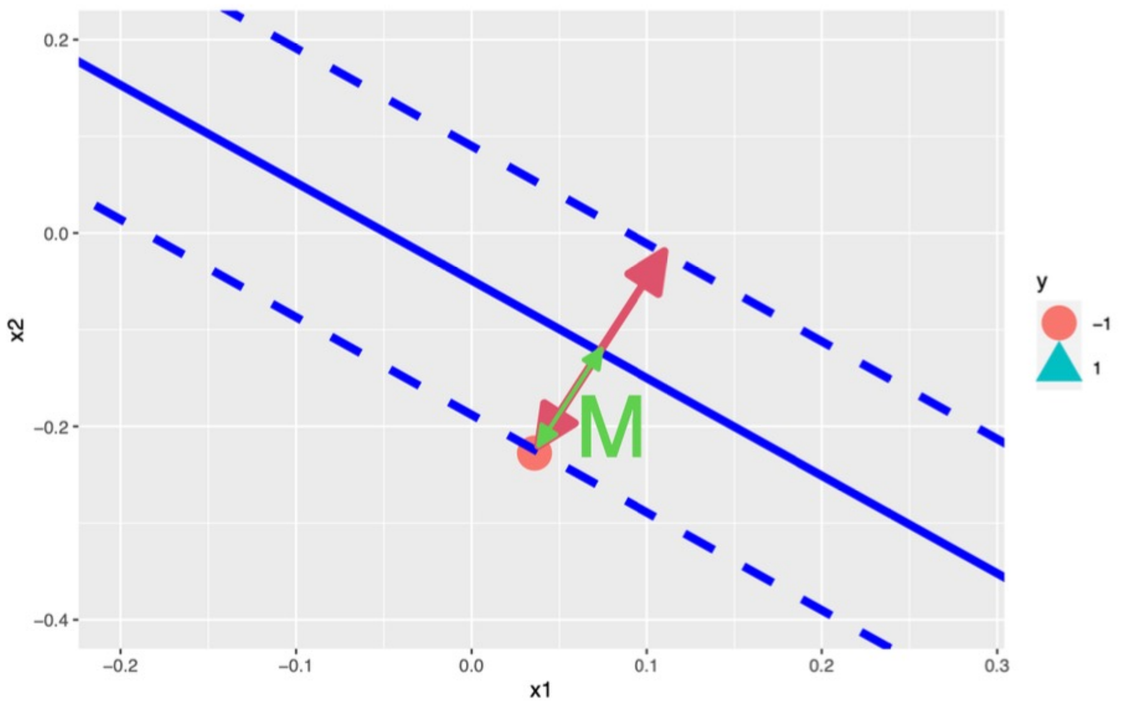
Optimal Hyperplane

- Decision Boundary : $g(x) = w^T x + w_0 = 0$

- $X = \{x^t, r^t\} \mid r^t = \begin{cases} +1 \\ -1 \end{cases}$

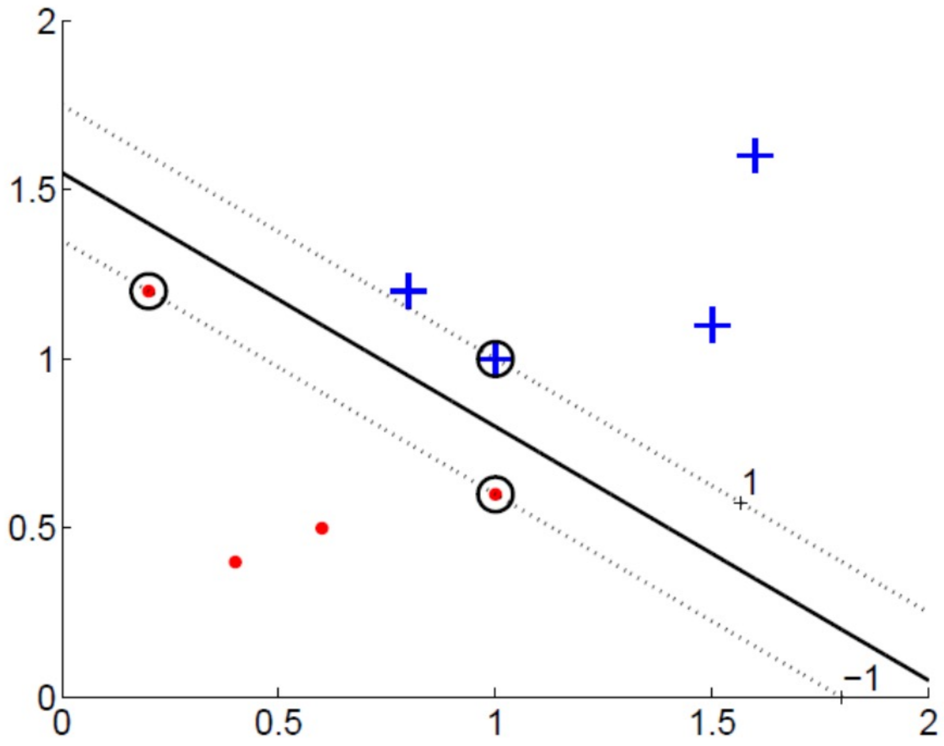
$\rightarrow r^t(w^T x + w_0) \geq +1$

[Margin]
Discriminant부터 양쪽 가장 가까운 instance 까지의 거리



Optimal Hyperplane(Discriminant) maximizes Margin

Objective of SVM



- Distance x to the hyperplane g(x)

- Margin

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

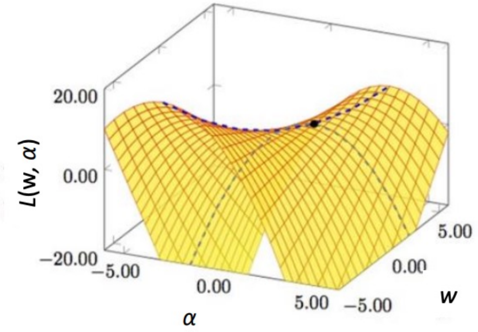
Lagrangian multiplier Method

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

Primal problem

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

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



KKT(Karush-Kuhn-Tucker Theorem)

1. Stationarity

2. Primal feasibility

3. Dual feasibility
4. Complementary slackness

Dual problem of SVM

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

Dual problem

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

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

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

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

KKT(Karush-Kuhn-Tucker Theorem)

1. Stationarity

2. Primal feasibility

3. Dual feasibility

4. Complementary slackness

Solution of SVM

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

We want optimal hyperplane $g(x) = w^T x + \ w_0$

We want optimal w^* & w_0^*

$$w = \sum_t \alpha^t r^t x^t \qquad\qquad w_0 = \frac{1}{N}\sum_t r^t \ - \ w^T x^t$$

$$g(x) = w_0 + \sum_t \alpha^t r^t x_t^T x$$

Solution of SVM

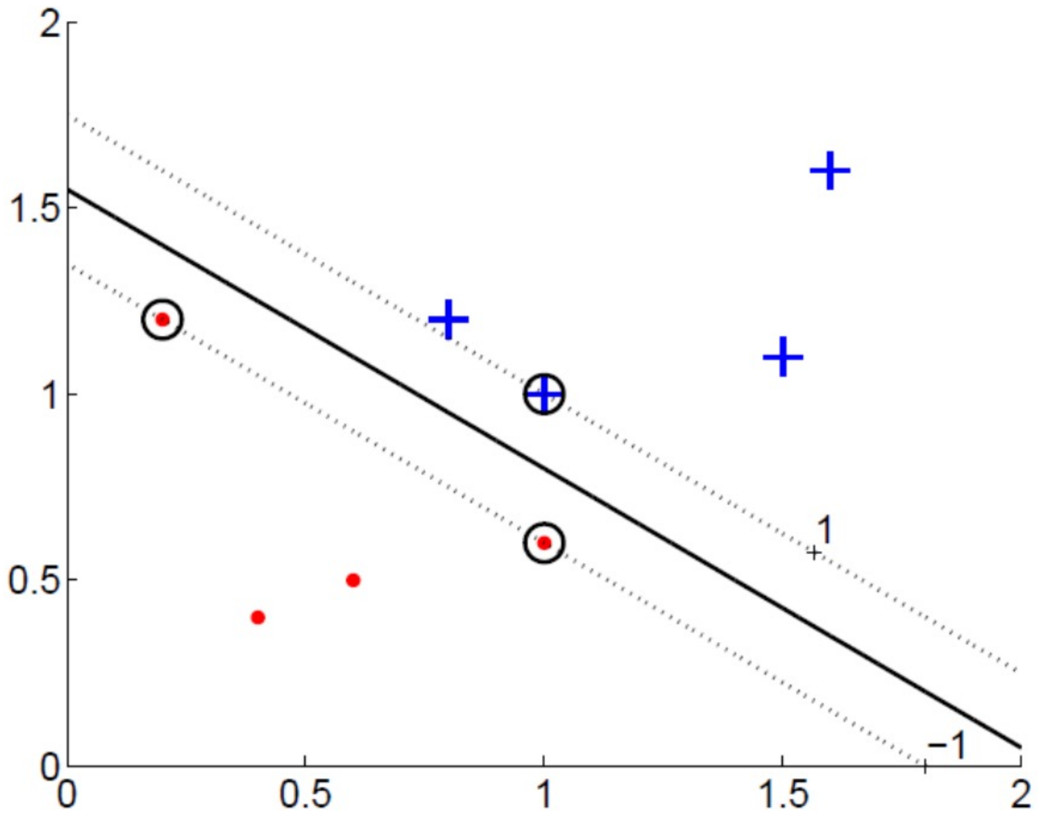
We want optimal hyperplane $g(x) = w^T x + w_0$

$$w = \sum_t \alpha^t r^t x^t \qquad w_0 = \frac{1}{N} \sum_t r^t - w^T x^t$$

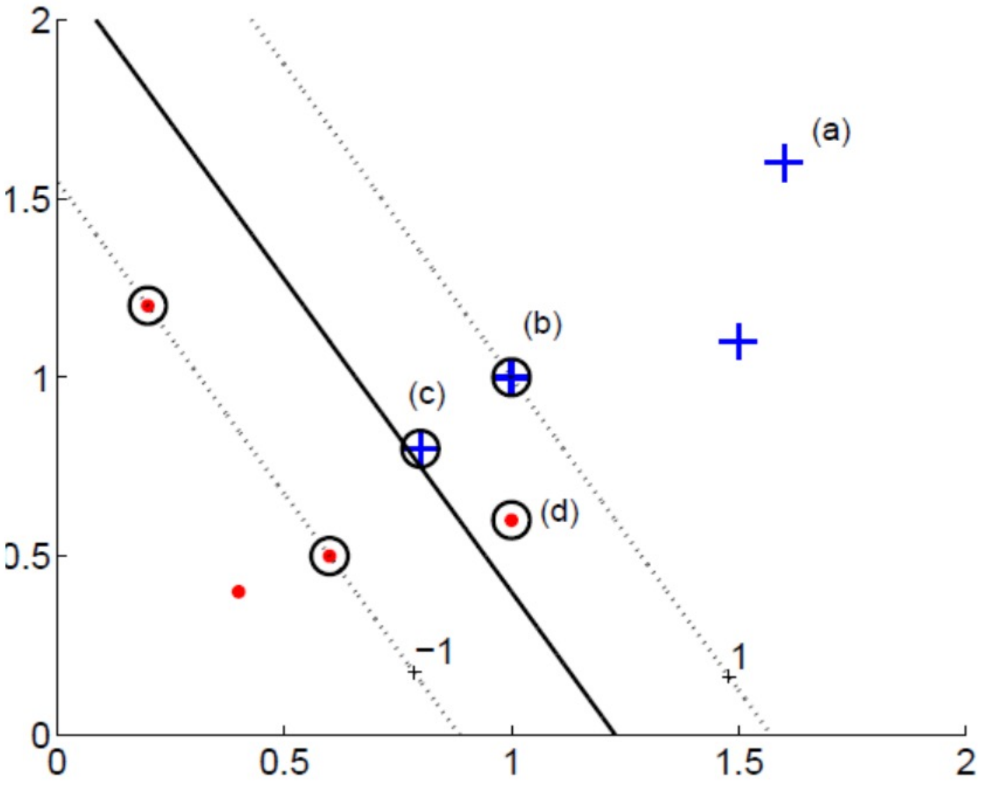
$$g(x) = w_0 + \sum_t \alpha^t r^t x_t^T x_t$$

"Most α^t = 0, only a small number have α^t>0" : **support vector**

SVM - Classification



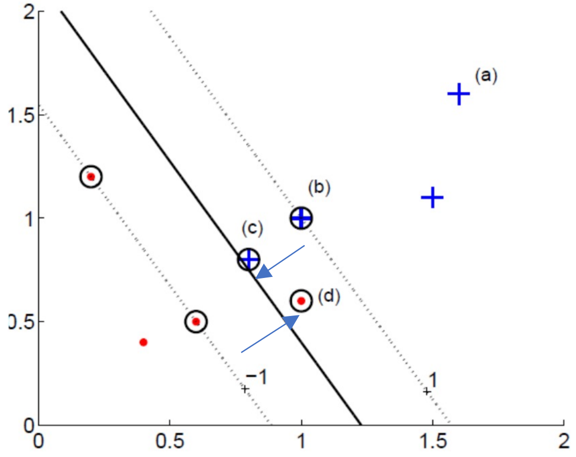
What if Non-Separable?



Soft Margin Hyperplane

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

Slack variable



- $soft\ error = \sum_t \xi^t$

$$\boxed{\min \frac{1}{2}\|w\|^2 + C\sum_t \xi^t \ \ subject\ to\ \ r^t(w^T x + w_0) \geq 1 - \xi^t\ , \xi^t \geq 0}$$

- New primal problem

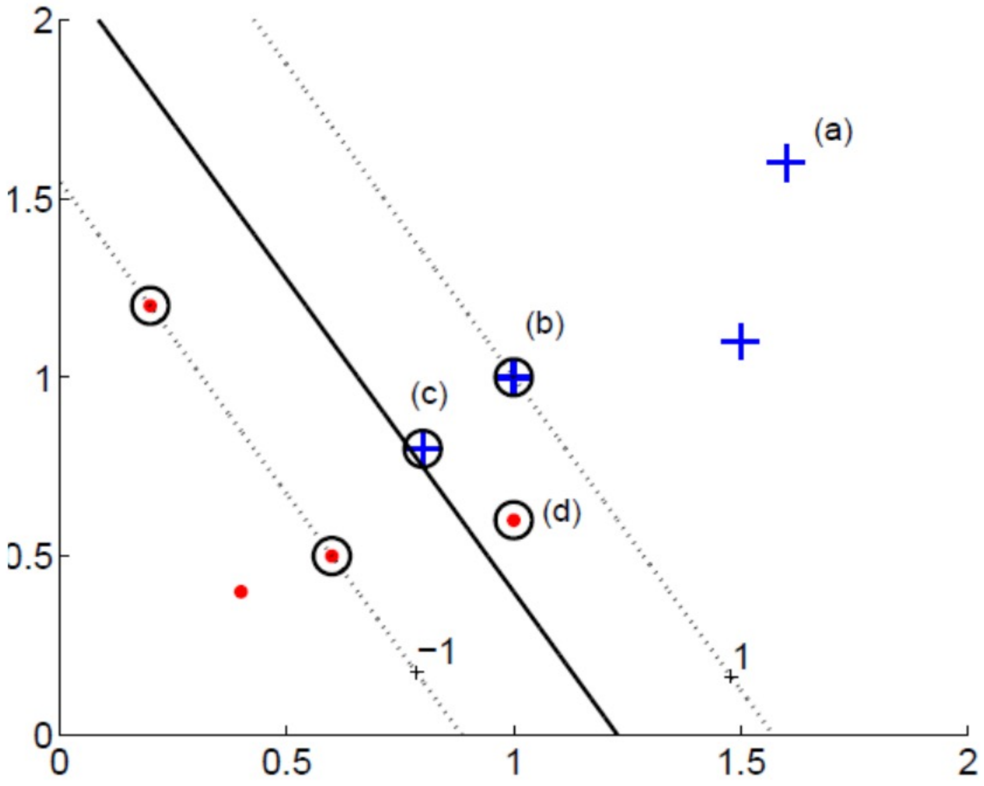
$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

- New Dual problem

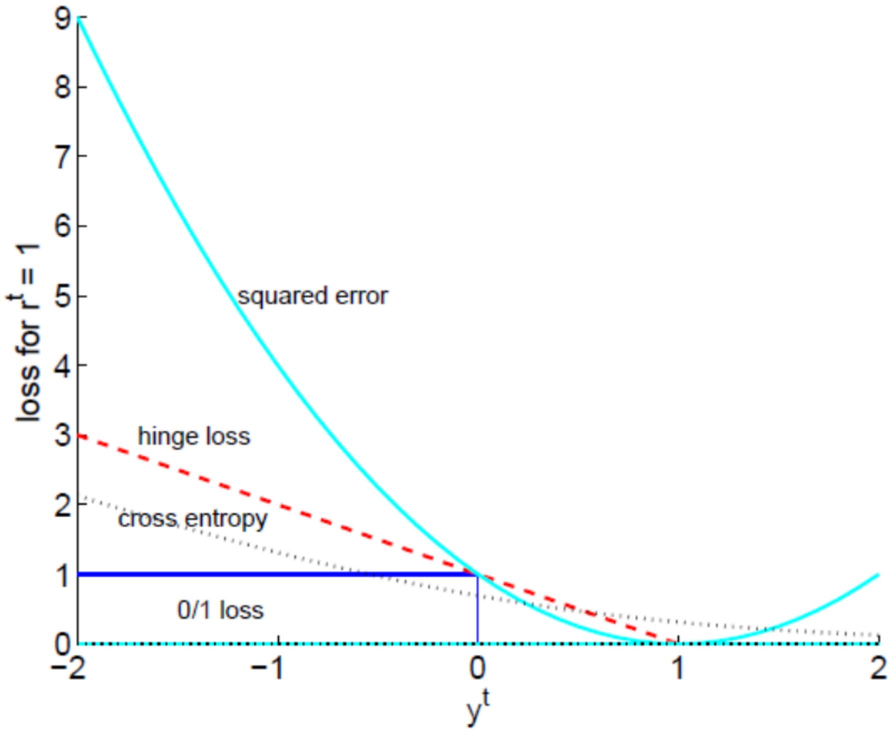
$$L_d(\alpha) = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s x_t^T x^s$$

$$subject\ to\ 0 \leq \alpha^t \leq C, \sum_t \alpha^t r^t = 0$$

Soft Margin Hyperplane



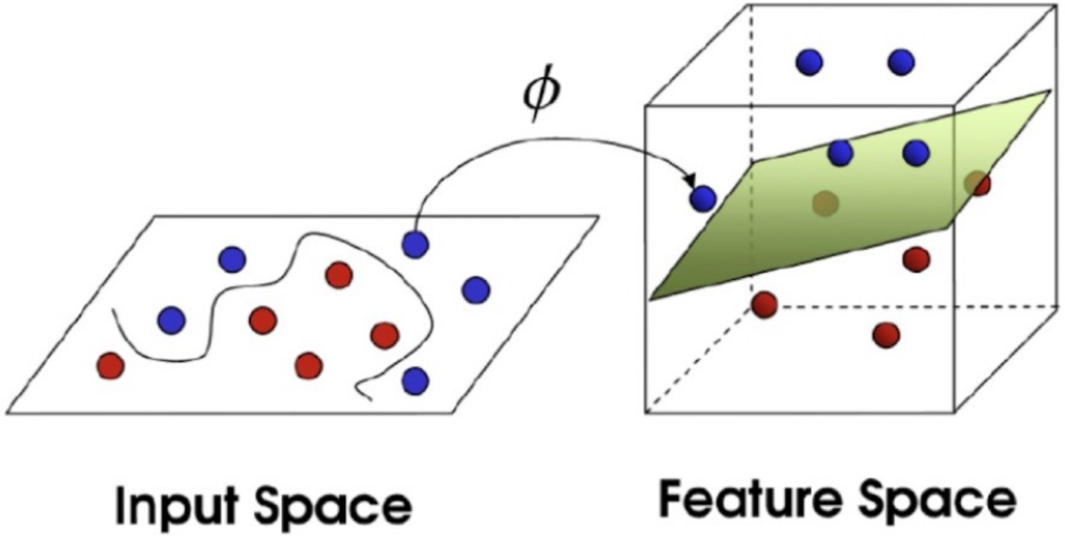
Hinge Loss



$$: \begin{cases} 0 & \text{if } y^t r^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$

2. Kernel SVM

Extension to non-linearity



$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \cdots, \phi_n(\mathbf{x}))$$

$$x = \{x_1, x_2\} \rightarrow z = \{1, \sqrt{2}x_1, \sqrt{2x_2}, \sqrt{2x_1x_2}, x_1^2, x_2^2\}$$

$$z = \varphi(x)$$

Feature mapping

Kernel Trick

$$z = \{1, \sqrt{2}x_1, \sqrt{2x_2}, \sqrt{2x_1x_2}, x_1^2, x_2^2\} = [z_1\ z_2\ \ z_3\ \ z_4\ z_5\ \ z_6\ ]$$

$$g(z) = w^T z + \ w_0$$

$$z = \ \varphi(x)$$

$$g(x) = w^T \varphi(x) + \ w_0$$

In linear SVM…

$$g(x) = w_0 + \sum_t \alpha^t r^t x_t^T x$$

$\rightarrow$

New feature space

$$g(z) = w_0 + \sum_t \alpha^t r^t z_t^T z$$

$$g(x) = w_0 + \sum_t \alpha^t r^t \varphi(x^t)^T \varphi(x)$$

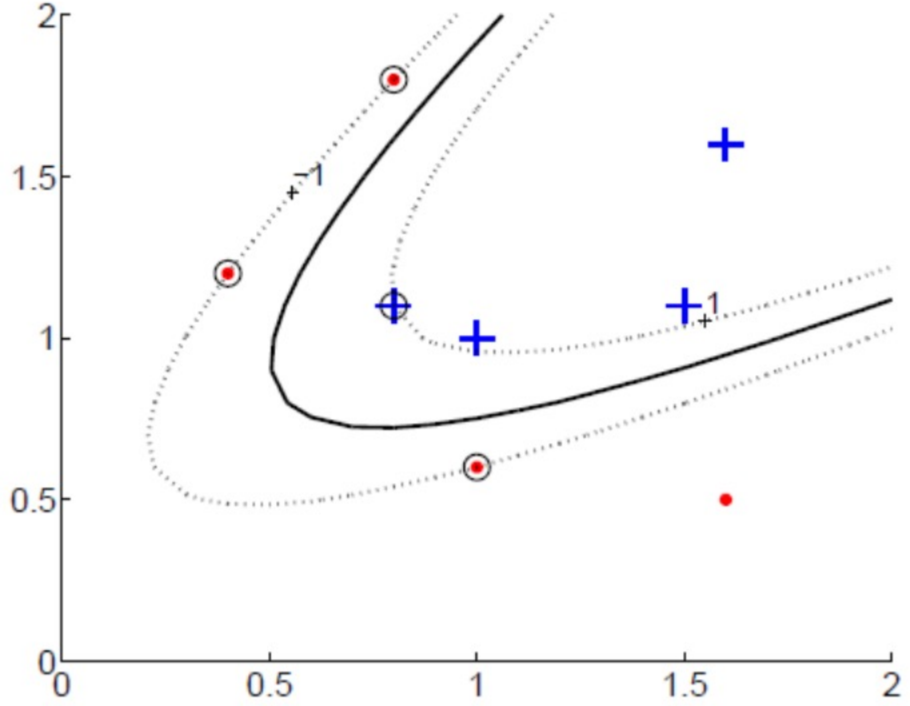
Using Kernel Trick : $K(x^t, x)$

Kernel Trick

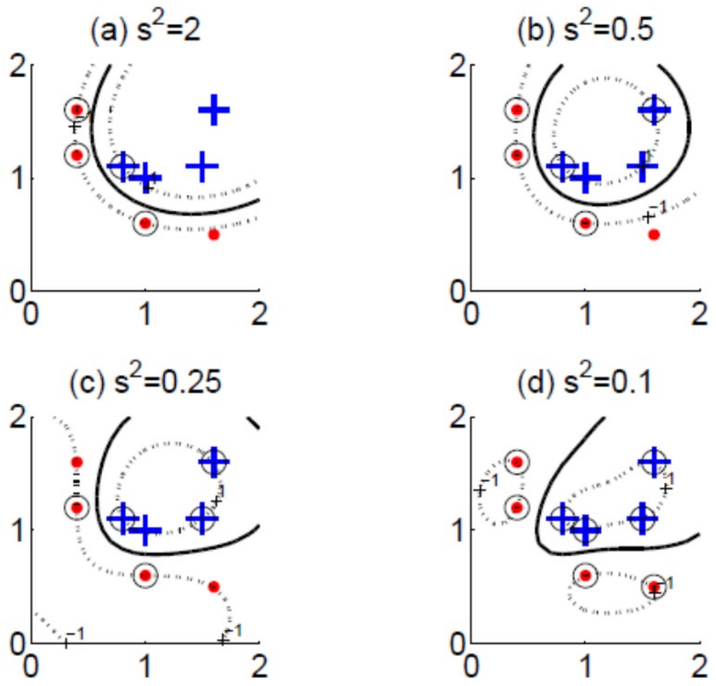
$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x})$$

$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$	*Linear Kernel*
$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j))$	*Gaussian Kernel (Radial Basis function)*
$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma + \gamma\, \mathbf{x}_i^T \mathbf{x}_j)^p$	*polynomial Kernel*
$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(k_1 \mathbf{x}_i^T \mathbf{x}_j + k_2)$	*Sigmoid Kernel*

Kernel SVM



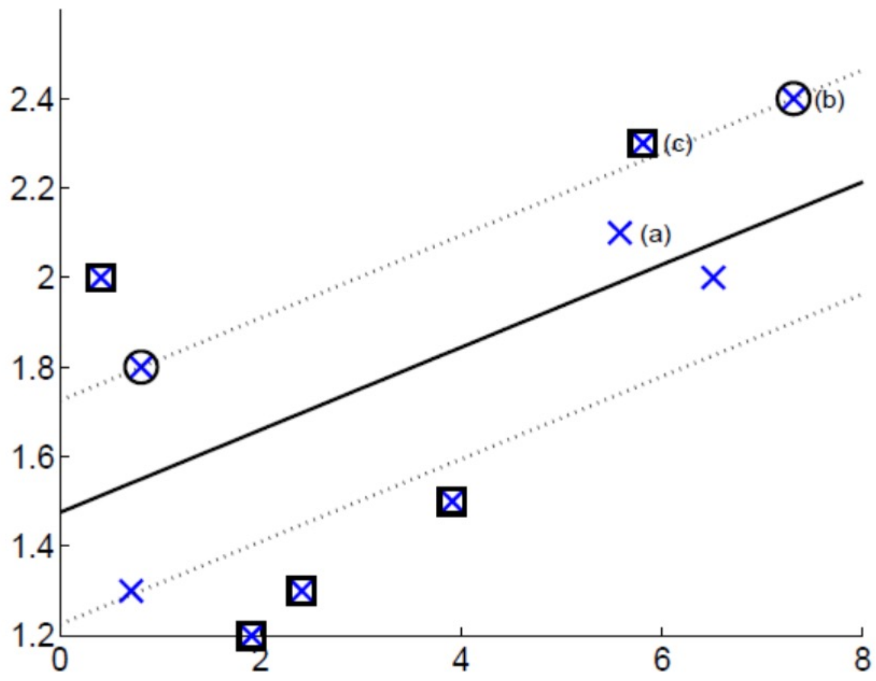
Polynomial Kernel



Gaussian(Radial-Basis function) Kernel

3. SVM - Regression

SVM- Regression



SVM- Regression

Let Assume linear model

$$f(x) = w^T x + w_0$$

- Error function(loss)

$$e = \begin{cases} 0 & if\ |r^t - f(x^t)| < \varepsilon \\ |r^t - f(x^t)| - \varepsilon \end{cases}$$

최대한 Margin 내로 들어오도록 학습 → Margin 밖에 있는 Error를 최소

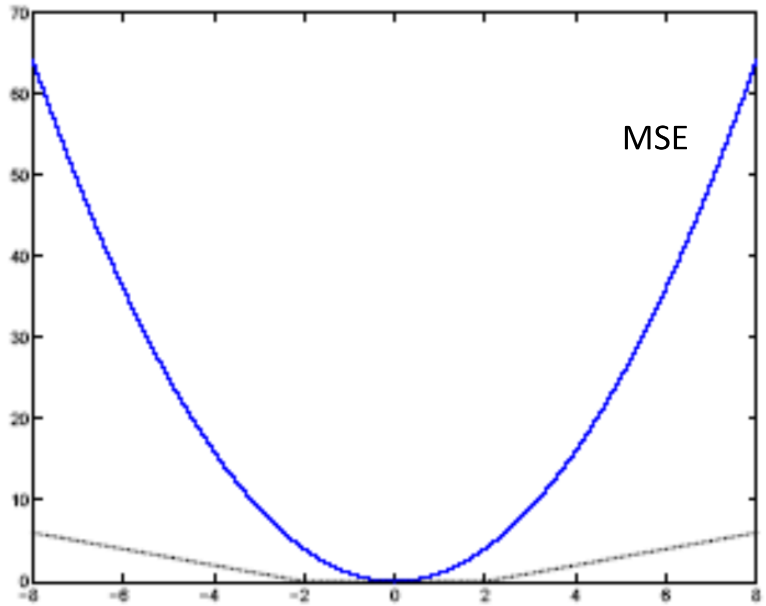
Lagragian Method

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

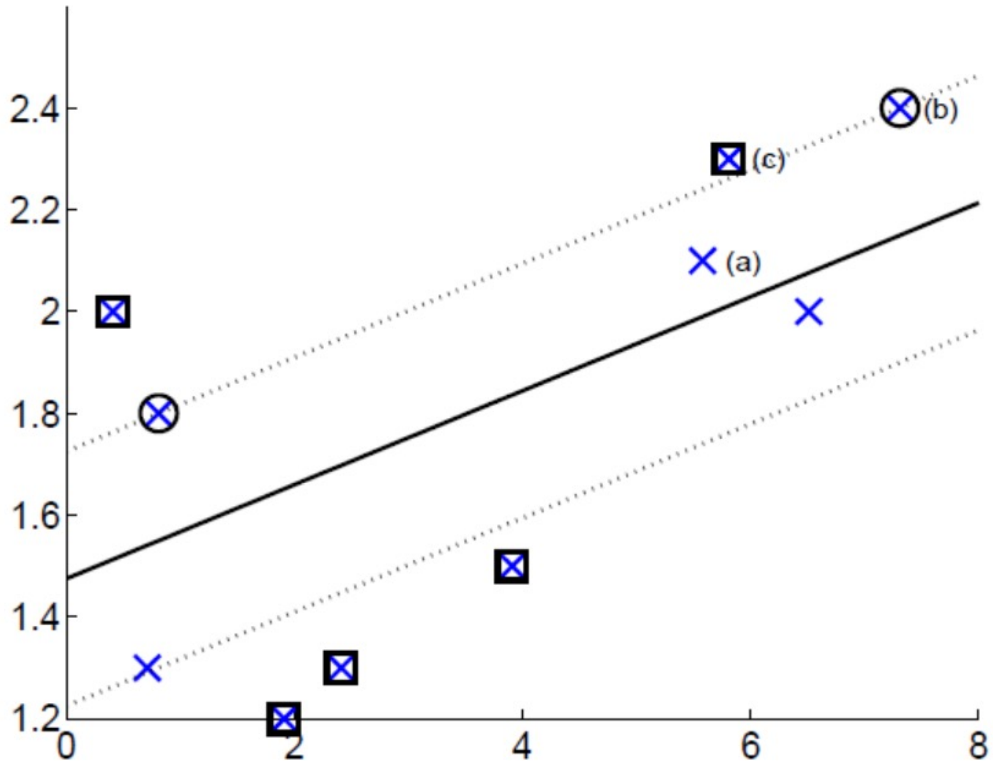
$$r^t - \left(\mathbf{w}^T\mathbf{x} + w_0\right) \le \varepsilon + \xi_+^t$$

$$\left(\mathbf{w}^T\mathbf{x} + w_0\right) - r^t \le \varepsilon + \xi_-^t$$

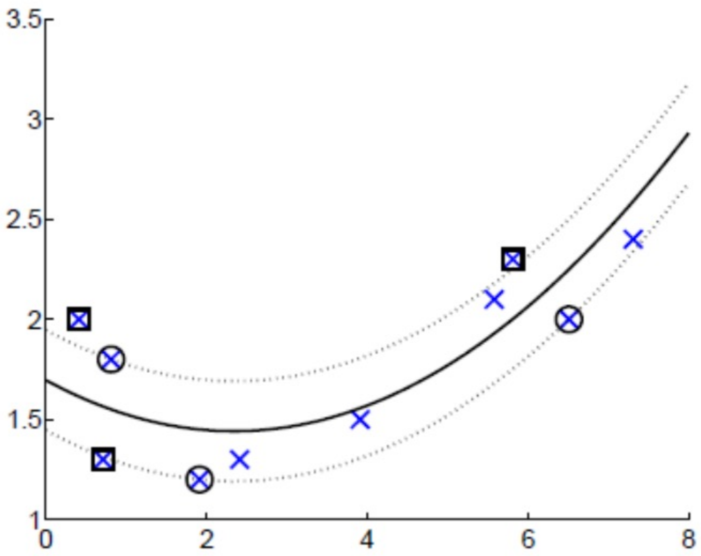
$$\xi_+^t, \xi_-^t \ge 0$$



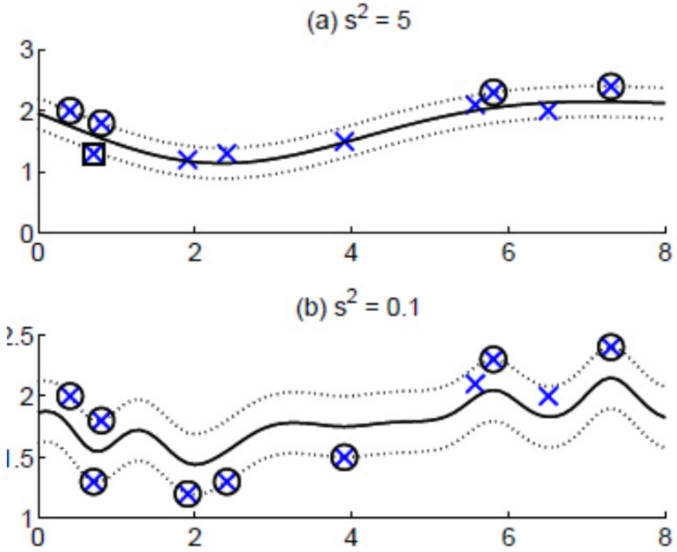
SVM- Regression



SVM Kernel Regression



Polynomial Kernel



Gaussian Kernel

SVM- Regression