Statistical Machine Learning

5주차

담당: 17기 이서연

1. Linear SVM

2. Kernel SVM

3. SVM-Regression

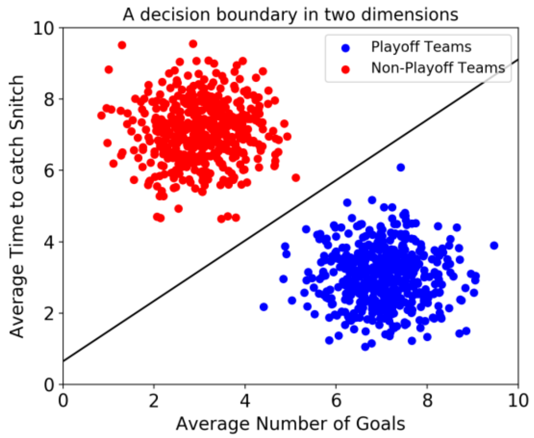
SVM(Support Vector Machine)

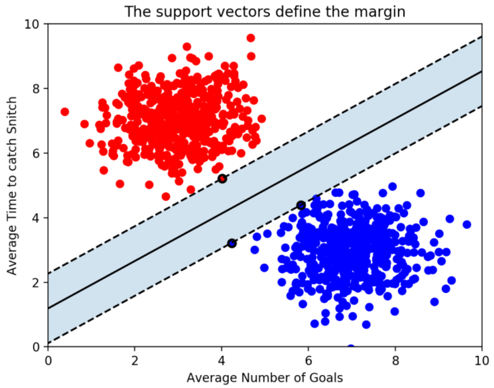
Q. What is SVM?

- very/ most powerful and versatile Machine Learning model

- linear / nonlinear classification, regression, outlier detection

-well suited for classification of complex but small/medium sized datasets





1. Linear SVM - Classification

Large margin classification

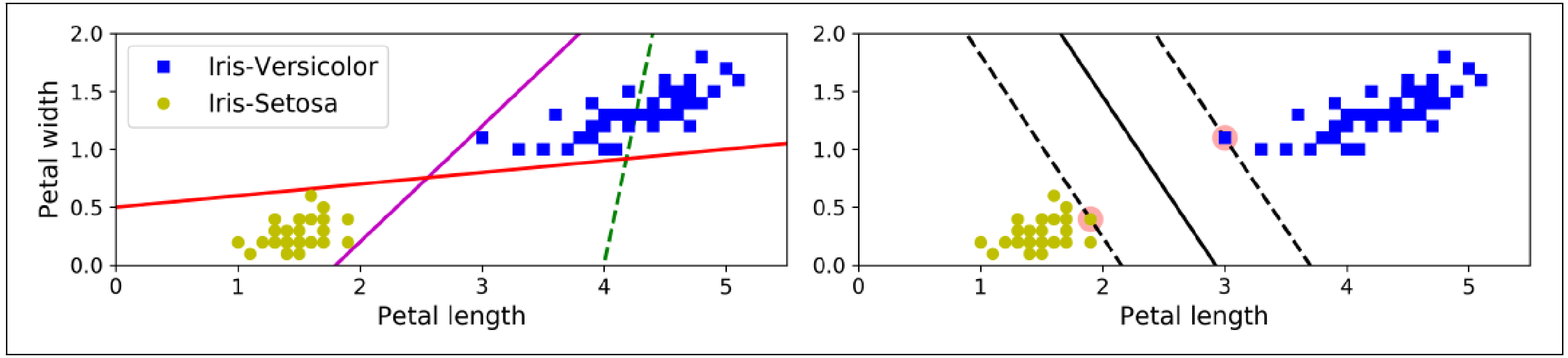


Figure 5-1. Large margin classification

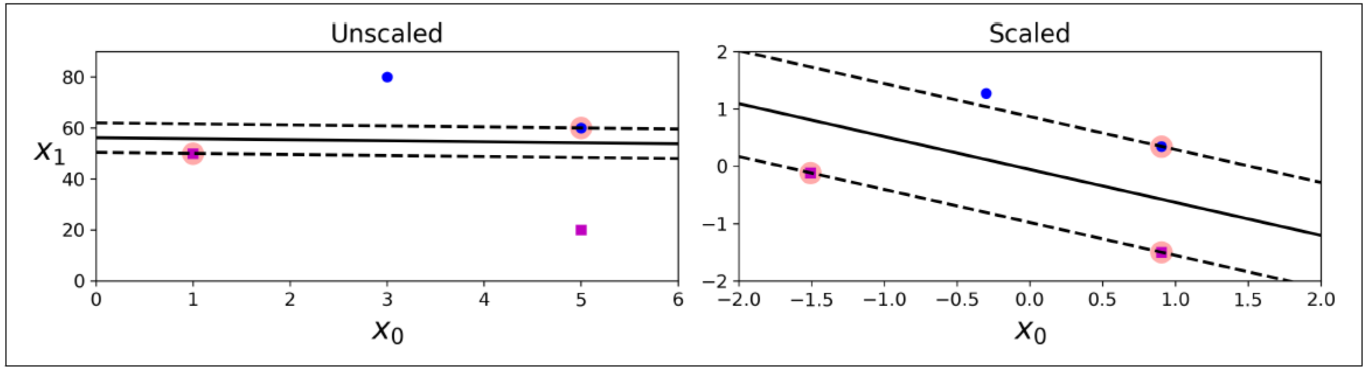
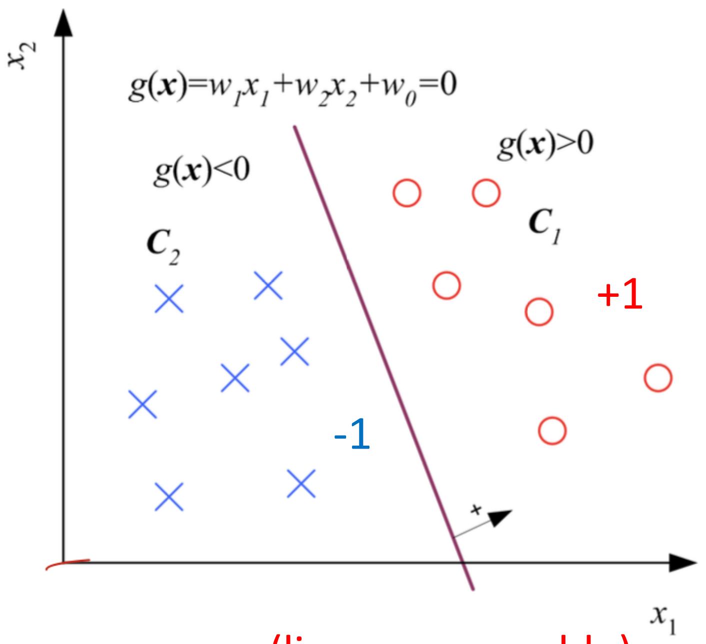


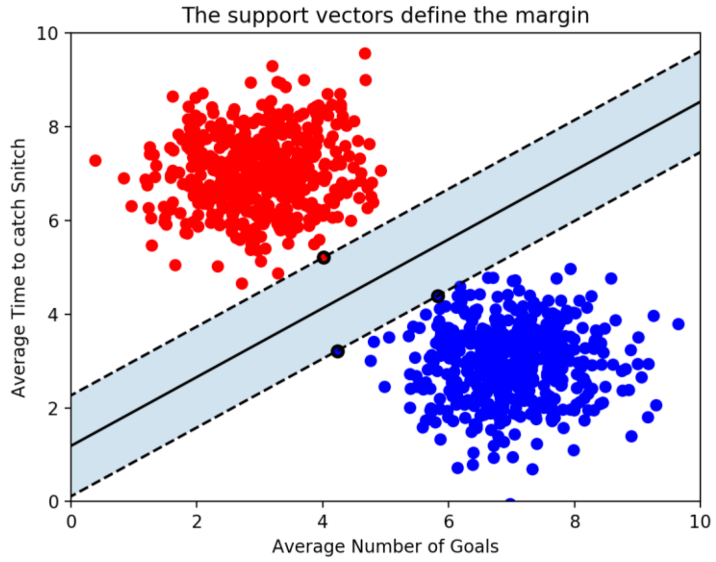
Figure 5-2. Sensitivity to feature scales

SVMs are sensitive to feature scales

Linear Discriminant



(lineary seperable)

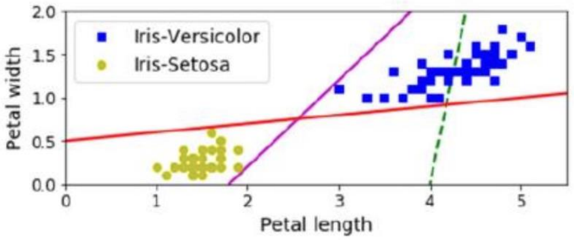


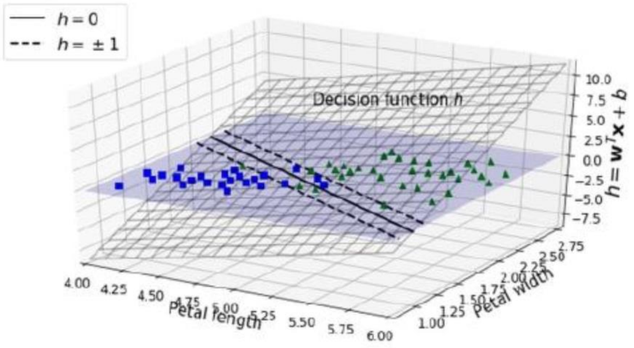
Decision Boundary or separating hyperplane

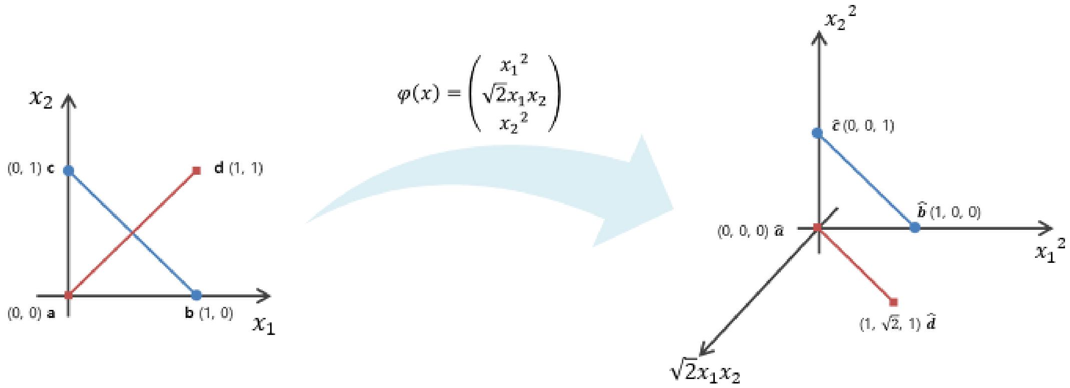
Decision Boundary : $g(x) = w^T x + w_0 = 0$

$$X = \{x^t, r^t\} \mid r^t = \begin{cases} +1 & w^T x + w_0 \geq +1, for\ r^t = +1 \\ -1 & w^T x + w_0 \leq -1, for\ r^t = -1 \end{cases}$$

Hyperplane

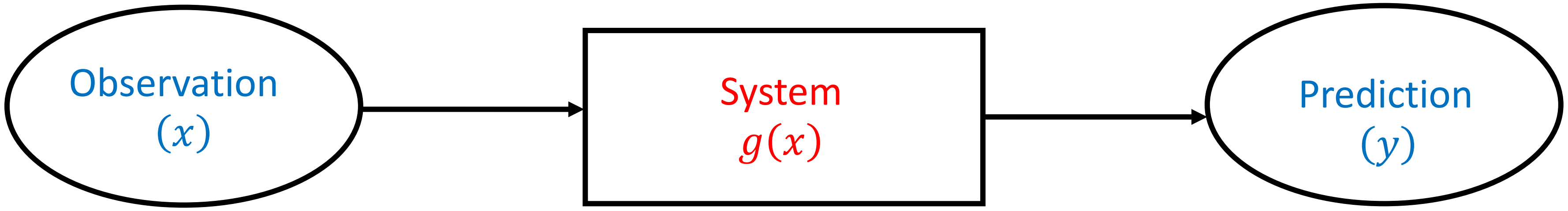






[그림 1]

Supervised Learning

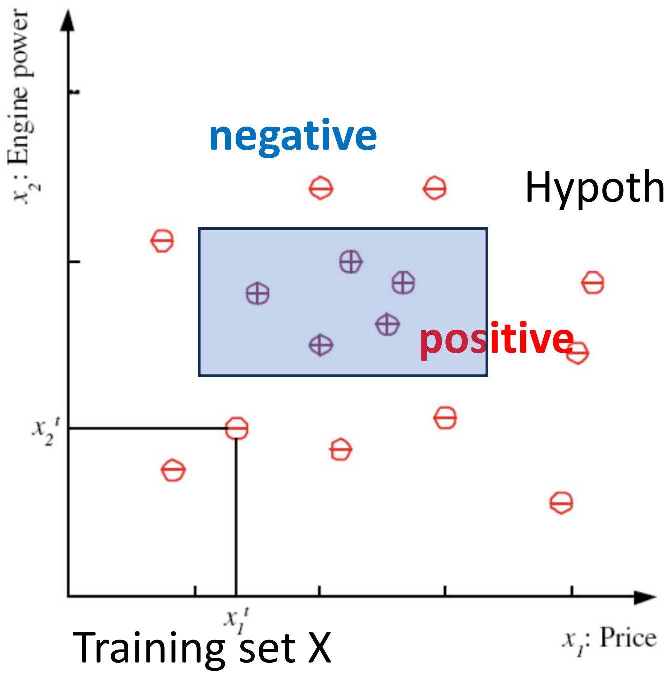


Given attributes of a car, we want to identify whether it is a family car or not

➔ Representation : price and engine power
➔ All other attributes (e.g., seating capacity, colour,,) are not under considerations

- X = $\begin{bmatrix} x_1 = price \\ x_2 = engine\ power \end{bmatrix}$ $\qquad y \in \{+, -\}$

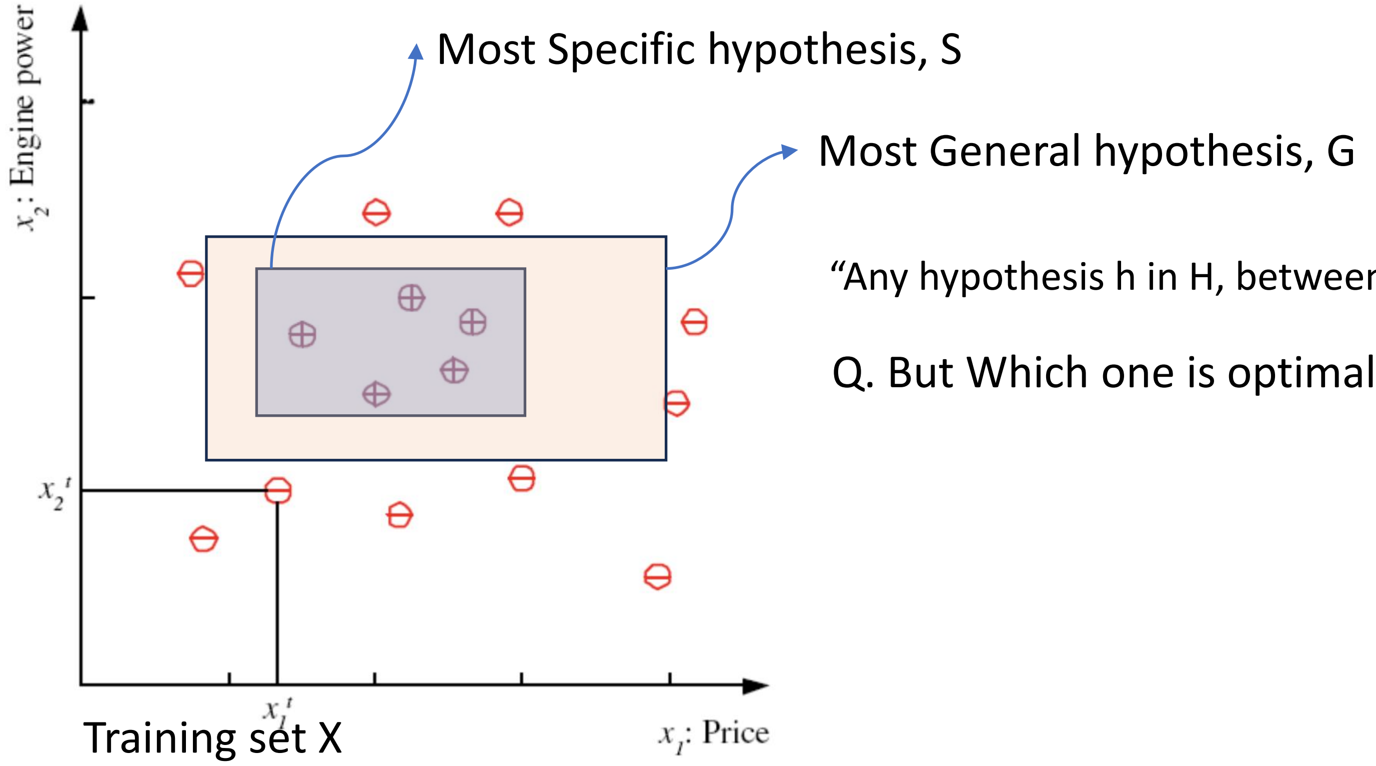
S, G and the Version Space



Training set X

Hypothesis H = Assumption = Model

S, G and the Version Space

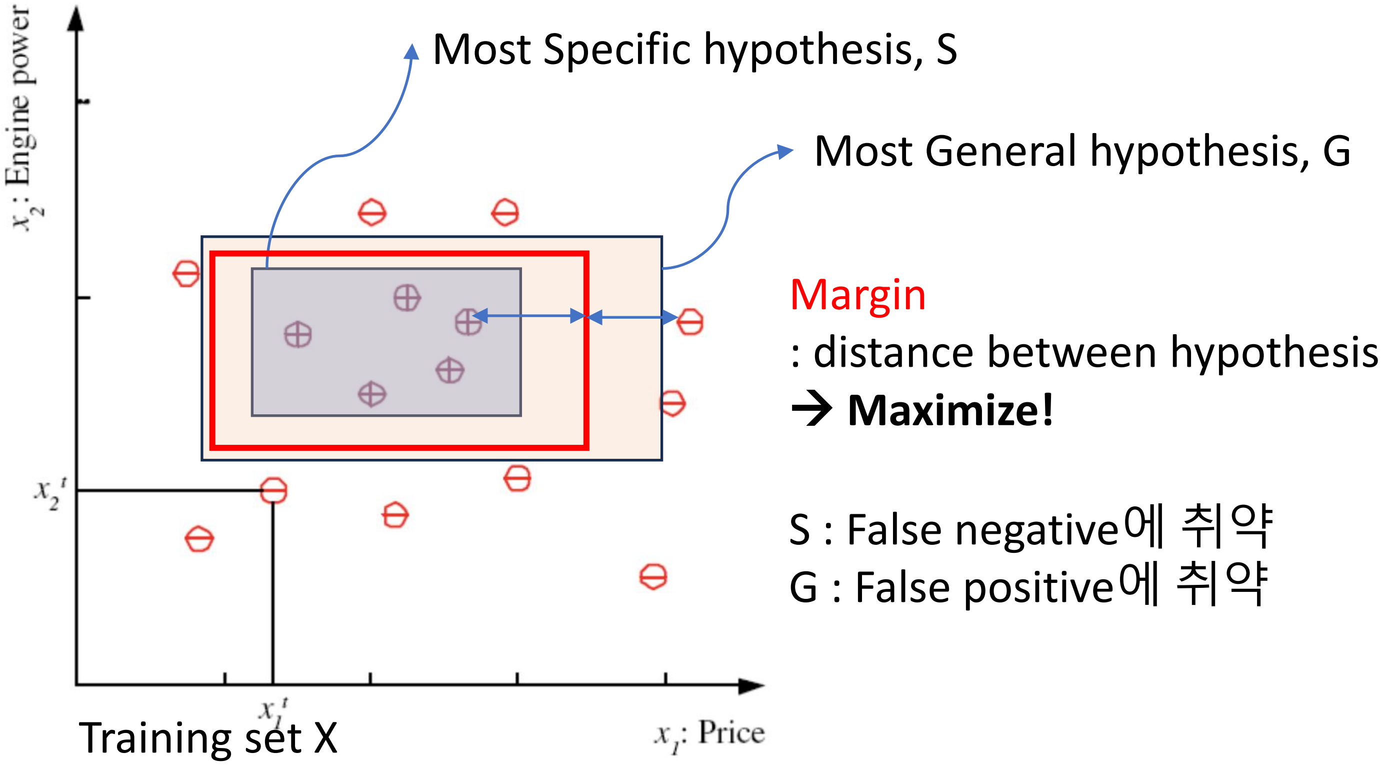


Training set X

"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!**

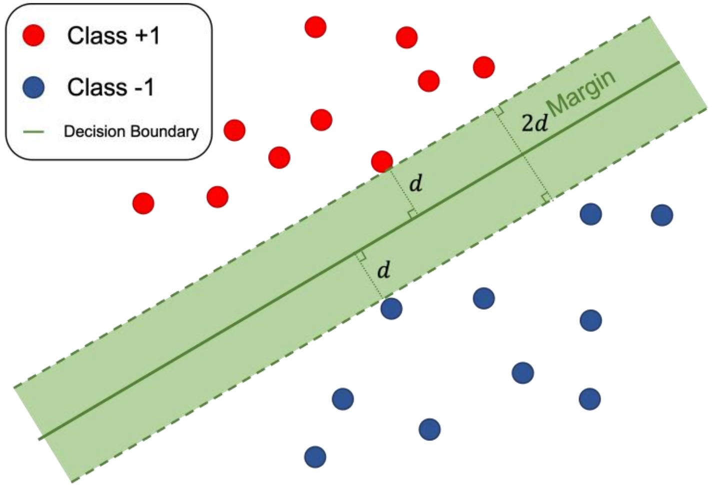
S : False negative에 취약
G : False positive에 취약

Optimal Hyperplane

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

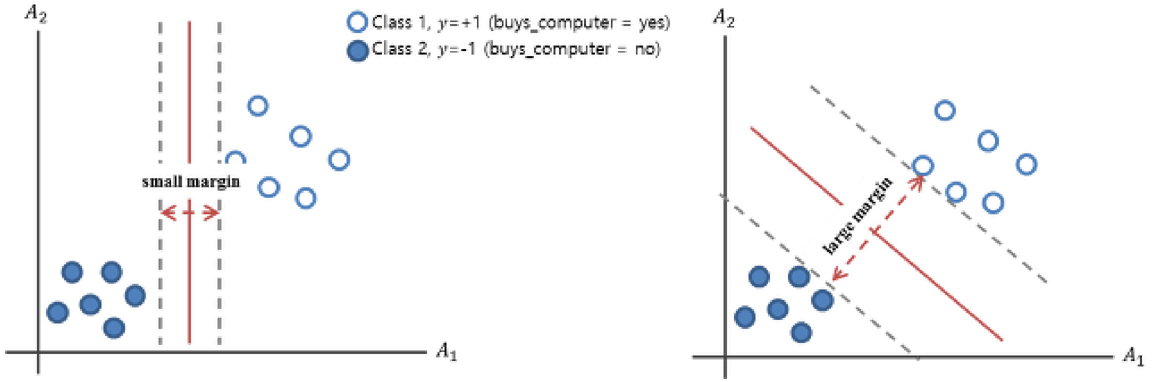
- $X = \{x^t, r^t\} \mid r^t = \begin{cases} +1 \\ -1 \end{cases}$

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

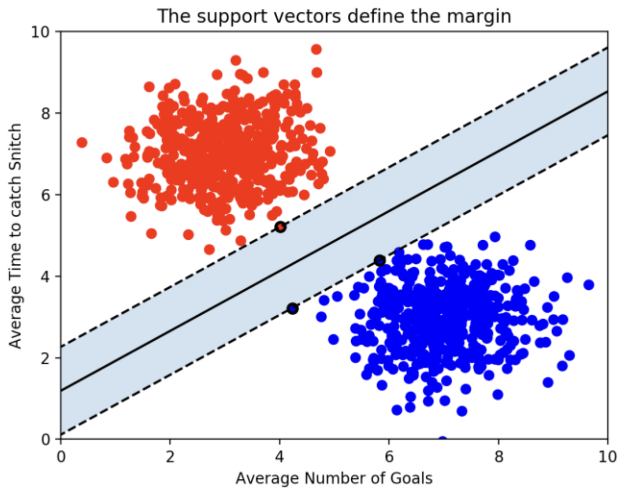


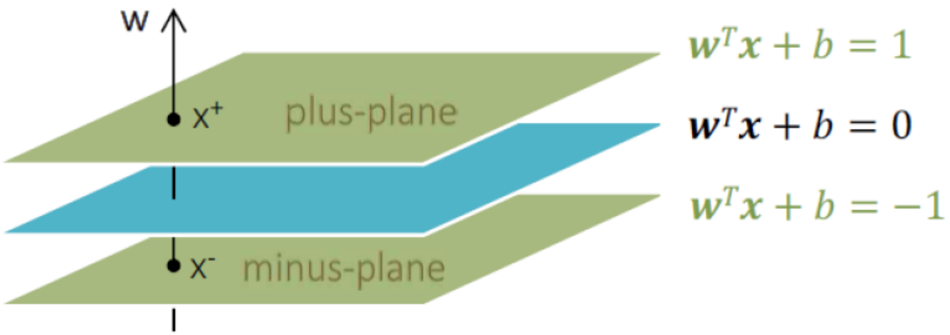
[Margin]
결정경계의 양의 방향과 음의 방향으로 d 만큼 떨어진 거리(영역)

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

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

* 라그랑주 승수법

SVM
- 목적함수 $\frac{1}{2}\|\vec{w}\|^2$
- 제약함수 $r^t(w^T x^t + w_0) \geq +1, \forall t$

· 기본 개념 : 제약 있는 최적화 문제에서 목적함수로 제약을 옮김으로서 제약이 없는 문제로 변환

ex) $\min f(x,y) = x^2+2y$ s.t. $3x+2y+1=0$ (등식 제약)

$$L(x,y,\alpha) = x^2+2y - \underbrace{\alpha}_{\text{라그랑주 승수}}(3x+2y+1)$$

→ 여기서 최적화 문제의 해라면, 변수 (x,y,α) 의 각 편도함수가 0 되는 지점을 찾으면 된다

$$\nabla f(x,y) + \alpha \nabla g(x,y) = 0$$

$$\left.\begin{aligned} \frac{\partial L}{\partial x} &= 2x - 3\alpha = 0 \\ \frac{\partial L}{\partial y} &= 2 - 2\alpha = 0 \\ \frac{\partial L}{\partial \alpha} &= -3x - 2y - 1 = 0 \end{aligned}\right) \Rightarrow \alpha = 1,\ x = \frac{3}{2},\ y = -\frac{11}{4}$$

Lagrangian multiplier Method

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

KKT(Karush-Kuhn-Tucker Theorem)

1. Stationarity

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial w_0} = 0$$

2. Primal feasibility

$r^t(w^T x^t + w_0) \geq 1$ → 제약식

3. Dual feasibility

$\alpha^t \geq 0 \quad t = 1, \ldots, n$

4. Complementary slackness

$\alpha^t = 0$ or $r^t(w^T x^t + w_0) - 1 = 0$ for all t

Primal problem

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

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

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

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

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

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

① 조건 적용. $\frac{\partial L}{\partial w} = w - \sum_{t=1}^{N} \alpha^t r^t x^t = 0 \quad \cdots \quad w = \sum_{t=1}^{N} \alpha^t r^t x^t$

$\frac{\partial L}{\partial w_0} = -\sum_{t=1}^{N} \alpha^t r^t = 0 \quad \therefore \sum_{t=1}^{N} \alpha^t r^t = 0$

$\rightarrow L_p = \frac{1}{2}\|w\|^2 - \sum_t \alpha^t (r^t(w^T x^t + w_0) - 1)$

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

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

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

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

α에 대한 식으로 간단해짐.
최고차항의 계수가 음수이므로 최솟값 문제에서 최대값 문제로 변환

Dual problem of SVM

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

Dual problem

- 원래(primal) 문제가 $f_i(\theta) \ge 0,\ i = 1,\dots,n$이라는 조건 하에 $J(\theta)$를 최소화 하는 문제라고 하면 쌍대(dual) 문제는 $\partial L(\theta,\alpha)/\partial\theta = 0$과 $\alpha_i \ge 0\ \ i = 1,\dots,n$이라는 두 가지 조건 하에 $L(\theta,\alpha) = J(\theta) - \sum_{i=1}^{n} \alpha_i f_i(\theta)$ 를 최대화 하는 문제로 표현할 수 있다.

Dual

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

subject to $\sum_t \alpha^t r^t = 0$, $\underbrace{\alpha^t \ge 0,\ \forall t}_{\text{KKT ③}}$ (제약식)

KKT(Karush-Kuhn-Tucker Theorem)

1. Stationarity $\quad \frac{\partial L}{\partial w} = \frac{\partial L}{\partial w_0} = 0$
2. Primal feasibility $\quad r^t(w^T x^t + w_0) \ge 1 \quad \rightarrow$ 제약식
3. Dual feasibility $\quad \alpha^t \ge 0 \quad t = 1, \dots, n$
4. Complementary slackness

$\alpha^t = 0$ or $r^t(w^T x^t + w_0) - 1 = 0$ for all t

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

☆ SVM의 해는? $w^Tx+w_0=0$에서 w_0 w_0 를 찾는 것이 목적이었고 이제 α만 알면 찾을 수 있다.

We want optimal w^* & w_0^*

KKT ① 조건: $\frac{dL}{dw}=0 \rightarrow w=\sum_{t=1}^{N}\alpha^t r^t x^t$ ∵ x, r data 값이므로 α만 알면 w 계산O

KKT ④ 조건: (∵ w식) $\underline{\alpha^t=0}$ or $r^t(w^Tx^t+w_0)-1=0$ for all t

이 경우는 w 생성에 영향X ∴ $r^t(w^Tx^t+w_0)-1=0$ 인 것들이 영향을 주는 것 = support vector

$r^t(w^Tx^t+w_0)-1=0$ 에서 x, r, w 값은 모두 알기 때문에 w_0 계산가능.

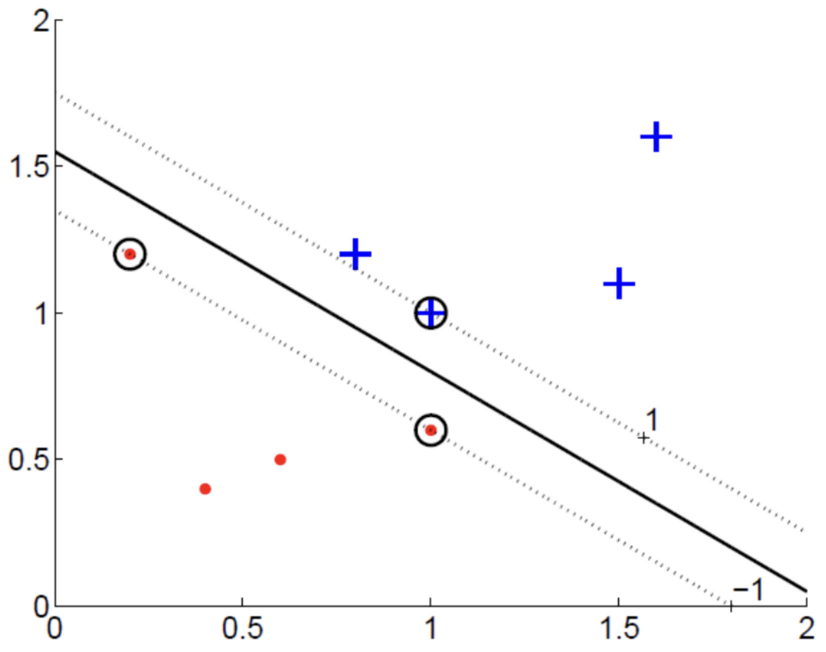
$$w = \sum_t \alpha^t r^t x^t$$

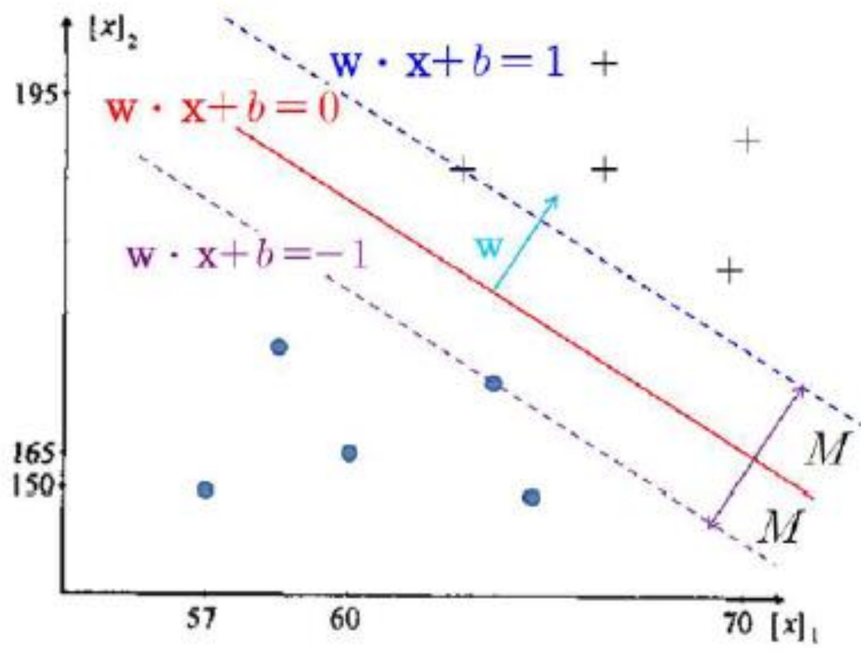
$$w_0 = \frac{1}{N}\sum_t r^t - w^T x^t$$

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

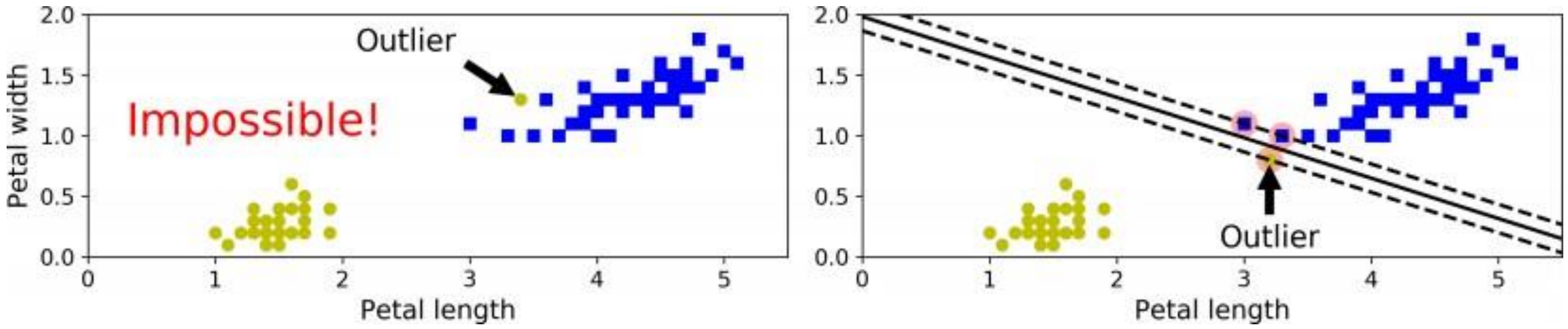
"Most α^t = 0, only a small number have α^t>0" : **support vector**

SVM - Classification





What if Non-Separable?



'Soft margin classification'

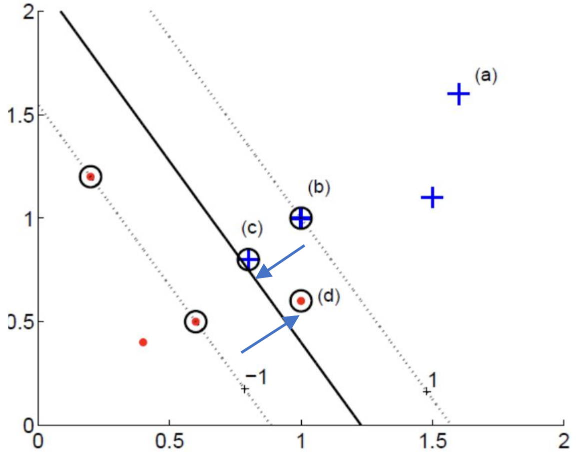
Find a good balance between keeping the street as large as possible vs limiting margin violations

*margin violations : instances that end up in the middle of the street or even on the wrong side

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

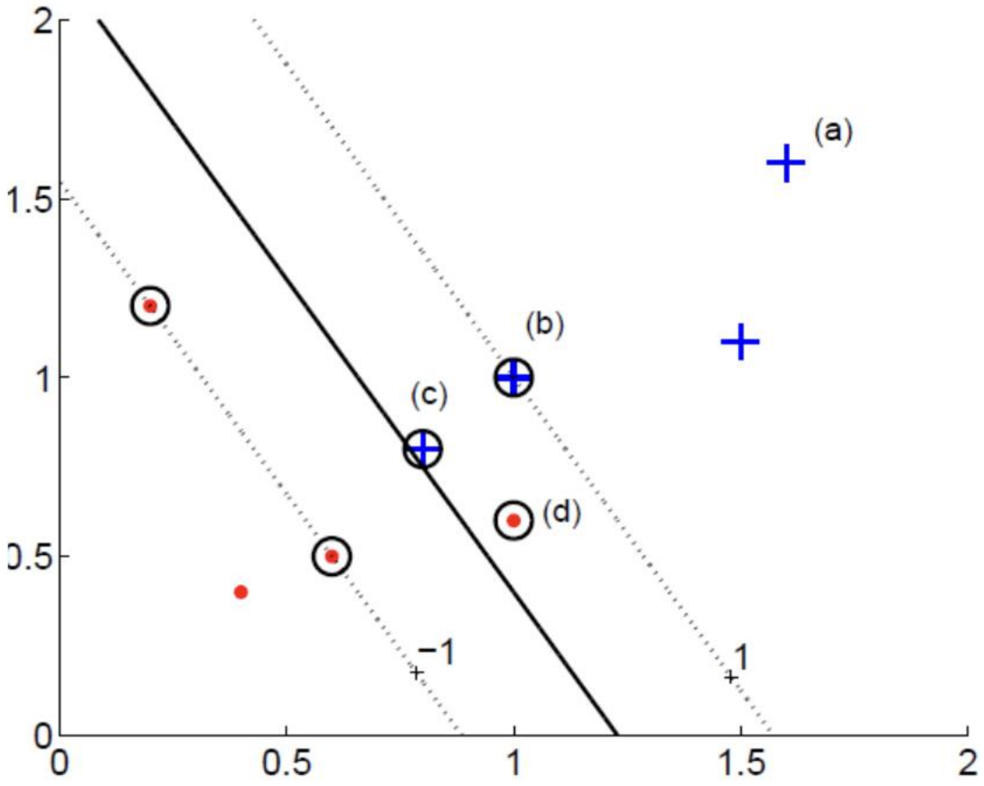
$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

- 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



Soft Margin Hyperplane

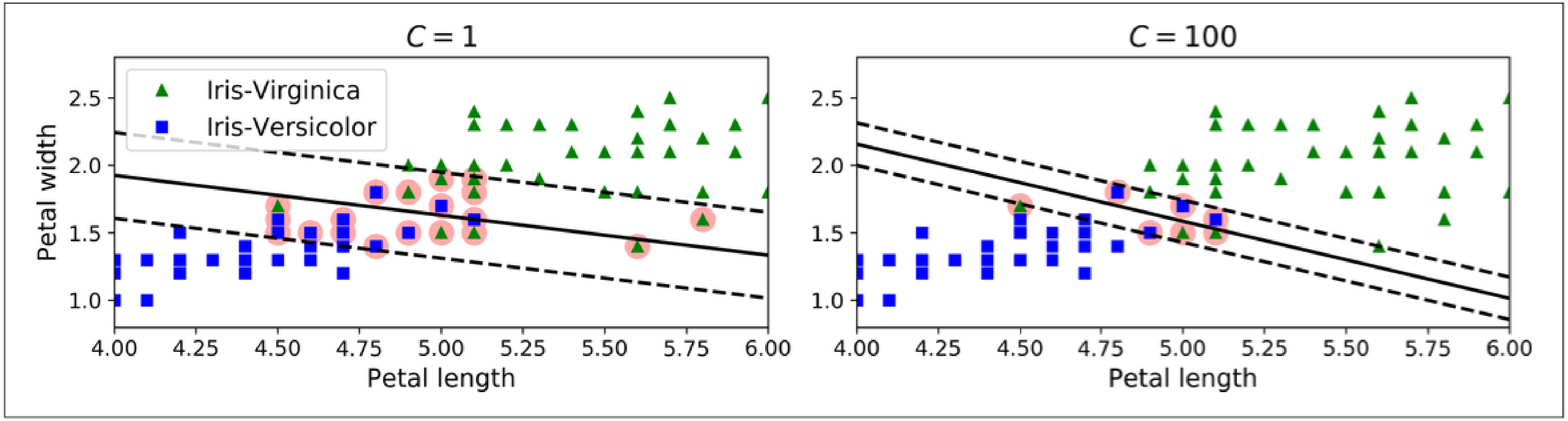
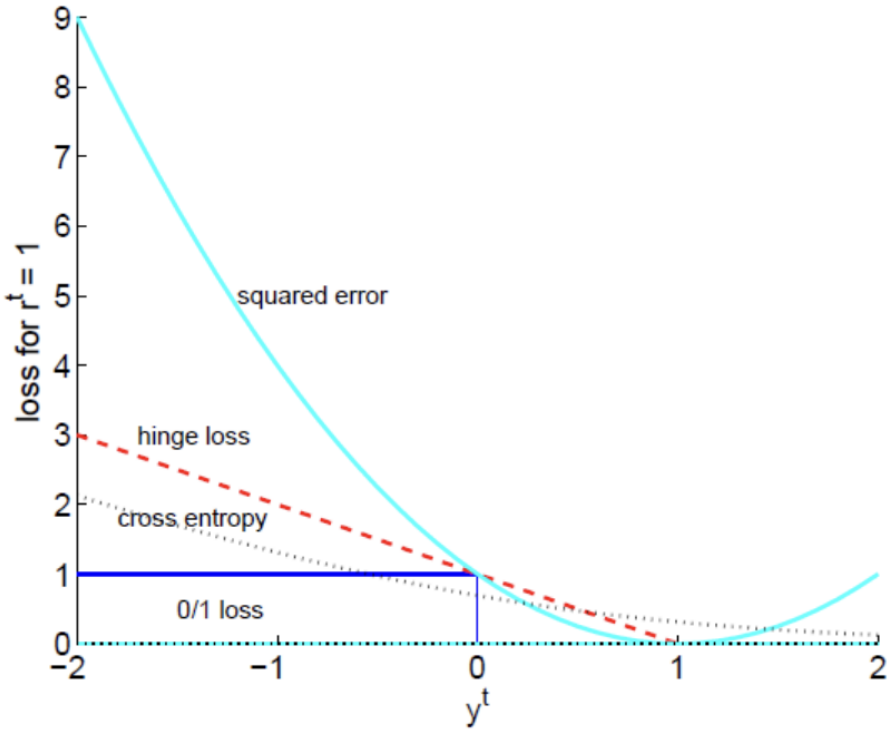


Figure 5-4. Large margin (left) versus fewer margin violations (right)

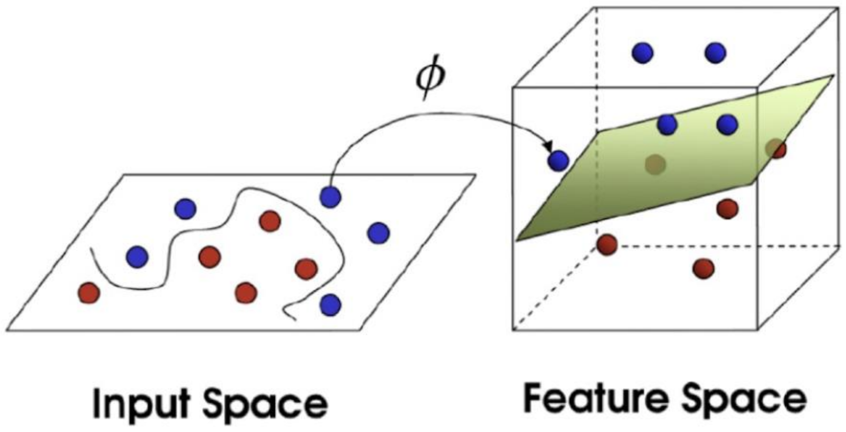
Hinge Loss



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

2. Kernel SVM

Extension to non-linearity

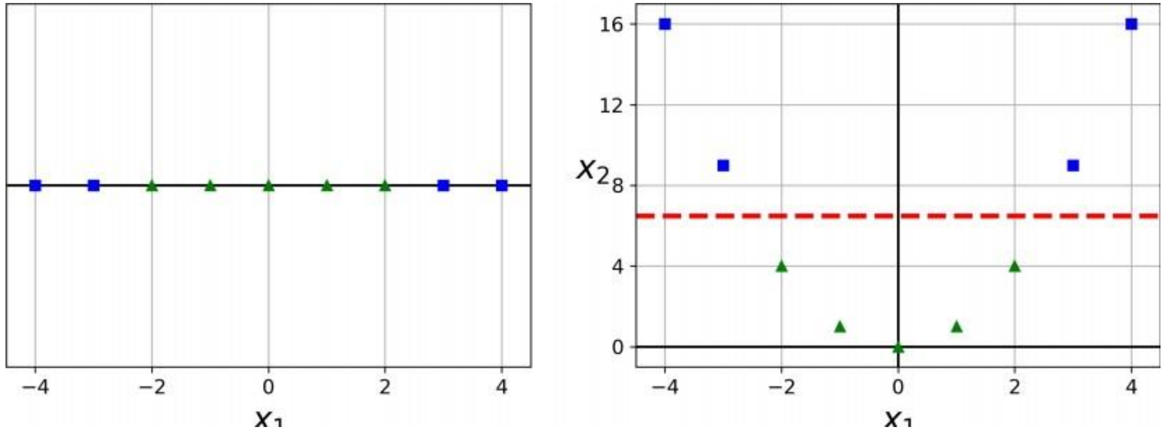


$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \cdots, \phi_n(\mathbf{x}))$$

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

$$z = \varphi(x)$$

Feature mapping



$$x_2 = (x_1)^2$$

$$x \rightarrow \{x, x^2\}$$

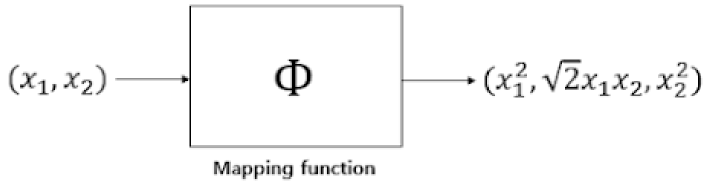
Kernel Trick

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

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

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

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



$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \langle \Phi(x_i) \Phi(x_j) \rangle$$

In linear SVM…

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

New feature space

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

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

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)

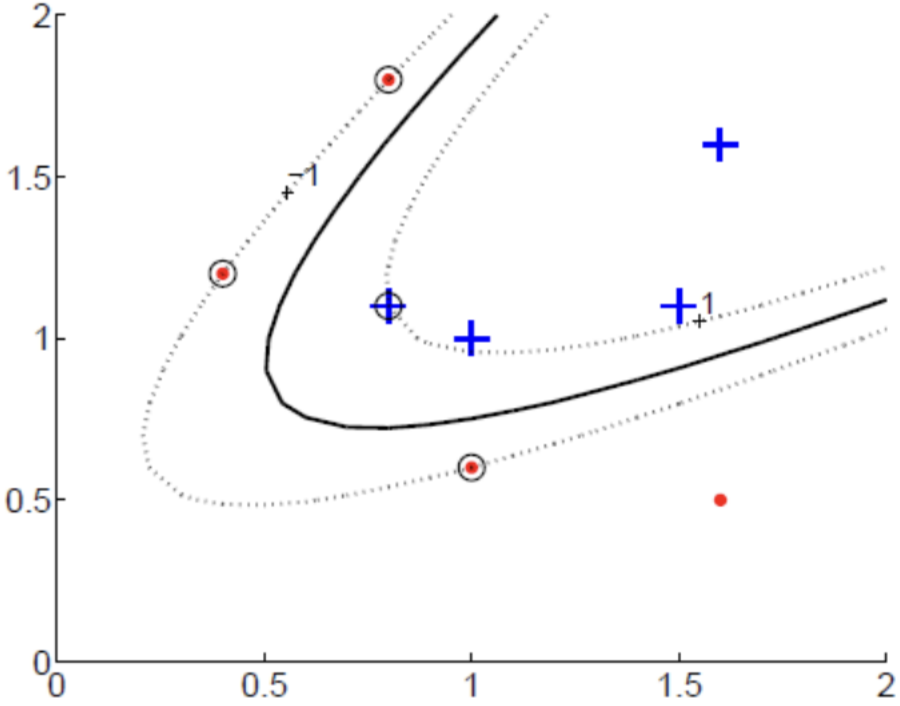
$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma + \gamma\, \mathbf{x}_i^T \mathbf{x}_j)^p$$

polynomial Kernel

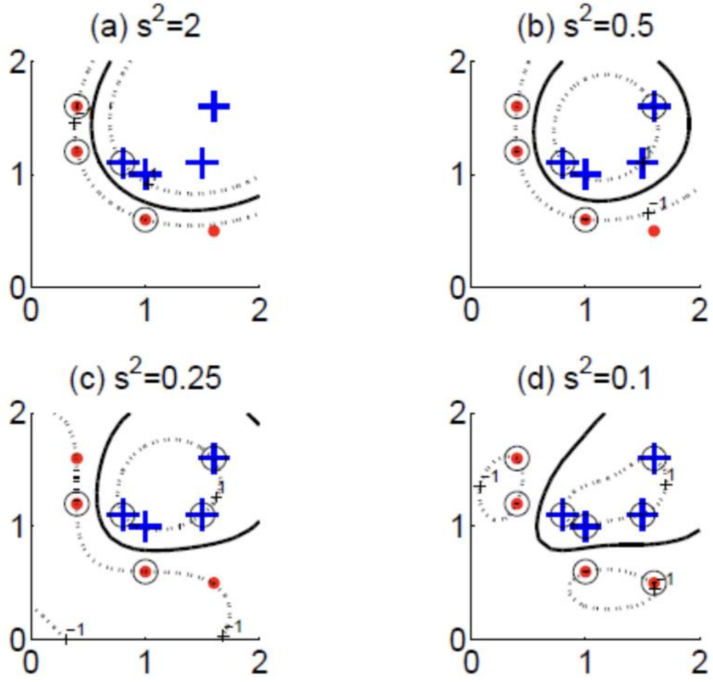
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(k_1 \mathbf{x}_i^T \mathbf{x}_j + k_2)$$

Sigmoid Kernel

Kernel SVM



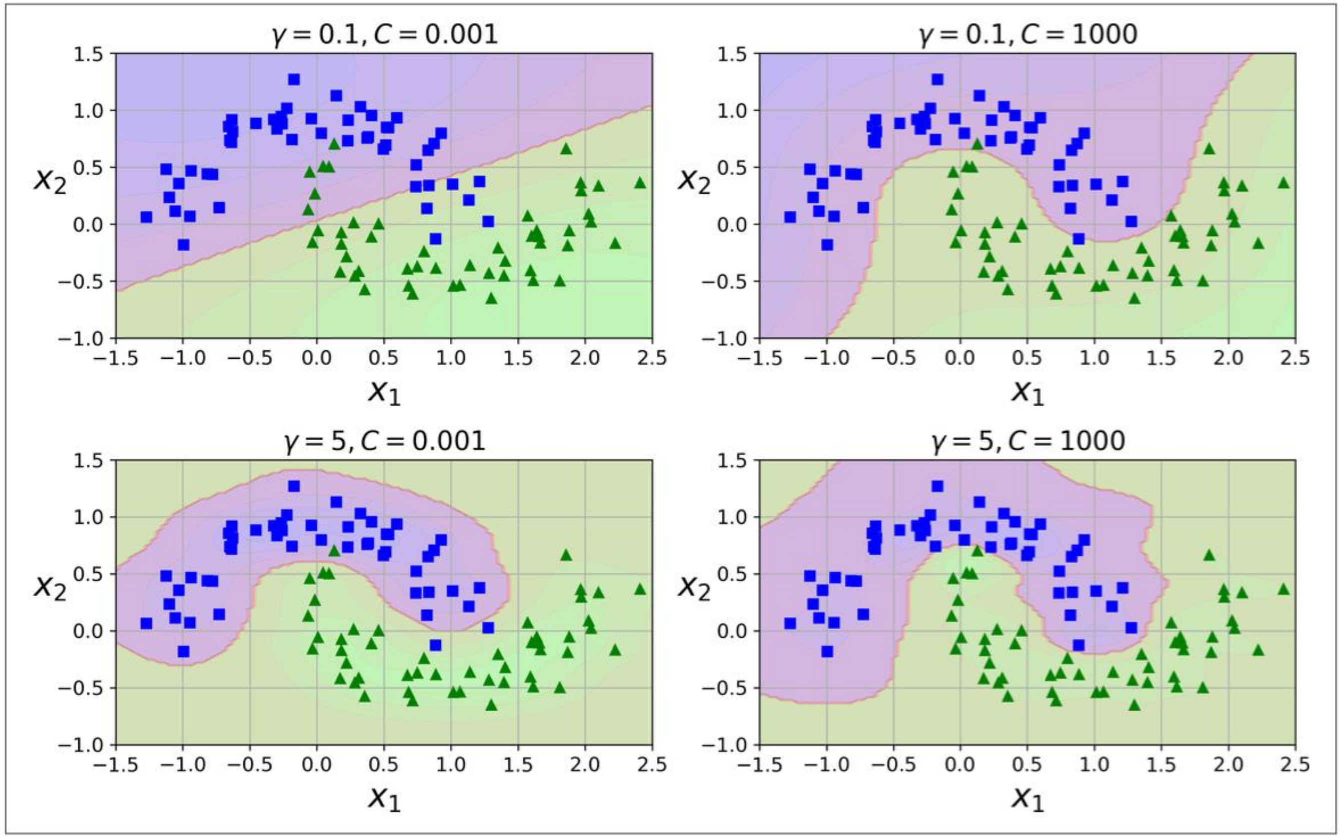
Polynomial Kernel



Gaussian(Radial-Basis function) Kernel

Kernel SVM

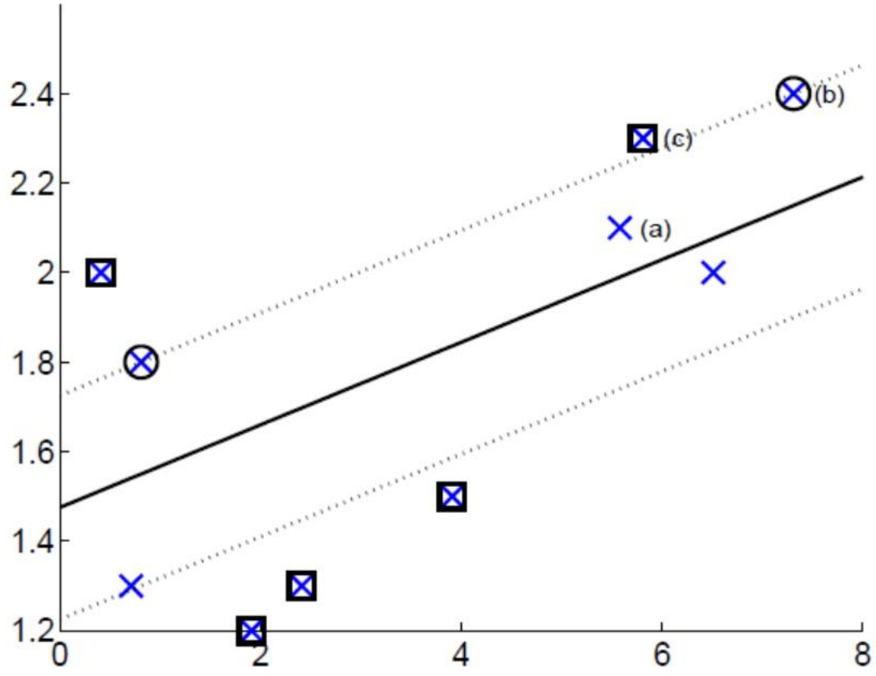
Gaussian(Radial-Basis function) Kernel



3. SVM - Regression

SVM- Regression

"Reverse the objective"



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

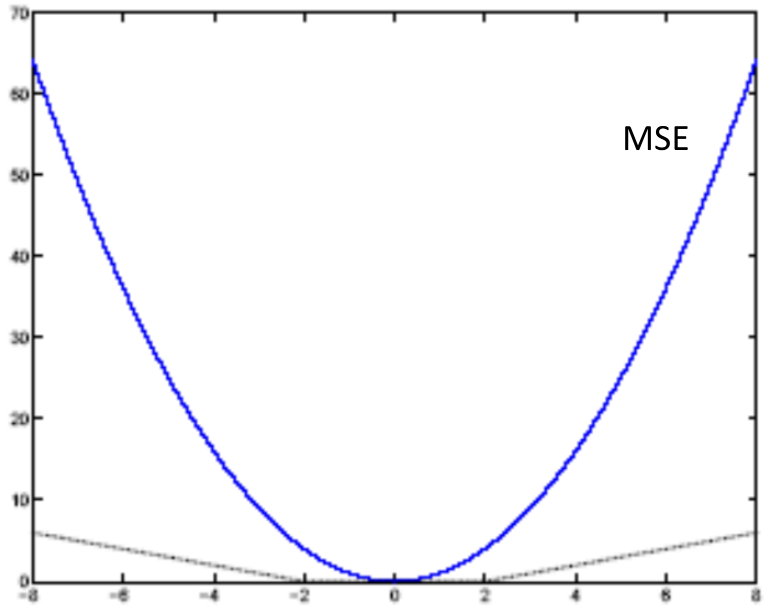
Lagragian Method

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

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

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

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



SVM- Regression

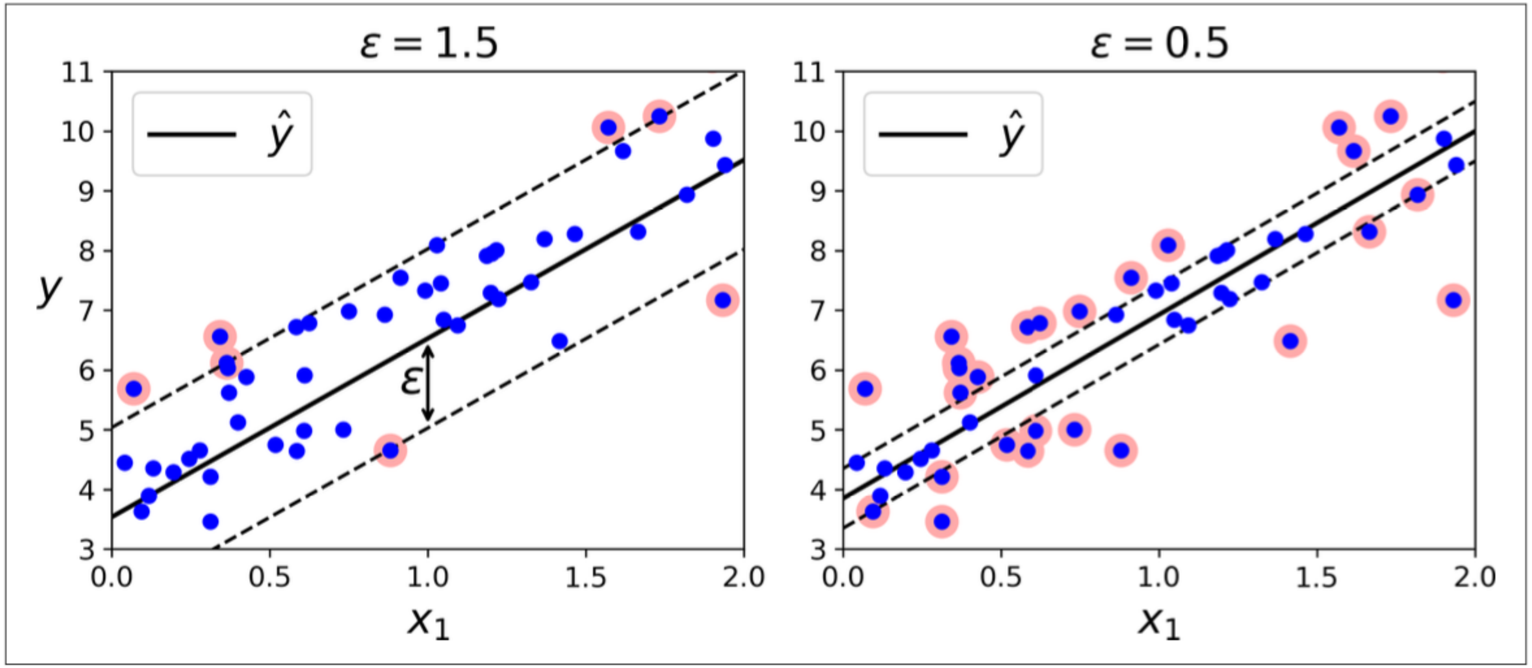
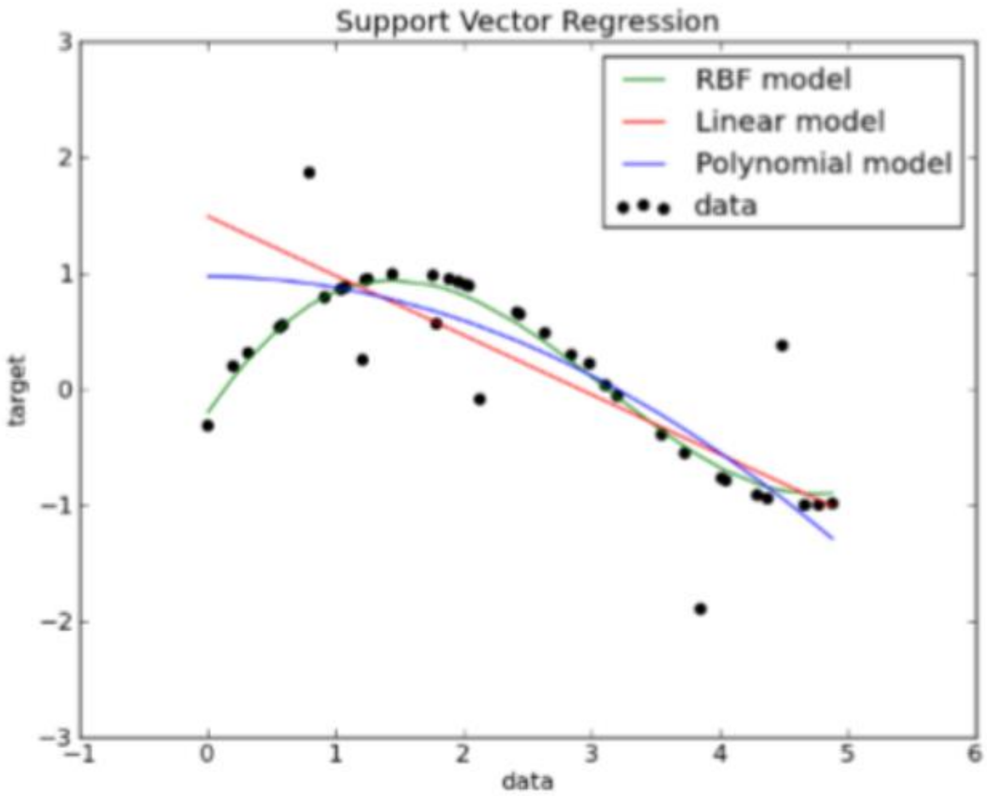
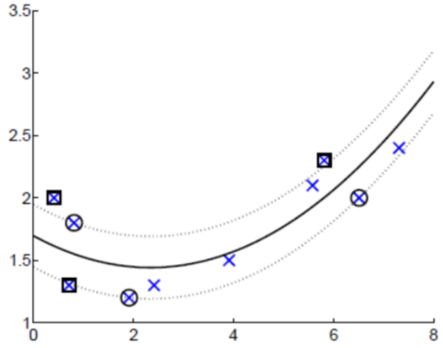


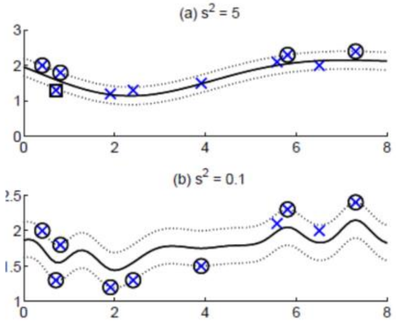
Figure 5-10. SVM Regression

SVM Kernel Regression



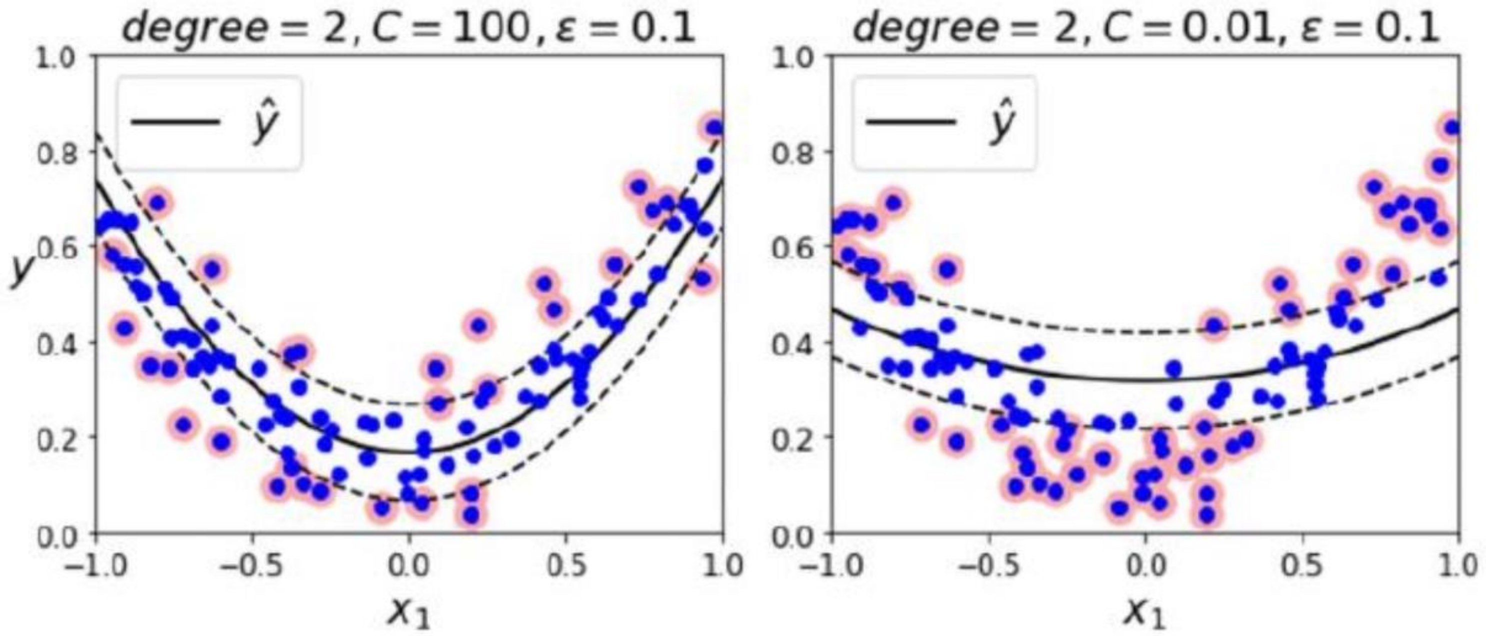


Polynomial Kernel



Gaussian Kernel

SVM- Regression



[6/7주차 개인별 프로젝트 발표 공지사항]

- 발표 형식 : 자유형식(ipynb,노션,ppt 등)
  - 발표 자료 제출 : KUBIG github > 1. 방학분반 > ML> 프로젝트 > 본인 이름)

- 발표 시간 : 6/7주차 세션 시작 전, 진행
  - 인당 5-10분 내외로 준비
  - 6주차 심서현, 임지우
  - 7주차 안태림, 하진우

* 2/29 쿠빅 콘테스트 예정 / 팀별 준비 진행!

수고하셨습니다!

해당 세션자료는 KUBIG Github에서 보실 수 있습니다!