Statistical Machine Learning

6주차

담당: 18기 방서연

1. Linear SVM

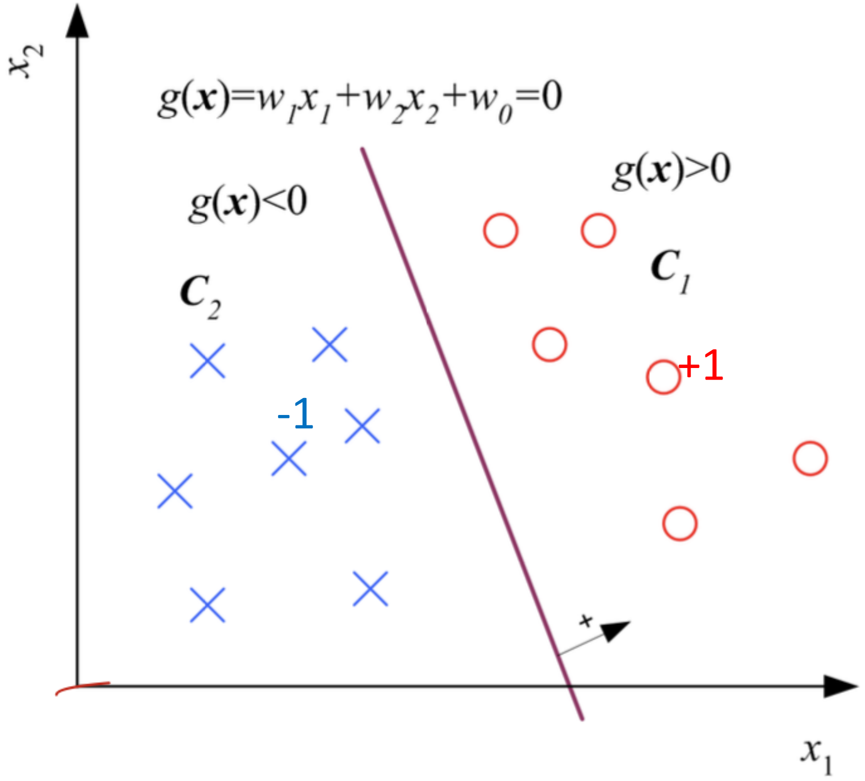
2. Kernel SVM

3. SVM-Regression

4. Decision Tree

1. Linear SVM - Classification

Linear Discriminant



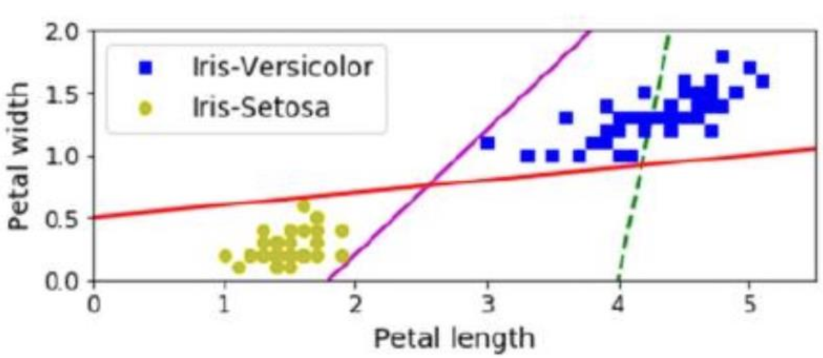
Decision Boundary or separating hyperplane

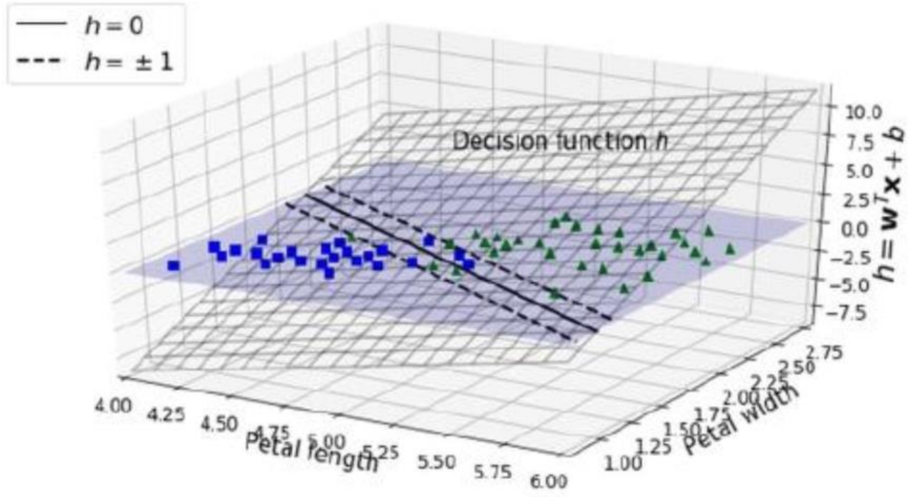
Decision Boundary : $g(x) = w^T x + w_0 = 0$

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

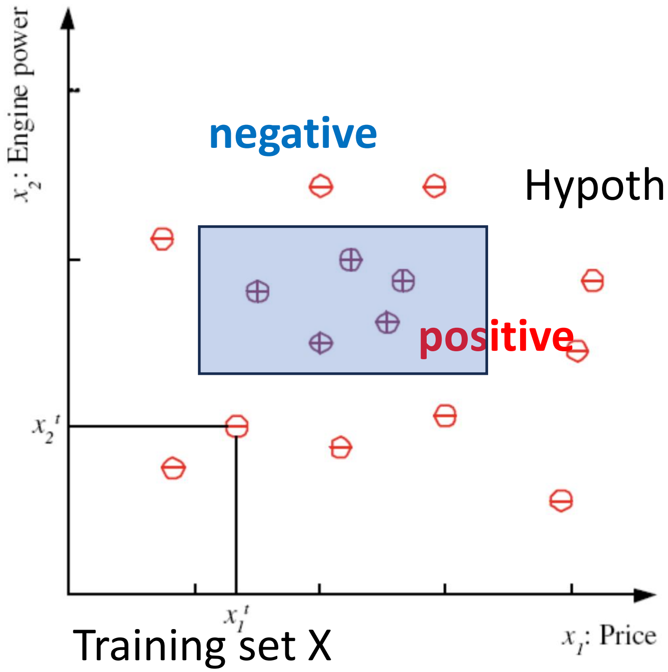
$w^T x + w_0 \geq +1, for\ r^t = +1$
$w^T x + w_0 \leq -1\ , for\ \ r^t = -1$

Hyperplane


Iris-Versicolor
Iris-Setosa
Petal width
Petal length
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5

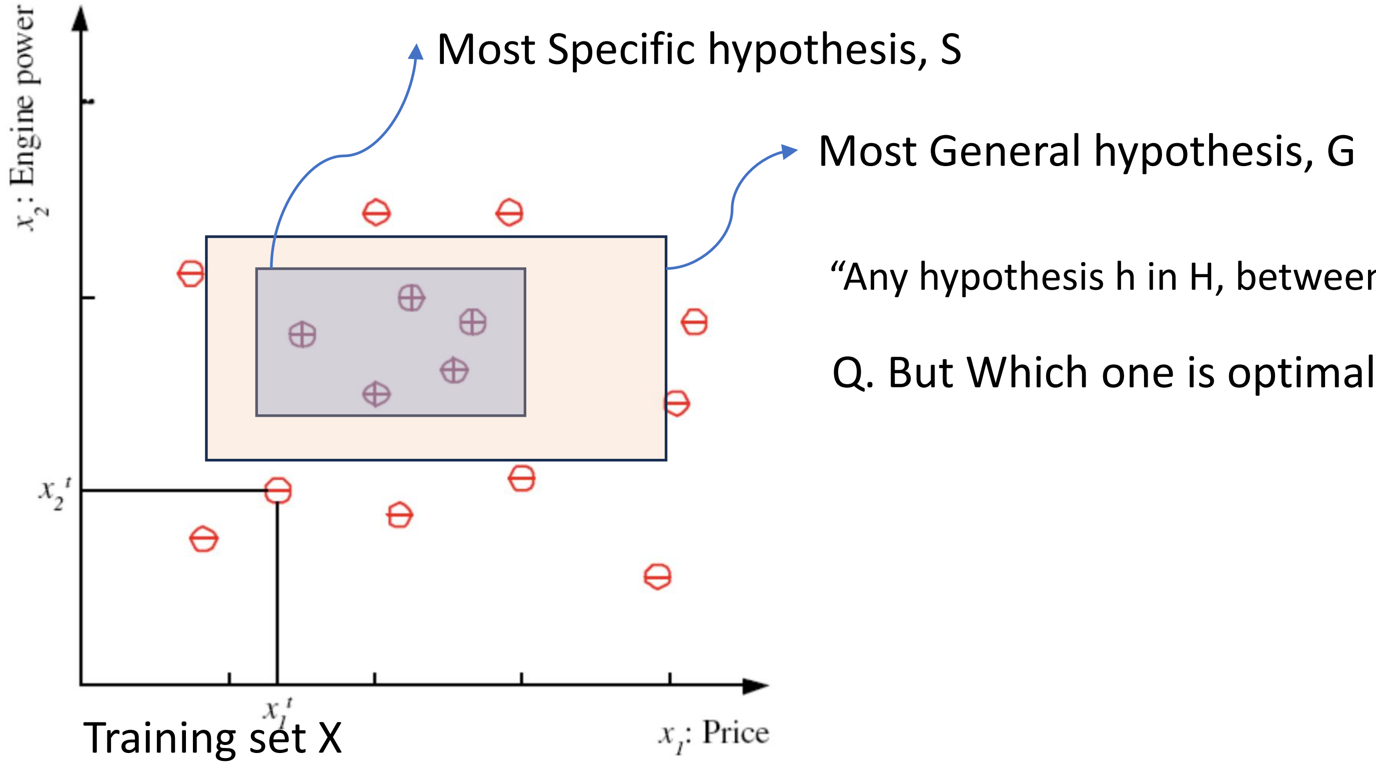

$h = 0$
$h = \pm 1$
Decision function h
$h = \mathbf{w}^T\mathbf{x} + b$
Petal length
Petal width
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
4.00
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75

S, G and the Version Space



Hypothesis H = Assumption = Model

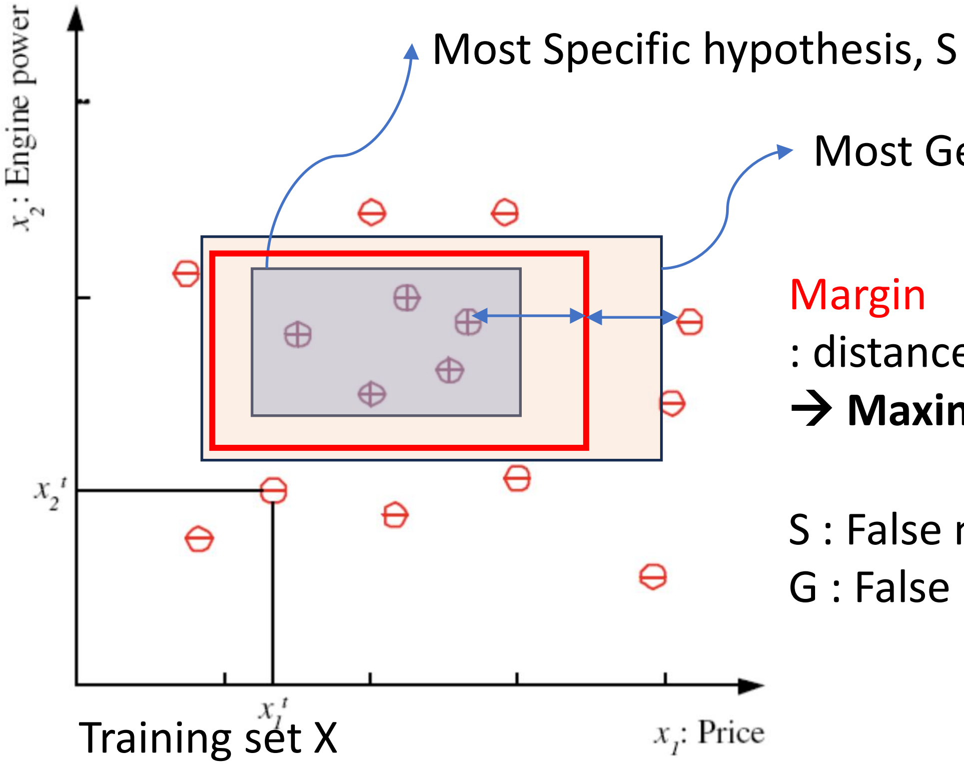
S, G and the Version Space



"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

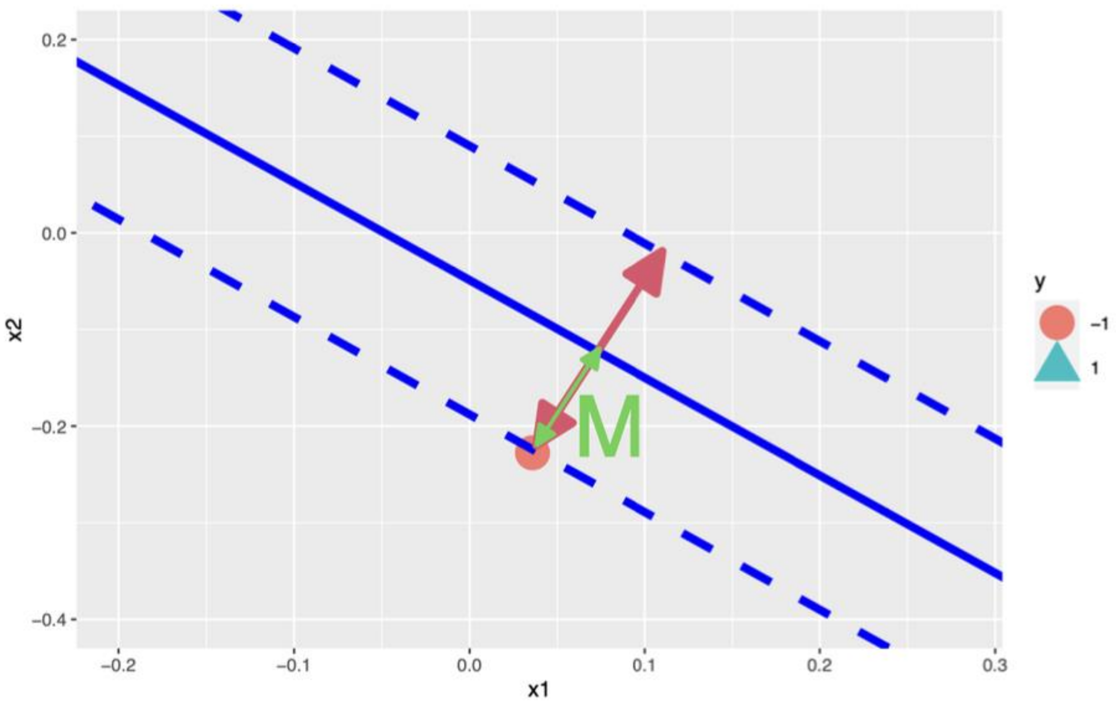
G : False positive에 취약

Optimal Hyperplane

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

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

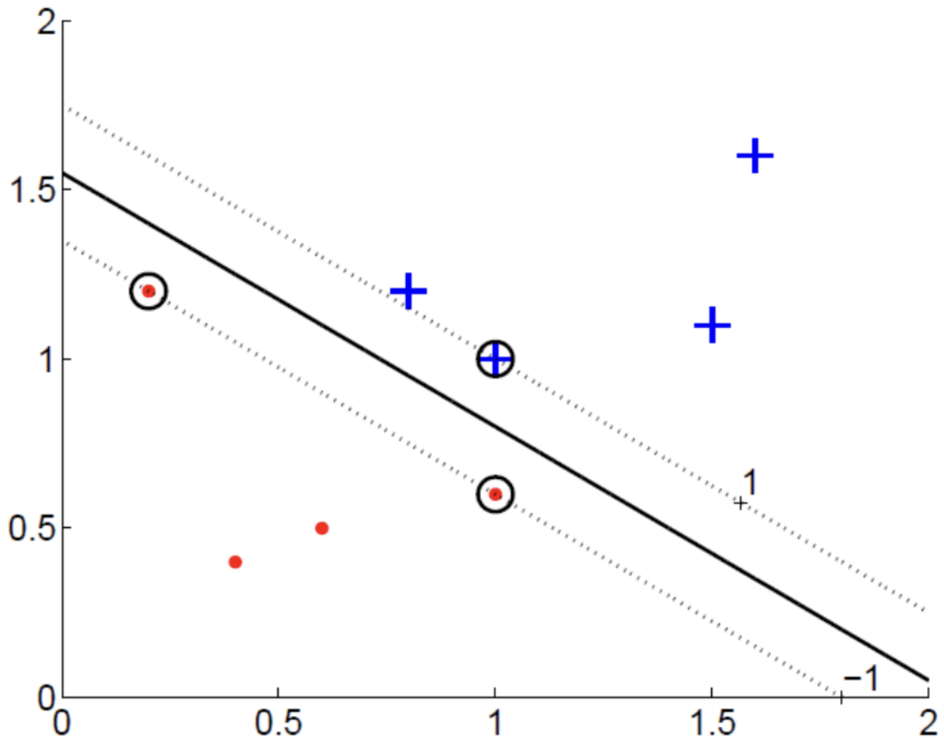
$\rightarrow\ r^t(w^T x + w_0)\ \geq\ +1$



[Margin]
Discriminant부터 양쪽 가장 가까운 instance 까지의 거리

Optimal Hyperplane(Discriminant) maximizes Margin

Objective of SVM



- Distance x to the hyperplane g(x)
- Margin

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

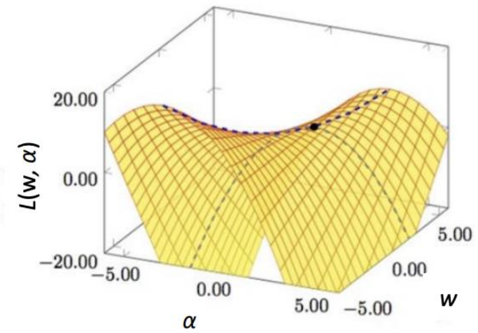
Lagrangian multiplier Method

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

Primal problem

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

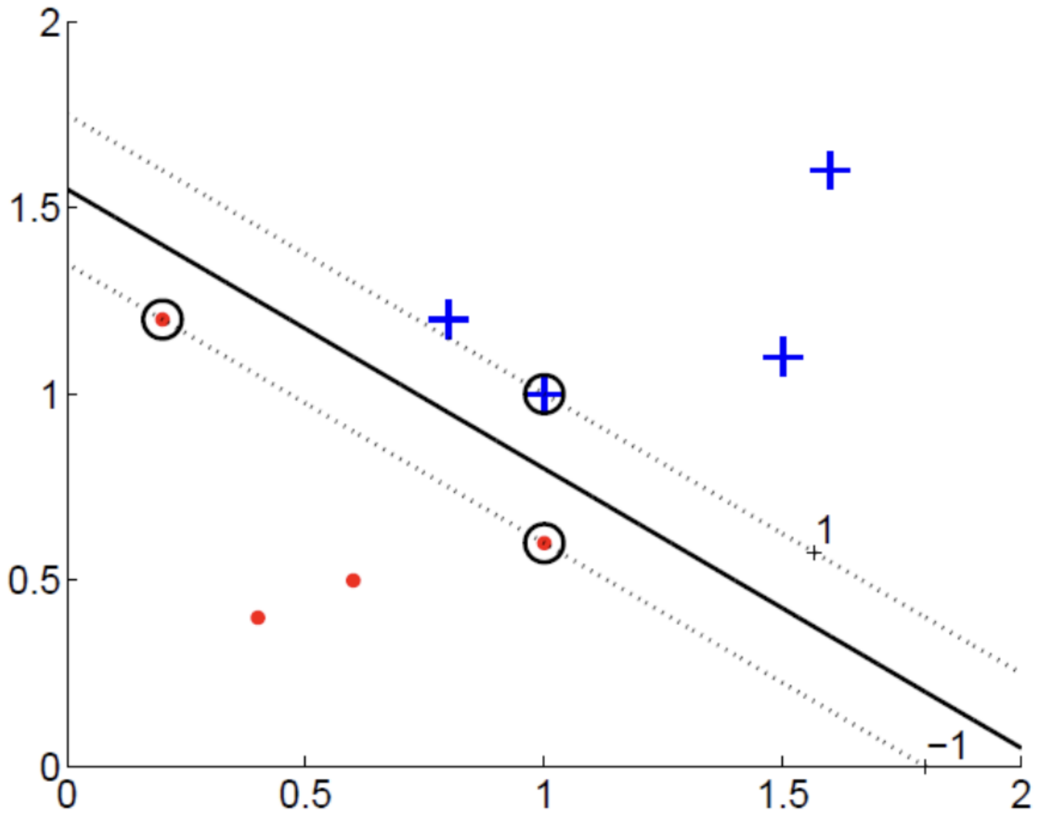
$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N} \alpha^t$$



KKT(Karush-Kuhn-Tucker Theorem)

1. Stationarity
2. Primal feasibility
3. Dual feasibility
4. Complementary slackness

SVM - Classification



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

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

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

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

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

Solution of SVM

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

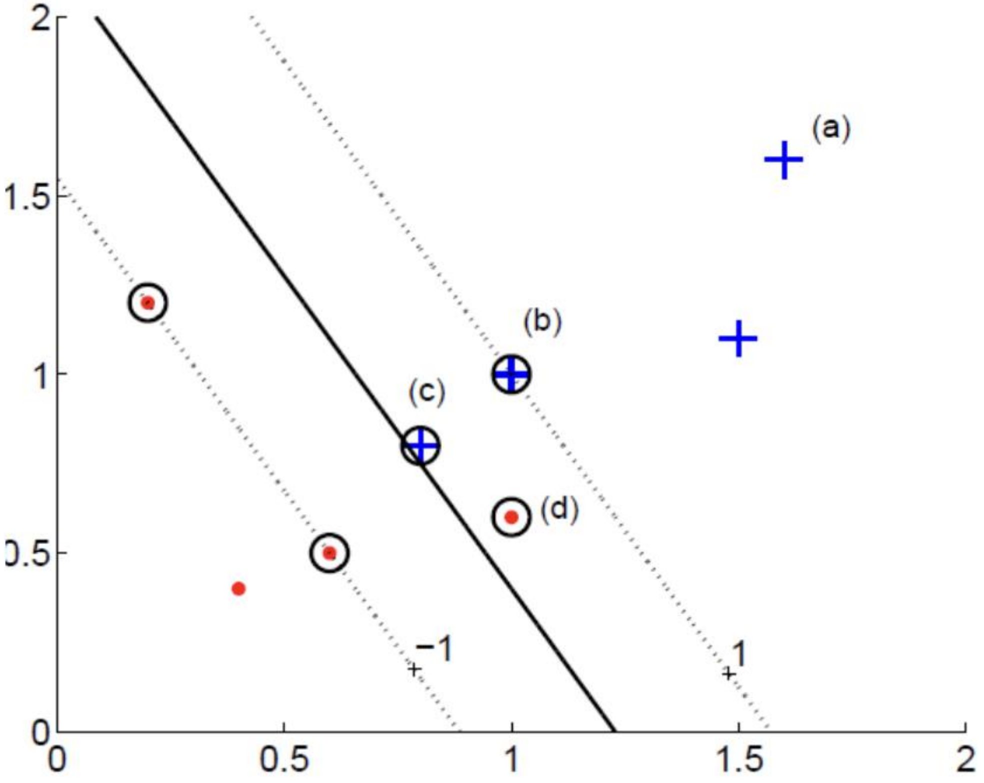
We want optimal hyperplane $g(x) = w^T x + w_0$

We want optimal w^* & w_0^*

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

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

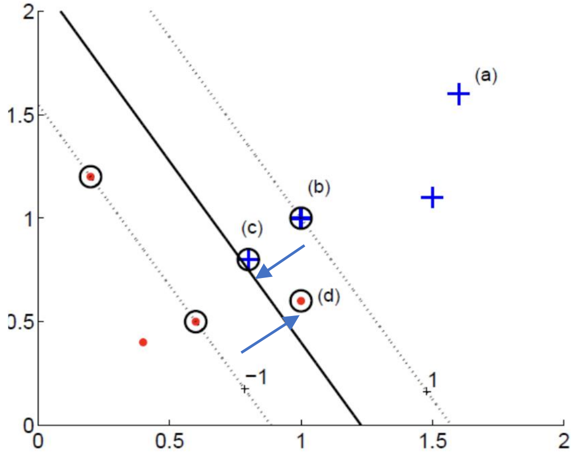
What if Non-Separable?



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

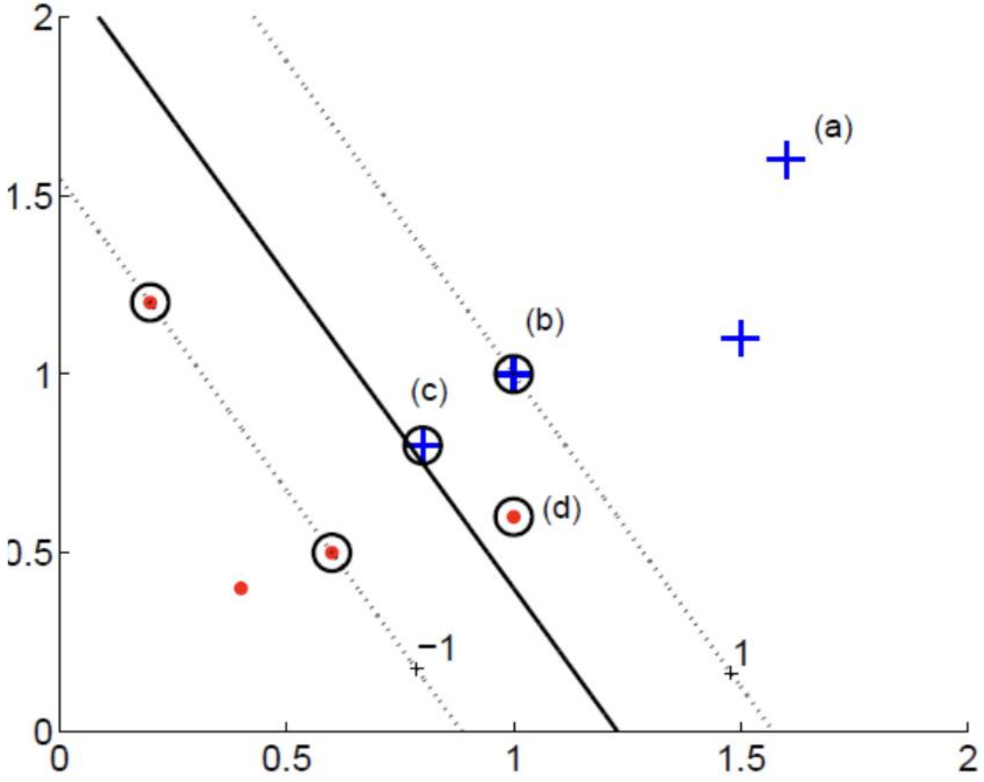
$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

- 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



Soft Margin Hyperplane

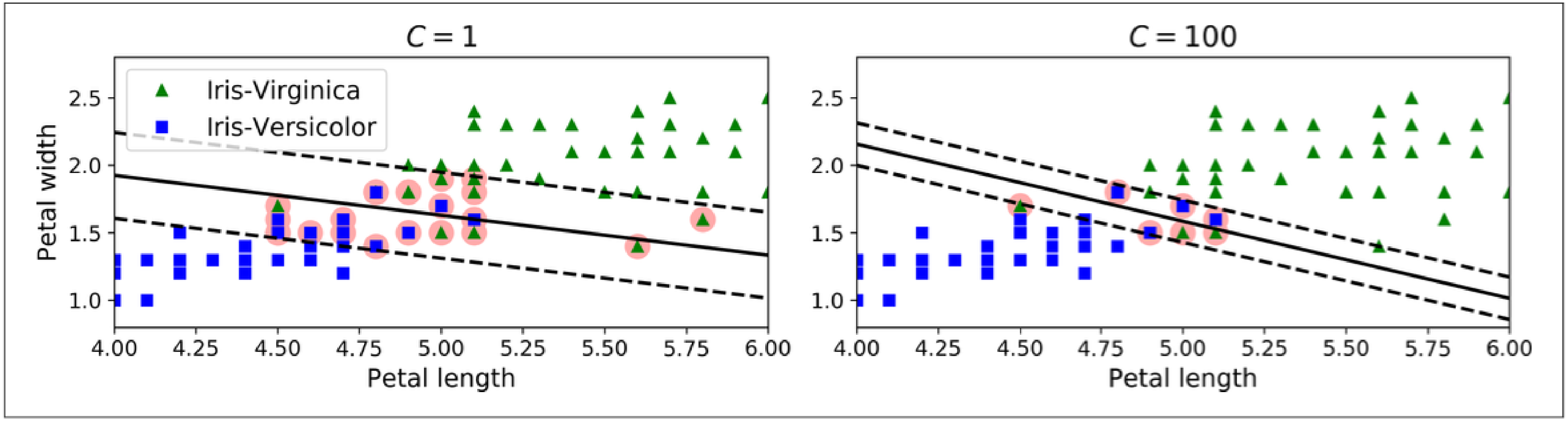
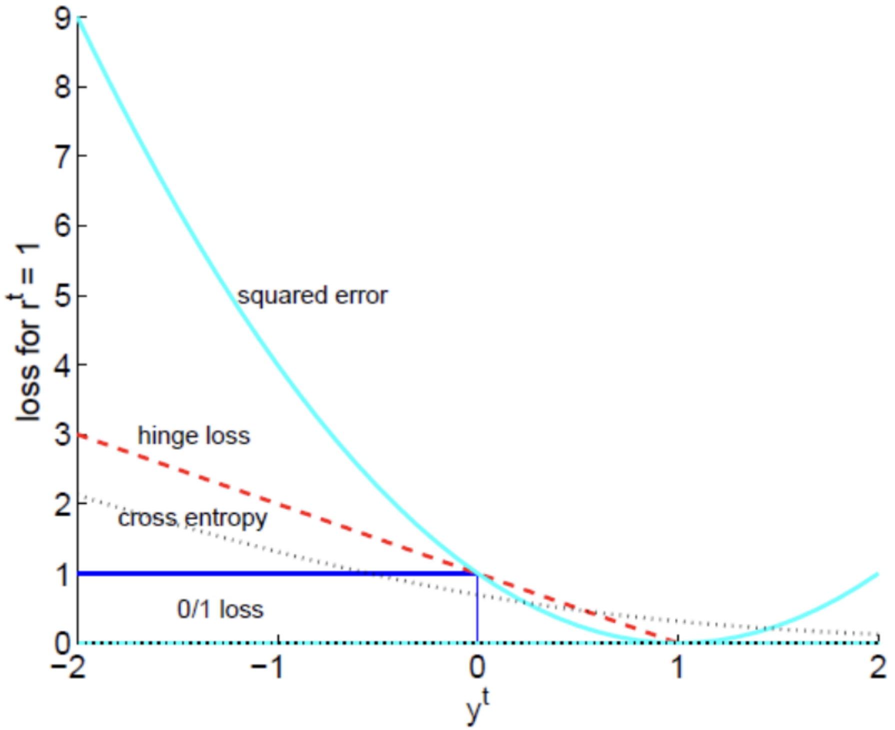


Figure 5-4. Large margin (left) versus fewer margin violations (right)

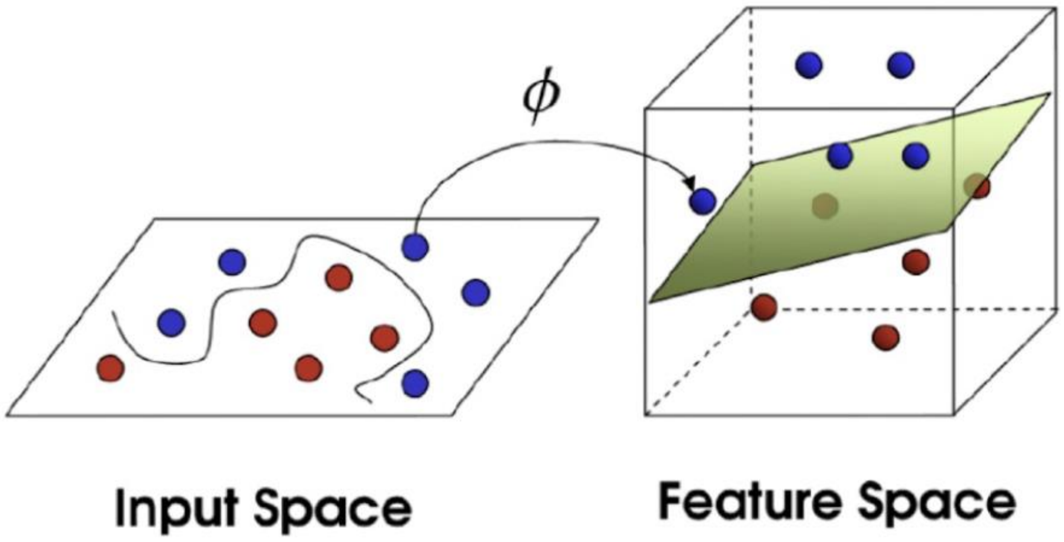
Hinge Loss



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

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

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…

New feature space

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

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

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

Kernel Trick

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

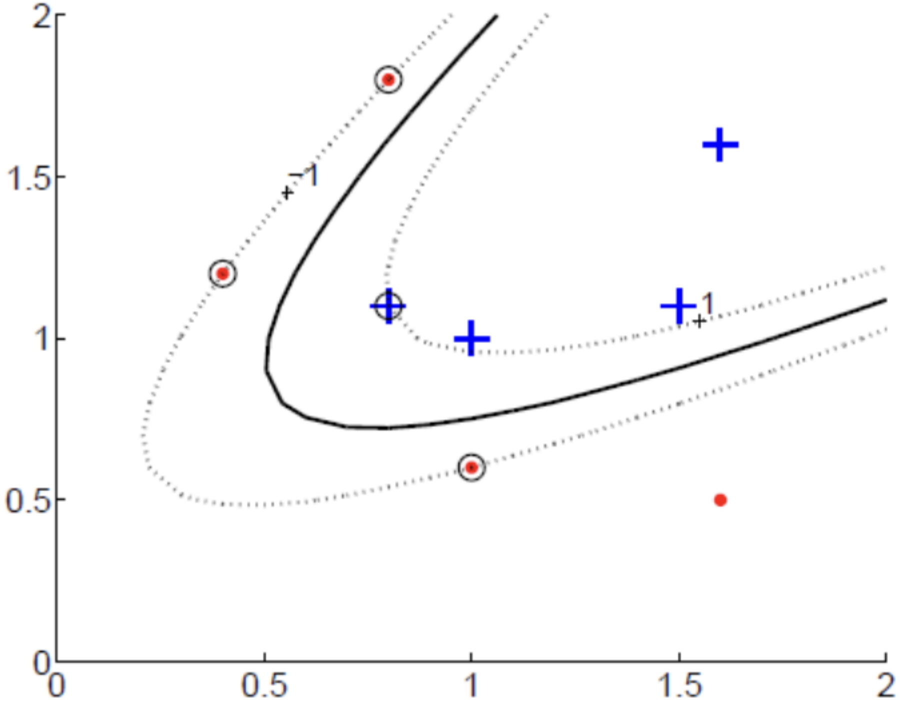
$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ — *Linear Kernel*

$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j))$ — *Gaussian Kernel (Radial Basis function)*

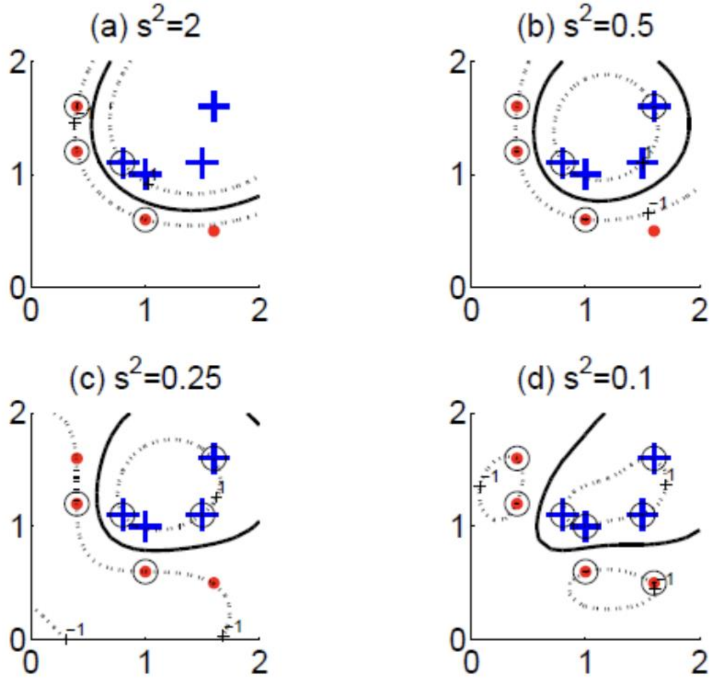
$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma + \gamma\, \mathbf{x}_i^T \mathbf{x}_j)^p$ — *polynomial Kernel*

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

Kernel SVM



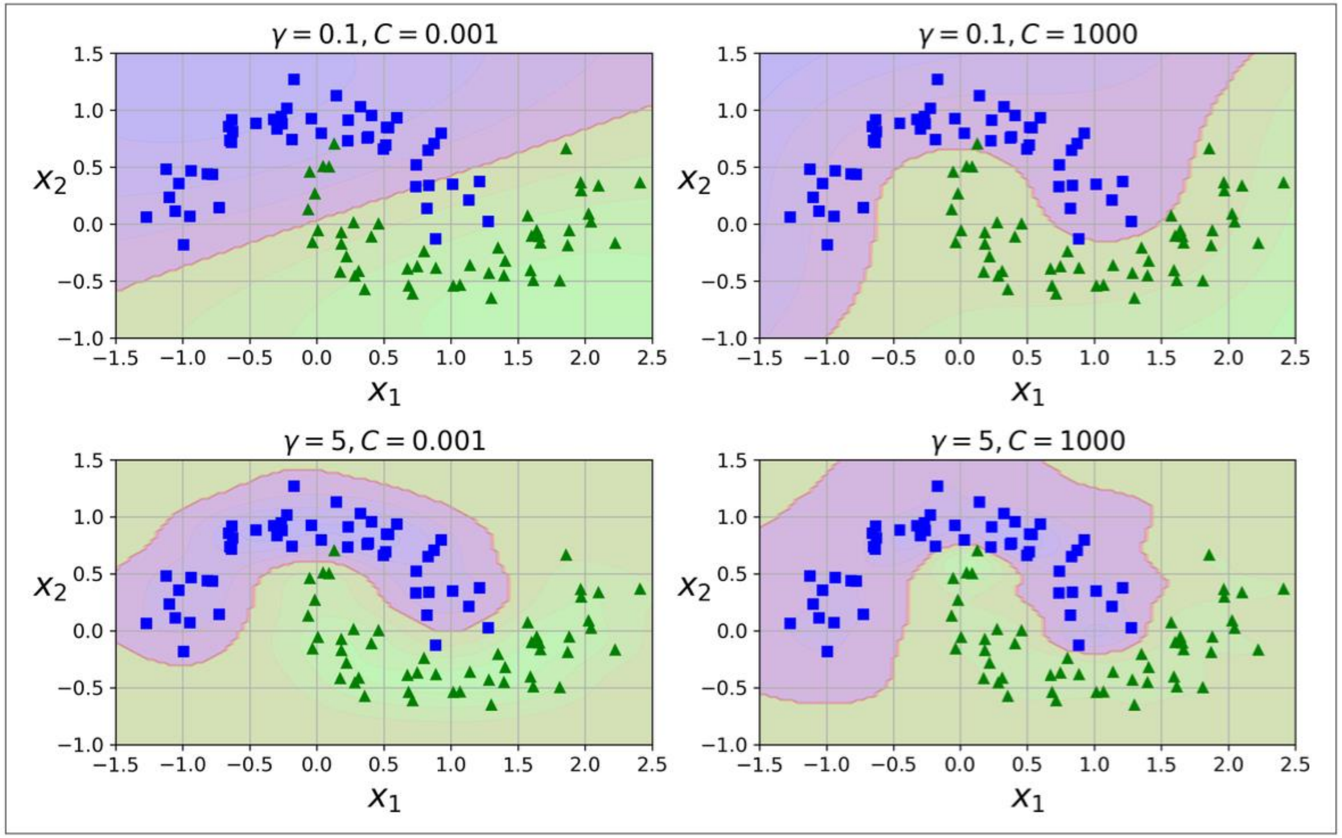
Polynomial Kernel



Gaussian(Radial-Basis function) Kernel

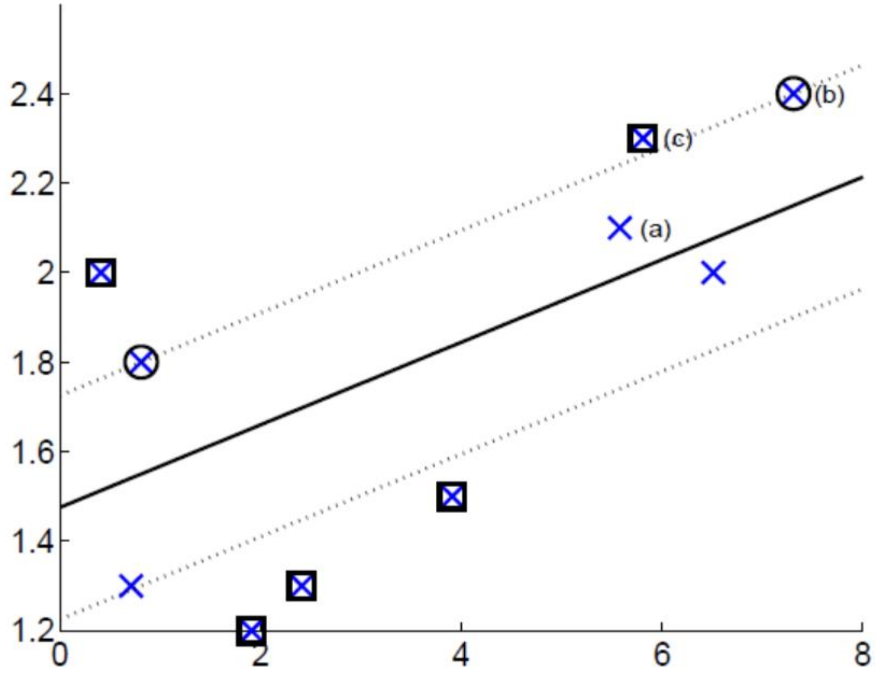
Kernel SVM

Gaussian(Radial-Basis function) Kernel



3. SVM - Regression

SVM- Regression



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

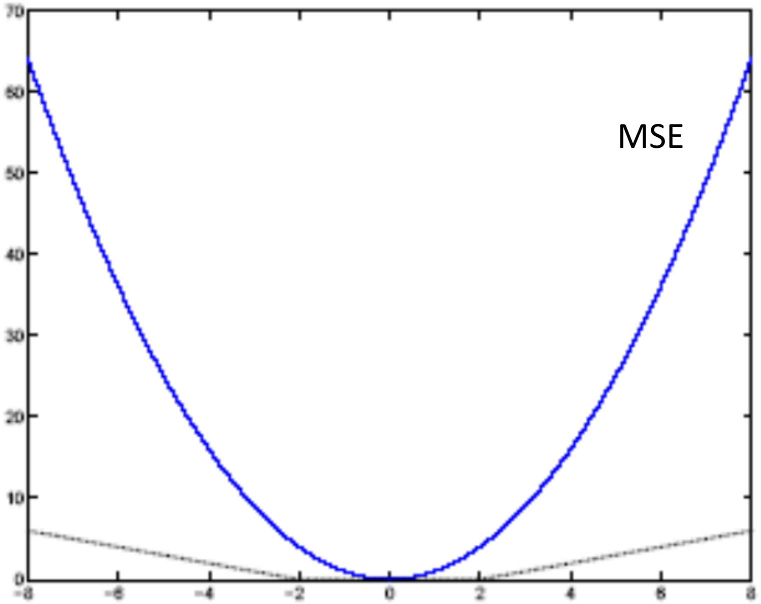
Lagragian Method

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

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

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

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



SVM- Regression

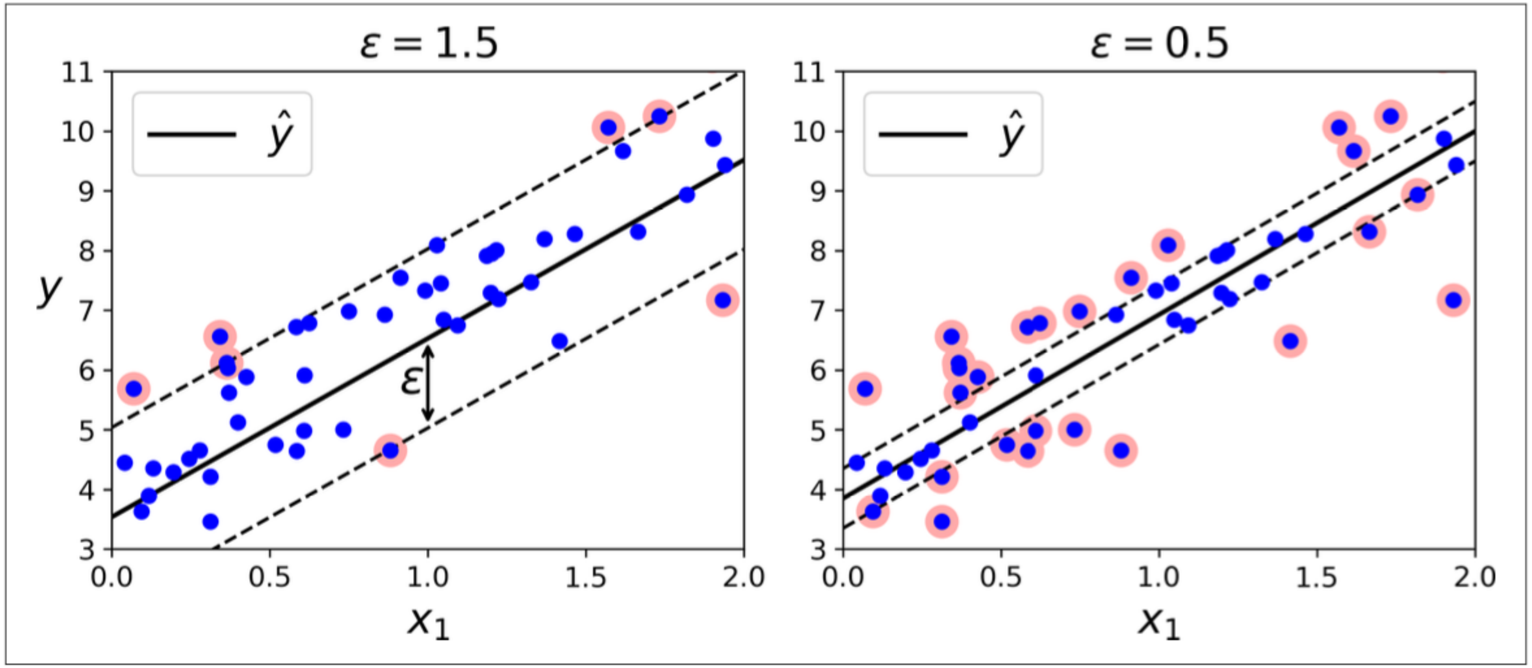
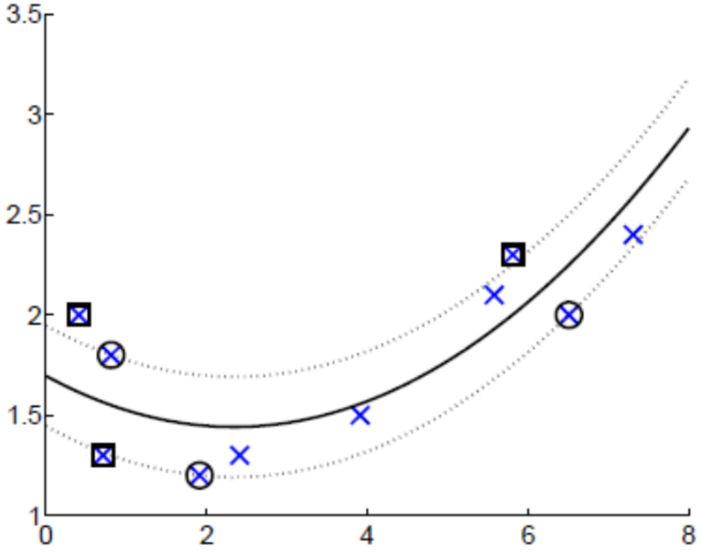
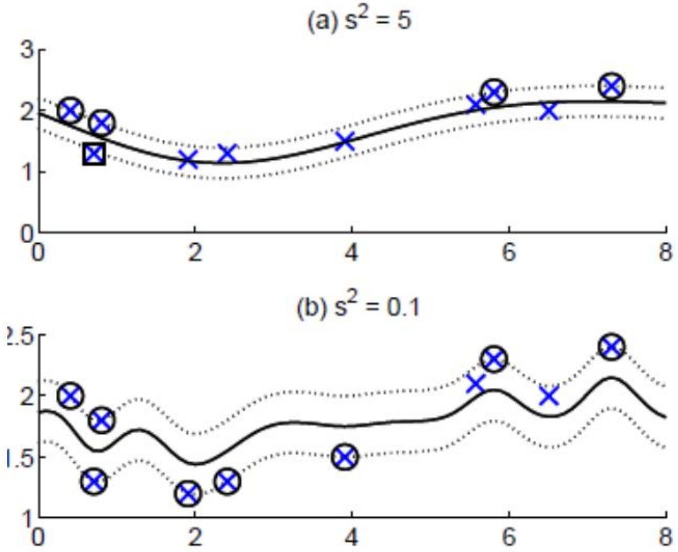


Figure 5-10. SVM Regression

SVM Kernel Regression

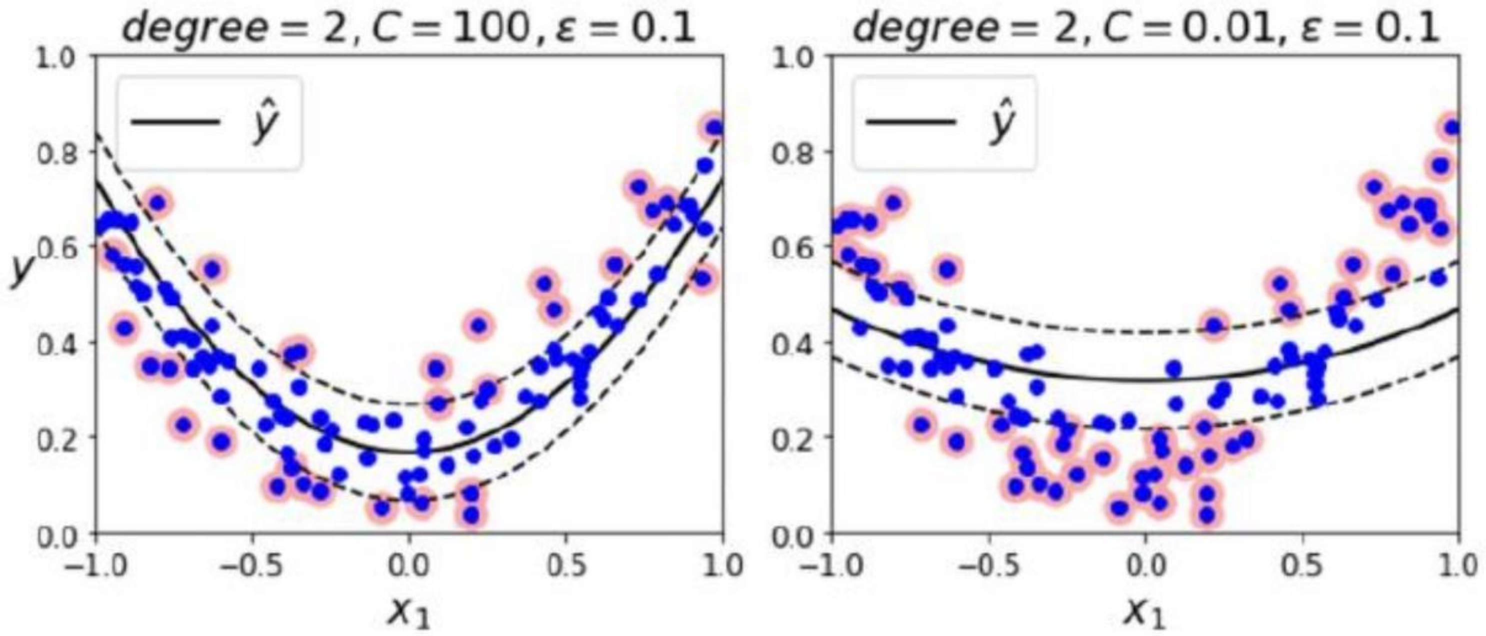


Polynomial Kernel



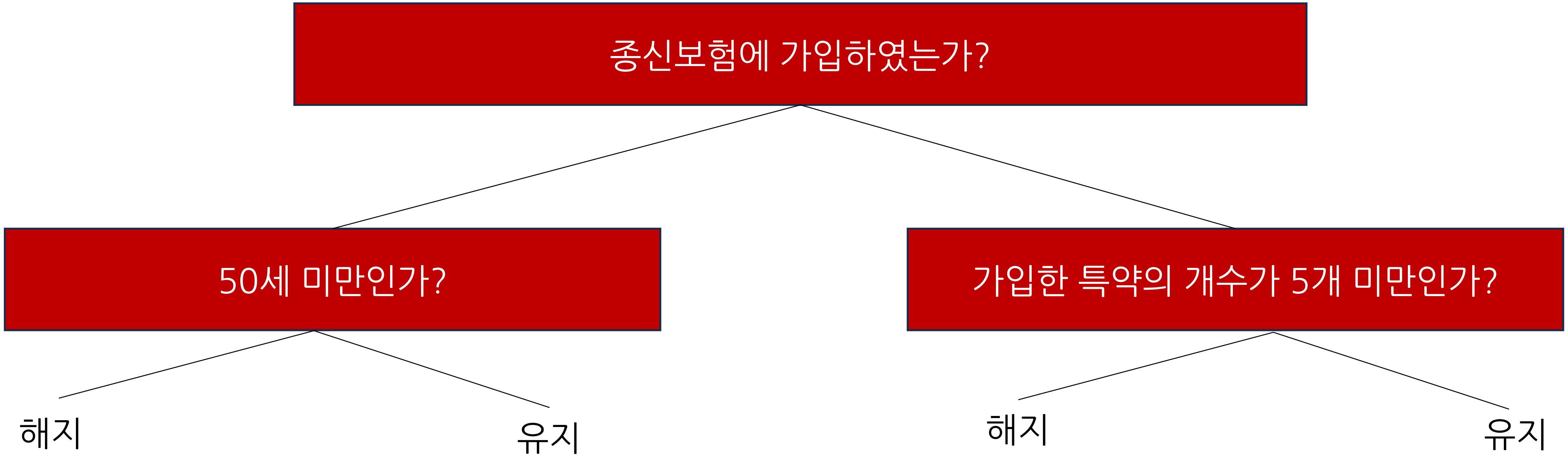
Gaussian Kernel

SVM- Regression



4. Decision Tree

Decision Tree



나무의 Root Node에서 출발하여 Leaf Node에 이르기까지 분기를 수행
-> 분류 성능은 분기의 기준에 달려있는데, 분기의 기준을 결정하는 기준은 무엇인가?

불순도

Gini Index	1 - (각 항목이 차지하는 비율의 제곱 합) 항목이 두 가지일 경우, 값의 범위는 0~0.5
Entropy	$-\Sigma p_i * \log_2 p_i$ 항목이 두 가지일 경우, 값의 범위는 0~1

Decision Tree

- 불순도가 낮아지는 방향으로 주어진 데이터를 분류하는 분석 방법
- 어떤 불순도 지표를 택할지, 어떻게 분류할지에 따라 생성 방식 상이

CART	Classification and Regression Tree
CHAID	Chi-squared Automatic Interaction Detection

CART DT - Classification

종신?	50세?	특약?	해지
O	X	X	O
O	O	X	X
O	X	X	X
X	X	X	X
O	O	X	O
O	O	O	X
X	X	O	X

CASE1 : '종신보험인가?'로 분류하는 것이 최적?

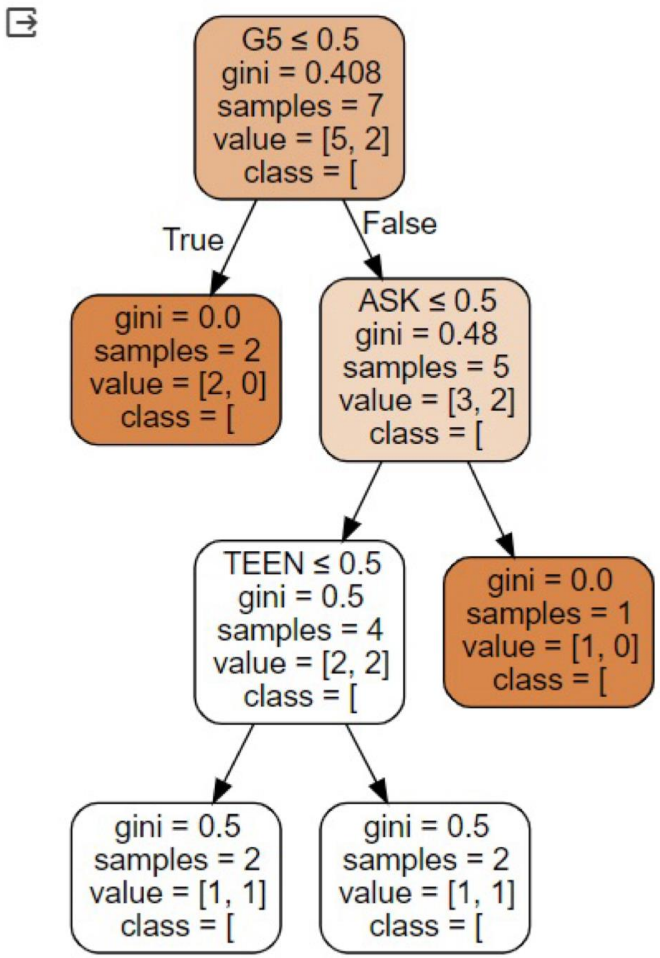
- (종신 보험 그룹) 해지 2 vs 유지 3 -> Gini =
- (종신 보험 x 그룹) 해지 0 vs 유지 2 -> Gini =

CASE2 : '50세 미만인가?'로 분류하는 것이 최적?

CASE3: '특약이 5개 미만인가?'로 분류하는 것이 최적?

CART DT - Classification

종신?	50세?	특약?	해지
O	X	X	O
O	O	X	X
O	X	X	X
X	X	X	X
O	O	X	O
O	O	O	X
X	X	O	X



CART DT - Regression

종신?	50세?	특약?	보험료
O	X	X	37k
O	O	X	43k
O	X	X	92k
X	X	X	15k
O	O	X	82k
O	O	O	83k
X	X	O	19k

CASE1 : '종신보험인가?'로 분류하는 것이 최적?

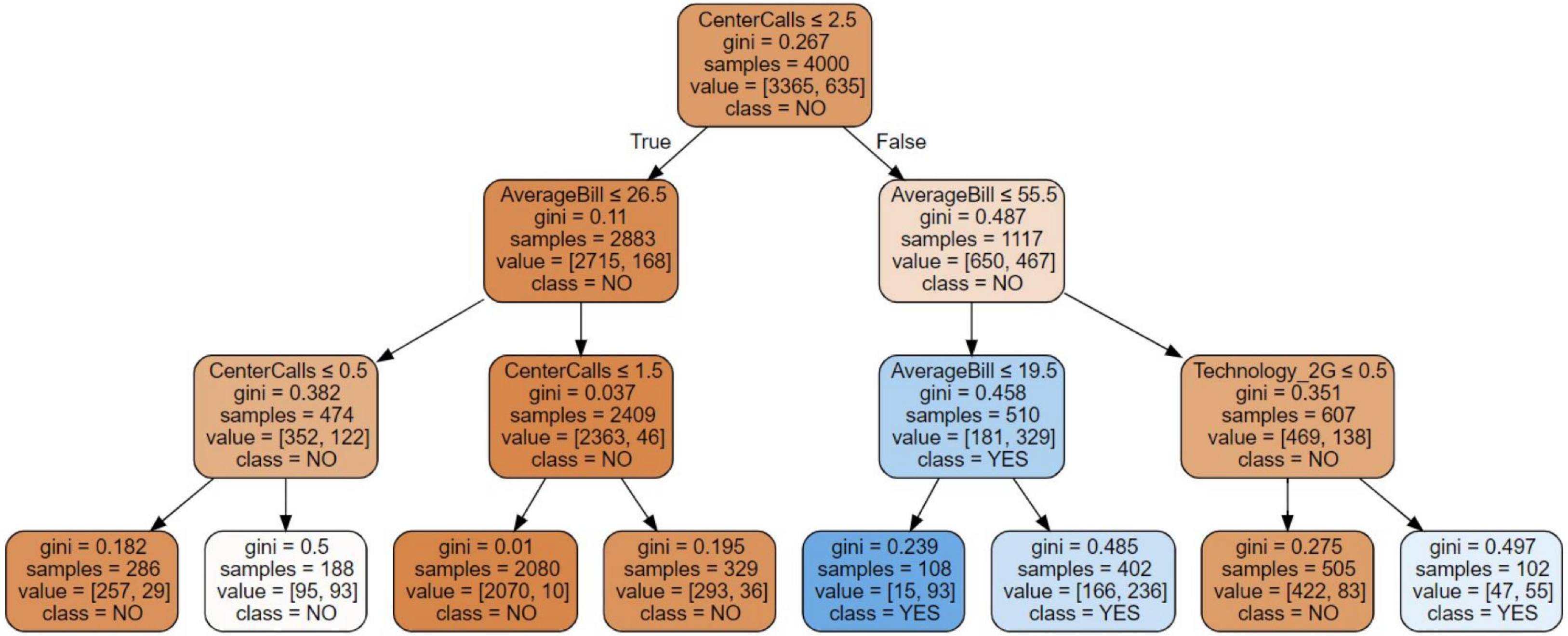
CASE2 : '50세 미만인가?'로 분류하는 것이 최적?

CASE3: '특약이 5개 미만인가?'로 분류하는 것이 최적?

Graphviz - 코드 실습

```
graph = graphviz.Source(dot_data)
graph

# value는 class의 범주 그룹 별로 몇개 씩 속해 있는지 NO에 3365개, YES에 635개
# 주황/파랑인 노드들이 많이 보임 (채색의 농도 차이는 있음) - 파랑 : yes / 주황 : no
```



- Graphviz 결과 해석?

7주차 과제 리마인드

- Week 6 과제 제출 (Github)
- 팀 별 Contest 중간 보고 요약 제출 (문서-노션 등, ppt 불필요) (Slack)

- 8월 14일 22:00까지 제출 요망

수고하셨습니다!

해당 세션자료는 KUBIG Github에서 보실 수 있습니다!