
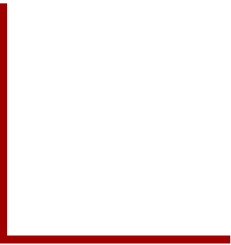


Support Vector Machine & Decision Tree

5주차
담당: 13기 박주영

- 
- 
1. Lagrange Multiplier Theorem
 2. Linear Support Vector Machine
 3. Kernel Support Vector Machine
 4. Decision Tree

Lagrange Multiplier Theorem

- Primal Problem

$$\min_{\mathbf{x}} f(\mathbf{x})$$

subject to $g_i(\mathbf{x}) \leq 0, \quad \text{for } i = 1, \dots, m$

$$h_j(\mathbf{x}) = 0, \quad \text{for } j = 1, \dots, k$$

Lagrange Multiplier Theorem

- Dual Problem

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) + \sum_i^m \alpha_i g_i(\mathbf{x}) + \sum_j^k \gamma_j h_j(\mathbf{x})$$

$$\alpha_i \geq 0, \quad \text{for } i = 1, \dots, m$$

$$\gamma_j \geq 0, \quad \text{for } j = 1, \dots, k$$

Karush-Kuhn-Tucker Conditions

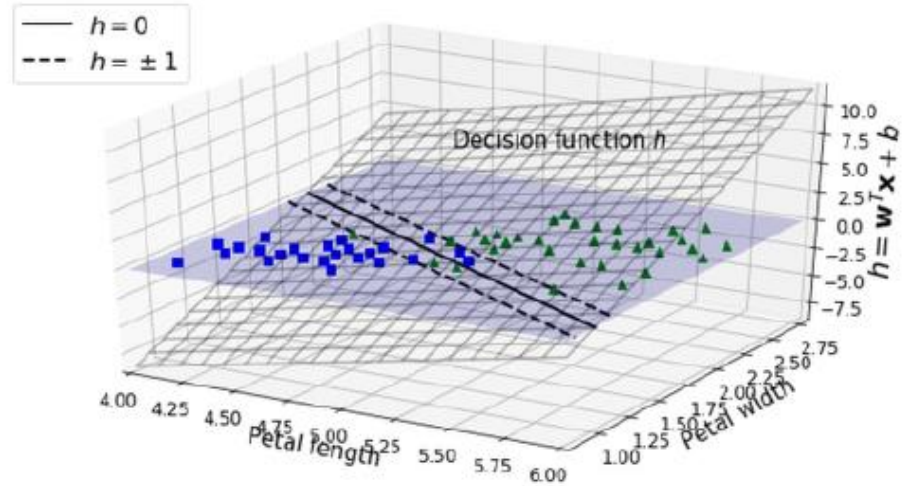
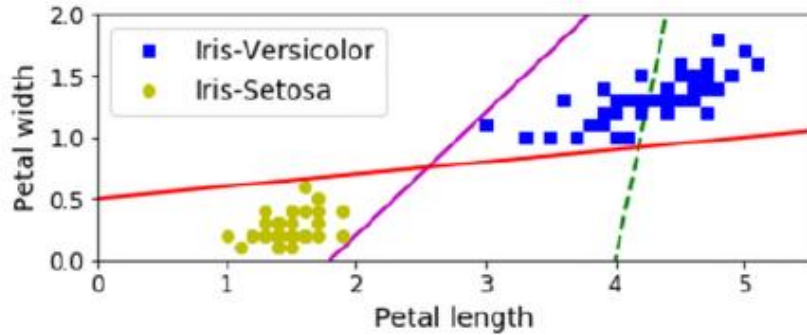
1. $\nabla f(\mathbf{x}) + \sum_i^m \alpha_i \nabla g_i(\mathbf{x}) + \sum_j^k \gamma_j \nabla h_j(\mathbf{x}) = 0$ (Stationary)

2. $\alpha_i g_i(\mathbf{x}) = 0$, for $i = 1, \dots, m$ (Complementary Slackness)

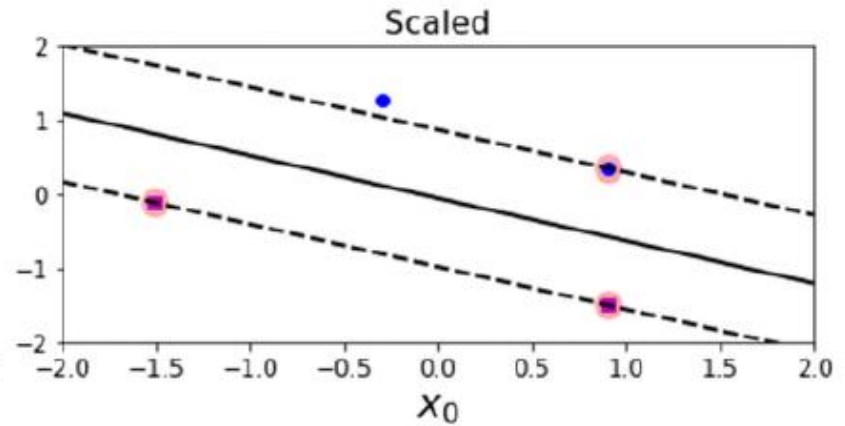
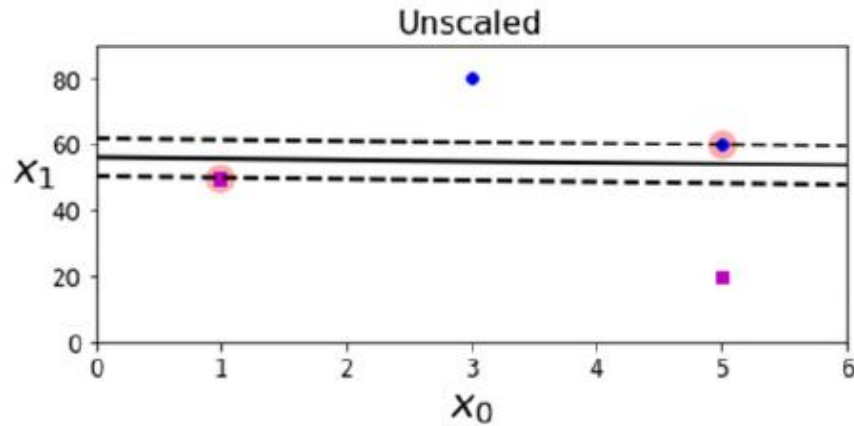
3. $g_i(\mathbf{x}) \leq 0$, for $i = 1, \dots, m$ and (Primal Feasibility)
 $h_j(\mathbf{x}) = 0$, for $j = 1, \dots, k$

4. $\alpha_i \geq 0$, for $i = 1, \dots, m$ (Dual Feasibility)

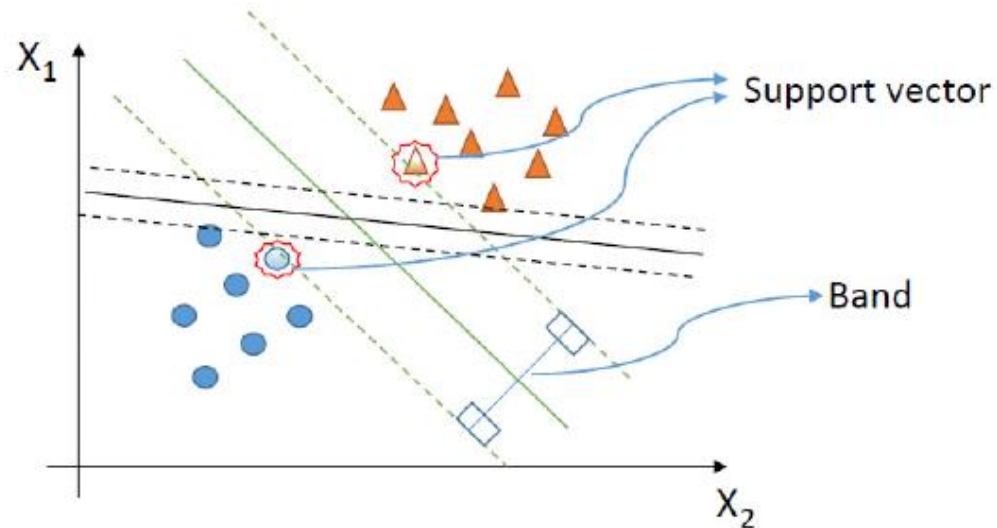
Hyperplane



Scaled? Unscaled?

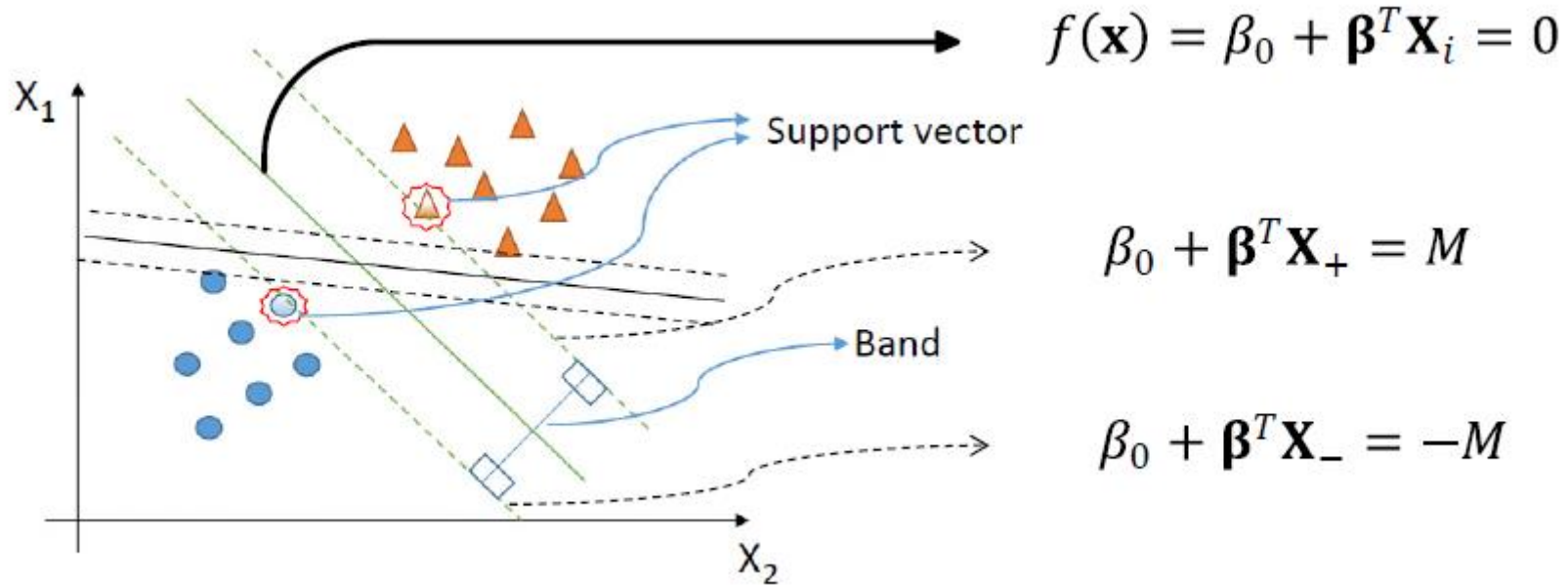


Linear Support Vector Machine

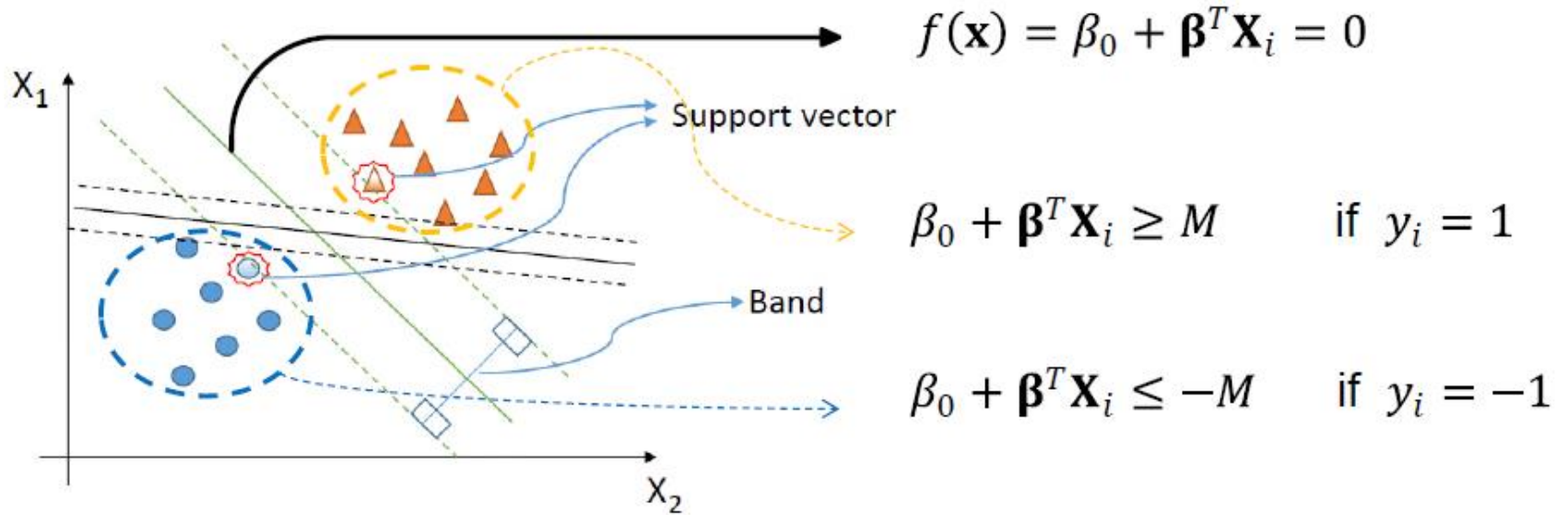


$$y = \{-1, 1\}$$

Linear Support Vector Machine

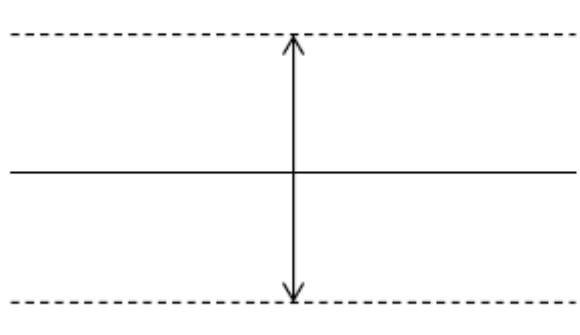


Linear Support Vector Machine



Linear Support Vector Machine

- We want to **maximize** the width of the band.

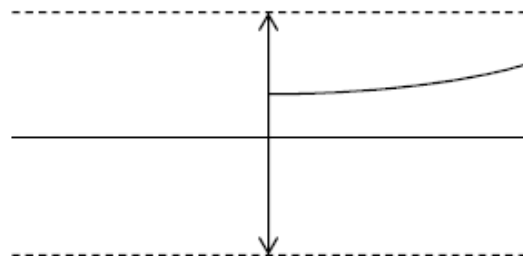


The diagram illustrates a linear Support Vector Machine (SVM) in a 2D space. It features three horizontal lines: a solid central line representing the decision boundary, and two dashed lines above and below it representing the margins. A vertical double-headed arrow spans the distance between the two dashed lines, indicating the width of the margin. To the right of each line is a corresponding equation:

$$\begin{aligned} \text{Top dashed line: } \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_+ &= M \\ \text{Solid line: } \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i &= 0 \\ \text{Bottom dashed line: } \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_- &= -M \end{aligned}$$

Linear Support Vector Machine

- We want to **maximize** the width of the band.



The diagram shows a 2D coordinate system with a solid horizontal line representing the decision boundary. Two dashed horizontal lines are parallel to the decision boundary, one above and one below, representing the margins. A vertical double-headed arrow indicates the distance between these two dashed lines, which is the width of the band. A curved arrow points from this vertical distance to the equation $\beta^T(\mathbf{x}_+ - \mathbf{x}_-) = 2M$.

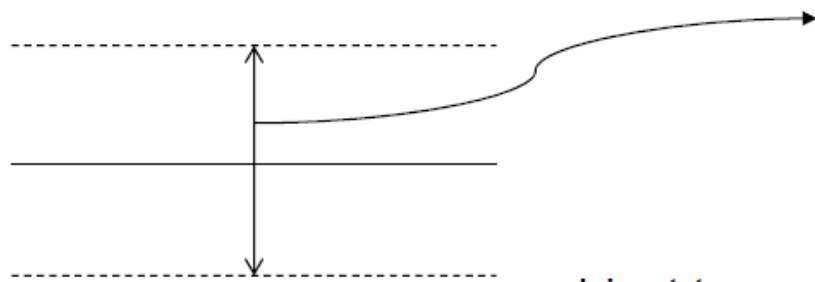
$$\beta^T(\mathbf{x}_+ - \mathbf{x}_-) = 2M$$
$$\max_{\beta_0, \beta} M$$

subject to $y_i(\beta_0 + \beta^T \mathbf{x}_i) \geq M$, for $i = 1, \dots, n$

$$\|\beta\| = 1$$

Linear Support Vector Machine

- We want to **maximize** the width of the band.



The diagram shows a 2D coordinate system with a solid horizontal line representing the decision boundary. Two dashed horizontal lines are parallel to the decision boundary, one above and one below, representing the margins. A vertical double-headed arrow indicates the distance between the decision boundary and the upper margin. A curved arrow points from this distance to the equation $\frac{\beta^T}{\|\beta\|}(\mathbf{x}_+ - \mathbf{x}_-) = \frac{2M}{\|\beta\|}$.

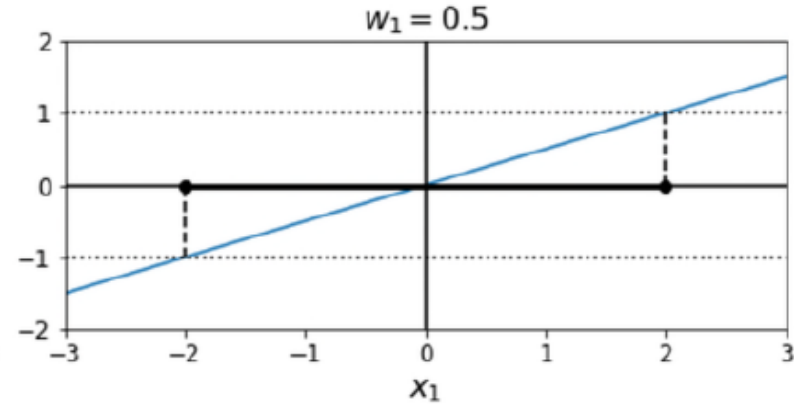
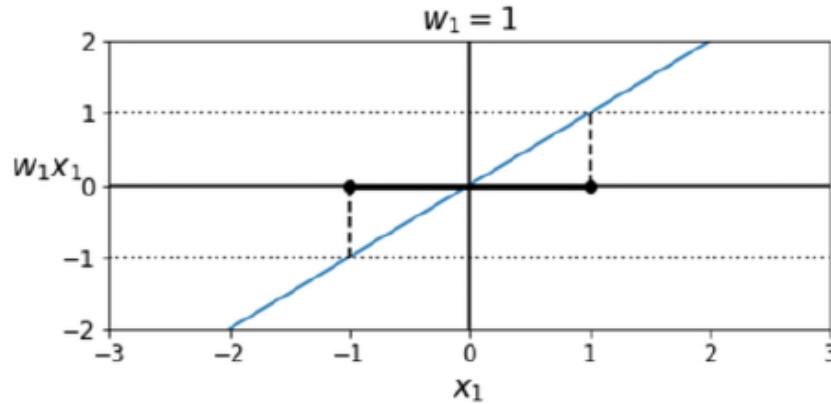
$$\frac{\beta^T}{\|\beta\|}(\mathbf{x}_+ - \mathbf{x}_-) = \frac{2M}{\|\beta\|}$$
$$\max_{\beta_0, \beta} M$$

subject to $y_i(\beta_0 + \beta^T \mathbf{x}_i) \geq M, \text{ for } i = 1, \dots, n$

$$\|\beta\| = 1$$

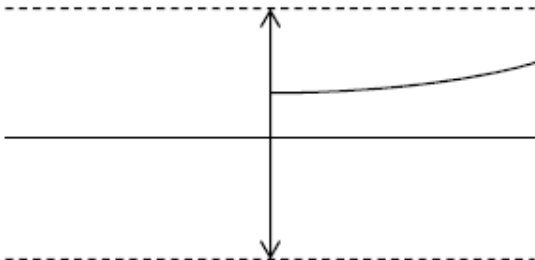
Linear Support Vector Machine

- A smaller weight vector results in a larger margin



Linear Support Vector Machine

- We want to **maximize** the width of the band.



The diagram shows a 2D coordinate system with a solid horizontal line representing the decision boundary. Above and below this line are two dashed horizontal lines representing the margins. A vertical double-headed arrow indicates the distance between these margins. A curved arrow points from this distance to the equation $\frac{\beta^T}{\|\beta\|}(\mathbf{x}_+ - \mathbf{x}_-) = \frac{2M}{\|\beta\|}$.

$$\frac{\beta^T}{\|\beta\|}(\mathbf{x}_+ - \mathbf{x}_-) = \frac{2M}{\|\beta\|}$$
$$\min_{\beta_0, \beta} \|\beta\|$$

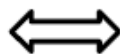
subject to $y_i(\beta_0 + \beta^T \mathbf{x}_i) \geq M$, for $i = 1, \dots, n$

$$M = ?$$

Linear Support Vector Machine

- We want to **maximize** the width of the band.

$$\max_{\beta_0, \beta} M$$



$$\min_{\beta_0, \beta} \|\beta\|^2$$

subject to $y_i(\beta_0 + \beta^T \mathbf{X}_i) \geq M$, for $i = 1, \dots, n$

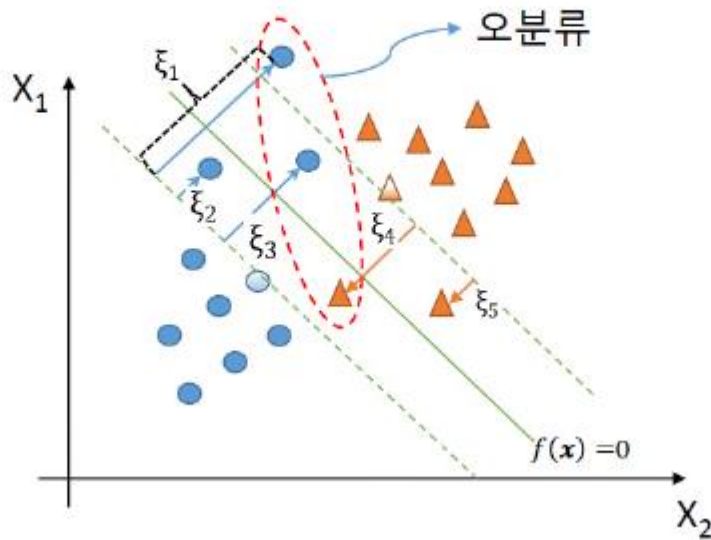
subject to $y_i(\beta_0 + \beta^T \mathbf{X}_i) \geq 1$, for $i = 1, \dots, n$

and

$$\|\beta\| = 1$$

Linear Support Vector Machine

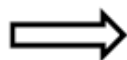
- If the data are not perfectly separable, no solution exists.



Linear Support Vector Machine

- Hard Margin Classifier

$$\min_{\beta_0, \boldsymbol{\beta}} ||\boldsymbol{\beta}'||^2$$



subject to $y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i) \geq 1$, for $i = 1, \dots, n$

- Soft Margin Classifier

$$\min_{\beta_0, \boldsymbol{\beta}} ||\boldsymbol{\beta}'||^2$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i) \geq 1 - \zeta_i$

and $\zeta_i \geq 0$,

and $\sum_i^n \zeta_i \leq \tilde{C}$, for $i = 1, \dots, n$

Linear Support Vector Machine

- Primal Problem

$$\min_{\beta_0, \boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i) \geq 1 - \zeta_i$

and $\zeta_i \geq 0$,

and $\sum_i \zeta_i \leq \tilde{C}$, for $i = 1, \dots, n$

- Dual Problem

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i) \geq 1 - \zeta_i$

and $\zeta_i \geq 0$, for $i = 1, \dots, n$

C is not a Lagrange multiplier

Linear Support Vector Machine

- Primal Problem

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i) \geq 1 - \zeta_i$

and $\zeta_i \geq 0$, for $i = 1, \dots, n$

- Dual Problem

$$\iff \min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i) \geq 1 - \zeta_i$

for $i = 1, \dots, n$

Linear Support Vector Machine

- Primal Problem

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i$$

subject to $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq 1 - \zeta_i, \quad \text{for } i = 1, \dots, n$

Linear Support Vector Machine

- Dual Problem

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

- Taking derivative w.r.t $\beta_0, \boldsymbol{\beta}, \zeta_i$
(Stationary)

Linear Support Vector Machine

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\begin{aligned} \text{(Stationary)} \left\{ \begin{array}{l} \frac{\partial}{\partial \beta_0} \mathcal{L}_p: \sum_i^n \alpha_i y_i = 0 \\ \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}_p: \boldsymbol{\beta} = \sum_i^n \alpha_i y_i \mathbf{x}_i \\ \frac{\partial}{\partial \zeta_i} \mathcal{L}_p: \alpha_i = C - \gamma_i \end{array} \right. \quad \begin{array}{l} \text{(Complementary} \\ \text{Slackness)} \end{array} \left\{ \begin{array}{l} \alpha_i [y_i f(\mathbf{x}_i) - (1 - \zeta_i)] = 0 \\ \gamma_i \zeta_i = 0 \end{array} \right. \end{aligned}$$

Linear Support Vector Machine

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\iff \max_{\alpha_i} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad \text{QP}$$

subject to $0 \leq \alpha_i \leq C$

and $\sum_i^n \alpha_i y_i = 0, \quad \text{for } i = 1, \dots, n$

Linear Support Vector Machine

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\Rightarrow \quad \hat{\boldsymbol{\beta}} = \sum_i^n \hat{\alpha}_i y_i \mathbf{x}_i$$

$$\widehat{\beta}_0 = y_i - \hat{\boldsymbol{\beta}}^T \mathbf{x}_k \quad \text{for any support vector } \mathbf{x}_k$$

$$\widehat{f(\mathbf{x}_i)} = \widehat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{x}_k$$

Kernel Support Vector Machine

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\iff \max_{\alpha_i} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to } 0 \leq \alpha_i \leq C$$

$$\text{and } \sum_i^n \alpha_i y_i = 0, \quad \text{for } i = 1, \dots, n$$

Kernel Support Vector Machine

- Kernel function

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

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)*

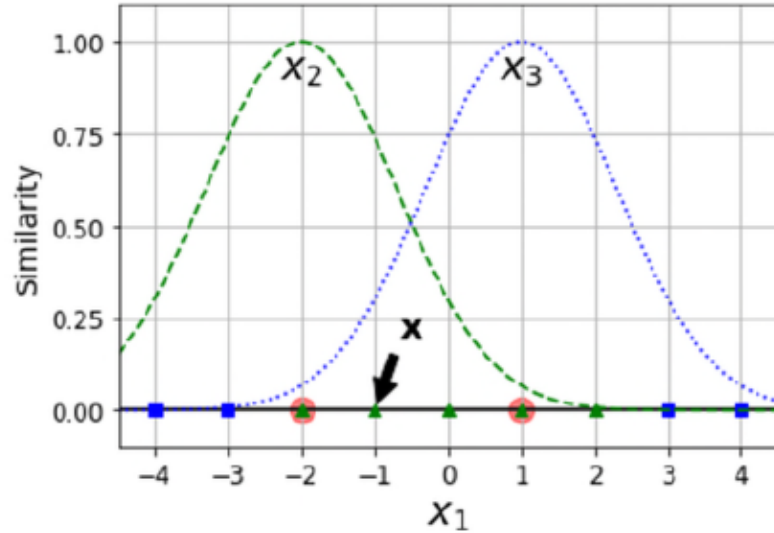
$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma + \gamma \mathbf{x}_i^T \mathbf{x}_j)^p$$

polynomial Kernel

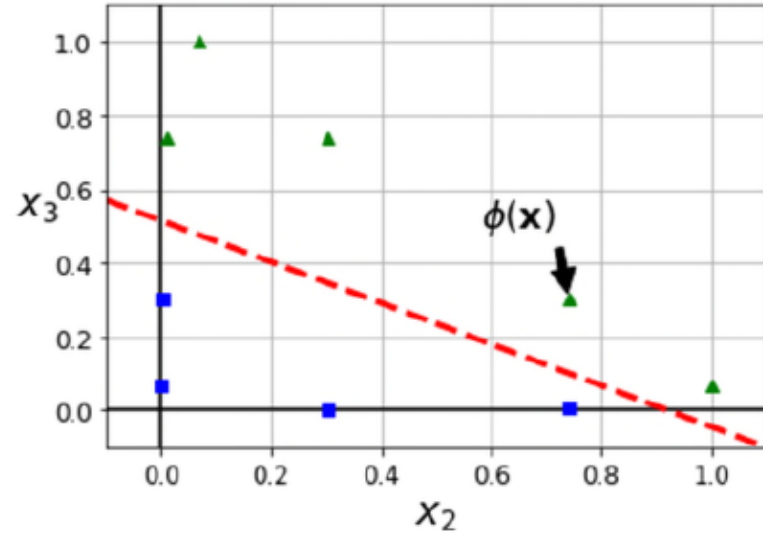
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(k_1 \mathbf{x}_i^T \mathbf{x}_j + k_2)$$

Sigmoid Kernel

Kernel Support Vector Machine

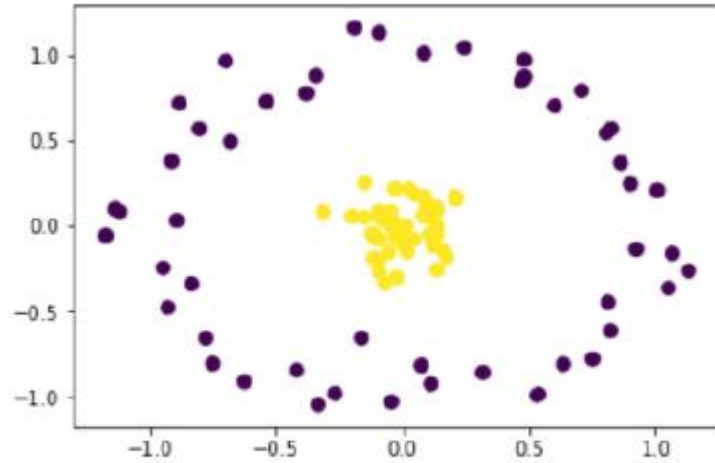


x_1

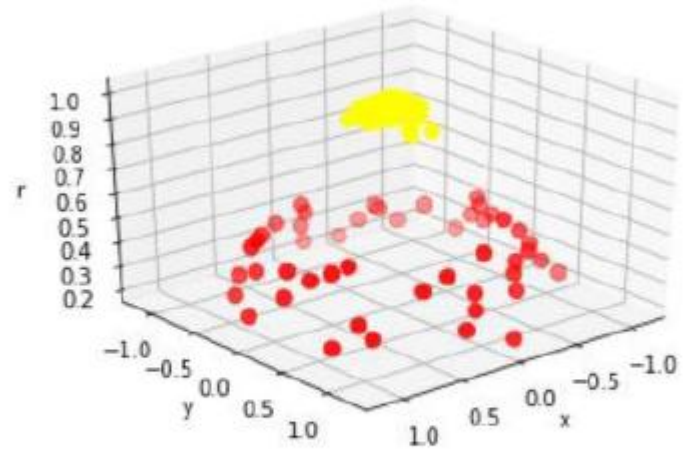


$$x_2 = \exp(-(x_1 - 2)^2)$$
$$x_3 = \exp(-(x_1 + 1)^2)$$

Kernel Support Vector Machine



(x_1, x_2)



$(x_1, x_2, \exp(-(x_1^2 + x_2^2)))$

Kernel Support Vector Machine

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\iff \max_{\alpha_i} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j \mathbf{h}(\mathbf{x}_i)^T \mathbf{h}(\mathbf{x}_j)$$

$$\text{subject to } 0 \leq \alpha_i \leq C$$

$$\text{and } \sum_i^n \alpha_i y_i = 0, \quad \text{for } i = 1, \dots, n$$

Kernel Trick

$$\min_{\beta_0, \boldsymbol{\beta}, \zeta_i} \|\boldsymbol{\beta}\|^2 + C \sum_i^n \zeta_i - \sum_i^n \gamma_i \zeta_i - \sum_i^n \alpha_i [y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - (1 - \zeta_i)]$$

$$\iff \max_{\alpha_i} \sum_i^n \alpha_i - \frac{1}{2} \sum_i^n \sum_j^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $0 \leq \alpha_i \leq C$

$$\text{and } \sum_i^n \alpha_i y_i = 0, \quad \text{for } i = 1, \dots, n$$

Kernel Trick

- 특성함수의 생성 어려움 + 고차원 확장시 차원의 저주 문제 발생.
- 2차 다항커널 : 입력변수 x_1 과 x_2 이고 i 번째 관측치와 j 번째 관측치일때,

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= (1 + \mathbf{x}_i^T \mathbf{x}_j)^2 \\ &= (1 + x_{i,1}x_{j,1} + x_{i,2}x_{j,2})^2 \\ &= 1 + 2x_{i,1}x_{j,1} + 2x_{i,2}x_{j,2} + (x_{i,1}x_{j,1})^2 + (x_{i,2}x_{j,2})^2 + 2x_{i,1}x_{j,1}x_{i,2}x_{j,2} \end{aligned} \quad (7.11)$$

- 이때 다음과 같이 정의하면,

$$h_1(x_1, x_2) = 1, \quad h_2(x_1, x_2) = \sqrt{2}x_1, \quad h_3(x_1, x_2) = \sqrt{2}x_2, \quad h_4(x_1, x_2) = x_1^2, \quad h_5(x_1, x_2) = x_2^2, \quad h_6(x_1, x_2) = \sqrt{2}x_1x_2$$

$$\mathbf{h}(x_1, x_2) = (h_1(x_1, x_2), h_2(x_1, x_2), \dots, h_6(x_1, x_2))^T$$

- 식 (7.11)은 $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^2 = \mathbf{h}(\mathbf{x}_i)^T \mathbf{h}(\mathbf{x}_j)$ 로 변형 가능.
- 특성함수를 정의하지 않고 커널 함수를 이용.
- 즉, $\hat{\beta}$ 이 $\mathbf{h}(\mathbf{x}_i)^T \mathbf{h}(\mathbf{x}_j)$ 의 형태이면, $K(\mathbf{x}_i, \mathbf{x}_j)$ 를 직접 이용하여 추정.

Kernel Trick

- 특성변수 x 로 부터 basis함수 $h(x)$ 로 차원을 증대시키면 커널 SVM 목적함수.

$$L_k = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j h(x_i)^T h(x_j) \quad (7.12)$$

- 선형 SVM 식 (7.11)은 $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}^T x = \hat{\beta}_0 + \sum_{i=1}^n \hat{\alpha}_i y_i x_i^T x$ 로 변형 가능.

- L_k 최소화한 모수 추정치를 $\hat{\beta}_0^*$ 와 $\hat{\beta}^*$ 라 할 때 커널 SVM의 예측치

$$\hat{f}(x) = \hat{\beta}_0^* + \sum_{i=1}^n \hat{\alpha}_i^* y_i h(x_i)^T h(x) \quad (7.13)$$

- 식(7.12)와 식(7.13) 모두 $h(x_i)^T h(x_j)$ 의 형태임.
- 식(7.12)에 $h(x_i)^T h(x_j)$ 대신 커널 함수 $K(x_i, x)$ 를 대체하여 $\hat{\beta}_0^*$ 와 $\hat{\beta}^*$ 를 추정.
- 식(7.13)도 $h(x_i)^T h(x_j)$ 를 이용하여 동일한 커널 SVM을 구함.

Support Vector Regression

