데이터마이닝
(**D**ata **M**ining)

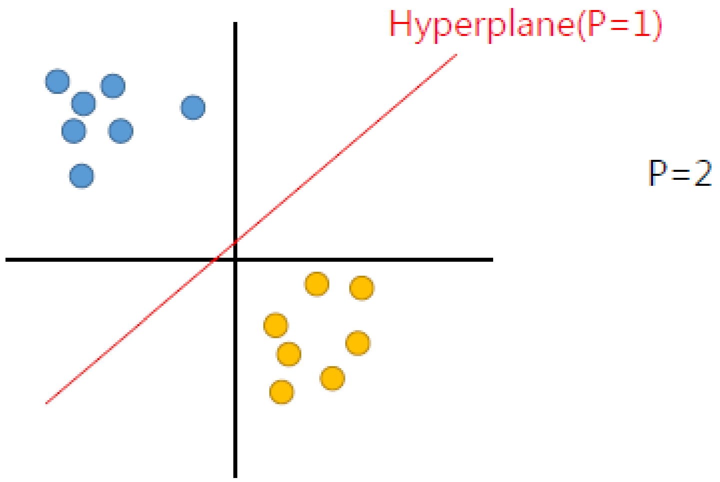
Chapter 2 Support Vector Machine

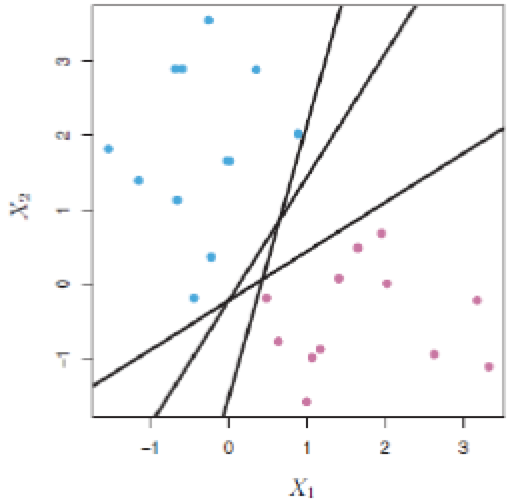
Classification Method

- 선형 판별분석
- 로지스틱 회귀분석
- 분류나무 (Bagging, Boosting)
- Support Vector Machines

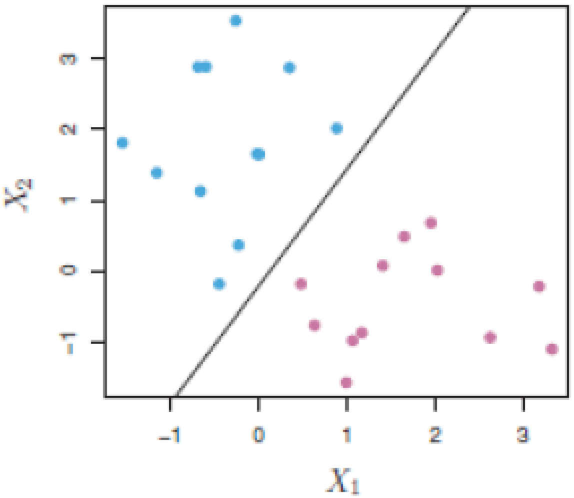
❖ Hyperplane?

: 초평면이란 P차원에서 Class를 구분하는 P-1차원의 Subspace





분류하는데 여러가지 경우의 수



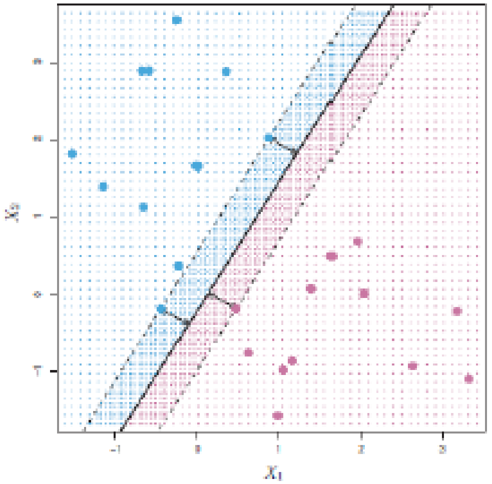
가장 Class를 잘 구분하는 경우 존재

마진(margin)

학습 데이터들 중에서 분류 경계에 가장 가까운 데이터로부터 분류경계까지의 거리

서포트벡터(support vector)

학습 데이터들 중에서 분류경계에 가장 가까운 곳에 위치한 데이터



"The furthest minimum distance observation"

$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$

$y_1, \ldots, y_n \in \{-1, 1\}$

일반화 오차를 작게 → 클래스간의 간격을 크게 → 마진을 최대

→ "최대 마진 분류기", "SVM"

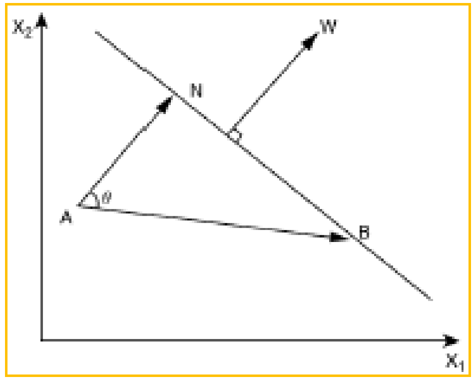
n차원의 공간상에서 초평면을 이루는 단위(normal) 법선 벡터가 $\mathbf{w} = (w_1, w_2, \cdots, w_n)$이고 중심에서 이 초평면까지의 거리가 b라고 할 때, 초평면상의 한 점이 $\mathbf{x} = (x_1, x_2, \cdots, x_n)$인 초평면의 방정식은 다음과 같다.

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n + b = 0$$

그리고 n차원 공간상의 임의의 점 $\mathbf{A} = (a_1, a_2, \cdots, a_n)$에서 이 초평면에 이르는 최소 거리($d$)는

다음과 같이 구한다.

$$d = \frac{|w_1a_1 + w_2a_2 + \cdots + w_na_n + b|}{\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}}$$



$\mathbf{A}=(x_{1A},\ x_{2A})$, $\mathbf{B}=(x_{1B},\ x_{2B})$인 경우, 점 $\mathbf{A}$에서 직선까지의 거리 $\|\mathbf{AN}\|$은 다음과 같이 표현된다.

$$\|\mathbf{AN}\|=\|\mathbf{AB}\|\cos\theta=\|\mathbf{AB}\|\frac{<\mathbf{AB},\ \mathbf{w}>}{\|\mathbf{AB}\|\|\mathbf{w}\|}$$

$$=\frac{(x_{1A}-x_{1B},\ x_{2A}-x_{2B})^T\cdot(w_1,\ w_2)}{\|\mathbf{w}\|}$$

$$=\frac{\mathbf{w}^T\mathbf{A}-\mathbf{w}^T\mathbf{B}}{\|\mathbf{w}\|}$$

$\mathbf{B}$는 직선상의 점이고 $\mathbf{w}^T\mathbf{B}=w_1x_{1B}+w_2x_{2B}+\ldots+w_nx_{nB}=-b$이므로, 다음과 같이 표현된다.

$$\|\mathbf{AN}\|=\frac{\mathbf{w}^T\mathbf{A}+b}{\|\mathbf{w}\|}$$

$\mathbf{w}$가 단위 법선 벡터면 $\|\mathbf{w}\|=1$이므로, 다음과 같이 표현된다.

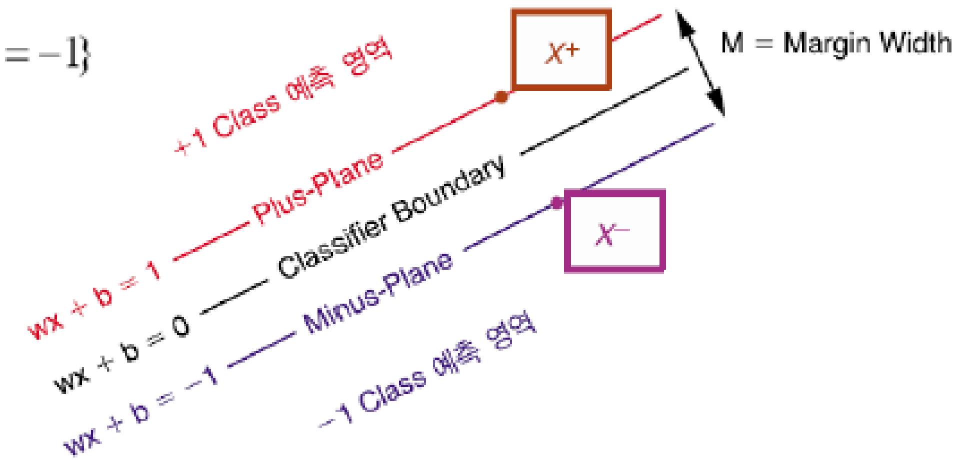
$$\|\mathbf{AN}\|=\mathbf{w}^T\mathbf{A}+b$$

- SVM은 입력이 m차원일 경우를 포함하여 최적 분류초평면인 결정경계와 마진을 최대화하는 최적 파라미터($\mathbf{w}$, b)를 찾아냄

$$d = \frac{\mathbf{w}^T\mathbf{A} + b}{\|\mathbf{w}\|}$$

$$\text{Plus-Plane} = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} + b = +1\}$$

$$\text{Minus-Plane} = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} + b = -1\}$$



$\mathbf{w}^T\mathbf{x}+b \geq 1$이면, $+1$

$\mathbf{w}^T\mathbf{x}+b \leq -1$이면, -1

$-1 < \mathbf{w}^T\mathbf{x}+b < 1$이면, 블랙홀 영역으로 결정한다.

$\mathbf{x}^-$를 Minus-Plane상의 어떤 점이라고 하고, $\mathbf{x}^+$를 $\mathbf{x}^-$와 가장 가까운 Plus-Plane상의 점이라고 하여 λ를 적절히 선택하면, $\mathbf{x}^+ = \mathbf{x}^- + \lambda\mathbf{w}$이 성립한다. 이것은 $\mathbf{x}^-$에서부터 $\mathbf{x}^+$까지의 선은 평면에 수직이기 때문에 $\mathbf{x}^+$에서 $\mathbf{w}$ 방향으로 얼마간의 거리가 떨어진 곳에 $\mathbf{x}^-$가 위치하기 때문이다.

$$\mathbf{w}\cdot\mathbf{x}^+ + b = +1$$

$$\mathbf{w}\cdot\mathbf{x}^- + b = -1$$

$$\mathbf{x}^+ = \mathbf{x}^- + \lambda\mathbf{w}$$

$$\left|\mathbf{x}^+ - \mathbf{x}^-\right| = M$$

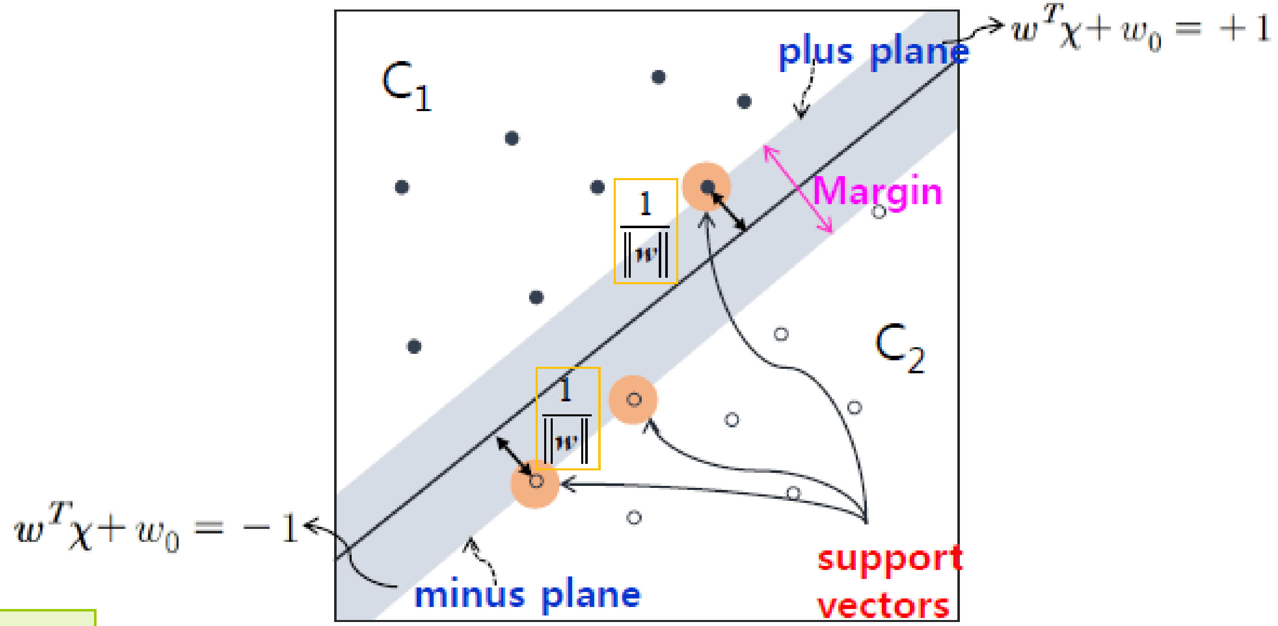
$$\mathbf{w} \cdot (\mathbf{x}^- + \lambda \mathbf{w}) + b = +1$$

$$\mathbf{w} \cdot \mathbf{x}^- + b + \lambda \mathbf{w}^T \mathbf{w} = +1$$

$$-1 + \lambda \mathbf{w}^T \mathbf{w} = +1$$

$$\therefore \lambda = \frac{2}{\mathbf{w}^T \mathbf{w}}$$

$$\mathbf{M} = \left|\mathbf{x}^+ - \mathbf{x}^-\right| = \left|\lambda \mathbf{w}\right| = \lambda \left|\mathbf{w}\right| = \lambda \sqrt{\mathbf{w}^T \mathbf{w}} = \frac{2\sqrt{\mathbf{w}^T \mathbf{w}}}{\mathbf{w}^T \mathbf{w}} = \frac{2}{\sqrt{\mathbf{w}^T \mathbf{w}}} = \frac{2}{\left\|\mathbf{w}\right\|}$$



마진 M

$$M = |\chi^+ - \chi^-| = \frac{1}{\| w \|}(w^T\chi^+ - w^T\chi^-) = \frac{2}{\| w \|}$$

$$\downarrow$$

마진의 최대화하려면, $\| w \|$의 최소화

학습 데이터 이용해
선형 분류기 파라미터 최적화

학습 데이터 집합 $\{(x_i, y_i)\}_{i=1\ldots N}$

$$\begin{cases} y_i = +1 & \text{if } x \in C_1 \\ y_i = -1 & \text{if } x \in C_2 \end{cases}$$

추정해야 할 파라미터 w, w_o가 만족해야 할 조건

$$\begin{cases} (w^T x_i + w_0) \ge +1 & \text{for } y_i = +1 \\ (w^T x_i + w_0) \le -1 & \text{for } y_i = -1 \end{cases} \longrightarrow \boxed{y_i(w^T x_i + w_0) - 1 \ge 0}$$

라그랑제 승수 $(\alpha_i \ge 0,\ i = 1, 2, \ldots, N)$

최소화할 목적함수 $\boxed{J(w) = \dfrac{\| w \|^2}{2}}$

"라그랑지안 함수"

$$\boxed{J(w, w_0, \alpha) = \frac{1}{2}\| w \|^2 - \sum_{i=1}^{N} \alpha_i \{ y_i (w^T x_i + w_0) - 1 \}}$$

파라미터 w, w_o에 대해 극소화하고, α_i에 대해 극대화

❖ KKT조건

1. 원 변수에 따른 라그랑지안 함수의 기울기는 0이 되어야 한다. $\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0$

2. 원 제약식은 다음 조건을 만족해야 한다. $\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0} = 0$

3. 부등제약식의 대한 라그랑제 승수는 다음 조건을 만족하여하 한다. $\alpha_i \geq 0 (i = 1, ..., N)$

4. 라그랑제 승수와 제약식의 곱은 0이 되어야 한다. $\alpha_i \{ y_i (\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 \} = 0$

$$J(\boldsymbol{w}, w_0, \alpha) = \frac{1}{2} \| \boldsymbol{w} \|^2 - \sum_{i=1}^{N} \alpha_i \{ y_i (\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 \}$$

$$\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i = 0 \quad \dashrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\rightarrow J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$J(\boldsymbol{w}, w_0, \alpha) = \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i - w_0 \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i$$

$$\boldsymbol{w}^T \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$의 dual problem
: $J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$ 최적화하는 대신,
$Q(\boldsymbol{\alpha})$의 최적화

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \, (i = 1, \ldots, N)$$

$Q(\boldsymbol{\alpha})$는 $\boldsymbol{\alpha}_i$에 대한 이차함수:
quadratic 프로그래밍 이용하여 간단히 해($\boldsymbol{\alpha}_i$) 구할 수 있음.
→ $\boldsymbol{w}$, w_0 추정

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

$$\hat{w}_0 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{\boldsymbol{w}}^T \boldsymbol{x}_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i\right)$$

$$J(\boldsymbol{w}, w_0, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N}\alpha_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\}$$

$$\begin{cases} (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq +1 & \text{for } y_i = +1 \\ (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \leq -1 & \text{for } y_i = -1 \end{cases}$$

대부분 데이터는 결정경계와 떨어져 있으므로, 조건식 $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 \geq 0$ 을 만족

→ $J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$을 최대화하는 음이 아닌 $\boldsymbol{\alpha}_i$ 는 0뿐임.

대부분의 학습 데이터에 대응되는 라그랑제 승수 $\hat{\alpha}_i$는 0

오직 $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 = 0$인 경우만 $\hat{\alpha}_i$가 0이 아닌 값

→ 서포트벡터 데이터의 $\hat{\alpha}_i \neq 0$

분류를 위해 저장할 데이터의 개수와 계산량의 현격한 감소

결정함수

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=1}^{n} w_i x_i + w_0 = 0$$

$$\hat{w} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

$$\widehat{w_0} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{w}^T x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i\right)$$

대입하면,

$$f(\boldsymbol{x}) = \mathrm{sign}\,(g(\boldsymbol{x})) = \mathrm{sign}(\hat{\boldsymbol{w}}^T\boldsymbol{x} + \hat{w}_0)$$

$$= \mathrm{sign}\left(\sum_{i=1}^{N} \widehat{\alpha_i} y_i \boldsymbol{x}_i^T \boldsymbol{x}_i + \widehat{w_0}\right)$$

$f(x) = 1 \rightarrow C_1$

$f(x) = -1 \rightarrow C_2$

① N개의 입출력 쌍으로 이루어진 학습 데이터 집합 $X=\{(\boldsymbol{x}_i, y_i)\}_{i=1\ldots N}$을 준비한다. 이때 목표 출력값은 $y_i \in \{-1,1\}$ $(i=1,\ldots,N)$을 만족한다.

② 다음과 같은 과정을 통해 SVM을 학습한다.

②-1. 학습 데이터를 이용하여 파라미터 추정을 위한 목적함수 $Q(\boldsymbol{\alpha})$를 정의한다.

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \ , \ \sum_{i=1}^{N} \alpha_i y_i = 0, \ \alpha_i \geq 0 (i=1,\ldots,N)$$

②-2. 주어진 조건을 만족하면서 $Q(\boldsymbol{\alpha})$를 최소화하는 추정치 $\hat{\alpha}_i$를 이차계획법에 의해 찾는다.

②-3. $\hat{\alpha}_i \neq 0$이 되는 서포트벡터를 찾아 집합 $X_S = \{\boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0\}$를 생성한다.

②-4. $\hat{\alpha}_i$와 서포트벡터를 이용하여 $\hat{w}_0$를 계산한다.

$$\widehat{w_0} = \frac{1}{N_S} \sum_{\boldsymbol{x}_i \in X_S} \left(y_i - \sum_{\boldsymbol{x}_j \in X_S} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i \right)$$

이때 N_S는 집합 X_S의 원소의 수이다.

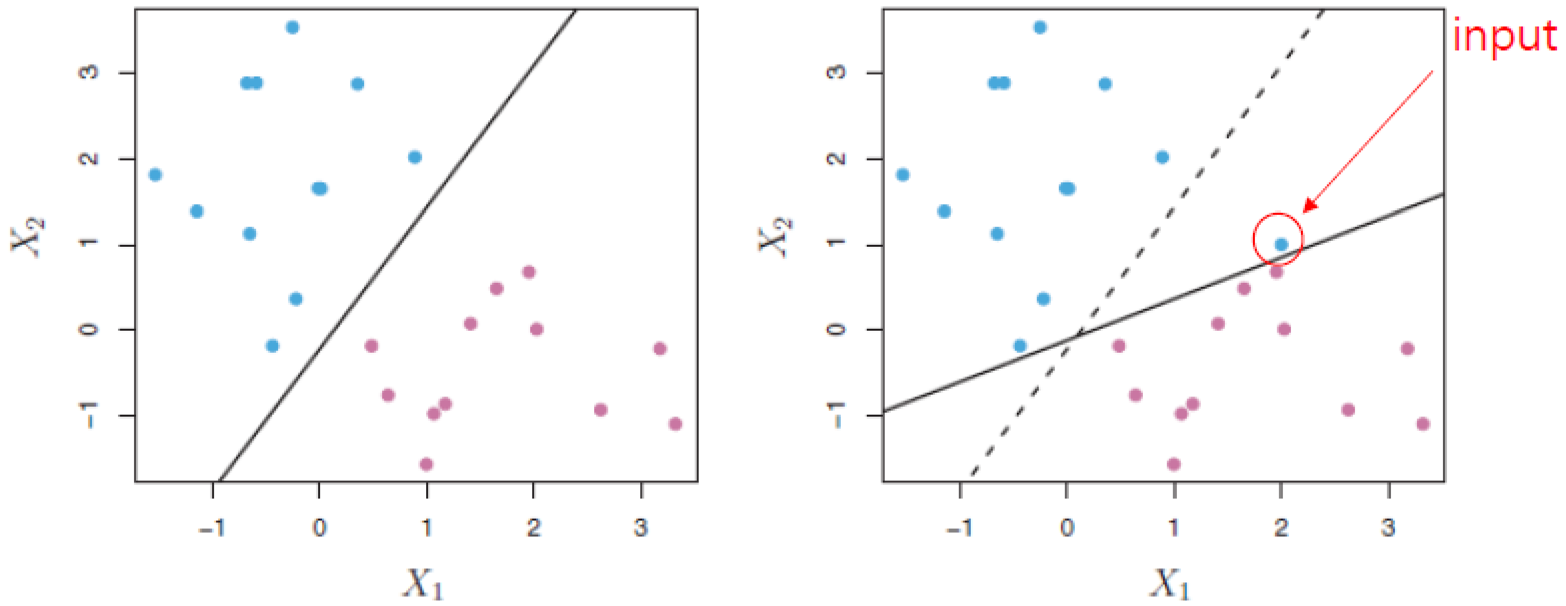
②-5. 서포트벡터 집합 $X_S = \left\{ \boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0 \right\}$와 파라미터 벡터 $\hat{\boldsymbol{\alpha}}$, 그리고 $\hat{w}_0$를 저장해 둔다.

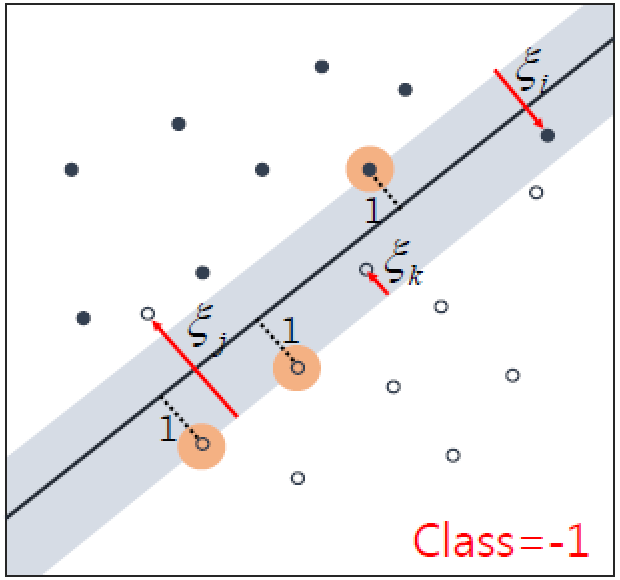
③ 새로운 데이터 $\boldsymbol{x}$가 주어지면, 저장해둔 서포트벡터와 파라미터를 이용하여 다음 판별함수로 분류를 수행한다.

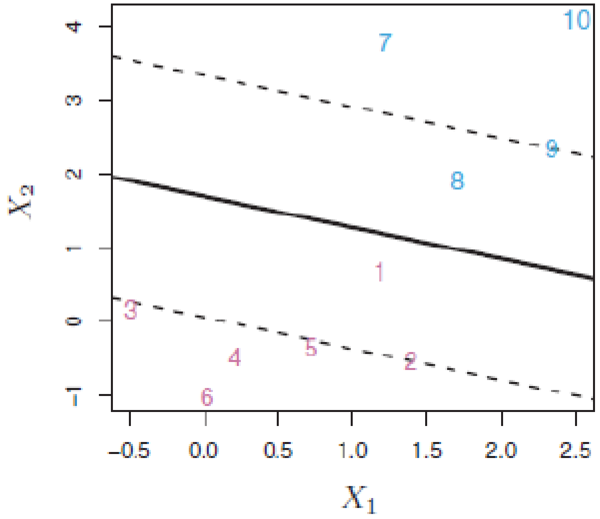
$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{\boldsymbol{x}_i \in X_S} \hat{\alpha}_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \widehat{w_0} \right)$$

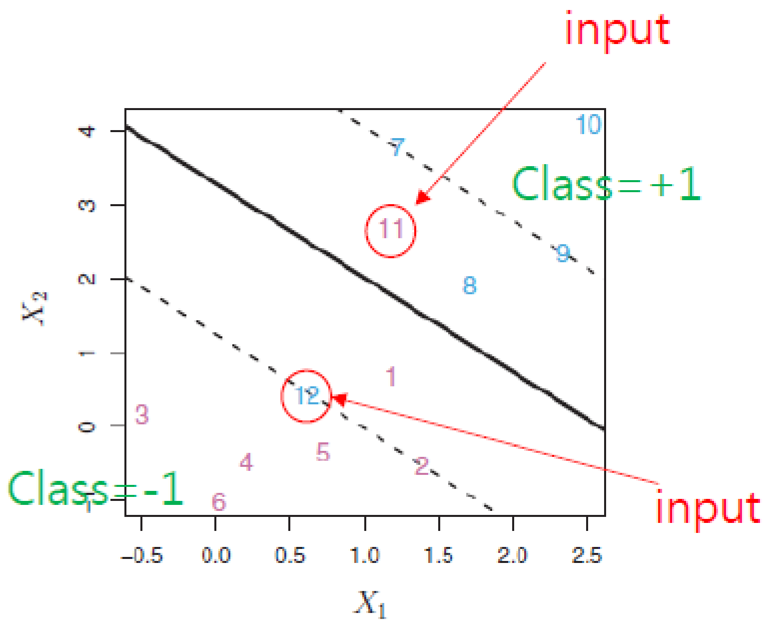
한계점1

❖ **최대 마진 분류기의 한계점**









$$\begin{cases}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \ge +1-\xi_i & \text{for } y_i = +1\\ (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \le -1+\xi_i & \text{for } y_i = -1\end{cases}$$

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \ge 1-\xi_i \quad (i = 1, \ldots, N)$$

: 오분류의 허용도 결정.
c 커지면, ξ값 커지는 것을 막으므로 오분류 오차 적어짐.

$$J(\boldsymbol{w}, \xi) = \frac{1}{2} \| \boldsymbol{w} \|^2 + c\sum_{i=1}^{N} \xi_i$$

목적함수 최소화

$$y_i(\boldsymbol{w}^T \boldsymbol{x_i} + w_0) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

라그랑지안 함수

$$J(\boldsymbol{w}, w_0, \boldsymbol{\alpha}, \xi, \beta) = \frac{1}{2} \| \boldsymbol{w} \|^2 + c\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \{ y_i(\boldsymbol{w}^T \boldsymbol{x_i} + w_0) - 1 + \xi_i \} - \sum_{i=1}^{N} \beta_i \xi_i$$

w, w_0에 대해 미분

슬랙변수를 양으로 유지하기 위한
새로운 라그랑제 승수 β_i 추가

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}^T \boldsymbol{x_j} \qquad \sum_{i=1}^{N} \alpha_i y_i = 0, \ \boxed{0 \leq \alpha_i < c} \ (i = 1, \ldots, N)$$

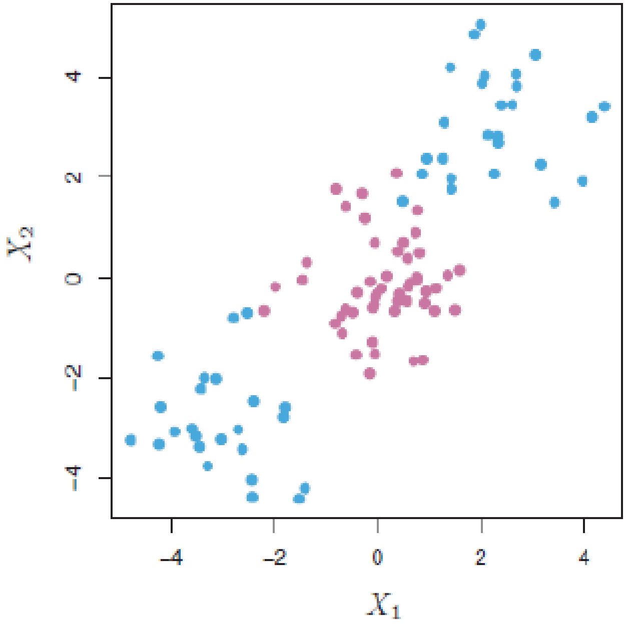
$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i}$$

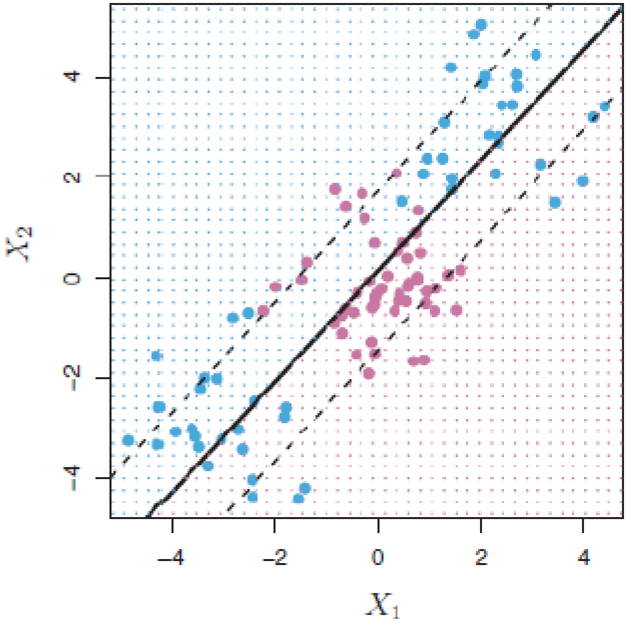
$$\hat{w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{\boldsymbol{w}}^T \boldsymbol{x_i}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{N} \alpha_j y_j \boldsymbol{x_j}^T \boldsymbol{x_i} \right)$$

α_i가 정해지면 $\boldsymbol{w}, w_0$의 값은 슬랙변수가 없는 경우와 완전히 동일

한계점2

❖ 최대 마진 분류기의 한계점



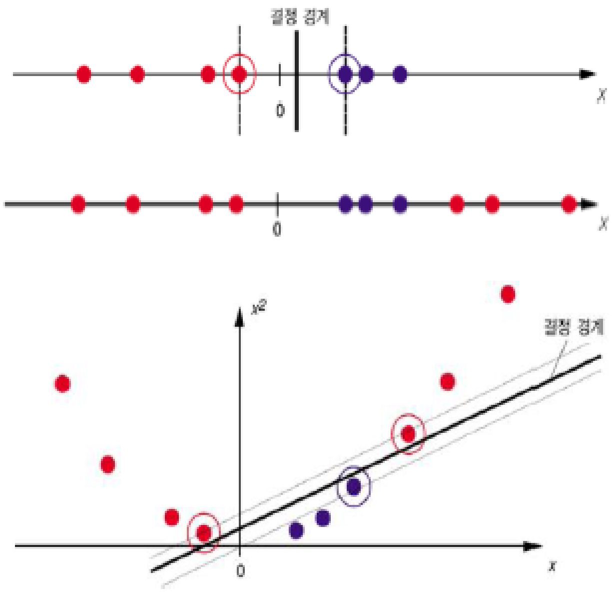


$$X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2.$$

다항식이나 비선형 함수를

표현하는 변수를 사용한다면?

❖ **Solution**



선형 분리 가능한 경우
선형 분리 불가능한 경우

2D 데이터로 사상