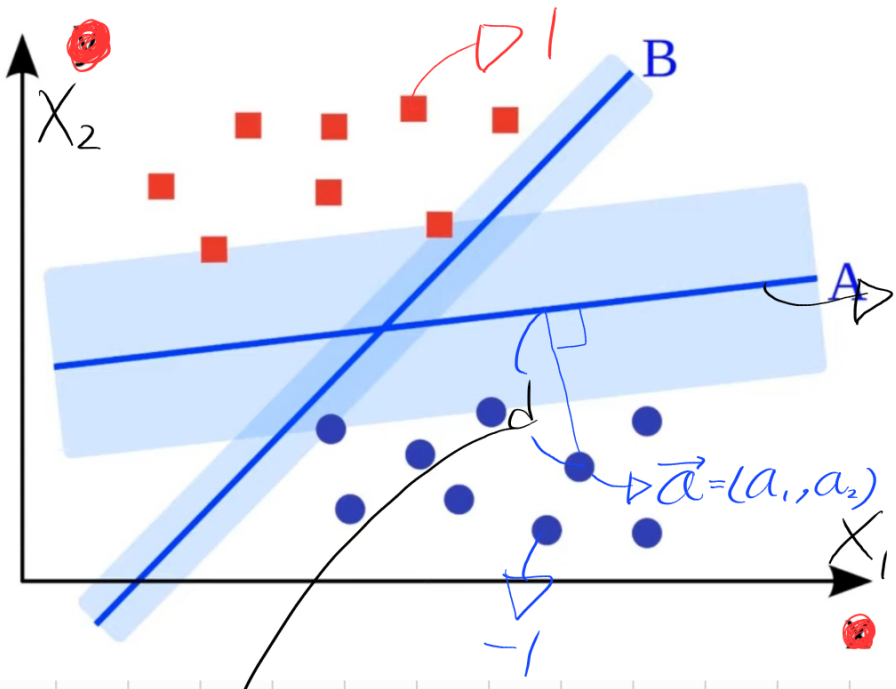
4. SVM

연세대학교 디지털애널리틱스 융합학과



$W_1x_1 + W_2x_2 + b = 0$

$\vec{w} = (W_1, W_2)$

$$d = \frac{|\vec{w}\cdot\vec{a} + b|}{\|\vec{w}\|_2} = \frac{|W_1a_1 + W_2a_2 + b|}{\sqrt{W_1^2 + W_2^2}}$$

점과 초평면 ($\vec{w}\cdot\vec{x} + b = 0$) 사이의 거리

데이터 $\vec{x}$ 가 d 차원이면

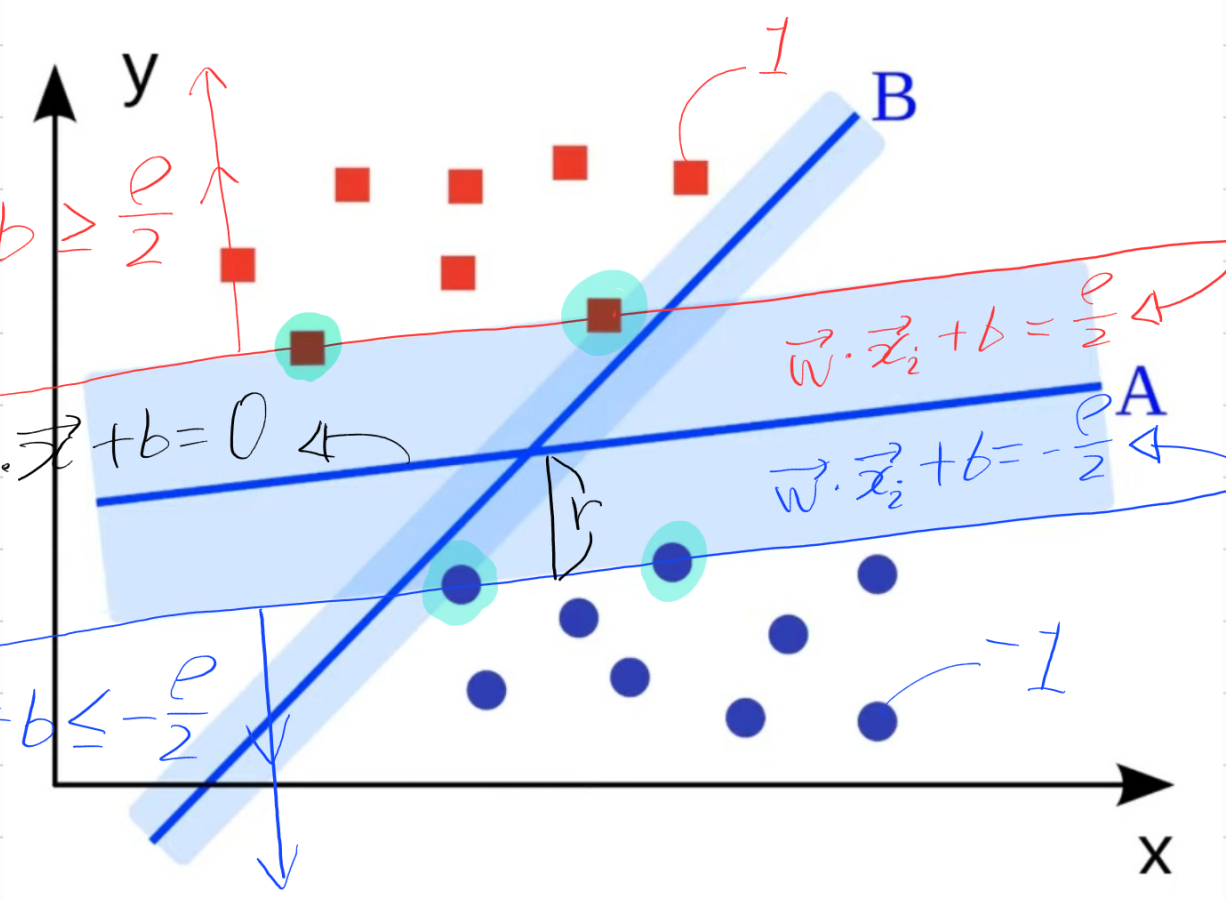
초평면은 몇 차원? → d-1 차원

ex) $\vec{x} = (x_1, x_2) \in \mathbb{R}^2$ → 2차원

초평면 → $W_1x_1 + W_2x_2 + b = 0$

$$\rightarrow x_1 = \frac{-W_2x_2 - b}{W_1}$$

→ 2-1 차원 초평면



$$y(\vec{w}\cdot\vec{x}'+b)=\frac{\rho}{2} \quad at \ x': S.V$$

초평면 $\vec{w}\cdot\vec{x}+b=0$ 과 S.V x' 사이의 거리 r?

$$r = \frac{|\vec{w}\cdot\vec{x}'+b|}{\|\vec{w}\|_2} = \frac{y(\vec{w}\cdot\vec{x}'+b)}{\|\vec{w}\|_2} = \frac{\frac{\rho}{2}}{\|\vec{w}\|_2}$$

$$= \frac{\frac{\rho}{2}\times\frac{2}{\rho}}{\|\vec{w}\|_2\times\frac{2}{\rho}} = \frac{1}{\|\vec{w}\times\frac{2}{\rho}\|_2} = \frac{1}{\|\vec{w}'\|_2}$$

margin r을 최대화 하는 $\vec{w}$와 b를 찾는 것이 SVM의 목표 ▷ 초평면 $\vec{w}\cdot\vec{x}+b=0$ 을 찾는 것

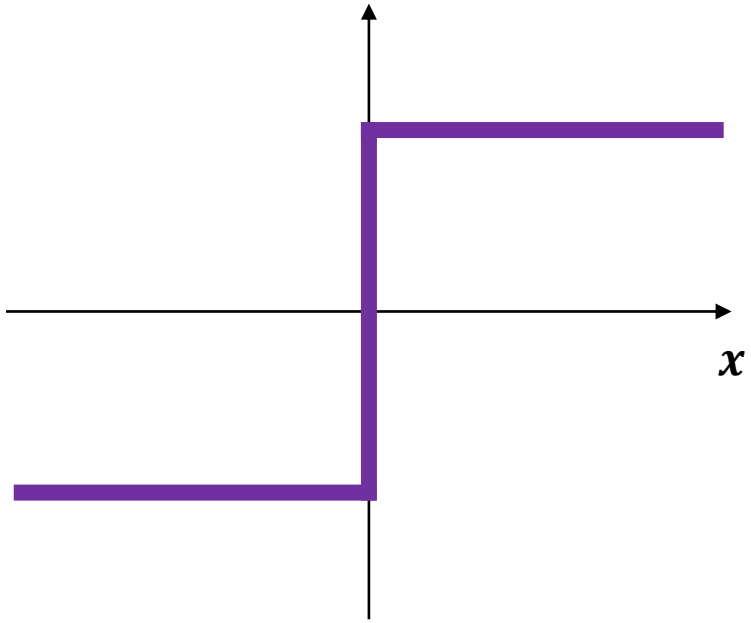
$$\max_{\vec{w},b} \frac{1}{\|\vec{w}\|_2} \quad s.t \ y_i(\vec{w}\cdot\vec{x}_i+b) \geq 1$$

What is Support Vector Machine?

Perceptron

SVM

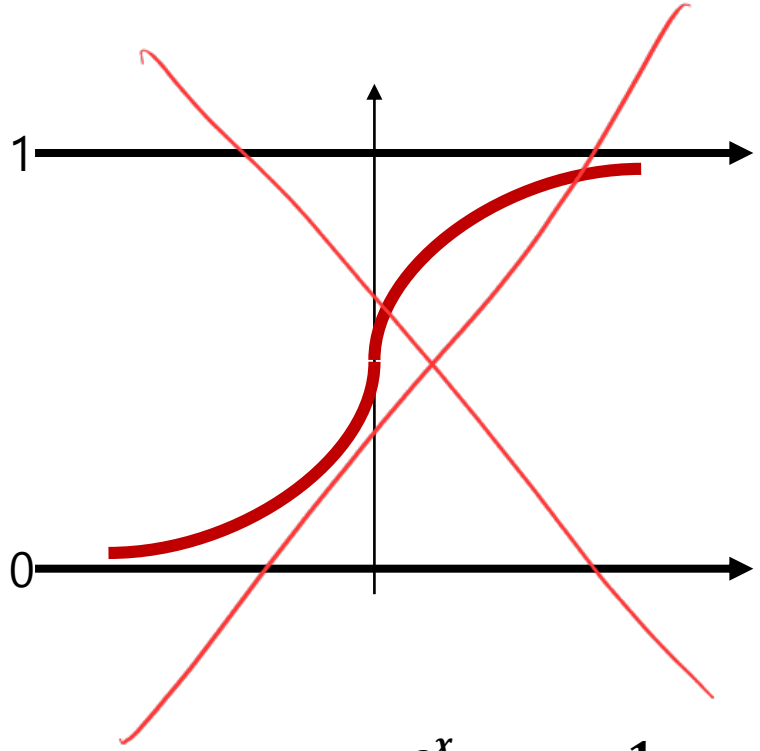
Sign 함수 :



$$sign(x) = \begin{cases} 1 & for\ x > 0 \\ -1 & else \end{cases}$$

Logistic Regression

Sigmoid 함수 :



$$sigmoid(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Hard Margin SVM

Primal

$$\min_{\vec{w},b,\xi_i} \frac{1}{2}\|\vec{w}\|_2^2$$

$$s.t \quad y_i(\vec{w}\cdot\vec{x}_i+b) \geq 1$$

KKT 조건을 만족하면

Dual

$$\Rightarrow \max_{\vec{\alpha},\vec{\beta}} \left[\min_{\vec{w},b,\xi_i} \frac{1}{2}\|\vec{w}\|_2^2 - \sum_{i=1}^{n}\alpha_i(y_i(\vec{w}\cdot\vec{x}_i+b)-1) \right]$$

$g(\vec{w},b)$ — Primal

대입

KKT

① $\nabla_{\vec{w}} g(\vec{w},b)=0 \rightarrow \vec{w}=\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i$

② $\nabla_b g(\vec{w},b)=0 \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$

③ $\alpha_i(y_i(\vec{w}\cdot\vec{x}_i+b)-1)=0$

④ $y_i(\vec{w}\cdot\vec{x}_i+b)-1 \geq 0$

⑤ $\alpha_i \geq 0$

Dual

$$\Rightarrow \max_{\vec{\alpha},\vec{\beta}} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \right]$$

$$s.t \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0$$

Solution

$$\Rightarrow f(\vec{x}_{new}) = sign(\vec{w}\cdot\vec{x}_{new}+b)$$

최종 Model

$$= sign\left(\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i^T \vec{x}_{new} + b\right)$$

KKT 조건에 따른 SVM의 특징

③ $\alpha_i(y_i(\vec{w}\cdot\vec{x}_i+b)-1)=0$

★ 둘 중 하나만 0

⑤ $\alpha_i \geq 0$

① $\vec{w}=\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i$

case 1 : $\alpha_i=0 \xrightarrow{①} y_i(\vec{w}\cdot\vec{x}_i+b)-1=0$

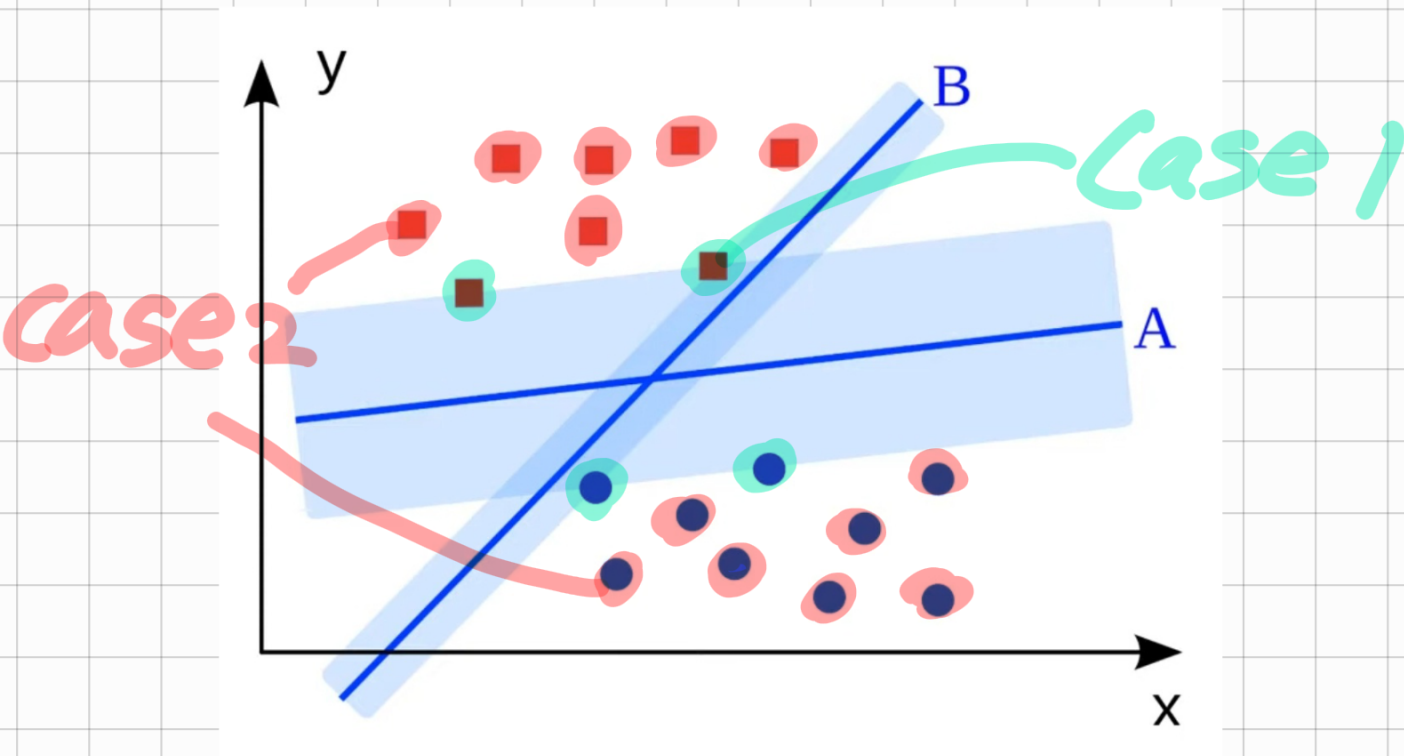
∴ S.V 이다

case 2 : $\alpha_i>0 \xrightarrow{①} y_i(\vec{w}\cdot\vec{x}_i+b)-1\neq 0$

∴ S.V 이 아니다

$$f(\vec{x}_{new}) = sign(\vec{w}\cdot\vec{x}_{new}+b)$$
$$= sign\left(\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i^{T}\vec{x}_{new}+b\right)$$

③

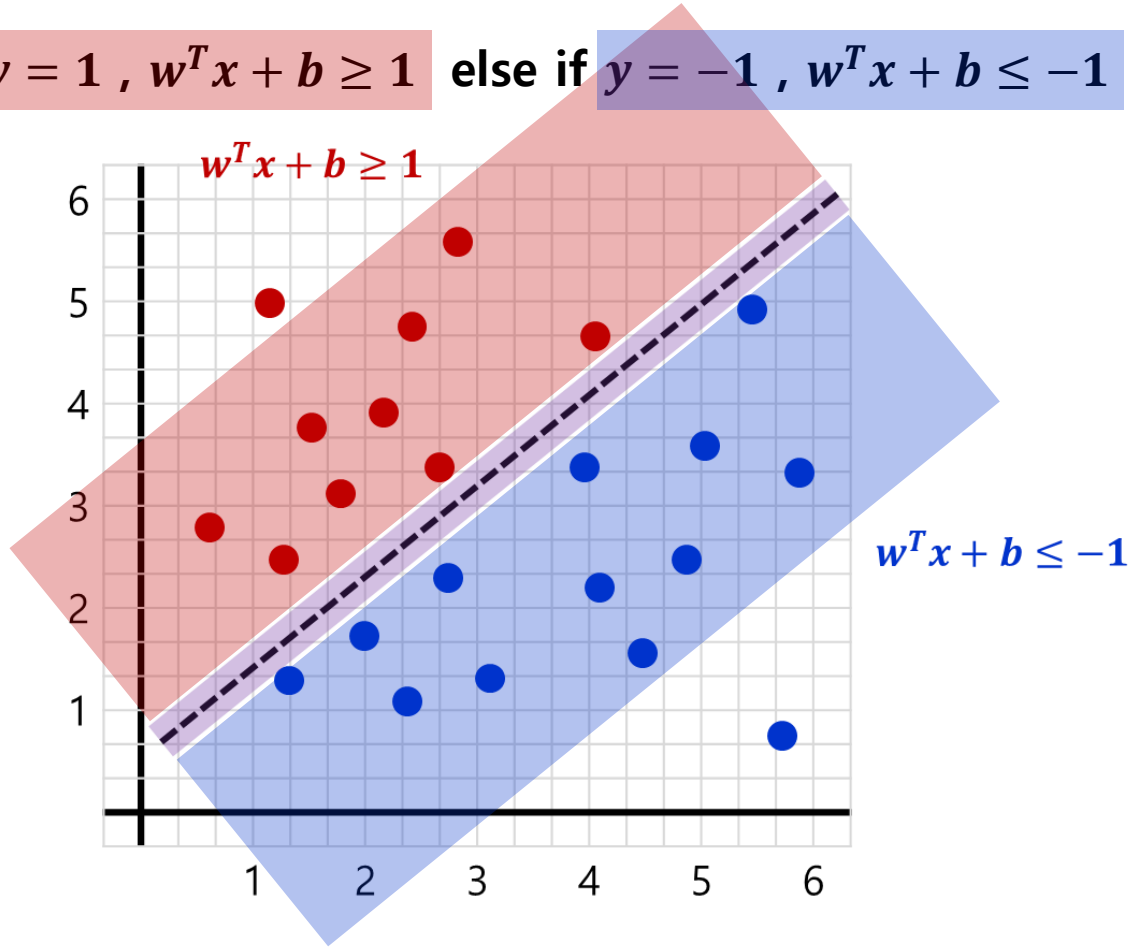


Support Vector Machine Concepts

Gradient Descent

Hinge loss $= \max\{0,\ 1 - y \times \hat{y}\}$

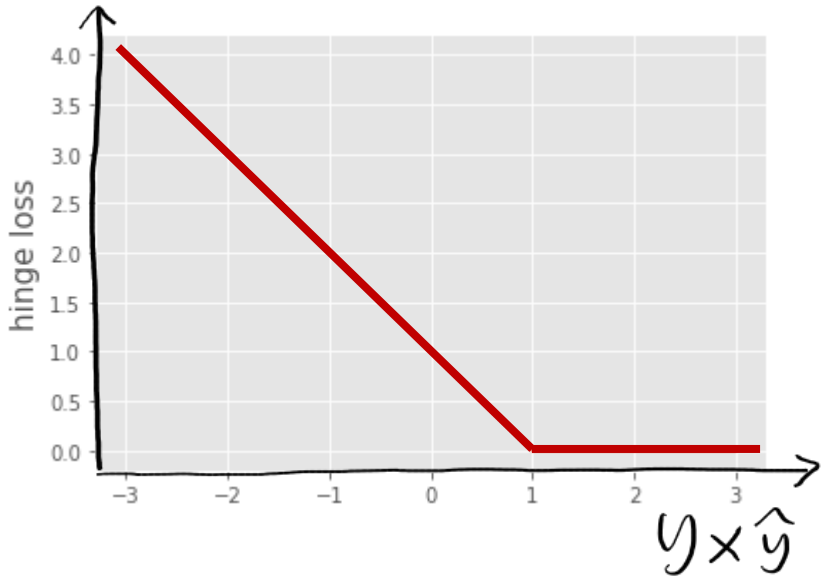
If $y = 1$, $w^T x + b \geq 1$ **else if** $y = -1$, $w^T x + b \leq -1$



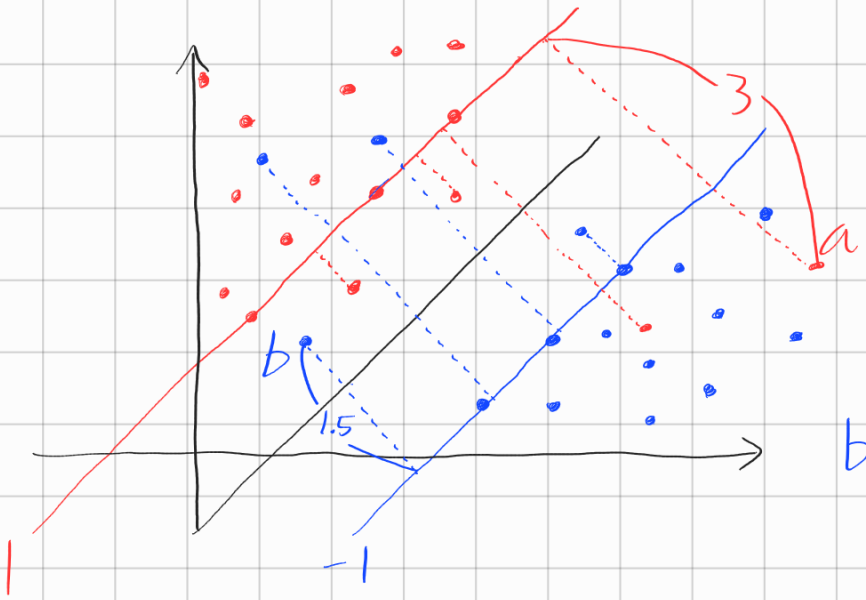
If y **and** $\hat{y}$ **are equal , then**

$$y \times \hat{y} = y(w^T x + b) \geq 1$$

$y \times \hat{y} < 1$ **means any observations exist in margin area**



$$\text{Hinge Loss} = \max\{0, 1 - y \times \hat{y}\}$$



a: $\max\{0, 1 - y \times \hat{y}\}$

$= \max\{0, 1 - y(\vec{w} \cdot \vec{x} + b)\}$

$= \max\{0, 1 - 1 \cdot (-2)\}$

$= \max\{0, 3\} = 3$

b: $\max\{0, 1 - y \times \hat{y}\}$

$= \max\{0, 1 - y(\vec{w} \cdot \vec{x} + b)\}$

$= \max\{0, 1 - (-1)(0.5)\}$

$= \max\{0, 1.5\} = 1.5$

즉, Hinge loss는 ($\vec{x}$ 자신의 참값의 S.V 초평면으로 부터 $\vec{x}$ 자신의 점 까지의 거리 -1)에 비례한다.

(단, 정답인 영역으로 $\hat{y}$을 예측 했을경우는 0이 된다)

$$\text{Hinge Loss}_2 = \max\{0, 1 - y \times \hat{y}\}$$

$$= \max\{0, 1 - y(\vec{w} \cdot \vec{x}_2 + b)\}$$

$$= \varepsilon_2$$

Soft Margin SVM

Primal $\min_{\vec{w}, b, \xi_i} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$

Hard Margin SVM과 다른 부분

$s.t \quad y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$

KKT 만족

$g(\vec{w}, b, \xi_i)$

Dual $\Rightarrow \max_{\vec{\alpha}, \vec{\beta}} \left[\min_{\vec{w}, b, \xi_i} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(\vec{w}\cdot\vec{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i \right]$

Primal

대입

KKT

① $\nabla_{\vec{w}} g(\vec{w}, b, \xi_i) = 0 \rightarrow \vec{w} = \sum_{i=1}^{n}\alpha_i y_i \vec{x}_i$

② $\nabla_b g(\vec{w}, b, \xi_i) = 0 \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$

③ $\nabla_{\xi_i} g(\vec{w}, b, \xi_i) = 0 \rightarrow C - \alpha_i - \beta_i = 0$

④ $\alpha_i(y_i(\vec{w}\cdot\vec{x}_i + b) - 1 + \xi_i) = 0, \quad \beta_i\xi_i = 0$

⑤ $y_i(\vec{w}\cdot\vec{x}_i + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0$

⑥ $\alpha_i \geq 0, \beta_i \geq 0$

$0 \leq \alpha_i \leq C$

Dual $\Rightarrow \max_{\vec{\alpha}, \vec{\beta}} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \right]$

$s.t \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$

Solution

$\Rightarrow f(\vec{x}_{new}) = sign(\vec{w}\cdot\vec{x}_{new} + b)$

최종 model $= sign\left(\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i^T \vec{x}_{new} + b\right)$

KKT 조건에 따른 SVM의 특징

④ $\alpha_i(y_i(\vec{w}\cdot\vec{x}_i+b)-1+\xi_i)=0$ ③ $C-\alpha_i-\beta_i=0$

둘 중하나만 0

④ $\beta_i\xi_i=0$

case 1 : $\alpha_i=0$ ③▷ $C=\beta_i$ ②▷ $\xi_i=0$ ①▷ $y_i(\vec{w}\cdot\vec{x}_i+b)-1+\xi_i\neq 0$ ▷ $y_i(\vec{w}\cdot\vec{x}_i+b)-1\neq 0$

∴ S.V가 아니다 and 오분류가 아니다

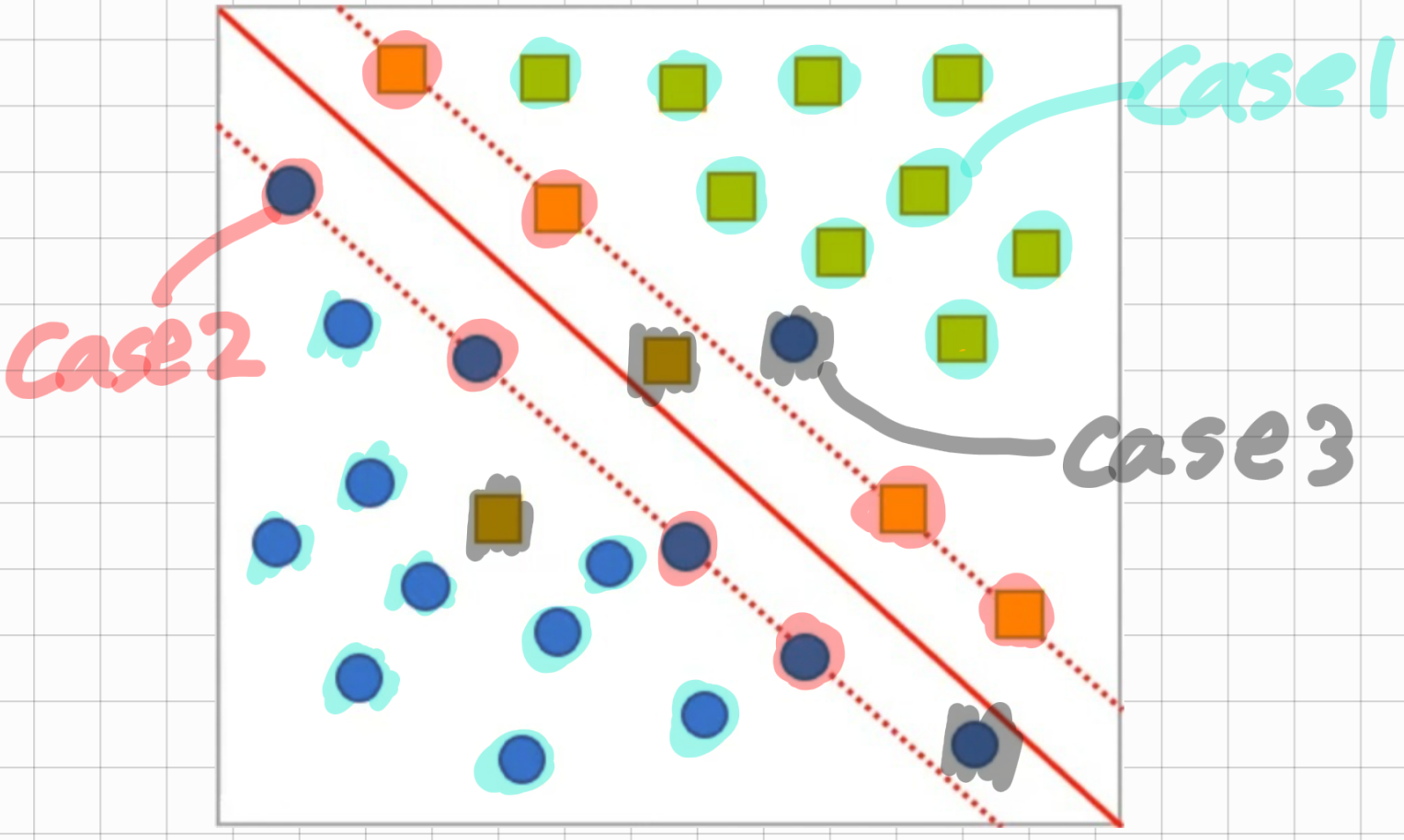
case 2 : $0<\alpha_i<C$ ③▷ $0<\beta_i<C$ ②▷ $\xi_i=0$ ①▷ $y_i(\vec{w}\cdot\vec{x}_i+b)-1+\xi_i=0$ ▷ $y_i(\vec{w}\cdot\vec{x}_i+b)-1=0$

S.V를 결정

∴ S.V이다 ▷ 당연히 오분류가 아니다

case 3 : $\alpha_i=C$ ③▷ $\beta_i=0$ ②▷ $\xi_i>0$ ①▷ $y_i(\vec{w}\cdot\vec{x}_i+b)-1+\xi_i=0$ ▷ $y_i(\vec{w}\cdot\vec{x}_i+b)=1-\xi_i\neq 1$

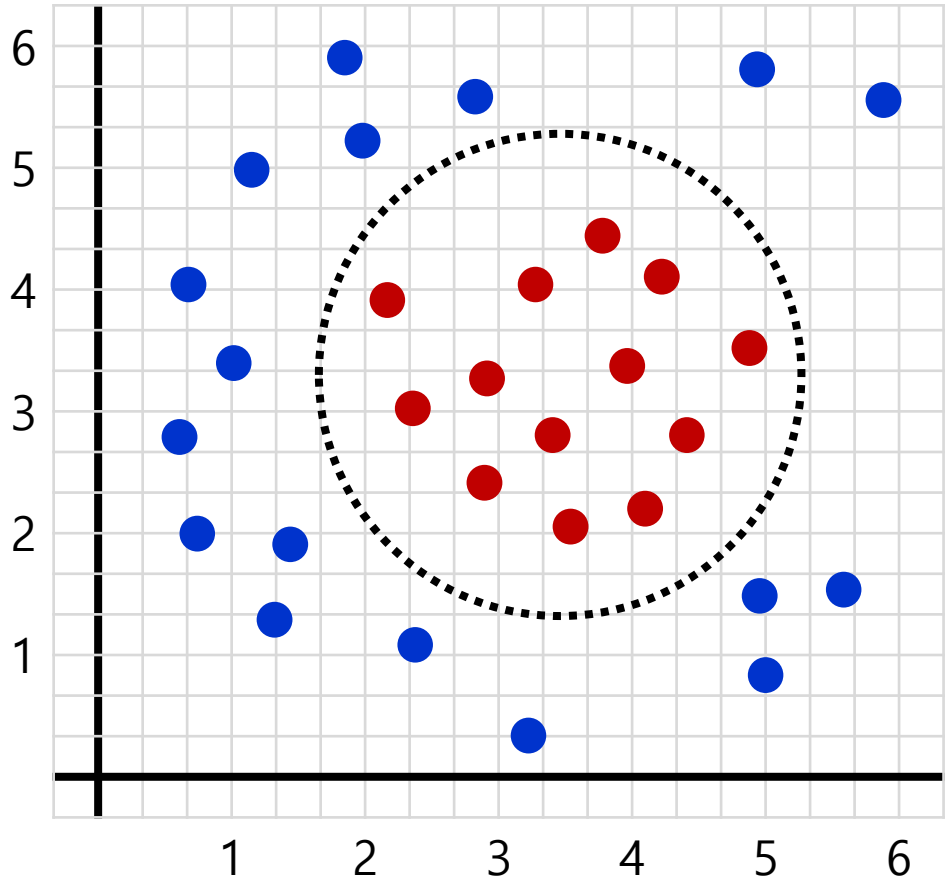
∴ S.V이 아니다 and 오분류 이다



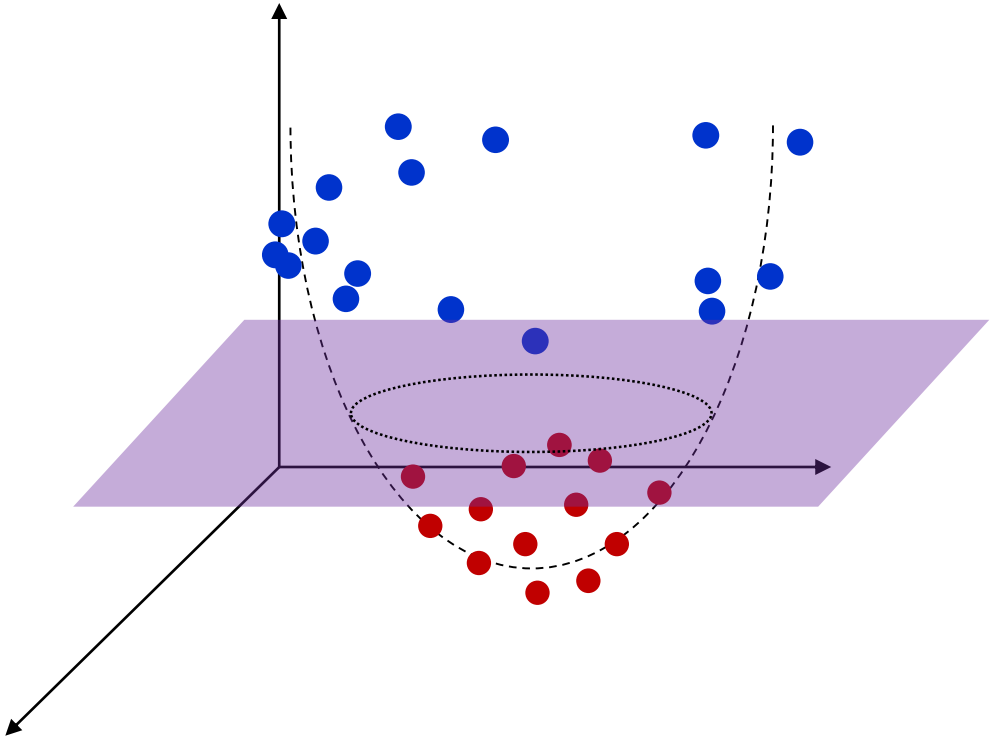
Kernel Function

Non Linear SVM

Non-Linear Support Vector Machine



$\Phi(x)$



$\vec{x} \in \mathbb{R}^2$ 2차원 → 확장 → 3차원 $\phi(\vec{x}) \in \mathbb{R}^3$

Non-Linear SVM의 목적함수

Hard Margin SVM과 다른 부분

Primal

$$\min_{\vec{w}, b, \xi_i} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t \quad y_i(\vec{w}\cdot\phi(\vec{x}_i)+b) \geq 1-\xi_i, \quad \xi_i \geq 0$$

Hard Margin SVM과 같은 방법으로 최적화를 하게되면

Dual

$$\Rightarrow \max_{\vec{\alpha}, \vec{\beta}} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \phi(x_i)^T\phi(x_j)\right]$$

$$s.t \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$$

Solution

$$\Rightarrow f(\vec{x}_{new}) = sign(\vec{w}^T\phi(\vec{x}_{new})+b)$$

최종 model

$$= sign\left(\sum_{i=1}^{n}\alpha_i y_i \phi(\vec{x}_i)^T\phi(\vec{x}_{new})+b\right)$$

$$= sign\left(\sum_{i=1}^{n}\alpha_i y_i K(\vec{x}_i, \vec{x}_{new})+b\right)$$

어떠한 $K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i)^T\phi(\vec{x}_j)$ 가 되는 함수 K를 찾아 대입

$\vec{x}_i, \vec{x}_j$ (d차원) $\xrightarrow{\phi}$ $\phi(\vec{x}_i), \phi(\vec{x}_j)$ (d_{new}차원) $\xrightarrow{\text{내적}}$ $\phi(\vec{x}_i)^T\phi(\vec{x}_j)$ (1차원)

$K(\vec{x}_i, \vec{x}_j)$

Kernel Function

kernel Trick 을 통해 연산량을 줄여줌

Non-Linear Support Vector Machine

1) Linear : $k(x_i, x_j) = x_i^T x_j$ $d \rightarrow d$

Mapping $\Phi: x \rightarrow \Phi(x)$, where $\Phi(x)$ is x itself

2) Polynomial of power p : $k(x_i, x_j) = (1 + x_i^T x_j)^p$ $d \rightarrow d+p C_p$

Mapping $\Phi: x \rightarrow \Phi(x)$, where $\Phi(x)$ has $\binom{d+p}{p}$ dimensions (d is the dimension of data)

3) Gaussian (radial-basis function) : $k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$ $d \rightarrow \infty$

Mapping $\Phi: x \rightarrow \Phi(x)$, where $\Phi(x)$ if infinite-dimensional

x 에 대한 mapping 차원이 증가
→비선형성이 증가함으로 복잡한 문제에 대한 분류 성능↑,
but, overfitting의 위험도 또한 ↑

4) Sigmoid : $k(x_i, x_j) = tanh(\kappa \cdot x_i^T x_j + c)$, $\kappa > 0$ and $c < 0$ $d \rightarrow \infty$

Kernel Function

Non-Linear Support Vector Machine

2) Polynomial of power p : $k(x_i, x_j) = (1 + {x_i}^T x_j)^p$

Mapping $\Phi: x \rightarrow \Phi(x)$, where $\Phi(x)$ has $\binom{d+p}{p}$ dimensions (d is the dimension of data)

Ex)

if $degree = 2$, $x_i = (x_{i_1}, x_{i_2})$, $x_j = (x_{j_1}, x_{j_2})$

$2 \xrightarrow{\phi} {}_{2+2}C_2 = 6$

$$k(x_i, x_j) = (1 + {x_i}^T x_j)^2 = (1 + x_{i_1}x_{j_1} + x_{i_2}x_{j_2})^2 = 1 + 2x_{i_1}x_{j_1} + 2x_{i_2}x_{j_2} + 2x_{i_1}x_{j_1}x_{i_2}x_{j_2} + {x_{i_1}}^2{x_{j_1}}^2 + {x_{i_2}}^2{x_{j_2}}^2$$

$$= (1, \sqrt{2}x_{i_1}, \sqrt{2}x_{i_2}, \sqrt{2}x_{i_1}x_{i_2}, {x_{i_1}}^2, {x_{i_2}}^2) \cdot (1, \sqrt{2}x_{j_1}, \sqrt{2}x_{j_2}, \sqrt{2}x_{j_1}x_{j_2}, {x_{j_1}}^2, {x_{j_2}}^2)$$

Gaussian (RBF) kernel

$\vec{x}, \vec{x}' \in \mathbb{R}^d$

d Dim $\xrightarrow{\phi(\vec{x}),\ \phi(\vec{x}')}$ ∞ Dim $\xrightarrow{\text{내적}}$ 1 Dim

$$K(\vec{x}, \vec{x}') = \exp\left(-\frac{(\vec{x}-\vec{x}')^2}{\sigma^2}\right)$$

$$= \left[\frac{1}{\exp((\vec{x}-\vec{x}')^2)}\right]^{\frac{1}{\sigma^2}} = \left[\frac{1}{\exp(\vec{x}^T\vec{x} + \vec{x}'^T\vec{x}' - 2\vec{x}^T\vec{x}')}\right]^{\frac{1}{\sigma^2}}$$

$$= \left[\frac{1}{\exp(\vec{x}^T\vec{x}) \times \exp(\vec{x}'^T\vec{x}') \times 1/\underbrace{\exp(2\vec{x}^T\vec{x}')}_{\text{☆}}}\right]^{\frac{1}{\sigma^2}}$$

Taylor expansion

$$\exp(2\vec{x}^T\vec{x}') = \sum_{k=0}^{\infty} \frac{(2\vec{x}^T\vec{x}')^k}{k!} = \sum_{k=0}^{\infty} \frac{2^k(\sum_{i=1}^{d} x_i x_i')^k}{k!}$$

$$= 1 + 2\left(\sum_{i=1}^{d} x_i x_i'\right) + \frac{2^2(\sum_{i=1}^{d} x_i x_i')^2}{2!} + \frac{2^3(\sum_{i=1}^{d} x_i x_i')^3}{3!} + \frac{2^4(\sum_{i=1}^{d} x_i x_i')^4}{4!} + \frac{2^5(\sum_{i=1}^{d} x_i x_i')^5}{5!} + \cdots$$

$\Rightarrow (\sum_{i=1}^{d} x_i x_i')^{\infty}$ 와 같이 ∞ 차원 까지 확장된다.

RBF 의 hyper parameter 은 σ^2 (분산) 으로

σ 가 작을수록 모델의 복잡도가 증가

Sigmoid kernel 도 taylor expansion 을 사용하여 ∞ 차 원으로 확장

$\rightarrow$ $\tan(K\vec{x}_i^T\vec{x}_j + C)$ 에서 C 가 클수록 K 가 클수록 모델의 복잡도 증가, 모델의 복잡도 감소

Kernel Function 이기위한 충분조건은 Mercer's condition을 만족하는 것

Mercer's condition

1. It is symmetric
   $\Rightarrow K(x_i, x_j) = K(x_j, x_i)$

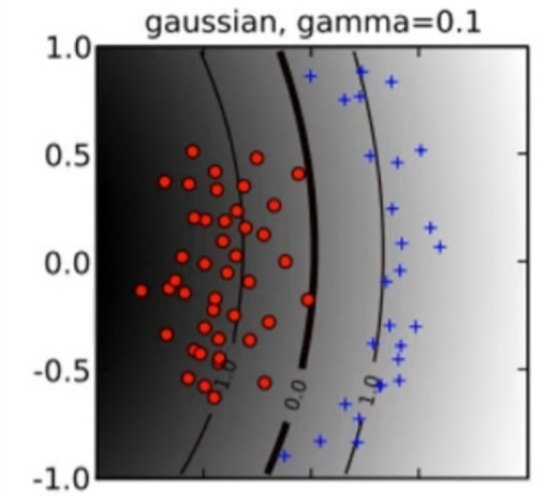
2. The matrix

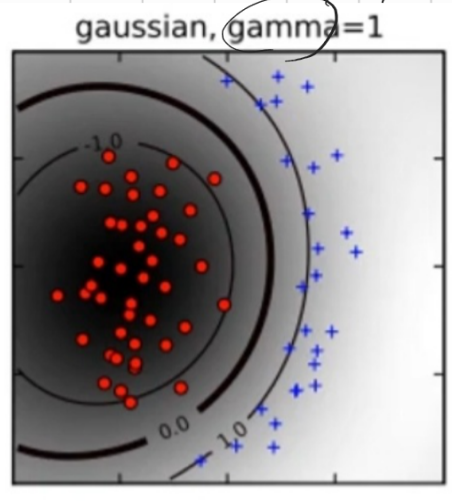
$$M = \begin{bmatrix} K(\vec{x}_1,\vec{x}_1) & K(\vec{x}_1,\vec{x}_2) & \cdots & K(\vec{x}_1,\vec{x}_n) \\ K(\vec{x}_2,\vec{x}_1) & K(\vec{x}_2,\vec{x}_2) & \cdots & K(\vec{x}_2,\vec{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(\vec{x}_n,\vec{x}_1) & K(\vec{x}_n,\vec{x}_2) & \cdots & K(\vec{x}_n,\vec{x}_n) \end{bmatrix}$$

   is positive semi-definite

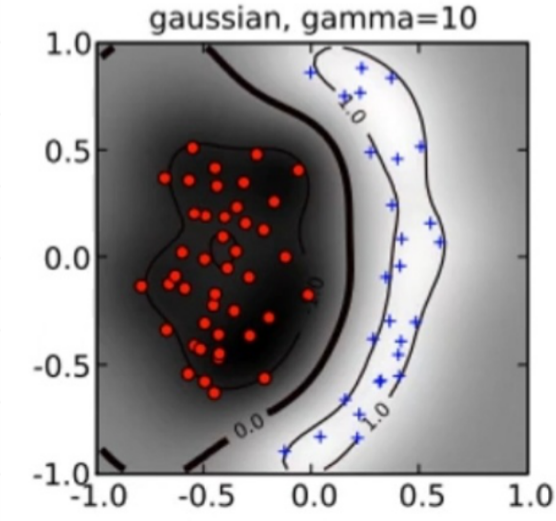
   $\Rightarrow$ 즉, $\vec{v}M\vec{v} \geq 0$ for all $\vec{v} \in (\mathbb{R}^n - \vec{0})$

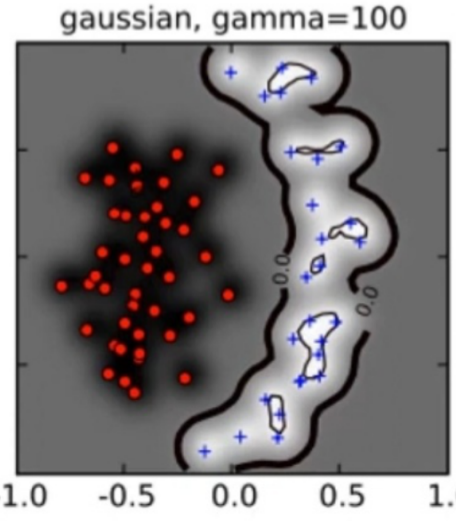
SVM의 hyper parameter $\frac{1}{\sigma^2}$





← 분산이 클때
경계면이 단순
분산↓





← 분산이 작을때
경계면이 복잡
분산↑

위의 목적함수 $\frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i$

분산↓
margin 늘어남
오분류에 관대해짐
작을수록

C

분산↑
margin 줄어듦
오분류에 엄격해짐
클수록