

# Support Vector Machine

## 1. SVM 소개

SVM(Support Vector Machine)은 두 가지 종류의 샘플들을 구분하는 최적의 결정 경계 (decision boundary)를 구하는 이진 분류기 (binary classifier)다.

어떤 n차원의 공간에 있을 때 그 공간을 두로 나눌 수 있는 기하체를 "hyper plane"이라 하고 다음과 같이 표현한다. geometric object

$$w^T x + b = 0 : \text{선형 hyper plane, 일반화하면 } f(x) = 0$$

예를 들어 그차원 공간은 직선으로 나눌 수 있고 ( $ax + by + c = 0$ )

3차원 공간은 평면으로 나눌 수 있다. ( $ax + by + cz + d = 0$ )

즉 n차원 공간에 n-1 차원의 subspace다.

어떤 이진 분류기를 만든다는 것은 샘플들을 분류할 수 있는 hyper plane을 결정경계로서 정하는 일이 된다.

SVM은 선형분류기로서 다음과 같은 선형 hyper plane을 결정 경계로 한다.

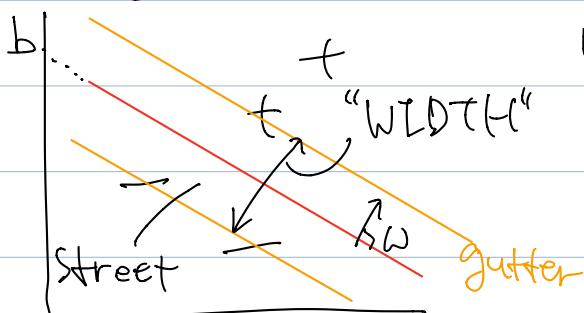
$$w^T x + b = 0, \quad x, w \in \mathbb{R}^n, \quad b \in \mathbb{R}$$

이 결정 경계에 따라 각 샘플은 다음과 같이 분류된다. (Decision Rule)

$$w^T x + b \geq 0 \text{ then } + \text{ (positive sample)}$$

$$w^T x + b < 0 \text{ then } - \text{ (negative sample)}$$

간단한 2차원 예시를 들면 다음과 같다.



빼간(?)은  $w^T x + b = 0$  결정경계를 그린

것이다. 결정경계와 가장 가까운 샘플들을

지나는 선을 "gutter" (내수구)라고 하고

그 사이를 "street"이라 하겠다.

그리고 gutter 위의 샘플들을 Support Vector라고 부른다.

+와 -를 분류할 수 있는  $(w, b)$ 는 무수히 많다. 가능한 직선의 방향도 무한히 많고  $(w, b)$ 에 같은 상수를 곱해준다. 결정 경계는 똑같다. (크기)

( $x_1 + x_2 + 1 = 0$ 이나  $2x_1 + 2x_2 + 2 = 0$ 이나 같다.)

SV(은)은 이러한 자유도를 없애고 최적의 해  $(\hat{w}, \hat{b})$  하나만 구할 수 있도록 조건을 추가한다.

## 2. SVM 문제 유도

SVM에서  $(w, b)$ 의 "방향"은 최적화해야 할 대상이고

"크기"는  $(w, b)$ 가 지켜야 할 조건으로 주어진다.

### 2.1 SVM의 조건

크기를 제약하기 위해 SVM에서는 결정 경계로 뉴런의 거리 조건을 준다.

$w^T x_i + b \geq +1$ , Support Vector에 대해서는 '='이 성립한다.

$$w^T x_i + b \leq -1$$

두 가지 쪽을 두는 것이 불편하므로  $y_i$ 라는 변수를 만든다

$$y_i = \begin{cases} +1 & \text{for } + \\ -1 & \text{for } - \end{cases}$$

$$\text{for } +, y_i(w^T x_i + b) = w^T x_i + b \geq 1$$

$$\text{for } -, y_i(w^T x_i + b) = -w^T x_i - b \geq 1$$

$$\Rightarrow y_i(w^T x_i + b) \geq 1 \quad (\text{for both } +, -)$$

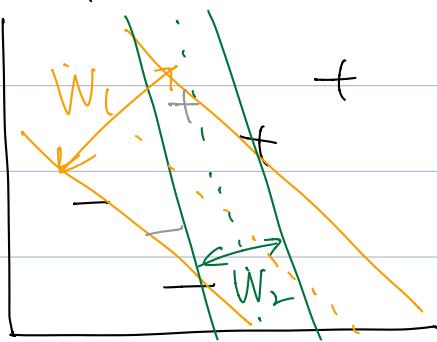
위 조건에 의해 이제  $(w, b)$ 는 무작정 커지거나 작아지지 못한다.

### 2.2 SVM의 목표

이제 "방향"에 대해 고민해보자.

"최적"의 결정 경계는 어떤 모양이어야 할까? SVM에서는 마진(margin)을

최대화하는 결정 경계를 찾는다. 예를 들어 설명해보자.



오른쪽 그림에서는 두 개의 결정 경계가 있다.

주행색 경계는 두 셀룰 사이를 큰 폭의 마진으로 구분하는 반면 녹색 경계는 상대적으로 미묘한 것이다.

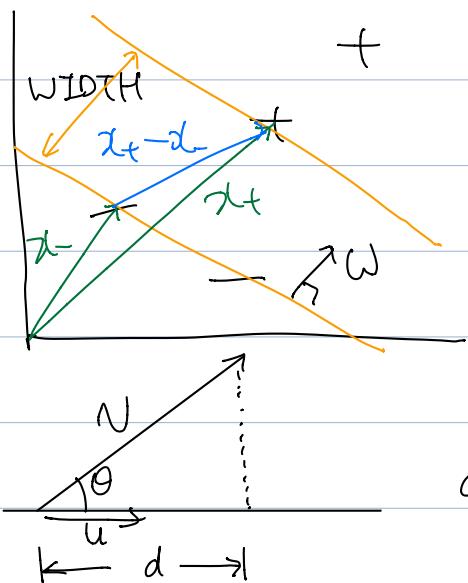
학습 셀룰에 대해 마진이 크면 좋은 이유는 학습 셀룰과 테스트 셀룰의 농도가 약간 다를 수 있기 때문이다. 위 그림에서 녹색 셀룰은 테스트 셀룰이다.

기준에 농도한 셀룰과 가까이 있긴 하지만 반대편 셀룰에 조금 가까워졌다.

학습 데이터가 아무리 많다고 해도 기존 학습 범위를 벗어난 테스트 데이터는 언제나 나타날 수 있고 이때 테스트 데이터가 학습 데이터와 비슷한 범위에 있다면 이를 최대한 바르게 분류해줘야 한다. 그러므로 학습 데이터 사이의 마진을 최대화해야 테스트 데이터가 약간 벗어나더라도 바르게 분류할 수 있다. 그렇다면 마진이 좋은 녹색 봉준기는 + 셀룰을 제대로 분류하지 못한다.

## 2.3 목적함수 정리

마진(margin)을 최대화하는 결정경계를 만들기 위해서는 일단 "margin"을 수학적으로 정리해야 한다. 마진은 "street"의 너비(width)라고 할 수 있다.



너비는 양쪽 gutter 사이의 두 점 사이의 벡터를 ( $x_+ - x_-$ ) 결정경계의 주 방향 ( $w$ )으로 projection 하여 구할 수 있다.

$$WIDTH = (x_+ - x_-)^T \frac{w}{\|w\|}$$

projection

$$d = \|v\| \cos \theta = \|v\| \cdot \frac{w}{\|w\|} = \|v\| \frac{\|w\|}{\|w\|} \cos \theta$$

위에서 만든 조건을 이용하면 넓이식을 간단히 정의할 수 있다.

$$w^\top \alpha_f + b = 1$$

$$w^\top Df = (\alpha_+ - \alpha_-)^\top \frac{w}{\|w\|}$$

$$\frac{w^\top \alpha_- - b = -1}{w^\top (\alpha_+ - \alpha_-) = 2} \rightarrow = \frac{2}{\|w\|}$$

드디어 풀어야 할 문제가 정의되었다. SVM 문제는 다음과 같이 표현할 수 있다.

$$(\hat{w}, \hat{b}) = \operatorname{argmax}_{(w, b)} \frac{2}{\|w\|} \text{ subject to } y_i(\alpha_i^\top w + b) \geq 1$$

그런데  $\max \frac{2}{\|w\|}$  를 쪼는 것은 어렵기 때문에 단순히 수학적 편의를 위해 문제를 다음과 같이 바꿀 수 있다.

$$\max \frac{2}{\|w\|} \rightarrow \min \frac{1}{2} \|w\|^2$$

SVM 문제를 다시 쓰면 다음과 같다.

$$(\hat{w}, \hat{b}) = \operatorname{argmin}_{(w, b)} \frac{1}{2} \|w\|^2 \text{ subject to } y_i(\alpha_i^\top w + b) \geq 1$$

### ⑥ 의미

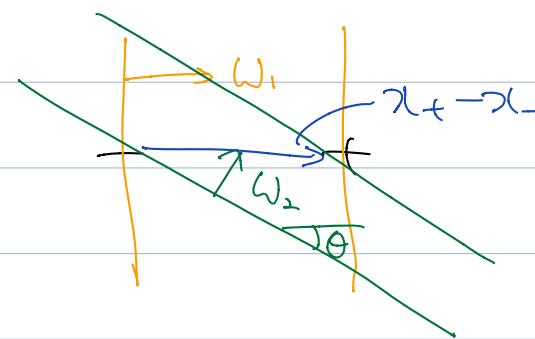
$w$ 의 크기를 왜 최소화시켜야 할까? ( $w$ 를 최소화시키는 것과 결점 경계의 빙향을 최적화하는 것과 무슨 관계가 있을까?)

마진을 유도하면서 중간에 다음식이 나온 적 있다. 이로부터 ( $w$ 를 유도해보자).

$$(\alpha_+ - \alpha_-) \cdot w = 2$$

$$\|\alpha_+ - \alpha_-\| \|w\| \cos \theta = 2$$

$$\|w\| = \frac{2}{\|\alpha_+ - \alpha_-\| \cos \theta}$$



$\|w\| \propto \frac{1}{\cos \theta} \propto \theta$  : ( $\alpha_+ - \alpha_-$ 는 상수이고  $w$ 와  $(\alpha_+ - \alpha_-)$ 의 각도에 따라  $\cos \theta$ 가 달라진다.  $\|w\|$ 는 각도가 벌어질수록 대각선 방향으로 거리 '2'를 만들어야 하기 때문에 커질수 밖에 없다.)

( $w$ 를 최소화 시키려면  $w$ 와  $(\alpha_+ - \alpha_-)$ 의 방향이 일치해야 한다.)

설제로는  $g+$  와  $g-$ 가 여러개 있기 때문에 평균적인 차이를 구하는데  
이는 이어지는 내용에서 설명한다.

### 3. SVM 풀이

#### 3.1 Lagrange Multiplier Method

우(여)에서 구한 SVM 정의식은 조건이 많아서 풀기 어렵다. 이럴 때 라그랑주 승수법 (Lagrange Multiplier Method)을 쓰면 조건식을 목적함수에 포함시켜서  
결과적으로 조건식없이 목적함수에 대한 최적화를 할 수 있다.

이 방법을 적용한 목적함수는 다음과 같다.

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \{y_i(w \cdot x_i + b) - 1\} : \text{Lagrangian}$$

subject to  $\alpha_i \geq 0$        $\alpha_i$  는 Lagrangian Multiplier function

조건식은 원래 목적함수에서 "빼기"가 되어야 한다.  $y_i(w \cdot x_i + b) \geq 1 \Rightarrow -\alpha_i(y_i(w \cdot x_i + b) - 1) \leq 0$

라그랑주 승수법에서 최적화의 결과는 항상  $\alpha_i(y_i(w \cdot x_i + b) - 1) = 0$  이어야 한다.

a 말의 의미는  $\alpha_i > 0$  or  $y_i(w \cdot x_i + b) - 1 = 0$  이라는 뜻이다.

즉, 서포트 벡터에서  $y_i(w \cdot x_i + b) - 1 = 0$  이므로  $\alpha_i$ 는 서포트 벡터에서만  
 $\alpha_i > 0$  값을 가질 수 있다.

라그랑주 승수법을 푸는 방법은 다음과 같다.

$$\frac{1}{2} \|\hat{w}\|^2 \geq \min_{(w,b)} L(w, b, \alpha) = g(\alpha)$$

$$\frac{1}{2} \|\hat{w}\|^2 = \max_{\alpha} g(\alpha) = \min_{\alpha} \max_{(w,b)} L(w, b, \alpha)$$

①  $L(w, b, \alpha)$  을  $(w, b)$ 에 대해 최대화시키는  $(\hat{w}, \hat{b})$  을 구하면  $(\hat{w}, \hat{b})$  이  
 $\alpha$ 에 관한 식으로 나올 것이다. ( $\hat{w} = \phi(\alpha)$ )

②  $(\hat{w}, \hat{b})$  을  $L(w, b, \alpha)$ 에 대입하고  $L(\hat{w}, \hat{b}, \alpha) = g(\alpha)$  을 최대화하는  
 $\alpha$  을 구한다.

③  $\alpha$  을  $(\hat{w}, \hat{b})$  식에 대입하여 최적의  $(\hat{w}, \hat{b})$  값을 계산한다.

### 3.2 극점의 조건

새로운 목적함수를  $(\omega, b)$ 에 대해 최적화 한다.  $(\omega, b)$ 에 대해  $\| \cdot \|$ 을 미분해서 0'이 나오는 극점 (extreme point)을 찾는다.

$$\begin{aligned}\frac{\partial L}{\partial \omega} &= \frac{\partial}{\partial \omega} \frac{1}{2} \omega^T \omega - \frac{\partial}{\partial \omega} \left( \sum_i \alpha_i (y_i (\omega \cdot x_i + b) - 1) \right) \\ &= \omega - \sum_i \left[ \frac{\partial}{\partial \omega} \alpha_i y_i (\omega \cdot x_i) + \frac{\partial}{\partial \omega} \alpha_i y_i b - \frac{\partial}{\partial \omega} \alpha_i \right] \\ &= \omega - \sum_i \alpha_i y_i x_i = 0\end{aligned}$$

$$\Rightarrow \hat{\omega} = \sum_i \alpha_i y_i x_i - \textcircled{1}$$

$$\begin{aligned}\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \frac{1}{2} \omega^T \omega - \sum_i \left[ \frac{\partial}{\partial b} \alpha_i y_i (\omega \cdot x_i) + \frac{\partial}{\partial b} \alpha_i y_i b - \frac{\partial}{\partial b} \alpha_i \right] \\ &= 0 - 0 - \sum_i \alpha_i = 0 = 0 \\ \Rightarrow \sum_i \alpha_i y_i &= 0 - \textcircled{2}\end{aligned}$$

여기서 수를 더 구현하기 전에 방금 구한 두式的 의미를 직관적으로 이해해보자.

식 ①의 의미는 최적의  $\hat{\omega}$ 가  $x_i$ 의 선형조합으로 구해진다는 것이다.

그런데  $y_i$ 는  $x_i$ 의 클래스에 따라 +1, -1 값을 가지므로 다음과 같이 쓸 수 있다.

$$\hat{\omega} = \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i$$

즉 최적의  $\omega$ 의 방향은 +샘플과 -샘플의 차이라는 것과 이는 일상 SUM 목적함수의 의미에서 살펴본 바와 같다.

최적화를 하게 되면 서로 다른 벡터에서만  $\alpha_i$ 가 값을 갖고 나머지  $\alpha_i = 0$  이 되므로  $\hat{\omega}$ 은 +, - 샘플 벡터 사이의 차이라고 볼 수 있다.

식 ②도 마찬가지로 +, - 샘플로 나눠서 생각해 볼 수 있다.

$$\sum_i \alpha_i y_i = \sum_{i \in +} \alpha_i - \sum_{i \in -} \alpha_i = 0 \Rightarrow \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i$$

이 식의 의미는 식 ①에서 +, - 두 클래스에 대해 서로 다른 벡터에 공유하는 가중치( $\alpha_i$ )의 합이 같다는 것이다. 이는 식 ①을 통해  $\hat{\omega}$ 이 -샘플에서 +샘플을 향하는 방향이 되기 위해 당연한 것이다.

### 3.3 SVM 문제 재정의

앞서 구한 극점의 조건을 이용해  $\mathcal{L}(\omega, b, \alpha) = \mathcal{L}(x)$ 로 단순화 할 수 있다.

$$\left\{ \begin{array}{l} \mathcal{L}(\omega, b, \alpha) = \frac{1}{2} \omega^T \omega - \sum_i \alpha_i \{ y_i (\gamma_i^T \omega + b) - 1 \} \\ \hat{\omega} = \sum_i \alpha_i y_i \gamma_i \\ \sum_i \alpha_i y_i = 0 \end{array} \right.$$

$$\begin{aligned} \mathcal{L}(\hat{\omega}, b, \alpha) &= \frac{1}{2} \left( \sum_i \alpha_i y_i \gamma_i \right)^T \left( \sum_j \alpha_j y_j \gamma_j \right) - \left( \sum_i \alpha_i y_i \gamma_i \right)^T \left( \sum_j \alpha_j y_j \gamma_j \right) \\ &\quad - \sum_i \alpha_i y_i b + \sum_i \alpha_i \\ &= \sum_i \alpha_i - \frac{1}{2} \left( \sum_i \alpha_i y_i \gamma_i \right)^T \left( \sum_j \alpha_j y_j \gamma_j \right) \\ \mathcal{L}(\alpha) &= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \gamma_i^T \gamma_j \quad (\alpha_i > 0) \end{aligned}$$

여기서  $\gamma_i, y_i$ 는 각각은 데이터고  $\alpha_i$ 만이 문제에서 최적화 해야 할 파라미터다.

그리고 MLT 장의에서도 강조했듯이  $\mathcal{L}$ 식은 그 자체가 아닌  $\gamma_i^T \gamma_j$ 의 식으로 볼 수 있다.

### 3.4 SVM의 해(Solution) ??

$\mathcal{L}(\alpha)$ 에 들어있는  $\sum$ 를 없애고 Matrix-Vector 형식으로 쓸 수도 있다.

$$\bar{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \quad (n \times 1) \quad X = \begin{bmatrix} y_1 \gamma_1 & \cdots & y_n \gamma_n \end{bmatrix} \quad (m \times n) \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad (n \times 1)$$

$$\begin{aligned} \mathcal{L}(\alpha) &= \sum_i \alpha_i - \frac{1}{2} \left( \sum_i \alpha_i y_i \gamma_i \right)^T \left( \sum_j \alpha_j y_j \gamma_j \right) \\ &= \bar{\alpha}^T \mathbf{1}_n - \frac{1}{2} (X \bar{\alpha})^T (X \bar{\alpha}) \\ &= - \left( \frac{1}{2} \underbrace{\bar{\alpha}^T X^T X \bar{\alpha}}_{[\alpha_1 \cdots \alpha_n] \begin{bmatrix} y_1 \gamma_1^T \\ \vdots \\ y_n \gamma_n^T \end{bmatrix} [\gamma_1^T \cdots \gamma_n^T]} - \bar{\alpha}^T \mathbf{1}_n \right) \end{aligned}$$

$$([\alpha_1 \cdots \alpha_n] \begin{bmatrix} y_1 \gamma_1^T \\ \vdots \\ y_n \gamma_n^T \end{bmatrix} [\gamma_1^T \cdots \gamma_n^T])^T \begin{bmatrix} \alpha_1 y_1 \\ \vdots \\ \alpha_n y_n \end{bmatrix} = ((X \bar{\alpha})^T \mathbf{1}_n) = ((X \bar{\alpha})) \text{ Scalar!}$$

벡터와 행렬에 걸먹지 않고 식을 단순하게 바라보면  $\alpha$ 에 대한 2차식 임을 알 수 있다. 단순하게  $\Delta$ 를 차로 표현하면 이런 것이다:  $-\frac{1}{2}(\alpha^T \alpha^2 - \alpha)$  이러한 2차식은 위로 복록한 Concave 함수라고 한다. Convex 함수가  $f'(\alpha) = 0$ 인 지점에서 전역 최소값을 갖듯이 Concave 함수는  $g'(\alpha) = 0$ 인 지점에서 전역 최대값을 갖는다.



최적값을 구하기 위해  $L(\alpha)$ 를  $\alpha$ 로 미분해 보자.

$$\frac{\partial L(\alpha)}{\partial \alpha} = -X^T X \bar{\alpha} + \mathbf{1}_n = 0_n$$

$$\underbrace{\bar{\alpha}^*}_{(n \times 1)} = \underbrace{(X^T X)^{-1}}_{(n \times n)} \underbrace{\mathbf{1}_n}_{(n \times 1)} = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix}$$

$\hat{\alpha}^*$ 을 알면 회전의  $\hat{\omega}, \hat{b}$ 도 계산 가능하다.

$$\hat{\omega} = \sum_i \hat{\alpha}_i^* y_i x_i$$

$$y_i (\hat{\omega} \cdot x_i + b) = 1, \quad \hat{b} = \hat{y}_i - \hat{\omega} \cdot x_i \quad \text{for support vector } (x_i, y_i)$$

$$y_i \in \{1, -1\} \text{ or } \hat{b} = y_i - \hat{\omega} \cdot x_i$$

하지만 SUM은 이렇게 풀리지 않는다!

나도 저렇게 풀어보려 했지만 원하는 답이 나오지 않아 한참 고민했다.

위 solution의 문제점은  $(X^T X)$ 의 역행렬을 구할 수 없다는 것이다.

$X$ 는  $(m \times n)$  행렬이다.  $m$ 은  $W$ 의 차원이고  $n$ 은 데이터 포인트 개수다.

보통은  $m < n$  이다.

$$\text{rank}(X) = m \Rightarrow \text{rank}(X^T X) = m, \quad \dim(X^T X) = (n \times n)$$

$\rightarrow X^T X$ 의 column들이 서로 linearly independent하지 않다.

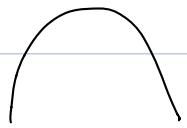
$= X^T X$  is Not full rank = eigen value 0'이 있다.

$= X^T X$  is Not invertible

좀 더 쉽게 이해할 수 있도록 기하학적으로 설명해보자.

간단한 두개의 다수가 함수를 비교해 보자.

$$f_1(x) = -x^2 + 1$$



$$f_2(x) = -x + 1$$

!

$$\frac{\partial^2}{\partial x^2} f_1(x) = -2 < 0 \quad \forall x$$

$$\frac{\partial^2}{\partial x^2} f_2(x) = 0$$

"strictly concave"

"(just) concave"

$$\frac{\partial}{\partial x} f(x) = 0 \text{에서 극대}$$

: 절댓값의 시작이나 끝에서 극대

다시  $f(\bar{x})$  문제로 돌아가보기

$$\frac{\partial}{\partial \bar{x}} f(\bar{x}) = -X^T X \bar{x} + \mathbf{1}, \quad \frac{\partial^2}{\partial \bar{x}^2} f(\bar{x}) = -X^T X \leq 0, \quad X^T X \geq 0$$

행렬이 양수라는 게 무슨 뜻일까?

$$X^T X = G \geq 0 \Leftrightarrow G \text{ is positive semi-definite}$$

$$\Leftrightarrow \text{Eigenvalues of } G \geq 0$$

$$G > 0 \Leftrightarrow G \text{ is (strictly) positive definite}$$

$$\Leftrightarrow \text{Eigenvalues of } G > 0$$

앞서 생각한  $\Delta$ 가 랑수 경우를 다시 생각해보자. 그간 미분이 완전히 음수인

경우에는 볼록한 모양이 나왔고 ○인 경우에는 직선이 나왔다. 이것 벡터에 적용하면 어떻게 될까? 간단한 예시를 들어보자

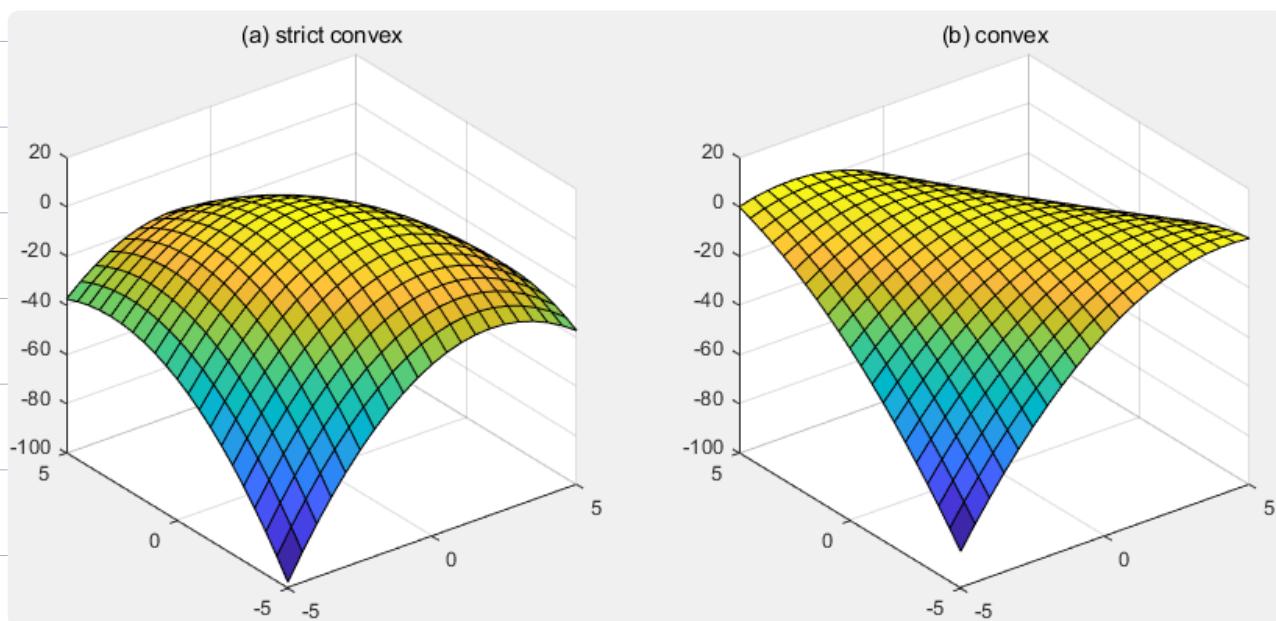
$$L(\bar{\alpha}) = -\frac{1}{2} \bar{\alpha}^T X^T X \bar{\alpha} - \bar{\alpha}^T \mathbf{1} = -\frac{1}{2} \bar{\alpha}^T G \bar{\alpha} + \bar{\alpha}^T \mathbf{1}$$

$\alpha$ 가 2차원 벡터인 경우  $G$ 는  $(2 \times 2)$  행렬이다.

$G$ 에 따라  $L(\bar{\alpha})$  함수가 어떻게 달라지는지 그래프로 확인해보자.

$$G_a = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

$$G_b = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$$



기로 세로 축은  $\bar{\alpha} = [\alpha_1, \alpha_2]$ 이고 높이가  $L(\bar{\alpha})$ 다.

두 그림을 보면 모양이 확연히 차이가 난다. (a)는 완전히 볼록한데 (b)는 한쪽 방향으로는 볼록한데 다른 방향으로는 직선적이다.

이렇게 된 이유는  $\text{rank}(G_a) = 2$ ,  $\text{rank}(G_b) = 1$ 이 때문이다.

즉  $G_a$ 는 2개의 eigenvalue 모두 양수가 때문에 어느 방향으로든

"Strictly Convex" 하지만  $G_b$ 의 eigenvalue는 하나만 양수고 하나는 0이다 때문에 한쪽으로만 볼록한 모양이 나오는 것이다.

결과적으로,

$L(\bar{\alpha}) = -\frac{1}{2} \bar{\alpha}^T X^T X \bar{\alpha} + \bar{\alpha}^T \mathbf{1}$ 의 해는  $\frac{\partial}{\partial \bar{\alpha}} L(\bar{\alpha}) = 0$  으로는 구할 수 없다.

하지만  $\frac{1}{2}(\alpha_i)$ 은 concave한 것은 맞기 때문에 경사하강법 등의 최적화 방법을 통해 해를 구할 수 있다. 그럼 (b)처럼 선형적인 측면 있기 때문에 일반적으로는 경사하강법을 쓰면 능성이 따라 무한히 올라가야 하지만

( $\frac{1}{2}(\alpha_i)$ 에 대해서는 최대화 문제이 때문에 엄밀히 말하면 경사 "상승"법이다.)

$\alpha_i \geq 0$ 이라는 조건이 있기 때문에 그렇게 되지는 않는다.

이 말은 결국 어떤 차원에서는  $\alpha_i = 0$  이되고 어떤 차원에서는  $\alpha_i > 0$ 이 된다는 것인데  $\alpha_i = 0$  이면 해당 샘플인  $x_i$ 가 Support Vector 된다.

## 4. kernel SVM

앞서 공부한 SVM을 요약하면 다음과 같다.

Decision Rule:  $w^T x + b \geq 0$

Primal Problem

$$(\hat{w}, \hat{b}) = \underset{(w, b)}{\arg \min} \frac{1}{2} \|w\|^2 \text{ subject to } y_i(\alpha_i^T w + b) \geq 1$$

Dual Problem

$$\hat{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\hat{\alpha} = \underset{\alpha}{\arg \max} \hat{L}(\alpha) \text{ subject to } \alpha_i \geq 0$$

Solution to Primal problem:  $\hat{w} = \sum_i \hat{\alpha}_i y_i x_i, \sum_i \hat{\alpha}_i y_i = 0$

Rewriting Decision Rule:  $\hat{w}^T x + b = \sum_i \hat{\alpha}_i y_i x_i^T x + b \geq 0$

## 4.1 Mapping Function

이러한 SVM은 결정 경계가 선형적인 hyper plane으로 나타나므로 학습데이터가 복잡하게 얹혀 있는 경우에는 적용하기 어렵다.

SVM으로 이런 문제를 해결하기 위해서는 학습데이터들을 선형적으로 분리 가능한 공간으로 mapping 해야 한다.

Mapping은 같은 차원으로 해도 도지만 보통 차원이 많아지면 뿐만 아니라 차원을 늘리는 방향으로 mapping 한다.

$\hat{\alpha}$ 라는 썬풀을 mapping 하는 함수를  $\phi(x)$ 라 하자. 가가 아닌 새로운  $\phi(x)$ 의 공간에서 SUM 문제를 다음과 같이 다시 쓸 수 있다.

Decision Rule:  $\sum_i \hat{\alpha}_i y_i \phi(x_i)^T \phi(x) + b \geq 0$

Dual Problem:

$$L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \quad (\alpha_i \geq 0)$$

이러한 kernel SVM을 적용하려면 먼저  $\phi(x)$ 에 대한 Dual Problem을 풀어서  $\hat{\alpha}$ 를 구한 다음  $\hat{\alpha}$ 를 Decision Rule에 대입하여 썬풀  $x$ 를 분류하는 것이다.

그런데 두 썬을 보면 둘다  $\phi(x)$  자체보다는  $\phi(x_i)^T \phi(x_j)$ 에 의존한다.

## 4.2 Kernel Function

mapping 된 두 벡터 사이의 내적을  $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 로 정의하면 썬을 다음과 같이 다시 쓸 수 있다.

Decision Rule:  $\sum_i \hat{\alpha}_i y_i K(x_i, x) + b \geq 0$

Dual Problem:  $L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (\alpha_i \geq 0)$

이렇게 보면 명시적으로  $\phi(x)$ 를 정의할 필요도 없어진다. 여기서 유일하게 필요한건 두 벡터 사이의 관계를 스칼라로 표현해 줄 커널 함수  $K(x_i, x_j)$ 다.

여기서도 마찬가지로 주어진  $K(x_i, x_j)$ 에 대해  $L(\alpha)$ 를 최대화하는  $\hat{\alpha}$ 를 구하고 이를 Decision Rule에 적용하면 임의의 썬풀  $x$ 를 분류할 수 있다.

## 4.3 Radial Basis Function (RBF)

비선형 분류문제를 풀기 위해 가장 흔하게 사용되는 커널이 RBF다.

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) = \exp(-\gamma \|x_i - x_j\|^2)$$

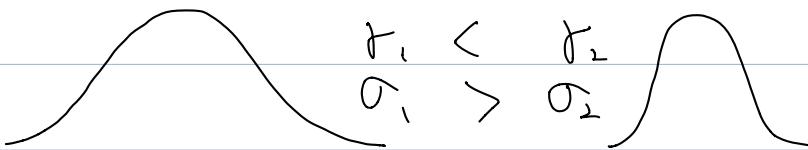
RBF는  $\sigma$  혹은  $\gamma$  값에 따라 결정 경계를 단순하게 만들 수도, 복잡하게 만들 수도 있다.

왜? 어떻게 그럴까?

기울함  $\alpha$ 의 모양을 보면 정확하지 않지만 꽤 만족인 이해는 할 수 있다.

$\gamma$ 가 작으면 ( $=\sigma$ 가 크면)  $x_i$  와  $x_j$  사이의 거리  $d = \|x_i - x_j\|^2$ 에 따른

RBF의 모양이 원만하게 나타나고  $\gamma$ 가 크면 ( $=\sigma$ 가 작으면) RBF의 모양이  
좁게 나타난다.



결정경계는 이러한 커널의 선형조합 (Linear Combination)이 때문에

$$\sum_i \hat{\alpha}_i y_i k(x_i, x) + b = 0$$

커널의 모양이 속도로를 주로 결정경계의 모양이 속도로워지는 것을 예상할 수 있다.

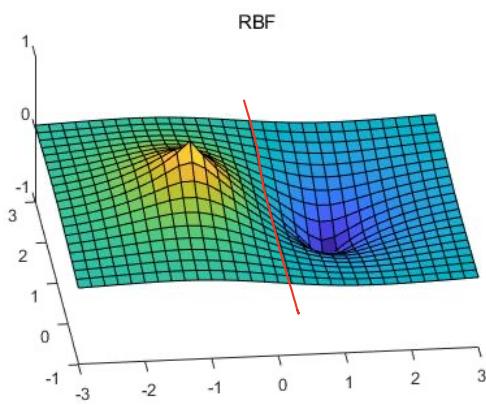
여기서  $\hat{\alpha}_i$  와  $y_i$  까지 고려해보면 결정경계는 직접 그릴 수도 있다.

$\hat{\alpha}_i$  는 시포트 벡터에서만 양수고 나머지 샘플에서선 0이다.  $y_i$ 는 샘플의  
클래스에 따라  $\pm 1$ 이다. 시포트 벡터 중 positive sample의 인덱스 집합을  
 $I^+$  라 하고 negative sample의 인덱스 집합을  $I^-$  라 하면 결정경계는

다음과 같이 쓸 수 있다.

$$\begin{aligned} D(x) &= \sum_{i \in I^+} k(x_i, x) - \sum_{j \in I^-} k(x_j, x) = 0 \\ &= \sum_{i \in I^+} \exp(-\gamma \|x_i - x\|^2) - \sum_{j \in I^-} \exp(-\gamma \|x_j - x\|^2) \end{aligned}$$

만약 2차원 공간에서 양쪽에 시포트 벡터가 하나씩 있다면  $D(x)$ 는 다음과 같은  
모양이 나온다. ( $x_+ = (-1, 1)$ ,  $x_- = (1, 1)$ )



즉 시포트 벡터마다 솟은곳 (+샘플)이나  
들어간곳 (-샘플)이 생기고 그 사이의  
높이가 0인 점들이 결정경계를 대룬다.  
그나마에 따라 솟은곳과 들어간곳의  
경사가 달라진다.