[SWCON253] Machine Learning – Lec.15a

Support Vector Machine

Fall 2025

김 휘 용

hykim.v@khu.ac.kr

경희대학교
KYUNG HEE UNIVERSITY

Contents

References

- *기계 학습* by 오일석, *패턴 인식* by 오일석

1. Introduction

- ✓ SVM (Support Vector Machine)
- ✓ (Revisit) Linear Decision Boundary
- ✓ **Concept of Margin (Motivation of SVM)**
- ✓ Distance from the Decision Boundary

SVM (Support Vector Machine)

SVM

- Classification을 위한 ML Algorithm의 하나로, 여백(margin)을 이용하여 일반화 능력을 향상시킴
- 1990년대 신경망의 성능을 능가하여 인기 있는 모델로 부상
- 2000년대 들어 DL에 밀려 시들한 편
- 최근 SVM과 DL을 결합하는 접근방법이 시도되고 있음

Linear SVM vs. Non-linear SVM

- 선형 SVM은 선형 분류문제를 오류 없이 풀 수 있음
- 비선형 SVM은 비선형 분류문제도 풀 수 있음

Vladmir Vapnik

Vladimir Vapnik
러시아

Vapnik은 SVM을 창안한 사람으로 유명하다. 현재 SVM은 일반화 능력이 ~~여러~~ 뛰어난 분류기로 인정 받고 있다. 패턴 인식에서 기술 돌파로 breakthrough 평가되고 있는 것이다. 그는 러시아에서 태어났으며 모스코바에 있는 제어 과학 연구원에서 Institute of Control Sciences 통계학으로 박사 학위를 취득하였다. 그 후 이 연구원에서 1990년까지 근무하였고 이후 미국 AT&T로 이적하였고 주로 미국에서 연구 활동을 하였다. 현재는 Columbia 대학과 London 대학의 교수이다. SVM을 포함하여 통계적 학습 이론을 정리한 책이 그의 대표적인 저서 중의 하나이다 [Vapnik98].

[Vapnik98] Vladimir Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.

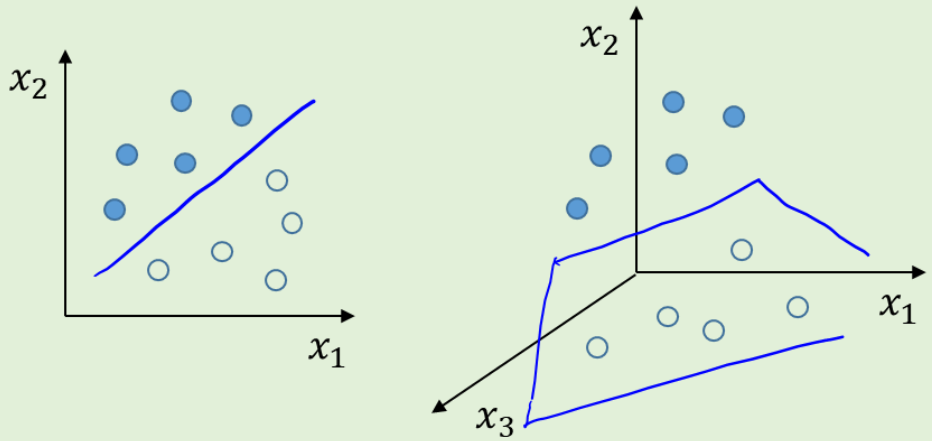
(Recap.) Linear Decision Boundary

Equation for *Linear Decision Boundary* (선형 결정 경계)

$$d(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_0 = 0$$

★ class 1 if $d(\mathbf{x}) > 0$, class 2 if $d(\mathbf{x}) < 0$

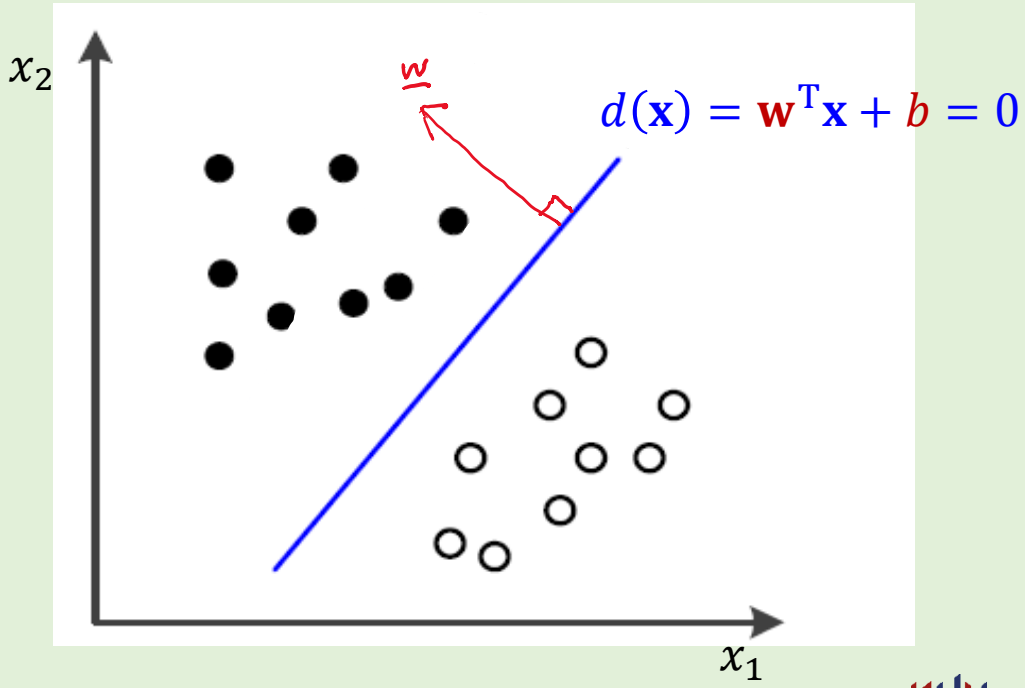
- Two types of vector representation:
  - ★ $\mathbf{x} = [x_0\ x_1\ \ldots x_d], \mathbf{w} = [w_0\ w_1\ \ldots w_d] : d(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = 0$
  - ★ $\mathbf{x} = [x_1\ \ldots x_d], \quad \mathbf{w} = [w_1\ \ldots w_d] \quad : d(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$



Geometric Interpretation

- $d(\mathbf{x}) = 0$ is a ***hyperplane*** in the feature space.
- $d(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$
  - ★ $\mathbf{w}$ is the surface ***normal vector*** of the hyperplane.
  - ★ b determines the ***position*** (i.e., the displacement from the origin) of the hyperplane.

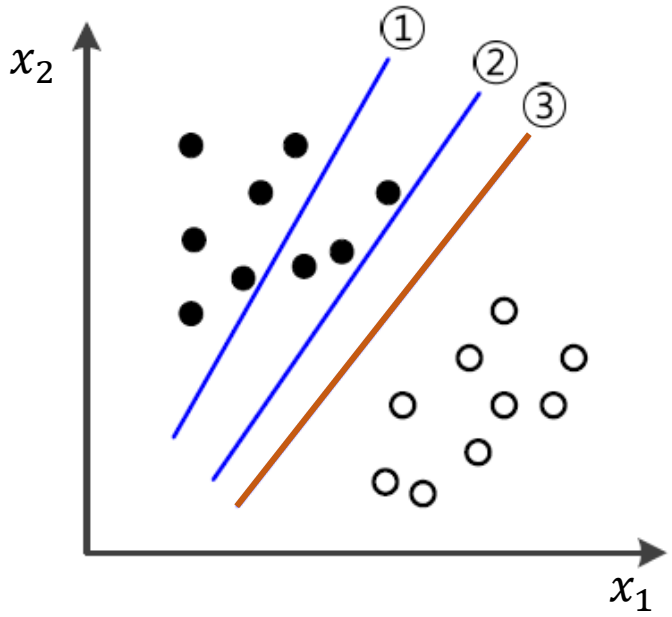
$\mathbf{w}$는 결정경계의 방향을 결정하고, b는 위치를 결정한다.



Concept of Margin

- The previously learned *linear classifiers* try to minimize some "*error*".
  - Cross entropy loss: $J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h(\mathbf{x}^{(i)})\right) + (1-y^{(i)})\log\left(1-h(\mathbf{x}^{(i)})\right)\right]$
  - MSE loss: $J(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(h(\mathbf{x}^{(i)}) - y^{(i)}\right)^2$

  ➔ *If all the samples are correctly classified, the optimization will stop.*



기존 선형분류 모델들:
①에서 출발하여
②에 도달하면 멈춤

- SVM의 Motivation: **오류**가 없으면 최선인가?
  - **결정 초평면으로부터 양쪽 class에 대한 "여백"을 최대화**할 수록 분류기의 **일반화 능력**이 향상될 것이다.

    ★ ② 보다 ③이 양쪽 class에 대해 여백이 커서 일반화 능력이 우월
  - 그러면, 여백(margin)이라는 개념을 어떻게 수학적으로 표현할 것인가?

    → ***Minimum distance from training sample to the decision boundary !***
  - 또한, 여백을 최대로 하는 결정 초평면을 어떻게 찾을 것인가?

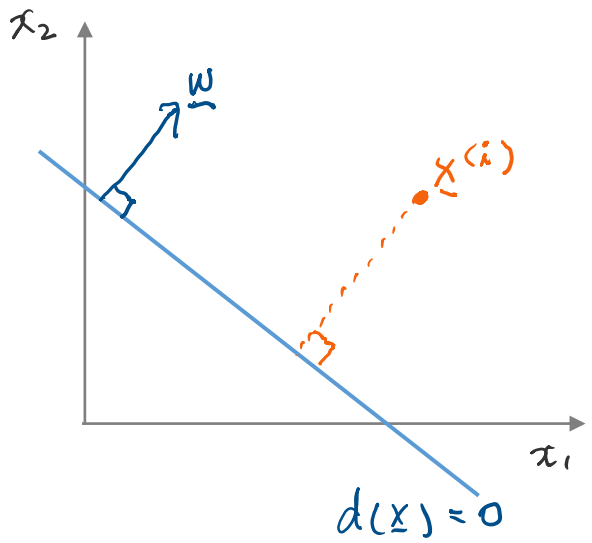
    → ***Constrained optimization !***

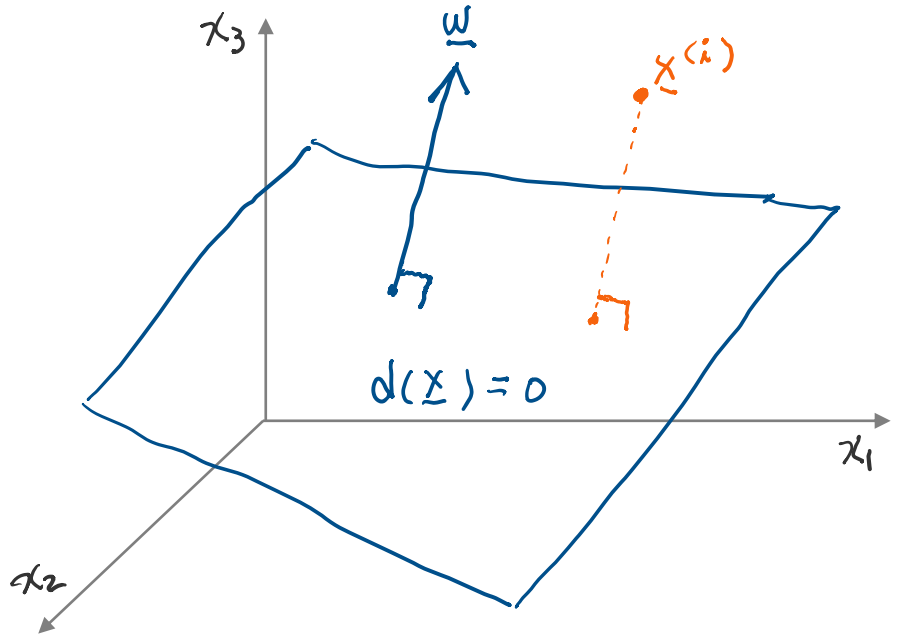
Distance From the Decision Boundary

- Consider
  - a *decision boundary*: $d(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$
    - ★ $\mathbf{x}$: a feature vector representation without $x_0 = 1$:
  - and a *sample point* to be classified: $\mathbf{x}^{(i)}$

- Then, the Euclidean distance from $\mathbf{x}^{(i)}$ to $d(\mathbf{x}) = 0$ is given by

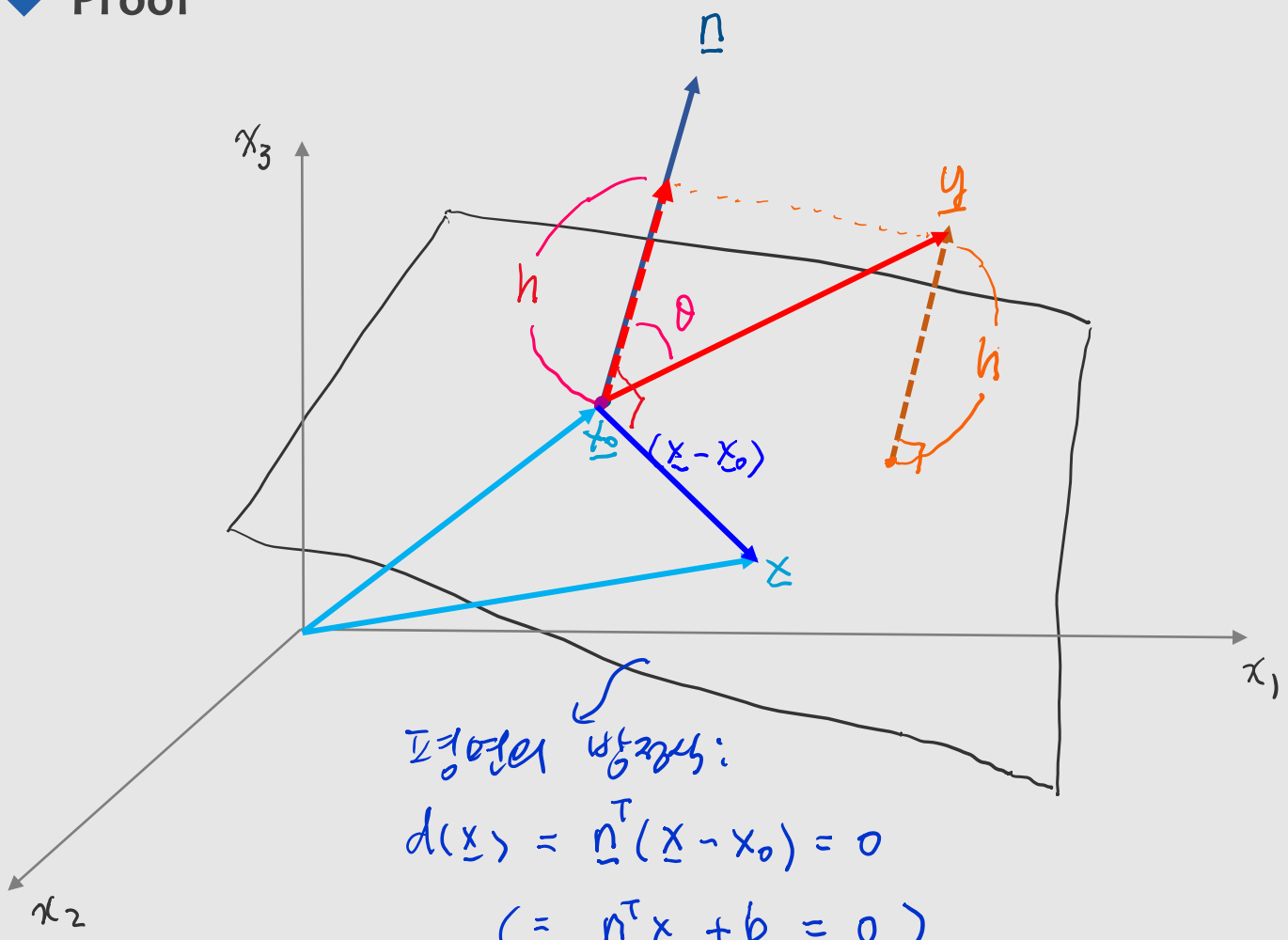
$$h = \frac{|d(\mathbf{x}^{(i)})|}{\|\mathbf{w}\|_2}$$





Distance From the Decision Boundary (cont'd)

- **Proof**



초평면의 방정식:

$$d(x) = n^T(x - x_0) = 0$$

$$(= n^T x + b = 0)$$

$$h = \|y - x_0\|_2 |\cos\theta|$$

$$= \|y - x_0\|_2 \cdot \frac{|n^T(y - x_0)|}{\|y - x_0\|_2 \|n\|_2}$$

$$= \frac{|n^T(y - x_0)|}{\|n\|_2}$$

Since $d(x) = n^T(x - x_0)$,

$$|n^T(y - x_0)| = |d(y)|.$$

$$\therefore h = \frac{|d(y)|}{\|n\|_2}$$

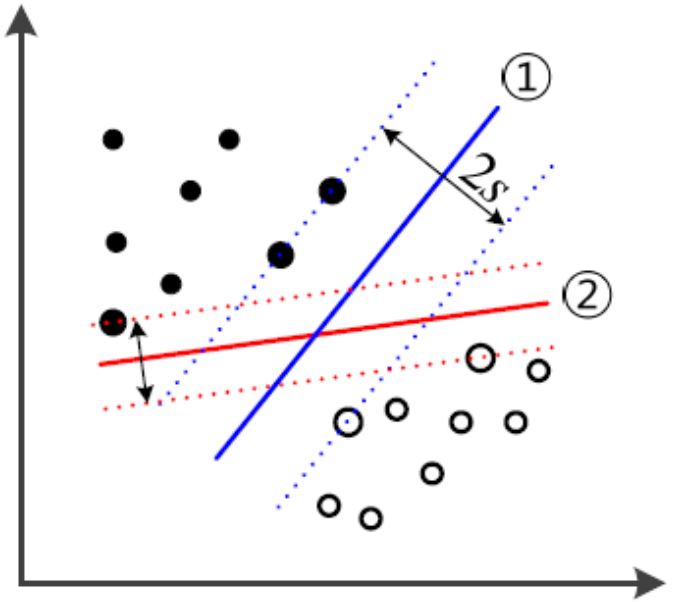
2. Linear SVM with Hard Margin

- ✓ Support Vector Machine
- ✓ Linear SVM – *Formulation*
- ✓ Linear SVM – *Lagrange Method*
- ✓ Linear SVM – *Lagrange Dual Problem*
- ✓ Linear SVM – *Illustrative Example*

Support Vector Machine

◆ 선형분리 가능한 분류 문제를 생각하자.

- 먼저 주어진 데이터를 오류없이 분류하도록 **w** 후보들을 정한다. (예: ①,②)
- 정해진 직선의 방향 **w**에 대해, 직선으로부터 가장 가까운 샘플까지의 거리가 같게 되도록 바이어스 b를 정한다.
- 그림의 ①과 ②는 위의 과정을 통해 얻은 결정 직선의 예를 나타낸다.



- 이때, 각 결정 직선의 분할 띠 너비 $2s$를 **여백**(***margin***)이라 부른다.
- 이때, 분할 띠의 경계에 있는 샘플을 **서포트 벡터**(***support vector***)라 한다.

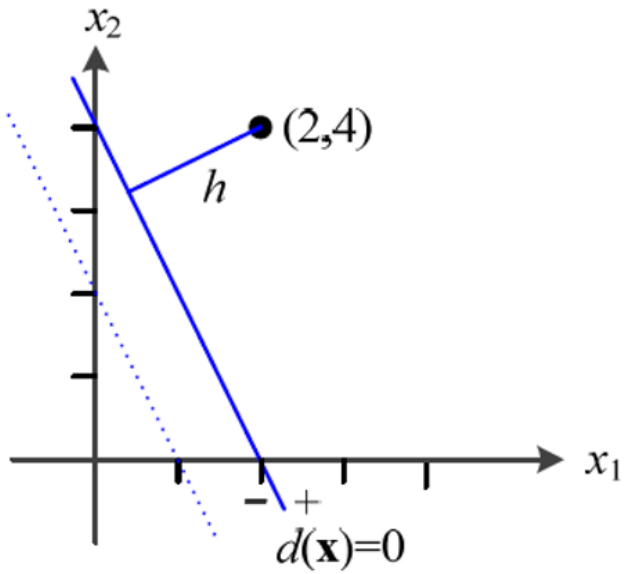
◆ SVM은 여백을 최대로 하는 결정 초평면을 구하는 알고리즘이다.

$$\text{여백} = 2s = \frac{2|d(\mathbf{x})|}{\|\mathbf{w}\|_2} = \frac{2}{\|\mathbf{w}\|_2}$$ (여기서 **x**는 support vector)

- $d(\mathbf{x})$에 상수를 곱해도 같은 초평면을 나타낸다는 성질을 이용하여, 서포트 벡터들에 대해 $|d(\mathbf{x})| = 1$ 이 되도록 $d(\mathbf{x})$ 식을 scale(즉, **w**를 scale)하였음에 유의하라. (이후 예시 참조)

Support Vector Machine (cont'd)

- **Example**
  - 특징 공간 상의 한 점: $\mathbf{x}_1 = [2\ 4]^T$
  - 결정 직선: $d(\mathbf{x}) = 2x_1 + x_2 - 4 = 0$



  - **1) 점과 직선 사이의 거리 구해 보기:**
    - ★ 결정 직선을 $d(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$ 하면, $\mathbf{w} = [2\ 1]^T,\ b = -4$.
    - ★ 점과 직선 사이의 거리:

$$h = \frac{|d(\mathbf{x}_1)|}{\|\mathbf{w}\|_2} = \frac{|2 \cdot 2 + 1 \cdot 4 - 4|}{\sqrt{2^2 + 1^2}} = \frac{4}{\sqrt{5}}$$

  - **2) $\mathbf{x}_1$이 서포트 벡터라 할 때, $|d'(\mathbf{x}_1)| = 1$ 이 되도록 결정 직선을 scale 하기:**
    - ★ $|d(\mathbf{x}_1)| = 4$이므로, 결정 직선 식 $d(\mathbf{x}) = 2x_1 + x_2 - 4 = 0$의 양변에 ¼을 곱하면,

$$d'(\mathbf{x}) = \frac{1}{2}x_1 + \frac{1}{4}x_2 - 1 = 0 \ \rightarrow\ d'(\mathbf{x}) = \mathbf{w}'^{\mathrm{T}}\mathbf{x} + b' = 0 \ \text{ with } \ \mathbf{w}' = [\tfrac{1}{2}\ \tfrac{1}{4}]^T,\ \ b' = -1.$$

  - **Note**
    - ★ $d'(\mathbf{x}) = c \times d(\mathbf{x})$라 하면, $d'(\mathbf{x}) = 0$ 과 $d(\mathbf{x}) = 0$ 은 동일한 결정경계를 나타낸다.
    - ★ 따라서, $d'(\mathbf{x})$를 사용 해도 위 거리 공식은 성립한다.

Support Vectors & Decision Boundary

Support Vectors

- 분할 띠의 **경계**에 있는 **샘플** 벡터를 말한다.
- 서포트 벡터들에 대해 $|d(\mathbf{x})| = 1$ 이 되도록 $d(\mathbf{x})$가 스케일되어 있다 가정하면, 다음 관계 식들이 성립한다.

★ $\mathbf{x}_i$가 서포트 벡터일 경우 (i.e., $|d(\mathbf{x}_i)| = 1$):

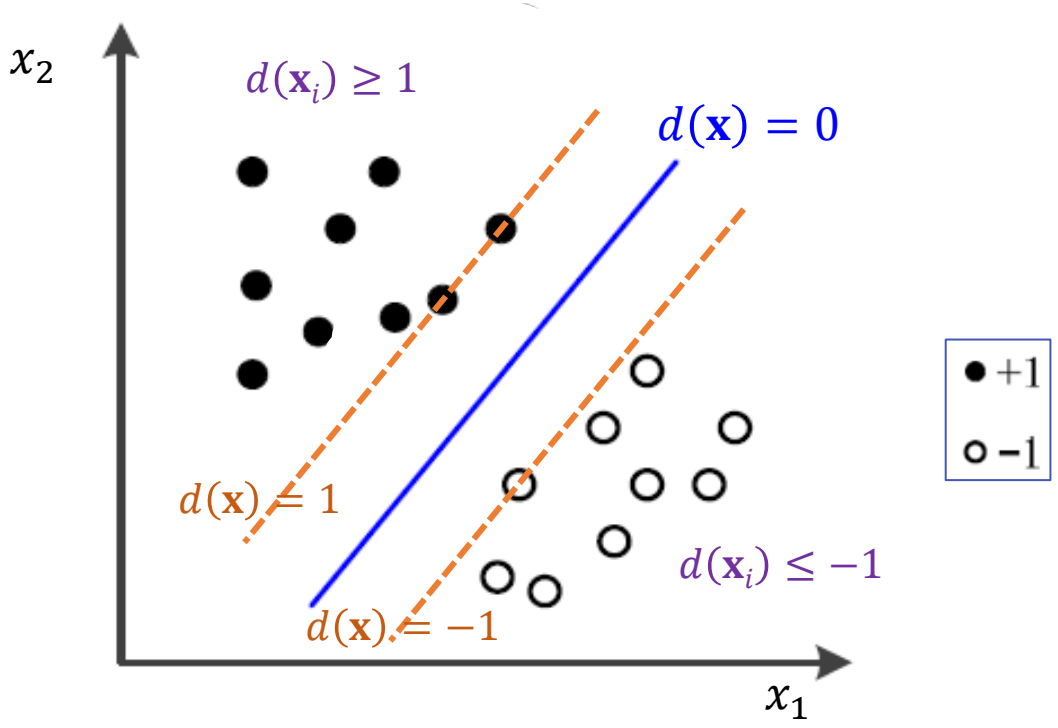
$$d(\mathbf{x}_i) = \begin{cases} 1, & if\ y_i = 1 \\ -1, & if\ y_i = -1 \end{cases}$$

→ $y_i d(\mathbf{x}_i) = 1$ → $y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$

★ $y_i = 1$인 샘플($\mathbf{x}_i$)의 경우: $d(\mathbf{x}_i) \geq 1$

★ $y_i = -1$인 샘플($\mathbf{x}_i$)의 경우: $d(\mathbf{x}_i) \leq -1$

★ 분할띠 내부 위치의 $\mathbf{x}$에 대해: $|d(\mathbf{x})| < 1$



Linear SVM – *Formulation*

Problem Definition

- 훈련 집합: $\mathbb{X} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- 주어진 데이터를 오류없이 분류하면서 여백을 최대로하는 파라메터 $\mathbf{w}$와 b를 찾아라.

Maximize: $$J(\mathbf{w}) = \frac{2}{\|\mathbf{w}\|_2}$$

Subject to: $$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \geq 1, \forall y_i = 1$$ ← $d(x_i) \geq 1$ for $\forall y_i = 1$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \leq -1, \forall y_i = -1$$ ← $d(x_i) \leq -1$ for $\forall y_i = -1$

Constrained L_2-Minimization 문제로 바꿔 쓰면,

Minimize: $$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$$

Subject to: $$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 \geq 0, i = 1,2,\cdots,n$$

해의 유일성과 전역성

- 선형 부등식 영역 내에서 L_2 Loss는 convex.
- 따라서 해는 유일하며 전역 최적해가 된다.

문제의 난이도

- n개의 선형 부등식을 제약조건으로 가진 2차 함수의 최적화 문제
- 조건부 최적화 문제는 Lagrange Multiplier 방법으로 푼다.

Linear SVM – *Lagrange Method*

minimize $f(\underset{\sim}{x})$
subject to $g(\underset{\sim}{x}) \leq 0$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$$

- $g(\boldsymbol{x}) = 0,\ \lambda \geq 0$ (*active*)
- $g(\boldsymbol{x}) < 0,\ \lambda = 0$ (*inactive*)

◆ **Lagrangian**

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\|\mathbf{w}\|_2^2}{2} - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1\right)$$

◆ **KKT Condition → 결국 α_i와 Support Vector를 구하는 문제**

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

→ α_i와 $\mathbf{x}_i$를 알면 $\mathbf{w}$를 구할 수 있음
($\alpha_i > 0$인 경우의 $\mathbf{x}_i$만 알면 됨)

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \qquad i = 1, 2, \cdots, n$$

- $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 > 0$ and $\alpha_i = 0$ (inactive)
- $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 = 0$ and $\alpha_i \geq 0$ (active)

(이 식을 만족하는 샘플($\mathbf{x}_i$)들이 support vector)

$\mathbf{w}$와 $\mathbf{x}_i$를 알면 b를 구할 수 있음:

$$b = y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) = 1$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = y_i$$

Linear SVM – *Dual Problem*

Wolfe Dual Problem

- Convex 성질을 만족하는 조건부 최적화 문제는 Wolfe Dual로 변형할 수 있다. (Lagrangian Duality 관련 강의 참조)
- Wolfe Dual로 바꾸면 부등식 조건이 등식 조건으로 바뀌어 풀기에 유리해 진다.

Max. $$\tilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$$

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

$$0 \le \alpha_i, \qquad i = 1,2,\cdots,n$$

- 2차 함수의 최대화 문제이다.
- **w**와 b가 사라지고, **α**를 찾는 문제가 되었다.
  (**α**를 찾으면 앞 슬라이드의 수식을 통해
  w와 b를 구할 수 있다.)
- 특징 벡터 $\mathbf{x}_i$가 내적 형태로 나타난다.
  (비선형으로 확장하는 발판)
- 목적 함수의 두번째 Σ항은 n^2개의 항을 갖는다.
  (따라서 효율적인 최적화 알고리즘이 필요하다.)
  - 예: 샘플이 6만 개인 MNIST는 36억개 항이 발생

Linear SVM – *Example*

- 훈련집합의 샘플 개수가 3개인 경우, Linear SVM의 해를 구해 보자.

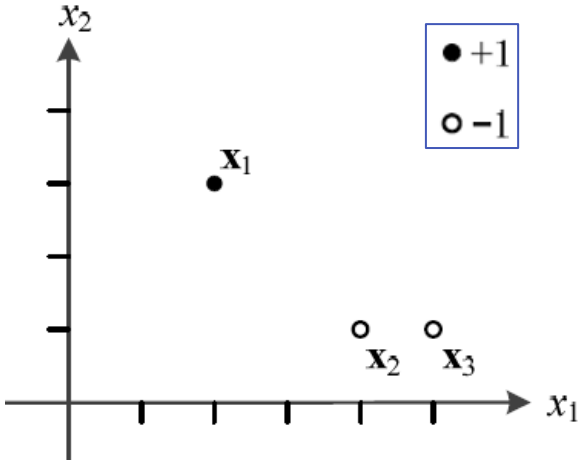
  - **훈련 집합:** $\mathbf{x}_1 = \begin{pmatrix}2\\3\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}4\\1\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}5\\1\end{pmatrix}, y_1 = 1, y_2 = -1, y_3 = -1$

  - **Primal Problem:**

$$\min \quad \underbrace{\frac{1}{2}\|\underline{w}\|_2^2}_{f(\underline{w},b)} - \sum_{i=1}^{3} \alpha_i \underbrace{\left(y_i(\underline{w}^T\underline{x}_i + b) - 1\right)}_{-g(\underline{w},b)} = \mathcal{L}(\underline{w}, b, \underline{\alpha})$$

$$\text{s.t.} \quad y_i(\underline{w}^T\underline{x}_i + b) \geq 1, \quad i = 1,2,3 \quad \leftarrow \text{등호}$$

$$\alpha_i \geq 0, \quad i = 1,2,3$$



KKT

Stationarity

$$\begin{cases} \nabla_{\underline{w}}\mathcal{L}(\underline{w}, b, \underline{\alpha}) = \underline{0} \Rightarrow \underline{w} = \sum_{i=1}^{3} \alpha_i y_i \underline{x}_i \\ \nabla_b \mathcal{L}(\underline{w}, b, \underline{\alpha}) = 0 \Rightarrow \sum_{i=1}^{3} \alpha_i y_i = 0 \end{cases}$$

Complementary Slackness

$$\begin{cases} * \text{ NSV}: \quad y_n(\underline{w}^T\underline{x}_i + b) > 1, \quad \alpha_i = 0 \\ * \text{ SV}: \quad y_i(\underline{w}^T\underline{x}_i + b) = 1, \quad \alpha_n \geq 0 \end{cases}$$

Linear SVM – *Example (cont'd)*

- Wolfe Dual Problem으로 바꿔서 해를 구해 보자.

Maximize: $\tilde{\mathcal{L}}(\boldsymbol{\alpha}) = (\alpha_1 + \alpha_2 + \alpha_3) - \frac{1}{2}(13\alpha_1^2 + 17\alpha_2^2 + 26\alpha_3^2 - 22\alpha_1\alpha_2 - 26\alpha_1\alpha_3 + 42\alpha_2\alpha_3)$

Subject to: $\alpha_1 - \alpha_2 - \alpha_3 = 0$

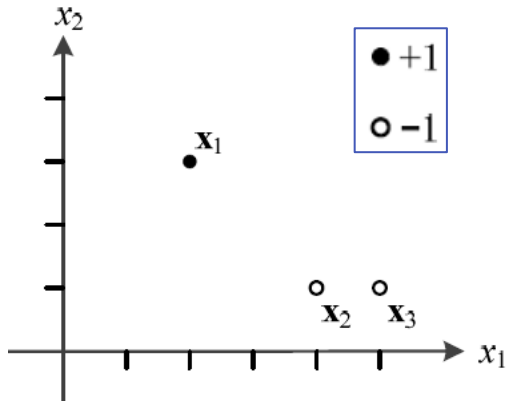
$0 \le \alpha_1, 0 \le \alpha_2, 0 \le \alpha_3$

- **풀이:**
  - ★ 구해야 하는 미지수가 3개뿐이므로 경우를 일일이 나열하여 풀 수 있다.
  - ★ Class별로 Support Vector가 하나 이상 있어야 하므로, 다음의 세 가지 경우만 가능하다.

(1) $\alpha_1 \neq 0, \alpha_2 = 0, \alpha_3 \neq 0$ (SV : x_1, x_3)

(2) $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 = 0$ (SV : x_1, x_2)

(3) $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 \neq 0$ (SV : x_1, x_2, x_3)



Linear SVM – *Example (cont'd)*

(1) $\alpha_1 \neq 0,\ \alpha_2 = 0,\ \alpha_3 \neq 0$

조건 $\alpha_1 - \alpha_2 - \alpha_3 = 0$으로부터 $\alpha_1 = \alpha_3$이다. 이것을 $\tilde{\mathcal{L}}(\boldsymbol{\alpha})$에 대입하여 정리하면 다음 식을 얻는데, 이 식은 $\alpha_1 = \frac{2}{13}$에서 최댓값을 가진다. 따라서 답은 $\alpha_1 = \frac{2}{13},\ \alpha_2 = 0,\ \alpha_3 = \frac{2}{13}$이다.

$$\tilde{\mathcal{L}}(\boldsymbol{\alpha}) = 2\alpha_1 - \frac{13}{2}\alpha_1^2 = -\frac{13}{2}\left(\left(\alpha_1 - \frac{2}{13}\right)^2 - \frac{4}{169}\right)$$

i=1과 3에 대해, $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 = 0$ 이므로, $\mathbf{w} = \left(-\frac{6}{13}, \frac{4}{13}\right)^{\mathrm{T}},\ b = 1$

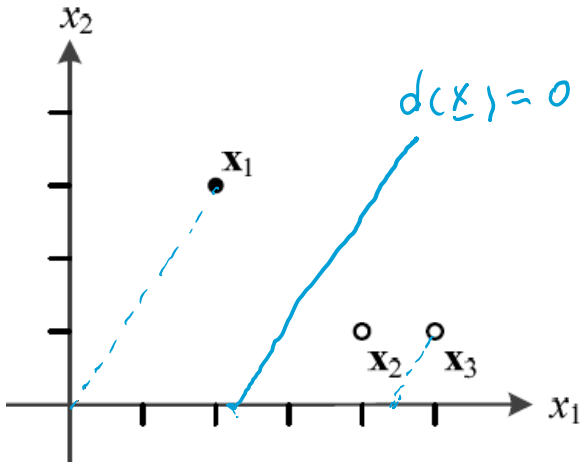
$d(\mathbf{x}) = -\frac{6}{13}x_1 + \frac{4}{13}x_2 + 1$ 이므로,

$d(\mathbf{x}_1) = 1,\ d(\mathbf{x}_2) = -\frac{7}{13},\ d(\mathbf{x}_3) = -1$

확인해 보면 $\mathbf{x}_2$는 분할 띠 안에 있다.

$y_i d(x_i) > 1$ 조건 만족 못함

(SVM의 해가 아님)



$x_2 = \frac{3}{2}x_1 - \frac{13}{4}$

x1과 x3가 support vector라는 조건 하에 문제를 푸는 것이므로 x2는 분할띠 바깥에 있어야함. 따라서 y2 * d(x2)가 1보다 커야 되는 데, 이 조건을 만족하지 못하므로 해가 되지 못함.

Linear SVM – *Example (cont'd)*

(2) $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 = 0$

(1)과 마찬가지 방법으로 풀면,

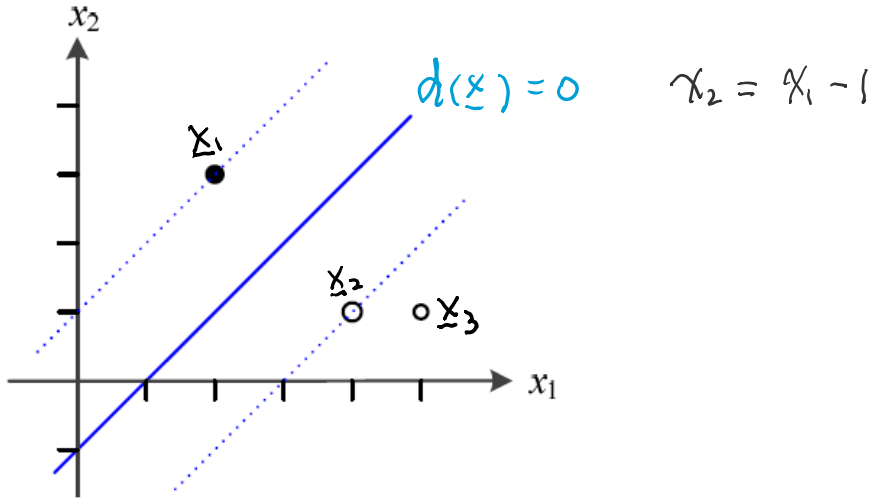
$\alpha_1 = \frac{1}{4}, \alpha_2 = \frac{1}{4}, \alpha_3 = 0$

$\mathbf{w} = \left(-\frac{1}{2}, \frac{1}{2}\right)^T$ 이고 $b = \frac{1}{2}$

$d(\mathbf{x}) = -\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}$

$|d(\mathbf{x}_1)| = |d(\mathbf{x}_2)| = 1$이므로 $\mathbf{x}_1$과 $\mathbf{x}_2$는 서포트 벡터

$d(\mathbf{x}_3) < -1$이므로 $\mathbf{x}_3$는 분할 띠 바깥 ⇒ $y_i d(x_i) > 1$ 조건 만족!



$x_2 = x_1 - 1$

(b) SVM으로 구한 결정 직선

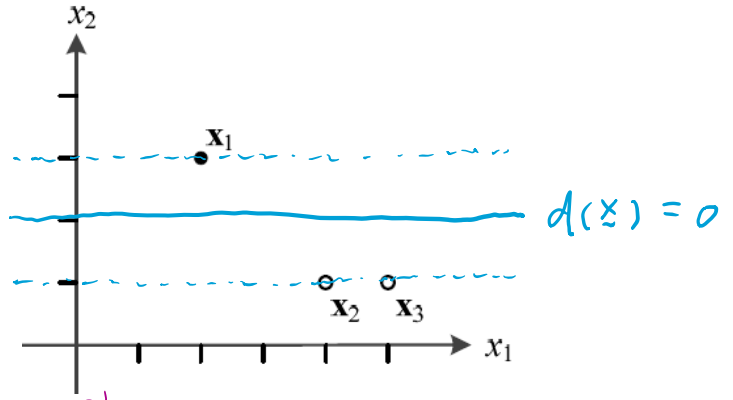
x1과 x2가 support vector라는 조건 하에 문제를 푸는 것이므로 x3가 분할띠 바깥에 있어야 함. 즉 y3 * d(y3)가 1보다 커야 되는데, 이 조건을 만족하므로 해가 될 수 있음

(3) $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 \neq 0$

(1)과 마찬가지 방법으로 풀면, … (연습해 보세요!)

$\alpha_1 = \frac{1}{2}, \alpha_2 = \frac{3}{2}, \alpha_3 = -1$ ⇒ $\alpha_3 > 0$ 조건만족 못함

(SVM의 해가 아님)



(Hint: $\alpha_1 = \alpha_2 + \alpha_3$ 이므로 $L(\alpha)$는 α_2와 α_3에 관한 식으로 바꿈. 이때 $\frac{\partial L}{\partial \alpha_2} = 0$, $\frac{\partial L}{\partial \alpha_3} = 0$ 을 풀면 α_2와 α_3를 구할 수 있음)

Q & A