I. Support vector machine이 뭐였인가?
II. Convex optimization
III. Convex optimization을 SVM을 이용하는 과정.

Lecture 11: Support Vector Machine I (SVM)

[SCS4049-02] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

AI Department, Dongguk University

Tentative schedule

week	topic	date (수 / 월)
1	Machine Learning Introduction & Basic Mathematics	09.02 / 09.07
2	Python Practice I & Regression	09.09 / 09.14
3	AI Department Seminar I & Clustering I	09.16 / 09.21
4	Clustering II & Classification I	09.23 / 09.28
5	Classification II	(추석) / 10.05
6	Python Practice II & Support Vector Machine I	10.07 / 10.12
7	Support Vector Machine II & Ensemble Learning and Random Forest	10.14 / 10.19
8	(휴강) & **Mid-term exam**	10.21 / **10.26**
9	Neural networks	10.28 / 11.02
10	Backpropagation	11.04 / 11.09
11	Convolutional Neural Network	11.11 / 11.16
12	Model Optimization	11.18 / 11.23
13	Recurrent Neural network	11.25 / 11.30
14	Autoencoders	12.02 / 12.07
15	**Final exam**	(휴강) / **12.14**

중간고사

10월 26일 월요일 오전 10시 ~ 12시. 수업시간.

비대면, webex에서 모두 접속한 후에 파일을 공유. 실험

준비물: 카메라(웹캠, 노트북, 혹은 대포)

문의 할 경우 -> 저한테 메일 s.park@dgu.edu

방법: 책상 근처, 본인의 얼굴이 나오도록 카메라 설정

시험 종료 후에는 카메라에 보이지 않도록 치우고

스캔앱으로 종료 후 바로 제출

답지 실물도 제 사무실로 제출 ~ 만해관 210호.

오픈북 X.

서술형 + 간단한 문제풀이. ~~Python 프로그래밍~~

강의자료만 보지마세요,

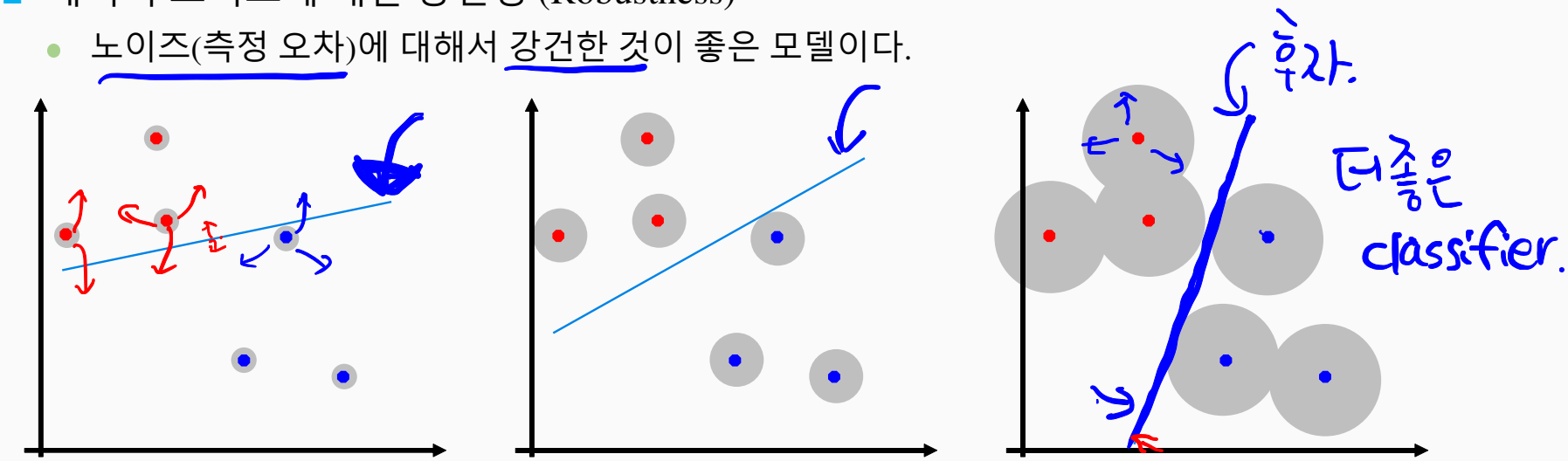
강의내용 전체가 시험범위

(강의자료 + 필기내용 + 강의내용 전부다).

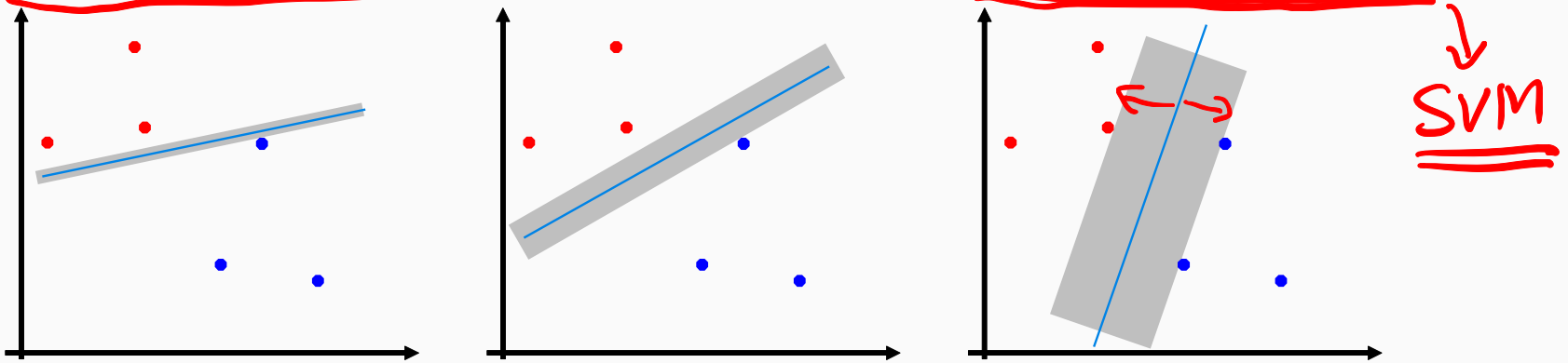
SVM: Large Maximum Margin Classifier

What is a good decision boundary?

- 데이터 노이즈에 대한 강건성 (Robustness)
  - 노이즈(측정 오차)에 대해서 강건한 것이 좋은 모델이다.



- 여유로운 것이 더 강건하다 ⇒ 넓은 통로가 좋다 ⇒ Large Margin Classification



통로 = margin

What is a good decision boundary?

decision boundary

의사 결정은 경계의 데이터(support vectors)에 의해서 결정됨

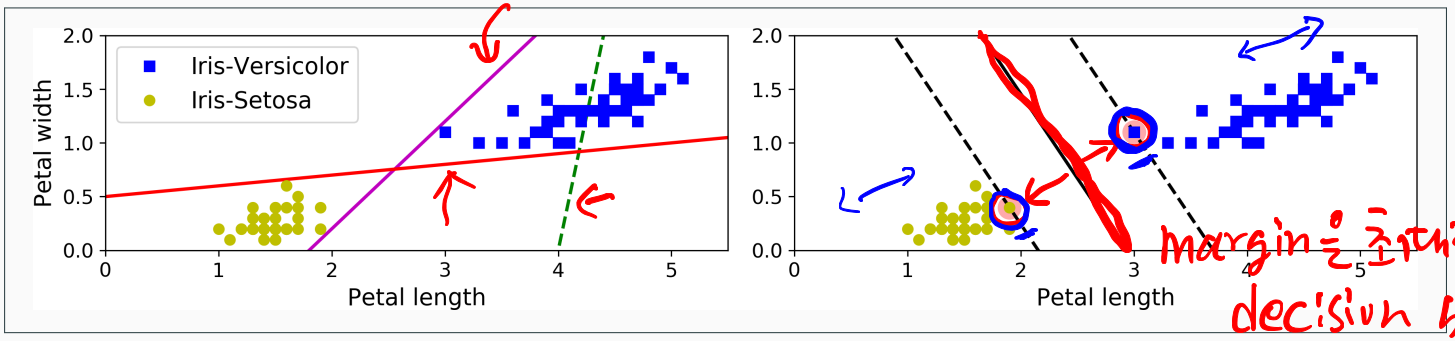
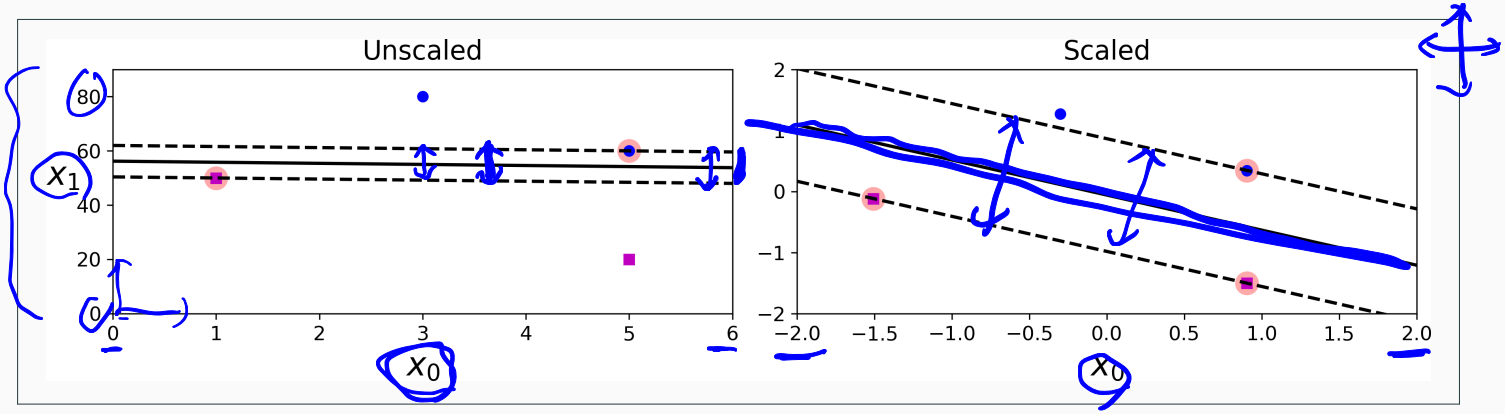


Figure 5-1. Large margin classification

margin을 최대화하는 decision boundary를 사용할 수 있다.

Input feature의 scale에 민감한 support vector machine

normalize, standardize가 필요하다·중요하다.



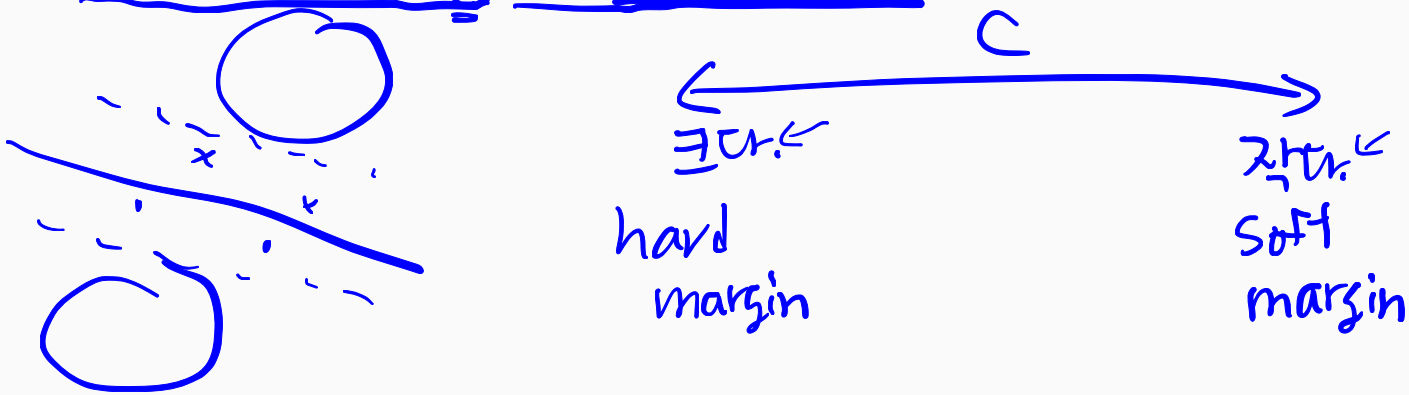
range

$x_1 > x_0$

Figure 5-2. Sensitivity to feature scales

Hard margin vs. soft margin

- Hard margin classification (hard-SVM) ← SVM I
  - 모든 데이터들이 margin 밖에 위치하도록 boundary를 설정
  - 데이터가 선형적으로 분리 가능(linearly separable)할 때만 적용 가능
  - outlier에 매우 민감
- Soft margin classification (soft-SVM) ← SVM II
  - margin을 가능한 넓게 하면서도 margin 안쪽으로 들어오는 것을 허용
  - hyperparameter C: 클수록 좁아짐 (엄격), 작을 수록 넓어짐 (위반 허용)



Hard margin vs. soft margin

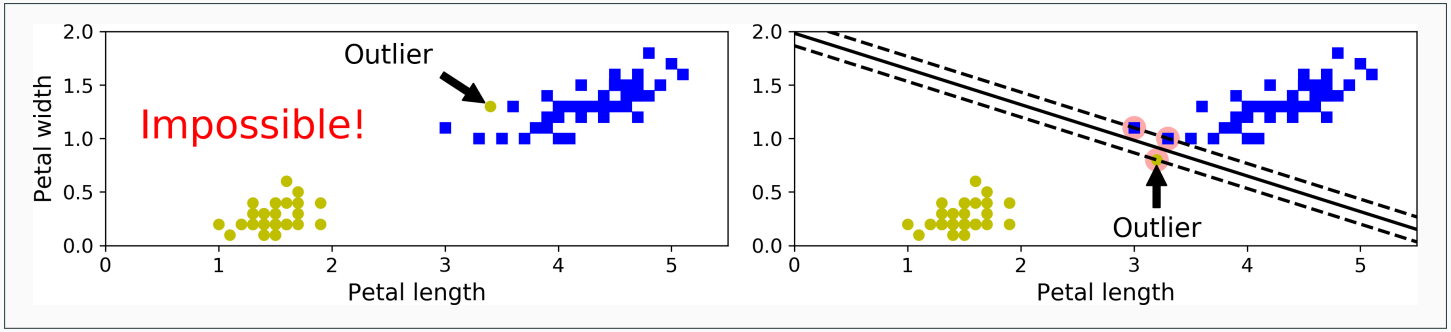


Figure 5-3. Hard margin sensitivity to outliers

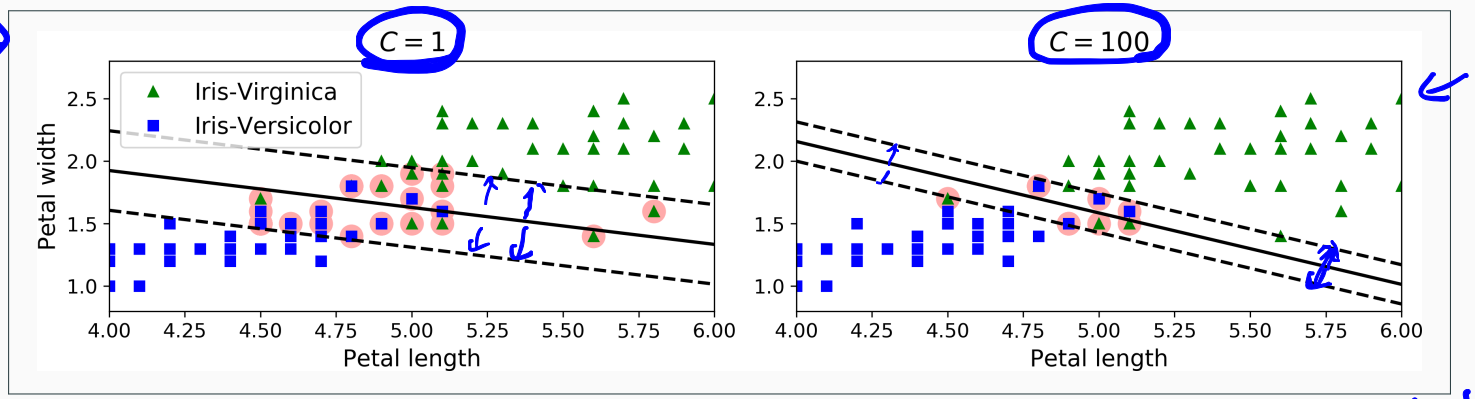


Figure 5-4. Large margin (left) versus fewer margin violations (right)

A brief history of SVM

- SVM은 1992년 Boser, Guyon and Vapnik에 의해서 소개됨
- Statistical Learning Theory에 이론적 바탕을 둔 알고리즘 (Vapnik Chervonenkis)
- 손글씨 숫자 인식에서 뛰어난 성능을 보이면서 널리 쓰이게 됨
- SVM으로 1.1% Test error rate ≈ 신중히 설계된 신경망(e.g., LeNet 4)과 맞먹음
- 실용적으로 우수한 성능
- bioinformatics, text, image recognition 등을 포함한 많은 성공 사례
- 강력하고 다재 다능한 머신 러닝 모델
- 선형/비선형 분류 뿐 아니라 회귀, outlier detection 도 수행
- 복잡한 소규모/중규모 데이터셋의 분류에 특히 잘 맞음
- 머신 러닝에서 중요한 기법 중 하나인 Kernel 방법을 사용하는 대표적 알고리즘
- 머신 러닝에서 가장 널리 쓰이는 모델이며 머신 러닝을 하며 반드시 알아야 할 기법 중 하나

→ Liblinear libsvm: Scikit-Learn 에서 liblinear 및 libsvm 을 사용하여 구현

Convex Optimization and Duality

convex optimization problem

↳ dual problem, duality : Lagrangian,

↳ KKT condition → support vector.

MoG: Lagrange multiplier

EM method.

optimization.

The method can be summarized as follows: in order to find the maximum or minimum of a function $f(x)$ subjected to the equality constraint $g(x) = 0$, form the Lagrangian function

constraint. 구속조건.

Lagrangian fn. → $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$ (1)

inequality $<0, >0$.

Lagrangian multiplier.

and find the stationary points of $\mathcal{L}$ considered as a function of x and the Lagrange multiplier λ. The solution corresponding to the original constrained optimization is always a saddle point of the Lagrangian function, which can be identified among the stationary points from the definiteness of the bordered Hessian matrix.

$\frac{\partial L}{\partial x} = 0$, $\frac{\partial L}{\partial \lambda} = 0$ ⇝ solution.

MoG: Lagrange multiplier

Minimize $f(x, y) = x + y$ subject to the constraint $x^2 + y^2 = 1$, i.e.,

$$g(x, y) = x^2 + y^2 - 1 = 0 \tag{2}$$

Hence,

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda g(x, y) = x + y + \lambda(x^2 + y^2 - 1) \tag{3}$$

Gradient

$$\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) = (1 + 2\lambda x, 1 + 2\lambda y, x^2 + y^2 - 1) \tag{4}$$

and therefore,

$$\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) \iff \begin{cases} 1 + 2\lambda x = 0 \\ 1 + 2\lambda y = 0 \\ x^2 + y^2 - 1 = 0 \end{cases} \tag{5}$$

MoG: Lagrange multiplier

$$\nabla_{x,y,\lambda}\mathcal{L}(x,y,\lambda) \iff \begin{cases} 1+2\lambda x = 0 \\ 1+2\lambda y = 0 \\ x^2+y^2-1=0 \end{cases} \tag{6}$$

This yields

$$x = y = -\frac{1}{2\lambda}, \quad \lambda \neq 0 \tag{7}$$

$$\frac{1}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 = 0 \tag{8}$$

So,

$$\lambda = \pm\frac{1}{\sqrt{2}} \tag{9}$$

$\nabla \mathcal{L} = 0.$

which implies that the stationary points of $\mathcal{L}$ are

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right), \quad \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) \tag{10}$$

Optimization problem in standard form

(general)

$$\begin{aligned} &\text{minimize} && f_0(x) && && (11)\\ &\text{subject to} && f_i(x) \le 0, && i = 1, 2, \ldots, m && (12)\\ &&& h_i(x) = 0, && i = 1, 2, \ldots, p && (13) \end{aligned}$$

- $x \in \mathcal{R}^n$ is the optimization variable
- $f_0 : \mathcal{R}^n \to \mathcal{R}$ is the objective or cost function
- $f_i : \mathcal{R}^n \to \mathcal{R}, i = 1, 2, \ldots, m$ are the inequality constraint functions
- $h_i : \mathcal{R}^n \to \mathcal{R}$ are the equality constraint functions

$x^* = \arg\min_{\text{s.t.}} f_0(x)$

Convex optimization problem

Standard form convex optimization problem

$$\text{minimize} \quad f_0(x) \tag{14}$$

$$\text{subject to} \quad f_i(x) \leq 0, \qquad i = 1, 2, ..., m \tag{15}$$

$$a_i^T x = b_i, \qquad i = 1, 2, ..., p \tag{16}$$

- $f_0, f_1, ..., f_m$ are convex
- equality constraints are affine

Often written as

$$\text{minimize} \quad f_0(x) \tag{17}$$

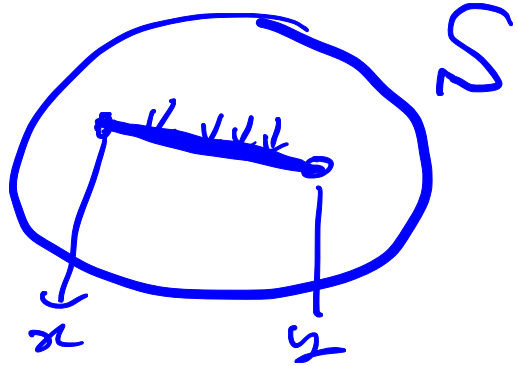
$$\text{subject to} \quad f_i(x) \leq 0, \qquad i = 1, 2, ..., m \tag{18}$$

$$Ax = b \tag{19}$$

Important property: feasible set of a convex optimization problem is convex

✡ Convex

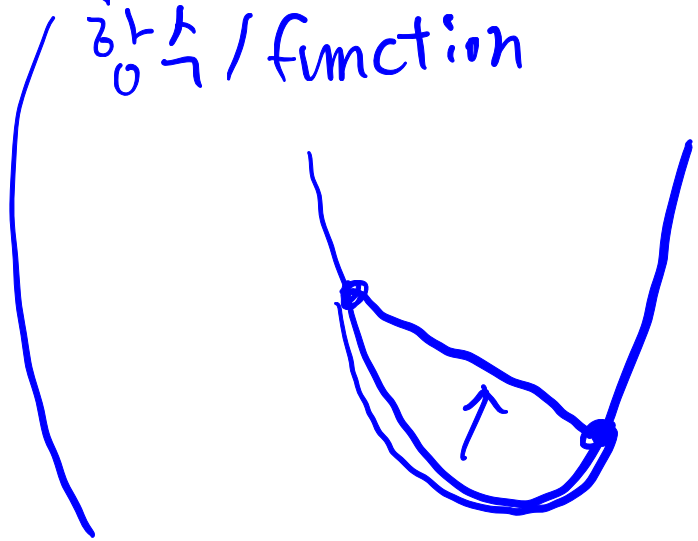
set / 집합.



Convex set : $x, y \in S$, $\underline{\theta x + (1-\theta) y \in S}$

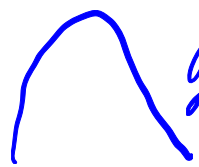
$\underline{0 \le \theta \le 1}$

함수 / function



$x, y \in \text{dom} f$.

$f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$



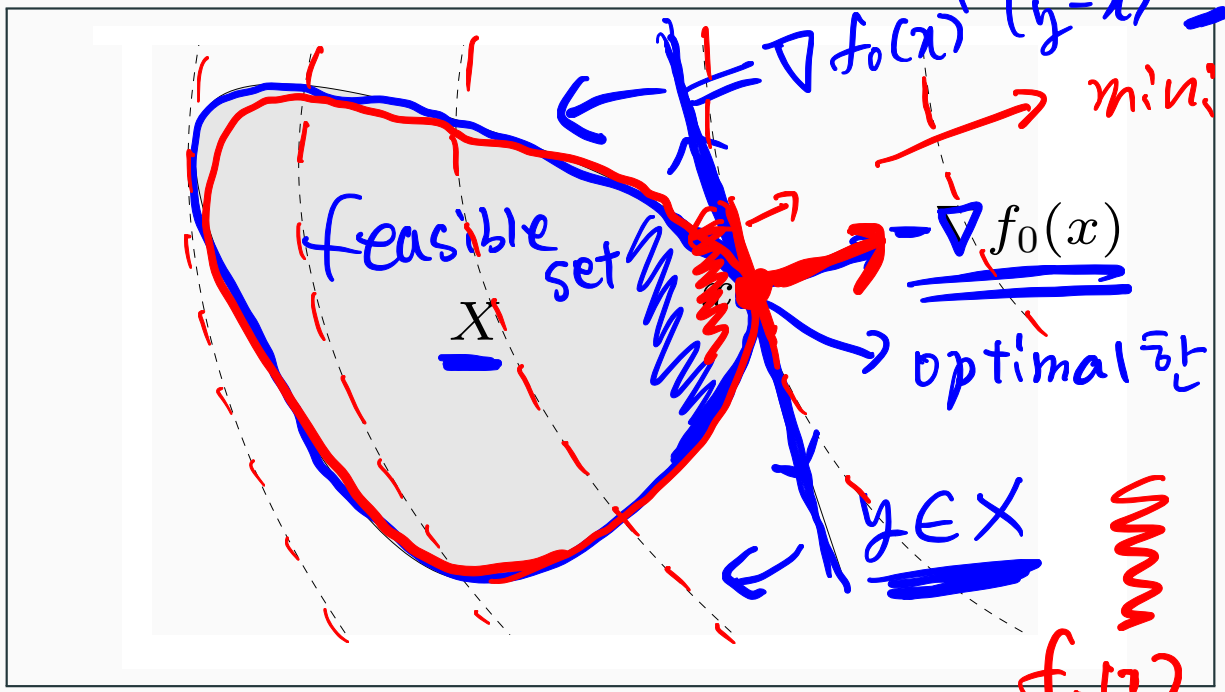
$g(x)$, $-g(x)$: convex

$g(x)$: concave.

Optimality criterion for differentiable f_0

x is optimal if and only if it is feasible and

$$\nabla f_0(x)^T(y-x) \geq 0 \quad \text{for all feasible } y \tag{20}$$



if nonzero, $\nabla f_0(x)$ defines a supporting hyperplane to feasible set X at x

Lagrangian

standard form problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) && (21)\\ \text{subject to} \quad & f_i(x) \le 0, \qquad i = 1, 2, ..., m && (22)\\ & h_i(x) = 0, \qquad i = 1, 2, ..., p && (23) \end{aligned}$$

variable $x \in \mathcal{R}^n$, domain $\mathcal{D}$, optimal value p^*

Lagrangian: $L : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{R}^p \to \mathcal{R}$ with $\mathbf{dom}\, L = \mathcal{D} \times \mathcal{R}^m \times \mathcal{R}^p$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \qquad (24)$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(x) \le 0$
- ν_i is Lagrange multiplier associated with $h_i(x) = 0$

Lagrange dual function

Lagrange dual function: $g : \mathcal{R}^m \times \mathcal{R}^p \to \mathcal{R}$

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \tag{25}$$

$$= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \tag{26}$$

g is concave, can be $-\infty$ for some λ, ν

lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq p^*$

proof: if $\tilde{x}$ is feasible and $\lambda \geq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu) \tag{27}$$

minimizing over all feasible $\tilde{x}$ gives $p^* \geq g(\lambda, \nu)$

The dual problem

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \quad (28) \\ \text{subject to} & \lambda \geq 0 \quad (29) \end{array}$$

- finds best lower bound on p^*, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted d^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \operatorname{dom} g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \operatorname{dom} g$ explicit

original optimization problem

$$\min f_0(x) \quad \text{s.t. } f_i(x) \le 0, \; h_i(x) = 0.$$

$\min f_0(x) = p^*$

Lagrangian dual problem

$$\max g(\lambda, \nu) \quad \text{s.t. } \lambda \ge 0$$

$\max g(\lambda, \nu) = d^*$

$$\mathcal{L}(x, \lambda, \nu) \leftarrow \text{Lagrangian}$$

$$= f_0 + \sum \lambda_i f_i + \sum \nu_i h_i$$

Lagrangian multiplier

$$g(\lambda, \nu) = \min_x \mathcal{L}(x, \lambda, \nu)$$

↳ Lagrangian dual function.

Weak and strong duality

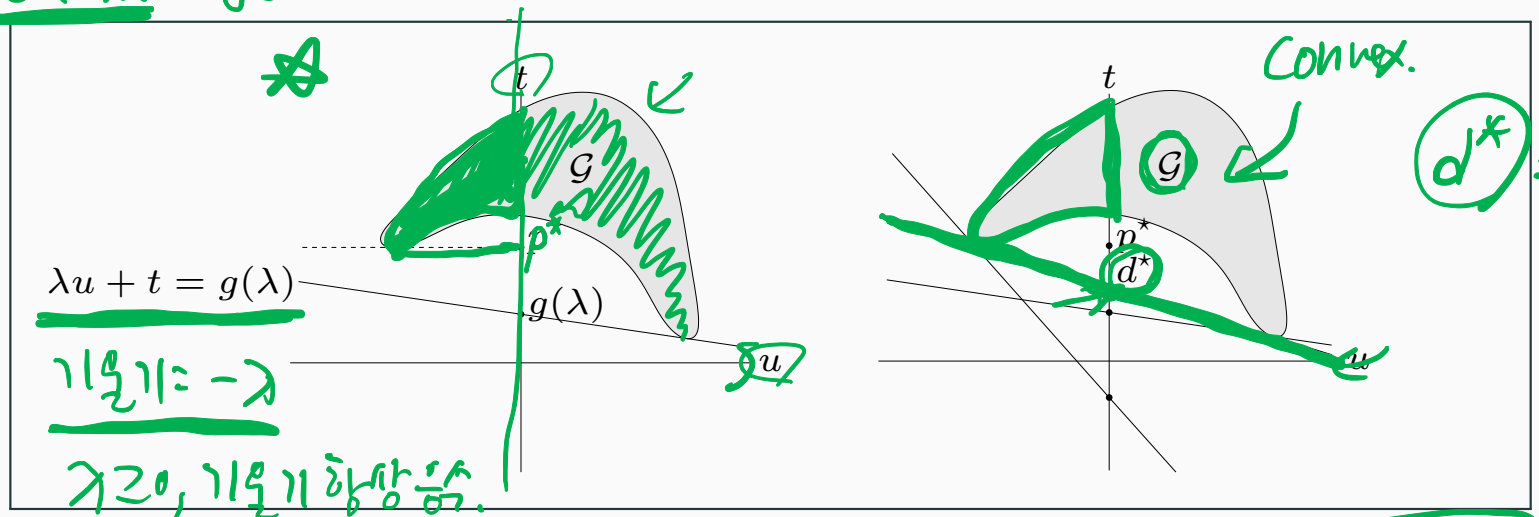
- weak duality: $d^* \leq p^*$
  - always holds (for convex and nonconvex problems)
  - can be used to find nontrivial lower bounds for difficult problems
- strong duality: $d^* = p^*$
  - does not hold in general
  - holds for convex problems
  - conditions that guarantee strong duality in convex problems are called constraint qualifications

Geometric interpretation (weak duality)

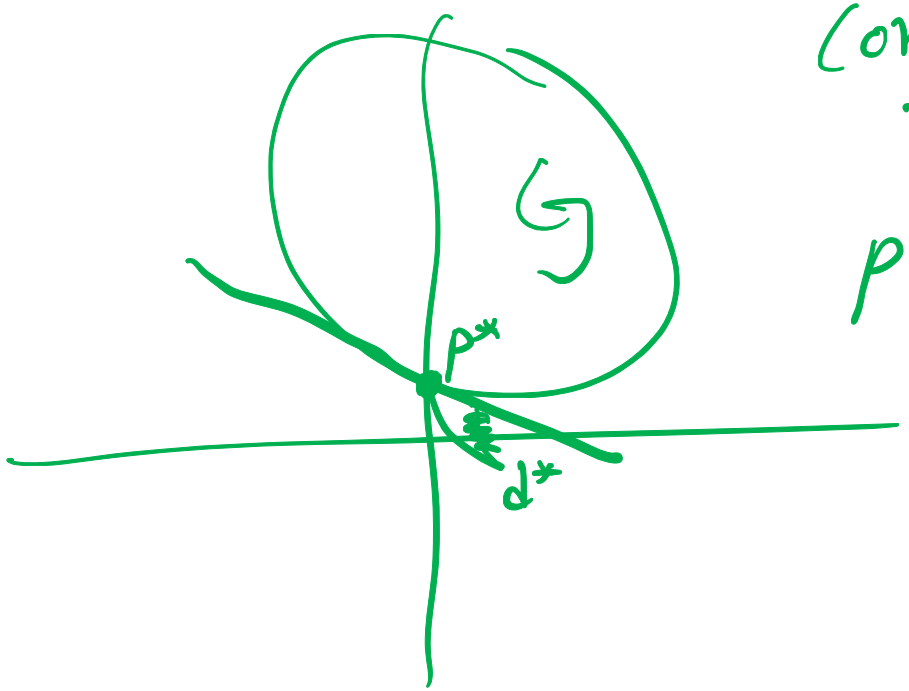
for simplicity, consider problem with one constraint $f_1(x) \leq 0$

interpretation of dual function (u, t)

$$g(\lambda) = \inf_{(u,t)\in\mathcal{G}} (t + \lambda u) \qquad \text{where } \mathcal{G} = \{(f_1(x), f_0(x)) \mid x \in \mathcal{D}\} \qquad (30)$$



- $\lambda u + t = g(\lambda)$ is supporting hyperplane to $\mathcal{G}$
- hyperplane intersects t-axis at $t = g(\lambda)$



Convex problem
과 같은 경우.

$p^* = d^*$

Geometric interpretation

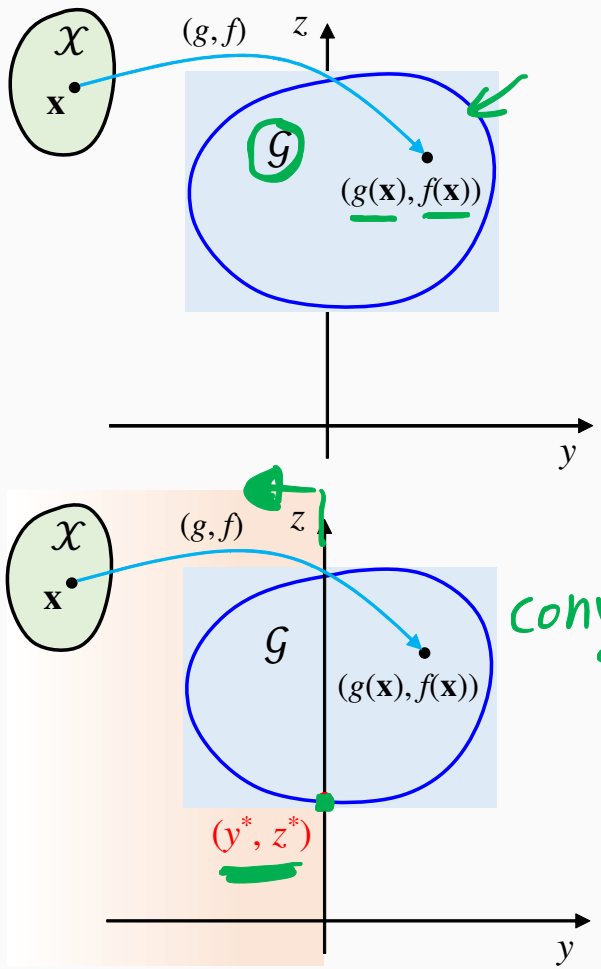
original

convex problem

- Primal problem:

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \\ &\text{subject to } g(\mathbf{x}) \leq 0, \\ &\mathbf{x} \in \mathcal{X} \end{aligned}$$

where $f: \mathbb{R}^n \to \mathbb{R}, \quad g: \mathbb{R}^n \to \mathbb{R}^m$



convex set

- $n = 2$ 에 대해서, 집합 $\mathcal{G}$ 를 다음과 같이 정의하자.

$$\mathcal{G} = \{(y, z) \mid y = g(\mathbf{x}), z = f(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$$

$\Longrightarrow$ 그러면, $\mathcal{G}$ 는 사상 (g, f) 하에서 $\mathcal{X}$ 의 치역(image)이다.

$$(g, f): \mathcal{X} \to \mathcal{G}$$

- 그러면 Primal solution은 최소 세로 좌표값 z를 갖는 $y \leq 0$ 인 $\mathcal{G}$ 내의 점이다.

$\Longrightarrow$ clearly (y^*, z^*)

$z^* = \min f(x)$
$\text{s.t. } g(x) \leq 0.$

Geometric interpretation

- Lagrange Dual Problem

$$\text{maximize}_u \ \theta(u)$$
$$\text{subject to} \ \ u \geq 0$$

where (Lagrangian subproblem):

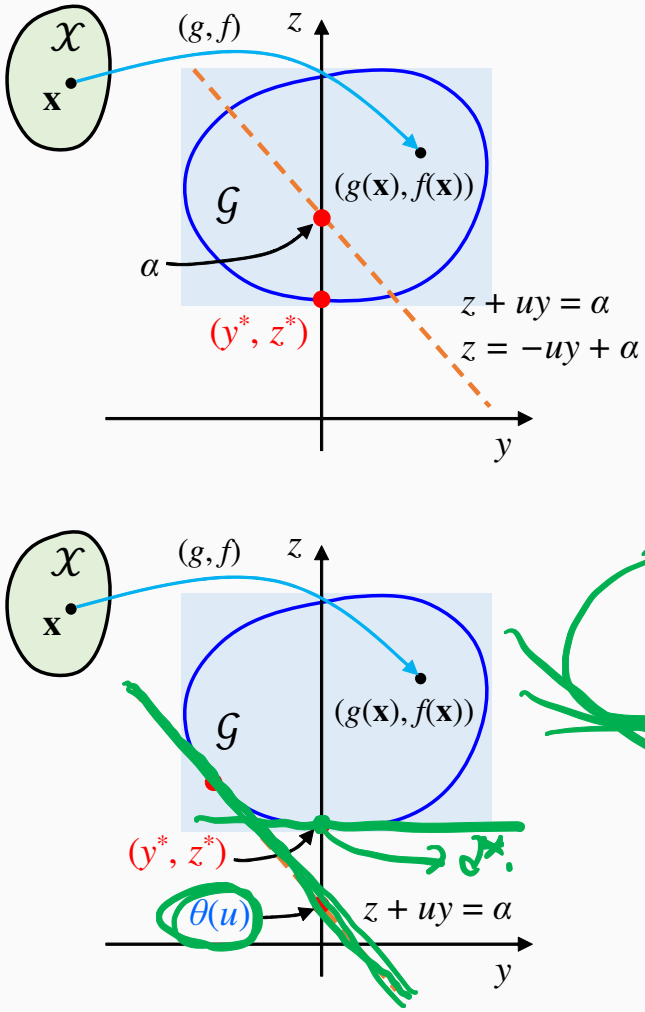
$$\theta(u) \ = \ \inf\{f(\mathbf{x}) + ug(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}.$$

1. $u \geq 0$ 일 때, Lagrangian dual subproblem 은 다음과 동등하다.

   minimize $z + uy$ over points (y, z) in $\mathcal{G}$,

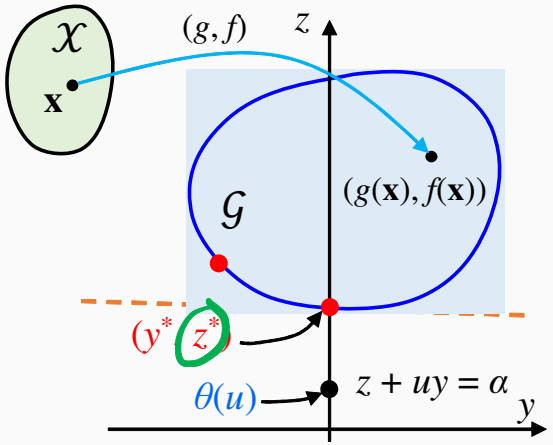
   여기서 $z + uy = \alpha$ 는 기울기가 $-u$ 이고 z 축과 만나는 점이 α 인 직선식이다.

2. $\mathcal{G}$ 에 대해서 $\theta(u) = z + uy$ 를 최소화하기 위해서 $\mathcal{G}$ 와의 접촉을 유지하면서 직선 $z + uy = \alpha$ 를 평행 이동해 내려가면, 최후로 얻어지는 z-축 절편값은 주어진 $u \geq 0$ 에 대응하는 $\theta(u)$ 값이다.



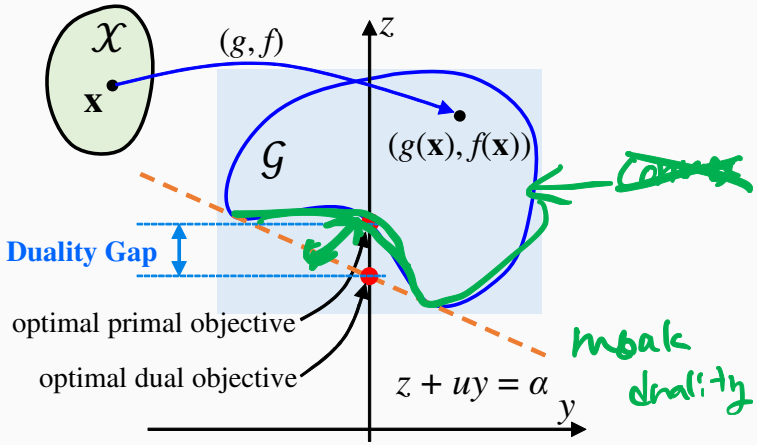
Geometric interpretation

3. 마지막으로 dual problem 을 풀기 위해서는, 최후의 z-축 절편 값 $\theta(u)$ 가 최대값이 되는 기울기 $-u$ $(u \geq 0)$ 를 찾아야 한다.
   이러한 직선은 기울기가 $-u^*$ 이고 점 (y^*, z^*) 에서 집합 $\mathcal{G}$ 를 지지(support)한다.
   그러므로, dual problem 의 해는 $-u^*$ 이며 최적 dual objective value는 z^* 이다.



- Primal problem 과 dual problem 의 최적해가 동일한 경우, duality gap 이 없다고 말한다 (**strong duality**).
- 동일하지 않은 경우는 duality gap 이 존재한다고 말한다 (**weak duality**).
- 적절한 convexity condition 들이 충족되면 primal 과 dual 최적화 문제에 duality gap 이 없다.
- [우측 그림] 집합 $\mathcal{G}$ 의 nonconvexity 로 인한 Duality Gap

$$f(\mathbf{x}) \geq \theta(\mathbf{u})$$



Karush-Kuhn-Tucker (KKT) conditions

convex 최적화 ⇒ 반드시 만족해야되는 조건. 4가지 = KKT condition.

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i)

1. primal constraints: $f_i(x) \leq 0$ i = 1, ..., m, $h_i(x) = 0$, i = 1, ..., p
2. dual constraints: $\lambda \geq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0, i = 1, ..., m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0 \tag{31}$$

if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

① $f_i(x) \leq 0$, ② $\lambda_i \geq 0$, ③ $\lambda_i f_i(x) = 0$ ($\lambda_i \geq 0$, $f_i(x) \leq 0$)

① $f_i(x) \le 0$

② $\lambda_i \ge 0$

③ $\lambda_i f_i(x) = 0$.

$\lambda_i \ge 0$, $f_i(x) \le 0$

$\lambda_i = 0$ or $f_i(x) = 0$

$f_i(x) \le 0$ / $\lambda_i > 0$

두 중에 하나는 0이다.

optimization.

$\min f_0$

s.t. $f_i \le 0$

complementary slackness.

↓

Support vector

SVM: max. margin classifier → convex optimization

SVM: Theory

Maximum margin classifier

We begin our discussion of support vector machines to the two-class classification problem using linear models of the form

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{32}$$

where $\phi(\mathbf{x})$ denotes a fixed feature-space transformation, and we have made the bias parameter b explicit.

The training data set comprises N input vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, with corresponding target values $t_1, t_2, ..., t_N$ where $t_n \in \{-1, 1\}$, and new data points $\mathbf{x}$ are classified according to the sign of $y(\mathbf{x})$

$y(\mathbf{x}) \geq 0 \longrightarrow t = 1$

$y(\mathbf{x}) < 0 \longrightarrow t = -1$

Maximum margin classifier

hard margin SVM

We shall assume that the training data set is linearly separable in feature space, so that by definition there exists at least one choice of the parameters $\mathbf{w}$ and b such that a function satisfies $y(\mathbf{x}_n) > 0$ for points having $t_n = +1$ and $y(\mathbf{x}_n) < 0$ for points having $t_n = -1$, so that $t_n y(\mathbf{x}_n) > 0$ for all training data points.

$t_n y(x_n) > 0$ 항상 만족.

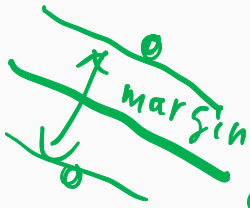
Maximum margin classifier: optimality criterion

b = location.

Thus the distance of a point $\mathbf{x}_n$ to the decision surface is given by

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}. \quad (33)$$

The margin is given by the perpendicular distance to the closest point $\mathbf{x}_n$ from the data set, and we wish to optimize the parameters $\mathbf{w}$ and b in order to maximize this distance. Thus the maximum margin solution is found by solving



$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n [t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + b)] \right\} \quad (34)$$

where we have taken the factor $1/\|\mathbf{w}\|$ outside the optimization over n because $\mathbf{w}$ does not depend on n.

Dual problem for convex optimiztion

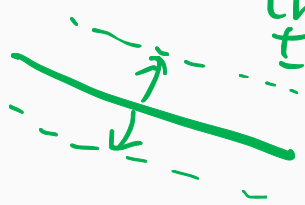
Direct solution of this optimization problem would be very complex, so we shall convert it into an equivalent problem that is much easier to solve.

→ Lagrangian dual problem

primal (original) problem

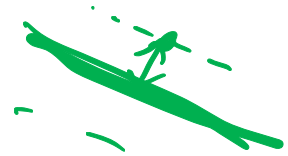
$$\max \frac{1}{\|w\|} \min \left[t_n (w^T \phi(x_n) + b) \right]$$

$$\text{s.t.} \quad t_n (w^T \phi(x_n) + b) > 0$$



$t_n (w^T \phi(x_n) + b) > 0$: small margin

$$\max \frac{1}{\|w\|} \min \left[t_n (w^T \phi(x_n) + b) \right]$$



$$\text{s.t.} \quad t_n (w^T \phi(x_n) + b) > \min \left[t_n (w^T \phi(x_n) + b) \right]$$

$$\Rightarrow \quad \max \frac{1}{\|w\|} = \max \frac{1}{\|w\|^2}$$

$$\text{s.t} \quad t_n (w^T \phi(x_n) + b) \geq 1.$$

$$-1 > 0.$$

$\Rightarrow$ Lagrangian dual problem

Lagrangian function with constraint

In order to solve this constrained optimization problem, we introduce Lagrange multipliers $a_n \geq 0$, with one multiplier a_n for each of the constraints, giving the Lagrangian function

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \{t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + b) - 1\} \tag{35}$$

where $\mathbf{a} = (a_1, ..., a_N)^T$. Note the minus sign in front of the Lagrange multiplier term, because we are minimizing with respect to $\mathbf{w}$ and b, and maximizing with respect to $\mathbf{a}$.

Setting the derivatives of $L(\mathbf{w}, b, \mathbf{a})$ with respect to $\mathbf{w}$ and b equal to zero, we obtain the following two conditions

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \tag{36}$$

$$0 = \sum_{n=1}^{N} a_n t_n \tag{37}$$

Lagrangian function with constraint

Eliminating $\mathbf{w}$ and b from $L(\mathbf{w}, b, \mathbf{a})$ using these conditions then gives the *dual representation* of the maximum margin problem in which we maximize

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \tag{38}$$

with respect to $\mathbf{a}$ subject to the constraints

$$a_n \geq 0, \qquad n = 1, ..., N \tag{39}$$

$$\sum_{n=1}^{N} a_n t_n = 0. \tag{40}$$

Here the kernel function is defined by $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$.

Prediction for a new sample: support vector machine

SVM: kernel method

$k(\mathbf{x}, \mathbf{x}_n) = \phi^T(\mathbf{x})\phi(\mathbf{x}_n)$ ⇒ 만족하는 함수인 커널

In order to classify new data points using the trained model, we evaluate the sign of $y(\mathbf{x})$. This can be expressed in terms of the parameter $\{a_n\}$ and the kernel function by substituting for $\mathbf{w}$ to give

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \tag{41}$$

Kernel

$k(\mathbf{x}, \mathbf{x}_n) = \phi^T(\mathbf{x})\phi(\mathbf{x}_n)$

$y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b.$

$= \sum a_n t_n \phi^T(\mathbf{x}_n)\phi(\mathbf{x}) + b$

KKT condition: complementary slackness

We show that a constrained optimization of this form satisfies the *Karush-Kuhn-Tucker* (KKT) conditions, which in this case require that the following three properties hold

$$a_n \geq 0 \tag{42}$$

$$t_n y(\mathbf{x}_n) - 1 \geq 0 \tag{43}$$

$$a_n\{t_n y(\mathbf{x}_n) - 1\} = 0 \tag{44}$$

Thus for every data point, either $a_n = 0$ or $t_n y(\mathbf{x}_n) = 1$. Any data point for which $a_n = 0$ will not appear in the sum and hence plays no role in making predictions for new data points. The remaining data points are called *support vectors* and because they satisfy $t_n y(\mathbf{x}_n) = 1$, they correspond to points that lie on the maximum margin hyperplanes in feature space.

KKT condition: complementary slackness

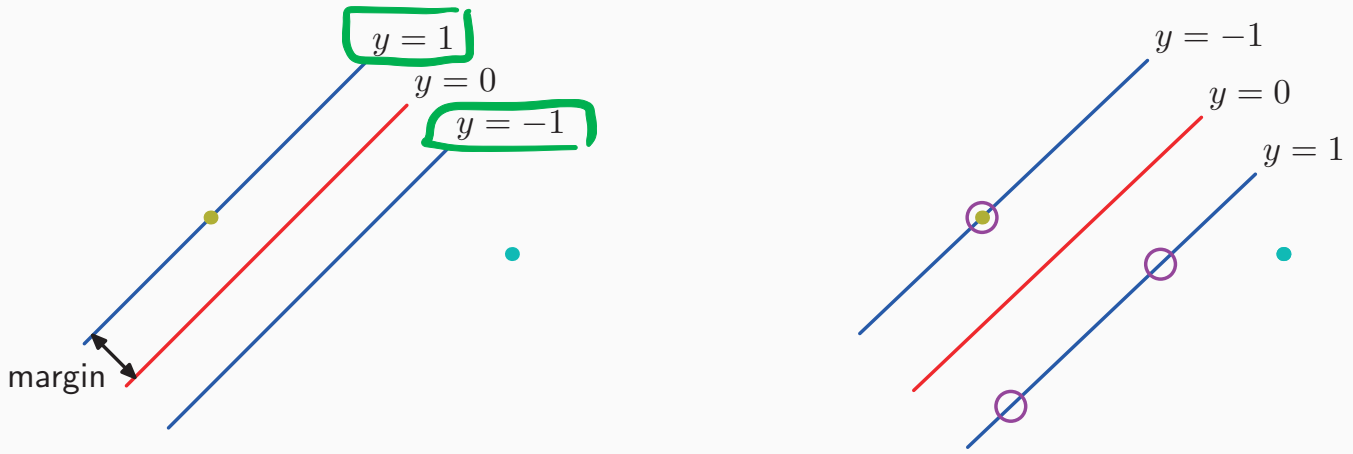
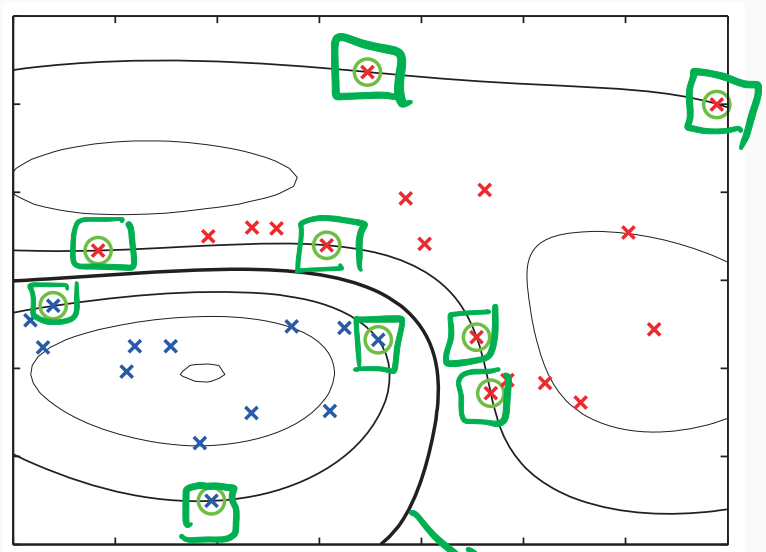


Figure 7.1 The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left figure. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right. The location of this boundary is determined by a subset of the data points, known as support vectors, which are indicated by the circles.

KKT condition: complementary slackness

Figure 7.2 Example of synthetic data from two classes in two dimensions showing contours of constant $y(\mathbf{x})$ obtained from a support vector machine having a Gaussian kernel function. Also shown are the decision boundary, the margin boundaries, and the support vectors.



$$y(x) = \sum_{n=1}^{N} t_n k(x_n, x) + b$$

Gaussian.

Appendix

Reference and further reading

- "Chap 7 | Sparse Kernel Machines" of C. Bishop, Pattern Recognition and Machine Learning
- "Chap 5 | Support Vector Machines" of A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- "Chap 4 | Convex Optimization Problems", "Chap 5 | Duality" of S. Boyd, Convex Optimization
- "Lecture 6 | Support Vector Machines" of Kwang Il Kim, Machine Learning (2019)