Lecture 11: Support Vector Machine I

[SCS4049-02] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

AI Department, Dongguk University

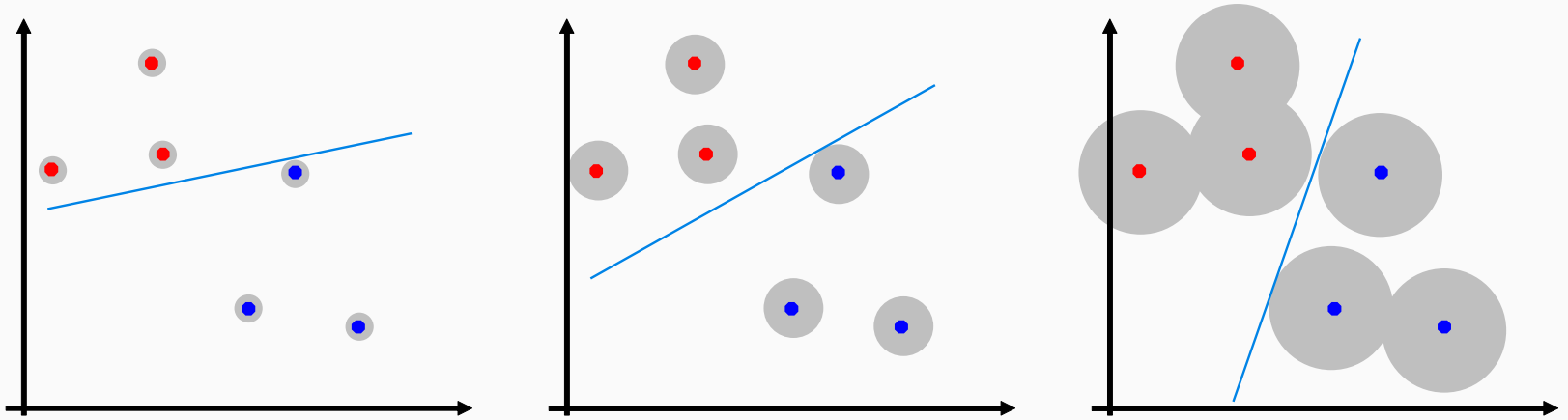
Tentative schedule

week	topic	date (ב / ה)
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	(יד) / 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	(ג סש) & **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**	(ג סש) / **12.14**

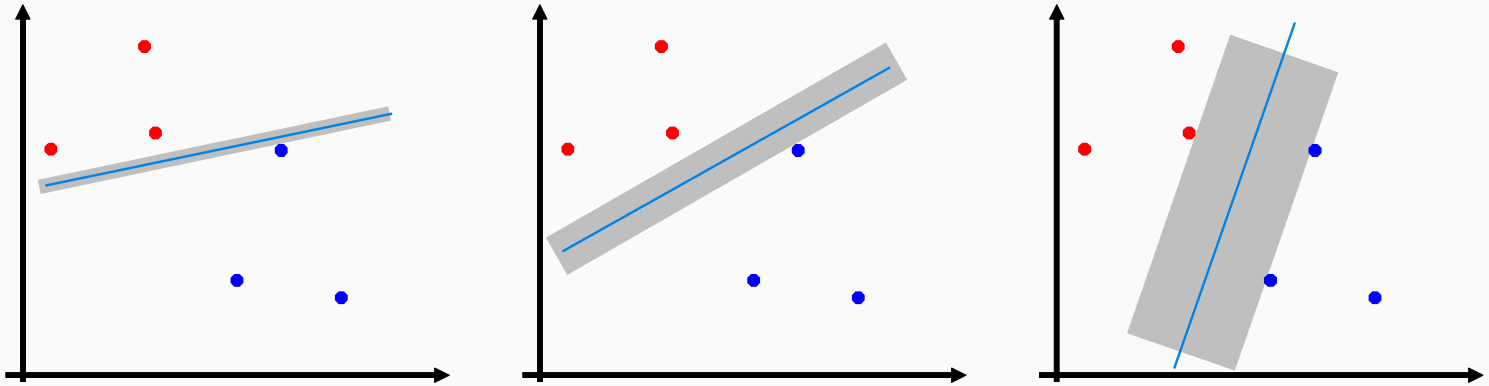
SVM: Maximum Margin Classifier

What is a good decision boundary?

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



- 여유로운 것이 더 강건하다 $\Rightarrow$ 넓은 통로가 좋다 $\Rightarrow$ Large Margin Classification



What is a good decision boundary?

u ▌Ḿð ẋ ÄÆu ḫ Ե Ⅼ (support vectors) Ⅎ u يم ƀ Ḿð Ȍ

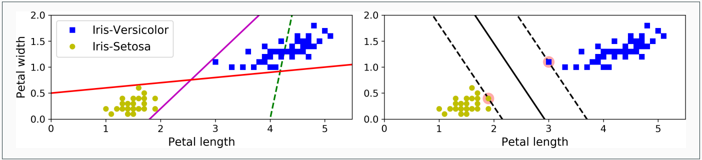


Figure 5-1. Large margin classi cation

Input feature u scale Ⅎ Ŷ Ƃ تز support vector machine

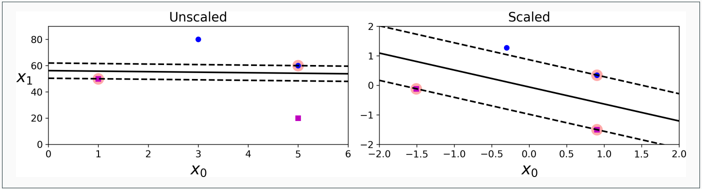


Figure 5-2. Sensitivity to feature scales

Hard margin vs. soft margin

Hard margin classification (hard-SVM)

- ῷ ṭ ḫ Ƅ ⅃ ṯ Ƅ margin Ò ↿ ˘ ݑ مم ḻ ˝E boundary ῦ ♥ ƌ
- ḫ Ƅ ⅃ ƅ ♠ضم Ƚ ݶ ̏E ₒ ῶ ƅ ɜ (linearly separable) نم ẋ ̏Ω Ƚ ἵ ƅ ɜ
- outlier ↿ ὸ ὓ Ў Ҕ

Soft margin classification (soft-SVM)

- margin ̤ ƅ ɜ نز ᵧ l مم ῴ ƅ ḻ margin ꟼ ݬ ݶ ̏E ṯ Ϡ ↿ ᵕ Ч ̤ بخ ἇ
- hyperparameter C: ضב ˝E Ֆ ꟼ ٢ (Ϡ Ḩ), Ư̤ ב ˝E ᵧ Ϡ ٢ (̄ ̀Ω بخ ἇ)

Hard margin vs. soft margin

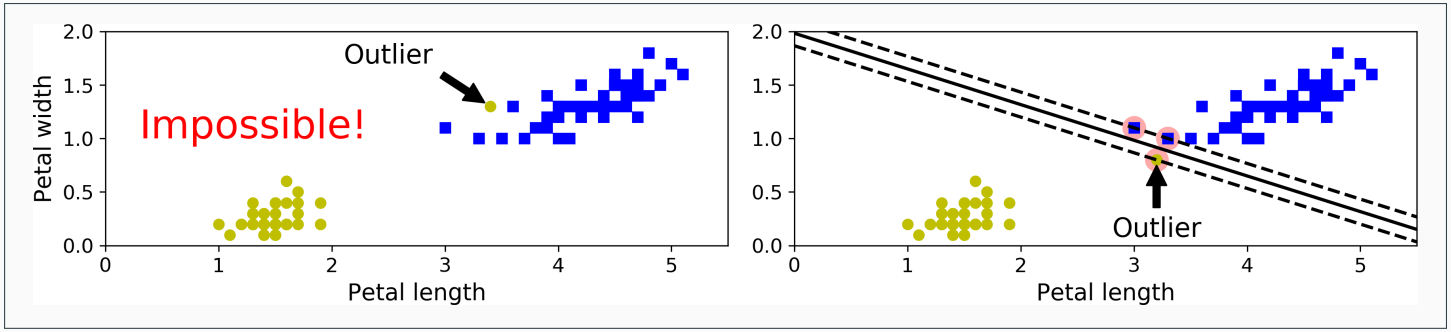


Figure 5-3. Hard margin sensitivity to outliers

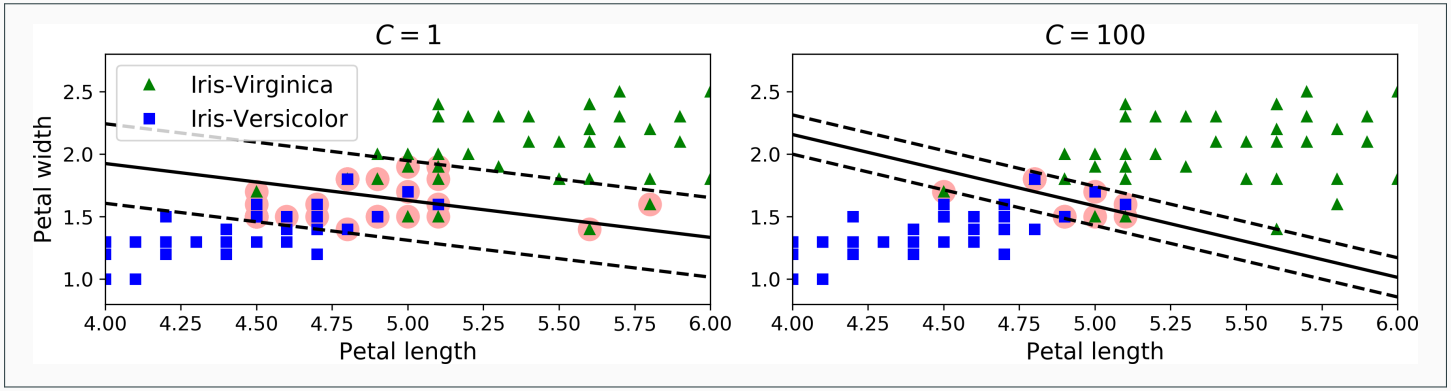


Figure 5-4. Large margin (le) versus fewer margin violations (right)

A brief history of SVM

- SVM א 1992 ҕ Boser, Guyon and Vapnik ɿ u يم ƅ ˉ ƙ Ồ
- Statistical Learning Theory ɿ Ƅ ̋E Ŀ ω̃ ڧ Ṗ Ⅵ ə ω̈ ͟ (Vapnik Chervonenkis)
- ˡ̋ ˥ ˑ Ⴍ Ŀ ` ɿ ƅ Ȅ ʮ ᵊ ŧ ʒ ̩ ̩ Ɨ Ƅ ώ̣ ƅ ᵥ ω̈ Ⅵ Ƅ I Ò
- SVM ٸ ̋E 1.1% Test error rate ≈ ` ظعي ♥ ÆŐ ` Ä ì (e.g., LeNet 4) Ö í ̋Ą̣
- ´ ĭ̋ Ŀ ٸ ̋E Ŭ̈ ב نز ŧ ʒ
- bioinformatics, text, image recognition Ų̩ ̩ نزنىمى ̋Ω א ŧ Й ▌ ɛ̆́
- ʒ̧ ̋A مم ə ą̊ ɼ ą̣ ʒ نز ̋Ą̣ ` Y̩̌ ω̣̃ Í
- ♠ضم / e ♠ضم ₒ ̋O ╝ Ⅵ ̃ Ớ مم w* outlier detection I̱ ב ئھ
- Ɪ Ω نز ˉ ˙ ω̣̃ / ظ˙ ω̣̃ ḫ Ƅ I ɧ u ₒ ̋O ɿ عيئح ∩ í̩̩
- ̋Ą̣ ` Y̩̌ ɿ ƅ ظ Ŭ̈ نز ͅ " ظ مم ᵁ Ŀ Kernel "̩̩ ▌ ĭ̋ مم ᵁ d̠ يج Ŀ Ⅵ ə ω̈ ͟
- ̋Ą̣ ` Y̩̌ ɿ ƅ ƀ Ч ᵥ ω̈ Ⅵ Ƅ ᵁ ω̣̃ Í Ƅ ̋Ḥ ̋Ą̣ ` Y̩̌ ̩̩ مم ̋Ḥ ̀Ω ṫ ` Ⅵ Ⅵ ˦ نم ͅ " ظ مم ᵁ
- Liblinear libsvm: Scikit-Learn ɿ ƅ liblinear ω̣ libsvm̩̩ ▌ ĭ̋ مم ɺ صحك

Convex Optimization and Duality

MoG: Lagrange multiplier

The method can be summarized as follows: in order to find the maximum or minimum of a function $\ddot{\imath}(\text{’})$ subjected to the equality constraint $\hat{O}(\text{’}) = 0$, form the Lagrangian function

$$L(\text{’};\ \ ) = \ddot{\imath}(\text{’}) \qquad \hat{O}(\text{’}) \tag{1}$$

and find the stationary points of L considered as a function of ’ and the Lagrange multiplier . 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.

MoG: Lagrange multiplier

Minimize $\ddot{\imath}(\text{’};\text{‘}) = \text{’} + \text{‘}$ subject to the constraint $\text{’}^2 + \text{‘}^2 = 1$, i.e.,

$$\hat{O}(\text{’};\text{‘}) = \text{’}^2 + \text{‘}^2 \quad 1 = 0 \tag{2}$$

Hence,

$$L(\text{’};\text{‘};\ ) = \ddot{\imath}(\text{’};\text{‘}) + \ \hat{O}(\text{’};\text{‘}) = \text{’} + \text{‘} + \ (\text{’}^2 + \text{‘}^2 \quad 1) \tag{3}$$

Gradient

$$r_{\text{’};\text{‘};\ } L(\text{’};\text{‘};\ ) = (1 + 2\ \text{’}; 1 + 2\ \text{‘}; \text{’}^2 + \text{‘}^2 \quad 1) \tag{4}$$

and therefore,

$$r_{\text{’};\text{‘};\ } L(\text{’};\text{‘};\ ) \quad (\ ) \quad \begin{cases} 1 + 2\ \text{’} = 0 \\ 1 + 2\ \text{‘} = 0 \\ \text{’}^2 + \text{‘}^2 \quad 1 = 0 \end{cases} \tag{5}$$

MoG: Lagrange multiplier

$$r_{\lambda,\alpha,\beta} L(\lambda,\alpha,\beta) \quad (*) \quad \begin{cases} 1 + 2\lambda\alpha = 0 \\ 1 + 2\lambda\beta = 0 \\ \alpha^2 + \beta^2 - 1 = 0 \end{cases} \tag{6}$$

This yields

$$\alpha = \beta = -\frac{1}{2\lambda}; \quad \lambda \neq 0 \tag{7}$$

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

So,

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

which implies that the stationary points of 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

$$\begin{aligned} \text{minimize} \quad & \tilde{\imath}_0(\text{'}) && (11) \\ \text{subject to} \quad & \tilde{\imath}_{\tilde{\text{0}}}(\text{'}) \quad 0; \qquad \tilde{\text{0}} = 1; 2; \ldots; g && (12) \\ & d_{\tilde{\text{0}}}(\text{'}) = 0; \qquad \tilde{\text{0}} = 1; 2; \ldots; \text{£} && (13) \end{aligned}$$

- $\text{'} \ 2 \ R^{\text{ɦ}}$ is the optimization variable
- $\tilde{\imath}_0 : R^{\text{ɦ}} \ ! \ R$ is the objective or cost function
- $\tilde{\imath}_{\tilde{\text{0}}} : R^{\text{ɦ}} \ ! \ R; \tilde{\text{0}} = 1; 2; \ldots; g$ are the inequality constraint functions
- $d_{\tilde{\text{0}}} : R^{\text{ɦ}} \ ! \ R$ are the equality constraint functions

Convex optimization problem

Standard form convex optimization problem

$$\begin{aligned} \text{minimize} \quad & \tilde{\imath}_0(\cdot) && && (14)\\ \text{subject to} \quad & \tilde{\imath}_{\tilde{\jmath}}(\cdot) \quad 0; && \tilde{\jmath} = 1; 2; \ldots; g && (15)\\ & h\upsilon_{\tilde{\jmath}}^{\dot{\imath}\cdot} = \varepsilon_{\tilde{\jmath}}; && \tilde{\jmath} = 1; 2; \ldots; \text{£} && (16) \end{aligned}$$

- $\tilde{\imath}_0; \tilde{\imath}_1; \ldots; \tilde{\imath}_g$ are convex
- equality constraints are affine

Often written as

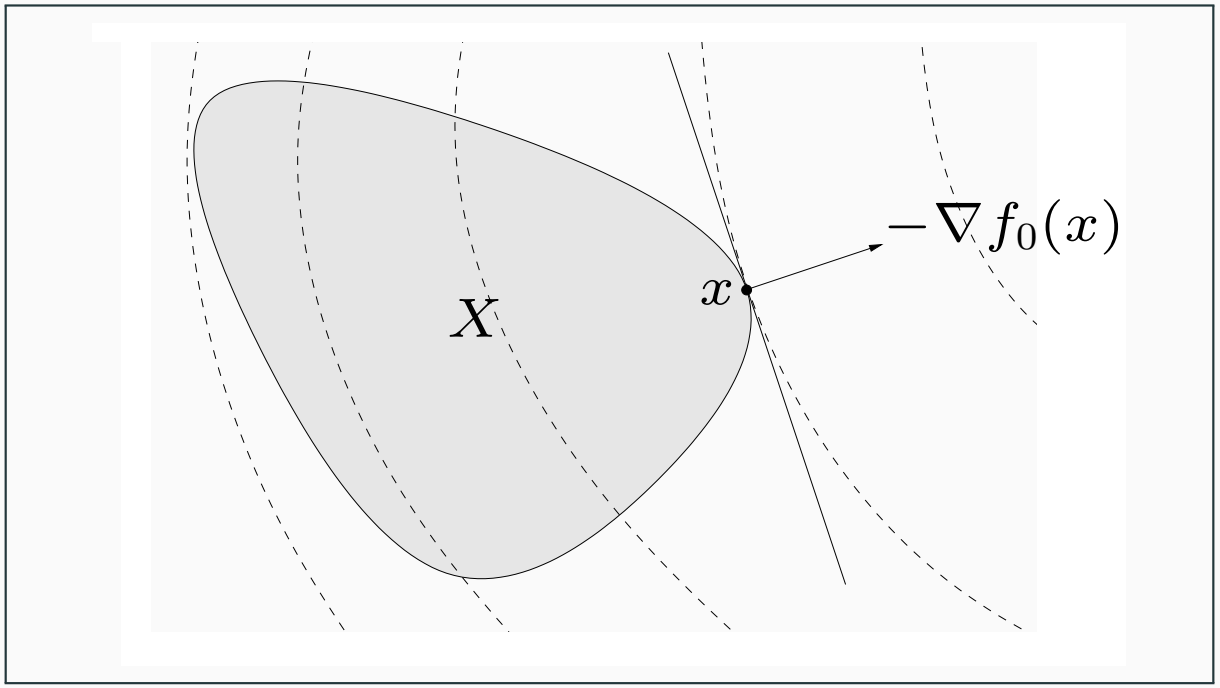
$$\begin{aligned} \text{minimize} \quad & \tilde{\imath}_0(\cdot) && && (17)\\ \text{subject to} \quad & \tilde{\imath}_{\tilde{\jmath}}(\cdot) \quad 0; && \tilde{\jmath} = 1; 2; \ldots; g && (18)\\ & \cdot = \varepsilon && && (19) \end{aligned}$$

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

Optimality criterion for differentiable $\tilde{\imath}_0$

’ is optimal if and only if it is feasible and

$$r\,\tilde{\imath}_0(\text{’})^{\dot{I}}(\text{‘} \quad \text{’}) \quad 0 \quad \text{for all feasible ‘} \tag{20}$$



if nonzero, $r\,\tilde{\imath}_0(\text{’})$ defines a supporting hyperplane to feasible set t at ’

Lagrangian

standard form problem

$$\text{minimize} \quad \tilde{\imath}_0(\text{'}) \tag{21}$$

$$\text{subject to} \quad \tilde{\imath}_{\text{i}}(\text{'}) \quad 0; \qquad \text{i} = 1; 2; \ldots; g \tag{22}$$

$$d_{\text{i}}(\text{'}) = 0; \qquad \text{i} = 1; 2; \ldots; \text{£} \tag{23}$$

variable $\text{'} \ 2 \ R^{\text{ɦ}}$, domain D, optimal value £

Lagrangian: $\acute{I} : R^{\text{ɦ}} \quad R^{g} \quad R^{\text{£}} \ ! \ R$ with $\operatorname{Lnh} \acute{I} = D \quad R^{g} \quad R^{\text{£}}$

$$\acute{I}(\text{'}; \ ; \ ) = \tilde{\imath}_0(\text{'}) + \sum_{\text{i}=1} \text{i}\tilde{\imath}_{\text{i}}(\text{'}) + \sum_{\text{i}=1} \text{i}d_{\text{i}}(\text{'}) \tag{24}$$

- weighted sum of objective and constraint functions
- i is Lagrange multiplier associated with $\tilde{\imath}_{\text{i}}(\text{'}) \quad 0$
- i is Lagrange multiplier associated with $d_{\text{i}}(\text{'}) = 0$

Lagrange dual function

Lagrange dual function: $\hat{O} : R^{g} \quad R^{\pounds} \; ! \; R$

$$\hat{O}(\; ; \;) = \inf_{2D} \acute{I}(' \; ; \; ; \;) \tag{25}$$

$$= \inf_{2D} \; \grave{\imath}_0(' \;) + \sum_{\ddot{0}=1} \; _{\ddot{0}}\grave{\imath}_{\ddot{0}}(' \;) + \sum_{\ddot{0}=1}^{\pounds} \; _{\ddot{0}}d_{\ddot{0}}(' \;) \tag{26}$$

$\hat{O}$ is concave, can be 1 for some $;$

lower bound property: if 0, then $\hat{O}(\; ; \;) \quad \pounds$

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

$$\grave{\imath}_0(\tilde{}) \quad \acute{I}(\tilde{}; \; ; \;) \quad \inf_{2D} \acute{I}(' \; ; \; ; \;) = \hat{O}(\; ; \;) \tag{27}$$

minimizing over all feasible $\tilde{}$ gives $\pounds \quad \hat{O}(\; ; \;)$

The dual problem

Lagrange dual problem

$$\text{maximize} \quad \hat{O}(\ ;\ ) \tag{28}$$

$$\text{subject to} \quad 0 \tag{29}$$

- finds best lower bound on $\mathfrak{L}$, obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted $\check{a}$
- $\ ;\ $ are dual feasible if $0, (\ ;\ )\ 2\ \mathrm{Lnh}\ \hat{O}$
- often simplified by making implicit constraint $(\ ;\ )\ 2\ \mathrm{Lnh}\ \hat{O}$ explicit

Weak and strong duality

weak duality: $\check{a} \quad \xi$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $\check{a} = \xi$

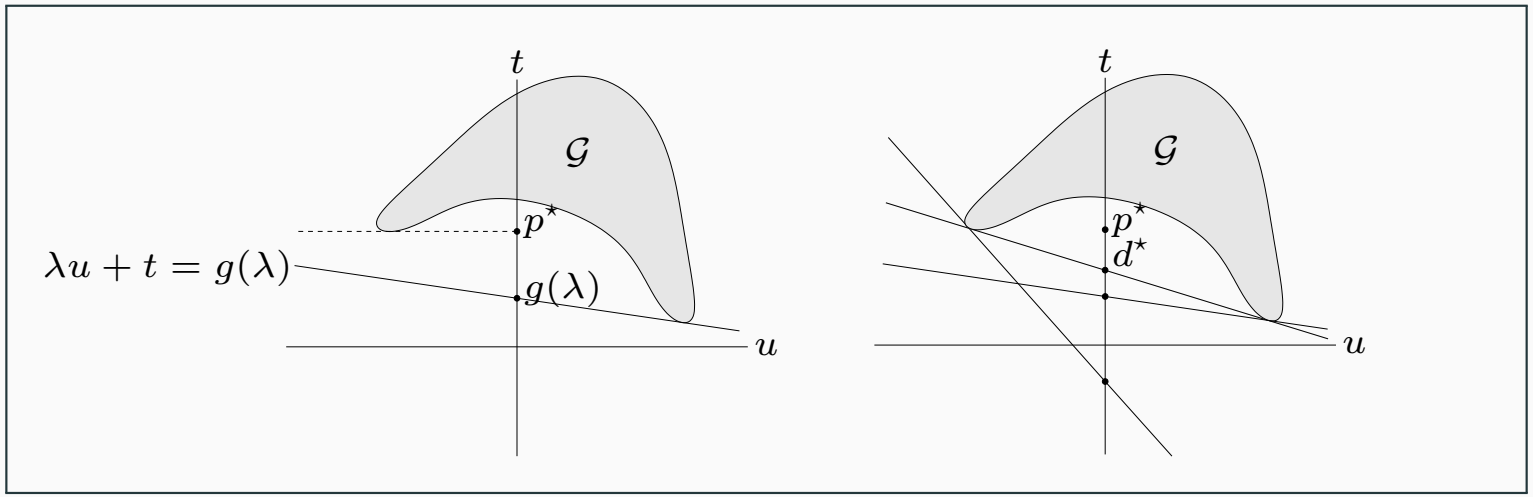
- does not hold in general
- holds for convex problems
- conditions that guarantee strong duality in convex problems are called constraint qualifications

Geometric interpretation

for simplicity, consider problem with one constraint $\tilde{\imath}_1(\cdot) \quad 0$

interpretation of dual function

$$\hat{O}(\ ) = \inf_{(\cdot;\bar{\ })2G} (\bar{\ } + \ \cdot) \qquad \text{where } G = f(\tilde{\imath}_1(\cdot); \tilde{\imath}_0(\cdot))\ j\ \cdot\ 2\ Dg \qquad (30)$$



- $\cdot + \bar{\ } = \hat{O}(\ )$ is supporting hyperplane to G
- hyperplane intersects $\bar{\ }$-axis at $\bar{\ } = \hat{O}(\ )$

Geometric interpretation

- Primal problem:

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

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

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

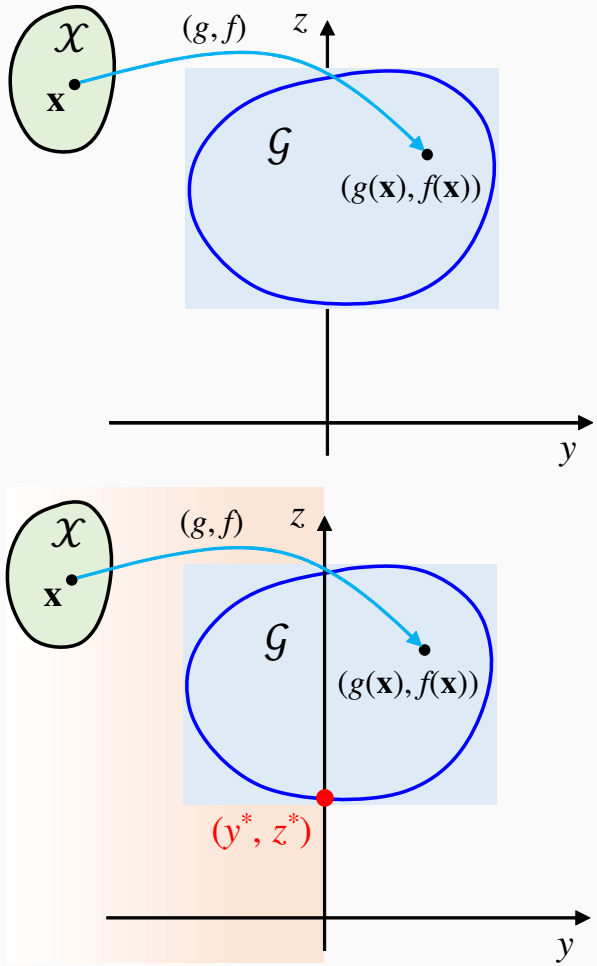
$$\mathcal{G} = \{(y, z) \mid y = g(\mathbf{x}), z = f(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$$

$\Longrightarrow$ 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^*)



Geometric interpretation

- Lagrange Dual Problem

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

where (Lagrangian subproblem):

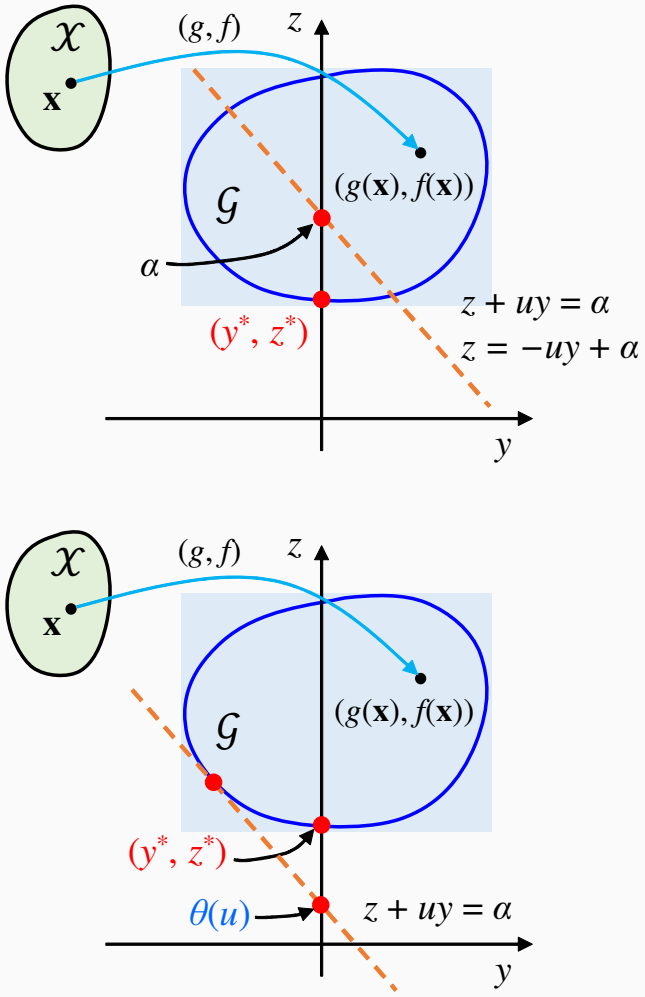
$$\theta(u) = \inf\{f(\mathbf{x}) + ug(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}.$$

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

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

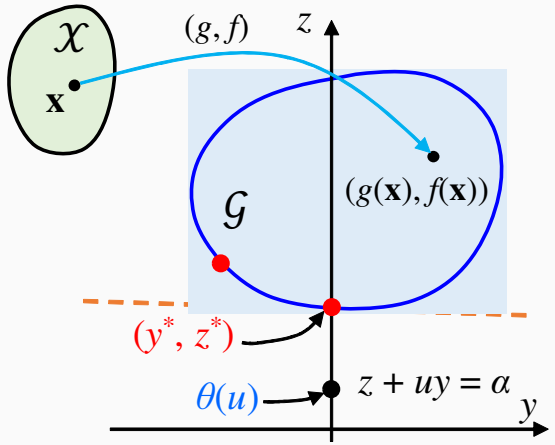
   여기서 $z + uy = \alpha$ 는 기울기가 $-u$ 이고 z 축과 만나는 점이 α 인 직선식이다.

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.



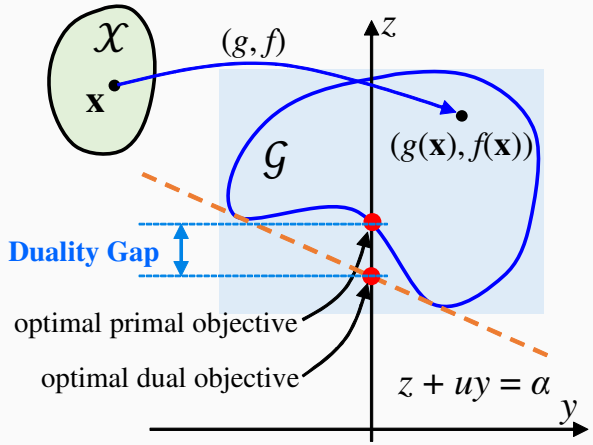
- Primal problem 과 dual problem 의 최적해가 동일한 경우, duality gap 이 없다고 말한다 (**strong duality**).

- 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

the following four conditions are called KKT conditions (for a problem with differentiable $\tilde{\imath}_{\tilde{\jmath}}; d_{\tilde{\jmath}}$)

1. primal constraints: $\tilde{\imath}_{\tilde{\jmath}}(\,') \quad 0$ i = 1, ..., m, $d_{\tilde{\jmath}}(\,') = 0$, i = 1, ..., p
2. dual constraints: 0
3. complementary slackness: $_{\tilde{\jmath}}\tilde{\imath}_{\tilde{\jmath}}(\,') = 0; \tilde{\jmath} = 1; :::; g$
4. gradient of Lagrangian with respect to $'$ vanishes:

$$r\,\tilde{\imath}_0(\,') + \sum_{\tilde{\jmath}=1} {}_{\tilde{\jmath}}\, r\,\tilde{\imath}_{\tilde{\jmath}}(\,') + \sum_{\tilde{\jmath}=1} {}_{\tilde{\jmath}}\, r\, d_{\tilde{\jmath}}(\,') = 0 \tag{31}$$

if strong duality holds and $'\;;\;\;;\;\;$ are optimal, then they must satisfy the KKT conditions

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

$$\text{˙}(²) = \circ^{\dot{I}} (²) + \mathcal{E} \tag{32}$$

where $(²)$ denotes a fixed feature-space transformation, and we have made the bias parameter $\mathcal{E}$ explicit.

The training data set comprises $\acute{Y}$ input vectors $²_1; ²_2; :::; ²_{\acute{Y}}$, with corresponding target values $¯_1; ¯_2; :::; ¯_{\acute{Y}}$ where $¯_{ĵ} \ 2 \ f \ \ 1; 1g$, and new data points $²$ are classified according to the sign of $\text{˙}(²)$

Maximum margin classifier

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 $\circ$ and ε such that a function satisfies $\grave{}(\mathbf{2}_{\jmath}) > 0$ for points having $\bar{}_{\jmath} = +1$ and $\grave{}(\mathbf{2}_{\jmath}) < 0$ for points having $\bar{}_{\jmath} = \quad 1$, so that $\bar{}_{\jmath}\grave{}(\mathbf{2}_{\jmath}) > 0$ for all training data points.

Maximum margin classifier: optimality criterion

Thus the distance of a point $²_ɧ$ to the decision surface is given by

$$\frac{¯_ɧ ‘(²_ɧ)}{k ° k} = \frac{¯_ɧ(° \dot{I} \ (²_ɧ) + \mathcal{E})}{k ° k}: \tag{33}$$

The margin is given by the perpendicular distance to the closest point $²_ɧ$ from the data set, and we wish to optimize the parameters $°$ and $\mathcal{E}$ in order to maximize this distance. Thus the maximum margin solution is found by solving

$$9\}X \underset{\mathbf{w};\mathcal{E}}{\mathrm{h}} 9² \quad \frac{1}{k ° k} \mathrm{h}\_\mathrm{j}[¯_ɧ(° \dot{I} \ (²_ɧ) + \mathcal{E})] \tag{34}$$

where we have taken the factor $1/k ° k$ outside the optimization over $ɧ$ because $°$ does not depend on $ɧ$.

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 function with constraint

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

$$L(\beta, \xi, \lambda) = \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i [y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] \quad (35)$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)^T$. Note the minus sign in front of the Lagrange multiplier term, because we are minimizing with respect to β and ξ, and maximizing with respect to λ.

Setting the derivatives of $L(\beta, \xi, \lambda)$ with respect to β and ξ equal to zero, we obtain the following two conditions

$$\beta = \sum_{i=1}^{n} \lambda_i y_i x_i \quad (36)$$

$$0 = \sum_{i=1}^{n} \lambda_i y_i \quad (37)$$

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

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

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

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

In order to classify new data points using the trained model, we evaluate the sign of ˋ (²). This can be expressed in terms of the parameter fw_ig and the kernel function by substituting for ° to give

$$ˋ(²) = \sum_{i=1}^{\acute{Y}} w_i \bar{}_i j(²; ²_i) + \mathcal{E} \tag{41}$$

KKT condition: complementary slackness

We show that a constrained optimization of this form satisfies the $\acute{E}\hbar\omega_{W}^{W}\check{}\ '\dot{q}_{,}\neg\acute{E}\check{}\ \dot{q}_{,}\hbar\neg\dot{I}\check{}\ !\jmath\bar{a}_{W}^{W}$ (KKT) conditions, which in this case require that the following three properties hold

$$\hbar\upsilon_{\jmath} \quad 0 \tag{42}$$

$$\bar{}_{\jmath}(^{2}{}_{\jmath}) \quad 1 \quad 0 \tag{43}$$

$$\hbar\upsilon_{\jmath} f\bar{}_{\jmath}\grave{}(^{2}{}_{\jmath}) \quad 1g = 0 \tag{44}$$

Thus for every data point, either $\hbar\upsilon_{\jmath} = 0$ or $\bar{}_{\jmath}\grave{}(^{2}{}_{\jmath}) = 1$. Any data point for which $\hbar\upsilon_{\jmath} = 0$ will not appear in the sum and hence plays no role in making predictions for new data points. The remaining data points are called $'\check{}\ \pounds\pounds\omega_{W}^{W-}\ \dot{}\bar{a}!\bar{}\omega_{W}^{W}'$, and because they satisfy $\bar{}_{\jmath}\grave{}(^{2}{}_{\jmath}) = 1$, they correspond to points that lie on the maximum margin hyperplanes in feature space.

KKT condition: complementary slackness

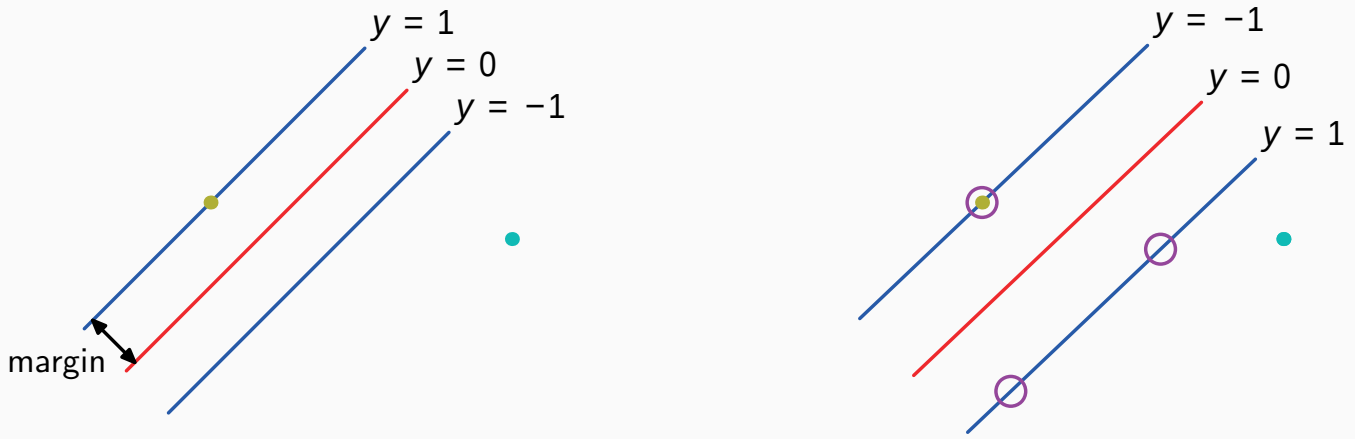
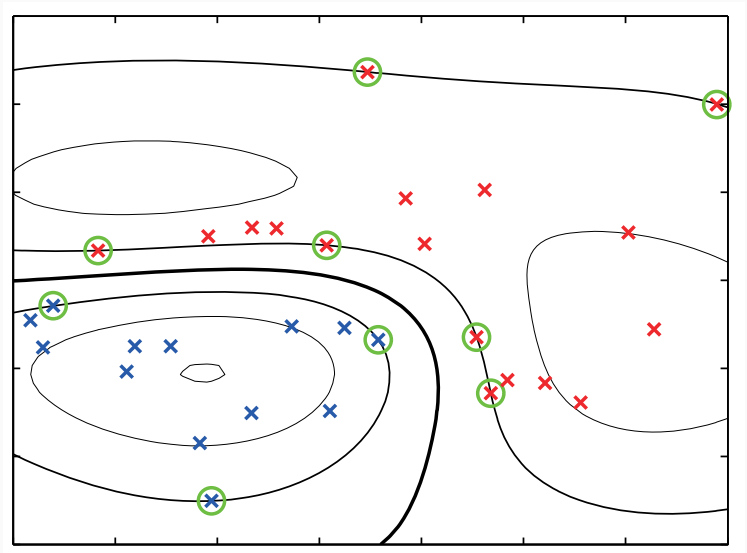


Figure 7.1 The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left figure. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right. The location of this boundary is determined by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



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)