Kernel Methods

90:30

패턴인식 및 기계학습 겨울학교 발표자료
2016. 01. 21.

고려대학교 제어계측공학과 박주영

Contents

- **Introduction**

- **Support Vector Machines:** Support vector classification (SVC), Support vector data description (SVDD), Reproducing kernel Hilbert space (RKHS), Representer theorem

- **Gaussian Processes:** Gaussian process regression (GPR), Gaussian process classification (GPC)

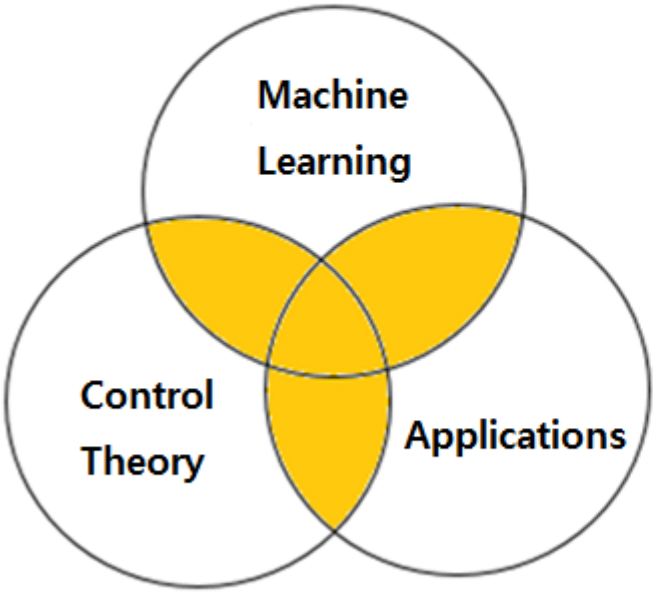
- **Concluding Remarks**

Outline

- **Introduction**
- Support Vector Machines
- Gaussian Processes
- Concluding Remarks

Introduction

My (Research) Interests:



✚ **Collaborations**

Introduction

기계학습 이론 연구의 3대 요소

- 수학이론
  - Probability
  - Stochastic process
  - Functional analysis
  - …
- 최적화 기술
  - Convex optimization
  - Calculus of variations
  - Dynamic programming, Approximate dynamic programming
  - …
- 전기전자 컴퓨터 분야의 패턴인식 기술
  - Statistical pattern recognition
  - Programming (Matlab, Python, Theano, Torch7, TensorFlow, …)
  - …

Introduction

한국정보과학회, 패턴인식 및 기계학습 여름학교, 연세대학교, 2014 (주제: 기계학습을 위한 최적화 이론).

한국정보과학회 인공지능소사이어티

패턴인식 및 기계학습 여름학교

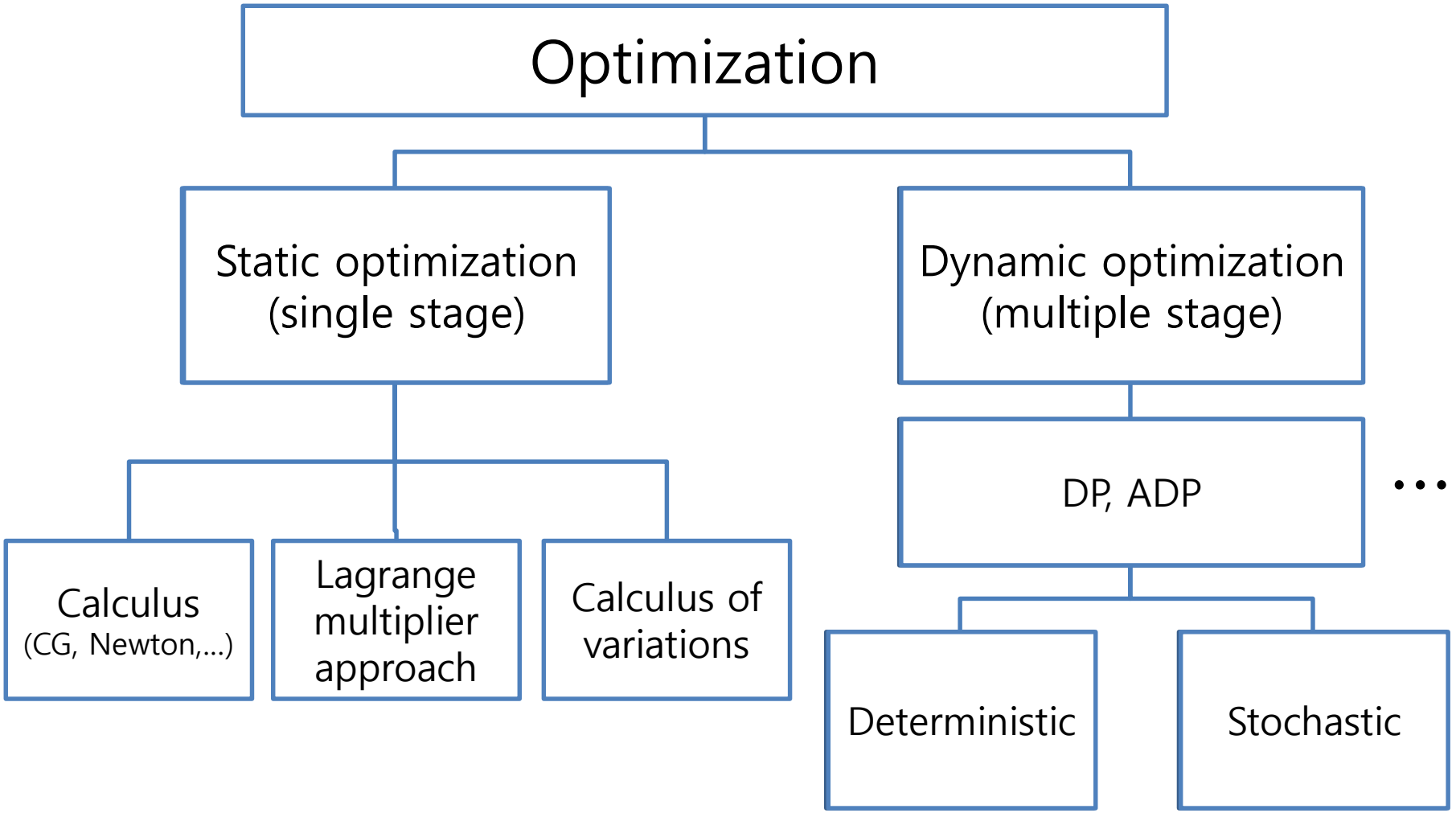
2014년 8월 7일(목)-8일(금)

▶ 일시 : 2014년 8월 7일(목) - 8일(금)
▶ 장소 : 연세대학교 제3공학관 지하 1층 C040
▶ 주최 : 한국정보과학회 인공지능소사이어티

〈 프로그램 〉

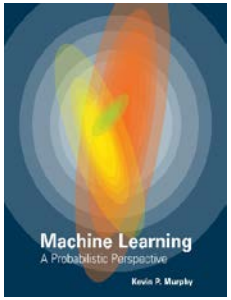
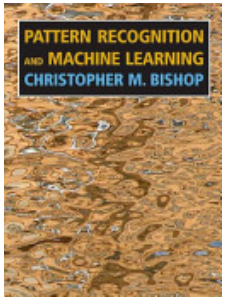
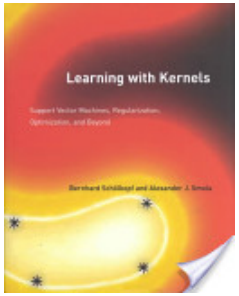
	7일(목)	8일(금)
09:00-10:45	Deep Learning for Computer Vision 김준모 교수, KAIST	Object Detection and Image Retrieval 윤성의 교수, KAIST
11:00-12:45	Spectral Clustering 한보형 교수, POSTECH	Visual Tracking 임종우 교수, 한양대
12:45-14:00	점심시간	
14:00-15:45	Optimization 박주영 교수, 고려대	Geometric Computer Vision 윤국진 교수, GIST
16:00-17:45	Topic Models for Image Annotation 최승진 교수, POSTECH	Computational Photography 김선주 교수, 연세대

Introduction



Introduction

Textbooks for Machine Learning

	Probabilistic approach	Non-probabilistic approach
Supervised learning	 	
Unsupervised learning		
Reinforcement learning		

Introduction to Kernel methods

- Kernel methods use $f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$

- The representer theorem (Schölkopf & Smola, 2002):
  - Let H be the RKHS (reproducing kernel Hilbert space) associated to the kernel k.
    Denote by $\Omega: [0, \infty) \to R$ a strictly monotonic increasing function, by X a compact set, and by $c: (X \times R^2)^m \to R \cup \{\infty\}$ an arbitrary loss function.

  - Then each minimizer $f \in H$ of the regularized risk
    $$c((x_1, y_1, f(x_1)), \cdots, (x_N, y_N, f(x_N))) + \Omega(\|f\|_H)$$
    admits a representation of the form
    $$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$$

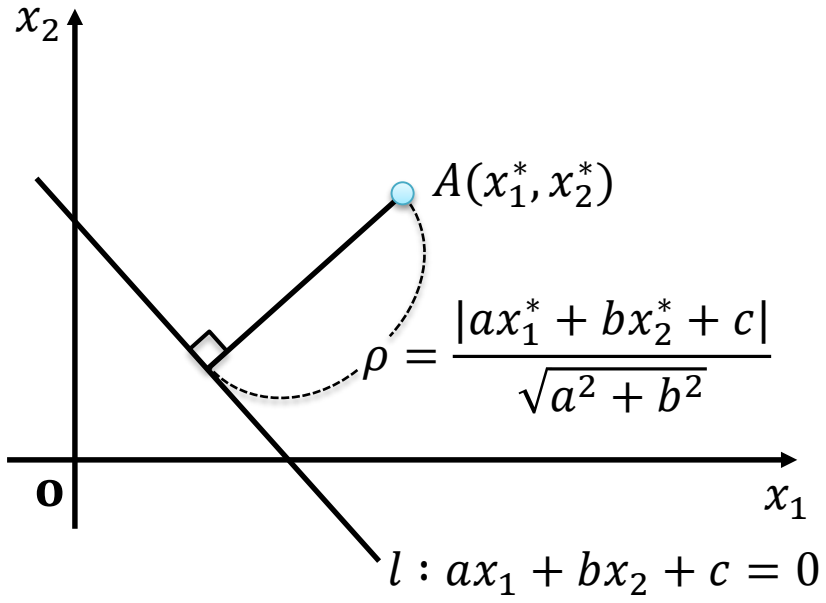
Outline

- Introduction
- **Support Vector Machines**
- Gaussian Processes
- Concluding Remarks

점과 직선 사이의 거리

점 (x_1, x_2)에서 직선 $ax_1 + bx_2 + c = 0$ 까지의 거리ρ는 $\rho = \frac{|ax_1+bx_2+c|}{\sqrt{a^2+b^2}}$

특히, 원점과 직선 $ax_1 + bx_2 + c = 0$ 사이의 거리 ρ는 $\rho = \frac{|c|}{\sqrt{a^2+b^2}}$



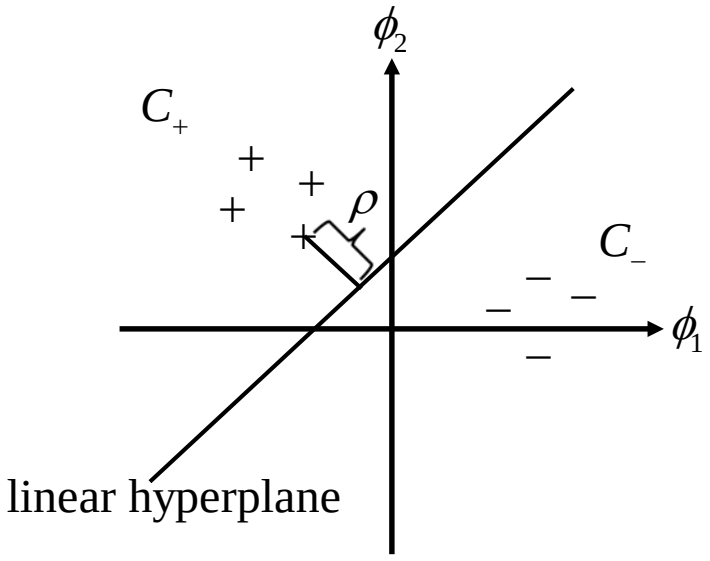
일반적인 경우: $\rho = \frac{|\omega^T \phi(x) + b|}{\|\omega\|}$

SVM

- **Support vector classification (SVC)**

Trn data: $D = \{(x_n, t_n)\}_{n=1}^{N}$, where $t_n \in \{+1, -1\}$

Decision fcn.: $y(x) = \omega^T \phi(x) + b$



Assumption: D is linearly separable, i.e., $\exists \omega, b$ s.t.

$$\omega^T \phi(x_n) + b > 0 \text{ for } \forall x_n \in C_+ \text{ (i.e., } t_n = +1)$$

$$\omega^T \phi(x_m) + b < 0 \text{ for } \forall x_m \in C_- \text{ (i.e., } t_n = -1)$$

Definition:

- The separation margin ρ : The distance between the separating hyperplane and a closest data point, $x_n \in SV$.

$$\rho = \frac{\left|\omega^T \phi(x_n) + b\right|}{\|\omega\|}$$

SVM

Canonical representation:

$$y(x_n) = \omega^T \phi(x_n) + b = +1 \text{ when } x_n \in C_+ \bigcap \text{SV}$$

$$y(x_n) = \omega^T \phi(x_n) + b = -1 \text{ when } x_n \in C_- \bigcap \text{SV}$$

Maximum margin classification for the linearly separable cases:

$$\max_{\omega,b} \rho = \frac{1}{\|\omega\|}$$

$$\text{s.t. } y(x_n) = \omega^T \phi(x_n) + b \geq +1 \text{ for } \forall x_n \in C_+$$

$$\text{and } y(x_n) = \omega^T \phi(x_n) + b \leq -1 \text{ for } \forall x_n \in C_-$$

$$\Leftrightarrow \quad \min_{\omega,b} J_p(\omega, b) = \frac{1}{2}\|\omega\|^2$$

$$\text{s.t. } t_n\left(\omega^T \phi(x_n) + b\right) \geq +1 \text{ for } \forall n \in \{1, \ldots, N\}$$

SVM

- **Theorem**: $f \in C^1$ has a min. at $x^* \Rightarrow \dfrac{\partial f}{\partial x}(x^*) = 0$

  This condition, together with convexity of f, is also a suff. cond.

- Example 1: $\min. \; f(x) = \dfrac{1}{2}\left(x_1^2 + x_2^2\right)$

  [solution] $\dfrac{\partial f}{\partial x} = 0 \Rightarrow \left[\dfrac{\partial f}{\partial x_1} \quad \dfrac{\partial f}{\partial x_2}\right] = \left[x_1 \quad x_2\right] = 0$

  $\therefore \; x^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

- In a constrained min. problem,

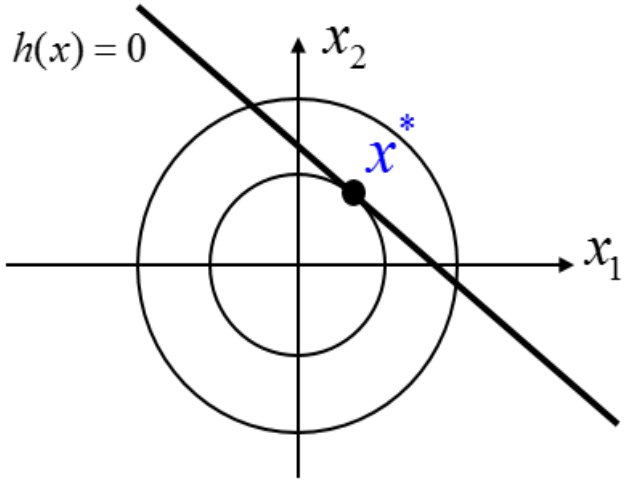
  $f \in C^1$ has a min. at $x^* \not\Rightarrow \dfrac{\partial f}{\partial x}(x^*) = 0$

SVM

- Example 2:

$$\min.\ f(x)=\frac{1}{2}\left(x_1^2+x_2^2\right)$$

$$s.t.\quad h(x)=1-x_1-x_2$$



[solution]

Define the Lagrange function

$$L(x,\lambda)=f(x)+\lambda h(x)$$

$$=\frac{1}{2}\left(x_1^2+x_2^2\right)+\lambda\left(1-x_1-x_2\right)$$

$$\frac{\partial L}{\partial x}=0 \Rightarrow \begin{bmatrix}x_1-\lambda & x_2-\lambda\end{bmatrix}=0. \quad \therefore\ x_1=x_2=\lambda$$

$$\frac{\partial L}{\partial \lambda}=0 \Rightarrow 1-x_1-x_2=0. \quad \therefore\ 1-2\lambda=0. \quad \therefore\ \lambda=\frac{1}{2}.$$

$$\therefore\ x^*=\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix}^T$$

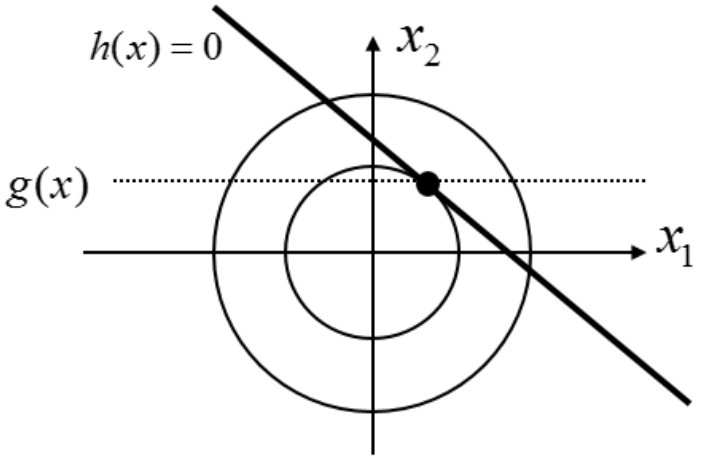
SVM

- Example 3:

$$\min.\ f(x)=\frac{1}{2}\left(x_1^2+x_2^2\right)$$

$$s.t.\quad h(x)=1-x_1-x_2$$

$$g(x)=\frac{3}{4}-x_2\le 0$$



[solution]

Define the Lagrange function

$$L(x,\lambda)=f(x)+\lambda h(x)+\alpha g(x)$$

$$=\frac{1}{2}\left(x_1^2+x_2^2\right)+\lambda\left(1-x_1-x_2\right)+\alpha\left(\frac{3}{4}-x_2\right),\quad \alpha\ge 0$$

$$\frac{\partial L}{\partial x}=0\ \Rightarrow\ \begin{bmatrix}x_1-\lambda & x_2-\lambda-\alpha\end{bmatrix}=0.\qquad \therefore\ x_1=\lambda,\ x_2=\lambda+\alpha$$

$$\frac{\partial L}{\partial \lambda}=0\ \Rightarrow\ 1-x_1-x_2=0.\qquad \therefore\ 2\lambda+\alpha=1.$$

SVM

Also, $\alpha \geq 0$ and $\frac{3}{4} - x_2 \leq 0$

One more condition is needed to solve the problem.

→ The Kuhn-Tucker complementarity condition

$$\alpha\left(\frac{3}{4} - x_2\right) = 0 \qquad \text{i.e.,} \quad \alpha = 0 \quad \text{or} \quad x_2 = \frac{3}{4}$$

i) If $\alpha = 0$, then $\lambda = \frac{1}{2}$; thus $x_1 = x_2 = \cancel{\frac{1}{2}}$ $\left(\because x_2 \geq \frac{3}{4}\right)$

ii) If $x_2 = \frac{3}{4}$, then $\begin{Bmatrix} \lambda + \alpha = 3/4 \\ 2\lambda + \alpha = 1 \end{Bmatrix}$ $\therefore \begin{Bmatrix} \lambda = 1/4 \,,\ \alpha = 1/2 \\ x_1 = 1/4 \,,\ x_2 = 3/4 \end{Bmatrix}$

$\therefore x^* = \begin{bmatrix} 1/4 & 3/4 \end{bmatrix}^T$

SVM

Theorem (Kuhn-Tucker Theorem)

Given an opt. prob. with convex domain

$$\left.\begin{array}{ll} \min. & f(x),\ x \in \Omega \ (x \text{ is primal var.}) \\ s.t. & g_i(x) \le 0, \quad i = 1,\ldots,I \\ & h_j(x) = 0, \quad j = 1,\ldots,J \end{array}\right\} \text{ primal opt. prob. (*)}$$

with $f \in C^1$ convex, and g_i, h_j affine, the following are necessary and sufficient condition for a point $x^* \in \Omega$ to be an opt. :

For $L(x,\ \alpha,\ \lambda) \triangleq f(x) + \sum_{i=1}^{I} \alpha_i g_i(x) + \sum_{j=1}^{J} \lambda_j h_j(x) = f + \alpha^T g + \lambda^T h,$

$\exists\ \alpha^*$ and λ^* s.t. $\dfrac{\partial L}{\partial x}(x^*,\ \alpha^*,\ \lambda^*) = 0, \quad \dfrac{\partial L}{\partial \lambda}(x^*,\ \alpha^*,\ \lambda^*) = 0$

$g_i(x^*) \le 0$ and $\alpha^* \ge 0$ for $i = 1,\ldots,I$

and $\alpha_i^*\ g_i(x^*) = 0\ , \quad i = 1,\ldots,I$

The KT complementarity cond.

SVM

Lagrange fcn.: $L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 + \sum_{n=1}^{N} \alpha_n \left\{1 - t_n\left(\omega^T \phi(x_n) + b\right)\right\}$,

where ω, b are primal variables and α is dual variable.

KKT conditions:

① $t_n\left(\omega^T \phi(x_n) + b\right) \geq +1 \;\; \forall n \in \{1, \ldots, N\}$

② $\alpha_n \geq 0 \;\; \forall n \in \{1, \ldots, N\}$

③ $\frac{\partial L}{\partial \omega} = \omega - \sum_{n=1}^{N} \alpha_n t_n \phi(x_n) = 0 \quad \therefore \omega = \sum_{n=1}^{N} \alpha_n t_n \phi(x_n)$

$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} \alpha_n t_n = 0 \quad \therefore \sum_{n=1}^{N} \alpha_n t_n = 0$

④ (Complementary slackness)

$\alpha_n \left\{1 - t_n\left(\omega^T \phi(x_n) + b\right)\right\} = 0$ for $\forall n \in \{1, \ldots, N\}$,

i.e., $\alpha_n = 0$ or $1 - t_n\left(\omega^T \phi(x_n) + b\right) = 0$ at the optimal solutions.

SVM

Dual problem:

$$J_D(\alpha) = \min_{\omega,b} L(\omega, b, \alpha)$$

$$= \frac{1}{2}\sum_n \sum_m \alpha_n \alpha_m t_n t_m \langle \phi(x_n), \phi(x_m) \rangle + \sum_{n=1}^{N} \alpha_n$$

$$- \sum_n \sum_m \alpha_n \alpha_m t_n t_m \langle \phi(x_n), \phi(x_m) \rangle - \left(\sum_{n=1}^{N} \alpha_n t_n \right) b$$

$$= -\frac{1}{2}\alpha^T Q \alpha + 1^T \alpha, \text{ where } Q = [Q_{nm}] = \left[t_n t_m \langle \phi(x_n), \phi(x_m) \rangle \right]$$

$$\therefore \max_{\alpha} J_D(\alpha) = -\frac{1}{2}\alpha^T Q \alpha + 1^T \alpha \quad \text{s.t. } \alpha \geq 0$$

Notes:

① Convex QP has strong duality. i.e., $d^* = p^*$. Its global opt. can be found efficiently.

② Kernel trick: $\langle \phi(x_n), \phi(x_m) \rangle = \phi^T(x_n)\phi(x_m) = k(x_n, x_m)$; Mercer Kernel

SVM

Examples of Mercer's kernels:

① RBF kernel: $k(x,y)=\exp\left(-\frac{1}{2}\frac{\|x-y\|^2}{\sigma^2}\right)$

② Polynomial kernel: $k(x,y)=\left(\langle x,y\rangle+c\right)^d$

SVM

Summary:

- Primal problem: $\min_{\omega,b} J_p(\omega,\mathrm{b}) = \frac{1}{2}\|\omega\|^2$ s.t. $1 - t_n\left(\omega^T\phi(x_n)+b\right) \le 0,\ \forall n$
- Lagrange function: $L(\omega,b,\alpha) = \frac{1}{2}\|\omega\|^2 + \sum_{n=1}^{N}\alpha_n\left\{1 - t_n\left(\omega^T\phi(x_n)+b\right)\right\}$
- Dual problem: $\max_{\alpha} J_D(\alpha) = -\frac{1}{2}\alpha^T Q\alpha + 1^T\alpha$ s.t. $\alpha \ge 0$
- KKT conditions:
  - ① $t_n\left(\omega^T\phi(x_n)+b\right) \ge +1\ \forall n \in \{1,\ldots,N\}$
  - ② $\alpha_n \ge 0\ \forall n \in \{1,\ldots,N\}$
  - ③ $\frac{\partial L}{\partial \omega} = 0,\ \frac{\partial L}{\partial b} = 0 \quad \therefore \omega = \sum_{n=1}^{N}\alpha_n t_n \phi(x_n),\ \sum_{n=1}^{N}\alpha_n t_n = 0$
  - ④ $\alpha_n\left\{1 - t_n\left(\omega^T\phi(x_n)+b\right)\right\} = 0$ for $\forall n \in \{1,\ldots,N\}$

SVM

Notes:

① If $1-t_n\left(\omega^T\phi(x_n)+b\right)\neq 0$ $\left(\text{i.e., } x_n \notin SV\right)$, then $\alpha_n = 0$

② If $\alpha_n \neq 0$, then $1-t_n\left(\omega^T\phi(x_n)+b\right)=0$ $\left(\text{i.e., } x_n \in SV\right)$

$$\therefore \omega_n = \sum_{x_n \in SV} \alpha_n t_n \phi(x_n)$$

③ Decision fcn.: $$y(x)=\omega^T\phi(x)+b=\left\langle \sum_{x_n \in SV} \alpha_n t_n \phi(x_n), \phi(x) \right\rangle + b$$

$$= \sum_{x_n \in SV} \alpha_n t_n k(x_n, x) + b \quad (\text{Kernel trick})$$

(Sparse solution, i.e., we need not all data points, but just support vectors)

④ How to find b?

For all $x_n \in SV$, we have $t_n\left\{\sum_{x_m \in SV} \alpha_m t_m k(x_m, x_n)+b\right\}=1.$ (*)

Hence, by solving the linear equation (*), we can obtain the bias term *b*.

SVM

Non-separable cases:

We start from

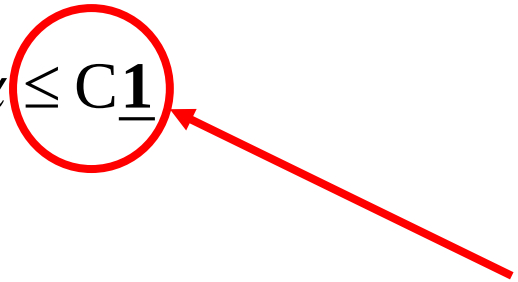
P $\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{n=1}^{N}\xi_n$, where C is trade-off constant.

$$\text{s.t. } t_n\left(\omega^T\phi(x_n)+b\right) \geq 1-\xi_n \quad \forall n \in \{1,\ldots,N\}$$

$$\xi_n \geq 0 \quad \forall n \in \{1,\ldots,N\}$$

Then we obtain

D $\max_{\alpha} -\frac{1}{2}\alpha^T Q\alpha + 1^T\alpha \quad \text{s.t. } 0 \leq \alpha \leq C\underline{1}$



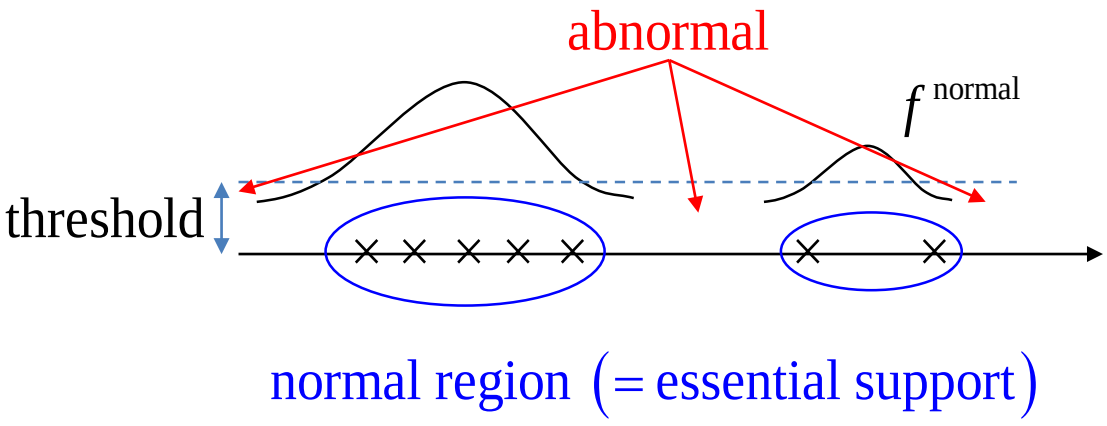
Note: The results remain almost the same except the upper bdd here.

One Class Problems

Goal: To distinguish one class of objects from other objects.

Idea: Given the trn points $\{x_i\}_{i=1}^{N}$ generated from some distribution, find its main support rather than its density.

Traditional approach:



SVM

Notes:

① What we need is the essential support of f^{normal}.

$\text{supp } f \equiv \text{closure}\{x \mid f(x) > 0\}$, $\text{essential supp } f \equiv \text{closure}\{x \mid f(x) > \text{threshold}\}$

② To find f^{normal} is NOT easy.

③ SVDD strategy: Directly focus on the essential support.

Applications:

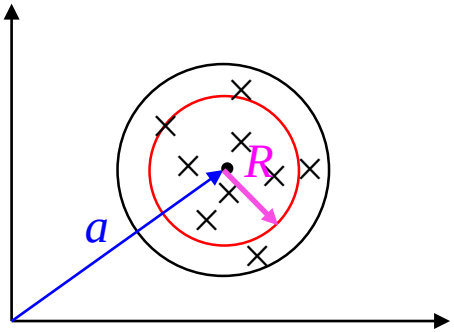
- Network intrusion detection
- Jet engine failure detection
- Fraud detection
- …

SVM

- **Support vector data description (SVDD) [Tax & Duin]**

Main idea:

Given the trn data $\{x_i\}_{i=1}^N$, find a hypersphere which contains as many as possible of the training data while keeping the radius small.



(Conflicting) Goals:

① Your SVDD ball should be small, i.e., $R^2 \downarrow$

② Penalty terms due to outliers should be small, i.e., $\sum_{i=1}^{N} \xi_i \downarrow$

SVM

Performance Index (primal problem):

$$\min_{R^2,a,\xi} J_p\left(R^2,a,\xi\right) = R^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \left\|x_i - a\right\|^2 \le R^2 + \xi_i, \;\; \xi_i \ge 0, \;\; \forall i.$$

Lagrange function:

$$L\left(R^2,a,\xi,\alpha,\eta\right) = R^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left\{\left\|x_i - a\right\|^2 - R^2 - \xi_i\right\} - \sum_{i=1}^{N}\eta_i\xi_i,$$

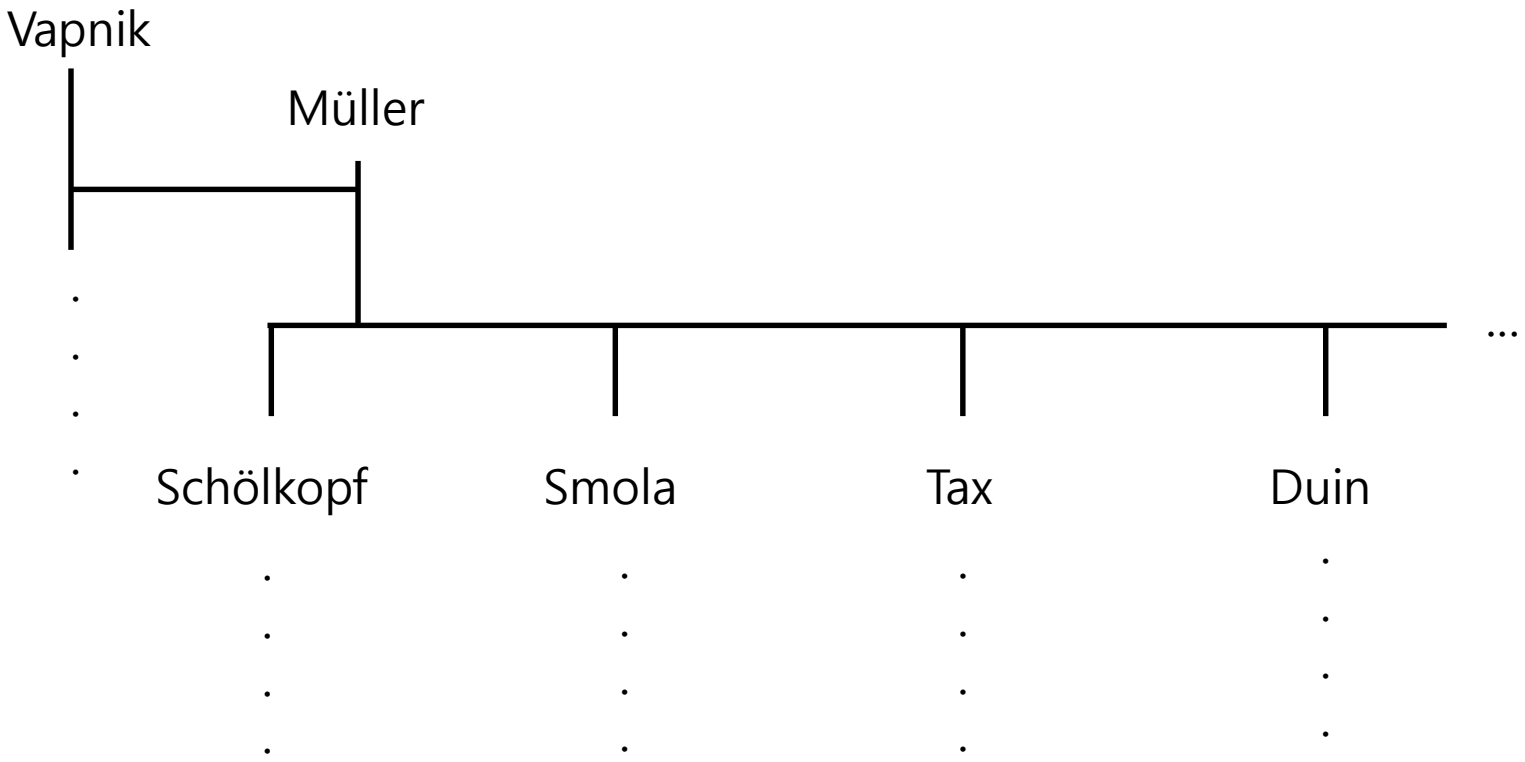
$$\text{where } \alpha_i \ge 0, \;\; \eta_i \ge 0, \;\; \forall i.$$

From the saddle point conditions:

$$\frac{\partial L}{\partial R^2} = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i = 1, \quad \frac{\partial L}{\partial a} = 0 \Rightarrow a = \sum_{i=1}^{N}\alpha_i x_i, \quad \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i \in \left[0,C\right], \; \forall i.$$

SVM

SVM 연구 관련 역사



SVM

- **Reproducing kernel Hilbert space (RKHS)**

▸ Definition (RKHS, Reproducing kernel Hilbert space): Let $X \neq \phi$, and let H be a Hilbert sp. of fcns $f: X \to R$. H is called an RKHS if there exists a fcn $k: X \times X \to R$ such that

① $< f, k(x, \cdot) >_H = f(x)$ for $^\forall f \in H$ (reproducing property)

② k spans H, i.e., $H = \overline{span\{k(x, \cdot) \mid x \in X\}}$

▸ **Mercer's theorem(1909):** Let $X \neq \varnothing$ be a compact set. If $k : X \times X \to R$ is a symmetric continuous kernel such that

$$\int \int_{X^2} k(x, x')f(x)f(x')dxdx' \geq 0 \quad \text{for any } f \in L_2(X),$$

then it can be expanded as

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(x'), \text{ where } \lambda_i > 0, \ ^\forall i.$$

(Here, the $\psi_i \in L_2(X)$ are the orthonormal eigenfunctions of operator $T_k : L_2(X) \to L_2(X)$ defined by $(T_k f)(x) = \int_X k(x, x')f(x')dx'$, and the λ_i are associated eigenvalues.)

Explicit construction of the RKHS with Mercer kernel

For a Mercer kernel k,

we can define the corresponding RKHS

$$H = \left\{ f \mid f(x) = \sum_i f_i \psi_i(x), \text{ and } \|f\|_H < \infty \right\},$$

where the RKHS norm $\|f\|_H^2 = \left\langle \sum_i f_i \psi_i(\cdot), \sum_i f_i \psi_i(\cdot) \right\rangle_H = \sum_i \frac{f_i^2}{\lambda_i}.$

Similiarly, its inner product is defined as

$$\langle f, g \rangle_H = \left\langle \sum_i f_i \psi_i(\cdot), \sum_i g_i \psi_i(\cdot) \right\rangle_H = \sum_i \frac{f_i g_i}{\lambda_i}.$$

Note:

$$\langle f, k(x,\cdot) \rangle_H = \left\langle \sum_i f_i \psi_i(\cdot), \sum_i \lambda_i \psi_i(x) \psi_i(\cdot) \right\rangle_H = \sum_i \frac{f_i \lambda_i \psi_i(x)}{\lambda_i} = \sum_i f_i \psi_i(x) = f(x).$$

SVM

- **Representer theorem**

Given a p.d. kernel k on $X \times X$, a training set $\{(x_i, y_i)\}_{i=1}^{N} \in X \times R$,

a strictly monotone increasing real-valued function Ω on $[0, \infty)$,

and an arbitrary cost function

$$c : \left(X \times R^2\right)^N \to R \cup \{\infty\},$$

any functional $f \in H^{RKHS}$ minimizing the regularized risk functional

$$c\left(\left(x_1, y_1, f(x_1)\right), \ldots, \left(x_N, y_N, f(x_N)\right)\right) + \Omega\left(\|f\|_H\right)$$

admits a representation of the form $f(\cdot) = \sum_{i=1}^{N} a_i k(x_i, \cdot)$

SVM

Proof:

Decompose $f \in H$ into

$$f = \sum_i \alpha_i k(x_i, \cdot) + f_\perp,$$

where for $\forall j$, $\left\langle f, k(x_j, \cdot) \right\rangle_H = 0.$

Then, application of f to an arbitrary training point x_j yields

$$\begin{aligned} f(x_j) &= \left\langle f_\perp, k(x_j, \cdot) \right\rangle_H \quad (\text{reproducing property}) \\ &= \left\langle \sum_i \alpha_i k(x_i, \cdot) + f_\perp, k(x_j, \cdot) \right\rangle_H \\ &= \sum_i \alpha_i \left\langle k(x_i, \cdot), k(x_j, \cdot) \right\rangle_H = \sum_i \alpha_i k(x_i, x_j), \end{aligned}$$

which is independent of $f_\perp$.

$(\therefore c$항은 $f_\perp$에 무관함$)$

SVM

Since $f_\perp \perp \sum_i \alpha_i k(x_i, \cdot)$, and Ω is strictly monotonic, we get

$$\Omega\left(\|f\|_H\right) = \Omega\left(\left\|\sum_i \alpha_i k(x_i, \cdot) + f_\perp\right\|_H\right)$$

$$= \Omega\left(\sqrt{\left\|\sum_i \alpha_i k(x_i, \cdot)\right\|_H^2 + \|f_\perp\|_H^2}\right) \quad \text{(Pythagorean theorem)}$$

$$\geq \Omega\left(\left\|\sum_i \alpha_i k(x_i, x)\right\|_H\right),$$

with equality occuring iff $f_\perp = 0$.

Hence, any minimizer must have $f_\perp = 0$.

$(\because c$항은 $f_\perp$에 무관하고, Ω 항은 $f_\perp$ 가 작아야 최소화됨$)$

Consequently, any solution takes the following form: $f(\cdot) = \sum_i \alpha_i k(x_i, \cdot)$

Outline

- Introduction
- Support Vector Machines
- **Gaussian Processes**
- Concluding Remarks

GP

- Def: Stochastic Process $X(t)$
  - An indexed family of random variables

- Def: Gaussian Process
  - A stochastic process $y(x)$ is a Gaussian process, if for ∀ finite index set $\{x_1, \dots, x_N\}$, $y(x_1), \dots, y(x_N)$ are jointly Gaussian.

*Note: $\begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix} \sim N(\underline{m}, K)$ where $\underline{m} = \underline{0}$, $K = [K_{ij}]$

$$K_{ij} = cov[y(x_i), y(x_j)] = E[y(x_i)y(x_j)] = k_\theta(x_i, x_j)$$

GP

- **Gaussian Process Regression (GPR)**

  Trn data : $D = \{(x_n, t_n)\}_{n=1}^{N}$, where $t_n \in R$

  Model : $t = y(x) + \epsilon$, where y(x) is a Gaussian process, and $\epsilon \sim N(0, \beta^{-1})$

  Note : $\underline{y} = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix} \sim N(\underline{0}, K)$, where $K_{ij} = k_\theta(x_i, x_j)$

  $\underline{t} = y(\underline{x}) + \underline{\epsilon}$, where $y(\underline{x}) \sim N(0, K)$, $\underline{\epsilon} \sim N(0, \beta^{-1} I_N)$

  Hence, $\underline{t} \sim N(0, K + \beta^{-1} I)$

  $\therefore t(\underline{x})$ is also a GP

GP

- Predictive distribution:

  Test input: x_*

  Target output: t_*

  Predictive distribution for t_*: $p(t_*|\underline{t}) = p(t_*|D, x_*)$

Note: $\begin{bmatrix} t_1 \\ \vdots \\ t_N \\ t_* \end{bmatrix} \sim N\left(\underline{0}, \begin{bmatrix} K + \beta^{-1}I & \underline{k_*} \\ \underline{k_*^T} & k_\theta(x_*, x_*) + \beta^{-1} \end{bmatrix}\right)$,

where $K + \beta^{-1}I = C$, $k_\theta(x_*, x_*) + \beta^{-1}$=c, $\underline{k_*} = \begin{bmatrix} k_\theta(x_1, x_*) \\ \vdots \\ k_\theta(x_N, x_*) \end{bmatrix}$

$$\therefore \begin{bmatrix} t_1 \\ \vdots \\ t_N \\ t_* \end{bmatrix} \sim N\left(\underline{0}, \begin{bmatrix} C & \underline{k_*} \\ \underline{k_*^T} & c \end{bmatrix}\right)$$

GP

Hence from the conditional dist. formula, we have

$p(t_* | \underline{t}) = N(m(x_*), \sigma^2(x_*))$, where

$m(\underline{x_*}) = k_*^T C^{-1} \underline{t} = \sum_{n=1}^{N} \alpha_n k(x_n, x_*)$, $\sigma^2(x_*) = c - k_*^T C^{-1} k_*$

※ Conditional distribution formula for Gaussians

- If $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$,

  then $X_{2|1} \sim N(\mu_{2|1}, \Sigma_{2|1})$,

  where $\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$

  $\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

GP

- **Gaussian Process Classification (GPC)**

Trn data: $D = \{(x_n, t_n)\}_{n=1}^{N}$, where $t_n = \{0,1\}$

Model:

$$x \longmapsto a(x) \longmapsto \sigma(a) \longmapsto t$$

$$x_* \longmapsto a_* = a(x_*),$$

$$\text{where } p(t|a) = \begin{cases} \sigma & if \quad t = 1 \\ 1 - \sigma & if \quad t = 0 \end{cases}$$

$$= \sigma^t (1 - \sigma)^{1-t}$$

GP

Test input x_*, then t_* ?

Goal : To find $p(t_*|\underline{t}) = p(t_*|D, x_*)$

Note : $\begin{bmatrix} a_1 \\ \vdots \\ a_N \\ a_* \end{bmatrix} \sim N(\underline{0}, C_{N+1})$, where $C_{N+1} = [C_{nm}]$

$C_{nm} = k_\theta(x_n, x_m)$

$$p(t_x = 1|\underline{t}) = p\left(t_* = 1 \middle| D, \underline{x_*}\right)$$

$$= \int p(t_* = 1, a_*|D, x_*)da_* = \int \underbrace{p(t_* = 1|a_*)}_{\sigma(a_*)} \overline{p(a_*|D, \underline{x_*})}^{\text{Posterior (Gaussian)}} da_*$$

$$= \int (sigmoid) * (gaussian) da_*$$

GP

- One can use approximate inference methods such as variational inference, sampling, etc.

- Methods of relating the GP output, $a(x)$, to success prob. :
  - Logistic sigmoid
  - Cdf of the std normal (probit)
  - Threshold ...

Outline

- Introduction
- Support Vector Machines
- Gaussian Processes
- **Concluding Remarks**

Concluding Remarks

- Kernel methods : SVM, SVDD, GP, …
- Application domains:
  - Pattern classification
  - Function approximation
  - Anomaly detection
  - Feature extraction
  - Pattern de-noising
  - Kernel-based reinforcement learning, etc.
- Deep vs. Sparse Kernel

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning.* (Springer-Verlag , 2007)

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* (MIT Press , 2012)

[3] B. Schölkopf and J. A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* (MIT Press , 2001)

[4] D. Tax and R. Duin, "Support vector data description," *Machine Learning*, Vol. 54, pp. 45-66, 2004

Acknowledgement: 본 자료 작성에 동참한 고려대학교 제어계측공학과 허성만, 김태환, 박정호학생과 고려대학교 수학과 김재인학생에게 감사.