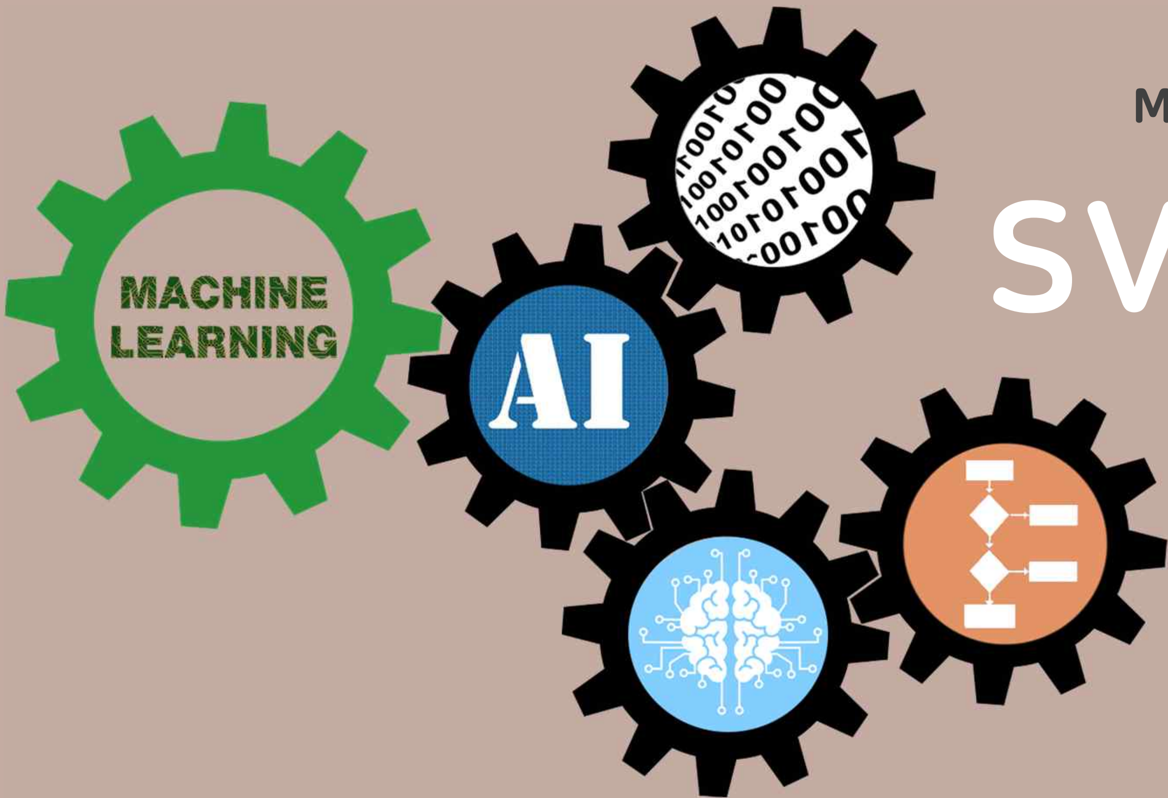


Machine Learning 8강

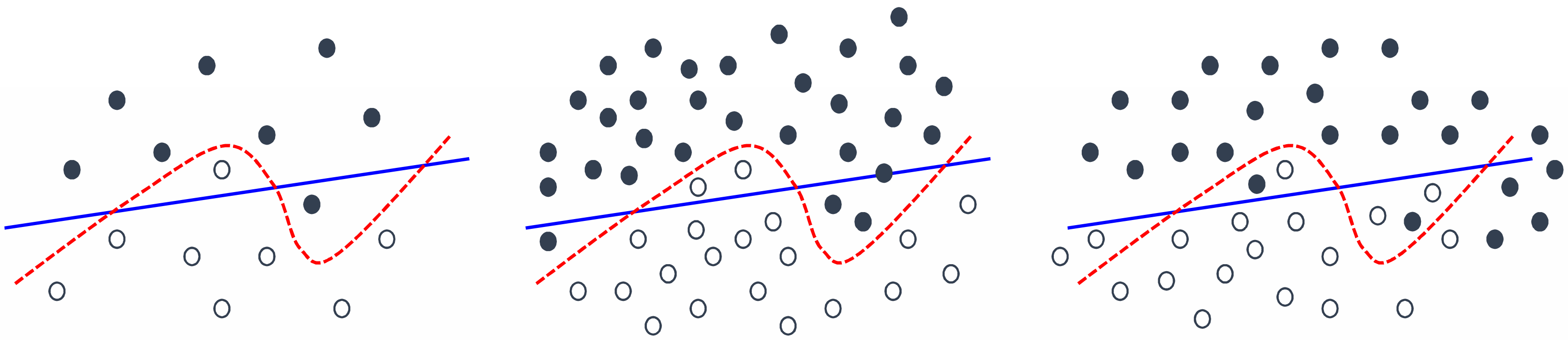
SVM과 커널법

컴퓨터과학과 이관용 교수

학습목차

1 선형 분류기

학습 시스템의 복잡도와 일반화 오차의 관계



학습 데이터의 경우

학습 오차

(선형) 2:0 (비선형)

학습에 사용되지 않은 데이터의 경우

일반화 오차

5:0

6:10

비선형 분류기가 더 큰 일반화 오차를 가짐

"과다적합"

과다학습을 피하고 일반화 오차를 줄이기 위해서는
학습 시스템의 복잡도를 적절히 조정하는 것이 중요

선형 초평면 분류기

선형 분류기 linear classifier

- 선형 판별함수를 기반으로 분류를 수행하는 학습 시스템
- 분류 시스템의 복잡도가 가장 낮으며, 분류 성능도 좋지 못함
- 과다적합의 발생을 피할 수 있음
- SVM → 일반화 오차를 최소화할 수 있는 방향으로 학습이 이루어지도록 설계된 선형 분류기

선형 초평면 분류기

입력 x에 대한 선형 초평면 판별함수

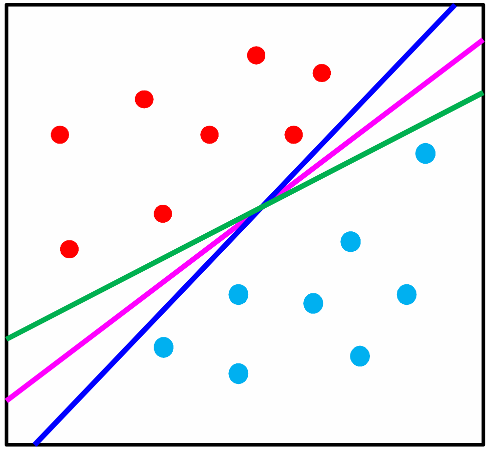
$$g(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + w_0 = \boldsymbol{w}^T \boldsymbol{x} + w_0 = \sum_{i=1}^{n} w_i x_i + w_0$$

결정규칙

$$f(\boldsymbol{x}) = \text{sign}(g(\boldsymbol{x}))$$

$$f(\boldsymbol{x}) = 1 \rightarrow \boldsymbol{x} \in C_1$$

$$f(\boldsymbol{x}) = -1 \rightarrow \boldsymbol{x} \in C_2$$



최소 학습 오차를 만족하는 여러 가지 선형 결정경계가 존재

↓

SVM에서는 여러 선형 결정경계 중

일반화 오차를 최소로 하는 최적의 경계를 찾기 위해

마진margin의 개념을 도입하여 학습의 목적함수를 정의

2

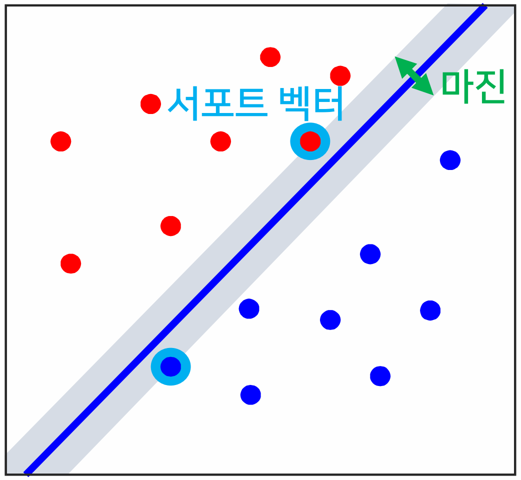
SVM 분류기

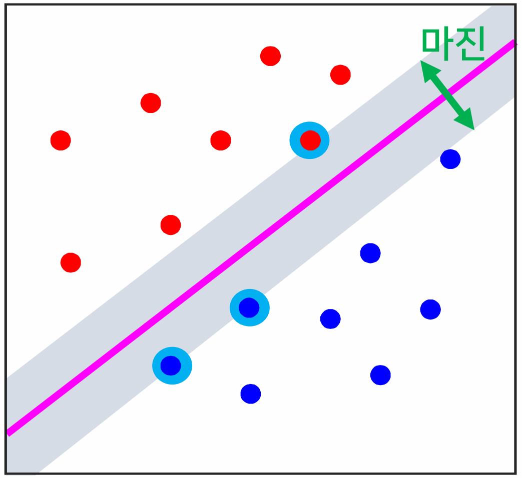
최대 마진 분류기 maximum margin classifier

마진 → 학습 데이터들 중에서 결정경계에 가장 가까운 데이터로부터 결정경계까지의 거리

서포트 벡터 support vector → 결정경계에 가장 가까운 곳에 위치한 데이터

결정경계에 따른 마진과 서포트 벡터의 차이





일반화 오차를 작게

↓

클래스 간의 간격을 최대로

↓

마진을 최대로

↓

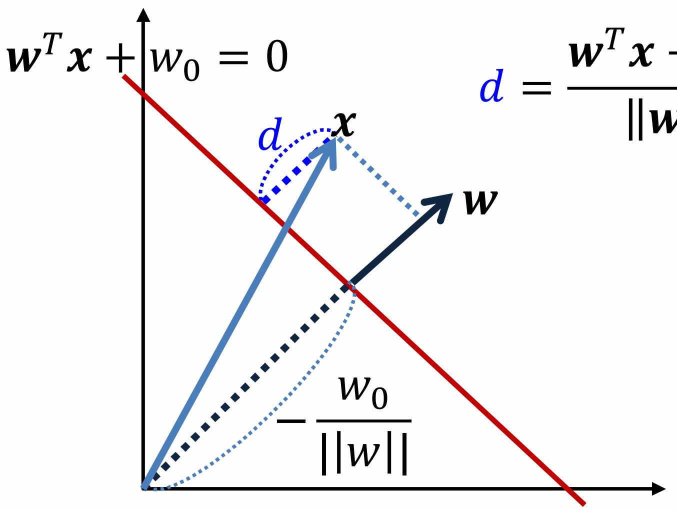
『최대 마진 분류기』, 『SVM』

최대 마진 분류기

최대 마진을 가진 선형 결정경계(초평면)를 얻기 위한 선형 판별함수

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=1}^{n} w_i x_i + w_0 = 0$$

한 점 x에서 결정경계까지의 거리



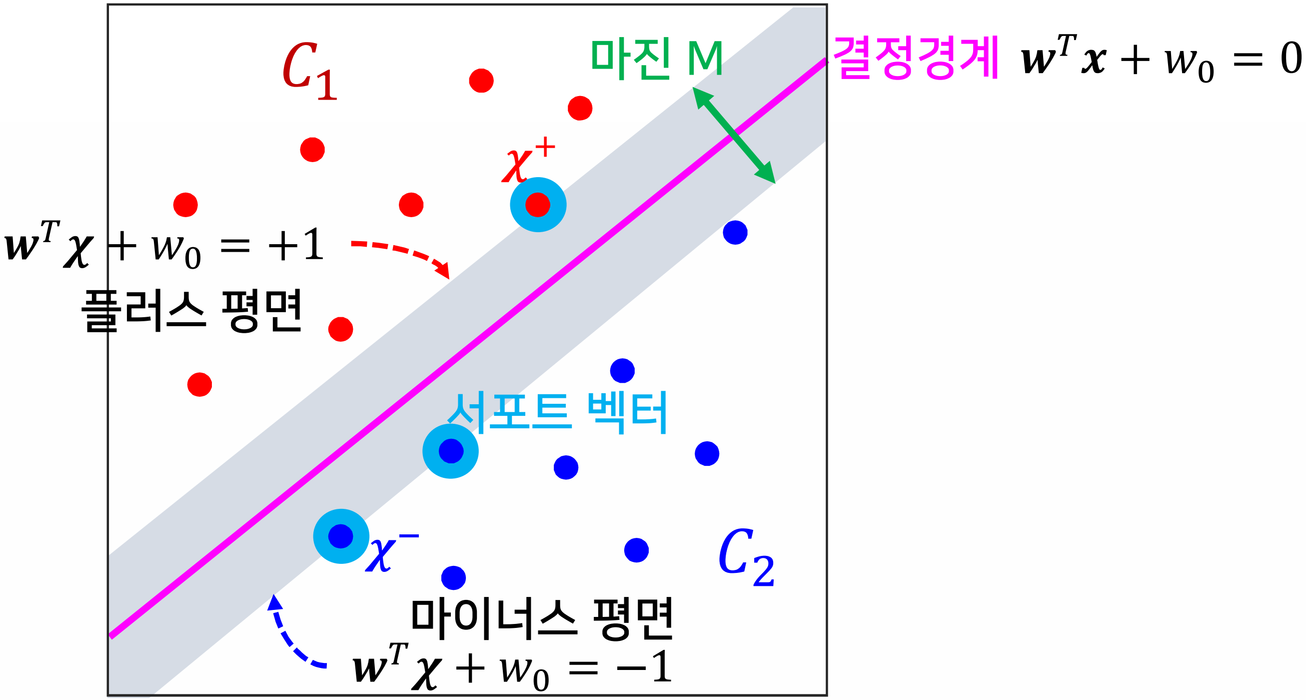
$$d = \frac{\boldsymbol{w}^T\boldsymbol{x} + w_0}{\|\boldsymbol{w}\|}$$

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x} + w_0 > 0 \rightarrow C_1 \text{ 영역} \\ \boldsymbol{w}^T\boldsymbol{x} + w_0 < 0 \rightarrow C_2 \text{ 영역} \end{cases}$$

서포트 벡터 χ

$$\begin{cases} \boldsymbol{w}^T\chi + w_0 = +1 \quad \text{if} \quad \chi \in C_1 \\ \boldsymbol{w}^T\chi + w_0 = -1 \quad \text{if} \quad \chi \in C_2 \end{cases}$$

마진 계산



마진 M

$$\mathrm{M} = |\boldsymbol{\chi}^+ - \boldsymbol{\chi}^-| = \frac{1}{\|\boldsymbol{w}\|}\left((\boldsymbol{w}^T\boldsymbol{\chi}^+ + w_0) - (\boldsymbol{w}^T\boldsymbol{\chi}^- + w_0)\right) = \frac{2}{\|\boldsymbol{w}\|}$$

마진의 최대화 $\dashrightarrow$ $\|\boldsymbol{w}\|$의 최소화

SVM의 학습

학습 데이터 집합 $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,N}$

$$\begin{cases} y_i = +1 & \text{if } \boldsymbol{x}_i \in C_1 \\ y_i = -1 & \text{if } \boldsymbol{x}_i \in C_2 \end{cases}$$

추정해야 할 파라미터 $\boldsymbol{w}, w_0$가 만족해야 하는 조건

$$\begin{cases} (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq +1 & \text{for } y_i = +1 \\ (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \leq -1 & \text{for } y_i = -1 \end{cases} \dashrightarrow y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 \geq 0$$

라그랑주 승수

$(\alpha_i \geq 0, i = 1, \cdots, N)$

최소화할 목적함수

$$J(\boldsymbol{w}) = \frac{\|\boldsymbol{w}\|^2}{2}$$

「라그랑주 함수」

$$J(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1\}$$

파라미터 $\boldsymbol{w}, w_0$ 에 대해 미분

SVM의 학습

$$\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i = 0 \quad \dashrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial J(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$J(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i \{y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1\}$$

$J(\boldsymbol{w}, w_0, \alpha)$에 대한 이원적 문제 dual problem

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j {\boldsymbol{x}_i}^T \boldsymbol{x}_j$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \ \ \alpha_i \geq 0 \ (i = 1, \cdots, N)$$

이차계획법을 이용하여 $\hat{\alpha}_i$를 찾아서 $\boldsymbol{w}, w_0$ 추정

SVM의 학습

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

$$\hat{w}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\boldsymbol{w}}^T \boldsymbol{x}_i) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{N} \hat{\alpha}_j y_j \boldsymbol{x}_j{}^T \boldsymbol{x}_i\right)$$

$$J(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i \{y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1\}$$

대부분의 학습 데이터에 대응되는 라그랑주 승수 $\hat{\alpha}_i$는 0이 됨

오직 $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1 = 0$인 경우만 $\hat{\alpha}_i$가 0이 아닌 값을 가짐

서포트 벡터 데이터의 $\hat{\alpha}_i \neq 0$

↓

서포트 벡터에 해당하는 $\hat{\alpha}_i$와 학습 데이터 $(\boldsymbol{x}_i, y_i)$ 만 필요

↓

분류를 위해 저장할 데이터의 개수와 계산량의 현격한 감소

SVM에 의한 분류

판별함수

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=1}^{n} w_i x_i + w_0 = 0$$

$$\widehat{\boldsymbol{w}} = \sum_{\boldsymbol{x}_i \in X_s} \hat{\alpha}_i y_i \boldsymbol{x}_i \qquad \widehat{w}_0 = \frac{1}{N_s} \sum_{\boldsymbol{x}_i \in X_s} \left(y_i - \sum_{\boldsymbol{x}_j \in X_s} \hat{\alpha}_j y_j {\boldsymbol{x}_j}^T \boldsymbol{x}_i \right)$$

$$f(\boldsymbol{x}) = \mathrm{sign}\big(g(\boldsymbol{x})\big) = \mathrm{sign}(\widehat{\boldsymbol{w}}^T x + \widehat{w}_0) = \mathrm{sign}\left(\sum_{i=1}^{N} \hat{\alpha}_i y_i {\boldsymbol{x}_i}^T \boldsymbol{x} + \widehat{w}_0 \right)$$

$$\begin{cases} f(\boldsymbol{x}) = 1 \rightarrow \boldsymbol{x} \in C_1 \\ f(\boldsymbol{x}) = -1 \rightarrow \boldsymbol{x} \in C_2 \end{cases}$$

선형 SVM 분류기의 학습과 인식 단계

① N개의 입출력 쌍으로 이루어진 학습 데이터 집합 $X = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,N}$ 을 준비함. 이때 목표 출력값은 $y_i \in \{-1,1\}$ $(i = 1,\cdots,N)$을 만족함.

② 다음과 같은 과정을 통해 SVM을 학습함

②-1 학습 데이터를 이용하여 파라미터 추정을 위한 목적함수 $Q(\boldsymbol{\alpha})$를 정의

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i{}^T \boldsymbol{x}_j \qquad \sum_{i=1}^{N} \alpha_i y_i = 0,\ \ \alpha_i \geq 0\ (i = 1,\cdots,N)$$

②-2 주어진 조건을 만족하면서 $Q(\boldsymbol{\alpha})$를 최대화하는 추정치 $\hat{\alpha}_i$를 이차계획법에 의해 찾음

②-3 $\hat{\alpha}_i \neq 0$이 되는 서포트 벡터를 찾아 집합 $X_s = \{\boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0\}$를 생성

선형 SVM 분류기의 학습과 인식 단계

②-4 $\hat{\alpha}_i$와 서포트 벡터를 이용하여 $\hat{w}_0$를 계산

$$\hat{w}_0 = \frac{1}{N_s} \sum_{\boldsymbol{x}_i \in X_s} \left(y_i - \sum_{\boldsymbol{x}_j \in X_s} \hat{\alpha}_j y_j {\boldsymbol{x}_j}^T \boldsymbol{x}_i \right)$$

N_s는 집합 X_s의 원소의 수임

②-5 서포트 벡터 집합 $X_s = \{\boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0\}$와 파라미터 벡터 $\hat{\boldsymbol{\alpha}}$, 그리고 $\hat{w}_0$를 저장

③ 새로운 데이터 $\boldsymbol{x}$가 주어지면, 저장해둔 서포트 벡터와 파라미터를 이용하여 다음 함수로 분류를 수행

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{\boldsymbol{x}_i \in X_s} \hat{\alpha}_i y_i {\boldsymbol{x}_i}^T \boldsymbol{x} + \hat{w}_0 \right)$$

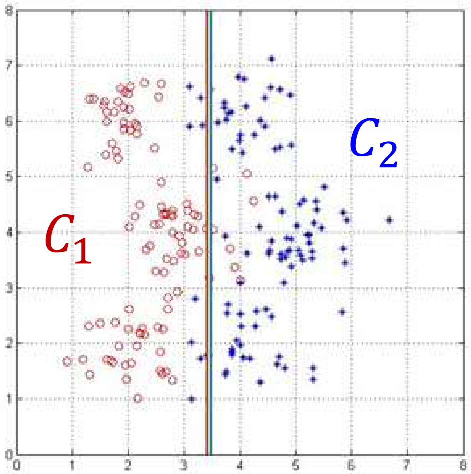
다중 클래스 분류 문제에의 적용

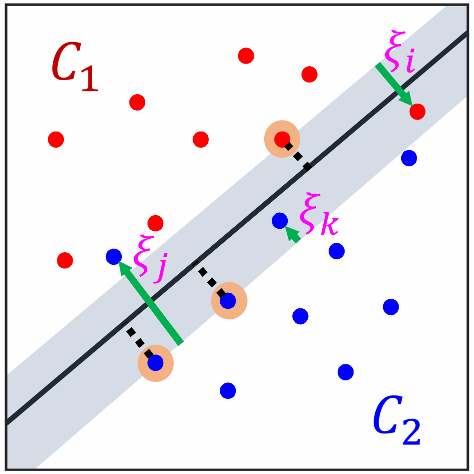
- 1대 나머지 방법 one-versus-the-rest
  - 가장 보편적인 방법 → k개의 개별적인 SVM 분류기 사용
  - k번째 SVM → k번째 클래스와 나머지 $k-1$개의 클래스를 분류
    - ✓ 클래스 C_k에 해당하는 데이터는 +1이 되도록 학습
    - ✓ 나머지 $k-1$개의 클래스의 데이터에 대해서는 -1이 되도록 학습
  - 문제 → 애매모호한 결정 영역, 학습 데이터 집합의 크기가 불균형적
- 1대1 방법 one-versus-one
  - 가능한 모든 클래스의 쌍에 대한 서로 다른 $\frac{k(k-1)}{2}$개의 SVM과 보팅
  - 문제 → 애매모호한 결정 영역, 학습/테스트를 위한 높은 계산 비용

슬랙변수를 가진 SVM

선형 분리가 불가능한 데이터 처리를 위해 슬랙변수 ξ 도입

잘못 분류된 데이터로부터
해당 클래스의 결정경계까지의 거리





슬랙변수를 포함한 분류 조건

$$\begin{cases} (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq +1 - \xi_i & \text{for } y_i = +1 \\ (\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \leq -1 + \xi_i & \text{for } y_i = -1 \end{cases} \longrightarrow y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \quad (i = 1, \cdots, N)$$

슬랙변수의 값이 클수록 더 심한 오분류를 허용!

슬랙변수를 가진 SVM의 파라미터 추정

$$J(\boldsymbol{w},\ \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + c\sum_{i=1}^{N}\xi_i \qquad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 - \xi_i,\ \xi_i \geq 0 \qquad (i = 1,\ \cdots,\ N)$$

$$J(\boldsymbol{w},\ w_0,\ \boldsymbol{\alpha},\ \boldsymbol{\xi},\ \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + c\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) - 1 + \xi_i\} - \sum_{i=1}^{N}\beta_i\xi_i$$

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \qquad \sum_{i=1}^{N}\alpha_i y_i = 0,\ 0 \leq \boxed{\alpha_i < c}$$

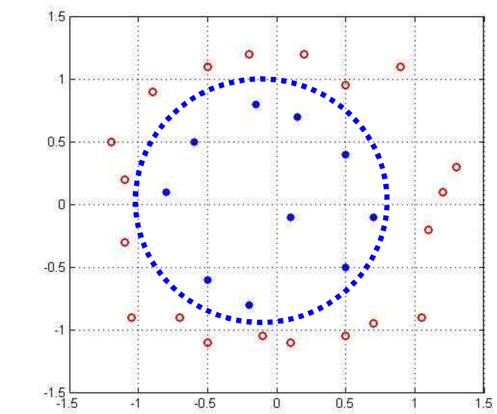
$$\widehat{\boldsymbol{w}} = \sum_{i=1}^{N}\hat{\alpha}_i y_i \boldsymbol{x}_i \qquad \widehat{w}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\boldsymbol{w}}^T\boldsymbol{x}_i) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{N}\hat{\alpha}_j y_j \boldsymbol{x}_j^T\boldsymbol{x}_i\right)$$

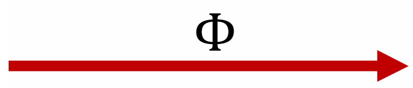
$\hat{\alpha}_i$가 정해지면 $\widehat{\boldsymbol{w}}, \widehat{w}_0$의 값은 슬랙변수가 없는 경우와 완전히 동일!

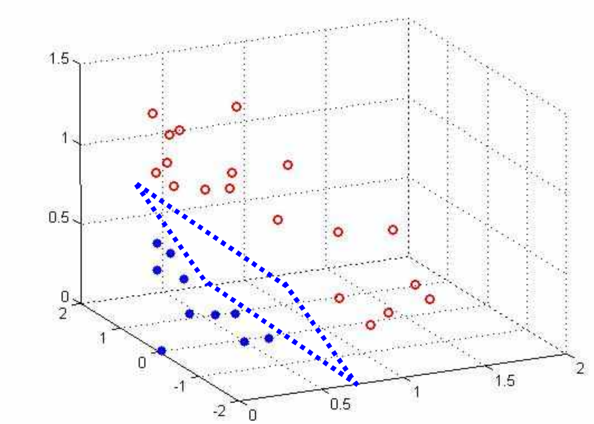
3 커널법

비선형 문제로의 확장

저차원의 입력 x를 고차원의 공간의 값 $\Phi(x)$로 매핑시키는 함수 Φ







$$\Phi : R^2 \rightarrow R^3$$

$$(x_1, x_2) \mapsto (x_1^2,\ \sqrt{2}\, x_1 x_2\ ,\ x_2^2)$$

- **고차원 문제로 선형화하면 간단한 선형 분류기를 사용한 분류가 가능**
- **계산량 증가와 같은 부작용 발생 → 「커널법」으로 해결**

고차원 매핑을 통해 비선형 문제를 선형화하여 해결하면서
커널 함수를 통해 계산량 증가의 문제를 해결하는 방법

커널법과 SVM

n차원의 입력 $\boldsymbol{x}$를 m차원의 특징 데이터 $\Phi(\boldsymbol{x})$로 매핑시킨 후 SVM으로 분류한다고 가정

SVM에서의 연산은 개개의 값 $\Phi(\boldsymbol{x})$가 아니라 두 벡터의 내적 $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$를 사용

고차원 매핑 $\Phi(\boldsymbol{x})$를 정의하는 대신에 $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$를 하나의 함수 $k(\boldsymbol{x}, \boldsymbol{y})$로 정의하여 사용

커널 함수 kernel function

$$
\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{y}) &= \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y}) \\
&= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2) \cdot (y_1^2, \sqrt{2}\, y_1 y_2, y_2^2) \\
&= (\boldsymbol{x} \cdot \boldsymbol{y})^2
\end{aligned}
$$

파라미터 추정을 위한 라그랑주 함수

$$
L(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i \{ y_i (\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0) - 1 \}
$$

이원적 문제의 함수 $Q(\boldsymbol{\alpha})$

$$
\begin{aligned}
Q(\boldsymbol{\alpha}) &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)
\end{aligned}
$$

커널법과 SVM

파라미터 추정 후 커널 함수만으로 표현된 분류 함수

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{i=1}^{N} \widehat{\alpha}_i y_i \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}_i) + \widehat{w}_0\right) = \mathrm{sign}\left(\sum_{i=1}^{N} \widehat{\alpha}_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + \widehat{w}_0\right)$$

대표적인 커널 함수

선형 커널	$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y})$
다항식 커널	$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y} + c)^d$
시그모이드 커널	$k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\theta_1 \boldsymbol{x} \cdot \boldsymbol{y} + \theta_2)$
가우시안 커널	$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\frac{\Vert\boldsymbol{x} - \boldsymbol{y}\Vert^2}{2\sigma^2}\right\}$

"사용자 정의 파라미터
(하이퍼파라미터)"

슬랙변수와 커널을 가진 SVM

① N개의 입출력 쌍으로 이루어진 학습 데이터 집합 $X = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,N}$ 을 준비하고, 하이퍼파라미터 c와 커널 함수 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$를 정의함. 이때 목표 출력값은 $y_i \in \{-1,1\}$ $(i = 1, \cdots, N)$을 만족함.

② 다음과 같은 과정을 통해 SVM을 학습함

②-1 학습 데이터를 이용하여 파라미터 추정을 위한 목적함수 $Q(\boldsymbol{\alpha})$를 정의

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \qquad \sum_{i=1}^{N} \alpha_i y_i = 0, \ \ 0 \le \alpha_i \le c \ \ (i = 1, \cdots, N)$$

②-2 주어진 조건을 만족하면서 $Q(\boldsymbol{\alpha})$를 최대화하는 추정치 $\hat{\alpha}_i$를 이차계획법에 의해 찾음

②-3 $\hat{\alpha}_i \neq 0$이 되는 서포트 벡터를 찾아 집합 $X_s = \{\boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0\}$를 생성

슬랙변수와 커널을 가진 SVM

②-4 $\hat{\alpha}_i$와 서포트 벡터를 이용하여 $\hat{w}_0$를 계산

$$\hat{w}_0 = \frac{1}{N_s} \sum_{\boldsymbol{x}_i \in X_s} \left(y_i - \sum_{\boldsymbol{x}_j \in X_s} \hat{\alpha}_j y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)$$

N_s는 집합 X_s의 원소의 수임

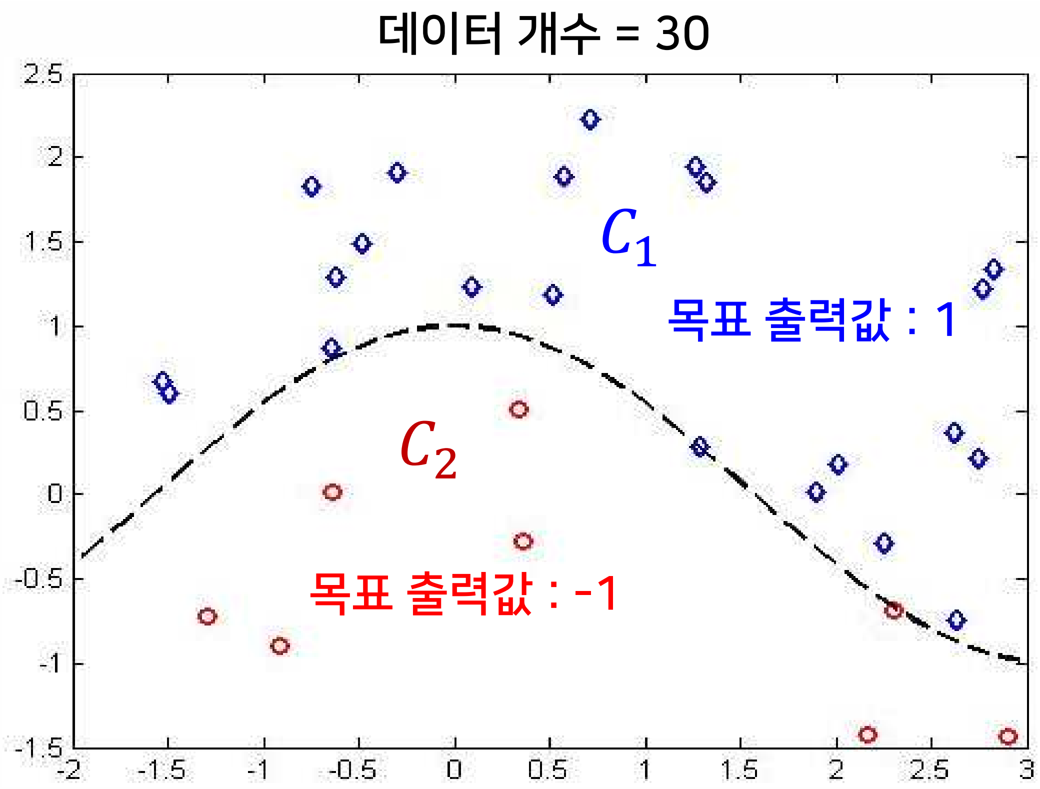
②-5 서포트 벡터 집합 $X_s = \{\boldsymbol{x}_i \in X \mid \hat{\alpha}_i \neq 0\}$와 파라미터 벡터 $\hat{\boldsymbol{\alpha}}$, 그리고 $\hat{w}_0$를 저장

③ 새로운 데이터 $\boldsymbol{x}$가 주어지면, 저장해둔 서포트 벡터와 파라미터를 이용하여 다음 판별함수로 분류를 수행

$$f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{\boldsymbol{x}_i \in X_s} \hat{\alpha}_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \hat{w}_0 \right)$$

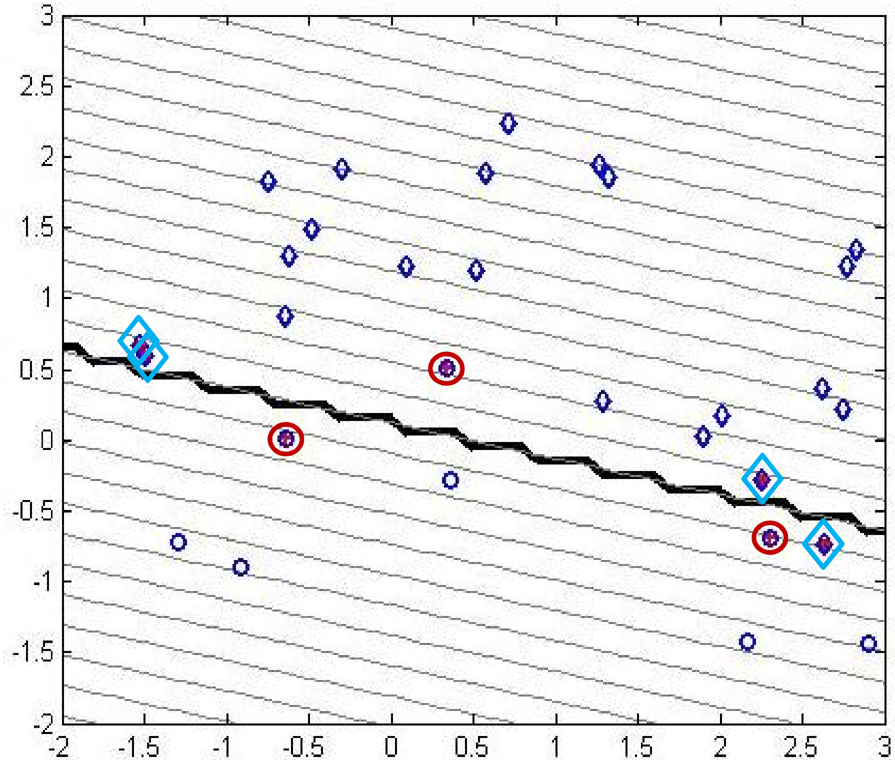
[예] 비선형 결정경계를 가진 이진 분류

○ 학습 데이터



실험 결과

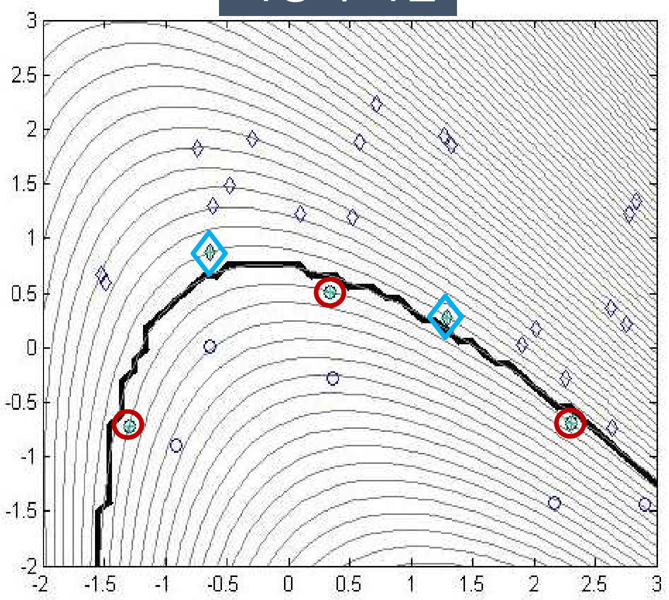
선형 커널 함수를 사용한 경우의 결정경계와 서포트벡터



실험 결과

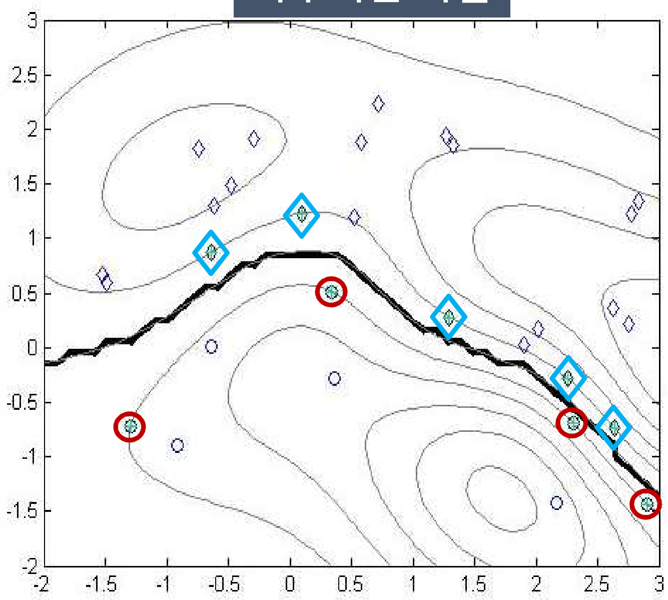
다항식 커널과 가우시안 커널을 사용한 경우의 결정경계와 서포트 벡터

다항식 커널



c=1, d=2, 5개

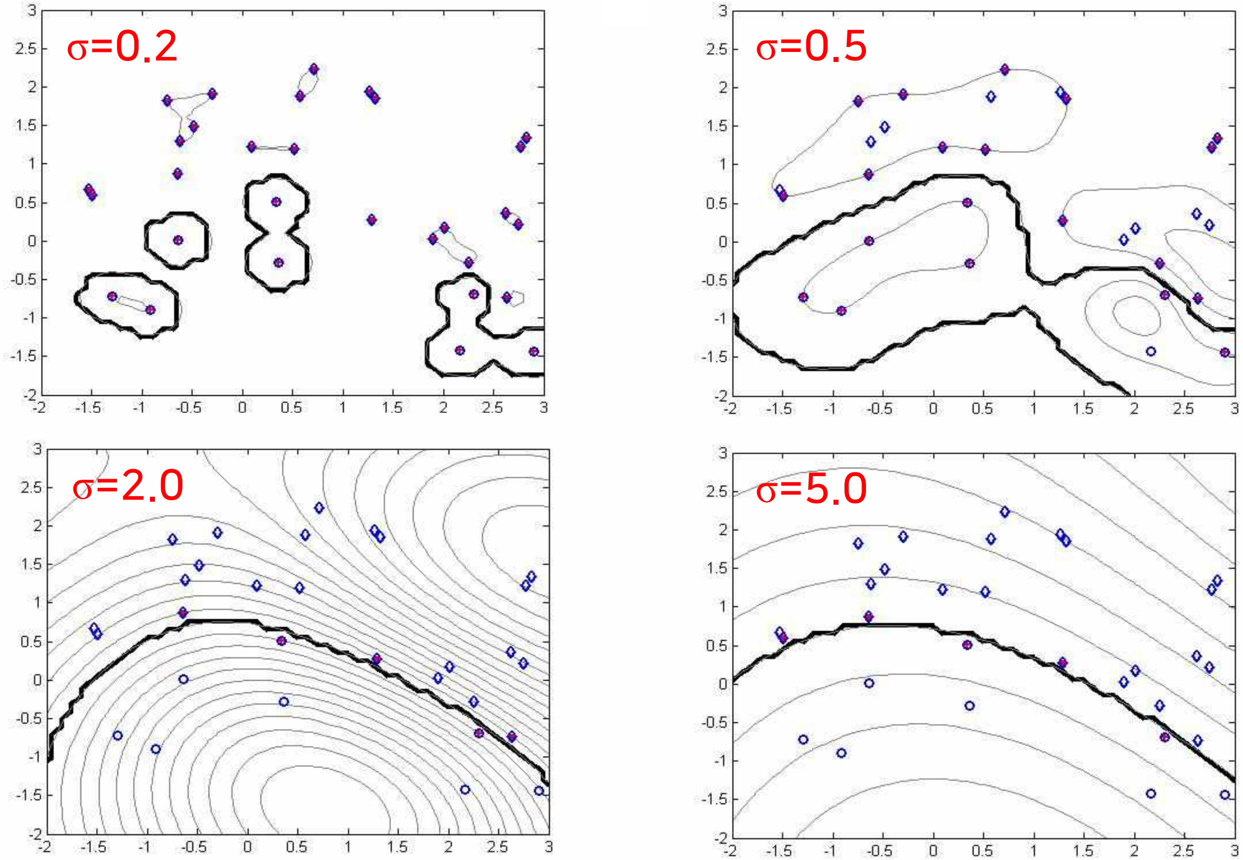
가우시안 커널

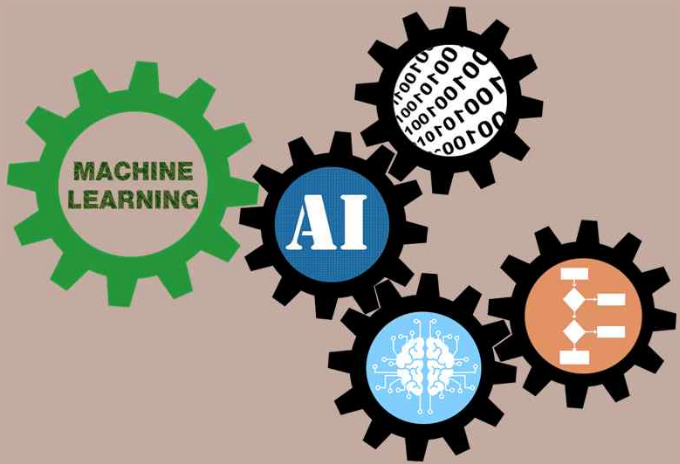


σ=1, 9개

실험 결과

가우시안 커널의 파라미터 σ값에 따른 결정경계의 변화





다음시간안내

제9강

신경망 (1)