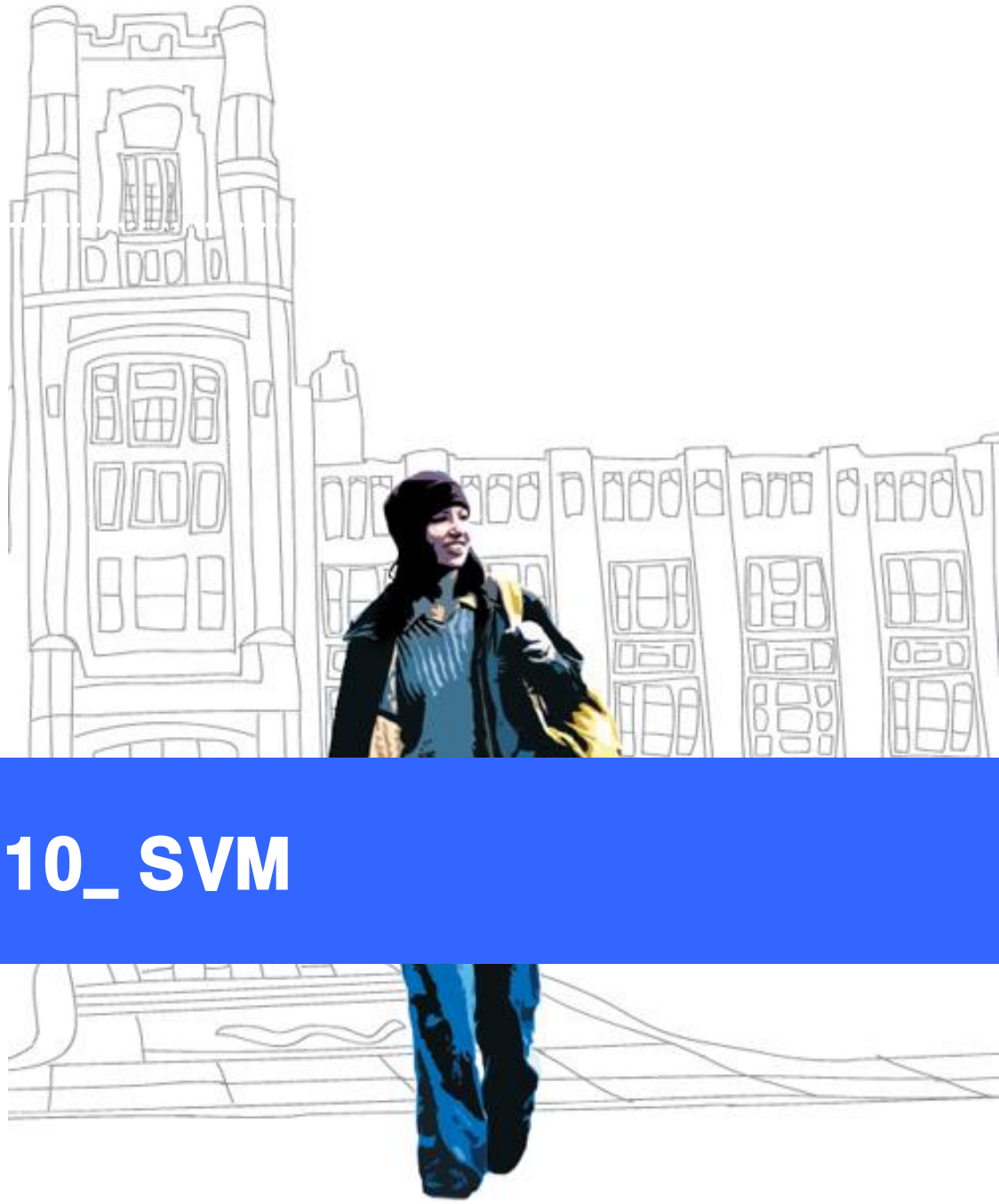
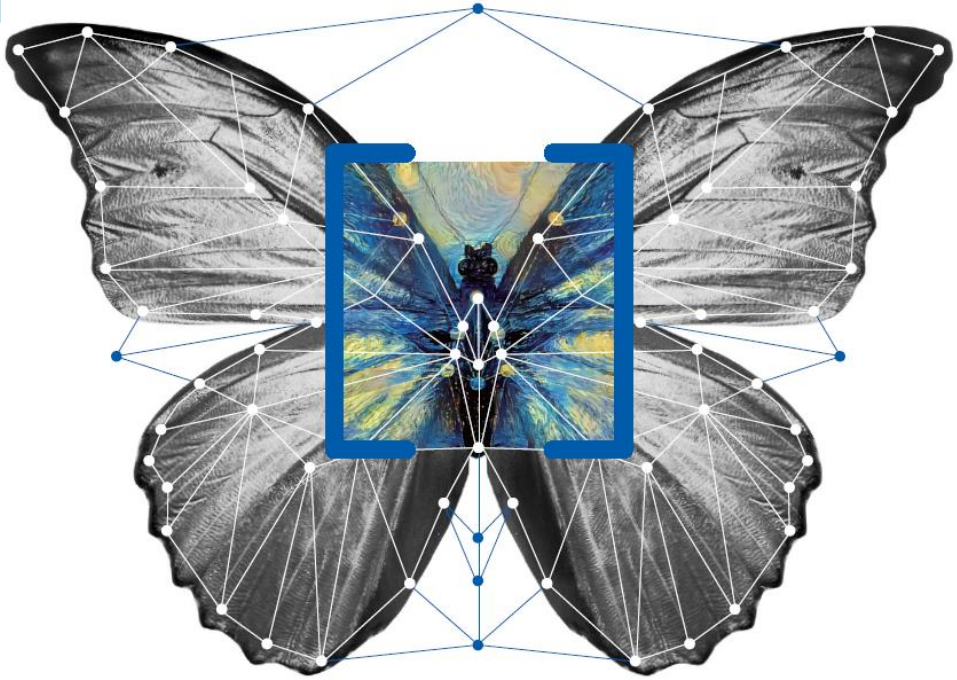


10_ SVM



MACHINE LEARNING 기계 학습

오일석 지음

11. 커널 기법

PREVIEW

11장은 커널 기법kernel method을 이용한 공간 변환

- 원래 특징 공간 $\mathcal{L}$을 새로운 특징 공간 $\mathcal{H}$로 변환하여 선형에 가까운 데이터 분포를 만듦
- $\mathcal{H}$는 매우 높은 차원이라 실제 변환은 불가능 → 커널 트릭을 사용하여 실제 변환하지 않고 변환 효과를 거둠

원래 특징 공간을 목적 달성에 더 유리한 새로운 공간으로 변환하는 작업

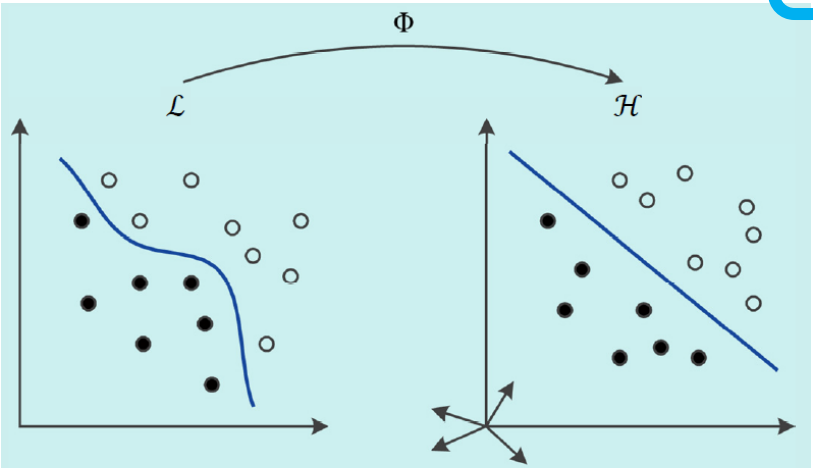


그림 11-1 원래 특징 공간 $\mathcal{L}$을 고차원 공간 $\mathcal{H}$로 변환하여 선형 분리 확보

커널 기법은 메모리 기반

- 학습이 끝난 후에 훈련집합 전체 또는 일부를 메모리에 저장하고 있다가 예측에 사용

공간 변환 과정

- 원래 특징 벡터 x(d 차원), 변환 후 벡터의 차원은 q (커널 기법에서는 보통 $q \gg d$)

$$\boldsymbol{\Phi}(\mathbf{x}) = \boldsymbol{\Phi}((x_1, x_2, \cdots, x_d)^{\mathrm{T}}) = \left(\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \cdots, \phi_q(\mathbf{x})\right)^{\mathrm{T}} \tag{11.1}$$

예제 11-1 XOR 분류를 위해 선형 분리 가능한 공간으로 변환

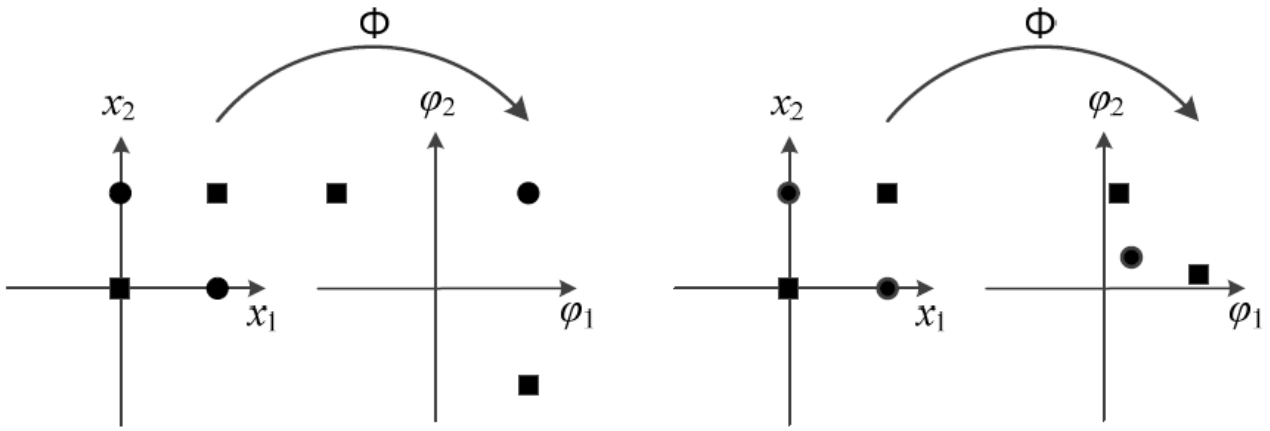
이 예제는 다음과 같은 훈련집합을 사용한다.

$$\mathbf{x}_1 = \begin{pmatrix}0\\0\end{pmatrix}, y_1 = -1, \mathbf{x}_2 = \begin{pmatrix}1\\0\end{pmatrix}, y_2 = 1, \mathbf{x}_3 = \begin{pmatrix}0\\1\end{pmatrix}, y_3 = 1, \mathbf{x}_4 = \begin{pmatrix}1\\1\end{pmatrix}, y_4 = -1$$

3장 [그림 3-9]의 퍼셉트론을 다시 살펴보자. 퍼셉트론이 수행하는 변환을 식 (11.1)에 맞춰 쓰면 식 (11.2)가 된다. 여기서 $\mathrm{sign}(z)$는 z가 양수면 1, 음수면 -1인 함수이다. [그림 11-2(a)]는 식 (11.2)의 변환을 보여 준다.

$$\boldsymbol{\Phi}(\mathbf{x}) = \left(\phi_1(\mathbf{x}), \phi_2(\mathbf{x})\right)^{\mathrm{T}} = (\mathrm{sign}(x_1 + x_2 - 0.5), \mathrm{sign}(-x_1 - x_2 + 1.5))^{\mathrm{T}} \tag{11.2}$$

11.1 커널 트릭



(a) 퍼셉트론을 이용한 공간 변환 (b) RBF를 이용한 공간 변환

그림 11-2 선형 분리가 불가능한 2차원 공간을 선형 분리가 가능한 2차원 공간으로 변환

[그림 11-2(b)]는 식 (11.3)을 이용한 또 다른 변환이다. 예를 들어, $\mathbf{x}_1 = \begin{pmatrix}0\\0\end{pmatrix}$은 $\begin{pmatrix}\exp\left(-\left\|\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}1\\1\end{pmatrix}\right\|_2^2\right)\\ \exp\left(-\left\|\begin{pmatrix}0\\0\end{pmatrix}-\begin{pmatrix}0\\0\end{pmatrix}\right\|_2^2\right)\end{pmatrix} = \begin{pmatrix}\exp(-2)\\ \exp(0)\end{pmatrix} = \begin{pmatrix}0.1353\\ 1.0\end{pmatrix}$으로 변환된다.

$$\mathbf{\Phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}))^T = \left(\exp\left(-\left\|\mathbf{x}-\begin{pmatrix}1\\1\end{pmatrix}\right\|_2^2\right),\ \exp\left(-\left\|\mathbf{x}-\begin{pmatrix}0\\0\end{pmatrix}\right\|_2^2\right)\right)^T \tag{11.3}$$

11.1 커널 트릭

[그림 11-2]는 2차원을 2차원으로 변환하는 예제이다. 이제 [그림 11-3]으로 관심을 옮겨 보자. 식 (11.4)는 2차원 공간을 3차원 공간으로 변환한다. 즉, d=20이고 q=30이다.

$$\mathbf{\Phi}(\mathbf{x}) = \left(\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \phi_3(\mathbf{x})\right)^{\mathrm{T}} = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)^{\mathrm{T}} \tag{11.4}$$

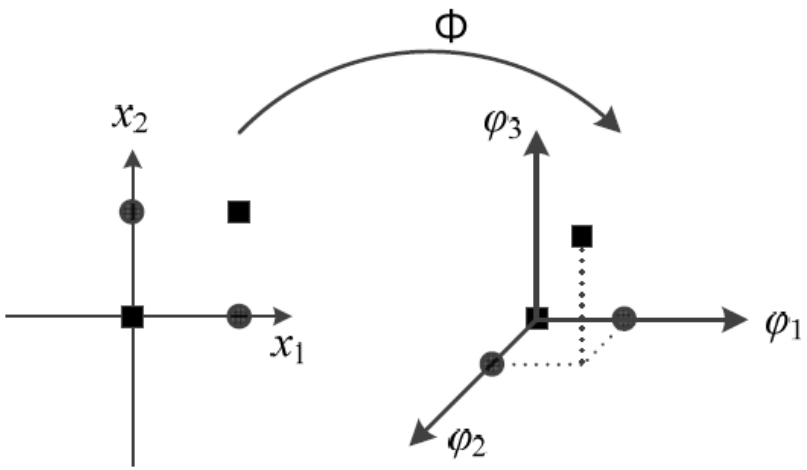


그림 11-3 선형 분리가 불가능한 2차원 공간을 선형 분리가 가능한 3차원 공간으로 변환

11.1 커널 트릭

커널함수와 커널 트릭

- 원래 특징 공간이 고차원인데 (예, MNIST는 784, ILSVRC 영상은 224*224=50176차원), 더 고차원으로 변환하는 일은 거의 불가능 → 커널함수를 이용하여 극복

정의 11-1 커널함수

원래 특징 공간 $\mathcal{L}$에 정의된 두 특징 벡터 $\mathbf{x}$와 $\mathbf{z}$에 대해 $K(\mathbf{x},\mathbf{z}) = \boldsymbol{\Phi}(\mathbf{x}) \cdot \boldsymbol{\Phi}(\mathbf{z})$인 변환함수 $\boldsymbol{\Phi}$가 존재하면 $K(\mathbf{x},\mathbf{z})$를 커널함수라 부른다.

널리 쓰이는 커널함수

- 1~2개의 하이퍼 매개변수를 가짐 (다항식의 p, RBF의 σ, 하이퍼볼릭 탄젠트의 α, β)

다항식 커널: $K(\mathbf{x},\mathbf{z}) = (\mathbf{x}\cdot\mathbf{z}+1)^p$ (11.6)

RBF 커널: $K(\mathbf{x},\mathbf{z}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}\right)$ (11.7)

하이퍼볼릭 탄젠트 커널: $K(\mathbf{x},\mathbf{z}) = \tanh(\alpha\mathbf{x}\cdot\mathbf{z}+\beta)$ (11.8)

11.4 SVM 분류

- **SVM(support vector machine)은 가장 성공적인 커널 기법**
  - 1990년대 신경망의 성능을 능가하여 인기 있는 모델로 부상(2000년대 들어 딥러닝에 밀려 시들한 편)
  - 최근 SVM과 딥러닝을 결합하는 접근방법이 시도되고 있음
- **SVM의 핵심 아이디어는 여백을 이용한 일반화 능력 향상**
  - 신경망은 ①에서 출발하여 ②에 도달하면 거기서 멈춤
  - SVM은 ② 대신 ③을 찾음 (③은 양쪽 부류에 대해 여백이 커서 일반화 능력이 우월)

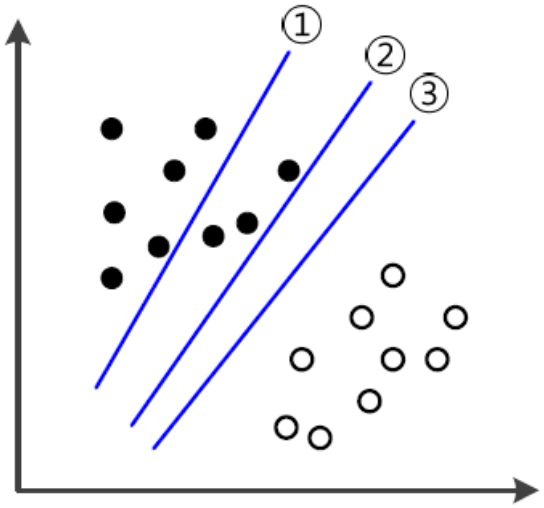


그림 11-6 최대 여백에 의한 일반화 능력 향상

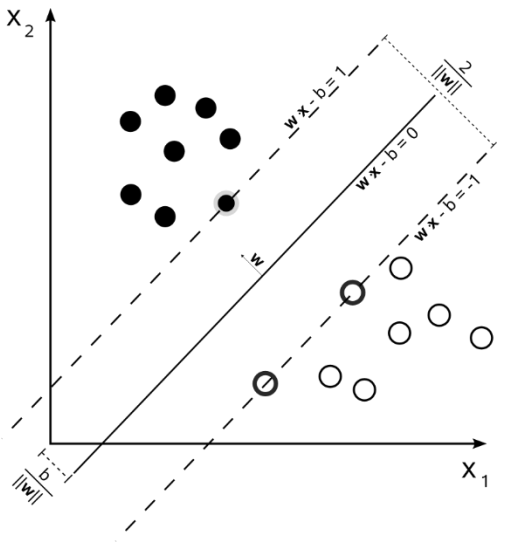
실제 데이터들이 훈련데이터에 없는 것들도 있을 수 있고 이에 대응가능하기 위해서는 최대한 최적의 경계를 찾는 것이 유리

11.4.1 선형 SVM

우선 2부류 선형 분리 가능한 상황으로 국한

- 이진 선형 분류기의 결정 초평면을 식 (11.23)으로 표현

$$d(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$



결정 초평면의 수학적 특성

- $d(\mathbf{x}) = 0$은 특징 공간을 부분 공간 2개로 분할하며, 한쪽 영역에 속한 점 $\mathbf{x}$는 $d(\mathbf{x}) > 0$이고, 다른 쪽 점 $\mathbf{x}$는 $d(\mathbf{x}) < 0$이다.
- 식 (11.23)에 상수 c를 곱한 $c\mathbf{w}^{\mathrm{T}}\mathbf{x} + cb = 0$도 동일한 초평면이다.
- $\mathbf{w}$는 초평면에 수직인 법선 벡터normal vector이고, 바이어스 b는 초평면의 위치를 지정한다.
- 점 $\mathbf{x}$에서 초평면까지 거리는 식 (11.24)로 구한다.

$$h = \frac{|d(\mathbf{x})|}{\|\mathbf{w}\|_2} \qquad (11.24)$$

11.4.1 선형 SVM

예제 11-3 결정 초평면의 수학적 특성

[그림 11-7]의 파란색 실선은 2차원의 결정 초평면인데, 2차원에서는 초평면이 직선이므로 결정 직선이라고 하자. $\mathbf{w} = (2,1)^T$이고 b=-4이다. 즉, $d(\mathbf{x}) = 2x_1 + x_2 - 4 = 0$이다.

$y = -2x + 4$
$2x + y - 4 = 0$

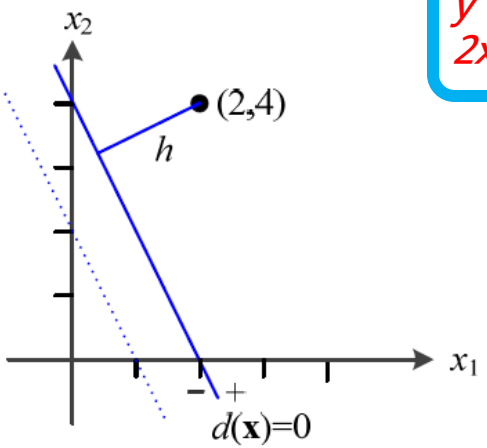


그림 11-7 결정 직선의 수학 특성

$(x0, y0)$와 직선 $ax+by+c=0$ 사이의 거리 d

$$d = \frac{ax_0+by_0+c}{\sqrt{a^2+b^2}}$$

+로 표시된 영역의 모든 점은 $d(\mathbf{x}) > 0$이고, −로 표시된 영역의 모든 점은 $d(\mathbf{x}) < 0$이다. 예를 들어, 원점 (0,0)을 입력하면 $d((0,0)^T) = -4 < 0$이다. $d(\mathbf{x})$에 2를 곱하면 $d(\mathbf{x}) = 4x_1 + 2x_2 - 8 = 0$이 되는데, 이 식도 [그림 11-7]의 파란 실선을 나타낸다. $d(\mathbf{x}) = 2x_1 + x_2 - 4$에서 b를 -4에서 -2로 바꾸면, 파란 점선이 된다. 즉, $\mathbf{w}$를 그대로 둔 채 b를 바꾸면 결정 직선은 방향을 유지한 채 위치만 바뀐다. 점 $(2,4)^T$에서 결정 직선까지 거리 h를 계산하면, $h = \frac{|2*2+4-4|}{\|(2,1)^T\|_2} = \frac{4}{\sqrt{5}} = 1.78885$이다.

11.4.1 선형 SVM

선형 분리 가능한 상황에서는

- 직선의 방향 w를 정하면 바이어스 b는 자동으로 정해짐(결정 직선이 두 부류를 완벽히 분류하면서 분할 띠의 중앙에 위치하도록 정함. 결정 직선 ①과 ②는 예제임)
- 분할 띠의 경계에 걸쳐있는 샘플이 서포트 벡터
- 분할 띠의 너비 $2s$를 여백margin이라 부름

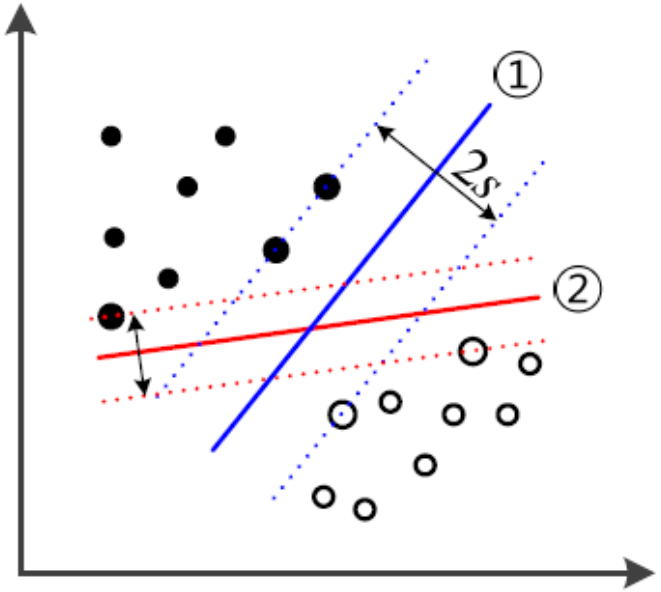


그림 11-8 서포트 벡터와 여백

11.4.1 선형 SVM

SVM 학습이 풀어야 하는 문제

문제 11-1: 여백이 가장 큰 결정 초평면의 방향 **w**를 찾아라.

- 방향 w를 찾으면 바이어스 b는 간단한 식으로 계산 가능
- [그림 11-8]의 경우 ①이 ②보다 우월
- ①이 최적일까? → 수학으로 최적을 찾아야 함

여백을 공식화

$$\text{여백} = 2s = \frac{2|d(\mathbf{x})|}{\|\mathbf{w}\|_2} = \frac{2}{\|\mathbf{w}\|_2} \qquad (11.25)$$

- 이 식에서 x는 서포트 벡터(즉 분할 띠의 경계에 있는 샘플)

11.4.1 선형 SVM

- **여백을 위한 식 (11.25)를 이용하여 문제를 구체화하면,**

문제 11-2: 아래 조건에서 $J(\mathbf{w}) = \frac{2}{\|\mathbf{w}\|_2}$를 최대화하라.

$$\left.\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \geq 1, \forall y_i = 1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \leq -1, \forall y_i = -1 \end{aligned}\right\}$$

  - 등호가 성립하는 샘플이 서포트 벡터

- **최소화 문제로 바꿔 쓰면,**

문제 11-3: 아래 조건에서 $J(\mathbf{w}) = \frac{\|\mathbf{w}\|_2^2}{2}$을 최소화하라.

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

  - 조건부 최적화 문제임 → 라그랑주 승수Lagrange multiplier로 해결

11.4.1 선형 SVM

[문제 11-3]에 해당하는 라그랑주함수

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\|\mathbf{w}\|_2^2}{2} - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1\right) \tag{11.26}$$

- **식 (11.26)의 KKT 조건**

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i \, y_i \mathbf{x}_i \tag{11.27}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0 \tag{11.28}$$

$$\alpha_i \geq 0, \qquad i = 1,2,\cdots,n \tag{11.29}$$

$$\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1) = 0, \qquad i = 1,2,\cdots,n \tag{11.30}$$

조건부 최적화 문제는 라그랑주함수에 KKT 조건을 적용하여 푼다. $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{n} \alpha_i f_i(\boldsymbol{\theta})$의 KKT 조건은 (1) $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \mathbf{0}$, (2) $0 \leq \alpha_i$, $i = 1, \cdots, n$, (3) $\alpha_i f_i(\boldsymbol{\theta}) = 0, i = 1, \cdots, n$이다. 식 (11.26)에서는 $\boldsymbol{\theta} = \{\mathbf{w}, b\}$이다.

11.4.1 선형 SVM

- 선형 분리 불가능한 상황 ([그림 11-10]의 예시)
  - 분할 띠 안에도 샘플을 허용 ← 소프트 여백soft margin 아이디어

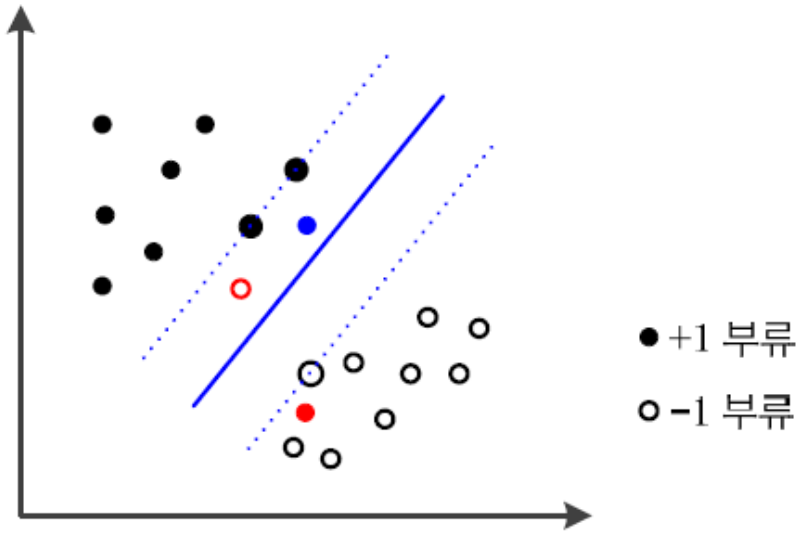


그림 11-10 선형 분리가 불가능한 상황에서 SVM

- 샘플은 세 가지 경우 중 하나

- 경우 1: 분할 띠 바깥에 있다. $1 \leq y(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b)$이다. 검정 샘플이 해당한다.
- 경우 2: 분할 띠 안에 있는데 자신이 속한 부류의 영역에 있다. $0 \leq y(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b) < 1$이다. 파란 샘플이 해당한다.
- 경우 3: 결정 경계를 넘어 다른 부류의 영역에 있다. $y(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b) < 0$이다. 빨간 샘플이 해당한다.

11.4.1 선형 SVM

- **슬랙 변수 ξ를 도입하여 세 가지 경우를 하나의 식으로 씀**

$$1 - \xi \leq y(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b) \qquad (11.31)$$

크시(크사이)

  - **경우 1은** $\xi = 0$, **경우 2는** $0 < \xi \leq 1$, **경우 3은** $1 < \xi$

- **선형 분리가 불가능한 상황의 SVM 문제**

문제 11-6: 여백을 될 수 있는 한 크게 하면서(목적 1), $0 < \xi$인 (즉, 경우 2와 경우 3에 속하는) 샘플의 수를 될 수 있는 한 적게 하는(목적 2) 결정 초평면의 방향 $\mathbf{w}$를 찾아라.

가장 현실적인 문제

11.4.1 선형 SVM

목적함수의 정의

$$J(\mathbf{w}, \boldsymbol{\xi}) = \underbrace{\frac{1}{2}\|\mathbf{w}\|_2^2}_{\text{목적 1}} + C \underbrace{\sum_{i=1}^{n} \xi_i}_{\text{목적 2}} \qquad (11.32)$$

- C는 어느 항에 비중을 더 둘 지를 결정하는 하이퍼 매개변수

목적 1과 2는 Trade-off 관계

C=0 이면 목적2는 무시
C가 클수록 목적2에 촛점

11.4.1 선형 SVM

예제 11–5 선형 분리가 불가능한 경우의 [문제 11-9] 풀이

[그림 11-9(a)]에서 $\mathbf{x}_3$의 소속을 1로 바꾸어 [그림 11-11]의 상황을 만들자.

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

훈련집합을 [문제 11-9]에 대입하고 정리하면 다음과 같다.

$$\left.\begin{array}{l}\alpha_1 - \alpha_2 + \alpha_3 = 0\\ 0 \le \alpha_1 \le C, 0 \le \alpha_2 \le C, 0 \le \alpha_3 \le C\end{array}\right\}\text{라는 조건에서}$$

$\tilde{\mathcal{L}}(\boldsymbol{\alpha}) = (\alpha_1 + \alpha_2 + \alpha_3) - \frac{1}{2}(13\alpha_1^2 + 17\alpha_2^2 + 26\alpha_3^2 - 22\alpha_1\alpha_2 + 26\alpha_1\alpha_3 - 42\alpha_2\alpha_3)$ 을 최대화하는 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$를 찾아라.

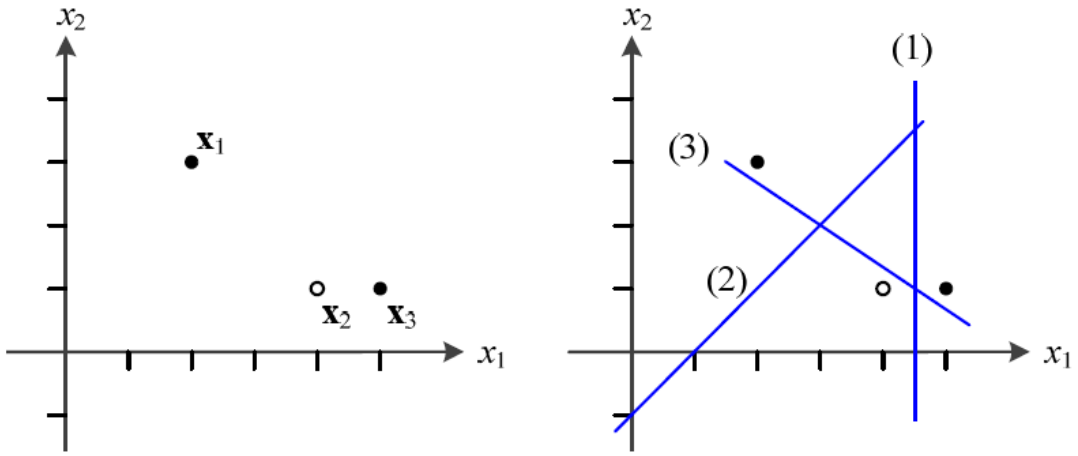


그림 11–11 문제 11–9의 풀이 예제

11.4.1 선형 SVM

- $C < 2$: 결정 직선 (2)만 유효하다. 이처럼 C를 작게 하면 식 (11.32)에서 목적 1, 즉 여백의 크기에 비중을 두겠다는 의도이다. 여백이 큰 (2)를 선택하였으니 의도대로 된 셈이다. $\mathbf{x}_3$의 오분류를 허용한다.
- $2 \leq C < 6.5$: 결정 직선 (1)과 (2)가 유효한데, (2)의 여백이 더 크므로 결정 직선 (2)를 선택한다. $\mathbf{x}_3$의 오분류를 허용한다.
- $6.5 \leq C$: 이처럼 C를 크게 하면 식 (11.32)에서 목적 2, 즉 분할 띠 바깥에 있어야 하는 규칙을 어기는 샘플 수를 줄이겠다는 의도이다.

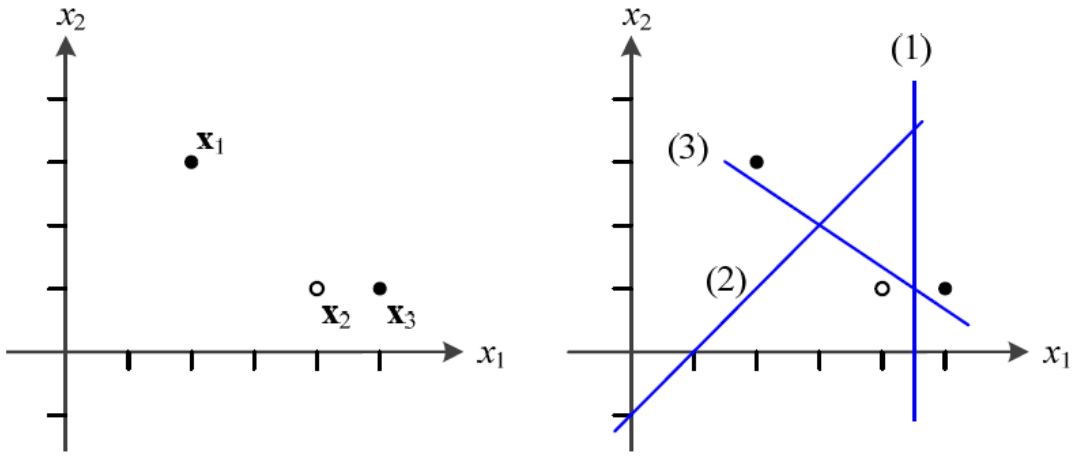


그림 11-11 문제 11-9의 풀이 예제