

05. SVM

통계적 분류방법 → 신경망 → 트리 분류기 → 마코프 모델

최근에 SVM이 등장하여 큰 주목을 받고 있음

아유는, 기존 방법들은 '오류율을 최소화' 하려는 목적으로 설계되었으나,

SVM은 한 방자 더 나아가 두 분류 사이에 존재하는

'여백(Margin)을 최대화' 하여 일반화 능력을 극대화 하므로.

5.1 분류

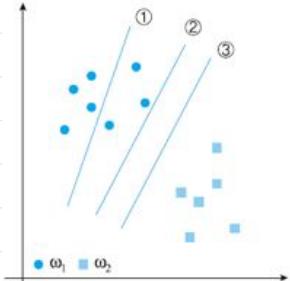


그림 5.1 분류기의 일반화 능력

분류기의 일반화 능력 (generalization)

- ②보다 ③이 여백이 더 크다.
- 즉, ③이 ②보다 일반화 능력이 뛰어나다

Margin(여백)이라는 개념을 어떻게
공식화 할 것인가? 여백을 최대로 하는
결정 초평면을 어떻게 찾을 것인가?

5.2 선형 SVM

이전 분류를 위한 결정 초평면과 그것의 수학적 특성

$$d(x) = w^T x + b = 0$$

(x 는 샘플을 나타내는 특징 벡터로서 $x = (x_1, \dots, x_d)^T$ 이다.)

- $d(x)$ 는 전체 특징 공간을 두 영역으로 분할하여 한쪽 영역에 속하는 점 x 는 $d(x) > 0$ 이고, 다른 쪽에 있는 점은 $d(x) < 0$ 이다.
- 하나의 초평면을 표현하는 식은 여럿 있다. 0이 아닌 임의의 상수 c 를 곱하여도 같은 초평면을 나타낸다.
- w 는 초평면의 법선 벡터 (normal vector) 초평면의 방향을 나타내고 b 는 위치를 나타낸다.
- 임의의 점 x 에서 초평면까지의 거리는 (5.2)와 같다.

$$h = \frac{|d(x)|}{\|w\|} \leftarrow \text{방향}$$

예제 5.1 결정 초평면의 수학적 특성

그림 5.2에 있는 결정 직선의 수학적 특성을 살펴보자. 이 직선의 마개변수는 $w=(2, 1)^T$ 이고, $b=-4$ 이다.

아래는 모두 같은 직선

$$d(x) = 2x_1 + x_2 - 4 = 0$$

$$d(x) = x_1 + 0.5x_2 - 2 = 0$$

$$d(x) = 6x_1 + 3x_2 - 12 = 0$$

점 $x=(2, 4)^T$ 에서 직선까지 거리

$$h = \frac{|2 \cdot 2 + 1 \cdot 4 - 4|}{\sqrt{2^2 + 1^2}} = \frac{4}{\sqrt{5}} = 1.78885$$

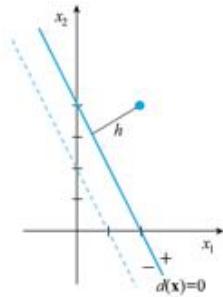


그림 5.2 직선의 수학적 특성

5.2-1 선형 분리 가능한 상황

- w (직선의 방향) 가 주어진 상황에서, '두 부류에 대해 직선으로부터 가장 가까운 샘플까지의 거리가 같게 되는 b 를 결정 (①과 ②는 그렇게 얻은 직선)

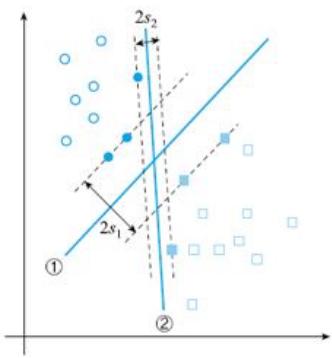


그림 5.3 선형 분리 가능한 상황

- 여백(margin)은 직선으로부터 가장 가까운 샘플까지의 거리의 두 배로 정의
- 점선으로 표시한 평행한 두 직선 사이의 영역을 분할 띠(separation band)
- 가장 가까운 샘플을 Support Vector라고 부름

SVM 목적: 여백을 가장 크게 하는 결정 초평면의 방향, 즉 w 를 찾는다.

$$\text{여백} = 2h = \frac{2|dc(y)|}{\|w\|} = \frac{2}{\|w\|}$$

\uparrow

w 와 b 에 적절한 상수 c 를 곱해도 같은 초평면을 나타내므로 $|dc(y)| = 1$ 이 되도록 적당한 c 를 곱함.

훈련집합을 $X = \{(x_1, t_1), (x_2, t_2), \dots, (x_N, t_N)\}$

N 은 훈련집합에 있는 샘플의 개수,
 t_i 는 부류를 표시, x_i 가 w 에 속하면 $t_i = 1$,
 x_i 가 w 에 속하면 $t_i = -1$

“모든 샘플을 옳게 분류한다는 조건 하에” 최대 여백을 갖는
정정 초평면을 찾으면 된다.

• 조건부 최적화 문제 (constrained optimization problem)

아래 조건 하에,

$$\begin{aligned} w^T x_i + b \geq 1, \quad & \forall x_i \in W_1 \\ w^T x_i + b \leq -1, \quad & \forall x_i \in W_2 \end{aligned}$$

$\frac{2}{\|w\|}$ 를 최대화 하라.



$$t_i(x^T x_i + b) - 1 \geq 0, \quad i = 1, \dots, N$$

$J(w) = \frac{1}{2} \|w\|^2$ 을 최소화 하라

① 해의 유일성

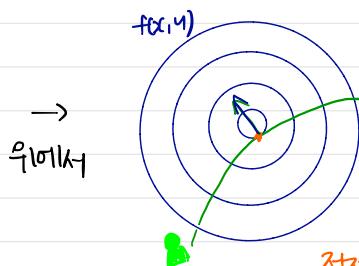
목적함수 $J(w)$ 는 w 의 2차항 만을 가지므로 볼록(convex) 함수이다. 또한 N 개의 조건식이 모두 선형이고 이들 모두를 만족하는 영역은 볼록이 된다. 구한 해는 반드시 전역 최적 점. 신경망과 달리 SVM은 지역 최적점에 빠지는 일이 결코 없다.

② 낸이도는?

N 개의 선형 부등식을 조건으로 가진 2차 항수의 최적화 문제
이므로 라그랑제 승수(Lagrange multiplier)를 도입한다.

라그랑제 승수란?

어떤 조건이 주어지고, 그 조건을 만족하는 상황에서 최고 또는 최저 값을 갖는 지점을 찾는 조건부 최적화 문제에서 사용.



이 경로를 따라 갔을 때
최고 높이는 얼마일까?

$$f(x,y) = 4 - x^2 - 2y^2$$

$$g(x,y) = 2(x-1)^2 - 10y + 30$$

점점 (a,b) 를 찾아 헤
대입하면 높이가 나온다. 이 점을
찾는 과정이 라그랑제 승수법

두 곡선의 공통접선에 수직인 벡터, 즉 기울기 벡터 (gradient vector)의 방향이 같으면 된다.

함수값의 변화량이 가장 큰 방향을
나타내는 벡터. 가장 가파른
방향? 등고선에 수직이므로...

$$g(a,b) = 0$$

$$\nabla f(a,b) = \lambda \nabla g(a,b)$$

↳ 두 기울기 벡터의 방향은 같아도, 그 크기는 다를 수
있으므로 그 차이를 맞춰주는 라그랑제 승수 필요.

기울기 벡터의 정의

$$\nabla \phi (\text{grad } \phi) = \left(\frac{d\phi}{dx}, \frac{d\phi}{dy}, \frac{d\phi}{dz} \right)$$

$$\phi = f(x, y, z) = c$$

$$d\phi = \frac{d\phi}{dx} dx + \frac{d\phi}{dy} dy + \frac{d\phi}{dz} dz = 0$$

$$\left(\frac{d\phi}{dx}, \frac{d\phi}{dy}, \frac{d\phi}{dz} \right) \cdot (dx, dy, dz) = 0 \Rightarrow \text{수직이라는 뜻}$$

기울기 벡터

↳ 미분한 방향, 즉 접선

다시 조건부 최적화 문제로 돌아가보자.

라그랑제 함수 $L(w, b, \alpha)$ 을 아래와 같이 정의할 수 있다.

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i (t_i (w^T x_i + b) - 1)$$

부등식 조건부 최적화 문제는 보통 Karush-Kuhn-Tucker KKT 조건도 만족시켜야 한다.

$$\frac{\partial L(w, b, \alpha)}{\partial w} = 0 \rightarrow w = \sum_{i=1}^N \alpha_i t_i x_i \quad (5.8)$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0 \rightarrow \sum_{i=1}^N \alpha_i t_i = 0 \quad (5.9)$$

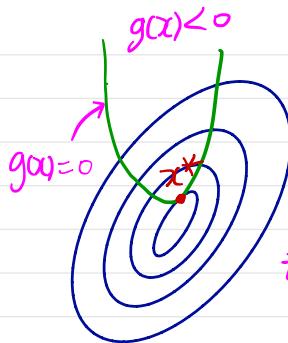
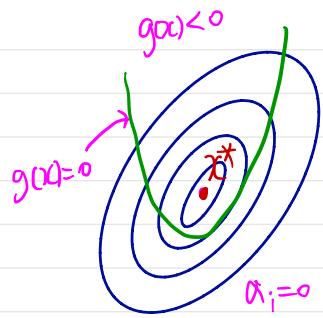
$$\alpha_i \geq 0, i=1, \dots, N \quad (5.10)$$

$$\alpha_i (t_i (w^T x_i + b) - 1) = 0, i=1, \dots, N \quad (5.11)$$

1. (5.8)로 결정 초평면의 매개변수 w 를 구할 수 있다.

t_i 와 x_i 는 훈련집합으로 주어진 것이고, 라그랑제 승수를 알고 있다는 말은 결정 초평면을 구했다는 말과 같다.

(5.4.2절에서 구체적으로 다시 언급)



2. 모든 샘플에 대해 $\alpha_i = 0$ 이거나 $t_i(w^T x_i + b) - 1 = 0$ 이어야 한다. (5.11) $\alpha_i \neq 0$ 인 샘플은 $t_i(w^T x_i + b) = 1$ 이어야 하는데, 이들이 바로 서포트 벡터이다. 그림 5.3에서 속이 찬 샘플이 이들에 해당하고 $\alpha_i = 0$ 인 샘플은 속이 빈 샘플들이다.

3. (5.11)로 결정 초평면의 매개변수 b 를 구할 수 있다. $\alpha_i \neq 0$ 인 샘플에 대해 $t_i(w^T x_i + b) = 1$ 을 풀면 된다. (5.8)로 w 를 구한 뒤에는 쉽게 b 를 구할 수 있다. (5.4.2절에서 자세히)

볼록성질을 만족하는 조건부 최적화 문제는 Wolfe dual 문제로 변형할 수 있다.

$$w = \sum_{i=1}^N \alpha_i t_i x_i, \quad \sum_{i=1}^N \alpha_i t_i = 0$$

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2} \left(\sum_{i=1}^N \alpha_i t_i x_i \right) \left(\sum_{j=1}^N \alpha_j t_j x_j \right) - \sum_{i=1}^N \alpha_i t_i x_i \cdot (\leq \alpha_j t_j x_j) \\ &\quad - \cancel{\sum_{i=1}^N \alpha_i b} + \sum \alpha_i \\ &= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j x_i^T x_j \end{aligned}$$

아래 조건 하에,

$$\sum_{i=1}^N \alpha_i t_i = 0$$

$$\alpha_i \geq 0, i=1, \dots, N$$

$$\tilde{L}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j x_i^T x_j$$

를 최대화하라

→ 2차 함수의 최대화 문제임

w 와 b 가 사라짐 (α 를 찾는 문제)

특징벡터 x_i 가 내적 형태로 비선형으로 학습하는 발판

목적 함수의 두번째 \leq 항은 N^2 개의 항을 갖는다.

(여전히 풀기 어려운 문제)

예제 5.2

훈련 집합은 아래와 같다.

$$x_1 = (2, 3)^T, t_1 = 1$$

$$x_2 = (4, 1)^T, t_2 = -1$$

\downarrow

$$\alpha_1 t_1 + \alpha_2 t_2 = 0$$

$$\alpha_1 \geq 0, \alpha_2 \geq 0$$

$$\begin{aligned} \tilde{L}(\alpha) &= (\alpha_1 + \alpha_2) - \frac{1}{2} (\alpha_1 \alpha_1 t_1 t_1 x_1^T x_1 + \alpha_1 \alpha_2 t_1 t_2 x_1^T x_2 \\ &\quad + \alpha_2 \alpha_1 t_2 t_1 x_2^T x_1 + \alpha_2 \alpha_2 t_2 t_2 x_2^T x_2) \end{aligned} \text{을 최대화하라.}$$

$$\alpha_1 - \alpha_2 = 0$$

$$\alpha_1 \geq 0, \alpha_2 \geq 0$$

$$\tilde{L}(\alpha) = (\alpha_1 + \alpha_2) - \frac{1}{2} (13\alpha_1^2 + 17\alpha_2^2 - 22\alpha_1 \alpha_2)$$
 를 최대화하는

$$\alpha = (\alpha_1, \alpha_2)^T$$
 를 찾아라.

$$\alpha_1 = \alpha_2 이므로,$$

$$\tilde{L}(\alpha) = -4\alpha_1^2 + 2\alpha_1 = -4(\alpha_1 - \frac{1}{4})^2 - \frac{1}{16}$$

$$\therefore (\frac{1}{4}, \frac{1}{4})^T$$
에서 최대값

이 해를 가지고 결정 초평면 $d(x)$ 을 위한 w 와 b 를 구해보자.

$$w = \sum_{i=1}^3 \alpha_i t_i x_i = \frac{1}{4} (2, 3)^T - \frac{1}{4} (4, 1)^T = (-\frac{1}{2}, \frac{1}{2})^T$$

$$\alpha_i(t_i(w^T x_i + b) - 1) = 0$$

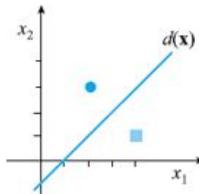
$$t_i(w^T x_i + b) = 1$$

$$\left(-\frac{1}{2}, \frac{1}{2}\right) \begin{pmatrix} 2 \\ 3 \end{pmatrix} + b = 1$$

$$-1 + \frac{3}{2} + b = 1$$

$$b = \frac{1}{2}$$

$$d(x) = -\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}$$



(b) SVM (속이 찬 샘플이 서포트 벡터)

예제 5.3

$$x_1 = (2, 3)^T, t_1 = 1$$

$$x_2 = (4, 1)^T, t_2 = -1$$

$$x_3 = (5, 1)^T, t_3 = -1$$

$$\alpha_1 - \alpha_2 - \alpha_3 = 0$$

$$\alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0$$

$$\tilde{L}(\alpha) = (\alpha_1 + \alpha_2 + \alpha_3) - \sum (\beta \alpha_1^2 + \gamma \alpha_2^2 + \delta \alpha_3^2 - 2\alpha_1 \beta - 2\alpha_2 \gamma - 2\alpha_3 \delta)$$

$\tilde{L}(\alpha)$ 을 최대화하는 $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ 를 찾으라.

훈련 샘플 세개 중에 최소한 두개는 서포트 벡터가 되어야 한다.

경우 ① $\alpha_1 = 0, \alpha_2 \neq 0, \alpha_3 \neq 0$

$$\alpha_2 = -\alpha_3$$

$\alpha_2 \geq 0, \alpha_3 \geq 0$ 인 조건에 맞는

경우 ② $\alpha_1 \neq 0, \alpha_2 = 0, \alpha_3 \neq 0$

$$\alpha_1 = \alpha_3$$

이것을 $\tilde{L}(\alpha)$ 에 대입하면 $\alpha_1 = 2/13$ 에서 최대값을 얻는다.

$$\alpha_1 = 2/13, \alpha_2 = 0, \alpha_3 = 2/13$$

$$\tilde{L}(\alpha) = 2\alpha_1 - \frac{13}{2}\alpha_1^2 = -\frac{13}{2}((\alpha_1 - \frac{2}{13})^2 - \frac{4}{169})$$

$$w = (-\frac{6}{13}, \frac{4}{13})^T, b = 1$$

$$d(x) = -\frac{6}{13}x_1 + \frac{4}{13}x_2 + 1 = -\frac{7}{13}$$

$$d(x_1) = \left(-\frac{6}{13}, \frac{4}{13}\right)(2, 3)^T + 1 = 1$$

$$d(x_2) = -\frac{7}{13}$$

$$d(x_3) = 1$$

$\alpha_2 = 0$ 이므로 $d(x_2) > 1$ 을 만족해야 하는데

그렇지 않으므로 맞는

경우 ③ $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 = 0$

$$\tilde{L}(\alpha) = 2\alpha_1 - 4\alpha_1^2 = -4\left(\alpha_1 - \frac{1}{4}\right)^2 - \frac{1}{16}$$

$$w = \left(-\frac{1}{2}, \frac{1}{2}\right)^T, b = \frac{1}{2}$$

$$d(x_1) = 1, d(x_2) = -1, d(x_3) = -\frac{3}{2}$$

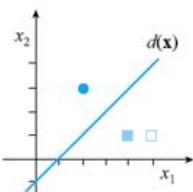
$t_3 d(x_3) > 1$ 만족

$$\therefore d(x) = -\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}$$

경우 ④ $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 \neq 0$

$$\alpha_1 = \alpha_2 + \alpha_3$$

$\tilde{L}(\alpha)$ 에 대입 후, $\partial \tilde{L}(\alpha) / \partial \alpha_2 = 0$ 과 $\partial \tilde{L}(\alpha) / \partial \alpha_3 = 0$ 을 풀면 $\alpha_2 = \frac{3}{2}, \alpha_3 = -1, \alpha_1$ 는 음수 이므로 이 경우는 버린다.



(b) SVM (속이 찬 샘플이 서포트 베터)

5.2.2 선형 분리 불가능한 상황

샘플 (x, t) 의 세 가지 상황

경우 1. 분할 띠의 바깥에 있다.

$t(w^T x + b) \leq 1$ 을 만족한다.

경우 2. 분할 띠의 한쪽에 있는데 자기가 속한 부류의 영역에 있다.

$0 \leq t(w^T x + b) < 1$ 을 만족한다.

경우 3. 결정 경계를 넘어 자신이 속하지 않은 부류의 영역에 놓여 있다. $t(w^T x + b) < 0$ 을 만족한다.

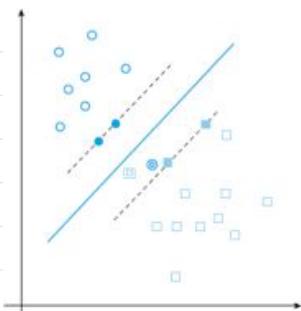


그림 5.6 선형 분리 불가능한 상황의 SVM

슬랙 변수 ξ 를 도입하여 하나의 식으로 쓰면,

$$t(w^T x + b) \geq 1 - \xi \quad (5.14)$$

표 5.2 선형 분리 불가능한 상황에서 샘플의 세 가지 경우

	샘플 위치	분류	$t(w^T x + b)$ 값	슬랙 변수	그림 5.6에서 기호
경우 1	분할 띠 바깥	옳게 분류	$1 \leq t(w^T x + b)$	$\xi = 0$	□○●■
경우 2	분할 띠 안쪽	옳게 분류	$0 \leq t(w^T x + b) < 1$	$0 < \xi \leq 1$	□
경우 3	결정 경계 넘음	틀리게 분류	$t(w^T x + b) < 0$	$1 < \xi$	◎

- 문제 공식화
- 길항(tradeoff) 관계를 갖는 두 가지 목적을 동시에 달성
 "여백을 될 수 있는 한 크게 하며(목적 1), 동시에 $0 < \xi_i$ 인
 (즉 경우 2 미지에 해당하는) 샘플의 수를 될 수 있는 한 적게)
 하는(목적 2) 결정 초평면의 방향 w 를 찾아라.

목적함수

$$J(w, \xi) = \underbrace{\frac{1}{2} \|w\|^2}_{\text{목적 1}} + c \underbrace{\sum_{i=1}^N \xi_i}_{\text{목적 2}}$$

이제 조건 하에,

$$t_i(w^T x_i + b) \geq 1 - \xi_i, \quad i=1, \dots, N$$

$$\xi \geq 0, \quad i=1, \dots, N$$

$$J(w, \xi) = \frac{1}{2} \|w\|^2 + c \sum_{i=1}^N \xi_i \text{ 를 최소화하라.}$$

$$L(w, b, \xi, \alpha, \beta) = \left(\frac{1}{2} \|w\|^2 + c \sum_{i=1}^N \xi_i \right)$$

$$- \left(\sum_{i=1}^N \alpha_i (t_i(w^T x_i + b) - 1 + \xi_i) + \sum_{i=1}^N \beta_i \xi_i \right)$$

kKT 조건에 의해

• 조건부 최적화 문제

아래 조건 하에,

$$w = \sum_{i=1}^N \alpha_i t_i x_i$$

$$\sum_{i=1}^N \alpha_i t_i = 0$$

$$C = \alpha_i + \beta_i$$

$$\alpha_i \geq 0, i=1, 2, \dots, N$$

$$\beta_i \geq 0, i=1, 2, \dots, N$$

$L(w, b, \xi, \alpha, \beta)$ 을 최대화하라.

$$\begin{array}{l} C = \alpha_i + \beta_i, \alpha_i \geq 0, \beta_i \geq 0 \text{ 이므로} \\ 0 \leq \alpha_i \leq C \end{array}$$

아래 조건 하에,

$$\sum_{i=1}^N \alpha_i t_i = 0$$

$$0 \leq \alpha_i \leq C, i=1, \dots, N$$

$\tilde{L}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j x_i^T x_j$ 를 최대화하라.

예제 5.4

$$x_1 = (2, 3)^T, t_1 = 1$$

$$x_2 = (4, 1)^T, t_2 = -1$$

$$x_3 = (5, 1)^T, t_3 = 1$$

$$\alpha_1 - \alpha_2 + \alpha_3 = 0$$

$$0 \leq \alpha_1 \leq C, 0 \leq \alpha_2 \leq C, 0 \leq \alpha_3 \leq C$$

$$\tilde{L}(\alpha) = (\alpha_1 + \alpha_2 + \alpha_3) - \frac{1}{2}(13\alpha_1^2 + 17\alpha_2^2 + 26\alpha_3^2 - 22\alpha_1\alpha_2 + 26\alpha_1\alpha_3 - 42\alpha_2\alpha_3)$$

을 최대화하는 $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ 를 찾아라.

경우 ①: $\alpha_1 = 0, \alpha_2 \neq 0, \alpha_3 \neq 0$

$$\alpha_2 = \alpha_3$$

$$\tilde{L}(\alpha) = 2\alpha_2 - \frac{1}{2}\alpha_2^2 = -\frac{1}{2}((\alpha_2 - 2)^2 - 4)$$

$$\rightarrow \alpha_2 = \alpha_3 = 2$$

$$w = (2, 0)^T, b = -9$$

$$d(x_1) = 2x_1 - 9 = 0$$

$$d(x_1) = (2, 0)(2, 3)^T - 9 = -5 \quad \text{오류}$$

$$d(x_2) = (2, 0)(4, 1)^T - 9 = -1 \quad \text{support vector}$$

$$d(x_3) = (2, 0)(5, 1)^T - 9 = 1$$

경우 ②: $\alpha_1 \neq 0, \alpha_2 = 0, \alpha_3 \neq 0$

$$\alpha_1 = -\alpha_3 ??$$

경우 ③ $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 = 0$

$\alpha_1 = \alpha_2$, $\tilde{L}(\alpha)$ 에 대입하면

$\alpha_1 = \frac{1}{4}$ 에서 최대값을 갖는다.

$$\alpha_1 = \frac{1}{4}, \alpha_2 = \frac{1}{4}, \alpha_3 = 0$$

$$W = \left(-\frac{1}{2}, \frac{1}{2} \right)^T, b = \frac{1}{2}$$

$$d(x) = -\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}$$

$$d(x_1) = 1$$

$$d(x_2) = -1$$

> support vector

$$d(x_3) = -\frac{3}{2}$$

- 오분류

경우 ④ $\alpha_1 \neq 0, \alpha_2 \neq 0, \alpha_3 \neq 0$

$\alpha_1 = \alpha_2 - \alpha_3$ $\tilde{L}(\alpha)$ 에 대입 후, $\partial \tilde{L}(\alpha) / \partial \alpha_2 = 0$ 과

$\partial \tilde{L}(\alpha) / \partial \alpha_3 = 0$ 을 풀면 $\alpha_2 = \frac{13}{2}, \alpha_3 = 5, \alpha_1 = \frac{3}{2}$

$$W = (2, 3)^T, b = -12$$

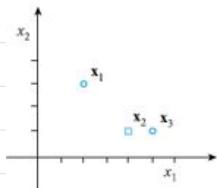
$$d(x) = 2x_1 + 3x_2 - 12$$

모두 뚫게 분류, 3개 모두 support vector

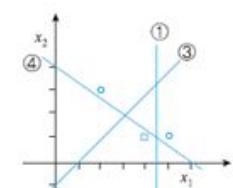
만일 라그랑제 승수의 상한 C 를 2보다 작게 주면 경우

①과 ④는 $0 \leq \alpha_i \leq C$ 의 조건을 만족시키지 못하므로

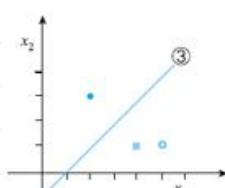
더 이상 유효하지 않다. 여백을 나타내겠다는 첫 번째 항을 더 중요하게 고려하겠다는 의도이다.



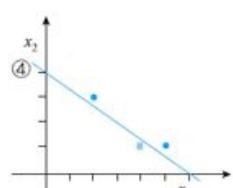
(a) 문제



(b) 경우 ①, ③, ④의 결정 직선



(c) $C < 6.5$ 일 때 SVM



(d) $C \geq 6.5$ 일 때 SVM

5.3 비선형 SVM

선형 SVM \rightarrow 비선형 SVM으로 확장하는데에 있어서의 핵심은

Kernel (특정 벡터가 내적 형태로만 나타나기 때문에 사용 가능)

5.3.1 커널 대치 (kernel Substitution)

더 높은 차원으로 매핑하여 선형 분리 불가능을 가능으로 만들 수 있다.

예제 5.5

원래 공간

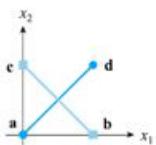
$$\begin{aligned} \mathbf{a} &= (0,0)^T, t_a = 1 \\ \mathbf{b} &= (1,0)^T, t_b = -1 \\ \mathbf{c} &= (0,1)^T, t_c = -1 \\ \mathbf{d} &= (1,1)^T, t_d = 1 \end{aligned}$$

매핑 함수

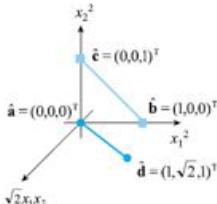
$$\Phi_1(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

매핑 결과

$$\begin{aligned} \mathbf{a} &= (0,0,0)^T \rightarrow \hat{\mathbf{a}} = (0,0,0)^T \\ \mathbf{b} &= (1,0)^T \rightarrow \hat{\mathbf{b}} = (1,0,0)^T \\ \mathbf{c} &= (0,1)^T \rightarrow \hat{\mathbf{c}} = (0,0,1)^T \\ \mathbf{d} &= (1,1)^T \rightarrow \hat{\mathbf{d}} = (1, \sqrt{2}, 1)^T \end{aligned}$$



(a) 원래 공간 L



(b) 매핑된 공간 H

그림 5.8 공간 매핑

$$\phi: L \rightarrow H$$

SVM이 사용하는 커널 함수의 성질: L 공간 상의 두 벡터 x 와 y 를 매개변수로 갖는 커널 함수를 $K(x, y)$ 라 하자. 그러면 $K(x, y) = \phi(x) \cdot \phi(y)$ 를 만족하는 매핑함수 $\phi(\cdot)$ 가 존재해야 한다. 즉, 커널 함수의 값과 H 공간 상으로 매핑된 두 점 $\phi(x)$ 와 $\phi(y)$ 의 내적이 같아야 한다.

예제 5.6 커널 함수의 성질

두 개의 벡터를 $x = (x_1, x_2)^T$, $y = (y_1, y_2)^T$ 라 하고,

커널함수 $K(x, y) = (x \cdot y)^2 = x_1^2 y_1^2 + 2x_1 y_1 x_2 y_2 + x_2^2 y_2^2$
이라 정의 하자.

$$\left. \begin{array}{l} K(\mathbf{a}, \mathbf{b}) = ((0, 0)^T \cdot (1, 0)^T)^2 = 0 \\ \Phi_1(\mathbf{a}) \cdot \Phi_1(\mathbf{b}) = ((0, 0, 0)^T \cdot (1, 0, 0)^T) = 0 \end{array} \right\} \rightarrow K(\mathbf{a}, \mathbf{b}) = \Phi_1(\mathbf{a}) \cdot \Phi_1(\mathbf{b})$$

$$\left. \begin{array}{l} K(\mathbf{c}, \mathbf{d}) = ((0, 1)^T \cdot (1, 1)^T)^2 = 1 \\ \Phi_1(\mathbf{c}) \cdot \Phi_1(\mathbf{d}) = ((0, 0, 1)^T \cdot (1, \sqrt{2}, 1)^T) = 1 \end{array} \right\} \rightarrow K(\mathbf{c}, \mathbf{d}) = \Phi_1(\mathbf{c}) \cdot \Phi_1(\mathbf{d})$$

증명. $K(x, y) = (x \cdot y)^2$

$$= x_1^2 y_1^2 + 2x_1 y_1 x_2 y_2 + x_2^2 y_2^2$$

$$= (x_1^2, \sqrt{2}x_1 x_2, x_2^2)^T \cdot (y_1^2, \sqrt{2}y_1 y_2, y_2^2)^T$$

$$= \phi_1(x) \cdot \phi_1(y)$$

이 외에도 이에 대응하는 매핑함수는 여럿 존재할 수 있다!

• kernel substitution (커널 대치 = 커널 트릭)

- 어떤 수식이 벡터 내적을 포함할 때, 그 내적을 커널 함수로 대체하여 계산하는 기법

- 실제 계산은 L 공간에서 $K(\cdot)$ 의 계산으로 이루어짐
- but, $\phi(\cdot)$ 로 매핑된 고차원 공간 H에서 작업하는 효과
- 적용 예. Fisher LD의 커널 LD로의 확장,

PCA를 커널 PCA로 확장

- SVM에 적용 가능 (벡터 내적만 나타나도록, (5.13), (5.21))

5. 3. 2 커널 대치를 이용한 비선형 SVM

목적함수를 다음과 같이 바꾸어 쓴다.

$$\tilde{L}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j \phi(x_i)^T \phi(x_j)$$

$$\tilde{L}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j k(x_i, x_j)$$

그렇다면, 어떤 커널을 사용할 것인가?

- SVM이 사용하는 대표적인 커널들

다항식 커널 $K(x, y) = (x \cdot y + 1)^p$ (5.32)

RBF (Radial Basis Function) 커널 $K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$ (5.33)

하이퍼볼릭 탄젠트 커널 $K(x, y) = \tanh(\alpha x \cdot y + \beta)$ (5.34)

커널함수에 대응하는 매핑함수는 몰라도 된다. SVM을 만들기 위해 풀어야 하는 식은 (5.30)이다.

→ 5.4.2절에서 자세히

5.4 구현

- SVM 학습이란?
 - (5.1)의 W 와 b 를 구하는 과정이다.
 - 라그랑제 승수 α 를 구하면 된다.
 - 비선형 SVM을 위한 조건부 최적화 문제(커널 토리 적용)
 - 조건부 최적화 문제 (비선형 SVM)
- 아래 조건 하에,
- $$\sum_{i=1}^N \alpha_i t_i = 0 \quad (5.35)$$
- $$0 \leq \alpha_i \leq C, i=1, \dots, N$$
- $$\tilde{L}(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j t_i t_j k(x_i, x_j) \text{를 최대화하라.}$$

5.4.1 학습

(5.35)를 어떻게 풀 것인가?

- 예를 들어, CEMPARMI

수자 DB의 경우 $N=4000$ 이다.

따라서, 목적함수는 8002000 개의

2차 항을 가진 매우 복잡한식

$(N^2+N)/2$ 개 항을 가짐

- N 개의 라그랑제 승수

활성화 집합 S 와 비활성화 집합 Z 로 나눈다. S 에 속한 라그랑제 승수는 고정된 상수값으로 간주하고 Z 에 대해 최적화를 수행한다. 반복~

알고리즘 [5.1] SVM 학습 (라그랑제 승수 구하기)

입력: 훈련 집합 $X = \{(x_1, t_1), (x_2, t_2), \dots, (x_N, t_N)\}$,

선형과 비선형의 선택.

비선형인 경우 커널 함수의 종류

출력: 라그랑제 승수 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$

알고리즘:

1. 라그랑제 승수 $\alpha_1, \alpha_2, \dots, \alpha_N$ 을 초기화 한다.
2. repeat {
 3. Y 에 속한 g 개의 라그랑제 승수를 선택한다. 나머지는 Z 에 속한다.
 4. $\alpha_i \in Z$ 는 상수, $\alpha_i \in Y$ 는 변수로 간주한다.
 5. if (선형 SVM) 문제 (5.27)를 분해한다.
 6. else 문제 (5.35)를 분해한다.
 7. 분해된 문제를 풀고 $\alpha_i \in Y$ 를 새로운 값으로 변경한다.
8. } until (최적화 조건이 만족).

- ① q 를 얼마로 할 것인지, ②를 어떻게 선택할지, ③ 분해된 문제를 푸는 방법을 고민해야하고, ④ 최적화 조건이 만족되었는지 여부를 판단하는 조건이 있어야 한다, 가장 널리 알려진 알고리즘은 SMO(sequential minimal optimization)으로, ④가 가질 수 있는 가장 작은 값 2를 사용한다.

5.4.2 인식

알고리즘 [5.1]을 수행하여 각각의 승수를 구했다고 하자.
미지의 패턴 x 인식은 어떻게 할 것인가?

(선형 SVM 분류기 :
 $d(x) > 0$ 이면 $x \in W_1$, $d(x) < 0$ 이면 $x \in W_2$ 로 분류하라.)
 이 때, $d(x) = w^T x + b$ (5.36)

• 각각제 승수 α 로 w 구하기

$$w = \sum_{i=1}^{N_s} \alpha_i t_i x_i = \sum_{k \in Y} \alpha_k t_k x_k \quad (5.37)$$

- $\alpha_i = 0$ 인 샘플은, $\alpha_i t_i x_i = 0$ 이 되므로 제거하여도 된다.
 x 는 $\alpha_i \neq 0$ 인 샘플의 합으로 이들이 바로 서포트 벡터이다. 분류기의 가중치 벡터 w 를 구하기 위해서는 서포트 벡터만 필요하다.

- 서포트 벡터는 $t_i(w^T x_i + b) - 1 = 0$ 이므로, 미지수가 b 뿐이나 쉽게 b 를 구할 수 있다. 서포트 벡터 중의 하나만 사용해도 b 를 구할 수 있다.
- 이렇게 w 와 b 를 미리 계산해 놓고, 미지의 패턴 x 가 들어오면 $\phi(x)$ 를 계산하여 그것의 부표를 보고 분류하면 된다.

• 비선형 SVM

\therefore 선형 SVM에서는 인식에 필요한 계산이 서포트 벡터의 수와 상관없이 일정하다.

$$d(x) = w^T \phi(x) + b$$

$$= \left(\sum_{x_k \in Y} \alpha_k t_k \phi(x_k) \right)^T \phi(x) + b$$

$$= \sum_{x_k \in Y} \alpha_k t_k \phi(x_k) \cdot \phi(x) + b$$

$$= \sum_{x_k \in Y} \alpha_k t_k k(x_k, x) + b$$

벡터 w 를 '미리 계산해
둘 수 없음'

새로운 테스트 샘플 x 를 인식하기 위해서 서포트 벡터 x_k 마다 x 와의 커널 핵수를 계산하고, 그 결과를 모두 더하는 계산이 필요.

비선형 SVM에서 인식 단계의 계산 시간은 서포트 벡터의 수에 비례.

5.4.3 M 부류 SVM

SVM 이진 분류기 \rightarrow M 부류 SVM으로 확장

① 1 대 M-1 방법 (실제 더 많이 사용)

M개의 이진 분류기 사용. j번째 이진 분류기는 부류 w_j 와, 나머지 M-1개 부류를 분류하는 역할을 한다.

결정 초평면 함수가 가장 큰 값을 갖는 부류로 분류한다.

M부류 SVM 분류기:

x 를 w_k 로 분류하라.

이 때 $k = \arg \max_j d_j(x)$

문제점: ① M개의 이진 분류기가 독립적으로 만들어 절로으로 $d_j(x)$ 가 갖는 값의 크기가 서로 다른 규모를 가질 수 있다.

② 이진 분류기의 훈련 집합이 불균형을 이룬다.

ex) 10부류라면 양의 영역 샘플 10%,

음의 영역 샘플 90% ↑

(모든 부류가 같은 개수의 훈련 샘플을 갖는다고 가정)

② 1 대 1 방법

부류 쌍을 분류하는 이진 분류기를 $M(M-1)/2$ 개 만든다.

부류 w_i 와 w_j 를 분류하는 이진 분류기의 결정 초평면을 $d_{ij}(x)$ 라 표기하자. 투표 개별을 도입하여 분류한다. $d_{ij}(x)$ 가 x 를 w_i 로

분류하면 뿐만 아니라 한 표를 얻고, 그 반대이면 ω_i 가 한 표를 얻는다. 이런 방식으로 $M(M-1)/2$ 개의 이진 분류기가 M 개 부류에 투표를 하고, 그 결과에 따라 가장 많은 표를 얻은 부류로 분류를 하면 된다. 이 방법은 1대 1-1 방법이 얻고 있는 문제를 가지지 않으나, 부류 수 M 이 커지면 이진 분류기 수가 많아져 학습과 인식에 시간이 많이 걸리는 단점을 알고 있다.

5.5 SVM의 특성

1. 사용자 설정 매개 변수가 적다. 어느 커널을 사용할지를 제외하면 (5.16)의 C 가 전부이다. SVM의 매개 변수 설정은 신경망에 비해 훨씬 쉽다.
2. 최적 커널을 자동 설정하는 방법이 없다. SVM의 커널은 일반화 능력을 최우하므로 신중하게 선택해야 한다. 실험에 의한 휴리스틱한 방법을 사용한다.
3. 일반화 능력이 뚜어나다.
4. 배선형 SVM에서는 인식 시간이 서포트 벡터의 개수에 비례한다. 따라서 많은 수의 서포트 벡터가 발생하는 경우 인식 시간이 길어질 수 있다.
5. (5.35)의 학습 알고리즘은 구현이 까다롭다. 다행스럽게도 많은 공개 소스 소프트웨어가 있다.