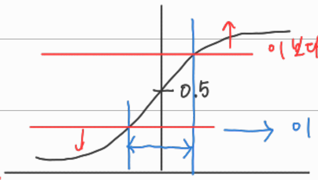
〈Ch. 3 : Support Vector Machines〉

↳ 이진 분류문제를 기하학적으로 해석하여 최적화시킨 알고리즘.

Logistic Regression

- logistic function



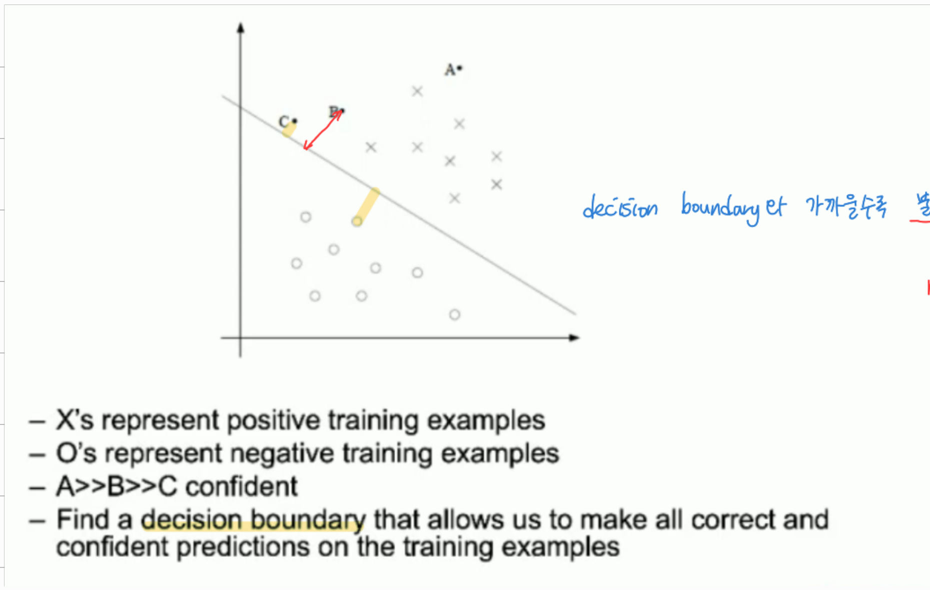
이보다 높으면 확실히 1.(positive label)

이보다 낮으면 확실히 0. (negative label)

→ 이 구간은 좀 애매함.

↳ 이 문제를 최적화를 통해 최대한 해결하는 linear classifier를 찾아보자.

Margins : Intuition



decision boundary와 가까울수록 불확정성이 높아진다.

↕

margin. (직선에서 벡터까지의 거리)

"그 거리 중 최소의 거리를 최대화하는 직선의 방정식을 찾는 이론."

⇓

Support Vector Machine.

(convex optimization 테크닉 사용)

목적 함수 뿐만 아니라 constraint 함수가 있는 ...

$$\theta^T x = \tilde{\theta}^T \tilde{x} + \tilde{\theta}_0 \Rightarrow w^T x + b = 0$$

θ_0 (상수항 추가).

$x_0 : 1$..

- 직선과의 거리 : $ax+by+c=0$ 직선과 (x_1, y_1) 의 거리.

$$\frac{|ax_1+by_1+c|}{\sqrt{a^2+b^2}} \Rightarrow \text{2차원에서.}$$

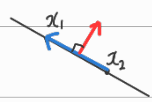
↓ n차원으로 확장하면. (유도는 다음 페이지에서)

$$\frac{|w^T x + b|}{\|w\|}$$

직선위의 두점 x_1, x_2에 대해

$w^T x_1 + b = 0$

$w^T x_2 + b = 0$) 두 식을 빼면 $w^T(x_1 - x_2) = 0$



w는 법선 벡터가 됨.

2025 / 1 / 8

Functional and geometric margins

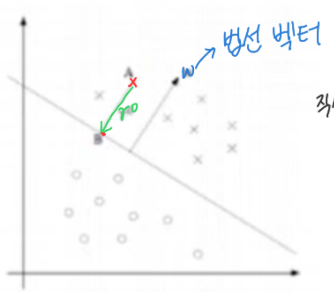
- Functional margin of (w, b)

- training set $S = \{(x^{(i)}, y^{(i)}) ;\ i = 1, \dots, m\}$

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

- (w, b)의 functional margin을 다음과 같이 정의.

$$\hat{\gamma} = \min_{i=1,\dots,m} \hat{\gamma}^{(i)}$$



γ

B is given by $x^{(i)} - \gamma^{(i)} \cdot w/||w||$.

직선 위에 존재하므로

$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{||w||}\right) + b = 0.$$

이 식을 $\gamma^{(i)}$에 대해

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}.$$

The optimal margin classifier

γ들 중 최솟값을 표현하는 방법.

$$\max_{\gamma,w,b} \ \gamma$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \dots, m$$

m개의 부등식을 모두 성립하는 γ

$$||w|| = 1.$$

→ normalized constraint는 다루기 힘들다.

- ||w||=1 is non convex

$$\max_{\hat{\gamma},w,b} \ \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \dots, m$$

Introduce scaling constraint to make $\hat{\gamma} = 1$.

$\hat{\gamma}/||w|| = 1/||w||$ is the same thing as minimizing $||w||^2$.

$$\min_{\gamma,w,b} \ \frac{1}{2}||w||^2$$

법선 벡터의 크기를 최소화한다? →

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m$$

Lagrange duality

- Primal optimization problem

convex optimization 이라는 이론에서 일반적으로 최적화하는 문제를 아래와 같이 표현 가능

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l. \end{aligned}$$

이 제약조건을 만족하는 w에 대하여 f(w)를 최소화하는 w를 찾는 것

어떠한 제약조건이 있는 것을 최대 최소하는 건 까다로운 일이다.

– To solve it, define generalized Lagrangian.

그 까다로운 문제를 해결하는 방법!

constrained → unconstrained

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta : \alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta).$$

$\alpha_i \ge 0$

If constraints are violated by w,

$$\begin{aligned} \theta_{\mathcal{P}}(w) &= \max_{\alpha, \beta : \alpha_i \ge 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) & (1) \\ &= \infty. & (2) \end{aligned}$$

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

– Consider minimization problem

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha, \beta : \alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta),$$

- It is the same problem with our original, primal problem

– Consider slightly different problem, we define

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta).$$

– Dual optimization problem

$$\max_{\alpha, \beta : \alpha_i \ge 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha, \beta : \alpha_i \ge 0} \min_w \mathcal{L}(w, \alpha, \beta).$$

$$d^* = \max_{\alpha, \beta : \alpha_i \ge 0} \min_w \mathcal{L}(w, \alpha, \beta) \le \min_w \max_{\alpha, \beta : \alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta) = p^*.$$

- In certain condition (KKT) $d^* = p^*$,

convex 문제이면서, KKT 조건을 만족시킬 경우.

original problem 이

– Karush-Kuhn-Tucker (KKT) conditions.

$$\begin{aligned} \frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, n & (3) \\ \frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, l & (4) \\ \alpha_i^* g_i(w^*) &= 0, \quad i = 1, \dots, k & (5) \\ g_i(w^*) &\le 0, \quad i = 1, \dots, k & (6) \\ \alpha^* &\ge 0, \quad i = 1, \dots, k & (7) \end{aligned}$$

→ KKT dual complementary condition이라 불림.

$\alpha_i^* > 0$ 라면, $g_i(w^*) = 0$ „

→ true only for support vectors.

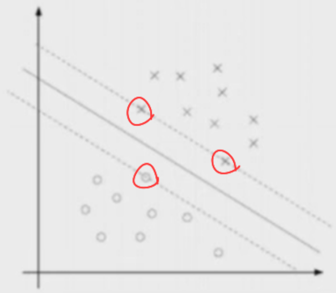
Optimal margin classifier

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

– Constraints as

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0.$$

= 를 만족하는 case



Three points are called
<support vectors>

– Construct the Lagrangian for our optimization problem

$$\mathcal{L}(w,b,\alpha) = \underbrace{\frac{1}{2}||w||^2}_{f(w)} - \sum_{i=1}^{m} \alpha_i \underbrace{\left[y^{(i)}(w^T x^{(i)} + b) - 1\right]}_{\alpha_i g_i(w)}. \qquad (8)$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}. \qquad (9)$$

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \qquad (10)$$

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j (x^{(i)})^T x^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)}.$$ by (9)

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j (x^{(i)})^T x^{(j)}.$$ by (10)

– Obtain the following dual optimization problem

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j \langle x^{(i)}, x^{(j)}\rangle.$$
$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

constrained 이기 때문에
단순 미분해서 0이 되는 해결 방법으로 못함.

– it is straightforward to find the optimal value for the intercept term b as

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}. \qquad (11)$$

$-1-b \;\; + \;\; 1-b \;\; = \;\; -2b$

Coordinate Descent
↳ 반복적으로 각 좌표축을 따라 움직이며
목적함수의 최솟값을 찾는 알고리즘

– Using (9)

m이 매우 클 경우... 연산량↑ → 안 좋은 알고리즘 아닌가? → 사실은 아니다. 좋은 알고리즘임!!

↳ 최적화 이론 중 (5) 조건을 보면, $g_i(w)$는 support vector일 경우만 0이다.
따라서 $\alpha_i^* g_i(w^*)$ 중 $g_i(w^*)$가 음수일 경우 무조건 α_i가 0이어야
(5) 조건을 만족하므로, 실제 (12) 계산은 support vector인 데이터만
연산하면 된다.

$$w^T x + b = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T x + b \qquad (12)$$
$$= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x\rangle + b. \qquad (13)$$

(문제점 ①)
decision boundary로 직선이 아닌 곡선이 더 적합하다면?
내적을 비선형함수로 바꿔서 훈련한다.
(kernel 테크닉)

CML

+) 직선을 훈련하는 것이므로 binary classification에만 사용.

2025 / 1 / 9

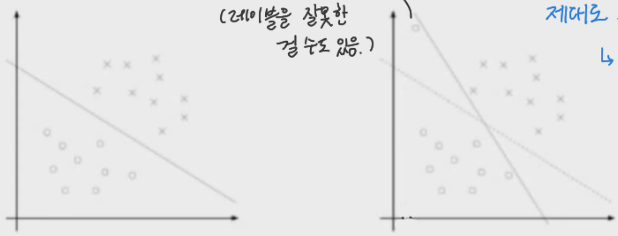
Regularization and the non-separable case

다른 문제점②: 데이터에 노이즈가 있는 경우.

이게 X일수도.
(레이블을 잘못한 걸 수도 있음.)

데이터 셋 자체가 제대로 분리돼있지 않은 것.
↳ entangled data

+) 이보다 더 얽혀있는 데이터를 그룹으로 나뉘게 하는 알고리즘
disentanglement
(생성형 AI에서 많이 쓰이는 테크닉)



- A single outlier is added in right figure

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m$$
$$\xi_i \geq 0, \quad i = 1, \ldots, m.$$

포기할 만큼 빼주기.

- Parameter C controls the relative weighting between the twin goals of making the $||w||^2$ small and of ensuring that most examples have functional margin at least 1

– We can form the Lagrangian:

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}(x^T w + b) - 1 + \xi_i\right] - \sum_{i=1}^{m} r_i \xi_i.$$

– Dual form

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m$$
$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0,$$

– KKT dual-complementarity conditions

$$\alpha_i = 0 \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) > 1 \qquad (14)$$ → Non-support vectors

α의 상한값. $$\alpha_i = C \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) < 1 \qquad (15)$$ → Relaxed cases

$$0 < \alpha_i < C \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) = 1. \qquad (16)$$ → Support vectors