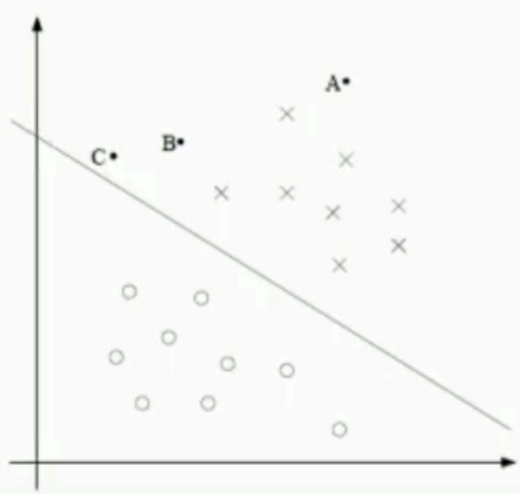
3강 Support vector machines

Margins: Intuition

- **Logistic Regression**
  - $p(y=1|x;\theta)$ is modeled by $h_\theta(x) = g(\theta^T x)$
  - We would then predict "1" on an input x if and only if $h_\theta(x) \geq 0.5$ or equivalently, if and only if $\theta^T x \geq 0$
  - y=1 if $\theta^T x \gg 0$ similarly y=0 if $\theta^T x \ll 0$
  - Given a training set, again informally it seems that we'd have found a good fit to the training data if we can find θ so that $\theta^T x^{(i)} \gg 0$ whenever $y^{(i)} = 1$, and $\theta^T x^{(i)} \ll 0$ whenever $y^{(i)} = 0$, since this would reflect a very confident set of classifications for all the training examples

Margins: Intuition



- X's represent positive training examples
- O's represent negative training examples
- A>>B>>C confident
- Find a decision boundary that allows us to make all correct and confident predictions on the training examples

결정 경계 = 2개의 class를 구분하는 직선의 방정식을 찾을 수 있다면, 클래스가 pos인지 neg인지 구분이 가능하다. 하지만 A,B,C의 경우 A와 같이 아주 멀리 떨어져있다면 확실히 pos class라고 주장할 수 있지만, 결정 경계에 가까이 위치한 데이터 포인트는 1) 노이즈에 의해 잘못 분류된 것인지, 2) 결정 경계가 잘못된 것인지 등 논란의 여지가 발생

-> 결정경계에 가까울수록 불확실성이 커진다.

결정 경계에 가까이 위치한 데이터 포인트는 높은 중요도를 가짐

불확정성에 반비례하는 개념이 Margin 이다.
그림에 직선으로부터 벡터까지의 거리를 Margin이라고 함

머신러닝의 기본 역할 = 결정 경계의 기울기와 절편을 구하는 것
각 데이터 포인트들의 중요도가 다르다

x가 1차원이면 직선의 방정식
2차원이면 평면의 방정식

일반적인 n차원의 평면 = hyperplane

Functional and geometric margins

- **Functional margin of (w,b)**

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

  - We can make the functional margin arbitrarily large without really changing anything meaningful
  - Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$
    - Define functional margin of (w,b)

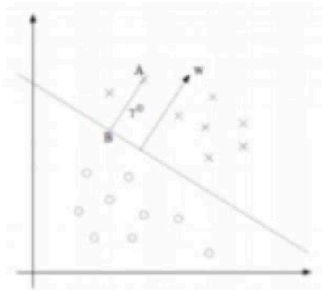
$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}.$$

모든 데이터 포인터들에 대해 가장 결정 경계에 가까운 값을 찾는다

그리고 마진을 최대화한다 = 가장 판단하기 어려운 데이터를 가장 좋게 판단하려는 노력(멀리 떨어져 있는 데이터 포인트는 쉽게 판단이 가능하지만, 결정 경계 근처의 데이터는 판단이 어렵기 때문에)

Functional and geometric margins

- **Functional margin of (w,b)**



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

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

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

– Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$
  - Define functional margin of (w,b)

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)}.$$

거리를 구하는 과정 유도

최종적인 목표 = 최솟값을 다시 최대화 시키는 것

The optimal margin classifier

- **Find the one that achieves maximum geometric margin**

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

– $||w||$=1 is non convex

$$\begin{aligned} \max_{\hat{\gamma},w,b} \quad & \frac{\hat{\gamma}}{||w||} \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m \end{aligned}$$

w, b를 적절히 scaling 해도 결정 경계는 변화가 없다.
역수를 minimize하는 것은 분모를 maximize하는 것과 같음

-> w를 minimize함, 백터의 크기 자체를 최소화하는 과정과 백터의 제곱의 크기를 최소화하는 과정은 동일하기 때문에 아래 식 나옴

The optimal margin classifier

- Find the one that achieves maximum geometric margin

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

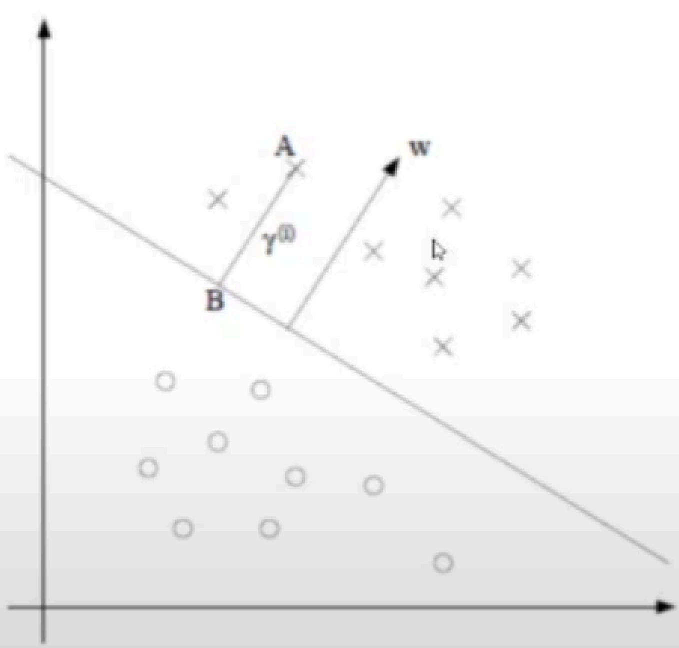
Introduce scaling constraint to make $\hat{\gamma} = 1.$

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

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

Functional and geometric margins

- Geometric margins



법선 백터의 크기를 최소화한다는 것의 의미= w, b가 정해지면 거리를 정확히 계산할 수 있음
우변을 1로 만들었던 이유 = 실제 거리를 계산했을 때 그 거리가 1보다 작으면 w를 적절히 조정하여 그 거리를 1로 키움, 실제 거리를 계산했을 때 그 거리가 1보다 크면 그 거리를 1로 축소

원래 거리가 멀었을수록 구별하기가 쉬움 -> 쉬운 문제

Lagrange duality

- **Primal optimization problem**

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

– To solve it, define generalized Lagrangian.

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

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

If constraints are violated by w,

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

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

constraint problem을 unconstraint problem으로 변경
변경된 문제를 해결하나 원래 문제를 해결하나 똑같다.

[라그랑주 승수법 참고]
https://velog.io/@nochesita/%EC%B5%9C%EC%A0%81%ED%99%94%EC%9D%B4%EB%A1%A0-%EB%9D%BC%EA%B7%B8%EB%9E%91%EC%A3%BC-%EC%8A%B9%EC%88%98%EB%B2%95-Lagrange-Multiplier-Method

Lagrange duality

– Consider minimization problem

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

- 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

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

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

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

original 문제가 안풀리는 경우 뒤집어서 풀면(min,max 순서를) 잘 풀리는 경우가 있다.

Lagrange duality

Karush-Kuhn-Tucker (KKT) conditions.

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

- If some w^*, α^*, β^* satisfy the KKT conditions, then it is also a solution to the primal and dual problems
- Equation (5), which is called KKT dual complementary condition.
- It implies that if $\alpha_i^* > 0$, then $g_i(w^*) = 0$ → true only for support vectors!

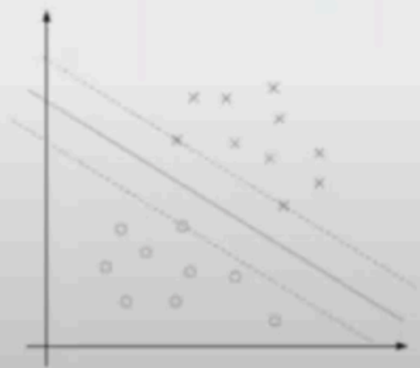
Optimal margin classifier

- **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

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



Three points are called **support vectors**

위 부등식을 등식으로 만족하는 데이터 포인트는 직선상의 데이터 포인트 3개이다.

Optimal margin classifier

– Construct the Lagrangian for our optimization problem

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

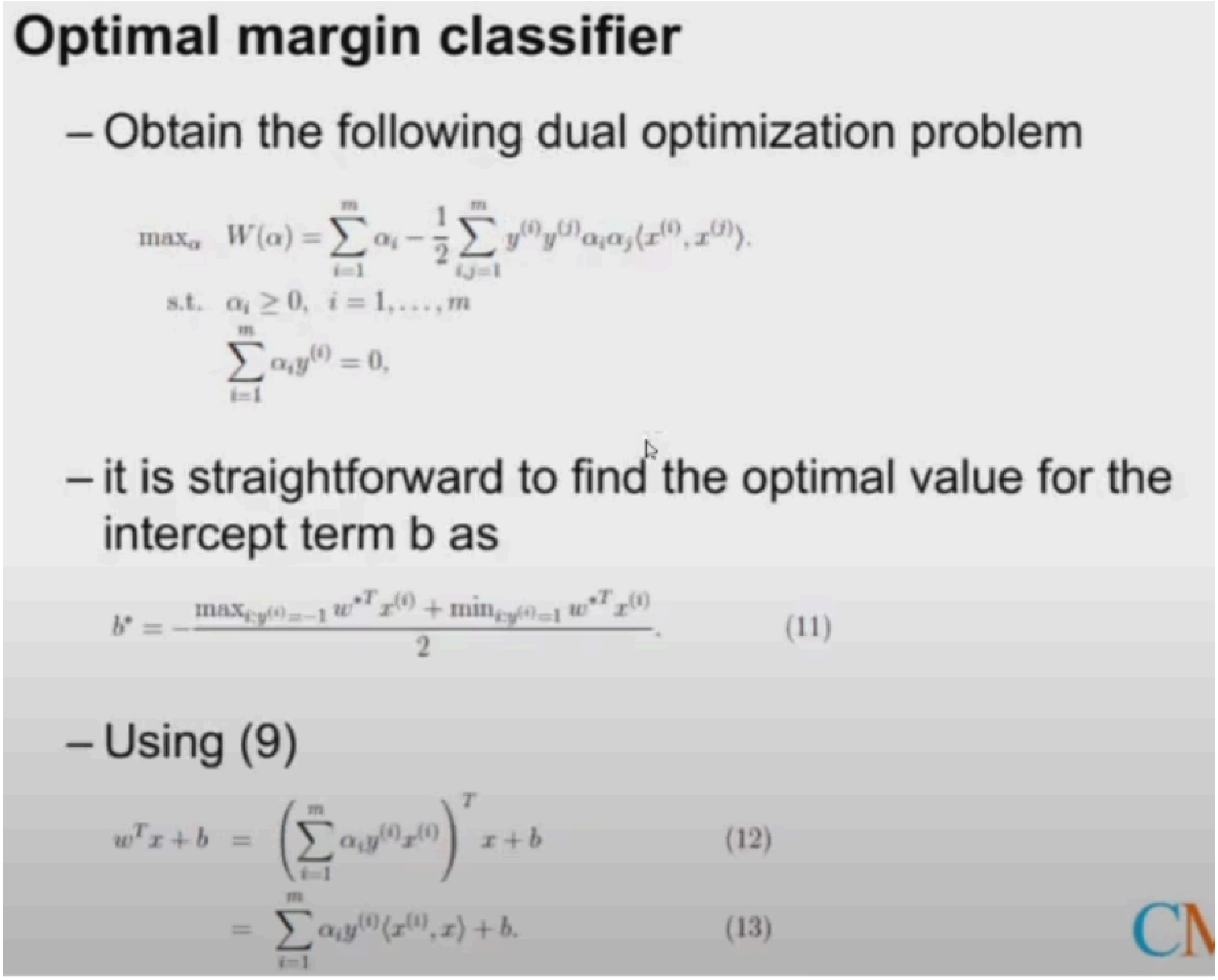
$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

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

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

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

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



SVM = binary classifier

training data의 수만큼 내적 연산을 해야함
그럼에도 좋은 알고리즘인 이유 = (5)의 조건에 의해 support vector의 연산만 남기고 나머지는 사라지고 한두개의 support vector의 연산만 진행하면 된다.

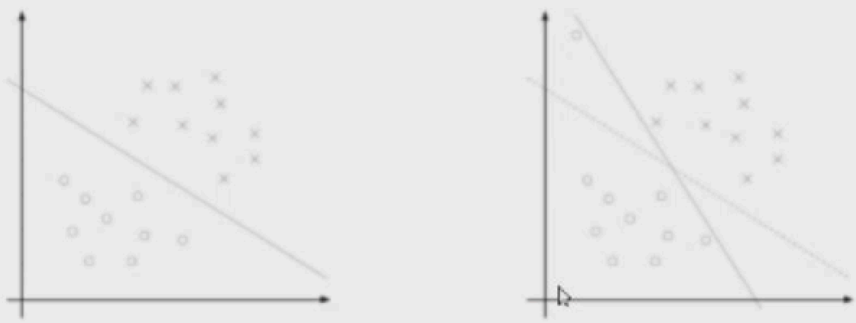
Support vector에 해당하는 부분에만 내적 연산을 진행하면 된다.

굉장히 좋은 최적화 알고리즘이라고 함

한계점:

1. 결정 경계가 직선형이라는 것, 많은 데이터들은 곡선형태로 뒤얽혀있음 -> 내적을 non-linear func로 변경하여 문제 해결(kernel)
2. 데이터 포인터들은 완벽하지 않다(노이즈, 이상치가 있다)

Regularization and the non-separable case



- A single outlier is added in right figure

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

- 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

C = 포기하는 정도

Regularization and the non-separable case

– 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

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

– KKT dual-complementarity conditions

$$\begin{aligned} \alpha_i = 0 \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) > 1 \qquad (14) \quad \rightarrow \text{Non-support vectors} \\ \alpha_i = C \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) < 1 \qquad (15) \quad \rightarrow \text{Relaxed cases} \\ 0 < \alpha_i < C \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) = 1. \qquad (16) \quad \rightarrow \text{Support vectors} \end{aligned}$$

alpha = C 인 경우 -> 포기하는 경우라서 1보다 작아지는 수식이 나온 것