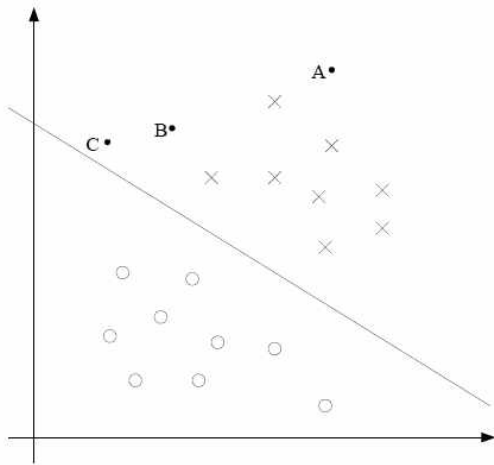


Machine Learning Lecture Note

This lecture note is based on the CS229 lecture materials from Stanford University.

Lecture 5. Support Vector Machine



A는 decision boundary에서 가장 멀리 떨어져 있고, A에서의 y 를 위한 예측값은 명확하게 1로 볼 수 있다.

그러나, C에서는 decision boundary에 대한 조금의 변화가 발생하게 될 때 y 의 예측값은 1이 아닌 0으로 판단할 수도 있다.

그래서, decision boundary (separating hyperplane)에서 멀리 떨어질수록 신뢰도가 높아짐을 알 수 있다.

∴ 훈련 데이터로부터 명확한 decision boundary를 찾는 것이 중요하다.

$$h_{\theta}(x) = g(\theta^T x) \text{ , predict 1 if } \theta^T x \gg 0; 0 \text{ otherwise}$$

So, If $y^i = 1$, hope that $\theta^T x^i \gg 0$

If $y^i = 0$, hope that $\theta^T x^i \ll 0$

Decision boundary를 찾기 위해 margin을 이용한다.

Notation

labels $y \in \{-1, 1\}$

$$g(z) = \begin{cases} 1 & \text{if } z \gg 0 \\ -1 & \text{otherwise} \end{cases}$$

$$h_{\theta}(x) = g(\theta^T x) \Rightarrow h_{w,b}(x) = g(w^T x + b) ; \sum_{i=1}^n w^T x^i + b$$

Functional margin

훈련 데이터 (x^i, y^i) 가 주어지면, (w, b) 의 functional margin은 다음과 같이 정의한다:

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

If $y^i = 1$, want $w^T x^i + b \gg 0$

If $y^i = -1$, want $w^T x^i + b \ll 0$

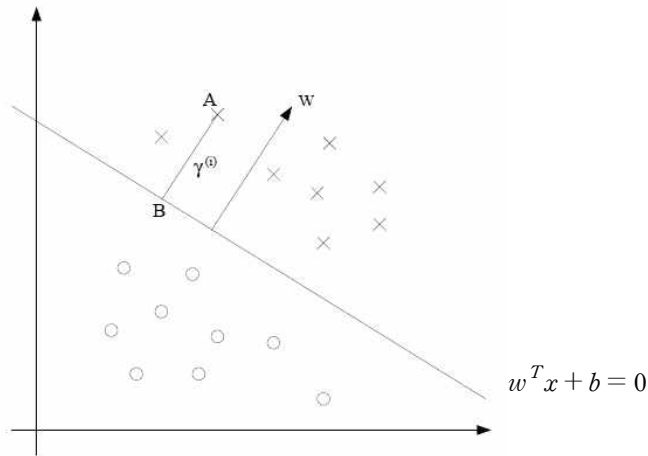
Want $\hat{\gamma}^i \gg 0$

If $\hat{\gamma}^i > 0$, that means $h(x^i) = y^i$

훈련 데이터 셋에 대하여 functional margin은 다음과 같이 정의한다:

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

Geometric margin



벡터 w 는 separating hyperplane에 수직이고, 훈련 데이터 x^i 에 대하여 레이블 $y^i = 1$ 인 지점 A에서 결정 경계까지의 거리에 해당하는 선분 AB가 geometric margin 이다. (즉, 포인트 A (x^i, y^i) 에서 $w^T x + b$ 와 수직인 거리 γ^i 가 geometric margin이 된다.)

선분 AB를 통해 γ^i 를 찾는다. 포인트 A에서 벡터 w 에 대한 단위길이 벡터를 적용한 거리 γ^i 를 빼면 포인트 B $(x^i - \gamma^i \frac{w}{\|w\|})$ 를 얻을 수 있다. 포인트 B는 $w^T x + b = 0$ 위에 놓여지기 때문에 다음과 같다:

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

그러면, γ^i 는 다음과 같이 나타낼 수 있다.

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

따라서, 일반적으로 (w, b) 를 가진 훈련 데이터 (x^i, y^i) 대한 geometric margin은 다음과 같이 정의한다.

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

만약, $\|w\| = 1$ 이라면, geometric margin은 functional margin과 같아진다.

훈련 데이터 셋 $S = \{(x^i, y^i); i = 1, \dots, m\}$ 이 주어지면, S의 개별 geometric margin 값들 중에서 가장 작은 값을 S의 geometric margin으로 정의한다.

$$\gamma = \min_i \gamma^i \quad (i = 1, \dots, m)$$

Optimal margin classifier : γ 를 최대화 하기 위한 w, b 를 선택

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

모든 훈련 데이터는 geometric margin 값보다 크거나 같아야 한다.

위 문제를 효율적으로 해결하기 위해, 선형 제약조건을 갖는 convex quadratic 함수에 대한 최적화 문제로 변환한다. 변환된 문제의 해는 최적 마진 분류기(optimal margin classifier)를 제공한다.

문제를 간단히 하기 위해 $\|w\| = \frac{1}{\gamma}$ 을 적용하면,

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

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

margin = distance(x^+, x^-)

$$= \|x^+ - x^-\|_2$$

$$= \|x^- + \alpha w - x^-\|_2$$

$$= \|\alpha w\|_2 = \alpha \sqrt{w^T w}$$

$$= \frac{2}{w^T w} \sqrt{w^T w} = \frac{2}{\sqrt{w^T w}} = \frac{2}{\|w\|_2}$$

$$x^+ = x^- + \alpha w$$

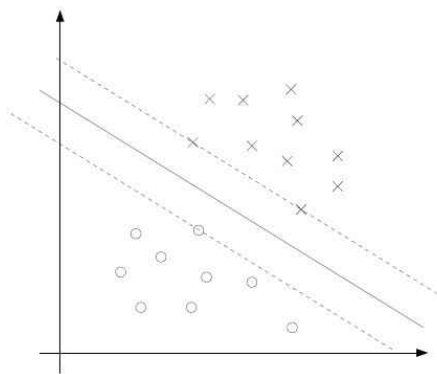
$$w^T x^+ + b = 1$$

$$w^T (x^- + \alpha w) + b = 1$$

$$w^T x^- + b + \alpha w^T w = 1$$

$$\alpha w^T w = 2$$

$$\alpha = \frac{2}{w^T w}$$



Support vectors: decision boundary에 가장 가까이 있는 margins (훈련 데이터들)

※ 최적화 문제: 라그랑지안 함수 + KKT 조건

라그랑지안 (Lagrangian)

$$\min f(x) \\ \text{s.t. } h(x) = 0 \\ g(x) \leq 0$$

제약 조건을 원래의
문제에 반영

$$L(x, \lambda, \mu) = f(x) + \lambda h(x) + \mu g(x)$$

KKT 조건

$$\nabla L(x^*, \lambda, \mu) = 0$$

$$\mu g(x^*) = 0$$

$$\mu \geq 0$$

$$h(x^*) = 0$$

$$g(x^*) \leq 0$$

Primal problem)

$$\min_{w, b} \frac{1}{2} \|w\|^2 \quad \Rightarrow \text{Lagrangian : } L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i [y^i(w^T x^i + b) - 1]$$

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

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

$$\nabla_b L(w, b, \alpha) = \sum_{i=1}^m \alpha_i y^i = 0 \quad \text{--②}$$

식 ①, ②를 라그랑지안에 대입하면,

$$\begin{aligned}
L(w, b, \alpha) &= \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i [y^i (w^T x^i + b) - 1] \\
&= \frac{1}{2} \left(\sum_{i=1}^m \alpha_i y^i x^i \right)^T \left(\sum_{j=1}^m \alpha_j y^j x^j \right) - \sum_{i=1}^m \alpha_i y^i \left(\sum_{j=1}^m \alpha_j y^j x^j \right)^T x^i - b \sum_{i=1}^m \alpha_i y^i + \sum_{i=1}^m \alpha_i \\
&= \sum_{i=1}^m \alpha_i - \frac{1}{2} \left(\sum_{i=1}^m \alpha_i y^i x^i \right)^T \left(\sum_{j=1}^m \alpha_j y^j x^j \right) \\
&= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^i y^j (x^i)^T x^j \\
&= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle
\end{aligned}$$

Dual problem)

α 에 대한 dual problem으로 변환하면,

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

dual problem을 통해 최적 α 를 구한다.

또한 b 는 primal problem으로부터 구할 수 있다.

$$b = - \frac{\max_{i: y^i = -1} w^T x^i + \min_{i: y^i = 1} w^T x^i}{2}$$

새로운 입력 데이터 x 에 대한 예측을 수행할 때는 $w^T x + b$ 의 계산 결과가 0보다 크면 $y = 1$ 로 예측한다.

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