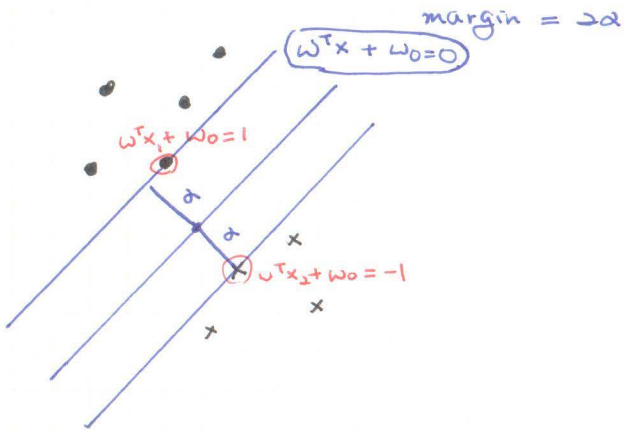
Support Vector Machines

1) Separable cases



- We want to find the equation $w^T x + w_0 = 0$ such that the margin $2d$ is maximized.
- Normalize $w^T x + w_0$ such that

$$\begin{cases} w^T x + w_0 \geq 1 & x \in \omega_1 \\ w^T x + w_0 \leq -1 & x \in \omega_2 \end{cases}$$

- Problem

$$\max \quad 2d = \max \frac{2}{\|w\|}$$

subject to

$$w^T x + w_0 \geq 1 \quad x \in \omega_1$$
$$w^T x + w_0 \leq -1 \quad x \in \omega_2$$

- Alternatively

$$\min \frac{1}{2}\|w\|^2$$
subject to $y_i (w^T x_i + w_0) \geq 1$ for all i.

where $y_i = 1$ if $x_i \in \omega_1$
$y_i = -1$ if $x_i \in \omega_2$

* It is a convex programing problem.
* We can use KKT conditions and the Wolfe dual representation technique.
* Recall
  - KKT Conditions

$$L(\theta, \lambda) = J(\theta) - \sum_i \lambda_i f_i(\theta)$$

① $\frac{\partial L}{\partial \theta} = 0$

② $\lambda \geq 0$

③ $\lambda_i f_i(\theta) = 0$

  - $\min_\theta \max_{\lambda \geq 0} L(\theta, \lambda)$

$$= \max_{\lambda \geq 0} L(\theta, \lambda) \quad \text{subject to} \quad \frac{\partial L}{\partial \theta} = 0.$$

$$L(w, w_0, \lambda) = \frac{1}{2} w^T w - \sum \lambda_i \left[y_i (w^T x_i + w_0) - 1 \right]$$

(a) KKT Conditions

① $\frac{\partial L}{\partial w} = w - \sum \lambda_i y_i x = 0$ $\quad \boxed{w = \sum \lambda_i y_i x}$

$\frac{\partial L}{\partial w_0} = - \sum \lambda_i y_i = 0$ $\quad \boxed{\sum \lambda_i y_i = 0}$

② $\lambda_i \geq 0$

③ $\lambda_i \left[y_i (w^T x + w_0) - 1 \right] = 0$

* For $\lambda_i \neq 0$ (i.e. $\lambda_i > 0$)

x_i is called a support vector.

$w^T x_i + w_0 = \pm 1$

Support vectors are the training vectors that are closest to the linear classifier.

(b) Wolfe Dual Representation Form

$$\max_{\lambda \geq 0} L(w, w_0, \lambda)$$

subject to $\frac{\partial L}{\partial w} = 0$ i.e. $w = \sum \lambda_i y_i x$

$\frac{\partial L}{\partial w_0} = 0$ i.e. $\sum \lambda_i y_i = 0$

or

$$\max L(w, w_0, \lambda)$$

subject to $\begin{cases} \lambda_i \geq 0 \\ w = \sum \lambda_i y_i x_i \\ \sum \lambda_i y_i = 0 \end{cases}$

Note that

$$L(w, w_0, \lambda)$$

$$= \frac{1}{2} \left(\sum_i \lambda_i y_i x_i \right)^T \sum_j \lambda_j y_j x_j - \sum \lambda_i \left\{ y_i \left[\left(\sum_j \lambda_j y_j x_j \right)^T x_i + w_0 \right] - 1 \right\}$$

$$= - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum \lambda_i y_i w_0 + \sum \lambda_i$$

$$= \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

Therefore, we solve

$$\max_{\lambda} \left(\sum \lambda_i - \frac{1}{2} \sum \lambda_i \lambda_j y_i y_j (x_i^T x_j) \right)$$

subject to $\begin{cases} \lambda_i \geq 0 \\ \sum \lambda_i y_i = 0 \end{cases}$

① λ 결정

It does not depend on the dimensionality of x !

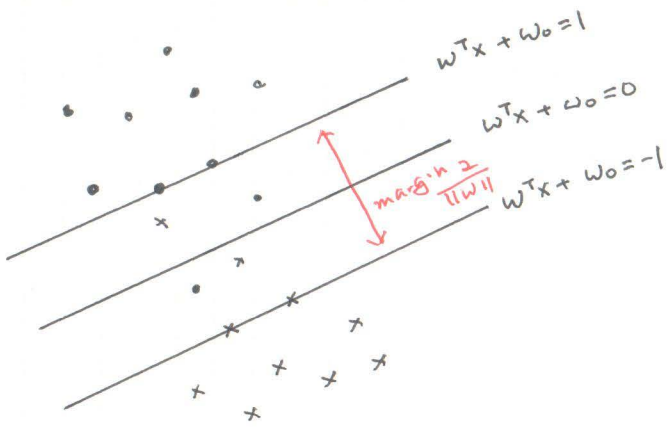
After solving this

$w = \sum_{i=1}^{N} \lambda_i y_i x_i$ ②

For $\lambda_i \neq 0$

$y_i (w^T x_i + w_0) = 1$. ③

2) Nonseparable cases



* Slack variables $\xi_i > 0$

$$y_i [w^T x + w_0] \geq 1 - \xi_i$$

* New formulation

$$\min \quad J(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to $y_i [w^T x + w_0] \geq 1 - \xi_i$ for each i

$\xi_i \geq 0$ "

- It is a convex programming problem

Lagrangian

$$f(w, w_0, \xi, \lambda, \mu) = \frac{1}{2}\|w\|^2 + C\sum \xi_i - \sum \mu_i \xi_i - \sum \lambda_i [y_i(w^T x + w_0) - 1 + \xi_i]$$

KKT Conditions

(1) $\frac{\partial f}{\partial w} = w - \sum \lambda_i y_i x = 0 \quad \therefore w = \sum \lambda_i y_i x_i$

$\frac{\partial f}{\partial w_0} = -\sum \lambda_i y_i = 0 \quad \therefore \sum \lambda_i y_i = 0$

$\frac{\partial f}{\partial \xi_i} = C - \mu_i - \lambda_i = 0$

(2) $\lambda_i \geq 0$, $\mu_i \geq 0$

(3) $\mu_i \xi_i = 0$, $\lambda_i [y_i (w^T x + w_0) - 1 + \xi_i] = 0$

Wolfe Dual Representation

$$\min \max_{\substack{\mu \geq 0 \\ \lambda \geq 0}} f = \max_{\substack{\mu \geq 0 \\ \lambda \geq 0}} \min f$$

$$= \max_{\substack{\mu \geq 0 \\ \lambda \geq 0}} f \quad \text{subject to} \quad w = \sum \lambda_i y_i x_i, \quad \sum \lambda_i y_i = 0, \quad C - \mu_i - \lambda_i = 0$$

This is simplified to

$$L = \frac{1}{2}\sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + C\sum \xi_i - \sum \mu_i \xi_i - \sum \lambda_i \xi_i$$
$$- \sum \lambda_i \left[y_i(w^T x_i + w_0) - 1 \right]$$

$$= -\frac{1}{2}\sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum \lambda_i (y_i w_0 - 1)$$

$$= \sum_i \lambda_i - \frac{1}{2}\sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j$$

subject to $\lambda_i \geq 0$

$\mu_i = C - \lambda_i \geq 0$

$\therefore$ $0 \leq \lambda_i \leq C$ ← The only difference.

$\sum \lambda_i y_i = 0$

- If x_i is inside the margin or wrongly classified

$\xi_i > 0 \Rightarrow \mu_i = 0 \Rightarrow \lambda_i = C.$