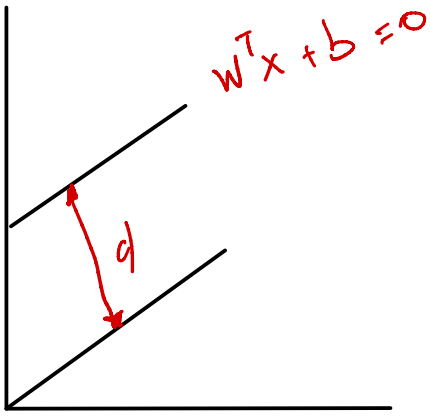

$$d = -\frac{b}{\|w\|}$$

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

$$\left.\begin{array}{ll} w^T x_i + b \geq +1 & \text{for } y_i = +1 \\ w^T x_i + b \leq -1 & \text{for } y_i = -1 \end{array}\right\} \forall i \; (w^T x_i + b) \geq 1$$

$$\min \frac{2}{\|w\|} = \max \frac{\|w\|}{2}$$

Lagrangian

$$L(\theta, \alpha) = f_0(\theta) + \sum_{i=1}^{M} \alpha_i f_i(\theta)$$

$$\min_{\substack{\theta \in \mathbb{R}^d \\ f_i(\theta) \leq 0}} f_0(\theta) = f_0(\theta^*) \geq f_0(\theta^*) + \underbrace{\sum_{i=1}^{M} \underbrace{\alpha_i}_{\geq 0} \underbrace{f_i(\theta^*)}_{\leq 0}}_{\leq 0} = L(\theta^*, \alpha) \geq \underbrace{\min_{\theta \in \mathbb{R}^d} L(\theta, \alpha)}_{g(\alpha)}$$

Hence, $\forall \alpha \; f_0(\theta^*) \geq g(\alpha)$ — $g(\alpha)$ is a lower bound

Lagrangian dual problem

maximize$_\alpha$ $g(\alpha)$

subject to $\alpha_i \geq 0, \; i = 1, \ldots, m$

(always)

weak duality $g(\alpha) \leq p^*$

$d^* \leq p^*$

strong duality (under certain conditions) $g(\alpha) = p^*$

$d^* = p^*$

SVM

1. Calculate the Lagrangian (constrained)

$$L(\underbrace{w, b, \alpha}_{\theta}) = \underbrace{\frac{1}{2} w^T w}_{f_0(\theta)} - \sum_{i=1}^{N} \alpha_i \underbrace{[y_i (w^T x_i + b) - 1]}_{f_i(\theta)}$$

2. Minimize $L(w, b, \alpha)$ w.r.t. w and b $\quad g(\alpha) = \min_{w,b} L(w, b, \alpha)$

$$\nabla_w L(w, b, \alpha) = w - \sum_{i=1}^{N} \alpha_i y_i x_i \stackrel{!}{=} 0 \quad ; \quad \underline{w = \sum_{i=1}^{N} \alpha_i y_i x_i}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i \stackrel{!}{=} 0$$

The weights are a linear combination of the training samples

Substituting both relations back into $L(w, b, \alpha)$ gives the Lagrange dual function $g(\alpha)$

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

$$= \frac{1}{2} w^T \left(\sum_{i=1}^{N} \alpha_i y_i x_i \right) - \sum_{i=1}^{N} \alpha_i y_i w^T x_i - \sum_{i=1}^{N} \alpha_i y_i b + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2} w^T \left(\sum_{i=1}^{N} \alpha_i y_i x_i \right) - b \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2} \sum_{j=1}^{N} \alpha_j y_j x_j^T \left(\sum_{i=1}^{N} \alpha_i y_i x_i \right) + \sum_{i=1}^{N} \alpha_i \quad , \text{ since } \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$= g(\alpha)$$

maximize $g(\alpha)$

subject to $\sum_{i=1}^{N} \alpha_i y_i = 0$

$$\alpha_i \geq 0, \quad \text{for } i = 1, \ldots, N$$

Rewrite $g(\alpha)$ in vector form

$$g(\alpha) = \frac{1}{2}\alpha^T Q \alpha + \alpha^T 1_N$$

$Q \rightarrow$ symmetric negative (semi) definite matrix

→ To solve this, there is an efficient algorithm Sequential minimal optimization (SMO)

Recover w and b from dual solution α^*

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

$\alpha_i^* \; f_i(\theta^*) = 0$

$$w^T x_i + b = y_i$$

$$b = y_i - w^T x_i$$

$y_i \begin{cases} -1 \\ 2 \end{cases}$

$f_i(\theta^*) = y_i (w^T x_i + b) - 1 \rightarrow y_i (w^T x_i + b) = 1 \; ; \; (w^T x_i + b) = \frac{1}{y_i} \; ;$

$w^T x_i + b = y_i \; ; \; b = y_i - w^T x_i$

A training sample x_i can only contribute to the weight vector ($\alpha_i \neq 0$) if it lies on the margin, that is $y_i (w^T x_i + b) = 1$

A training sample x_i with $\alpha_i \neq 0$ is called support vector

Slack variable $\xi_i \geq 0$ to control noisy data points

$\xi_i = 0 \Rightarrow$ No violation

$$\begin{cases} w^T x_i + b \geq +1 - \xi_i & \text{for } y_i = +1 \\ w^T x_i + b \leq -1 + \xi_i & \text{for } y_i = -1 \end{cases} \rightarrow y_i (w^T x_i + b) \geq 1 - \xi_i$$

The new cost function is, $f_0(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$

The factor $C > 0$ determines how heavy a violation is punished

$C \to \infty$ recovers hard-margin SVM

$$\min f_0(w, b, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

$$\text{subject to } \begin{cases} y_i (w^T x_i + b) - 1 + \xi_i \geq 0 \\ \xi_i \geq 0 \end{cases}$$

1. Lagrangian

$$L(w, b, \xi, \alpha, \mu) = \overbrace{\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i}^{f_0(\theta)} - \sum_{i=1}^{N} \alpha_i \underbrace{[y_i (w^T x_i + b) - 1 + \xi_i]}_{f_i(\theta)} - \sum_{i=1}^{N} \mu_i \xi_i$$

$f_i(\theta)$

2. Minimize

$$\nabla_w L(w, b, \xi, \alpha, \mu) = w - \sum_{i=1}^{N} \alpha_i y_i x_i \stackrel{!}{=} 0 \quad ; \quad \frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i \stackrel{!}{=} 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i \stackrel{!}{=} 0 \quad \text{for } i = 1, \ldots, N$$

From $\alpha_i = C - \mu_i$, and dual feasibility $\mu_i \geq 0$, $\alpha_i \geq 0$ we get $\boxed{0 \leq \alpha_i \leq C}$

$$\max \; g(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \, x_i^T x_j$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \quad i = 1, \dots, N \end{cases}$$

Only the constraint $\alpha_i \le C$ is new, it ensures that α_i is bounded and cannot go to infinity

C big → small margin (less datapoints violating in margin)

C small → big margin (more datapoints violating in margin)

$$\xi_i = \begin{cases} 1 - y_i (w^T x_i + b) & \text{if } \overbrace{y_i (w^T x_i + b) < 1}^{\text{margin is violated}} \\ 0 & \text{else} \end{cases}$$

since we are minimizing over ξ

$$\max \left(0, \; 1 - y_i \left(w^T x_i + b \right) \right)$$

↓ we can rewrite the objective function as an unconstrained optimization problem known as the **hinge loss** formulation

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \max \{ 0, \; 1 - y_i (w^T x_i + b) \}$$

Kernel trick → to construct non linear classifiers

$$g(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \underbrace{\phi(x_i)^T \phi(x_j)}_{K(x_i, x_j)}$$

Kernel function $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$

$$K(x_i, x_j) := \phi(x_i)^T \phi(x_j)$$

Kernel is valid if it corresponds to an inner product in some feature space

Mercer's theorem → A kernel is valid if it gives rise to a symmetric, positive semi definite kernel matrix K for any input data X

a new point can be classified as

$$h(x) = \text{sign}\left(\sum_{\xi_j \mid x_j \in S_1} a_i y_j k(x_j, x) + b \right)$$