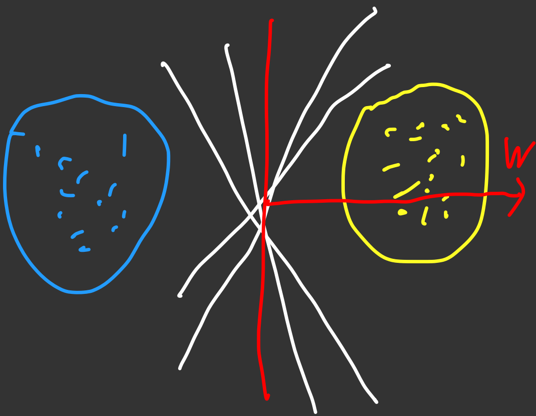
SVM II



pick one maximize the closet distance

primal optimum

$$\min \frac{1}{2}\|w\|_2^2$$

subj to $w \in \mathbb{R}^d$

$$y_i x_i^T w \geq 1$$

Inf classifiers to chose

Primal Dual

$$L(w, \alpha)$$

Original form

equivalent when feasible.

$$\min \boxed{\frac{1}{2}\|w\|_2^2}$$

subj to $w \in \mathbb{R}^n$

$$y_i x_i^T w \geq 1 \ \forall i$$

$$1 - y_i x_i^T w \leq 0$$

Form without constraints

$$\min_{w} \sup_{\alpha \geq 0} \left[\frac{1}{2}\|w\|_2^2 + \sum_i \alpha_i (1 - y_i x_i^T w) \right]$$

sup: max can take ∞

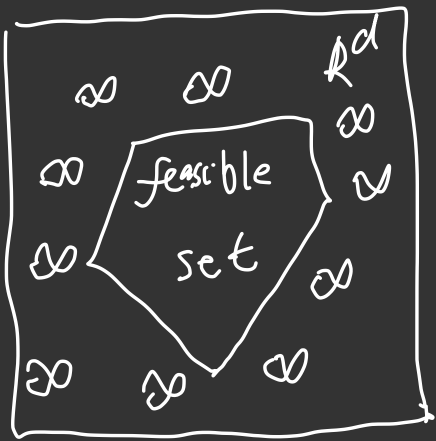
if $\exists j$ $y_i x_j^T w < 1$

α can take ∞

so the eq

$\to \infty$

$$\min_{w} \begin{cases} \infty & \text{constraint violate} \\ \frac{1}{2}\|w\|_2^2 & \text{feasible} \end{cases}$$



Lagrangian

$$L(w,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha (1 - y_i x_i^T w)$$

dual variable

$$\sup_{\alpha \geq 0} L(w,\alpha)$$

$$p(w)$$

dual $\min_{w \in R^d} L(w,\alpha)$

$$D(\alpha)$$

Hard margin Dual

$$\min_w L(w,\alpha) = \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha(1 - y_i x_i^T w)$$

$$\nabla w = 0 = w - \sum_i \alpha_i y_i x_i \Rightarrow w = \sum \alpha_i y_i x_i$$

$$= \frac{1}{2}\|\underbrace{\sum_i \alpha_i y_i x_i}_{w_\alpha}\|^2 + \sum \alpha_i - \underbrace{w_\alpha^T \sum \alpha_i y_i x_i}_{\|w_\alpha\|^2}$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\|\sum \alpha_i y_i x_i\|^2$$

$$= \underbrace{D(\alpha)}_{\text{concave}} \quad \alpha \geq 0$$

Weak duality: $P(w^*) \geq D(\alpha^*)$

With convexity $\longrightarrow$ **Strong convexity**

$$\bar{w} = \sum_{i=1}^{n} \bar{\alpha} y_i x_i = \arg\min_w L(w, \bar{\alpha})$$

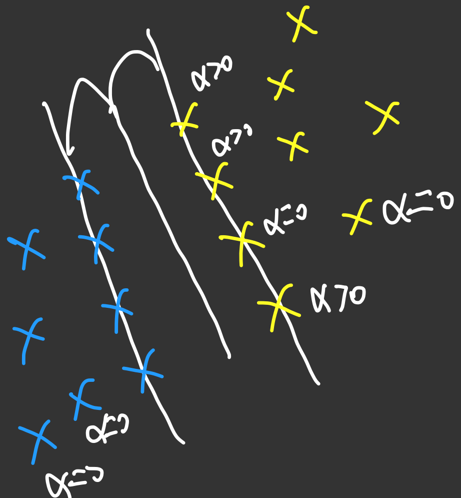
$$\min P(w) = \min_w \max_{\alpha \geq 0} L(w,\alpha) = \max_{\alpha \geq 0} \min_w L(w,\alpha) = \max_{\alpha \geq 0} D(\alpha)$$

$\boxed{\text{Complementary Slackness}}$

$$\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i x_i$$

$$\bar{\alpha}_i > 0 \Rightarrow y_i x_i^T \bar{w} = 1 \text{ for all } i = 1 \cdots n$$

$$\bar{\alpha}_i > 0 \Rightarrow y_i x_i^T \bar{x} = 1$$



Proof:

Let $p(\bar{w}) = D(\bar{\alpha})$

$$= \min_{w \in \mathbb{R}^d} L(w, \bar{\alpha})$$

$$\le L(\bar{w}, \bar{\alpha})$$

$$= \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i x_i^T \bar{w})$$

$$\le \frac{1}{2}\|\bar{w}\|_2^2$$

$$= \boxed{p(\bar{w})}$$

So $\bar{\alpha}_i (1 - y_i x_i^T \bar{w}) = 0 \quad \forall i$

So $\bar{\alpha}_i (1 - y_i x_i^T \bar{w}) = 0 \quad \forall i$

So $\bar{\alpha}_i > 0 \Rightarrow 1 - y_i x_i^T \bar{w} = 0$

$\boxed{\text{SVM Soft-margin Dual}}$ ☆

$$L(w, \xi, \alpha) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1-\xi_i - y_i x_i^T w\right)$$

$$P(w, \xi) = \sup_{\alpha \geq 0} L(w, \xi, \alpha)$$

$$= \begin{cases} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i & \forall i \; 1-\xi_i - y_i x_i^T w \leq 0 \\ \infty \end{cases}$$

$$D(\alpha) = \min_{w\in\mathbb{R}^d \; \xi\in\mathbb{R}^n \geq 0} L(w, \xi, \alpha) \qquad \bar{w} = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$= \frac{1}{2}\left\|\sum_i \alpha_i y_i x_i\right\|^2 + \sum_{i=1}^{n}\alpha_i - w_\alpha^T\sum \alpha_i y_i x_i + C\sum_{i=1}^{n}\xi_i$$

$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$= \max_{\substack{\alpha\in\mathbb{R}^n \\ 0\leq\alpha_i\leq C}}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i y_i x_i\right\|^2\right]$$

$\boxed{\text{when } C\to\infty \rightarrow \text{hard margin}}$

Nonlinear SVM

$w \to \phi(x)^T w$ $\qquad \phi: \mathbb{R}^d \to \mathbb{R}^p$

hard: $\min \{ \frac{1}{2}\|w\|^2 : w \in \mathbb{R}^p, \forall i, \phi(x_i)^T (w \geq 1 \}$

Dual: $\max_{\alpha_1 \cdots \alpha_n \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{\phi(x_i)^T \phi(x_j)}_{\phi(x_i)^T \phi(x_j)}$

$\nabla w = 0 \implies \bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(x_i)$

future: $x \to \phi(x)^T \bar{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(x)^T \phi(x)$

Kernel trick: $K(\cdot, \cdot) = \phi(x)^T \phi(x')$

Eg.

① Affine features $\phi: \mathbb{R}^d \to \mathbb{R}^{1+d}$

$\phi(x) = (1, x_1 \cdots x_d)$

$\phi(x)^T \phi(x') = 1 + x^T x'$

② Qudratic features: $\phi(x) = (1, \sqrt{2}x_1 \cdots \sqrt{2}x_d, x_1^2 \cdots x_d^2$

$\sqrt{2}x_1x_2 \cdots \sqrt{2}x_{d-1}x_d)$

$\phi(x)^T \phi(x') = (1 + x^T x')^2$

③ RBF kernel (Gaussian kernel)

For any $\alpha > 0$ $\phi: \mathbb{R}^d \to \mathbb{R}^\infty$

$$K(x, x') \quad \phi(x)^T \phi(x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$$