MA 4270 Lecture 8

Last Time: Kernel

- Arise naturally when we consider Regularized cost Func.

$\theta \in \mathrm{span}(\{\phi(x_t)\})_1^n$.

Kernel func: $K: \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ $\exists$ a feature map. $\phi: \mathbb{R}^d \longrightarrow \mathbb{R}^D$

s.t $K(\underline{x}, \underline{x}') = \langle \underline{\phi}(\underline{x}), \underline{\phi}(\underline{x}') \rangle$

General Convex Optimization & KKT Conditions

Consider $\min_{x} g_0(x)$ s.t $\underbrace{g_j(x) \le 0}_{\text{inequality }(k)}$ $j=1,\dots,k$

$\underbrace{Ax = b}_{\text{equality }(\ell)}$

Feasible set $\mathcal{D} = \{ \underline{x} : g_j(\underline{x}) \le 0, j=1,\dots,k.\ A\underline{x} = b \}$

Def: Primal Optimal Value

$$p^* = \inf \{ g_0(x) : x \in \mathcal{D} \}$$

Def: $\inf \phi = +\infty \Rightarrow$ infeasible.

Def: Lagrangian $\underbrace{\lambda \in \mathbb{R}^k \quad \nu \in \mathbb{R}^\ell}$.

$$L(\underline{x}, \lambda, \mu) = g_0(x) + \sum_{i=1}^{k} \lambda_i g_i(x) + \nu^T (Ax - b)$$

$\boxed{\text{Claim}}$ $p^* = \inf_x \left[\max_{\lambda \geq 0} L(x, \lambda, \nu) \right]$. $\rightsquigarrow$ min max Problem

Pf. ① if x violates some inequl. constraints

then we have $g_j(x) > 0$ for some j

$\hookrightarrow$ $\max_{\lambda \geq 0} L(x, \lambda, \mu) \to +\infty$. for x violates the constraints

② if x does not violate any constraint

we all have $g_j(x) \leq 0$ $j = 1, \dots, k$

$\hookrightarrow$ $\max_{\lambda \geq 0} L(x, \lambda, \mu) = g_0(x)$ $\hookrightarrow$ set all $\underline{\lambda}$ to $\underline{0}$

All in all, $\max_{\lambda \geq 0, \nu} L(x, \lambda, \nu) = \begin{cases} +\infty, & x \notin \mathcal{D}. \\ g_0(x), & x \in \mathcal{D} \end{cases}$

Therefore $p^* = \inf_x \max_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$

Def: Lagrangian Dual Function.

$$\phi(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$$

Observation (Rmk): $\boxed{\lambda \geq 0}$, ν, $x \in \mathcal{D}$ feasible

$$\phi(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$$

$$\leq g_0(x) + \underbrace{\sum_{i=1}^{k} \lambda_i g_i(x)}_{\leq 0} + \underbrace{\nu^T (Ax - b)}_{=0} \quad \forall x$$

More Condition

$$\leq g_0(x) \quad \text{for } x \in \mathcal{D},\ \lambda_i \geq 0,\ \nu$$

$$\Rightarrow \phi(\lambda, \mu) \leq g_0(x) \leq \inf_{x \in \mathcal{D}} g_0(x) = p^*$$

Def: Lagrangian Dual Prob.

$$d^* = \max_{\lambda \geq 0, \nu} \phi(\lambda, \nu)$$

$$= \max_{\lambda \geq 0, \nu} \inf_{x} \mathcal{L}(x, \lambda, \nu)$$

[Lemma]: **Weak Duality.**

$\cancel{p^* \leq d^*}$

$d^* \leq p^*$

Pf: $\phi(\lambda, \nu) \leq p^*$

$\Rightarrow \max_{\lambda \geq 0, \mu} \phi(\lambda, \nu) \leq p^*$

$\Rightarrow d^* \leq p^*$

[Rmk]: dual opt. value $\leq$ primal opt. value.

$$\Updownarrow$$

$$\max_{\lambda \geq 0, \nu} \inf_{x} L(x, \nu, \lambda) \leq \inf_{x} \max_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

Strong Duality : $d^* = p^*$

$\Uparrow$

Need more condition to guarantee

(Slater Condition)

$\hookrightarrow$ If PRIMAL Prob. is convex,

and $\exists x$ s.t $g_i(x) < 0$ for $i = 1, 2, \ldots, k$

$Ax = b$,

then Strong Duality holds!

Strict feasibility!

Recall: $\lambda^* \geq 0,\ \nu^*$

$$d^* = \max_{\lambda \geq 0, \nu} \inf_{x} L(x, \lambda, \nu) = \max_{\lambda \geq 0, \nu} \phi(\lambda, \nu) \longrightarrow \lambda^*, \nu^*$$

$$p^* = \inf \max L(x, \lambda, \nu) = \inf g_0(x) \longrightarrow \boxed{x^*}$$

$x \in \mathcal{D}$

If Strong Duality holds,

$x^* \in \mathcal{D}$

$\Rightarrow \begin{cases} g_i(x^*) \le 0 \\ Ax^* = b \end{cases}$

$$\begin{aligned} p^* = g_0(x^*) &= \phi(\lambda^*, \nu^*) \\ &= \inf_x L(x, \lambda^*, \nu^*) \\ &= \inf_x \left\{ g_0(x) + \sum_{i=1}^{k} \lambda_i^* g_i(x) + \nu^{*T}(Ax-b) \right\} \\ &\boxed{\le} g_0(x^*) + \underbrace{\sum_{i=1}^{k} \lambda_i^* g_i(x^*)}_{\lambda_i^* \ge 0,\ g_i(x^*) \le 0} + \underbrace{\nu^{*T}(Ax^*-b)}_{=0} \\ &\boxed{\le} g_0(x^*) \end{aligned}$$

Equality !

If Strong Duality holds, we have the following $\boxed{\text{necessary}}$ conditions for the optimality of (x^*, λ^*, ν^*)

KKT Condition

i) Stationarity $\boxed{\nabla_x L(x, \lambda^*, \nu^*) = 0}$

$$\nabla g_0(x^*) + \sum \lambda_i^* \nabla g_i(x^*) + A^T \nu^* = 0$$

☆ if we have $x^* \to (P)$
$(\lambda^*, \nu^*) \to (D)$
it must hold KKT Condition

ii) Primal Feasibility

$$g_i(x^*) \le 0$$

$$Ax^* = b$$

iii) Dual Feasibility

$$\lambda_i^* \ge 0 \quad \text{for } i = 1, 2, \dots, k.$$

iv) Complementary Slackness

$$\lambda_i^* g_i(x^*) = 0 \quad \text{for } i = 1, 2, \dots, k$$

E.g. SVM without slack

$$\min \frac{1}{2}\|\theta\|^2 \quad \text{s.t} \quad y_t(\langle\theta, x_t\rangle + \theta_0) \geq 1 \quad \text{for } t = 1, 2, \dots, n$$

Strong Duality $\Leftrightarrow$ we can find (θ, θ_0) s.t

$$y_t(\langle x_t, \theta\rangle + \theta_0) > 1$$

$\Leftrightarrow$ $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^n$ is affinely separable!

E.g. SVM with slack

$$\min \frac{1}{2}\|\theta\|^2 + C\sum \xi_t \quad \text{s.t} \quad y_t(\langle x_t, \theta\rangle + \theta_0) \geq 1 - \xi_t$$
$$\xi_t \geq 0$$

Note:

(strict feasibility)

We can always find (θ, θ_0, ξ) s.t

this means Slater condition always hold

$$y_t(\langle x_t, \theta\rangle + \theta_0) > 1 - \xi_t \quad t = 1, \dots, n$$
$$\xi_t > 0 \quad t = 1, 2, \dots, n$$

Reason: we can take ξ_t to be BIG enough

$\Downarrow$

Strong Duality always holds!

SVM duality

$$\min \frac{1}{2}\|\theta\|^2 + C\sum \xi_t$$

s.t $y_t(\langle x_t, \theta\rangle + \theta_0) \geq 1 - \xi_t \rightarrow \alpha_t \Leftrightarrow 1 - \xi_t - y_t(\langle\theta, \phi(x_t)\rangle + \theta_0) \leq 0$

$\xi_t \geq 0 \rightarrow \lambda_t \Leftrightarrow -\xi_t \leq 0$

Important:

Lagrangian: $[\underline{\alpha} = (\alpha_1, \dots, \alpha_n) \quad \underline{\lambda} = (\lambda_1, \dots, \lambda_n)] \rightarrow$ dual variables

$\underline{\theta} = (\theta_1, \dots, \theta_d)$ θ_0 $\underline{\xi} = (\xi_1, \dots, \xi_n) \rightarrow$ primal variables

Lag. v. Lag. v.

g_j g_j

$\mathcal{L}(\underline{\theta}, \theta_0, \underline{\xi} ; \underline{\alpha}, \underline{\lambda})$

$= \frac{1}{2}\|\theta\|^2 + C \sum \xi_t + \sum_{t=1}^{n} \boxed{\alpha_t} \left(1 - \xi_t - y_t (\langle \underline{\theta}, \phi(x_t) \rangle + \theta_0)\right) - \sum_{t=1}^{n} \boxed{\lambda_t} \xi_t$

primal variable dual variable

For $(\theta^*, \theta_0^*, \xi^* ; \alpha^*, \lambda^*)$ to be Primal- Dual Optimal,

we check the KKT (Necessary) condition.

Actually, all variables here are $(\underline{\theta}^*, \theta_0^*, \underline{\xi}^* ; \alpha^*, \lambda^*)$

① Stationarity.

$$\begin{cases} \frac{\partial L}{\partial \underline{\theta}} = 0 \Leftrightarrow \underline{\theta} - \sum \alpha_t y_t \phi(x_t) = 0 \\ \frac{\partial L}{\partial \theta_0} = 0 \Leftrightarrow \sum \alpha_t y_t = 0 \\ \frac{\partial L}{\partial \xi_t} = 0 \Leftrightarrow C - \alpha_t - \lambda_t = 0 \Leftrightarrow \alpha_t + \lambda_t = C \quad \text{For all } t \end{cases}$$

② Primal Feasibility

$y_t(\langle \underline{\theta}, \phi(x_t) \rangle + \theta_0) \geq 1 - \xi_t$ $\forall t = 1, 2, \dots, n$

$\xi_t \geq 0$

③ Dual Feasibility

$\boxed{\alpha_t \geq 0, \quad \lambda_t \geq 0}$ $\forall t = 1, 2, \dots, n$.

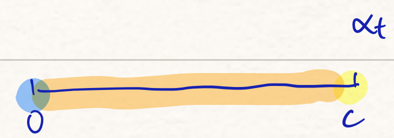
④ Complementary Slackness

$\alpha_t \left[1 - \xi_t - y_t(\langle \underline{\theta}, \phi(x_t) \rangle + \theta_0)\right] = 0$ $\forall t = 1, 2, \dots, n$

$\lambda_t \xi_t = 0$

What is SVs?

Some Results:



By combining , we have: $\alpha_t \in [0, C]$

$\Downarrow$

(Based on α_t)

Partition the $\{(\underline{x}_t, y_t)\}_{t=1}^{n}$ into 3 disjoint subsets!

① Non-margin SVs: $\alpha_t = C > 0$ $\leftrightarrow$ CS1

i) $\Rightarrow 1 - \xi_t - y_t(\langle \underline{\theta}, \underline{\varphi}(x_t) \rangle + \theta_0) = 0$

$y_t(\langle \underline{\theta}, \underline{\varphi}(\underline{x}_t) \rangle + \theta_0) = 1 - \xi_t$

ii) $\Rightarrow \lambda_t = 0 \longleftrightarrow \boxed{\alpha_t + \lambda_t = C}$

$\Rightarrow \xi_t$ can be positive

② Margin SVs: $\alpha_t \in (0, C)$ $\leftrightarrow$ CS1.

i) $\Rightarrow 1 - \xi_t - y_t(\langle \underline{\theta}, \underline{\varphi}(x_t) \rangle + \theta_0) = 0$

$y_t(\langle \theta, \underline{\varphi}(\underline{x}_t) \rangle + \theta_0) = 1 - \xi_t$

ii) $\Rightarrow \lambda_t \neq 0$

$\Rightarrow \xi_t = 0$ $\Leftarrow$ CS2.

i)+ii) $\Rightarrow y_t(\langle \underline{\theta}, \underline{\varphi}(\underline{x}_t) \rangle + \theta_0) = 1$.

$\hookrightarrow$ lies in the margin boundary

③ No SVs: $\alpha_t = 0$

i) $\lambda_t = C$

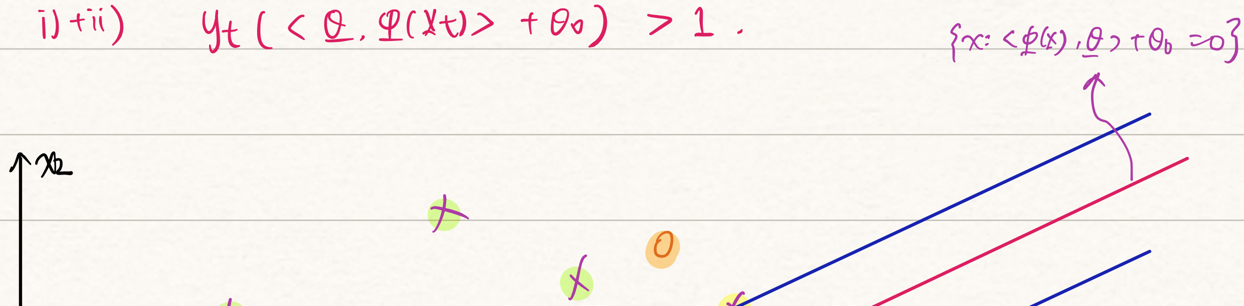
$\Rightarrow \xi_t = 0$ $\rightarrow$ CS2.

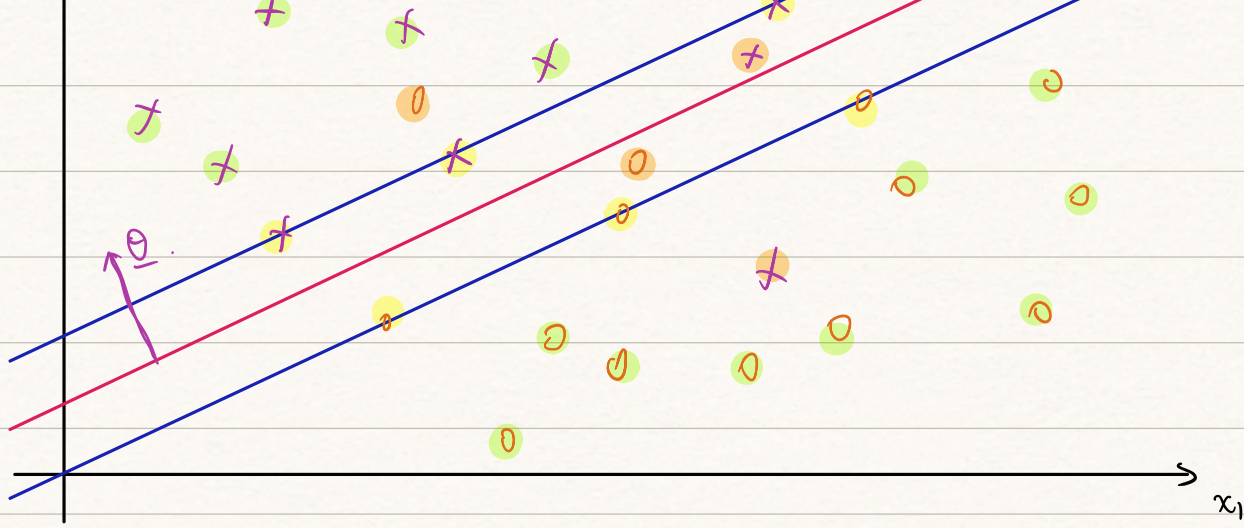
ii) $\alpha_t = 0$ $\rightarrow$ CS1

$1 - \xi_t - y_t(\langle \underline{\theta}, \underline{\varphi}(x_t) \rangle + \theta_0) < 0$

$y_t(\langle \theta, \underline{\varphi}(\underline{x}_t) \rangle + \theta_0) > 1 - \xi_t$

i)+ii) $y_t(\langle \underline{\theta}, \underline{\varphi}(\underline{x}_t) \rangle + \theta_0) > 1$.





- (green) Non SVs $\longrightarrow$ $\alpha_t = 0$
- (yellow) Margin SVs $\longrightarrow$ $\alpha_t \in (0, C)$ } SVs.
- (orange) Non-Margin SVs. $\longrightarrow$ $\alpha_t = C$ } SVs.

Defn: $SV = \{ (\underline{x}_t, y_t) : \alpha_t \in (0, C] \}$

Rmk: i) The solution is sparse, i.e., many points have $\alpha_t = 0$
$\downarrow$
Non SVs.

ii) Only the points on margins & those that result in Margin Errors contribute to the decision of a new test sample x'

$\Downarrow$

$\underline{\theta} \in \operatorname{span}\{\underline{\phi}(\underline{x}_t)\}_{t=1}^{n}$ $\longrightarrow$ $\boxed{\text{like kernel}}$

$$\hat{y}(x') = \langle \underline{\theta}, \underline{\phi}(\underline{x}') \rangle + \theta_0$$

$$= \langle \sum_{t=1}^{n} \alpha_t y_t \underline{\phi}(\underline{x}_t), \underline{\phi}(\underline{x}') \rangle + \theta_0$$

$\Downarrow$

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

$$= \sum_{t=1}^{n} \alpha_t y_t K(\underline{x}_t, \underline{x}') + \theta_0$$

$\rightarrow$ High dimension space

not need to $\Phi: \mathbb{R}^d \to \boxed{\mathbb{R}}$

$$= \sum_{t \in SV} \alpha_t y_t K(\underline{x}_t, \underline{x}') + \theta_0$$

$\boxed{\text{Offset: } \theta_0 \text{ ? How do we estimate ?}}$

$\Downarrow$

Method: pick a margin SV (i.e. $\alpha_t \in (0, C)$)

$\Downarrow$

$$y_t(\langle \underline{\theta}, \underline{\Phi}(\underline{x}_t) \rangle + \theta_0) = 1$$

$$\Rightarrow y_t\left(\left\langle \sum_{s=1}^{n} \alpha_s y_s \underline{\Phi}(\underline{x}_s), \underline{\Phi}(\underline{x}_t) \right\rangle + \theta_0\right) = 1$$

$$\Rightarrow y_t\left(\sum_{s=1}^{n} \alpha_s y_s K(\underline{x}_s, \underline{x}_t) + \theta_0\right) = 1$$

$$\Rightarrow \theta_0 = y_t - \sum_{s=1}^{n} \alpha_s y_s K(\underline{x}_s, \underline{x}_t)$$

$$= y_t - \sum_{s \in SV} \alpha_s y_s K(\underline{x}_s, \underline{x}_t)$$

都不一样

Question RBF $K(x, x') = \exp(-\frac{\beta}{2}\|x - x'\|^2)$ $\rightarrow$ $\boxed{\begin{matrix} \beta\uparrow & SV\downarrow \\ \beta\downarrow & SV\uparrow \end{matrix}}$

都一样