DSA5103 LEC5

$\boxed{\text{SVM}}$

---

1. **Idea of SVM**

$$\boxed{\max_{w,b} : \text{margin}(w,b)}$$

assume correctly separate these points

margin(w,b) : given a (hyperplane $H: w^T x + b = 0$),

the minimum distance between hyperplane and point

Observation : if hyperplane $\tilde{H}$ is such an optimal hyperplane,

then it must satisfy :

min distance between $\tilde{H}$ and positive class points

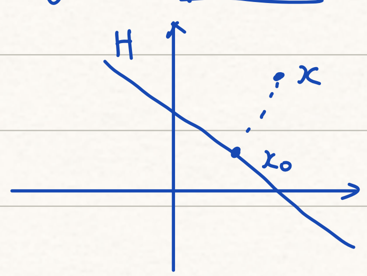
= min distance between $\tilde{H}$ and negative class points

---

2. **Question : How to calculate Distance between Hyperplane and Point ?**

a) Self-proposed method

given point x & hyperplane $H: w^T x + b = 0$



$$\text{distance} = |\langle x - x_0, w \rangle| / \|w\|$$
$$= \frac{|w^T x + b|}{\|w\|}$$

b) More sophisticated way

① consider $\forall x_0 \in H$, the normal cone

$$\boxed{N_H(x_0) = \{\lambda w : \lambda \in \mathbb{R}\}}$$

$$N_H(x) := \{ z : \langle z, y - x \rangle \leq 0, \forall y \in H \}$$

② then, for $\forall x \in \mathbb{R}^n$, its projection $\bar{x}$ on H satisfy:

$$\bar{x} \in \Pi_H(x)$$

$$\Leftrightarrow x - \bar{x} \in N_H(x)$$

$$\Leftrightarrow x - \bar{x} = \lambda \cdot w \quad \text{for some } \lambda$$

③ $\bar{x} \in H \Rightarrow w^T\bar{x} + b = 0$

$$\Rightarrow w^T(x - \lambda w) + b = 0$$

$$\Rightarrow \lambda = \frac{w^T x + b}{w^T w}$$

$$\Rightarrow \text{distance} = \| x - \bar{x} \| = |\lambda \cdot \|w\||$$

$$= \frac{|w^T x + b|}{\|w\|_2}$$

---

3. Formulation of SVM

$$\rightarrow \begin{cases} \max\limits_{w,b} \boxed{\min\limits_{i \in [n]} \dfrac{|w^T x_i + b|}{\|w\|}} \rightarrow \text{margin for dataset} \\ \text{s.t} \boxed{y_i (w^T x_i + b) \geq 0 \quad \forall i \in [n]} \end{cases}$$

Correctly separate all data

$$\Longleftrightarrow \begin{cases} \max\limits_{w,b} \boxed{\dfrac{1}{\|w\|} \min\limits_{i\in[n]} |w^T x_i + b|} = f(w,b) \\ \text{s.t} \quad y_i(w^T x_i + b) \geq 0 \quad i \in [n] \end{cases}$$

$\boxed{\text{observation}}$: $f(w,b) = f(kw, kb)$ for arbitrary k !

maintain same object, but shrink the maximizer set

$$\begin{cases} \max\limits_{w,b} \dfrac{1}{\|w\|} \\ \text{s.t} \quad y_i(w^T x_i + b) \geq 0 \quad i \in [n] \\ \qquad \min\limits_{i\in[n]} y_i(w^T x_i + b) = 1 \end{cases}$$

$$\Longleftrightarrow \begin{cases} \min\limits_{w,b} \frac{1}{2} w^T w \\ \text{s.t} \quad y_i(w^T x_i + b) \geq 1 \quad i \in [n] \end{cases} \longrightarrow \boxed{\text{Primal Form}}$$

(it must exists some k, s.t $y_k(w^T x_k + b) = 1$)

---

4. Lagrange Duality / KKT Condition

① $\boxed{\text{Non-Linear Programming}}$

$$\longrightarrow \begin{cases} \min\limits_{x} f(x) \\ \text{s.t} \quad g_i(x) = 0 \quad , i \in [m] \\ \qquad h_j(x) \leq 0 \quad , j \in [\ell] \\ \qquad x \in \mathcal{X} \quad (\mathcal{X} \subseteq \mathbb{R}^p) \end{cases}$$

Generally speaking, we can have KKT optimality condition

a) Necessary part : if $\bar{x}$ is a minimizer, then there exists $\{u_i : i \in [m]\}$ $\{v_j : j \in [\ell]\}$, such that :

$$\begin{cases} \nabla f(\bar{x}) + \sum\limits_i u_i \nabla g_i(\bar{x}) + \sum\limits_j v_j \nabla h_j(\bar{x}) = 0 \\ g_i(\bar{x}) = 0, \quad h_j(\bar{x}) \leq 0 \\ v_j \geq 0 \\ v_j h_j(\bar{x}) = 0 \quad \forall j \in [\ell] \end{cases}$$

b) Sufficient Part : under convexity condition of f, g_i, h_j, & there exists $\{u_i : i \in [m]\}$ $\{v_j : j \in [\ell]\}$ s.t KKT Condition holds at $\bar{x} \in \mathbb{R}^n$.

then $\bar{x}$ is a global minimizer

② From the duality perspective :

Consider Lagrangian $L(x; u, v)$

$$:= f(x) + \sum_i u_i g_i(x) + \sum_j v_j h_j(x)$$

Lagrangian Dual function $\Theta(u,v)$

$$\Theta(u,v) := \inf_{x \in X} L(x; u, v)$$

$\boxed{\text{concave func}}$

Note : $\Theta(u,v) = \inf_x L(x; u,v)$

$\leq L(\bar{x}; u,v)$ → $\boxed{\text{weak duality}}$

$\leq f(\bar{x})$ $\boxed{(D) \leq (P)}$

$=$ primal objective

generally holds

Thus, our interest is $\begin{cases} \max_{u,v} \Theta(u,v) \\ \text{s.t } v_j \geq 0 \quad j \in [\ell] \end{cases} \leq f(\bar{x})$

↓

the largest lower bound of primal problem

(Regularity)

Note : under Sufficient Condition (e.g., Slater's condition + convexity)

Strong Duality holds !

That means, there exists $\begin{cases} \hat{x} \in \text{Primal Solution} \\ (\hat{u}, \hat{v}) \in \text{Dual Solution} \end{cases}$

such that $\boxed{\Theta(\hat{u}, \hat{v}) = f(\hat{x})}$ → $\boxed{(D) = (P)}$

Rmk: If Strong Duality holds, then

① directly, it implies the $\left\{ \begin{array}{l} \text{primal solution } \hat{x} \\ \text{dual solution } (\hat{u}, \hat{v}) \end{array} \right.$ will satisfy

KKT condition system (necessary condition)

② if the program is convex program,

then it will imply:

$\{\hat{x} \,;\, \hat{u}, \hat{v}\}$ is the primal, dual solution

$\Leftrightarrow$ $\{\hat{x} \,;\, \hat{u}, \hat{v}\}$ satisfy KKT condition system

---

5. Dual of Hard / Soft Margin SVM

a) Conclusion:

$\boxed{\text{hard-margin}}$

$$\begin{cases} \max\limits_{\alpha} \; -\frac{1}{2} \sum\limits_{i}^{n} \sum\limits_{j}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum\limits_{i=1}^{n} \alpha_i \\ \text{s.t} \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \qquad \alpha_i \geq 0 \quad i \in [n] \end{cases}$$

$\boxed{\text{soft-margin}}$

$$\begin{cases} \max\limits_{\alpha} \; -\frac{1}{2} \sum\limits_{i}^{n} \sum\limits_{j}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum\limits_{i=1}^{n} \alpha_i \\ \text{s.t} \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \qquad 0 \leq \alpha_i \leq C \quad i \in [n] \end{cases}$$

b) derivation for soft-margin case

$\boxed{\text{Primal Formulation}}$ $\begin{cases} \min\limits_{w, b, \xi} \; \frac{1}{2} w^T w + C \sum\limits_{i=1}^{n} \xi_i \\ \text{s.t} \quad y_i (w^T x_i + b) \geq 1 - \xi_i \quad i \in [n] \\ \qquad \xi_i \geq 0 \quad i \in [n] \end{cases}$

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

① Stationary

$$\begin{cases} \nabla_w L = w - \sum_{i=1}^{n} \alpha_i y_i x_i \\ \nabla_b L = \sum_{i=1}^{n} \alpha_i y_i \\ \nabla_{\xi_i} L = C - \alpha_i - \beta_i \end{cases}$$

② Feasibility

③ Complementary Condition $\begin{cases} \alpha_i (1 - \xi_i - y_i (w^T x_i + b)) = 0 \\ \beta_i \xi_i = 0 \end{cases}$

$\rightarrow$ KKT optimality $\rightarrow$ $\boxed{\text{for solution analysis}}$

$$\Theta(\alpha, \beta) = \min_{w, b, \xi} L(w, b, \xi; \alpha, \beta)$$

$$= \begin{cases} \frac{1}{2} \hat{w}^T \hat{w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i x_i^T \hat{w}, & \begin{cases} \sum \alpha_i y_i = 0 \\ \alpha_i + \beta_i = C \end{cases} \\ -\infty, & \text{otherwise} \end{cases}$$

$$\Rightarrow \begin{cases} \max_{\alpha, \beta} & \Theta(\alpha, \beta) \\ \text{s.t} & \alpha \geq 0, \ \beta \geq 0 \end{cases} \Leftrightarrow \begin{cases} \max_{\alpha, \beta} & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i \\ \text{s.t} & \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \alpha \geq 0, \ \beta \geq 0 \\ & \alpha_i + \beta_i = C \qquad i \in [n] \end{cases}$$

$$\Leftrightarrow \begin{cases} \max_{\alpha} & -\frac{1}{2} \sum_{i} \sum_{j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i \\ \text{s.t} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \qquad i \in [n] \end{cases}$$

#

Rmk: (Dual Solution Analysis)

From KKT Condition, we are able to reconstruct $(\hat{w}, \hat{b})$ from the $\boxed{\text{dual solution}}$ $\hat{\alpha}$

① $\hat{w} = \sum_i \alpha_i y_i x_i$

② to derive $\hat{b}$, we have to determine:

$\rightarrow$ index sv such that $0 < \alpha_{sv} < C$

$\rightarrow$ it forces $\begin{cases} \xi_{sv} = 0 \\ 1 - \xi_{sv} - y_{sv} (\hat{w}^T x_{sv} + \hat{b}) = 0 \end{cases}$

$$\Rightarrow y_{sv}(\hat{w}^T x_{sv} + \hat{b}) = 1$$

$$\Rightarrow \hat{w}^T x_{sv} + \hat{b} = y_{sv}$$

$$\Rightarrow \hat{b} = y_{sv} - \sum_{i=1}^{n} \hat{\alpha}_i y_i \langle x_i, x_{sv} \rangle$$

$$= y_{sv} - \sum_{i \in SV} \hat{\alpha}_i y_i \langle x_i, x_{sv} \rangle$$

c) dual $\Longrightarrow$ kernel

just by replace all $\langle x_i, x_j \rangle$ by $k(x_i, x_j)$

$$= \langle \phi(x_i), \phi(x_j) \rangle$$

---

6. Re-think Soft-Margin SVM

$\longrightarrow$ Primal Formulation

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ \text{s.t} & y_i (w^T x_i + b) \geq 1 - \xi_i \quad i \in [n] \\ & \xi_i \geq 0 \quad i \in [n] \end{cases}$$

$$\Rightarrow \begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ \text{s.t} & \xi_i \geq \max\{0, 1 - y_i (w^T x_i + b)\} \quad i \in [n] \end{cases}$$

$$\downarrow \begin{cases} \min\limits_{w,b,\xi} & f(w,b) + C \sum_{i=1}^{n} \xi_i \\ \text{s.t} & \xi_i \geq g_i(w,b) \quad i \in [n] \end{cases}$$

$$= \min_{w,b} f(w,b) + C \sum_{i=1}^{n} g_i(w,b)$$

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

regularization

loss term

$$\Rightarrow \boxed{\min_{w,b} \sum_{i=1}^{n} \text{Loss}_{\text{hinge}}(z_i) + \frac{1}{2C}\|w\|_2^2} \rightarrow \boxed{\text{soft-margin SVM}}$$

$$\begin{cases} z_i = y_i(w^T x_i + b) \rightarrow \underline{\text{agreement}} \\ \text{Loss}_{\text{hinge}}(z) = \max\{0, 1 - y_i(w^T x_i + b)\} \end{cases}$$

Re-cap: what is Logistic Regression?

$$\boxed{\begin{aligned} \hat{y}_i &= \text{sigmoid}(w^T x + b) \\ &= \frac{1}{1 + \exp(-(w^T x + b))} \end{aligned}}$$

→ optimization:

cross-entropy loss

$\text{Loss}(w, b)$

$y_i \in \{0,1\}$

$$= -\sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+\exp(-w^T x_i - b)}\right) + (1-y_i)\log\left(\frac{\exp(-w^T x_i - b)}{1+\exp(-w^T x_i - b)}\right) \right]$$

$$= -\sum_{i=1}^{n} \left[\log\left(\frac{1}{1+\exp(w^T x_i + b)}\right) + y_i(w^T x_i + b) \right]$$

$$= -\sum_{i=1}^{n} \log\left(\frac{1}{1+\exp(-\tilde{y}_i(w^T x_i + b))}\right) \qquad \tilde{y}_i \in \{-1, 1\}$$

$$= \sum_{i=1}^{n} \log\left(1 + \exp(-\tilde{y}_i(w^T x_i + b))\right)$$

$$:= \sum_{i=1}^{n} \text{Loss}_{\text{logistic}}(z_i) \qquad \begin{cases} z_i = \hat{y}_i(w^T x_i + b) \\ \text{Loss}_{\text{logistic}}(z) = \log(1 + \exp(-z)) \end{cases}$$

7. Optimization for Dual Problem

→ $\boxed{\text{Block Coordinate Descent}}$

$$\min_{x \in \mathbb{R}^n} \quad f(x_1, \ldots, x_\ell)$$
$$\text{s.t} \quad x \in \mathcal{X}$$

here $\sum_{j=1}^{\ell} n_j = n$

$x_j \in \mathbb{R}^{n_j}$

$\boxed{\text{Algo}}$:

1) set $x^{(0)}$
2) for $j = 1, 2, \ldots, \ell$

$$x_j^{(k+1)} = \operatorname*{argmin}_{x_j} f(\underline{x_{<j}^{(k+1)}}, \underline{x_j}, \underline{x_{>j}^{(k)}})$$

3) $k \to k+1$.

Apply to $\boxed{\text{Dual SVM}}$ ⟹ <u>SMO Algo</u>

<u>Recap</u> : $\boxed{\text{Dual SVM}}$ ⟹

$$\begin{cases} \min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K_{ij} - \sum_{i=1}^{n} \alpha_i \\ \text{s.t} \quad 0 \le \alpha_i \le C \\ \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

<u>SMO Algo</u> :

1) <u>initialize</u> α
2) choose (i,j) pairs in $[n]$.
3)

$$(\hat{\alpha}_i, \hat{\alpha}_j) = \begin{cases} \min_{\alpha} \frac{1}{2} K_{ii} \alpha_i^2 + \frac{1}{2} K_{jj} \alpha_j^2 - y_i y_j K_{ij} \alpha_i \alpha_j \\ \qquad - \alpha_i - \alpha_j \\ \text{s.t} \quad 0 \le \alpha_i \le C, \quad 0 \le \alpha_j \le C \\ \qquad \alpha_i y_i + \alpha_j y_j = \zeta \end{cases}$$

$$(\alpha_i, \alpha_j) \leftarrow (\hat{\alpha}_i, \hat{\alpha}_j)$$

4) <u>Repeat</u>