Support Vector Machine

2021年5月23日 0:48

P28 Hard-Margin SVM – Model Defination

SVM three features : Margin ; Duality ; kernel

① hard-margin SVM ② Soft-margin SVM ③ Kernel SVM

$f(w) = \mathrm{sign}(w^T x + b)$ Let distance (w, b, x_i) be the distance of x_i to hyperplane $w^T x + b$

$$distance(w, b, x_i) = \frac{1}{\|w\|} |w^T x_i + b|$$

Maximum interval classifier
$$\begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmax}}\ margin(w,b) \\ margin(w,b) = \underset{\substack{x_i \\ i=1\cdots N}}{\min}\ distance(w,b,x_i) \Rightarrow \text{The distance from the nearest point to the hyperplane among all points} \\ s.t.\ y_i(w^T x_i + b) > 0 \quad i = 1 \cdots N \end{cases}$$

$$\Rightarrow \begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmax}} \frac{1}{\|w\|} \underset{\substack{x_i \\ i=1\cdots N}}{\min} y_i(w^T x_i + b) \\ s.t.\ y_i(w^T x_i + b) > 0 \Rightarrow \exists\, r > 0,\ s.t.\ \underset{\substack{x_i, y_i \\ i=1\cdots N}}{\min} y_i(w^T x_i + b) = r \end{cases}$$

$\because 2w^T x + 2b$ and $w^T x + b$ are the same plane $\quad \therefore r$ can take any value. After r is fixed, w, b are also determined.

设 $r = 1 \Rightarrow \begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmax}} \frac{1}{\|w\|} \\ s.t.\ \underset{\substack{x_i, y_i \\ i=1\cdots N}}{\min} y_i(w^T x_i + b) = 1 \end{cases} \Rightarrow \begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmin}} \frac{1}{2} w^T w \\ s.t.\ y_i(w^T x_i + b) \geq 1 \quad, i = 1 \cdots N \end{cases}$

P29 Model Solution, Duality Problem

$$\mathcal{L}(w,b,\lambda) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i [1 - y_i(w^T x_i + b)] \Rightarrow \begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\min} \underset{\lambda}{\max} \mathcal{L}(w,b,\lambda) \\ s.t.\ \lambda_i \geq 0 \end{cases}$$

Lagrange Multiplier Method hides the $y_i(w^T x_i + b) \geq 1$ condition in the filtering process of finding minimum

$$\Rightarrow \begin{cases} \text{If } 1 - y_i(w^T x_i + b) > 0,\ \underset{\lambda}{\max} \mathcal{L}(w,b,\lambda) = \frac{1}{2} w^T w + \infty = \infty \\ \text{If } 1 - y_i(w^T x_i + b) \leq 0,\ \underset{\lambda}{\max} \mathcal{L}(w,b,\lambda) = \frac{1}{2} w^T w \end{cases}$$

$$\therefore \underset{w,b}{\min} \underset{\lambda}{\max} \mathcal{L}(w,b,\lambda) = \underset{w,b}{\min} (\infty, \frac{1}{2} w^T w) = \underset{w,b}{\min} \frac{1}{2} w^T w$$

Duality Problem : Weak Duality : $\min\max f \geq \max\min f$

Strong Duality : $\min\max f = \max\min f$

$$\therefore \begin{cases} \hat{\lambda} = \underset{\lambda}{\max} \underset{w,b}{\min} \mathcal{L}(w,b,\lambda) \\ s.t.\ \lambda_i \geq 0 \quad i = 1 \cdots N \end{cases}$$

$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0$ 代入 $\mathcal{L}(w,b,\lambda) \Rightarrow$ Find it with support vectors in the next section.

$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0 \quad \hat{w} = \sum_{i=1}^{N} \lambda_i y_i x_i$ 代入 $\mathcal{L}(w,b,\lambda)$

$$\begin{aligned} \therefore \mathcal{L}(w,b,\lambda) &= \frac{1}{2} \left(\sum_{i=1}^{N} \lambda_i y_i x_i\right)^T \left(\sum_{j=1}^{N} \lambda_j y_j x_j\right) + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j x_j\right)^T x_i \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_j^T x_i \\ &= \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j \end{aligned}$$

$$\therefore \begin{cases} \hat{\lambda} = \underset{\lambda}{\mathrm{argmin}} \left(\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_{i=1}^{N} \lambda_i\right) \\ s.t.\ \lambda_i \geq 0,\ \sum_{i=1}^{N} \lambda_i y_i = 0 \quad i = 1 \cdots N \end{cases}$$

P30 Model Solution, KKT Condition (Karush-Kuhn-Tucker condition)

KKT condition ? :
$$\begin{cases} \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \frac{\partial \mathcal{L}}{\partial w} = 0 \\ \lambda_i [1 - y_i(w^T x_i + b)] = 0 \longrightarrow \text{Complementary Slackness (互补松弛)} \\ \lambda_i \geq 0,\ 1 - y_i(w^T x_i + b) \leq 0 \quad \text{Support Vectors} : \lambda \neq 0 \text{ ; Non-support Vectors} : \lambda = 0 \end{cases}$$

Find $\hat{b}$: $\exists (x_k, y_k)$, s.t. $1 - y_k(w^T x_k + b) = 0$ (i.e. support vectors)

$\therefore y_k(w^T x_k + b) = 1 \Rightarrow y_k^2(w^T x_k + b) = y_k \quad \because y_k = \pm 1 \quad \therefore w^T x_k + b = y_k$

$\therefore \hat{w} = \sum_{i=1}^{N} \lambda_i y_i x_i \quad \therefore \hat{b} = y_k - \sum_{i=1}^{N} \lambda_i y_i x_i^T x_k$

P31 Soft-margin SVM

Allow classifer to make mistakes : $\hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmin}} \frac{1}{2} w^T w + loss$

① Count : $loss = \sum_{i=1}^{N} I\{y_i(w^T x_i + b) < 1\}$ But the function is't continuous

② Distance : If $y_i(w^T x_i + b) \geq 1$, $loss_i = 0$; If $y_i(w^T x_i + b) < 1$, $loss_i = 1 - y_i(w^T x_i + b)$

$\therefore loss_i = \max\{0, 1 - y_i(w^T x_i + b)\}$

Simplify to $\Rightarrow \xi_i = 1 - y_i(w^T x_i + b)$, $\xi_i \geq 0$
$$\begin{cases} \hat{w}, \hat{b} = \underset{w,b}{\mathrm{argmin}} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \\ s.t.\ y_i(w^T x + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad i = 1 \cdots N \end{cases}$$

P32 Weak Duality

Generalized constrained optimization problem :
$$\begin{cases} \underset{x}{\min} f(x) \\ s.t.\ m_i(x) < 0,\ n_j(x) = 0 \quad i = 1 \cdots M,\ j = 1 \cdots N \end{cases}$$

$$\mathcal{L}(x,\lambda,\eta) = f(x) + \sum_{i=1}^{M} \lambda_i m_i(x) + \sum_{j=1}^{N} \eta_j n_j(x) \Rightarrow \begin{cases} \underset{x}{\min} \underset{\lambda,\eta}{\max} \mathcal{L}(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0 \quad i = 1 \cdots M \end{cases}$$

If $m_i(x) > 0$, coefficient is negative

$\approx$ Unconstrained form of the original problem $\Downarrow$ In fact there are still constraints

Duality Problem $\leq$ Original Problem

If $m_i(x) \geq 0$, coefficient is negative

$\begin{cases} \\ s.t.\ \lambda_i \geq 0 \quad i=1\cdots M \end{cases}$ In fact there are still constraints

Duality Problem ≤ Original Problem

$\max\limits_{\lambda,\eta} \min\limits_{x} \mathcal{L}(x,\lambda,\eta) \leq \min\limits_{x} \max\limits_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)$

证: $\min\limits_{x} \mathcal{L}(x,\lambda,\eta) \leq \mathcal{L}(x,\lambda,\eta) \leq \max\limits_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)$

x is constant now, i.e. $A(\lambda,\eta) \leq B(x)$

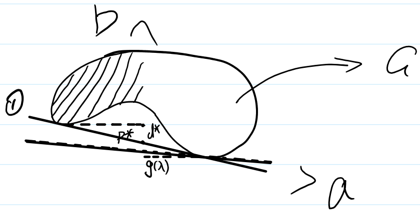
$\therefore \max A(\lambda,\eta) \leq \min B(x)$

P33 Duality Problem – Geometric Interpretation

Simplified optimization problem: $\begin{cases} \min f(x) & x \in \mathbb{R}^p \\ s.t.\ m(x) \leq 0 \end{cases} \Rightarrow \mathcal{L}(x,\lambda) = f(x) + \lambda m(x),\ \lambda \geq 0$

$p^* = \min f(x), m(x) \leq 0$ (Original problem optimum solution) $\quad d^* = \max\limits_{\lambda} \min\limits_{x} \mathcal{L}(x,\lambda)$ (Duality problem optimum solution)

$G = \{ (m(x), f(x)) \mid x \in D \}$ 设 $m(x)=a \quad f(x)=b$



$p^* = \inf\{ b \mid (a,b) \in G,\ a \leq 0 \}$ (inf is infimum (下确界))

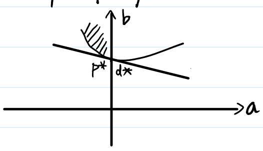
$d^* = \max\limits_{\lambda} \min\limits_{x} \mathcal{L}(x,\lambda)$ Set $g(\lambda) = \min\limits_{x} (b + \lambda a) \qquad d^* = \max\limits_{\lambda} g(\lambda)$

$g(\lambda) = \inf\{ b + \lambda a \mid (a,b) \in G \}$ i.e. the minimum of the intersection of the line $b+\lambda a = ?$ and G

d^* is the point that maximum $g(\lambda)$ when λ (slope) changes. i.e. line ①

$\therefore d^* \leq p^*$ When the part of the function G that intersects the b-axis is a convex function + slater condition

$p^* = d^*$



Slater is a sufficient non-essential condition for strong duality

P34 Slater Condition

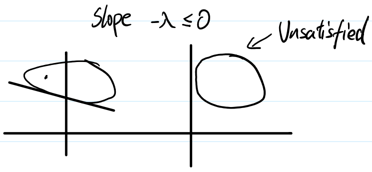
$\begin{cases} \min f(x) \\ s.t.\ m_i(x) \leq 0,\ i=1\cdots M \\ \quad n_j(x) = 0,\ j=1\cdots N \end{cases}$ $\quad D = \{ \mathrm{dom} f \cap \mathrm{dom} \bigcap\limits_{i=1}^{M} m_i \cap \bigcap\limits_{i=1}^{N} n_i \}$

Domain of definition

Slater condition:

$\exists \hat{x} \in \mathrm{relint}\ D$ The set is the inner part after removing the boundary

s.t. $\forall\ i = 1 \cdots M,\ m_i(\hat{x}) < 0 \Longrightarrow$ The set G must have points at $m_i(x) \leq 0$



① Most convex optimazation problems satisfy Slater ② Affine function no need to proof $m_i(\hat{x}) < 0$

Affine function: The polynomial function with the highest order is 1 e.g. $f(x) = Ax + b$

Affine function with a constant term of zero is linear function e.g. $f(x) = Ax$

From ②, if $f(x)$ is a convex function and m_i, n_j are affine functions, then it satisfies Slater. ∴ Satisfy strong duality

P35 KKT Condition

$\begin{cases} \min f(x) \\ s.t.\ m_i(x) \leq 0 \quad i=1\cdots N \\ \quad n_j(x) = 0 \quad j=1\cdots M \end{cases} \Rightarrow \begin{matrix} \mathcal{L}(x,\lambda,\eta) = f(x) + \sum\limits_{i=1}^{N} \lambda_i m_i(x) + \sum\limits_{j=1}^{M} \eta_j n_j(x) \\ g(\lambda,\eta) = \min\limits_{x} \mathcal{L}(x,\lambda,\eta) \end{matrix}$

Convex + Slater ⟹ Strong Duality ⟺ KKT

KKT $\begin{cases} \text{Passable condition:} \begin{cases} m_i(x^*) \leq 0 \\ n_j(x^*) = 0 \\ \lambda^* \geq 0 \end{cases} \\ \text{complementary slackness}: \lambda_i m_i = 0 \\ \text{gradient is } 0: \left. \dfrac{\partial \mathcal{L}(x,\lambda^*,\eta^*)}{\partial x} \right|_{x=x^*} = 0 \end{cases}$

optimal solution, optimal value

Set $d^* \to \lambda^*, \eta^*$

$p^* \to x^*$

It's also used in SVM:

Support vectors with $\lambda = 0$, and non-support vectors with $\lambda \neq 0$

Proof KKT by Strong Duality

$d^* \overset{①}{=} \max\limits_{\lambda,\eta} g(\lambda,\eta) \overset{①}{=} g(\lambda^*,\eta^*) = \min\limits_{x} \mathcal{L}(x,\lambda^*,\eta^*) \overset{②}{\leq} \mathcal{L}(x^*,\lambda^*,\eta^*) \geq$

$= f(x^*) + \sum\limits_{i=1}^{N} \lambda^* m_i(x) + \sum\limits_{j=1}^{M} \eta^* n_j(x) \qquad \because \lambda^* m_i(x) \leq 0 \quad n_j(x) = 0$

$\therefore$ Original function $\overset{③}{\leq} f(x^*) \overset{\text{like ①}}{=} p^*$

$d^* \overset{①}{=} \max_{\lambda,\eta} g(\lambda,\eta) \overset{①}{=} g(\lambda^*,\eta^*) = \min_x \mathcal{L}(x,\lambda^*,\eta^*) \overset{②}{\le} \mathcal{L}(x^*,\lambda^*,\eta^*) \rightleftharpoons$

$= f(x^*) + \sum_{i=1}^{N} \lambda^* m_i(x) + \sum_{j=1}^{M} \eta^* n_j(x)$ $\quad \because \lambda^* m_i(x) \le 0 \quad n_j(x) = 0$

$\therefore$ Original function $\overset{③}{\le} f(x^*) \overset{\text{like } ①}{=\!=} p^*$

$\because$ Strong Duality $\Rightarrow d^* = p^*$ $\quad \therefore$ equations at ②③ hold

$\therefore$ ② $\min_x \mathcal{L}(x,\lambda^*,\eta^*) = \mathcal{L}(x^*,\lambda^*,\eta^*)$ $\quad \therefore \mathcal{L}(x,\lambda^*,\eta^*)$ takes the minimum value at x^*, it can be proved the gradient is 0

$\therefore$ ③ $f(x^*) + \sum_{i=1}^{N} \lambda^* m_i(x) = f(x^*)$ $\quad \therefore$ Complementary slackness can be proved