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

$f(w) = sign(w^Tx+b)$ Let distance (w,b,x_i) be the distance of x_i to hyperplane w^Tx+b

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

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

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

∵ $2w^Tx+2b$ and w^Tx+b are the same plane ∴ r can take any value. After r is fixed, w,b are also determined.

如果不做限定，则w可以任意缩放，无限种取值。所以此处只取能使r=1的w的值

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

P29 Model Solution , Duality Problem

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

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

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

∴ $\min\limits_{w,b}\max\limits_{\lambda}\mathcal{L}(w,b,\lambda) = \min\limits_{w,b}(\infty, \frac{1}{2}w^Tw) = \min\limits_{w,b}\frac{1}{2}w^Tw$

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

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

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

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

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

∴ $\mathcal{L}(w,b,\lambda) = \frac{1}{2}\left(\sum_{i=1}^{N}\lambda_iy_ix_i\right)^T\left(\sum_{j=1}^{N}\lambda_jy_jx_j\right) + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_iy_i\left(\sum_{j=1}^{N}\lambda_jy_jx_j\right)^Tx_i$

$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_j^Tx_i$

$= \sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j$

∴ $\begin{cases} \hat{\lambda} = \underset{\lambda}{argmin}\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j - \sum_{i=1}^{N}\lambda_i\right) \\ s.t.\ \lambda_i\geq 0,\ \sum_{i=1}^{N}\lambda_iy_i = 0 \quad i=1\cdots N \end{cases}$

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

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

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

∴ $y_k(w^Tx_k+b)=1 \Rightarrow y_k^2(w^Tx_k+b)=y_k$ ∵ $y_k=\pm 1$ ∴ $w^Tx_k+b=y_k$

∴ $\hat{w} = \sum_{i=1}^{N}\lambda_iy_ix_i$ ∴ $\hat{b} = y_k - \sum_{i=1}^{N}\lambda_iy_ix_ix_k$

P31 Soft-margin SVM

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

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

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

∴ $loss_i = \max\{0, 1-y_i(w^Tx_i+b)\}$

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

P32 Weak Duality

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

$\mathcal{L}(x,\lambda,\eta) = f(x) + \sum_{i=1}^{M}\lambda_im_i(x) + \sum_{j=1}^{N}\eta_jn_j(x)$

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

$\Rightarrow \begin{cases} \min\limits_{x}\max\limits_{\lambda,\eta}\mathcal{L}(x,\lambda,\eta) \approx \text{Unconstrained form of the original problem} \\ s.t.\ \lambda_i\geq 0 \quad i=1\cdots M \quad \text{In fact there are still constraints} \end{cases}$

Duality Problem ≤ Original Problem

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

$\begin{cases} \min_x \mathcal{L}(x,\lambda,\eta) \sim \text{unconstrained form of the original problem} \\ \text{s.t. } \lambda_i \geq 0 \quad i=1\cdots M \end{cases}$ ↓ In fact there are still constraints

Duality Problem ≤ Original Problem

$\max_{\lambda,\eta} \min_x \mathcal{L}(x,\lambda,\eta) \leq \min_x \max_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)$

即: $\min_x \mathcal{L}(x,\lambda,\eta) \leq \mathcal{L}(x,\lambda,\eta) \leq \max_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)$

$\min_x \mathcal{L}(x,\lambda,\eta)$: 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: $\begin{cases} \min f(x) \quad x \in \mathbb{R}^n \\ \text{s.t. } m(x) \leq 0 \end{cases} \Rightarrow \mathcal{L}(x,\lambda) = f(x) + \lambda m(x), \ \lambda \geq 0$

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

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

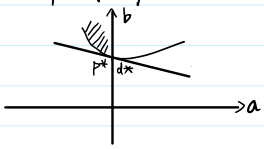
$p^* = \inf\{b \mid (a,b) \in G, a \leq 0\}$ (inf is infimum 下确界)

$d^* = \max_\lambda \min_x \mathcal{L}(x,\lambda)$ Set $g(\lambda) = \min_x (b+\lambda a) \quad d^* = \max_\lambda g(\lambda)$

$g(\lambda) = \inf\{b+\lambda a \mid (a,b) \in G\}$ 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 interserts the b-axis is a convex function + slater condition



$p^* = d^*$

Slater is a sufficient non-essential condition for strong duality

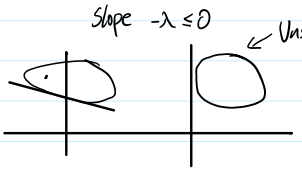
P34 Slater Condition

$\begin{cases} \min f(x) \\ \text{s.t. } m_i(x) \leq 0, \ i=1\cdots M \\ n_j(x) = 0, \ j=1\cdots N \end{cases} \quad D = \{\text{dom} f \cap \text{dom} \bigcap_{i=1}^M m_i \cap \bigcap_{i=1}^N n_i\}$ ← Domain of definition

Slater condition:

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

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



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

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

Convex + Slater ⇒ Strong Duality ⟺ kkT

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

optimal solution ↓ $d^* \to \lambda^*, \eta^*$ optimal value

$p^* \to x^*$

It's also used in SVM!

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

Proof kkT by Strong Duality

$d^* \overset{①}{=} \max_{\lambda,\eta} g(\lambda,\eta) \overset{②}{=} g(\lambda^*,\eta^*) = \min_x \mathcal{L}(x,\lambda^*,\eta^*) \overset{③}{\leq} \mathcal{L}(x^*,\lambda^*,\eta^*) \geq$

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

∴ Original function ③ $f(x^*) \overset{\text{like } ①}{=} D^*$

$d^* \triangleq \max_{\lambda,\eta} g(\lambda,\eta) \triangleq g(\lambda^*,\eta^*) = \min_x L(x,\lambda^*,\eta^*) \leq L(x^*,\lambda^*,\eta^*)$

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

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

$\because$ Strong Duality $\Rightarrow d^* = p^* \quad \therefore$ equations at ①③ hold

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

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