Support Vector Machines

Peng Lingwei

July 30, 2019

Contents

15 Support Vector Machine **2**
15.1 MARGIN AND HARD-SVM 2
15.2 GENERALIZATION BOUNDS FOR SVM 2
15.3 SOFT-SVM AND NORM REGULARIZATION 3

15 Support Vector Machine

15.1 MARGIN AND HARD-SVM

Claim 1. *The distance between the hyperplane $\langle \vec{w}, \vec{x} \rangle + b = 0$ and the point $\vec{x}$ is*

$$\frac{|\langle \vec{w}, \vec{x} \rangle + b|}{\|\vec{w}\|}$$

Definition 1. ***(Hard -SVM rule).***

$$\arg \max_{(\vec{w},b):\|\vec{w}\|=1} \min_{i\in[m]} |\langle \vec{w}, \vec{x}_i \rangle + b| \quad s.t. \quad \forall i, y_i(\langle \vec{w}, \vec{x}_i \rangle + b) > 0$$

We can change it into

$$\min_{\vec{w}} \frac{1}{2}\|\vec{w}\|^2 \quad s.t. \quad \forall i, \quad y_i\langle \vec{w}, \vec{x}_i \rangle + b \geq 1.$$

If we add one dimension into sample space, we can use this rule

$$\min_{\vec{w}} \frac{1}{2}\|\vec{w}\|^2 \quad s.t. \quad \forall i, \quad y_i\langle \vec{w}, \vec{x}_i \rangle \geq 1.$$

The regularizing b usually does not make a significant difference to the sample complexity.

15.2 GENERALIZATION BOUNDS FOR SVM

Definition 2. ***(Loss function).*** *Let $\mathcal{H} = \{\vec{w} : \|\vec{w}\|_2 \leq B\}$, $Z = \mathcal{X} \times \mathcal{Y}$ be the examples domain. Then, the loss function: $l : \mathcal{H} \times Z \to \mathbb{R}$ is*

$$l(\vec{w}, (\vec{x}, y)) = \phi(\langle \vec{w}, \vec{x} \rangle, y) \tag{1}$$

1. *Hinge-loss function: $\phi(a, y) = \max\{0, 1 - ya\}$;*
2. *Absolute loss function: $\phi(a, y) = |a - y|$.*

Theorem 1. *Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that w.p.1 we have $\|\vec{x}\|_2 \leq R$. Let $\mathcal{H} = \{\vec{w} : \|\vec{w}\|_2 \leq B\}$ and let $l : \mathcal{H} \times Z \to \mathbb{R}$ be a loss function of the form $\phi(a, y)$ and it's a $\rho-$Lipschitz function and $\max_{a\in[-BR,BR]} |\phi(a, y)| \leq c$, so*

$$\mathbb{P}\left\{\forall \vec{w} \in \mathcal{H}, L_{\mathcal{D}}(\vec{w}) \leq L_S(\vec{w}) + \frac{2\rho BR}{\sqrt{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}}\right\} \geq 1 - \delta$$

(Chapter 26)

Theorem 2. *In Hard-SVM, we assume that $\exists \vec{w}^*$ with $\mathbb{P}_{(\vec{x},y)\sim\mathcal{D}}[y\langle \vec{w}^*, \vec{x} \rangle \geq 1] = 1$ and $\mathbb{P}\{\|\vec{x}\|_2 \leq R\} = 1$. Let the SVM rule's output is $\vec{w}_S$.*

$$\mathbb{P}\left\{L_{\mathcal{D}}^{0-1}(\vec{w}_S) \leq L_{\mathcal{D}}^{ramp}(\vec{w}_S) \leq \frac{2R\|\vec{w}^*\|_2}{\sqrt{m}} + \sqrt{\frac{2\ln(2/\delta)}{m}}\right\} \geq 1 - \delta$$

The preceding theorem depends on $\|\vec{w}^*\|_2$, which is unknow. In the following we derive a bound that depends on the norm of the output of SVM.

Theorem 3.

$$\mathbb{P}\left\{L_{\mathcal{D}}^{0-1}(\vec{w}_S) \leq \frac{4R\|\vec{w}_S\|_2}{\sqrt{m}} + \sqrt{\frac{\ln\left(4\log_2\|\vec{w}_S\|_2/\delta\right)}{m}}\right\} \geq 1-\delta \tag{2}$$

The proof is similar to the SRM.

Proof. For $i \in \mathbb{N}^+$, let $B_i = 2^i, \mathcal{H}_i = \{\vec{w} : \|\vec{w}\|_2 \leq B_i\}$, and let $\delta_i = \frac{\delta}{2i^2}$, then we have

$$\mathbb{P}\left\{\forall \vec{w} \in \mathcal{H}_i, L_{\mathcal{D}}(\vec{w}) \leq L_S(\vec{w}) + \frac{2B_iR}{\sqrt{m}} + c\sqrt{\frac{2\ln(2/\delta_i)}{m}}\right\} \geq 1-\delta_i$$

Applying the union bound and using $\sum_{i=1}^{\infty}\delta_i \leq \delta$, so the union event happens with probability of at least $1-\delta$. $\forall \vec{w}$, we let $\vec{w} \in \mathcal{H}_{\lceil\log_2(\|\vec{w}\|_2)\rceil}$. Then $B_i \leq 2\|\vec{w}\|_2$ and $\frac{2}{\delta} = \frac{(2i)^2}{\delta} \leq \frac{(4\log_2(\|\vec{w}\|_2))^2}{\delta}$. □

Theorem 4. *Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that w.p.1 we have $\|\vec{x}\|_\infty \leq R$. Let $\mathcal{H} = \left\{\vec{w} \in \mathbb{R}^d : \|\vec{w}\|_1 \leq B\right\}$ and let $l : \mathcal{H} \times Z \to \mathbb{R}$ be a loss function of the form $\phi(a, y)$ and it's a $\rho -$ Lipschitz function and $\max_{a\in[-BR,BR]}|\phi(a,y)| \leq c$, so*

$$\mathbb{P}\left\{\forall \vec{w} \in \mathcal{H}, L_{\mathcal{D}}(\vec{w}) \leq L_S(\vec{w}) + 2\rho BR\sqrt{\frac{2\log(2d)}{m}} + c\sqrt{\frac{2\ln(2/\delta)}{m}}\right\} \geq 1-\delta$$

(Also following Chapter26).

15.3 SOFT-SVM AND NORM REGULARIZATION

Definition 3. ***(Soft-SVM).***

$$\min_{\vec{w},b,\xi}\left(\lambda\|\vec{w}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i\right) \quad s.t. \quad \forall i, y_i(\langle\vec{w},\vec{x}_i\rangle) + b \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Recall the definition of the hinge loss:

$$l^{hinge}((\vec{w},b),(\vec{x},y)) = \max\{0, 1 - y(\langle\vec{w},\vec{x}\rangle + b)\}$$

Then, the Soft-SVM rule changes into:

$$\min_{\vec{w},b}\left(\lambda\|\vec{w}\|_2^2 + L_S^{hinge}((\vec{w},b))\right)$$

If considering Soft-SVM for learning a homogenous halfspace, it's convenient to optimize

$$\min_{\vec{w}}\left(\lambda\|\vec{w}\|_2^2 + L_S^{hinge}(\vec{w})\right), \quad L_S^{hinge}(\vec{w}) = \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - y\langle\vec{w},\vec{x}_i\rangle\}$$