Support Vector Machines

Peng Lingwei

August 2, 2019

Contents

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).***

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

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.

15.1.1 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 $\frac{2}{\delta} = \frac{(2i)^2}{\delta} \leq \frac{(4\log_2(\|\vec{w}\|_2))^2}{\delta}$. $\square$

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).

15.2 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

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

15.2.1 The Sample Complexity of Soft-SVM

Corollary 1. *Let* $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq \rho\}$. *Then* $L_S^{hinge}(\mathbf{w})$ *is* $\|\mathbf{x}\| - $ *Lipschitz.*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_D^{0-1}(A(S))] \leq \mathbb{E}_{S\sim\mathcal{D}^m}[L_D^{hinge}(A(S))] \leq L_D^{hinge}(\mathbf{u})+\lambda\|\mathbf{u}\|^2+\frac{2\rho^2}{\lambda m} \leq L_D^{hinge}(\mathbf{u})+\sqrt{\frac{8\rho^2 B^2}{m}}$$

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_D^{0-1}(A(S))] \leq \min_{\mathbf{w}:\|\mathbf{w}\|\leq B} L_D^{hinge}(\mathbf{u}) + \sqrt{\frac{8\rho^2 B^2}{m}}$$

15.2.2 The Ramp Loss

$$l^{ramp}(\mathbf{w},(\mathbf{x},y)) = \min\left\{1, l^{hinge}(\mathbf{w},(\mathbf{x},y))\right\}$$

15.3 IMPLEMENTING SOFT-SVM USING SGD

Algorithm 1 SGD for Solving Soft-SVM

Require: T
Ensure: $\vec{\theta}^{(1)} = \vec{0}$
  for $t = 1, \ldots, T$ **do**
    Let $\vec{w}^{(t)} = \frac{1}{\lambda t}\vec{\theta}^{(t)}$
    Uniformly choose i at random from [m]:
    $\vec{\theta}^{(t+1)} += (y_i\langle\vec{w}^{(t)}, x_i\rangle \leq 1) ? y_i\vec{x}_i : 0$
  end for.
  return $\bar{\vec{w}} = \frac{1}{T}\sum_{t=1}^{T}\vec{w}^{(t)}$

15.4 Revisit SVM

15.4.1 The optimal problem of hard-SVM

1. Original:

$$\max_{\vec{w},b}\min_{(\vec{x},y)\in S}\frac{|\langle\vec{w},\vec{x}\rangle + b|}{\|\vec{w}\|}, \quad s.t.\forall y(\vec{x},y)\in S, y(\langle\vec{w},\vec{x}\rangle + b) > 0$$

2. Equal Problem1:

$$\max_{\vec{w},b:\|\vec{w}\|=1}\min_{(\vec{x},y)\in S}|\langle\vec{w},\vec{x}\rangle + b|, \quad s.t.\forall(\vec{x},y)\in S, y(\langle\vec{w},\vec{x}\rangle + b) > 0$$

3. Equal Problem2:

$$\max_{\vec{w},b:\|\vec{w}\|=1}\min_{(\vec{x},y)\in S} y\left(\langle\vec{w},\vec{x}\rangle + b\right),$$

4. Equal Problem3:

$$\min_{\vec{w},b}\frac{1}{2}\|\vec{w}\|^2, \quad s.t.\forall(\vec{x},y)\in S, y(\langle\vec{w},\vec{x}\rangle + b) > 1$$

5. Lagrangian Problem:

$$\min_{\vec{w},b}\max_{\vec{\alpha}\succeq\vec{0}}\left(L(\vec{w},b,\vec{\alpha}) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{m}\alpha_i\left[y_i(\langle\vec{w},\vec{x}_i\rangle + b) - 1\right]\right)$$

15.4.2 Support Vector

In hard-SVM, we can guarrantees that (KKT conditions.):

1. $\forall i, \sum_{i=1}^{m} \alpha_i \left[y_i(\langle \vec{w}^*, \vec{x}_i \rangle + b^*) - 1 \right] = 0$
2. $\nabla_{\vec{w}} L(\vec{w}^*) = \vec{w}^* - \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i = 0 \Rightarrow \vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i$
3. $\nabla_b L(b^*) = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$

For α_i is 0 when x_i isn't on the bound hyperplane, so we call bound points support vector, and $\vec{w}$ is in the support vectors's linear spaces.

15.4.3 Analysis Hard-SVM Problem

$$
\begin{aligned}
&\min_{\vec{w},b} \max_{\vec{\alpha} \succeq \vec{0}} \left(L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left[y_i(\langle \vec{w}, \vec{x}_i \rangle + b) - 1 \right] \right) \\
\geq &\max_{\vec{\alpha} \succeq \vec{0}} \min_{\vec{w},b} \left(L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left[y_i(\langle \vec{w}, \vec{x}_i \rangle + b) - 1 \right] \right) \\
= &\max_{\vec{\alpha} \succeq \vec{0}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle \\
= &\max_{\vec{\alpha} \succeq \vec{0}} \langle \vec{\alpha}, \vec{1} \rangle - \frac{1}{2} \vec{\alpha}^T D_y^T X^T X D_y \vec{\alpha}, \quad s.t. \forall i \in [m], \sum_{i=1}^{m} \alpha_i y_i = 0.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\vec{\alpha} =& \left(D_y^T X^T X D_y \right)^{-1} \vec{1} \\
\vec{w} =& X D_y \vec{\alpha} \\
b =& y_i - \sum_{j=1}^{m} \alpha_j y_j \langle \vec{x}_j, \vec{x}_i \rangle \\
\|\vec{w}\|^2 =& \|X D_y \vec{\alpha}\|^2 = \vec{1}^T \left(D_y^T X^T X D_y \right)^{-1} \vec{1} = \|\vec{\alpha}\|_1
\end{aligned}
$$

15.4.4 Analysis Soft-SVM Problem

$$
\min_{\vec{w},b,\vec{\xi}} \max_{\vec{\alpha},\vec{\beta}} \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \left\{ y_i \left(\langle \vec{w}, \vec{x}_i \rangle + b \right) + \xi_i - 1 \right\} - \sum_{i=1}^{m} \beta_i \xi_i
$$

The dual problem can also be changed into

$$
\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle, \quad s.t. 0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [m]
$$

which is almost analogue to Hard-SVM.

15.5 Margin Theorem

Definition 4. ***(Confidence margin).*** $\rho_{con} = yh(\vec{x})$

Definition 5. ***(Margin loss function).***

$$l^{\rho-ramp}(h, (\vec{x}, y)) = \Phi_\rho(yh(\vec{x})), \quad \Phi_\rho(u) = \min\left(1, \max\left(0, 1 - \frac{u}{\rho}\right)\right).$$

Comparing 0–1 loss: $l^{0-1}(h, (\vec{x}, y)) = \mathbf{1}\left\{yh(x) \leq 0\right\}$