Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology
Technical University of Munich

Support Vector Machine

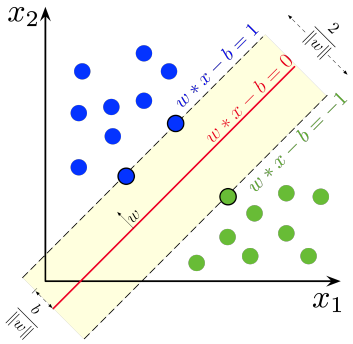
Context for this topic

- Context 1: Compare different generalisation error bounds for SVM
  - Bounds for *Hard SVM* using VC dimension, Rademacher complexity
  - Bounds for *Soft SVM* using algorithmic stability
- Context 2: Limitations for VC dimension based bounds
  - *Rademacher complexity* gives problem-dependant bounds
  - *Stability* allows analysis beyond ERM, including regularised loss minimisation

Hard SVM

$$\underset{w,b}{\text{minimise}} \ \|w\|^2$$

$$\text{s.t.} \quad y_i(\langle w, x_i \rangle + b) \geq 1 \quad \forall\ i \in [m] \qquad\qquad w, x_i \in \mathbb{R}^p$$



Is hard SVM an ERM?

Generalisation error

- Hard SVM learns from function class

$$\mathcal{H}_{lin} = \{\text{sign}(\langle w, x\rangle + b) \ : \ w \in \mathbb{R}^p, b \in \mathbb{R}\}$$

- $\text{VCdim}(\mathcal{H}) = p + 1$
- Hard SVM work for linearly separable data (realisable setting)
- What is the generalisation error w.r.t. 0-1 loss?

Generalisation error

Theorem SVM.1 (Generalisation error for hard SVM)

Let $\mathcal{D}$ be realisable w.r.t $\mathcal{H}_{lin}$

- *High probability bound: With probability $1-\delta$*

$$L_{\mathcal{D}}^{0\text{-}1}(\mathcal{A}_{\textit{hard-SVM}}(S)) \leq \sqrt{\frac{8(p+1)\ln(\frac{2me}{p+1}) + \ln(\frac{4}{\delta})}{m}}$$

- *Expected generalisation error: For a constant $C > 0$,*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}^{0\text{-}1}(\mathcal{A}_{\textit{hard-SVM}}(S))\right] \leq C\sqrt{\frac{(p+1)\ln(\frac{2me}{p+1})}{m}}$$

$L_{\mathcal{D}}^{0\text{-}1}$ denotes generalisation error w.r.t 0-1 loss

Generalisation error under margin assumption

Theorem SVM.2 (Generalisation error for hard SVM)

Assume $\mathcal{D}$ satisfies separation with (ρ, γ)-margin, that is,

- $\|x\| \leq \rho$ *for all* x
- $y(\langle w^*, x\rangle + b^*) \geq \gamma$ *for some* w^*, b^* *with* $\|w^*\| = 1$

Then, for some constant $C > 0$,

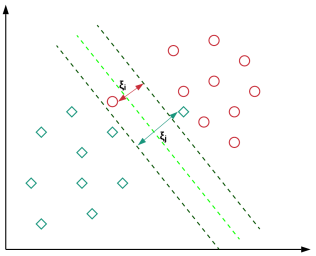
$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}^{0\text{-}1}(\mathcal{A}_{\textit{hard-SVM}}(S))\right] \leq C\frac{\rho}{\gamma}\sqrt{\frac{(1+(b^*)^2)\ln m}{m}}$$

- Proof skipped. Based on Rademacher complexity (discussed later)
- For large margin, bound better than VC / data dimension based bound

Soft SVM

$$\underset{w,b}{\text{minimise}} \ \|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \ , \quad \xi_i \geq 0 \quad \forall\ i \in [m]$$



Is soft SVM an ERM?

Soft SVM as regularised loss minimisation

- Rewrite $C = \dfrac{1}{m\lambda}$

$$\begin{aligned} \underset{w,b,\xi}{\text{minimise}} \ & \frac{1}{m}\sum_{i=1}^{m}\xi_i \ + \ \lambda\|w\|^2 \\ \text{s.t.} \ & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \forall \ i \in [m] \end{aligned}$$

- Rewrite $\xi_i = \max\{0, 1 - y_i(\langle w, x_i\rangle + b)\}$

$$\underset{w,b}{\text{minimise}} \ \frac{1}{m}\sum_{i=1}^{m} \underbrace{\max\{0, 1 - y_i(\langle w, x_i\rangle + b)\}}_{\text{hinge loss}} + \underbrace{\lambda\|w\|^2}_{\text{regulariser}}$$

Generalisation error

- Soft SVM is Tikhonov RLM with hinge loss

$$\underset{w}{\text{minimise}}\ L_S^{hinge}(w) + \lambda\|w\|^2$$

- Use generalisation error bound for Tikhonov RLM

Theorem SVM.3 (Generalisation error for Soft SVM)

Assume $\mathcal{X} = \{x \in \mathbb{R}^p \ : \ \|x\| \leq \rho\}$

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}^{\text{0-1}}(\mathcal{A}_{\textit{soft-SVM}}(S))\right] &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}^{\text{hinge}}(\mathcal{A}_{\textit{soft-SVM}}(S))\right] \\
&\leq L_{\mathcal{D}}^{\text{hinge}}(w) + \lambda\|w\|^2 + \frac{2\rho^2}{\lambda m} \qquad \textit{for all } w \\
&\leq \min_{w:\|w\|\leq B} L_{\mathcal{D}}^{\text{hinge}}(w) + \sqrt{\frac{8\rho^2B^2}{m}}\,.
\end{aligned}$$

Proof

- First inequality holds since hinge loss is a surrogate for 0-1 loss
- For second, write

$$\begin{aligned}\mathbb{E}_S\left[L_{\mathcal{D}}^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] &= \mathbb{E}_S\left[L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] \\ &\quad + \mathbb{E}_S\left[L_{\mathcal{D}}^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S)) - L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] \\ &\leq \mathbb{E}_S\left[L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] + \frac{2\rho^2}{\lambda m}\end{aligned}$$

$\ldots$ bound for Tikhonov RLM

Proof

- Since $\mathcal{A}_{\text{soft-SVM}}(S)$ is regularised hinge risk

$$\begin{aligned} L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S)) &\leq L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S)) + \lambda\|\mathcal{A}_{\text{soft-SVM}}(S)\|^2 \\ &\leq L_S^{\text{hinge}}(w) + \lambda\|w\|^2 \qquad \text{for all } w \end{aligned}$$

- Take expectation on both sides

$$\mathbb{E}_S\left[L_S^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] \leq L_{\mathcal{D}}^{\text{hinge}}(w) + \lambda\|w\|^2 \qquad \text{for all } w$$

- This gives 2^{nd} inequality

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}^{\text{hinge}}(\mathcal{A}_{\text{soft-SVM}}(S))\right] \leq L_{\mathcal{D}}^{\text{hinge}}(w) + \lambda\|w\|^2 + \frac{2\rho^2}{\lambda m} \quad \text{for all } w$$

Proof

- For 3^{rd} inequality, note

$$\begin{aligned} \min_{\lambda} \left(\lambda \|w\|^2 + \frac{2\rho^2}{\lambda m} \right) &= \sqrt{\frac{8\rho^2 \|w\|^2}{m}} \\ &\leq \sqrt{\frac{8\rho^2 B^2}{m}} \qquad \text{if } \|w\| \leq B \end{aligned}$$

- Practical choice for λ or C
  - Choose λ so that above minimum is achieved
  - If we want to restrict to $\|w\| \leq B$ (think in terms of SRM)

$$\text{choose } \lambda = \sqrt{\frac{2\rho^2}{B^2 m}} \qquad \text{or} \qquad C = \frac{1}{\lambda m} = \sqrt{\frac{B^2}{2\rho^2 m}}$$

Rademacher complexity (data dependent)

- Let $\mathcal{Z}$ be some space, and consider a finite set $Z = \{z_1, \ldots, z_m\} \subset \mathcal{Z}$
- Let $\mathcal{F}$ be class of real-valued functions defined on $\mathcal{Z}$, that is $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$
  - Example 1: $\mathcal{Z} = \mathcal{X}$ and $\mathcal{F} = \mathcal{H}$ for some hypothesis class $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$
  - Example 2: $\mathcal{Z} = \mathbb{R}^p$ and $\mathcal{F} = \{f_w(z) = \langle w, z \rangle \ : \ w \in \mathbb{R}^p\}$
  - Example 3: $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ and $\mathcal{F} = \left\{f_h(z) = f_h(x, y) = \mathbf{1}\left\{h(x) \neq y\right\} \ : \ h \in \mathcal{H}\right\}$
- Rademacher complexity of $\mathcal{F}$ with respect to set Z

$$R(\mathcal{F} \circ Z) = \underbrace{\mathbb{E}_{\sigma_1, \ldots, \sigma_m \sim_{iid} \text{Unif}\{\pm 1\}}}_{\text{Rademacher variables}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, f(z_i) \right]$$

Rademacher complexity: Intuition

$$R(\mathcal{F} \circ Z) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim_{iid} \text{Uniform}\{\pm 1\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, f(z_i) \right]$$

- Consider Example 1: $Z = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ and $\mathcal{F} = \mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$
- $\mathcal{F} \circ Z = \{(f(x_1), \ldots, f(x_m)) \ : \ f \in \mathcal{F}\} = \mathcal{H}_{|Z}$
- Think of $\sigma_1, \ldots, \sigma_m \sim_{iid} \text{Uniform}\{\pm 1\}$ as random labels
- $\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, h(z_i)$ is max alignment with $\sigma_1, \ldots, \sigma_m$ that we can get using any $h \in \mathcal{H}$
- $R(\mathcal{F} \circ Z) =$ expected maximum alignment with random labels

Rademacher complexity: Another intuition

$$R(\mathcal{F} \circ Z) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim_{iid} \text{Uniform}\{\pm 1\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

- Consider Example 3: $Z = \{(x_1, y_1), \ldots, (x_{2m}, y_{2m})\}$ and $\mathcal{F} = \left\{ f_h(z) = f_h(x, y) = \mathbf{1}\{h(x) \neq y\} \;:\; h \in \mathcal{H} \right\}$
- Split training data $Z = S \cup S'$ into $S = \{(x_i, y_i) : \sigma_i = +1\}$, $S' = \{(x_i, y_i) : \sigma_i = -1\}$
- For random $\sigma_1, \ldots, \sigma_{2m} \sim_{iid} \text{Uniform}\{\pm 1\}$, above is a random split
- $\sup_{h \in \mathcal{H}} \sum_{i \in S} \mathbf{1}\{h(x_i) \neq y_i\} - \sum_{i \in S'} \mathbf{1}\{h(x_i) \neq y_i\}$ similar to $\sup_{h \in \mathcal{H}} (L_S(h) - L_{S'}(h))$
- Recall from uniform convergence proof:
  If $\sup_{h \in \mathcal{H}} (L_S(h) - L_{S'}(h))$ is small, then $\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)|$ is also small

Rademacher complexity (distribution dependent)

$$R(\mathcal{F} \circ Z) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim_{iid} \text{Uniform}\{\pm 1\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, f(z_i) \right]$$

- Above definition depends on data $Z = \{z_1, \ldots, z_m\} \subset \mathcal{Z}$
- Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$, and $z_1, \ldots, z_m \sim_{iid} \mathcal{D}$, that is, $Z \sim \mathcal{D}^m$
- Rademacher complexity of $\mathcal{F}$ with respect to set $\mathcal{D}$

$$R_{\mathcal{D},m}(\mathcal{F}) = \mathbb{E}_{Z \sim \mathcal{D}^m}[R(\mathcal{F} \circ Z)]$$

Generalisation error bound (proof skipped)

Theorem SVM.4 (Generalisation bound using Rademacher complexity)

Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class (not necessarily binary)

Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss that satisfies $|\ell(h(x), y)| \leq c$ for all $h \in \mathcal{H}, (x, y) \in \mathcal{X} \times \mathcal{Y}$

Define $\mathcal{F} = \{\ell(h(x), y) \ : \ h \in \mathcal{H}\}$. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ and $\delta \in (0, 1)$,

- *with probability $1 - \delta$ over training samples $S \sim \mathcal{D}^m$,*

$$\sup_{h \in \mathcal{H}} \big(L_{\mathcal{D}}(h) - L_S(h)\big) \ \leq \ 2R_{\mathcal{D},m}(\mathcal{F}) + c\sqrt{\frac{2\ln(\frac{2}{\delta})}{m}}$$

- *with probability $1 - \delta$ over training samples $S \sim \mathcal{D}^m$,*

$$\sup_{h \in \mathcal{H}} \big(L_{\mathcal{D}}(h) - L_S(h)\big) \ \leq \ 2R(\mathcal{F} \circ S) + 4c\sqrt{\frac{2\ln(\frac{4}{\delta})}{m}}$$

Generalisation error bound: Discussions

- Rademacher complexity based bounds are applicable for all setting/loss (unlike VC)
- 2nd bound is practical since one can compute the bound from training data S
- Rademacher bounds could be more tight than VC/growth function based bounds, since for $\mathcal{F} = \{\mathbf{1}\{h(x) \neq y\} \ : \ h \in \mathcal{H}\}$,

$$\text{for every } \mathcal{D}, \qquad R_{\mathcal{D},m}(\mathcal{F}) \leq \sqrt{\frac{2\ln(\tau_{\mathcal{H}}(m))}{m}} \leq \sqrt{\frac{2d\ln(\frac{em}{d})}{m}}$$

  - Above follows from Massart's lemma:
    For any finite set $A \subset \mathbb{R}^m$, $\mathbb{E}_{\boldsymbol{\sigma} \sim \text{Uniform}(\{\pm 1\}^m)} \left[\max_{\mathbf{a} \in A} \boldsymbol{\sigma}^\top \mathbf{a} \right] \leq \frac{\sqrt{2\ln(|A|)}}{m} \cdot \max_{\mathbf{z} \in A} \|\mathbf{a}\|_2$

(how?)

Rademacher complexity for linear hypothesis class

Theorem SVM.5 (Rademacher complexity for linear hypothesis class)

Let $\mathcal{X} = \{x \in \mathbb{R}^p \ : \ \|x\|_2 \leq \rho\}$ *and* $\mathcal{F} = \{f_w(x) = \langle w, x\rangle \ : \ \|w\|_2 \leq B\}$

For any $X = \{x_1, \ldots, x_m\} \subset \mathcal{X}$, $R(\mathcal{F} \circ X) \leq \dfrac{B\rho}{\sqrt{m}}$

Implication:

- Assume there is w^* such that $\|w^*\|_2 = 1$ and $y\langle w^*, x\rangle \geq \gamma$ (linearly separable with margin γ)
- Hard SVM can be rephrased as finding w such that $\|w\|_2 \leq 1/\gamma$ and $y_i\langle w, x_i\rangle \geq 1 \ \forall i$
- Hence, $R(\mathcal{F} \circ X) \leq \dfrac{\rho}{\gamma\sqrt{m}}$ and $L_{\mathcal{D}}(h_{SVM}) = O\left(\dfrac{\rho}{\gamma\sqrt{m}} + \sqrt{\dfrac{\ln(\frac{4}{\delta})}{m}}\right)$,with prob. $1 - \delta$

Proof: Bound on Rademacher complexity

$$
\begin{aligned}
R(\mathcal{F} \circ X) &= \frac{1}{m} \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim \text{Uniform}\{\pm 1\}} \left[\sup_{w:\|w\|_2 \leq B} \sum_{i=1}^{m} \sigma_i \langle w, x_i \rangle \right] \\
&= \frac{1}{m} \mathbb{E}_{\sigma_1,\ldots,\sigma_m} \left[\sup_{w:\|w\|_2 \leq B} \left\langle w, \sum_{i=1}^{m} \sigma_i x_i \right\rangle \right] \\
&\leq \frac{1}{m} \mathbb{E}_{\sigma_1,\ldots,\sigma_m} \left[\sup_{w:\|w\|_2 \leq B} \|w\|_2 \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_2 \right] && \text{Cauchy-Schwarz inequality} \\
&\leq \frac{B}{m} \mathbb{E}_{\sigma_1,\ldots,\sigma_m} \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_2 \\
&\leq \frac{B}{m} \left(\mathbb{E}_{\sigma_1,\ldots,\sigma_m} \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_2^2 \right)^{1/2} && \text{Jensen's inequality}
\end{aligned}
$$

Proof (contd.)

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim \text{Uniform}\{\pm 1\}} \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|_2^2 = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim \text{Uniform}\{\pm 1\}} \sum_{i,j=1}^{m} \sigma_i \sigma_j \langle x_i, x_j \rangle$$

cross terms are 0 as σ_i's are independent, zero mean

$$= \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim \text{Uniform}\{\pm 1\}} \sum_{i=1}^{m} \sigma_i^2 \|x_i\|_2^2$$

$$= \sum_{i=1}^{m} \|x_i\|_2^2$$

$$\leq m\rho^2$$

Hence, $R(\mathcal{F} \circ X) \leq \frac{B}{m}(m\rho^2)^{1/2} = \frac{B\rho}{\sqrt{m}}$