CSC 588: Machine learning theory **Spring 2021**

Lecture 14: Support Vector Machine (SVM)

Lecturer: Chicheng Zhang *Scribe: Ruby Abrams*

Margin-based generalization error bounds for SVMs

Theorem 1. *(The More Abstract Version) Suppose $\mathcal{D}$ is supported on $\{x \in \mathbb{R}^d : ||x||_\infty \le R_\infty\} \times \{\pm 1\}$. Fix the margin value $\theta \in (0, B_1 R_\infty]$. Then with probability $1-\delta$ over m samples in S, for any predictor w such that $||w||_1 \le B_1$,*

$$\mathbb{P}_{\mathcal{D}}(y\langle w, x\rangle \le 0) \le \mathbb{P}_S(y\langle w, x\rangle \le \theta) + \mathcal{O}\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln(d/\delta)}{m}}\right)$$

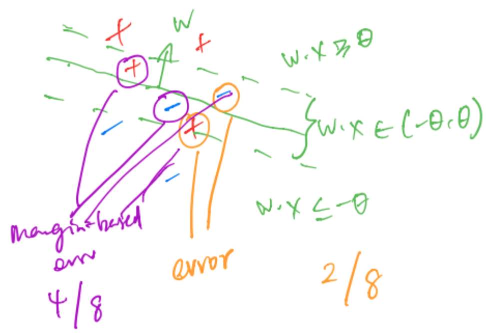

Figure 1: Illustration of binary classifier finding the line of best fit to separate data points + and −. For the given line, the classification error is 2/8.

Theorem 2. *Define the family of loss functions $\mathcal{F}$ by*

$$\mathcal{F} = \{l_{\theta,w} : ||w||_1 \le B_1\}$$

where $l_{\theta,w} = \phi_\theta(y\langle w, x\rangle)$ is the ramp loss function (see figure below). Then

$$\mathrm{Rad}_n(\mathcal{F}) \le \mathcal{O}\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d}{m}}\right)$$

where Rad_n is the Rademacher complexity.

Proof. Let's use some intuition:

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^m} \mathrm{Rad}_S(\mathcal{F})$$

By definition of Rademacher complexity

$$\begin{aligned}\mathrm{Rad}_S(\mathcal{F}) &= \mathbb{E}_{\sigma\sim U(\pm 1)^m} \frac{1}{m} \sup_{f\in\mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i, y_i) \\ &= \frac{1}{m}\mathbb{E}_{\sigma\sim U(\pm 1)^m} \sup_{w:||w||_1\le B_1} \sum_{i=1}^m \sigma_i \phi_\theta(y_i\langle w, x\rangle)\end{aligned}$$

The first step is to use the contraction inequality to "remove" ϕ_θ.

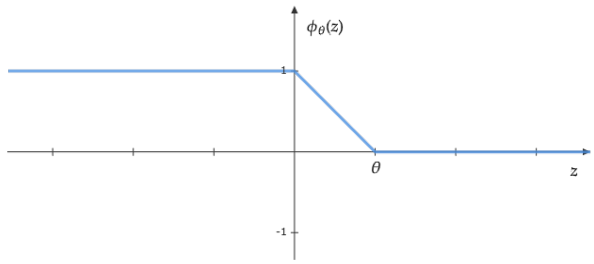

Figure 2: The Ramp Loss function. The ramp has slope $-1/\theta$. Note this is a Lipschitz function with Lipschitz constant $1/\theta$.

Lemma 3 (Contraction Inequality)**.** *Suppose $S = \{z_1, ..., z_m\}$, $\mathcal{G}$ is a function class, and ϕ is a Lipschitz function ($\forall a, b, |\phi(a) - \phi(b)| \leq L|a-b|$ with Lipschitz constant L). If we define $\mathcal{F}$*

$$\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$$

Then

$$\mathrm{Rad}_S(\mathcal{F}) \leq L\,\mathrm{Rad}_S(\mathcal{G}).$$

Applying the contraction inequality to $\mathcal{G}$

$$\mathcal{G} = \{m_w : ||w||_1 \leq B_1\}$$

where $m_w(x, y) = y\langle w, x\rangle$. Choose $\phi = \phi_\theta$ as defined by the ramp loss function and define the class of functions

$$\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$$

We obtain

$$\mathrm{Rad}_S(\mathcal{F}) \leq L_{\phi_\theta}\,\mathrm{Rad}_S(\mathcal{G})$$

with the Lipschitz constant $L_{\phi_\theta} = 1/\theta$. Now to bound $\mathrm{Rad}_S(\mathcal{G})$

$$\mathrm{Rad}_S(\mathcal{G}) = \frac{1}{m}\,\mathbb{E}_\sigma \sup_{||w||_1 \leq B_1} \sum_{i=1}^m \sigma_i y_i \langle w, x_i\rangle \tag{1}$$

$$= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{||w||_1 \leq B_1} \sum_{i=1}^m \sigma_i \langle w, x_i\rangle \tag{2}$$

$$= \frac{1}{m}\,\mathbb{E}_\sigma \sup_{||w||_1 \leq B_1} \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle \tag{3}$$

The second equality is due to the fact that σ_i is equivalent in distribution to $\sigma_i y_i$. i.e. $(\sigma_1, ..., \sigma_m) =^d (\sigma_1 y_1, ..., \sigma_m y_m)$. The third equality is by linearity of expectation. To bound this last term, we briefly discuss Hölder's Inequality. Note the following fact: given $\beta = (\beta_1, ..., \beta_d)$,

$$\max_{\alpha : ||\alpha||_1 \leq A} \langle \alpha, \beta \rangle = A||\beta||_\infty$$

$$\max_{\alpha : ||\alpha||_2 \leq A} \langle \alpha, \beta \rangle = A||\beta||_2$$

These are particular consequences of Hölder's Inequality for conjugate pairs (p, q) that satisfy $\frac{1}{p} + \frac{1}{q} = 1$. The second case above, $p = 2, q = 2$, is also known as the Cauchy-Schwartz inequality. We prove the first statement.

Proof. First we show that $A||\beta||_\infty$ is an upper bound. Suppose

$$\forall \alpha, \sum_i |\alpha_i| \leq A$$

Then

$$\begin{aligned}
\langle \alpha, \beta \rangle &= \sum_i \alpha_i \beta_i \\
&\leq \sum_i |\alpha_i||\beta_i| \\
&\leq \sum_i |\alpha_i| \max_i |\beta_i| \\
&= \sum_i |\alpha_i| ||\beta||_\infty \\
&= ||\beta||_\infty ||\alpha||_1 \leq A||\beta||_\infty
\end{aligned}$$

Now we show that there exists a value of α, say α^* so that $\langle \beta, \alpha^* \rangle = A||\beta||_\infty$. Choose α^* as follows

$$\alpha^* = \begin{cases} Ae_{i^*} & \beta_{i^*} > 0 \\ -Ae_{i^*} & \beta_{i^*} \leq 0 \end{cases}$$

where

$$i^* = \underset{i}{\operatorname{argmax}} |\beta_i|$$

and e_i is the i^{th} standard basis vector. Then $||\alpha||_1 \leq A$ and

$$\langle \alpha^*, \beta \rangle = A|\beta_{i^*}| = A||\beta||_\infty$$

□

Continuing with our bounding of $\mathrm{Rad}_S(\mathcal{G})$, we can apply the Hölder inequality to equation (3) to obtain

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{G}) &\leq \frac{B_1}{m} \mathbb{E}_\sigma \left|\left| \sum_{i=1}^m \sigma_i x_i \right|\right|_\infty && (4) \\
&= \frac{B_1}{m} \mathbb{E}_\sigma \max \left(\max_{j=1}^d \sum_{i=1}^m \sigma_i x_{ij}, \max_{j=1}^d \sum_{i=1}^m \sigma_i (-x_{ij}) \right) && (5)
\end{aligned}$$

The above is true since $||U||_\infty = \max(u_1, -u_1, u_2, -u_2, ..., u_d, -u_d)$ and recall that the j^{th} entry of the i^{th} data point, $x_{ij} \in [-R_\infty, R_\infty]$.
Now we apply **Massart's Lemma:** If N σ^2-sub-Gaussian random variables $X_1, ..., X_N$, then

$$\mathbb{E} \max_i X_i \leq \sigma \sqrt{2 \ln N}$$

Letting $N = 2d$, $\sigma^2 = mR_\infty^2$, we can upper bound equation (5) by

$$\begin{aligned}
&\leq \frac{B_1}{m} \sqrt{mR_\infty^2 2 \ln(2d)} \\
&= B_1 R_\infty \sqrt{\frac{2 \ln(2d)}{m}}
\end{aligned}$$

Now to complete the proof of the contraction inequality. Consider the family of sets

$$\mathcal{F} = \{\phi \circ g : g \in \mathcal{G}\}$$

with ϕ being Lipschitz with Lipschitz constant L. Then we'd like to present following argument.

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{F}) &= \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^m \sigma_i \phi(g(z_i)) \\
&\leq \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} L\sigma_1 g(z_1) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) \\
&\leq \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} L\sigma_1 g(z_1) + L\sigma_2 g(z_2) + \sum_{i=3}^m \sigma_i \phi(g(z_i)) \\
&\dots \\
&\leq \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^m L\sigma_i g(z_i)
\end{aligned}$$

We prove the first inequality. Note:

$$\begin{aligned}
\mathrm{Rad}_S(\mathcal{F}) = \mathbb{E}_{\sigma_{2:n}} \Bigg[&\frac{1}{2} \sup_{g\in\mathcal{G}} \Big(\phi(g(z_1)) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) \Big) \\
&+ \frac{1}{2} \sup_{g'\in\mathcal{G}} \Big(-\phi(g'(z_1)) + \sum_{i=2}^m \sigma_i \phi(g'(z_i)) \Big) \Bigg]
\end{aligned}$$

so that

$$\mathrm{Rad}_S(\mathcal{F}) = \mathbb{E}_{\sigma_{2:n}} \frac{1}{2} \sup_{g,g'\in\mathcal{G}} \left[\phi(g(z_1)) - \phi(g'(z_1)) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) + \sum_{i=2}^m \sigma_i \phi(g'(z_i)) \right]$$

since this upper bound is symmetric with respect to g and g', we can apply the Lipschitz property of ϕ to get

$$\phi(g(z_1)) - \phi(g'(z_1)) \leq L|g(z_1) - g'(z_1)|.$$

And without loss of generality, we can consider $g(z_1) \geq g'(z_1)$

$$\begin{aligned}
&\leq \mathbb{E}_{\sigma_{2:n}} \frac{1}{2} \sup_{g,g'\in\mathcal{G}, g(z_1)\geq g'(z_1)} L(g(z_1) - g'(z_1)) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) + \sum_{i=2}^m \sigma_i \phi(g'(z_i)) \\
&= \mathbb{E}_{\sigma_{2:n}} \left[\frac{1}{2} \sup_{g\in\mathcal{G}} \Big(Lg(z_1) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) \Big) - \frac{1}{2} \sup_{g'\in\mathcal{G}} \Big(Lg'(z_1) + \sum_{i=2}^m \sigma_i \phi(g'(z_i)) \Big) \right] \\
&= \mathbb{E}_{\sigma_{2:n}} \left[\mathbb{E}_{\sigma_1} \left[\sup_{g\in\mathcal{G}} L\sigma_1 g(z_1) + \sum_{i=2}^m \sigma_i \phi(g(z_i)) \right] \right]
\end{aligned}$$

Repeated application of this procedure will allow us to obtain

$$\mathrm{Rad}_S(\mathcal{F}) \leq \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \sum_{i=1}^m L\sigma_i g(z_1)$$

□

The algorithm inspired by the margin-based generalization bound: Fix $\theta = 1$. We'd like to find a weights vector w such that

1. $\mathbb{P}_S(y\langle w, x\rangle \le 1) = 0$
2. $||w||_1$ is as small as possible

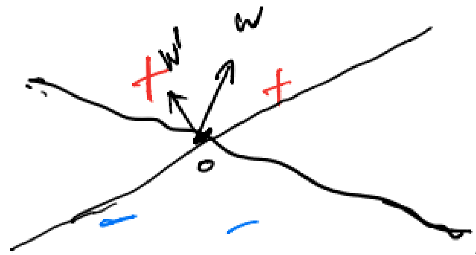

Figure 3: w and w' are possible classifications of the data. w is the better classifier because we don't need a large scaling factor to ensure that all examples have margin of error ≥ 1.

So this can be formulated as follows

$$\min ||w||_1$$

subject to

$$y_i\langle w, x_i\rangle \ge 1, \qquad \forall i \in \{1, ..., m\}$$

This is known as the l_1-Support Vector Machine (SVM). This is a convex optimization problem of the form

$$\min_x f(x)$$
$$\text{s.t. } x \in K$$

where K is a convex set. l_2-SVM is formulated as follows

$$\min_w ||w||_2$$
$$\text{s.t. } y_i\langle w, x_i\rangle \ge 1, \qquad \forall i \in \{1, ..., m\}$$

In the next class we will discuss margin-based generalization error bounds for l_2-SVMs.