FoSML: Lecture #15

- We now develop learning bounds for SVM. These provide good mathematical justification of the SVM method, even in the non-separable case.
- Recall that the VC-dimension of a family of linear hypotheses in $\mathbb{R}^D$ is $D+1$.
- Apply our learning bound based on VC-dimension immediately yields that for any linear hypothesis h, and $\forall \delta > 0$, then with probability exceeding $1-\delta$,

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2(D+1)\log\left(\frac{em}{D+1}\right)}{m}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}},$$

where $m = |S|$.

- Note that when D is large, this bound is poor, unless m is very large.
- We shall develop learning bounds that are <u>independent of D</u>, i.e. are not cursed by dimensionality.
- Our results will hold for general $\mathbb{R}$-valued functions, as opposed to classifiers outputting ± 1.

<u>Defn (margin loss function)</u>: For any $\rho > 0$, the ρ-margin loss is the function $L_\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ defined as

$$L_\rho(y, y') = \Phi_\rho(yy')$$

where $\Phi_\rho(x) = \min\left\{1, \max\left\{0, 1 - \frac{x}{\rho}\right\}\right\}$

$$= \begin{cases} 1, & \text{if } x \leq 0 \\ 1 - \frac{x}{\rho}, & \text{if } x \in (0, \rho) \\ 0, & \text{if } x \geq \rho \end{cases}$$

Defn (empirical margin loss): Given a sample $S=\{x_i\}_{i=1}^m$ and a hypothesis h, the empirical margin loss is

$$\hat{R}_{S,\rho}(h) = \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)).$$

Remark: $\forall i \in \{1,\dots,m\}$, $\Phi_\rho(y_i h(x_i)) \le \mathbb{1}_{y_i h(x_i) \le \rho}$

$$\Rightarrow \quad \hat{R}_{S,\rho}(h) \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i h(x_i)\le \rho}.$$

- Note that Φ_ρ is $\frac{1}{\rho}$ Lipschitz, unlike the 0-1 loss which isn't even continuous.

Lemma (Talagrand): Let $\{\Phi_i\}_{i=1}^m$ be L-Lipschitz functions from $\mathbb{R}\to\mathbb{R}$. Let $\{\sigma_i\}_{i=1}^m$ be Rademacher r.v. Then for any hypothesis class, H, consisting of $\mathbb{R}$-valued functions, the following inequality holds:

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left(\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right) \le \frac{L}{m}\mathop{\mathbb{E}}_{\sigma}\left(\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(x_i)\right)$$

$$= L\cdot\hat{\mathcal{R}}_S(H)$$

In particular, if $\Phi_i \equiv \Phi$ $\forall i$, then $\hat{\mathcal{R}}_S(\Phi\circ H) \le L\hat{\mathcal{R}}_S(H)$.

Proof: Fix a sample $S=\{x_i\}_{i=1}^m$. Then

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left(\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\Phi_i\circ h)(x_i)\right) = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,\dots,\sigma_{m-1}}\left(\mathop{\mathbb{E}}_{\sigma_m}\left(\sup_{h\in H}\left(u_{m-1}(h)+\sigma_m(\Phi_m\circ h)(x_m)\right)\right)\right)$$

where $u_{m-1}(h) = \sum_{i=1}^{m-1}\sigma_i(\Phi_i\circ h)(x_i)$. By the definition of the supremum,

$\forall \varepsilon > 0$, $\exists h_1, h_2 \in \mathcal{H}$ s.t.

$$u_{m-1}(h_1) + \Phi_m \circ h_1(x_1) \geq (1-\varepsilon)\left(\sup_{h\in\mathcal{H}} u_{m-1}(h) + (\Phi_m \circ h)(x_m)\right)$$

$$u_{m-1}(h_2) - \Phi_m \circ h_2(x_m) \geq (1-\varepsilon)\left(\sup_{h\in\mathcal{H}} u_{m-1}(h) - (\Phi_m \circ h)(x_m)\right).$$

Then for any $\varepsilon > 0$, by definition of $\mathop{\mathbb{E}}_{\sigma_m}$, we have

$$(1-\varepsilon)\mathop{\mathbb{E}}_{\sigma_m}\left(\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m(\Phi_m \circ h)(x_m)\right)$$

$$= (1-\varepsilon)\left[\frac{1}{2}\sup_{h\in\mathcal{H}} u_{m-1}(h) + (\Phi_m \circ h)(x_m) + \frac{1}{2}\sup_{h\in\mathcal{H}} u_{m-1}(h) - (\Phi_m \circ h)(x_m)\right]$$

$$\leq \frac{1}{2}\left[u_{m-1}(h_1) + (\Phi_m \circ h_1)(x_m) + u_{m-1}(h_2) - (\Phi_m \circ h_2)(x_m)\right] \quad (\star)$$

Letting $s = \operatorname{sgn}\left(h_1(x_m) - h_2(x_m)\right)$, $(\star)$ implies

$$(1-\varepsilon)\mathop{\mathbb{E}}_{\sigma_m}\left[\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m(\Phi_m \circ h)(x_m)\right]$$

$$\leq \frac{1}{2}\left[u_{m-1}(h_1) + u_{m-1}(h_2) + sL\left(h_1(x_m) - h_2(x_m)\right)\right]$$

$$= \frac{1}{2}\left(u_{m-1}(h_1) + sL\,h_1(x_m)\right) + \frac{1}{2}\left(u_{m-1}(h_2) - sL\,h_2(x_m)\right)$$

$$\leq \frac{1}{2}\sup_{h\in\mathcal{H}}\left(u_{m-1}(h) + sLh(x_m)\right) + \frac{1}{2}\sup_{h\in\mathcal{H}}\left(u_{m-1}(h) - sLh(x_m)\right)$$

$$= \mathop{\mathbb{E}}_{\sigma_m}\left(\sup_{h\in\mathcal{H}} u_{m-1}(h) + \sigma_m L h(x_m)\right)$$

Since this holds $\forall \varepsilon > 0$, we get

$$\mathop{\mathbb{E}}_{z_m}\left(\sup_{h\in H} a_{m-1}(h) + z_m - (\Phi_\rho \circ h)(x_m)\right) \le \mathop{\mathbb{E}}_{z_m}\left(\sup_{h\in H} a_{m-1}(h) + z_m - L\,h(x_m)\right). \quad (9)$$

The result follows by doing the same thing for the expectations over z_i, $i \le m-1$. ∎

We can use Talagrand's lemma to prove the following learning bound:

Theorem (margin bound for binary classification): Let H be a set of $\mathbb{R}$-valued functions. Fix $\rho > 0$. Then $\forall \delta > 0$, with probability exceeding $1-\delta$, both of the following hold $\forall h \in H$:

(a.) $$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathcal{R}_m(H) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

(b.) $$R(h) \le \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\hat{\mathcal{R}}_m(H) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}}$$

Proof: We begin with (a.); (b.) follows immediately by earlier results relating RC and empirical RC.

Let $\tilde{H} = \{z = (x,y) \mapsto y\,h(x)\}_{h\in H}$. Consider the following family of functions taking values in $[0,1]$:

$$H_* = \{\Phi_\rho \circ f \mid f \in \tilde{H}\}.$$

By our earlier results on RC, with probability exceeding $1-\delta$, $\forall g \in H_*$,

$$\mathbb{E}(g(z)) \le \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(H_*) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\Leftrightarrow \mathbb{E}\left(\Phi_\rho(y\,h(x))\right) \le \hat{R}_{S,\rho}(h) + 2\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}.$$

Since Φ_ρ is $\frac{1}{\rho}$-Lipschitz, Talagrand's Lemma implies

$$\mathcal{R}_m(\Phi_{\rho} \circ \tilde{H}) \leq \frac{1}{\rho} \mathcal{R}_m(\tilde{H})$$

$$= \frac{1}{\rho} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{S,\sigma} \left(\sup_{h \in H} \sum_{i=1}^{m} \sigma_i y_i h(x_i) \right)$$

$$= \frac{1}{\rho^m} \mathop{\mathbb{E}}_{S,\sigma} \left(\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right)$$

$$= \frac{1}{\rho} \mathcal{R}_m(H).$$

This gives (9.1), as needed. ∎