FoSML: Lecture 16

- Our theorem from last time gave bounds on generalization error in terms of the margin ρ. Recall:

Theorem (margin bound for binary classification): Let $\mathcal{H}$ 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 $\forall h \in \mathcal{H}$:

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

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

- Combined with the following bound on $\hat{R}_S(\mathcal{H})$ for $\mathcal{H}$ the class of linear hypotheses, we will get a generalization bound for SVM.

Theorem: Let $S \subseteq \{x \mid \|x\|_2 \leq r\}$ be a sample of size m. Let $\mathcal{H} = \{x \mapsto w \cdot x \mid \|w\|_2 \leq \Lambda\}$. Then the empirical Rademacher complexity of $\mathcal{H}$ is bounded as

$$\hat{R}_S(\mathcal{H}) \leq \sqrt{\frac{r^2 \Lambda^2}{m}}.$$

Proof: We estimate

$$\begin{aligned}
\hat{R}_S(\mathcal{H}) &= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\|w\| \leq \Lambda} \sum_{i=1}^{m} \sigma_i w \cdot x_i \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\|w\| \leq \Lambda} w \cdot \sum_{i=1}^{m} \sigma_i x_i \right] \\
&\leq \frac{\Lambda}{m} \mathop{\mathbb{E}}_{\sigma} \left[\left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \right] \quad \text{1/3} \\
&\leq \frac{\Lambda}{m} \mathop{\mathbb{E}}_{\sigma} \left[\left(\left\| \sum_{i=1}^{m} \sigma_i x_i \right\|^2 \right) \right]
\end{aligned}$$

$$= \frac{1}{m} \mathop{\mathbb{E}}_{z} \left[\sum_{i,j=1}^{m} z_i z_j (x_i \cdot x_j) \right]^{1/2}$$

$$\leq \frac{1}{m} \left(\sum_{i=1}^{m} \|x_i\|^2 \right)^{1/2}$$

$$\leq \frac{1}{m} \cdot \sqrt{m r^2 \Lambda^2}$$

$$= \sqrt{\frac{r^2 \Lambda^2}{m}} \quad \blacksquare$$

Corollary (margin bound for linear hypotheses): Let $\mathcal{H} = \{ x \mapsto w \cdot x \mid \|w\|_2 \leq \Lambda \}$ and suppose that $X \subseteq \{ x \mid \|x\|_2 \leq r \}$. Fix $\rho > 0$. Then $\forall \delta > 0$, with probability at least $1-\delta$ over samples S of size m, the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{r^2 \Lambda^2}{m \rho^2}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

Remark: This bound depends on r, Λ and also m (naturally) and δ (naturally), as well as the margin ρ. But not D!

. In particular, if $\hat{R}_{S,\rho}(h)$ is small, then $\frac{\rho}{r\Lambda}$ large implies good generalization error.

. Note that the structure of the data goes a long way to determining if it is possible to choose ρ s.t. (a.) $\hat{R}_{S,\rho}(h)$ is small
(b.) $\frac{\rho}{r\Lambda}$ is large.

} in tension, and determined by how well-separated the data is

Remark: We can get this result to hold uniformly over ρ by the following theorem

Theorem: Let $\mathcal{H}$ be a set of real-valued functions. Fix $r > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds $\forall h \in \mathcal{H}$ and $\forall \rho \in (0, r]$:

(a.) $$R(h) \le \hat{R}_{S,\rho}(h) + \frac{4}{\rho}\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log(\log_2(\frac{2r}{\rho}))}{m}} + \sqrt{\frac{\log(\frac{2}{\delta})}{m}}$$

(b.) $$R(h) \le \hat{R}_{S,\rho}(h) + \frac{4}{\rho}\hat{\mathcal{R}}_S(\mathcal{H}) + \sqrt{\frac{\log(\log_2(\frac{2r}{\rho}))}{m}} + 3\sqrt{\frac{\log(\frac{4}{\delta})}{2m}}.$$

Proof: We prove (a.), since (b) just requires replacing the RC with the empirical RC. Consider sequences $\{\rho_k\}_{k\ge 1}$, $\{\varepsilon_k\}_{k\ge 1}$ with $\varepsilon_k \in [0,1]$. By our theorem on margin bounds for binary classification, we know that for k fixed,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}_{S,\rho_k}(h) > \frac{2}{\rho_k}\mathcal{R}_m(\mathcal{H}) + \varepsilon_k \right) \le \exp(-2m\varepsilon_k^2).$$

Let $\varepsilon_k = \varepsilon + \sqrt{\frac{\log k}{m}}$ for a fixed $\varepsilon > 0$. Then by the union bound,

$$\mathbb{P}\left(\sup_{\substack{h\in\mathcal{H} \\ k\ge 1}} R(h) - \hat{R}_{S,\rho_k}(h) - \frac{2}{\rho_k}\mathcal{R}_m(\mathcal{H}) - \varepsilon_k > 0 \right)$$

$$\le \sum_{k=1}^{\infty} \exp(-2m\varepsilon_k^2)$$

$$= \sum_{k=1}^{\infty} \exp\left(-2m\left[\varepsilon + \sqrt{\frac{\log k}{m}}\right]^2\right)$$

$$\le \sum_{k=1}^{\infty} \exp(-2m\varepsilon^2)\exp(-2m\log k/m)$$

$$= \sum_{k=1}^{\infty} \exp(-2m\varepsilon^2)\exp(-2\log k)$$

$$= \exp(-2m\varepsilon^2) \sum_{k=1}^{\infty} \frac{1}{k^2}$$

$$\leq 2\exp(-2m\varepsilon^2).$$

Now, choose $p_k = \frac{r}{2^k}$. Then for any $p \in (0, r]$, $\exists\, k$ s.t. $p \in (p_k, p_{k-1}]$ with $p_0 = r$. For that k, $p \leq p_{k-1} = 2p_k$, so that $\frac{1}{p_k} \leq \frac{2}{p}$ and hence

$$\sqrt{\log(\log_2(r/p_k))} \leq \sqrt{\log(\log_2(2r/p))}$$

Furthermore, for any $h \in H$, $\hat{R}_{S,p_k}(h) \leq \hat{R}_{S,p}(h)$. Thus,

$$\mathbb{P}\left(\sup_{\substack{h \in H \\ p \in (0,r]}} R(h) - \hat{R}_{S,p}(h) - \frac{4}{p} R_m(H) - \sqrt{\frac{\log(\log_2(2r/p))}{m}} - \varepsilon > 0 \right)$$

$$\leq 2\exp(-2m\varepsilon^2). \blacksquare$$

This allows us to make a uniform bound on the generalization error of SVM: setting $\Lambda = 1$, $\forall \delta > 0$ with probability at least $1-\delta$, the following estimate holds for all $h \in \{ x \mapsto w \cdot x \mid \|w\|_2 \leq 1 \}$, and $p \in (0, r]$:

$$R(h) \leq \hat{R}_{S,p}(h) + 4\sqrt{\frac{r^2}{mp^2}} + \sqrt{\frac{\log(\log_2(\frac{2r}{p}))}{m}} + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} \max\left\{0, 1 - \frac{y_i (w \cdot x_i)}{p}\right\} + 4\sqrt{\frac{r^2}{mp^2}} + \sqrt{\frac{\log(\log_2(2r/p))}{m}} + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$$