FoSML: Lecture 9

- Last time, we showed ERM works well as long as $\hat{R}_S(h) \simeq R(h)$. But, in practice we struggle to know $|\hat{R}_S(h) - R(h)|$ in practice.

- An alternative is to consider a family of hypothesis classes $\{\mathcal{H}_\gamma\}_{\gamma \in \Gamma}$ with increasing richness. We can then manage the approximation–estimation tradeoff by selecting a "good" choice $\gamma^* \in \Gamma$, not too large and not too small.

- This method is called structural risk minimization (SRM), and it works as follows:

Suppose our hypothesis set $\mathcal{H}$ (understood as very large, in most cases) decomposes as $\mathcal{H} = \bigcup_{k=1}^{\infty} \mathcal{H}_k$, where $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ for all $k \geq 1$. The SRM method ↳ given a sample of size m

Chooses $$h_S^{SRM} = \underset{k \geq 1,\, h \in \mathcal{H}_k}{\arg\min} \; \underbrace{\hat{R}_S(h)}_{\substack{\text{best} \\ \text{empirical} \\ \text{error achievable} \\ \text{in } \mathcal{H}_k}} + \underbrace{R_m(\mathcal{H}_k)}_{\substack{\text{Rademacher} \\ \text{complexity} \\ \text{of } \mathcal{H}_k}} + \underbrace{\sqrt{\frac{\log(k)}{m}}}_{\substack{\text{penalty for working with} \\ \text{a larger hypothesis class} \\ \mathcal{H}_k}}$$

- Let $\mathcal{H}_{k(h)}$ be the smallest $\mathcal{H}_k$ s.t. $h \in \mathcal{H}_{k(h)}$.

Theorem (SRM learning guarantees): For any $\delta > 0$, with probability exceeding $1-\delta$, a sample of size m, call it S, is s.t.

$$R(h_S^{SRM}) \leq \inf_{h \in \mathcal{H}} \left(R(h) + 2\mathcal{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log(k)}{m}} \right) + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{m}}.$$

Proof: For notational simplicity, let

$$F_k(h) = \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log(k)}{m}}.$$

Then by a union bound, we may estimate

$$\begin{aligned}
\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left[R(h) - F_{k(h)}(h)\right] > \varepsilon\right) &= \mathbb{P}\left(\sup_{k\geq 1}\sup_{h\in\mathcal{H}_k}\left[R(h) - F_k(h)\right] > \varepsilon\right) \\
&\leq \sum_{k=1}^{\infty} \mathbb{P}\left(\sup_{h\in\mathcal{H}_k}\left[R(h) - F_k(h)\right] > \varepsilon\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(\sup_{h\in\mathcal{H}_k}\left[R(h) - \hat{R}_S(h) - \mathcal{R}_m(\mathcal{H}_k)\right] > \varepsilon + \sqrt{\frac{\log(k)}{m}}\right) \\
&\leq \sum_{k=1}^{\infty} \exp\left(-2m\left[\varepsilon + \sqrt{\frac{\log k}{m}}\right]^2\right) \\
&\leq \sum_{k=1}^{\infty} \exp\left(-2m\varepsilon^2\right)\exp\left(-2\log(k)\right) \\
&= \sum_{k=1}^{\infty} \exp\left(-2m\varepsilon^2\right)\frac{1}{k^2} \\
&= \frac{\pi^2}{6}\cdot\exp\left(-2m\varepsilon^2\right) \\
&\leq 2\exp\left(-2m\varepsilon^2\right).
\end{aligned}$$

Notice moreover that for any two random variables X_1, X_2, a union bound argument yields

$\mathbb{P}(X_1 + X_2 > \varepsilon) \leq \mathbb{P}(X_1 > \frac{\varepsilon}{2}) + \mathbb{P}(X_2 > \frac{\varepsilon}{2})$. Now, $\forall h \in \mathcal{H}_1$ (3)

$$\mathbb{P}\left(R(h_S^{SRM}) - R(h) - 2\mathcal{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\frac{\log k(h)}{m}} > \varepsilon \right)$$

$$\leq \mathbb{P}\left(R(h_S^{SRM}) - F_{k(h_S^{SRM})}(h_S^{SRM}) > \frac{\varepsilon}{2} \right)$$

$$+ \mathbb{P}\left(F_{k(h_S^{SRM})}(h_S^{SRM}) - R(h) - 2\mathcal{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\frac{\log k(h)}{m}} > \frac{\varepsilon}{2} \right)$$

$$\leq 2\exp(-m\varepsilon^2/2) + \mathbb{P}\left(\underbrace{F_{k(h)}(h)}_{= F_R(h) \ \forall h \in \mathcal{H}, \text{ by definition of } h_S^{SRM}} - R(h) - 2\mathcal{R}_m(\mathcal{H}_{k(h)}) - \sqrt{\frac{\log k(h)}{m}} > \frac{\varepsilon}{2} \right)$$

$$= 2\exp(-m\varepsilon^2/2) + \mathbb{P}\left(\hat{R}_S(h) - R(h) - \mathcal{R}_m(\mathcal{H}_{k(h)}) > \frac{\varepsilon}{2} \right)$$

$$\leq 2\exp(-m\varepsilon^2/2) + \exp(-m\varepsilon^2/2)$$

$$= 3\exp(-m\varepsilon^2/2).$$

Setting the RHS $= \delta$ gets the job done.

Remarks: Suppose $\exists h^* \in \mathcal{H}$ s.t. $R(h^*) = \inf_{h \in \mathcal{H}} R(h)$. Then this result implies that $\forall h \in \mathcal{H}$, with probability exceeding $1 - \delta$,

$$R(h_S^{SRM}) \leq R(h^*) + 2\mathcal{R}_m(\mathcal{H}_{k(h^*)}) + \sqrt{\frac{\log(k(h^*))}{m}} + \sqrt{\frac{2\log(3/\delta)}{m}}.$$

This differs from the estimate we would have if we had known the "best" hypothesis class index $k(h^*)$ a priori... The only additional penalty is the

$$\sqrt{\frac{\log(k(h^*))}{m}} \text{ term.}$$

- On the other hand, we may not be able to write $\mathcal{H} = \bigcup_{k=1}^{\infty} \mathcal{H}_k$ with $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ and $R_m(\mathcal{H}_k)$ converging.

- Even worse from a practical standpoint, actually computing the SRM solution, namely

$$h_S^{SRM} = \underset{\substack{k \geq 1 \\ h \in \mathcal{H}_k}}{\operatorname{arg\,min}} \hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\frac{\log(k)}{m}}$$

is NP-hard, and an exhaustive search over k must be performed.

- In practice, one of the most common ways of selecting a predictor is to use cross-validation. In this case, we use some of our training labels to select the hypothesis set $\mathcal{H}_k$.

- As before, let $\{\mathcal{H}_k\}_{k \geq 1}$ be a family of nested hypothesis classes.

- Given a sample of size m, S, and $\alpha \in (0,1)$, we partition S into a training set of size $(1-\alpha)m$, S_1, and a validation set of size αm, S_2.

- For any $k \in \mathbb{Z}_+$, let $h_{S_1,k}^{ERM}$ be the ERM solution on S_1 using the hypothesis set $\mathcal{H}_k$. Then

$$h_S^{CV} = \underset{h \in \{h_{S_1,k}^{ERM}, k \geq 1\}}{\operatorname{arg\,min}} \hat{R}_{S_2}(h),$$

i.e. we pick the ERM solution which predicts best on S_2.

Proposition: For any $\alpha > 0$ and any $m \geq 1$,

$$\mathbb{P}\left(\sup_{k \geq 1} \left| R(h_{S_1,k}^{ERM}) - \hat{R}_{S_2}(h_{S_1,k}^{ERM}) \right| > \varepsilon + \sqrt{\frac{\log k}{\alpha m}} \right) \leq 4 \exp(-2\alpha m \varepsilon^2).$$

Proof: Next time.