EoSML: Lecture 5

- Recall the notion of Rademacher complexity.
- Our goal is to use the Rademacher complexity to bound the gap between $\hat{R}_S(h)$ (empirical classification error) and $R(h)$ (generalization error). Intuitively, we expect the principle of parsimony / Occam's razor to hold, i.e. lower Rademacher complexity $\Rightarrow$ better bounds on generalization error.

(★)

Thm: Let G be a family of functions mapping Z to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$, an iid sample S of size m satisfies:

$$\forall g \in G, \quad \text{(a.)} \quad \mathbb{E}(g(z)) \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2R_m(G) + \sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\text{(b.)} \quad \mathbb{E}(g(z)) \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{R}_S(G) + 3\sqrt{\frac{\log(2/\delta)}{2m}}$$

- This may be understood as a uniform concentration estimate in the mean, i.e. a quantitative law of large numbers. As $m \to \infty$, $\sqrt{\frac{\log(1/\delta)}{2m}} \sim \frac{1}{\sqrt{m}}$, so we have the expected convergence rate in m. But, the $R_m(G)/\hat{R}_S(G)$ terms account for the fact that if G is "complex", then there may be some unlucky $g \in G$ that are poorly estimated from the random sample S.

- To prove this theorem we shall make use of another concentration inequality, namely McDiarmid's inequality:

Theorem (McDiarmid's Inequality): Let $\{X_i\}_{i=1}^m$ be a set of independent r.v.'s and $X_i \in \mathcal{X}$ $\forall i$

Suppose $\exists c_1, \dots, c_m > 0$ s.t. $f: \mathcal{X} \to \mathbb{R}$ satisfies

$$|f(x_1, \dots, x_i, \dots, x_m) - f(x_1, \dots, x_i', \dots, x_m)| \leq c_i, \quad \forall i = 1, \dots, m$$
$$\forall x_1, \dots, x_m, x_i' \in \mathcal{X}.$$

Let $f(S) = f(X_1, \dots, X_m)$. Then, $\forall \varepsilon > 0$,

(a.) $\mathbb{P}\left(f(S) - \mathbb{E}f(S) \geq \varepsilon\right) \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^m c_i^2\right)$.

(b.) $\mathbb{P}\left(f(S) - \mathbb{E}(f(S)) \leq -\varepsilon\right) \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^m c_i^2\right)$.

- This says that if the range of f is pointwise bounded (quantified by the c_i's), then $f(S)$ concentrates (exponentially bounded error estimates) around its mean.

Proof of Thm ★: Let $\hat{\mathbb{E}}_S(g)$ be the empirical average of g over S, i.e. $\hat{\mathbb{E}}_S(g) = \frac{1}{m}\sum_{i=1}^m g(z_i)$.

We shall apply McDiarmid's inequality to the function

$$\Phi(S) = \sup_{g \in G} \left(\mathbb{E}(g) - \hat{\mathbb{E}}_S(g)\right).$$

- Let S, S' be two samples differing in exactly one point, WLOG z_m in S and z_m' in S'.

Then: $\Phi(S') - \Phi(S)$

$$= \sup_{g \in G} \left(\mathbb{E}(g) - \hat{\mathbb{E}}_{S'}(g)\right| - \sup_{g \in G} \left(\mathbb{E}(g) - \hat{\mathbb{E}}_S(g)\right)$$

$$\leq \sup_{g \in G} \left(\hat{\mathbb{E}}_S(g) - \hat{\mathbb{E}}_{S'}(g)\right)$$

$$= \sup_{g \in G} \left(\frac{1}{m} \sum_{i=1}^{m} g(z_i) - \frac{1}{m} \sum_{i=1}^{m} g(z_i') \right)$$

$$= \sup_{g \in G} \left(\frac{g(z_m) - g(z_m')}{m} \right)$$

$$\leq \frac{1}{m}, \quad \text{since } g : Z \to [0,1].$$

By an identical argument, $\Phi(S) - \Phi(S') \leq \frac{1}{m}$, so $|\Phi(S) - \Phi(S')| \leq \frac{1}{m}$. Then,

by McDiarmid, $\Phi(S) \leq \mathbb{E}_S(\Phi(S)) + \sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$ with probability at least $1 - \frac{\delta}{2}$.

We now estimate $\mathbb{E}_S(\Phi(S))$ in terms of Rademacher complexity:

$$\mathbb{E}_S(\Phi(S)) = \mathbb{E}_S\left(\sup_{g \in G} \left(\mathbb{E}(g) - \hat{\mathbb{E}}_S(g) \right) \right)$$

$$= \mathbb{E}_S\left(\sup_{g \in G} \left(\mathbb{E}_{S'}\hat{\mathbb{E}}_{S'}(g) - \hat{\mathbb{E}}_S(g) \right) \right)$$

$$\leq \mathop{\mathbb{E}}_{S,S'} \left(\sup_{g \in G} \left(\hat{\mathbb{E}}_{S'}(g) - \hat{\mathbb{E}}_S(g) \right) \right)$$

$$= \mathop{\mathbb{E}}_{S,S'} \left(\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} g(z_i') - g(z_i) \right)$$

$$= \mathop{\mathbb{E}}_{\sigma,S,S'} \left(\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z_i') - g(z_i)) \right) \quad \text{for a Rademacher r.v. } \sigma = (\sigma_1, \dots, \sigma_m)$$

$$\leq \mathop{\mathbb{E}}_{\sigma,S'} \left(\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i') \right) + \mathop{\mathbb{E}}_{\sigma,S} \left(\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \right)$$

$$= 2 \mathop{\mathbb{E}}_{z,S} \left(\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} z_i g(z_i) \right)$$

$= 2 \mathcal{R}_m(G)$. This gives (a).

To see (b), we again use McDiarmid's inequality to replace $\mathcal{R}_m(G)$ with $\hat{\mathcal{R}}_S(G)$: $\mathcal{R}_m(G) \leq \hat{\mathcal{R}}_S(G) + \sqrt{\frac{\log(1/\delta)}{2m}}$. The result then follows from a union bound, together with $\hat{\mathcal{R}}(S) \leq \mathbb{E}_S(\hat{\mathcal{R}}(S)) + \sqrt{\log(2/\delta)/2m}$ and (a). ∎

So far, our results have held for a broad class G. To bring things back to the question of binary classification and generalization error, we need a lemma.

Lemma: Let H be a family of functions taking values in $\{-1, 1\}$. Let G be the associated family of loss functions with the 0-1 loss:

$$G = \{ (x, y) \mapsto \mathbb{1}_{h(x) \neq y} \}_{h \in H}.$$

For any sample $S = \{(x_i, y_i)\}_{i=1}^m \subset X \times \{-1, 1\}$, let S_X denote the projection onto the first coordinate. Then:

$$\hat{\mathcal{R}}_S(G) = \frac{1}{2} \hat{\mathcal{R}}_{S_X}(H).$$

Proof: By definition, $\hat{\mathcal{R}}_S(G) = \mathbb{E}_z \left(\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} z_i \mathbb{1}_{h(x_i) \neq y_i} \right)$

$$= \mathbb{E}_z \left(\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} z_i \frac{1 - y_i h(x_i)}{2} \right)$$

$$= \frac{1}{2} \mathbb{E}_z \left(\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} -z_i y_i h(x_i) \right)$$

$$= \frac{1}{2}\mathbb{E}_\sigma\left(\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i y_i h(x_i)\right)$$

$$= \frac{1}{2}\mathbb{E}_\sigma\left(\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right)$$

$$= \frac{1}{2}\widehat{R}_{S_X}(H). \blacksquare$$

- From Thm(☆) and this lemma, it immediately follows that

Thm (Rademacher complexity bound)- binary classification): Let H be a family of functions taking values in $\{-1,1\}$. Let D be the distribution over the input space X. Then $\forall \delta > 0$, with probability exceeding $1-\delta$ a sample S of size m drawn from D satisfies:

(a) $R(h) \le \hat{R}_S(h) + R_m(H) + \sqrt{\log(\frac{1}{\delta})/2m}$

(b) $R(h) \le \hat{R}_S(h) + \widehat{R}_S(H) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$