FoML: Lecture 8

- A similar result holds when generalization error of 0 is not achievable, i.e. in the non-realizable case. We will not prove this result:

Theorem (lower bound, non-realizable case): Let $\mathcal{H}$ be a hypothesis set with VC-dimension $d > 1$. Then, $\forall m \geq 1$, for any learning algorithm $\mathcal{A}$, there exists a distribution over $\mathcal{X} \times \{0,1\}$ such that

$$\mathbb{P}_{S \sim \mathcal{D}}\left(R_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} R_{\mathcal{D}}(h) > \sqrt{\frac{d}{320m}} \right) \geq \frac{1}{64}.$$

Q: How should one construct a good class of hypotheses? If $\mathcal{H}$ is too small, we may be unable to fit the data well. If $\mathcal{H}$ is too big, we may generalize poorly, as shown by our generalization bounds which depend crucially on some notion of size of $\mathcal{H}$ (e.g. $|\mathcal{H}|$, $VCdim(\mathcal{H})$, $\mathfrak{R}(\mathcal{H})$, ...).

- We can quantify this tradeoff with the bias/variance or approximation/estimation tradeoff.

- More precisely, let $\mathcal{H}$ be a family of functions mapping $\mathcal{X}$ to $\{-1, +1\}$. For a hypothesis $h \in \mathcal{H}$, we can quantify its generalization performance by comparing to the Bayes error R^*: $R(h) - R^* \geq 0$.

We can decompose this into two terms, reflecting the approximation (dictated by how well $\mathcal{H}$ approximates the Bayes classifier) and the estimation (reflected by how well h does compared to the best $\mathcal{H}$ can do) aspects of a statistical learning problem:

$$R(h) - R^* = \underbrace{\left(R(h) - \inf_{h \in \mathcal{H}} R(h)\right)}_{\text{estimation}} + \overbrace{\left(\inf_{h \in \mathcal{H}} R(h) - R^*\right)}^{\text{approximation}}$$

- In general, the approximation error is hard to know, since the underlying distribution D on the data is needed. This is something in the tradition of approximation theory and optimal representations. We will discuss the estimation error instead, since it is closely linked to generalization bounds.

- A natural approach is to choose a hypothesis h which ~~minimizes~~ minimizes the empirical error on a training sample, then hope (or prove) this will also have low generalization error.

Defn: Let $\mathcal{H}$ be a hypothesis class and S a sample. Empirical risk minimization computes the hypothesis with minimal training error: $h_S^{ERM} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_S(h)$.

- We can relate the quality of ERM to generalization error as follows:

Proposition: For any sample S, the ERM estimator h_S^{ERM} satisfies

$$\mathbb{P}\left(R(h_S^{ERM}) - \inf_{h\in H} R(h) > \varepsilon\right) \le \mathbb{P}\left(\sup_{h\in H} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2}\right).$$

Proof: For any $\varepsilon > 0$, $\exists h_\varepsilon$ s.t. $R(h_\varepsilon) \le \inf_{h\in H} R(h) + \varepsilon$. Then, using $\hat{R}_S(h_S^{ERM}) \le \hat{R}_S(h_\varepsilon)$, we see

$$\begin{aligned}
R(h_S^{ERM}) - \inf_{h\in H} R(h) &= R(h_S^{ERM}) - R(h_\varepsilon) + R(h_\varepsilon) - \inf_{h\in H} R(h) \\
&\le R(h_S^{ERM}) - R(h_\varepsilon) + \varepsilon \\
&= R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_S^{ERM}) - R(h_\varepsilon) + \varepsilon \\
&\le R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_\varepsilon) - R(h_\varepsilon) + \varepsilon \\
&\le 2 \cdot \sup_{h\in H} |R(h) - \hat{R}_S(h)| + \varepsilon
\end{aligned}$$

Sending $\varepsilon \to 0^+$ gives $R(h_S^{ERM}) - \inf_{h\in H} R(h) \le 2 \cdot \sup_{h\in H} |R(h) - \hat{R}_S(h)|$.

The result follows immediately. ∎

Remark: We can use ~~concent~~ generalization bounds to control

$$\mathbb{P}\left(\sup_{h\in H} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2}\right),$$

thus acquiring bounds on $\mathbb{P}\left(R(h_S^{ERM}) - \inf_{h\in H} R(h) > \varepsilon\right)$, i.e. on the quality of ERM.