Prof. Yiping Lu
IEMS 402 Statistical Learning
November 3, 2024

Homework 6: Generalization Error

Question 1. (Fast Rate Generalization Error in the Realizable Setting) Let $\mathcal{H}$ be a hypothesis class, where each hypothesis $h \in \mathcal{H}$ maps some $\mathcal{X}$ to $\mathcal{Y}$. ℓ be the zero-one loss: $\ell((x, y), h) = \mathbb{I}[y \neq h(x)]$. p^* be any distribution over $\mathcal{X} \times \mathcal{Y}$.

Let $\hat{L}(h)$ be the **empirical risk** (training error) of a hypothesis $h \in \mathcal{H}$ as the average loss over the training examples:

$$\hat{L}(h) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), h).$$

Define an **empirical risk minimizer** (ERM) be any hypothesis that minimizes the empirical risk:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h).$$

Now we assume

- Hypothesis class $\mathcal{H}$ is finite.
- Assume there exists a hypothesis $h^* \in \mathcal{H}$ that obtains zero expected risk, that is:

$$L(h^*) = \mathbb{E}_{(x,y)\sim p^*} \left[\ell((x, y), h^*)\right] = 0.$$

Prove that with probability at least $1 - \delta$,

$$L(\hat{h}) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{n}.$$

Question 2. (Generalization Error near Interpolate) In the realizable setting with binary classification (where the expected risk minimizer h^* satisfies $L(h^*) = 0$ for the 0-1 error), we obtained excess risk bounds of $O(1/n)$, but in the unrealizable setting, we had $O(\sqrt{1/n})$. What if the learning problem is almost realizable, in that $L(h^*)$ is small? This problem explores ways to interpolate between $1/n$ and $1/\sqrt{n}$ rates, showing that (roughly) $\sqrt{L(h^*)/n}+1/n$ rates are possible by developing generalization bounds that depend on the *variance of losses* (recall Question 12).

(a) Assume that the loss function $\ell(y, t)$ takes values in $[0, 1]$, where $L(h) = \mathbb{E}[\ell(Y, h(X))]$, and let $\hat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i))$. Show that for all $\varepsilon \geq 0$ we have

$$\mathbb{P}\left(\hat{L}_n(h) - L(h) \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(L(h) + \varepsilon/3)}\right).$$

(Note that if $L(h) = 0$, this bound scales as $e^{-n\varepsilon} \ll e^{-n\varepsilon^2}$ for $\varepsilon \approx 0$.)

(b) We now show that bad hypotheses usually look pretty bad. Fix any $\varepsilon(h), \varepsilon \geq 0$, and assume

$$L(h) \geq \varepsilon(h) + \varepsilon.$$

Show that

$$\mathbb{P}\left(\hat{L}_n(h) \leq \varepsilon(h)\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon(h) + 4\varepsilon/3)}\right).$$

(c) Assume $\text{card}(\mathcal{H}) < \infty$ and let h^* satisfy $L(h^*) = \min_{h \in \mathcal{H}} L(h)$. Using the preceding parts, conclude that if $\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \hat{L}_n(h)$, then

$$\mathbb{P}\left(\hat{L}_n(h) - L(h^*) \geq 2\varepsilon\right) \leq \text{card}(\mathcal{H}) \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right).$$

Show that this implies (for appropriate numerical constants c_1, c_2) that with probability at least $1-\delta$, we have

$$L(\hat{h}_n) \leq L(h^*) + c_1\sqrt{\frac{L(h^*)\log\frac{\mathrm{card}(\mathcal{H})}{\delta}}{n}} + c_2\frac{\log\frac{\mathrm{card}(\mathcal{H})}{\delta}}{n}.$$

(d) How does this bound compare with a more naive strategy based on applying Hoeffding's inequality and a union bound?

Question 3. (Random Matrix) (Random matrix) Let A be an $m \times n$ matrix of iid $\mathcal{N}(0,1)$ entries. Denote its operator norm by

$$\|A\|_{\mathrm{op}} = \max_{v\in S^{n-1}} \|Av\|,$$

which is also the largest singular value of A.

(a) Show that

$$\|A\|_{\mathrm{op}} = \max_{u\in S^{m-1}, v\in S^{n-1}} \langle Au, v'\rangle.$$

(b) Let $\mathcal{U} = \{u_1, \ldots, u_M\}$ and $\mathcal{V} = \{v_1, \ldots, v_M\}$ be an ϵ-net for the spheres S^{m-1} and S^{n-1} respectively. Show that

$$\|A\|_{\mathrm{op}} \leq \frac{1}{(1-\epsilon)^2}\max_{u\in\mathcal{U}, v\in\mathcal{V}} \langle Au, v\rangle.$$

(c) Use (a) and (b) to conclude that

$$\mathbb{E}[\|A\|] \lesssim \sqrt{n} + \sqrt{m}.$$

(*hint*: You can also Rademacher Complexity for the Uniform Bound.)

(d) By choosing u and v in (5) smartly, show a matching lower bound and conclude that

$$\mathbb{E}[\|A\|] \approx \sqrt{n} + \sqrt{m}.$$

Question 4. (Rademacher Complexity Leads to Suboptimal Bounds) Suppose we aim to estimate a parameter θ based on i.i.d. samples $X_i \sim N(\theta, I)$ for $i = 1, 2, \ldots, n$. We use the estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

1) Derive the asymptotic distribution of $\hat{\theta}$ using the Central Limit Theorem. Additionally, compute $\mathbb{E}\|\hat{\theta} - \theta\|^2$ and discuss how this expectation behaves as n grows.
2) Consider $\hat{\theta}$ as the minimizer of the empirical risk: $\hat{\theta} = \arg\min_\theta \mathbb{E}_{\hat{P}}\|\theta - X\|^2$, where $\mathbb{E}_{\hat{P}}$ denotes the empirical expectation based on the sample. Use Rademacher complexity to derive an upper bound for the error of $\hat{\theta}$ in estimating θ, and assess whether this bound is optimal.

Northwestern University