3.8 Rademacher Complexity

Rademacher complexity helps to measure the complexity of a hypothesis class.

① Set up:

Define $G_n := \sup_{h\in\mathcal{H}} L(h) - \hat{L}(h)$, then G_n is r.v. depending on $Z_1, \cdots, Z_n$ (data).

Define $G_n' := \sup_{h\in\mathcal{H}} \hat{L}(h) - L(h)$, so that

$$P\{L(\hat{h}) - L(h^*) \geq \varepsilon\} \leq P\{\sup_{h\in\mathcal{H}} |L(h) - \hat{L}(h)| \geq \frac{\varepsilon}{2}\}$$
$$\leq P\{G_n \geq \frac{\varepsilon}{2}\} + P\{G_n' \geq \frac{\varepsilon}{2}\}$$

② Concentration:

Let g be the deterministic function s.t. $G_n = g(Z_1, \cdots, Z_n)$, Then g satisfies the bounded differences condition:

$$|g(Z_1, \cdots, Z_i, \cdots, Z_n) - g(Z_1, \cdots, Z_i', \cdots, Z_n)| \leq \frac{1}{n}$$

Pf: $\hat{L}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, h)$, we have

$$\left|\sup_{h\in\mathcal{H}}\{L(h) - \hat{L}(h)\} - \sup_{h\in\mathcal{H}}\{L(h) - \hat{L}(h) + \frac{1}{n}[\ell(Z_i,h) - \ell(Z_i',h)]\}\right| \leq \frac{1}{n}$$

$$\Rightarrow \quad |g(Z_1, \cdots, Z_i, \cdots, Z_n) - g(Z_1, \cdots, Z_i', \cdots, Z_n)| \leq \frac{1}{n} \qquad \square$$

Now apply McDiarmid's inequality:

$$P[G_n \geq E[G_n] + \varepsilon] \leq \exp\{-2n\varepsilon^2\}.$$

③ Symmetrization. Bound $E[G_n]$.

Introduce ghost dataset $Z_1', \cdots, Z_n'$ drawn i.i.d. from p^*.

Let $\hat{L}'(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i', h)$

Rewriting $L(h)$ in terms of the ghost dataset.

$$E[G_n] = E[\sup E[\hat{L}'(h)] - \hat{L}(h)]$$

$$E[G_n] = E[\sup_{h\in H} E[\hat{L}'(h)] - \hat{L}(h)]$$
$$= E[\sup_{h\in H} E[\hat{L}'(h) - \hat{L}(h) \mid Z_{1:n}]]$$
$$\le E[E[\sup_{h\in H} \hat{L}'(h) - \hat{L}(h) \mid Z_{1:n}]]$$
$$= E[\sup_{h\in H} \hat{L}'(h) - \hat{L}(h)]$$

$$\Rightarrow \quad E[G_n] \le E[\sup_{h\in H} \frac{1}{n}\sum_{i=1}^{n}[\ell(Z_i', h) - \ell(Z_i, h)]]$$

To remove the dependence on the ghost dataset $Z_{1:n}'$, we introduce i.i.d. Rademacher variables $\sigma_1, \cdots, \sigma_n$ independent of $Z_{1:n}, Z_{1:n}'$, where σ_i is uniform over $\{-1, +1\}$. Since $\ell(Z_i', h) - \ell(Z_i, h)$ is symmetric around 0, multiplying by σ_i doesn't change its distribution. Then without dependence of σ_i's distribution:

$$E[G_n] \le E[\sup_{h\in H} \frac{1}{n}\sum_{i=1}^{n}\sigma_i[\ell(Z_i', h) - \ell(Z_i, h)]]$$
$$\le E[\sup_{h\in H} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(Z_i', h) + \sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}(-\sigma_i)\ell(Z_i, h)]$$
$$= 2E[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(Z_i, h)]$$

here we consider $\sigma_i \overset{d}{=} -\sigma_i$

RHS is defined as Rademacher complexity.

[Definition 9] Rademacher complexity.

Let F be a class of real-valued functions $f: Z \to \mathbb{R}$.

① Define the Rademacher complexity of F to be

$$R_n(F) := E[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)]$$

where (i) $Z_1, \cdots, Z_n \sim p^*$, i.i.d.

(ii) $\sigma_1, \cdots, \sigma_n \sim U\{-1, +1\}$ uniform.

② Define the empirical Rademacher complexity of F to be

$$\hat{R}_n(F) := E[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i) \mid Z_{1:n}]$$

[Theorem 9] generalization bounds based on Rademacher complexity.

Define $A := \{z \mapsto \ell(z, h) : h \in H\}$ to be the loss class. With prob. $\geq 1-\delta$, $L(\hat{h}) - L(h^*) \leq 4R_n(A) + \sqrt{\frac{2\log(4/\delta)}{n}}$

Pf: Note that $E[G_n] \leq 2R_n(A)$, and $R_n(A) = R_n(-A)$

$E[G_n'] \leq 2R_n(-A)$

$$P\{G_n \geq \tfrac{\varepsilon}{2}\} \leq \exp(-2n(\tfrac{\varepsilon}{2} - E[G_n])^2)$$

$$\leq \exp(-2n(\tfrac{\varepsilon}{2} - 2R_n(A))^2) \quad \text{for } \varepsilon \geq 4R_n(A)$$

$$:= \frac{\delta}{2}$$

Similarly, $P\{G_n' \geq \frac{\varepsilon}{2}\} \leq \frac{\delta}{2}$. And we have

$$\varepsilon = 2\sqrt{\log(2/\delta)/2n} + 4R_n(A) \qquad (\varepsilon \geq 4R_n(A)). \quad \square$$

Then we just need to study $R_n(A)$ over different A. now we first discuss some basic propeties:

[Basic Propeties of Rademacher complexity]

- ✓ (i) Boundedness: $R_n(\mathcal{F}) \leq \max_{f \in \mathcal{F}} \max_{z} f(z)$
- ✓ (ii) Singleton: $R_n(\{f\}) = 0$
- ✓ (iii) Monotonicity: $R_n(\mathcal{F}_1) \leq R_n(\mathcal{F}_2)$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$.
- ✓ (iv) Linear combination: $R_n(\mathcal{F}_1 + \mathcal{F}_2) = R_n(\mathcal{F}_1) + R_n(\mathcal{F}_2)$
- ✓ (v) Scaling: $R_n(c\mathcal{F}) = |c| R_n(\mathcal{F})$
- * (vi) Lipschitz composition: $R_n(\phi \circ \mathcal{F}) \leq C_\phi R_n(\mathcal{F})$, C_ϕ is the L-constant.
- * (vii) Convex hull: $R_n(\text{convex-hull}(\mathcal{F})) = R_n(\mathcal{F})$ for finite $\mathcal{F}$.