9. VC dimension and Rademacher Complexity

Sunday, January 23, 2022 5:44 PM

§6 in [SS] w/ Rademacher Complexity taken from §3.1 in Mohri et al.

We've covered ① $|\mathcal{H}| < \infty$ (restrictive!)

② HW: axis-aligned rectangles, $|\mathcal{H}| = \infty$. $X = \mathbb{R}^d$, $\dim(\mathcal{H}) = 2d$

Can we generalize?

we'll cover

① Rademacher Complexity (Mohri, and essentially used in later chapters of [SS])

Simple proofs, but computing R.C. may be impossible (eg, NP-Hard)

especially if $ERM_{\mathcal{H}}$ is difficult to compute

② Growth Function

③ VC-dimension, a way to bound the growth function, and easier to compute or bound

④ Result: for binary classification, finite VC-dim is necessary and sufficient for PAC learnability

"Fundamental Thm. of ML"

Rademacher Complexity

(§3.1 Mohri, notation adapted a bit)

Will depend on $\mathcal{H}$ and loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}$ in [SS]

$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ in [Mohri et al.]

e.g. $\ell(h, (x,y)) = \ell(h(x), y)$

we'll apply to a family of functions

$$\mathcal{F} = \{ f : (x,y) \mapsto \ell(h, (x,y)) \;\; \forall \; h \in \mathcal{H} \}$$
$$= \ell \circ \mathcal{H}$$

but it'll work for any family of functions $\mathcal{F}$, not just $\ell \circ \mathcal{H}$

$\mathcal{F} \subseteq \mathbb{R}^Z$ $\quad Z = X \times \mathcal{Y}$

Idea

Rademacher Complexity (RC) measures the richness / expressivenes of $\mathcal{F}$ by measuring how well it can fit noise

Def Empirical Rademacher Complexity

$\mathcal{F}$ a family of fcn $f: Z \to [a,b]$. Fix $S = (z_1, \dots, z_m)$

then empirical R.C. of $\mathcal{F}$ w.r.t. S is

$$\hat{R}_S(\mathcal{F}) = \mathbb{E}_{\sigma}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \cdot f(z_i) \right]$$

where $\sigma = (\sigma_1, \dots, \sigma_m)$, σ_i iid Rademacher variables

i.e. $\sigma_i = \begin{cases} +1 & \text{w.p. } .5 \\ -1 & \text{w.p. } .5 \end{cases}$ aka symmetric Bernoulli or uniform on $\{-1,1\}$

i.e., let $f_S := \begin{bmatrix} f(z_1) \\ \vdots \\ f(z_m) \end{bmatrix} \in \mathbb{R}^m$ then

$$\hat{R}_S(\mathcal{F}) = \frac{1}{m} \mathbb{E}_{\sigma} \underbrace{\sup_{f \in \mathcal{F}} \langle \sigma, f_S \rangle}_{\text{i.e., best correlation w/ noise}}$$

Extremes: $\mathcal{F} = \{f\}$, $\hat{R}_S(\{f\}) = \frac{1}{m} \mathbb{E}_{\sigma} \langle \sigma, f_S \rangle = 0$. Best possible

vs. $\mathcal{F}$ = all functions, say $f: Z \to \{0,1\}$, then possible for some S

for $\{f_S : f \in \mathcal{F}\} = \{0,1\}^m$

Then $\sup_{f \in \mathcal{F}} \langle \sigma, f_S \rangle = m$, so $\hat{R}_S(\mathcal{F}) = \frac{1}{m} \mathbb{E}_{\sigma} m = \underline{\underline{1}}$

worst-possible (if $[a,b] = [0,1]$)

Def Rademacher Complexity (not "empirical")

$$R_m(\mathcal{F}) := \mathbb{E}_{S \sim \mathcal{D}^m} \hat{R}_S(\mathcal{F})$$

Careful: [SS] uses different terminology:

we'll follow Mohri

	concept	concept
Mohri et al.	$\hat{R}_S(\mathcal{F})$ "Empirical R.C."	$R_m(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{D}^m} R_S(\mathcal{F})$ "R.C."

Shalev-Shwartz & Ben-David $R(\mathcal{F} \circ S)$ and $\mathcal{F} = \ell \circ \mathcal{H}$

"R.C."

$\mathbb{E}_S \, R(\mathcal{F} \circ S)$ (no special notation)

"Expected R.C."

How to use?

Recall for uniform convergence, S was "ε-representative" if

$$\sup_{h \in \mathcal{H}} \left| \underbrace{L_D(h) - \hat{L}_S(h)} \right| \leq \varepsilon$$

(which implied ERM worked: $L_D(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \varepsilon + \min_{h \in \mathcal{H}} L_D(h)$)

Something very similar is the "representativeness" of S (w.r.t. $\mathcal{H}, \ell$) as

$$\mathrm{Rep}_D((\mathcal{H},\ell), S) := \sup_{h \in \mathcal{H}} \overbrace{\underbrace{\mathbb{E}_{z \sim D} \ell(h,z)}_{L_D(h)} - \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)}_{\hat{L}_S(h)}}$$

or more generally

$$\underset{\text{in Mohri}}{\Phi(S)} = \mathrm{Rep}_D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \underbrace{\mathbb{E}_{z \sim D}}_{\mathbb{E}} f(z) - \underbrace{1/m \sum_{i=1}^{m}}_{\hat{\mathbb{E}}_S} f(z_i)$$

$\underbrace{\text{want this small}}_{\text{clear}}$ $= \sup\left(\mathbb{E} f - \hat{\mathbb{E}}_S f\right)$ in shorthand notation.

Intuitively, $\hat{R}_S(\mathcal{F})$ is a reasonable estimate for $\mathrm{Rep}_D(\mathcal{F}, S)$

Why? in $\mathrm{Rep}_D(\mathcal{F}, S)$ we have $\mathbb{E} f - \hat{\mathbb{E}}_S f$. Split $S = S_1 \cup S_2$ at random

estimate $\mathbb{E} f - \hat{\mathbb{E}}_S f$ by $\underbrace{\hat{\mathbb{E}}_{S_1} f - \hat{\mathbb{E}}_{S_2} f}_{\text{rewrite}}$

Let $S_1 = \{ i \in [m] : \sigma_i = +1 \}$, $S_2 = S \setminus S_1$, and suppose $|S_1| = m/2$ exactly

then $$\hat{\mathbb{E}}_{S_1} f - \hat{\mathbb{E}}_{S_2} f = \frac{1}{m/2} \sum_{i \in S_1} f(z_i) - \frac{1}{m/2} \sum_{i \in S_2} f(z_i)$$

$$= \frac{1}{m/2} \sum_i \sigma_i f(z_i)$$

Thus taking a $\sup_{f \in \mathcal{F}}(\ldots)$, as we do in Rep_D and $\hat{R}_S$,

we get $\mathrm{Rep}_D(\mathcal{F}, S) \approx 2 \cdot \hat{R}_S(\mathcal{F})$

Now, let's be slightly more careful and formalize the above:

Lemma 26.2 (Mohri) $\mathbb{E}_{S\sim D^m} \operatorname{Rep}_D(\mathcal{F}, S) \le 2 \cdot \mathbb{E}_{S\sim D^m} \hat{\mathcal{R}}_S(\mathcal{F})$

$= 2 \cdot \mathcal{R}_m(\mathcal{F})$.

proof:

$$\mathbb{E}_{S\sim D^m} \operatorname{Rep}_D(\mathcal{F}, S) := \mathbb{E}_{S\sim D^m} \sup_{f\in\mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_S f$$

no effect

$$= \mathbb{E}_{S\sim D^m} \sup_{f\in\mathcal{F}} \underbrace{\mathbb{E}_{S'\sim D^m} \left(\hat{\mathbb{E}}_{S'} f\right.}_{\mathbb{E} f} \left. - \hat{\mathbb{E}}_S f \right)$$

$$\le \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_{S'} f - \hat{\mathbb{E}}_S f \right)$$

sup is sub-additive [see details later]

$$:= \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m f(z_i') - f(z_i)$$

$$= \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(f(z_i') - f(z_i) \right)$$

for any σ_i

$\sigma_i = 1$ ok

$\sigma_i = -1$ flips z_i', z_i ... but same distribution

true $\forall \sigma$ so true for $\mathbb{E}$

$$= \mathbb{E}_{S,S',\sigma} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(f(z_i') - f(z_i) \right)$$

$$\le \mathbb{E}_{S',\sigma} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i') + \mathbb{E}_{S,\sigma} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i (z_i)$$

$\sup(x+y) \le \sup(x) + \sup(y)$

$$= 2\, \mathbb{E}_{\sigma,S} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i) = 2\, \mathbb{E}_S \hat{\mathcal{R}}_S(\mathcal{F})$$

$$= 2\, \mathcal{R}_m(\mathcal{F}) \quad \square$$

Sub-additive:

$\forall a, b \qquad g(a,b) \le \sup_{b'} g(a,b')$

so $\forall b \qquad \mathbb{E}_a\, g(a,b) \le \mathbb{E}_a \sup_{b'} g(a,b')$

so

$$\underline{\sup_b \mathbb{E}_a\, g(a,b)} \le \underline{\mathbb{E}_a \sup_{b'} g(a,b')}$$