12. More on Rademacher Complexity

Wednesday, January 26, 2022 9:36 AM

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

or, to generalize (simplify): Rademacher Complexity of a set:

def $\hat{\mathcal{R}}(A) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{a\in A} \langle a, \sigma\rangle$, $A \subseteq \mathbb{R}^m$

(By setting $A = \left\{ \begin{bmatrix} f(z_1) \\ \vdots \\ f(z_m) \end{bmatrix} : f \in \mathcal{F} \right\}$

we recover $\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}_S(\mathcal{F})$

i.e. $\hat{\mathcal{R}}_S(\mathcal{F}) := \hat{\mathcal{R}}(\mathcal{F}\circ S)$)

N.B. Very related ideas have been discovered / proposed in other subfields

eg. Gaussian width of a set A is $G(A) = \mathop{\mathbb{E}}_{g\sim N(0,I_m)} \sup_{a\in A} \langle g, a\rangle$ ← sometimes in abs. value

Also "Local Rademacher Complexity"

and Statistical Dimension $= \mathop{\mathbb{E}}_{g\sim N(0,I_m)} \|\mathrm{Proj}_A(g)\|^2$

Calculus / tools

Lemma 26.2 [SS] $\hat{\mathcal{R}}(c\cdot A + a_0) \le |c|\cdot\hat{\mathcal{R}}(A)$

"shift-invariant", scales as expected

Lemma 26.7 [SS] $\hat{\mathcal{R}}(\mathrm{conv}(A)) = \hat{\mathcal{R}}(A)$

where $\mathrm{conv}(A)$ = convex hull of A

= smallest convex set containing A

= intersection of all convex sets containing A

$= \{ b = \sum_{i=1}^{p} \alpha_i \cdot a_i$ s.t. $p\in\mathbb{N}$, $a_i \in A$, $\alpha_i \ge 0$, $\sum \alpha_i \le 1 \}$

Lemma 26.8 [SS] Massart Lemma ← Pascal Massart (b. 1958) "t" is silent

If $|A| = N < \infty$ then $\hat{\mathcal{R}}(A) \le \max_{a\in A} \|a - \bar{a}\| \cdot \frac{1}{m}\sqrt{2\cdot\log(N)}$

$\bar{a} := \frac{1}{N}\sum_{i=1}^{N} a_i$ is center of mass

(wlog can shift to 0)

eg. if $A = \ell \circ \mathcal{H} \circ S$,
and ℓ is bounded, then $\max \|\cdots\|$ is bounded

and $\hat{R}_S(\ell \circ \mathcal{H}) = O(1/m \sqrt{\log|\mathcal{H}|})$. Combine w/ thm to bound L_D

Lemma 26.9 [SS] If $\phi: \mathbb{R}^m \to \mathbb{R}^n$ is ρ-Lipschitz in each component $i \in [m]$
then $\hat{R}(\phi \circ A) \leq \rho \cdot \hat{R}(A)$

Using these tools you can sometimes compute / bound R for simple/nice enough sets

Ex $\mathcal{H}_2 := \{ x \mapsto \langle w, x \rangle : \|w\|_2 \leq 1 \}$, an ∞ set of linear classifiers

By Lemma 26.10 [SS], if $S_x = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^n$

$$\hat{R}(\mathcal{H}_2 \circ S_x) \leq 1/\sqrt{m} \cdot \max_i \|x_i\|_2$$

Ex $\mathcal{H}_1 := \{ x \mapsto \langle w, x \rangle : \|w\|_1 \leq 1 \}$

note $\|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty$
and $\|x\|_\infty \leq \|x\|_2$

Lemma 26.11 [SS],

$$\hat{R}(\mathcal{H}_1 \circ S_1) \leq 1/\sqrt{m} \max_i \sqrt{2 \cdot \log(2n)} \|x_i\|_\infty$$

(we might do in detail for Htw)

Perspective

R.C. used for uniform convergence ~2000
(often tools have been around, but only "popular" once ML discovers)

R.C. mainly for classification

- For regression, if loss is bounded (less common), R.C. can work
- otherwise need specialized measures

Alternatives: (ideas like VC-dim are inherently combinatorial and won't work)

- Pseudo-dimension (like VC-dim, needs boundedness)
- Fat-shattering dim. (still needs boundedness)

see Mohri for brief discussion

Another notion of complexity:

Covering Numbers §27 [SS], §3.5 Mohri

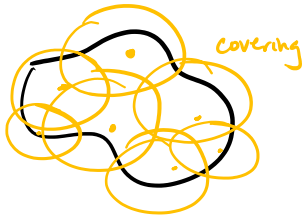
Def The $N_p(\varepsilon, A)$ covering number of a set A is the minimum number of points needed to cover A, i.e., s.t. A is contained in the union of ε-balls centered at these points

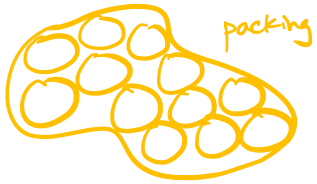
$$B_\varepsilon(y) = \{x : \|x-y\|_p \le \varepsilon\}$$

(If p not specified, assume $p=2$)

related to... packing numbers $M_p(\varepsilon, A)$, the max # balls you can fit in A w/o overlapping. (of radius ε wrt ℓ^p norm)

Fact: $M_p(2\varepsilon, A) \le N_p(\varepsilon, A) \le M_p(\varepsilon, A)$





we'll prove shortly that you can bound $\hat{\mathcal{R}}(A)$ by $N_2(\varepsilon, A)$ "Dudley's Thm" ('67, '87) ... and can bound N_2 by VC-dim.

uses "chaining technique" which we'll cover

Facts

log(covering number) is sometimes called "metric entropy"

- $N_2(\varepsilon, cA + a_0) = N_2(c\cdot\varepsilon, A)$ translation / scaling
- ϕ componentwise Lipschitz, w/ constant ρ, then $N_2(\varepsilon, \phi \circ A) \le N_2(\varepsilon/\rho, A)$
- Lemma 27.1 [SS] If $A \subseteq \mathbb{R}^m$ is in a d-dimensional subspace and $A \subseteq B_2(0)$ (ie. radius $\le c$ wrt ℓ^2 norm) then $N_2(\varepsilon, A) \le \left(2c\frac{\sqrt{d}}{\varepsilon}\right)^d$
- Lemma 2.2 [Woodruff] $A \subseteq \mathbb{R}^m$ is in a d-dim. subspace and $\forall a \in A$, $\|a\|_2 = 1$ (ie. subspace $\cap$ sphere) then $N_2(\varepsilon, A) \le (1 + 2/\varepsilon)^d$

proof (fun)

wlog work in unit sphere inside $\mathbb{R}^d$.

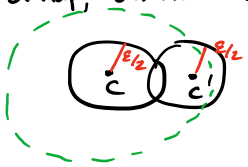
We'll find a covering (maybe not optimal) in a greedy fashion that will at least be maximal! nothing redundant

Let C be the centers, so covering $= \bigcup_{c \in C} B_\varepsilon(c)$

Pick $c \in A$ arbitrarily. Check if A is covered.
- If so, stop
- If not, $\exists\, c' \in A$ that hasn't been covered (ie. not with ε of any $c \in C$). Add c' to C

Repeat. This is maximal.
(This is the intuitive way) How big is $|C|$?

Balls of radius ε may overlap, but balls of radius $\varepsilon/2$ (around centers) cannot overlap, otherwise non maximal.



if $\varepsilon/2$ balls overlap
$\Rightarrow \|c - c'\| < \varepsilon$
$\Rightarrow$ not maximal : either c (or c') was already covered.

Furthermore, all these $\varepsilon/2$ balls are within $1+\varepsilon/2$ of origin (centers on unit sphere)

Recall in dimension d, volume (ball radius r) $\propto r^d$

1D: $2r$
2D: πr^2
3D: $4/3\,\pi r^3$

By disjointness of $\varepsilon/2$ balls,

$$\text{Volume}(\text{all } \varepsilon/2 \text{ balls}) = \sum_{\text{all } \varepsilon/2 \text{ balls}} \text{Vol.}(\varepsilon/2 \text{ ball})$$
$$= |C| \cdot \text{const} \cdot (\varepsilon/2)^d$$

but also
$$\text{Vol.}(\text{all } \varepsilon/2 \text{ balls}) \le \text{Vol}(1+\varepsilon/2 \text{ ball}) = \text{const} \cdot (1+\varepsilon/2)^d$$

$$\Rightarrow |C| \cdot (\varepsilon/2)^d \le (1+\varepsilon/2)^d, \quad |C| \le \left(\frac{1+\varepsilon/2}{\varepsilon/2}\right)^d = (1+2/\varepsilon)^d$$

$$\Rightarrow N_2(\varepsilon, A) \le (1+2/\varepsilon)^d. \quad \square$$

(seems loose since we didn't exploit $\|a\| = 1$ only $\|a\| \le 1$,
but in high dim. the surface is where all the volume is, so not a big deal)