10. Rademacher, part 2

Sunday, January 23, 2022 7:49 PM

Interlude: McDiarmid's Inequality

· <u>Thm</u> (McDiarmid)

Let f be a function s.t. $\exists\, c_i < \infty$ s.t.

$\forall i \in [m]$, $\quad | f(x_1, \dots, x_i, \dots, x_m) - f(x_1, \dots, x_i', \dots, x_m) | \le c_i$

"bounded difference"

for arbitrary $(x_1, \dots, x_m)$ and x_i'

Let $S = (\underbrace{X_1, \dots, X_m}_{\text{independent r.v.}})$ ← prob. notation

Then $\mathbb{P}[\, f(S) - \mathbb{E} f(S) \ge \varepsilon \,] \le \exp\left(\frac{-2\varepsilon^2}{\sum_i c_i^2} \right)$

$\mathbb{P}[\, f(S) - \mathbb{E} f(S) \le -\varepsilon \,] \le$ " ______ " .

(in Hoeffding, $f(S) = 1/m \sum_i X_i$ ... f is linear in S)

<u>Thm</u> (main R.C. generalization bound) Thm 3.3 in Mohri

Let $\mathcal{F}$ be a family of functions from X to $[0,1]$ (or really $[a,b]$ works)

then $\forall\, \delta > 0$, w.p. $\ge 1-\delta$ (w.r.t. $S \sim D^m$ iid) [any D]

$$(\forall f \in \mathcal{F}) \quad \underbrace{\mathbb{E}_{z \sim D} f(z)}_{\text{risk if } f = \ell \circ h} - \underbrace{\frac{1}{m} \sum_{i=1}^m f(z_i)}_{\text{empirical risk}} \le \begin{cases} \text{Ⓐ} \; 2\overbrace{\mathcal{R}_m(\mathcal{F})}^{D \text{ shows up here}} + \sqrt{\frac{\log(1/\delta)}{2m}} \\ \text{Ⓑ} \; \underbrace{2\hat{\mathcal{R}}_S(\mathcal{F})}_{\substack{\text{data dependent} \\ \text{(sometimes nice, sometimes not)}}} + 3 \cdot \sqrt{\frac{\log(2/\delta)}{2m}} \end{cases} \Bigg\} \text{ Right Hand Side RHS}$$

(see also Thm. 26.5 in [SS] ...

[SS] usually sets RHS $= \varepsilon(m)$, solves for $m(\varepsilon)$

which is nice to have but requires many approximations / bounds

to make it tractable)

<u>Proof</u> Recall $\mathrm{Rep}_D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \mathbb{E}_{z \sim D} f(z) - \underbrace{\frac{1}{m} \sum_{i=1}^m f(z_i)}_{\mathbb{E}_S f \text{ shorthand}}$

so by observation, we just need $\mathrm{Rep}_D(\mathcal{F}, S) \le$ RHS.

Idea: use McDiarmid's, applied to $\mathrm{Rep}_D(\mathcal{F}, S) = \mathrm{Rep}(S)$ drop D, $\mathcal{F}$ notation since those are fixed

Check the assumptions:

Let $S = S'$ except some $z_i \neq z_i'$

[We'll use $\sup_a g(a) = \sup_a \left(g(a) - h(a) + h(a) \right)$
$\leq \sup_a \left(g(a) - h(a) \right) + \sup_a h(a)$ like earlier]

$$\mathrm{Rep}(S) - \mathrm{Rep}(S') := \left(\sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_S f \right) - \left(\sup_{f \in \mathcal{F}} \mathbb{E} f - \hat{\mathbb{E}}_{S'} f \right)$$

$$\leq \sup_{f \in \mathcal{F}} \left(\left(\cancel{\mathbb{E} f} - \hat{\mathbb{E}}_S f \right) - \left(\cancel{\mathbb{E} f} - \hat{\mathbb{E}}_{S'} f \right) \right) \quad \text{via}$$

$$= \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \underbrace{\left(f(z_i') - f(z_i) \right)}_{\leq 1 \text{ if } f : X \to [0,1]} \right)$$

$$\leq 1/m .$$

So $\forall i \in [m]$, the "c_i" in McDiarmid's is $1/m$

Applying McDiarmid's

$$\mathbb{P}\left[\mathrm{Rep}(S) - \mathbb{E}\,\mathrm{Rep}(S) \geq \varepsilon \right] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_i (1/m)^2} \right) = \exp(-2\varepsilon^2 m) =: \delta$$

Now fix δ, solve for ε

$$-2\varepsilon^2 m = \log(\delta), \quad \varepsilon = \sqrt{-\frac{\log(\delta)}{2m}} = \sqrt{\frac{\log(1/\delta)}{2m}}$$

So w.p. $\geq 1 - \delta$,

$$\mathrm{Rep}(S) \leq \mathop{\mathbb{E}}_{S \sim D^m} \mathrm{Rep}(S) + \varepsilon$$

$$\leq 2 R_m(\mathcal{F}) + \varepsilon \quad \text{via previous lemma.}$$

which proves A

To prove B, restate A using $\delta/2$:

w.p. $\geq 1 - \delta/2$, $\mathrm{Rep}(S) \leq 2 R_m(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2m}}$ $(**)$

Recall $R_m(\mathcal{F}) := \mathop{\mathbb{E}}_S \underbrace{\hat{R}_S(\mathcal{F})}_{\text{write as } \hat{R}(S) \text{ since } \mathcal{F} \text{ fixed for now}}$

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

Let's see if we can't apply McDiarmid's again, but this time the $\mathbb{P}[\ldots \leq -\varepsilon] \leq \ldots$ side instead of $\mathbb{P}[\ldots \geq \varepsilon] \leq \ldots$

Turns out $\hat{R}(S) - \hat{R}(S') \leq 1/m$ $\quad S = (z_1, \ldots, z_m)$
$S' = (z_1, \ldots, z_i', \ldots, z_m)$

just like for $\text{Rep}(S)$ (only difference is now $\mathbb{E}_\sigma(\ldots)$ which is linear and doesn't affect anything)

So

McDiarmid: $\mathbb{P}[\hat{R}(S) - \underbrace{\mathbb{E}\hat{R}(S)}_{R_m} \leq -\varepsilon] \leq \underbrace{\exp(-2\varepsilon^2 m) =: \delta/2}_{\text{so } \varepsilon = \sqrt{\frac{\log(2/\delta)}{2m}}}$

i.e. $\mathbb{P}[\hat{R}(S) - R_m(\mathcal{F}) > -\varepsilon] \geq 1 - \delta/2$

So (**)

w.p. $\geq 1-\delta/2$ $\quad \text{Rep}(S) \leq 2 \cdot R_m(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2m}}$

and

w.p. $\geq 1-\delta/2$ $\quad R_m(\mathcal{F}) < \hat{R}_S(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2m}}$

So combining :

w.p. $\geq 1-\delta$, $\quad \text{Rep}(S) \leq 2\hat{R}_S(\mathcal{F}) + 3 \cdot \sqrt{\frac{\log(2/\delta)}{2m}}$ B $\square$

Specific Application : Binary classification, 0-1 loss (Lemma 3.4 Mohri)

$\mathcal{H}$ a set of functions (hypotheses) of the form $h : X \to \mathcal{Y} := \{\pm 1\}$

and use 0-1 loss as usual, $\ell(h,(x,y)) := \mathbb{1}_{h(x)\neq y}$

$S = ((x_1,y_1), \ldots, (x_m,y_m))$

$S_x = (x_1, \ldots, x_m)$

Observe $\underbrace{\hat{R}_S(\ell \circ \mathcal{H})}_{\text{what we care about}} := \mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \cdot \underbrace{\mathbb{1}_{h(x_i)\neq y_i}}_{= \frac{\cancel{1} - y_i h(x_i)}{2}}$

check: $y_i h(x_i) = \pm 1$ +1 if agree -1 else

$= \frac{1}{2}\mathbb{E}_\sigma \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m h(x_i)$

$\mathbb{E}\sigma_i \cdot 1 = 0$

$\mathbb{E} - \sigma_i y_i = \mathbb{E}\sigma_i y_i$

$\mathbb{E}\sigma_i y_i h(x_i) = \mathbb{E}\sigma_i h(x_i)$

since $y_i \in \{\pm 1\} \Rightarrow \sigma_i y_i \sim$ Rad.

$= \frac{1}{2}\underbrace{\hat{R}_{S_x}(\mathcal{H})}$

No reference to labels or loss (0-1 is implicit)

only $\mathcal{H}$, D_x

$$\boxed{\begin{aligned} &\text{So } \hat{R}_S(\ell\circ\mathcal{H}) = \tfrac{1}{2}\hat{R}(\mathcal{H}) \\ &\text{and similarly} \\ &\quad R_m(\ell\circ\mathcal{H}) = \tfrac{1}{2}R_m(\mathcal{H}) \end{aligned}}$$

[implicit: labels $y = f(x_i)$] Not agnostic

and now apply our Thm to this case

Thm 3.5 [Mohri] Rademacher Complexity thm for binary classif. w/ 0-1 loss

Let $\mathcal{H} \subseteq \{\pm 1\}^X$, D any distribution X, Then $\forall \delta \in (0,1)$, w.p. $\geq 1-\delta$ (wrt $S_x \sim \mathcal{D}^m$ iid), $\forall h \in \mathcal{H}$

$$L_{\mathcal{D}}(h) \leq \hat{L}_S(h) + \begin{cases} R_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{m}} \\ \hat{R}_{S_x}(\mathcal{H}) + 3 \cdot \sqrt{\frac{\log(2/\delta)}{2m}} \end{cases}$$

proof via previous lemma.

Typically $R_m(\mathcal{H}) = O(1/m)$ or $O(1/\sqrt{m})$. Either way,

we're bounding our estimation error by $O(1/\sqrt{m})$

so need $m = \Omega(1/\varepsilon_{est}^2)$

$R_m(\mathcal{H})$ is one measure of complexity of $\mathcal{H}$

Also, we could possibly compute it via optimization:

$$\hat{R}_S(\mathcal{H}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{H}} \frac{1}{m} \sum_i \sigma_i f(z_i)$$

$$= -\mathbb{E}_\sigma \inf_{f \in \mathcal{H}} \frac{1}{m} \sum_i \tilde{\sigma}_i f(z_i) \qquad \tilde{\sigma}_i = -\sigma_i, \text{ both} \sim \text{Rademach}$$

but at least as hard as solving

$$\inf_{f \in \mathcal{H}} \frac{1}{m} \sum_i f(z_i) = ERM_{\mathcal{H}}$$

and as we'll see, that's often intractable.

But for simple $\mathcal{H}$, we can often compute (or at least upper bound) $\hat{R}_S$ or R_m by hand