4. Finite hypothesis classes, PAC

Saturday, January 8, 2022 5:00 PM

How to analyze and pick good hypothesis classes $\mathcal{H}$

Ex. §2.3.1 SS Finite hypothesis class, i.e., $|\mathcal{H}| < \infty$
(so <u>not</u> axis-aligned rectangles: that had finite <u>dimension</u>)

Motivation: $\mathcal{H} = \{$ all functions we can implement in 100 lines (80 char. wide) of C++ code $\}$

or

$\mathcal{H} = \{$ axis-aligned rectangles in $\mathbb{R}^2$, within a <u>bounded set</u>, and discretized (eg. $\underbrace{\text{IEEE double precision float}}_{\text{64 bits so } |\mathcal{H}| \le 2^{64}}$) $\}$

We're going to start with the assumption of realizability:

Assume $\exists\, h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D},f}(h^*) = 0$, i.e., $\mathcal{H}$ is large enough
or
our <u>bias</u> is harmless.

Q: could we find this h^* via ERM?
(let's restrict to $\underbrace{\text{binary}}$ classification)

true labeling function

so $L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}\left(h(x) \neq f(x)\right)$

so $$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\hat{L}_S(h^*) = 0\right) = \prod_{i=1}^m \mathbb{P}_{x_i\sim\mathcal{D}}\left(h^*(x_i) = f(x_i)\right) \quad \text{since iid}$$
$$= \prod_{i=1}^m 1 = 1$$

so $\hat{L}_S(h^*) = 0$ w.p. 1, and since $\hat{L}_S(h) \geq 0 \ \forall h$,
this means, with probability 1, $h^* \in \operatorname{argmin}_h \hat{L}_S(h)$

... but, there may be more than 1 minimizer, so that's a problem.
(it does tell us that if $h \in \mathrm{ERM}_{\mathcal{H}}(S)$ then $\hat{L}_S(h) = 0$)

So we'll pick <u>some</u> $h_S \in \mathrm{ERM}_{\mathcal{H}}(S) := \operatorname{argmin}_h \hat{L}_S(h)$, which is a <u>random variable</u>. What kind of analysis can we do?

Can we hope to say $L_{\mathcal{D},f}(h_S) = 0$

or $L_{\mathcal{D},f}(h_S) \leq \varepsilon$? No! It's random.

We're instead going to ask for a

$\underbrace{\text{probably}}_{\delta}$, $\underbrace{\text{approximately}}_{\varepsilon}$ correct (PAC) learner:

$\forall\ \varepsilon, \delta \in (0,1)$, $\underbrace{\text{w.p.}}_{\text{with probability}} \geq 1-\delta$,

$$L_{\mathcal{D},f}(h_S) \leq \varepsilon$$

i.e., $\mathcal{D}^m(\{S : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \leq \delta$

"PAC" sounds fancy but it's already common in statistics,
e.g., a 95% confidence interval means $\delta = 0.05$
and ε is half the width of the confidence interval

Our first analysis: a (specialized) analysis of when $|\mathcal{H}| < \infty$