Ch 20 Neural Networks: sample complexity

Friday, March 27, 2020 3:05 PM

i.e. II Estimation / generalization error

In direct competition w/ approximation error, now we want $\mathcal{H}$ to be small (low-complexity)

We'll analyze by bounding VC dimension.

We learn $|E|$ parameters (weights), so via discretization trick, we can think of

$\mathcal{H}_{V,E,\sigma}$ as having finite size $|E|\cdot$ constant, so expect $\text{VCdim}(\mathcal{H}_{V,E,\sigma}) \approx O(|E|)$

but large, like 64

Case 1: $\sigma = \text{sign}$

Thm 20.6 $\text{VCdim}(\overbrace{\mathcal{H}_{V,E,\text{sign}}}^{\mathcal{H}}) = O(|E| \cdot \log(|E|))$

Proof Recall growth fcn $\tau_{\mathcal{H}}(m) = \max_{\substack{C \subseteq X \\ |C|=m}} |\mathcal{H}_C| \le 2^m$ i.e. # dichotomies.

Can extend to any finite $\mathcal{Y}$, not just $\mathcal{Y}=\{0,1\}$, if want to

Exercise 20.4: showed that if $\mathcal{H} = \{f \circ g,\ f \in \mathcal{H}_1,\ g \in \mathcal{H}_2\}$ then

$$\tau_{\mathcal{H}}(m) \le \tau_{\mathcal{H}_1}(m) \cdot \tau_{\mathcal{H}_2}(m)$$

Using the ANN's layered structure, $\mathcal{H} = \mathcal{H}^{(T)} \circ \ldots \circ \mathcal{H}^{(2)} \circ \mathcal{H}^{(1)}$

and each layer maps $\mathbb{R}^{|V_{t-1}|} \to \{\pm 1\}^{|V_t|}$

since sign(...)

Also, each layer $\mathcal{H}^{(t)}$ is a direct product of its individual neurons (all independent)

$$\mathcal{H}^{(t)} = \mathcal{H}^{(t,1)} \times \mathcal{H}^{(t,2)} \times \ldots \times \mathcal{H}^{(t,|V_t|)}$$

and using exercise 20.3,

$$\tau_{\mathcal{H}^{(t)}}(m) \le \prod_{i \in [V_t]} \tau_{\mathcal{H}^{(t,i)}}(m)$$

halfspace classifier, so $\text{VCdim}(\mathcal{H}^{(t,i)})$ = dimension of input

$= d_{t,i}$ (# edges arriving to i^{th} neuron in layer t)

and Sauer's Lemma:

$$\tau_{\mathcal{H}^{(t,i)}}(m) \le \left(\frac{e \cdot m}{d_{t,i}}\right)^{d_{t,i}} \le (e \cdot m)^{d_{t,i}}$$

So, combining everything,

$$\tau_{\mathcal{H}}(m) \le \prod_t \prod_{i \in [V_t]} (e \cdot m)^{d_{t,i}} = (e \cdot m)^{\left(\sum_t \sum_i d_{t,i}\right)} = (em)^{\overset{\#\text{edges}}{|E|}}$$

now "reverse Sauer's lemma" using Lemma A.2:

If $VCdim(\mathcal{H}) = m \Rightarrow \tau_{\mathcal{H}}(m) \geq 2^m$, i.e., $2^m \leq (em)^{|E|}$

i.e. $m \leq |E| \cdot \log_2(em)$

So Lemma A.2 $\Rightarrow$ $\underline{m \leq 4 \cdot |E| \cdot \log(2|E|) + \text{const.}}$ $\Rightarrow VCdim(\mathcal{H}_{V,E,\text{sign}}) \leq O(|E| \cdot \log(|E|))$ $\square$

Case 2: σ = sigmoid

Exercise 20.5: $VCdim(\mathcal{H}_{V,E,\sigma}) = \Omega(|E|^2)$

↑ i.e. at least. BAD!

and (not proven) $VCdim(\mathcal{H}_{V,E,\sigma}) \leq O(|E|^2 \cdot |V|^2)$ also bad

though in practice $VCdim(\mathcal{H}_{V,E,\sigma}) \simeq 64 \cdot O(|E|)$

via discretization trick w/ 64 bits

Case 3: σ = ReLU (not in our book)

Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piece-wise linear neural networks. COLT 2017

let $\mathcal{H}_{V,E,\sigma}$ have $|E|$ edges (weights), T layers

Thm (lower bound) $VCdim(\mathcal{H}_{V,E,r}) \geq \Omega\left(|E| \cdot T \cdot \log\left(\frac{|E|}{T}\right)\right)$

Thm (upper bound) $VCdim(\mathcal{H}_{V,E,r}) \leq O(|E| \cdot T \cdot \log(|E|))$

Since T is mild (even for a "deep network", $T = O(10)$),

this is saying $VCdim \approx |E| \cdot \log(|E|)$, same result we derived for σ=sign case