ML Theory Lecture 4

Matus Telgarsky

1 Miscellaneous

- Last time we mentioned *decision lists* — decision trees where all internal nodes form a single path. We haven't discussed learning algorithms yet, but as a cautionary note: while decision lists can be learned in polynomial time when there exists a decision list consistent with a data set (meaning it labels the data perfectly), when there is no such perfect decision list, the problem is NP-hard.

2 Box representation: 3 layer ReLU networks

A simple model of neural networks is functions of the form

$$x \mapsto \sigma_L(A_L\sigma_{L-1}(A_{L-1}\cdots\sigma_1(A_1x+b_1)\cdots+b_{L-1})+b_L),$$

where weights and biases $((A_i,b_i))_{i=1}^n$ are the parameters found during training, and $(\sigma_1,\ldots,\sigma_L)$ are fixed nonlinearities (they are not modified during training).

Remark 2.1. Some conventions.

- Old choice for σ_i is coordinate-wise it applies $r \mapsto \mathbb{1}[r \geq 0]$ or $r \mapsto {}^1/_{1+\exp(-r)}$.
- Contemporary choice is Lipschitz and continuous. E.g., either coordinate-wise *ReLU* $r \mapsto \max\{0,r\}$, or *max-pooling*, which replaces groups of coordinates with their maximum *[picture dawn in class]*.

  AFAIK: popularized by ImageNet paper; ReLU might be fundamental to resurgence?
- σ_L typically identity (or softmax, which equivalently can be included in the (cross entropy) loss).
- A_i really is just a linear operator. It may be written in a funny way though (e.g., *convnet*).
- Sometimes we'll drop b_i for convenience; maybe have hwk question on this.
- Can also be interpreted as a graph/network. *[Picture drawn in class; "layers" defined, layer 0.]*
- Obvious nice property: easy to adapt to complicated input/output domains.
- Real question ("non-mathematical"): why is this the function class that's taking over? Not just a representation question.

◇

Theorem 2.2. *Consider standard 3 layer ReLU networks, meaning $\sigma_L(r) = r$ (last layer has no nonlinearity), whereas other nodes use the ReLU nonlinearity $r \mapsto \max\{0,r\}$. Then for every continuous function g and every $\epsilon > 0$, there exists a function f written as a 3 layer ReLU network such that $\|f-g\|_1 \leq \epsilon$.*

Remark 2.3. Next lecture we'll strengthen this result in various ways: we'll use $\|\cdot\|_{\mathrm{u}}$ not $\|\cdot\|_1$, we'll allow many choices of σ, and we'll use only 2 layers. But the proof will be nonconstructive and unenlightening. . . ◇

To prove Theorem 2.2, we'll use the following lemma.

Lemma 2.4. *Suppose a function class $\mathcal{F}$ is given such that for any rectangle R and $\tau > 0$, there exists $g \in \mathcal{F}$ with $\|g - \mathbb{1}_R\|_1 \le \tau$. Then for every continuous function f and $\epsilon > 0$, there exists $h \in \operatorname{span}(\mathcal{F})$ with $\|f - h\|_1 \le \epsilon$.*

Remark 2.5. We're going to use this lemma only for 3 layer networks, but it applies to any class of functions that can approximate "bumps" (and linear combinations thereof). ◇

Proof. Applying the piecewise constant approximation lemma from last lecture, let $(R_1, \dots, R_N)$ be a partition of $[0,1]^d$ and $h_0 = \sum_{i=1}^N \alpha_i \mathbb{1}_{R_i}$ be a piecewise constant function so that $\|h_0 - f\|_{\mathrm{u}} \le \epsilon/2$. Define $A := \sum_i |\alpha_i|$; if $A = 0$, then $h = 0 \in \operatorname{span}(\mathcal{F})$ satisfies

$$\|h - f\|_1 = \|h_0 - f\|_1 = \int_{[0,1]^d} |h_0(x) - f(x)|\,\mathrm{d}x \le \int_{[0,1]^d} \|h_0 - f\|_{\mathrm{u}}\,\mathrm{d}x \le \frac{\epsilon}{2},$$

therefore suppose $A > 0$. For each $i \in \{1, \dots, N\}$, by the assumption on $\mathcal{F}$, choose g_i so that $\|g_i - \mathbb{1}_{R_i}\|_1 \le \frac{\epsilon}{2A}$. Setting $h := \sum_i \alpha_i g_i$,

$$\begin{aligned}
\|h - f\|_1 &\le \|h - h_0\|_1 + \|h_0 - f\|_1 \\
&= \int_{[0,1]^d} \Big| \sum_i \alpha_i g_i - \sum_i \alpha_i \mathbb{1}_{R_i} \Big|\,\mathrm{d}x + \int_{[0,1]^d} \big| h_0(x) - f(x) \big|\,\mathrm{d}x \\
&\le \int_{[0,1]^d} \sum_i |\alpha_i| \big| g_i - \mathbb{1}_{R_i} \big|\,\mathrm{d}x + \int_{[0,1]^d} \big\| h_0 - f \big\|_{\mathrm{u}}\,\mathrm{d}x \\
&\le \sum_i |\alpha_i| \int_{[0,1]^d} \big| g_i - \mathbb{1}_{R_i} \big|\,\mathrm{d}x + \frac{\epsilon}{2} \le \sum_i |\alpha_i| \left(\frac{\epsilon}{2A} \right) + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

□

Proof of Theorem 2.2. [Proof had tons of pictures in class.]

The proof proceeds in two steps.

1. First we show that for any rectangle $R \subseteq [0,1]^d$, there exists a network with a single ReLU layer, meaning a function of the form $x \mapsto \sigma(A_2\sigma(A_1 x + b_1) + b_2)$, which approximates $\mathbb{1}_R$ in the $\|\cdot\|_1$ norm.
2. By Lemma 2.4, a linear combination of functions of the preceding form approximates continuous functions. But we can use the last affine combination layer to compute this linear combination, thus completing the proof.

So let's suppose a rectangle $R := \times_{i=1}^d [a_i, b_i] \subseteq [0,1]^d$ and scalar $\tau > 0$ are given. The idea of the proof is as follows. It is easy to build an indicator for an interval (univariate rectangle) using a linear combination of nodes. If we try combining these for each dimension, we don't get quite what we want, and we need to do some cleanup with another layer.

In more detail, let $\delta > 0$ be artibrary (we'll pick a value later), fix a dimension $i \in \{1, \dots, d\}$, and define

$$f_i(x) := \sigma\left(\frac{x_i - a_i}{\delta} + 1\right) - \sigma\left(\frac{x_i - a_i}{\delta}\right) - \sigma\left(\frac{x_i - b_i}{\delta}\right) + \sigma\left(\frac{x_i - b_i - \delta}{\delta}\right).$$

Thus $f_i(x) = 1$ when $x_i \in [a_i, b_i]$, $f_i(x) = 0$ when $x_i \le a_i - \delta$ or $x_i \ge b_i + \delta$, and for the remaining strips around $[a_i, b_i]$, f_i linearly interpolates (and thus lies with $[0,1]$.

To start with the multivariate case, consider what's wrong with the mapping $x \mapsto \sum_i f_i(x)$. This is equal to d within R, but it is large elsewhere. Note however that at least one f_i is 0 whenever we are δ away from R along any axis, and therefore $\sum_i f_i$ is at most $d - 1$ whenever we are at least δ away. Thus define

$$f_R(x) := \sigma\left(\left(\sum_i f_i(x)\right) - (d - 1)\right).$$

(Picture drawn in $\mathbb{R}^2$ in class: without the outer σ, the function $\sum_i f_i$ is correct on R, but lots of slop elsewhere; subtracting $(d-1)$ and applying a ReLU cleans this up.) Summarizing what we said before,

$$f_R(x) \begin{cases} = 1 & x \in R, \\ = 0 & \inf_{y \in R} \|x - y\|_\infty \geq \delta, \\ \in [0,1) & \text{otherwise.} \end{cases}$$

Setting $s_i := b_i - a_i$ for convenience,

$$\begin{aligned} \|f_R - \mathbb{1}_R\|_1 &\leq \operatorname{vol}\left(\times_{i=1}^d [a_i - \delta, b_i + \delta] \ \setminus \ \times_{i=1}^d [a_i, b_i]\right) \\ &= \operatorname{vol}\left(\times_{i=1}^d [a_i - \delta, b_i + \delta]\right) - \operatorname{vol}\left(\times_{i=1}^d [a_i, b_i]\right) \\ &= \prod_{i=1}^d (s_i + 2\delta) - \prod_{i=1}^d s_i \\ &\leq \sum_{i=1}^d \binom{d}{i} (2\delta)^i =: \star. \end{aligned}$$

This $\star \to 0$ as $\delta \to 0$, there exists δ sufficiently small so that $\star \leq \tau$. □

3 Polynomial fit

[We finished the lecture by introducing polynomial fit and how we will use it. More next time…]

References