Machine Learning Michaelmas 2018

Neural Nets

James Worrell

The following notes are based on Chapter 2 of Wolf [1].

1 Layered Feedforward Neural Nets

A *ridge function* $f : \mathbb{R}^d \to \mathbb{R}$ is given by $f(x) = \sigma(w_0 + \sum_{i=1}^d w_i x_i)$, where $w_0, \ldots, w_d \in \mathbb{R}$ are *weights* and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear *activation function*. The following are some typical activation functions:

1. Step function $\sigma = \mathbb{I}_{z \geq 0}$.
2. Logistic sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$.
3. Hyperbolic tangent $\sigma(z) = \tanh(z) = \frac{e^{2z}-1}{e^{2z}+1}$.
4. Rectified linear unit (ReLU) $\sigma(z) = \max(0, z)$.

Given positive integers d and $m_0, \ldots, m_d$, a function $f : \mathbb{R}^{m_0} \to \mathbb{R}^{m_d}$ is computed by a *layered feedforward neural network* with $d-1$ *hidden layers* and activation function σ if we can write

$$f = g_d \circ \sigma^{m_{d-1}} \circ g_{d-1} \circ \cdots \circ \sigma^{m_1} \circ g_1$$

where $g_i : \mathbb{R}^{m_{i-1}} \to \mathbb{R}^{m_i}$ is an affine map for $i = 1, \ldots, d$ and, for all $m \in \mathbb{N}$, $\sigma^m : \mathbb{R}^m \to \mathbb{R}^m$ is defined by applying σ pointwise. Note that each map g_i can be written in the form $g_i(x) = W^{(i)}x + w^{(i)}$ for $W^{(i)}$ an $m_i \times m_{i-1}$ *weight matrix* and $w^{(i)}$ a m_i-dimensional vector of *biases*. We refer to [1] for a presentation of neural networks as layered directed weighted graphs.

2 Expressiveness

In this section we show how a feedforward neural network with a single hidden layer can approximate continuous functions on compact subsets of $\mathbb{R}^d$. We work under the simplifying assumption that the activation function is bounded and such that $\lim_{x \to -\infty} \sigma(x)$ and $\lim_{x \to +\infty} \sigma(x)$ both exist and are distinct. This assumption applies to the logistic sigmoid and hyperbolic tangent functions. We refer to [1] for an argument that works in case σ is merely continuous and not equal to a polynomial (e.g., the ReLU activation function).

Our starting point is the following result, which essentially says that the class of feedforward neural nets with a single hidden layer and the exponential function as activation function can approximate continuous functions on compact sets.

Given a compact set $K \subseteq \mathbb{R}^d$, we write $C(K)$ for the set of continuous functions $K \to \mathbb{R}$.

Proposition 1. Let $K \subseteq \mathbb{R}^d$ be a compact set for some $d \in \mathbb{N}$. Then the algebra

$$\mathcal{E} = \operatorname{span}\left\{ f : K \to \mathbb{R} : f(x) = \exp\left(\sum_{i=1}^d w_i x_i\right), w \in \mathbb{R}^d \right\}$$

is dense in $C(K)$ with respect to the norm $\| \cdot \|_\infty$.

Proof. Note that $\mathcal{E}$ is indeed an $\mathbb{R}$-algebra (i.e., it is closed under pointwise products and $\mathbb{R}$-linear sums). The result now follows from the Stone-Weierstrass Theorem, which states that any subalgebra of $C(K)$ that contains a non-zero constant function and separates points (i.e., such that for all $x \neq y \in K$ there exists $f \in \mathcal{E}$ such that $f(x) \neq f(y)$) is dense in $C(K)$ with respect to $\|\cdot\|_\infty$. Separation of points is clear in the case at hand: if $x \neq y \in K$ then setting $w = x - y$ we have that $e^{wx} \neq e^{wy}$. □

The idea in the rest of the section is to show how to approximate the exponential function $\exp : \mathbb{R} \to \mathbb{R}$ by a neural net with a single hidden layer, using only the above-stated assumptions on the activation function.

Let $\mathcal{F}_m$ be the set of functions $\mathbb{R} \to \mathbb{R}$ that can be represented by a feedforward neural network with a single hidden layer containing m neurons with σ as activation function. Formally we have

$$\mathcal{F}_m = \left\{ f : \mathbb{R} \to \mathbb{R} : f(x) = \sum_{i=1}^{m} a_i \sigma(w_i x + b_i),\ a_i, b_i, w_i \in \mathbb{R} \right\} .$$

Proposition 2. Let $I \subseteq \mathbb{R}$ be a closed bounded interval. Then there exists a constant c such that for every L-Lipschitz function $f : I \to \mathbb{R}$ and all $m \in \mathbb{N}$ we have

$$\inf_{g \in \mathcal{F}_m} \|f - g\|_\infty \leq c \left(\frac{L}{m} \right) .$$

Proof. We give the proof in the case that $I = [0, 1]$, $\lim_{x \to -\infty} \sigma(x) = 0$, and $\lim_{x \to +\infty} \sigma(x) = 1$. These assumptions simplify the details, but the same argument works in the general case.

The idea is to approximate f by a step function g and then to approximate g by a function $h \in \mathcal{F}_m$. To this end, write $x_i := i/m$ for $i = 0, 1, \ldots, m$ and define $g : [0, 1] \to \mathbb{R}$ by $g(x) = f(x_i)$ for all $x \in [x_{i-1}, x_i)$. Clearly $\|f - g\|_\infty \leq \frac{L}{m}$.

To motivate the definition of the function h, notice that for all $x \in [0, 1]$ we have

$$\begin{aligned} g(x) &= f(x_1) + \sum_{i=1}^{\lfloor mx \rfloor} (f(x_{i+1}) - f(x_i)) \\ &= f(x_1) + \sum_{i=1}^{m-1} (f(x_{i+1}) - f(x_i)) \mathbb{I}_{i \leq \lfloor mx \rfloor} . \end{aligned}$$

We are thus led to define

$$h(x) := f(x_1) + \sum_{i=1}^{m-1} (f(x_{i+1}) - f(x_i)) \sigma(\alpha(mx - i)) ,$$

for some "large" constant $\alpha \in \mathbb{R}$ to be chosen momentarily. Notice that $h \in \mathcal{F}_m$, as desired.

Fix $\varepsilon > 0$ and choose $\alpha > 0$ such that $|1 - \sigma(z)| \leq \varepsilon$ for all $z \geq \alpha$ and $|\sigma(z)| \leq \varepsilon$ for all $z \leq -\alpha$. Fix $x \in [0, 1]$ and write $i_0 := \lfloor mx \rfloor$ Then if $i \notin \{i_0, i_0 + 1\}$ we have that $|mx - i| \geq 1$ and hence

$|\sigma(\alpha(mx-i))-\mathbb{I}_{i\leq\lfloor mx\rfloor}|\leq\varepsilon$. It follows that

$$\begin{aligned}|g(x)-h(x)|\leq&\varepsilon(m-3)\frac{L}{m}\\&+|f(x_{i_0+1})-f(x_{i_0})|\,|1-\sigma(\alpha(mx-i_0))|\\&+|f(x_{i_0+2})-f(x_{i_0+1})|\,|\sigma(\alpha(mx-i_0-1))|\\\leq&\varepsilon L+(\|\sigma\|_\infty+1)\frac{L}{m}\,.\end{aligned}$$

Since ε can be chosen arbitrarily small the result holds. □

Theorem 3. Let $d\in\mathbb{N}$ and $K\subseteq\mathbb{R}^d$ be compact. Given a given continuous function $f:\mathbb{K}\to\mathbb{R}$, we can approximate f can arbitrarily closely with respect to $\|\cdot\|_\infty$ by feedforward neural networks with a single hidden layer.

Proof. Given $f:K\to\mathbb{R}$ we first approximate f by exponentials and then approximate the exponentials by linear combinations of ridge functions.

By Proposition 1, for every $\varepsilon>0$ there exists $k\in\mathbb{N}$, a set of vectors $v_1,\ldots,v_k\in\mathbb{R}^d$, and a weight vector $s\in\mathbb{R}^k$ such that the function $g:\mathbb{R}^d\to\mathbb{R}$ defined by $g(x):=\sum_{i=1}^k s_i e^{v_i\cdot x}$ satisfies $\|f-g\|_\infty\leq\varepsilon/2$.

Define $K_1:=\bigcup_{i=1}^k\{v_i\cdot x:x\in K\}$ and note that K_1 is a compact subset of $\mathbb{R}$. By Proposition 2 (noting that $\exp$ is Lipschitz on some bounded interval containing K_1) there exist $\ell\in\mathbb{N}$ and real numbers a_j,w_j,b_j for $j=1,\ldots,\ell$ such that

$$\left|e^y-\textstyle\sum_{j=1}^\ell a_j\sigma(w_jy-b_j)\right|\leq\dfrac{\varepsilon}{2\|s\|_1}$$

for all $y\in K_1$.

Now given $x\in K$, we have

$$\begin{aligned}&\left|f(x)-\textstyle\sum_{i=1}^k\sum_{j=1}^\ell s_ia_j\sigma(w_jv_i\cdot x-b_j)\right|\\&=\left|f(x)-\textstyle\sum_{i=1}^k s_ie^{v_ix}+\sum_{i=1}^k s_ie^{v_ix}-\sum_{i=1}^k\sum_{j=1}^\ell s_ia_j\sigma(w_jv_i\cdot x-b_j)\right|\\&\leq\left|f(x)-\sum_{i=1}^k s_ie^{v_i\cdot x}\right|+\sum_{i=1}^k|s_i|\left|e^y-\sum_{j=1}^\ell a_j\sigma(w_jy-b_j)\right|\\&\leq\frac{\varepsilon}{2}+\|s\|_1\frac{\varepsilon}{2\|s\|_1}\\&=\varepsilon\,.\end{aligned}$$

□

3 Deeper Networks Can Be Exponentially More Succinct

Given the results of Section 2, one may wonder why to consider neural networks with more than one hidden layer. In this section we give a simple example in which adding depth to a neural network achieves an exponential gain in succinctness. Throughout this section we work with the rectified linear activation function.

Let $\mathcal{F}(\ell, m)$ be the class of functions $\mathbb{R} \to \mathbb{R}$ that can be represented by feedforward neural networks with ℓ layers and m neurons per hidden layer.

Given $f \in \mathcal{F}(\ell, m)$, define as associated classifier $\widetilde{f} : \mathbb{R} \to \{0, 1\}$ by $\widetilde{f}(x) = \mathbb{I}_{f(x) \geq 1/2}$. Given a finite set $S \subseteq \mathbb{R} \times \{0, 1\}$, the empirical risk of f on S is

$$\widehat{R}_S(f) := \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{I}_{\widetilde{f}(x) \neq y} \,.$$

Theorem 4. Let $k \in \mathbb{N}$ and $n = 2^k$. Define $S := \{(x_i, y_i) : i = 0, \ldots, n-1\}$ with $x_i = i/n$ and $y_i = i \bmod 2$. Then

1. There exists $f \in \mathcal{F}(2k, 2)$ such that $\widehat{R}_S(f) = 0$.
2. For every $f \in \mathcal{F}(\ell, m)$ with $m \leq 2^{\frac{k-1}{\ell} - 1}$, we have $\widehat{R}_S(f) \geq 1/4$.

Proof. Define $g(x) = \sigma(2\sigma(x) - 4\sigma(x - \frac{1}{2}))$. Then

$$g(x) = \begin{cases} 0 & \text{if } x \notin [0, 1] \\ 2x & \text{if } x \in [0, 1/2] \\ 1 - 2x & \text{if } x \in [1/2, 1]. \end{cases}$$

Furthermore, write $f := \underbrace{g \circ \cdots \circ g}_{k}$ (see [1] for the graph of f). Then $f \in \mathcal{F}(2k, 2)$ and $f(x_i) = y_i$ for $i = 0, \ldots, n-1$ and hence $\widehat{R}_S(f) = 0$.

On the other hand consider $f \in \mathcal{F}(\ell, m)$. We claim that such an f is piecewise affine, with at most $t := (2m)^\ell$ pieces. This claim can be proved by induction on ℓ. The induction step follows by the following claim.
Claim. If $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ are each piecewise affine with at most p pieces, then given $w_0, \ldots, w_m \in \mathbb{R}$ the function $g(x) := \sigma(w_0 + \sum_{i=1}^m w_i f_i(x))$ is piecewise affine with at most $2mp$ pieces. To prove the claim one notes that there are at most $m(p-1)$ endpoints of the intervals defining the pieces of $f_1, \ldots, f_m$ and hence $x \mapsto w_0 + \sum_{i=1}^m w_i f_i$ is piecewise affine (and hence piecewise monotone) with at most mp pieces. By splitting each of these pieces into at most two parts one obtains a decomposition of $\mathbb{R}$ into at most $2mp$ pieces such that g is affine on each piece.

Since $f \in \mathcal{F}(\ell, m)$ is piecewise monotone with at most t pieces, it crosses the line $y = 1/2$ at most t times. Hence $\widetilde{f}$ is piecewise constant with at most $t + 1$ pieces. It follows that there are at least $\frac{n-t-1}{2}$ consecutive pairs of points in S that both lie in an interval in which $\widetilde{f}$ is constant. By the assumption $m \leq 2^{\frac{k-1}{\ell} - 1}$, we have $\frac{t/2-1}{2n} \leq \frac{1}{4}$. We conclude that

$$\widehat{R}_S(f) \geq \frac{n - t - 1}{2n} = \frac{1}{2} - \frac{t/2 - 1}{2n} \geq \frac{1}{4} \,.$$

□

4 Rademacher Complexity of Neural Networks

We first recall a bound on the Rademacher complexity of classes of linear functions. Then we combine this analysis with general properties of Rademacher complexity to bound the Rademacher complexity of feedforward neural networks by induction on the number of layers.

Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a class of functions. Recall that the *empirical Rademacher complexity* of $\mathcal{H}$ with respect to $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is

$$\widehat{\mathsf{RAD}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

where the Rademacher random variables σ_i are independent and uniformly distributed over $\{-1, +1\}$.

We recall the following properties of Rademacher complexity, where $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ are subsets of $\mathbb{R}^{\mathcal{X}}$ and $\mathrm{conv}(\mathcal{H})$ denotes the convex hull of $\mathcal{H}$:

1. $\widehat{\mathsf{RAD}}_S(c\mathcal{H}) = |c|\widehat{\mathsf{RAD}}_S(\mathcal{H})$ for $c \in \mathbb{R}$.
2. $\widehat{\mathsf{RAD}}_S(\mathcal{H}_1 + \mathcal{H}_2) = \widehat{\mathsf{RAD}}_S(\mathcal{H}_1) + \widehat{\mathsf{RAD}}_S(\mathcal{H}_2)$.
3. $\widehat{\mathsf{RAD}}_S(\mathrm{conv}(\mathcal{H})) = \widehat{\mathsf{RAD}}_S(\mathcal{H})$.
4. Talagrand's Lemma: if $\varphi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz $\widehat{\mathsf{RAD}}_S(\varphi \circ \mathcal{H}) \leq L \cdot \widehat{\mathsf{RAD}}_S(\mathcal{H})$.

In the rest of this section let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty = 1\}$.

Proposition 5 (Rademacher complexity of linear maps)**.** Given $b > 0$, define $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ by $\mathcal{G} = \{x \mapsto x \cdot w : \|w\|_1 \leq b\}$. Then for $S \in \mathcal{X}^n$ we have $\widehat{\mathsf{RAD}}_S(\mathcal{G}) \leq \frac{b}{\sqrt{n}}$.

Proof. See the lecture on Rademacher complexity. □

Proposition 6 (Adding a layer)**.** Let $a, b \in \mathbb{R}$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz. Suppose that $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ contains the zero function and satisfies $\mathcal{F} = -\mathcal{F}$. Define $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ by

$$\mathcal{G} := \{x \mapsto \sigma(w_0 + \textstyle\sum_{i=1}^{m} w_j f_j(x)) : |w_0| \leq a, \|w\|_1 \leq b, f_j \in \mathcal{F}\}.$$

With respect to $S = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ we have

$$\widehat{\mathsf{RAD}}_S(\mathcal{G}) \leq L \left(\frac{a}{\sqrt{n}} + b\widehat{\mathsf{RAD}}_S(\mathcal{F}) \right).$$

Proof. Define $\mathcal{G}_1 := b\,\mathrm{conv}(\mathcal{F})$ and $\mathcal{G}_2 := \{x \mapsto w_0 : |w_0| \leq a\}$. Then

$$\begin{aligned} \widehat{\mathsf{RAD}}_S(\mathcal{G}_2) &= \frac{a}{n}\mathbb{E}_\sigma[|\sigma_1 + \cdots + \sigma_n|] \\ &\leq \frac{a}{n}\left(\mathbb{E}_\sigma[(\sigma_1 + \cdots + \sigma_n)^2]\right)^{1/2} \qquad \text{Jensen's inequality} \\ &= \frac{a}{\sqrt{n}}. \end{aligned}$$

By the assumptions that $0 \in \mathcal{F}$ and $\mathcal{F} = -\mathcal{F}$, we have that

$$\left\{ x \mapsto \sum_{i=1}^{m} w_j f_j(x)) : \|w\|_1 \leq b, f_j \in \mathcal{F} \right\} = b\,\mathrm{conv}(\mathcal{F}).$$

Then from Properties 1,3 of $\widehat{\mathsf{RAD}}_S$ we have $\widehat{\mathsf{RAD}}_S(\mathcal{G}_1) = b\,\widehat{\mathsf{RAD}}_S(\mathcal{H})$.

We conclude that

$$\begin{aligned}\widehat{\mathsf{RAD}}_S(\mathcal{G}) &\leq L(\widehat{\mathsf{RAD}}_S(\mathcal{G}_1 + \mathcal{G}_2)) \quad \text{by Talagrand's Lemma} \\ &= L(\widehat{\mathsf{RAD}}_S(\mathcal{G}_1) + \widehat{\mathsf{RAD}}_S(\mathcal{G}_2)) \quad \text{by Property 2 of } \widehat{\mathsf{RAD}}_S. \\ &\leq L\left(\frac{a}{\sqrt{n}} + b\widehat{\mathsf{RAD}}_S(\mathcal{F})\right).\end{aligned}$$

□

Theorem 7. Let $a, b > 0$. Fix a neural network architecture with $\ell \geq 1$ layers and assume that (i) the activation function σ is 1-Lipschitz and anti-symmetric ($\sigma(-x) = -\sigma(x)$), (ii) the weight vector w and bias v for every at every node in the network satisfies $\|w\|_1 \leq b$ and $|v| \leq a$. Then the class of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ defined by such a network satisfies

$$\widehat{\mathsf{RAD}}_S(\mathcal{F}) \leq \frac{1}{\sqrt{n}}\left(b^\ell + a\sum_{i-0}^{\ell-2} b^i\right).$$

Proof. The theorem follows by a straightforward induction on the number ℓ of layers. The base case $\ell = 1$ reduces to the bound on the Rademacher complexity of linear classifiers in Proposition 5. The induction step uses Proposition 6. Note that thanks to the requirement that σ be anti-symmetric the set of functions computed at each node of the neural network satisfies the assumptions of Proposition 6. □

Note that the above bound has an exponential dependence on the number of layers.

References

[1] Michael A. Wolf. *Mathematical Foundations of Supervised Learning*.

`https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2016S/ML_notes_main.pdf`