Notes on MATH 260 AA

Haosheng Zhou

Sept, 2022

Contents

Introduction to Neural Network

A handwriting recognition example: a dataset consisting of pictures of numbers to train the model. Human's recognition of number 9 is that it's a loop on the top with a vertical stroke in bottom right. The machine recognizes number using neural network, receive pictures as training samples and automatically infer the rules for recognition.

NN has neurons that can be activated or deactivated. It has input variables and for each input variable $x_i \in \mathbb{R}$ there is a weight parameter w_i and the linear combination $w \cdot x$, compared with a certain threshold, produces the output. For example, for sigmoid neurons, the output of a neuron is $\sigma(w \cdot x + b)$, with b as the bias working as the intercept in the linear combination. The sigmoid function is given by $\sigma(z) = \frac{1}{1+e^{-z}}$.

Different NN Structures

The simplest NN is given by the feedforward NN ('feedforward' refers to the fact that there's no edge going back, i.e. the values in hidden layer fed into input layer or hidden layer again), with 1 input layer, 1 hidden layer and 1 output layer, all neurons connected. The NN propogates forward, with the neurons in the hidden layer has the activated result of linear combinations of the real numbers in the input layer neurons. The number of neurons in a layer is called the width. In this specific example, the pictures are formed as 64×64 pixels with grey scales, so that will be the input of the NN.

On the other hand, recurrent NN (RNN) in NLP has loops and can deal with sequential data better and memorize previous inputs. The hidden layers can feed data to itself using a different activation function. Consider a 3-layer RNN and let h_i denote the number in the hidden neuron at time i, x_i as the input at time i into the hidden layer, y_i for the output at time i, then

$$\begin{cases} h_k = f(\beta^1 + w^{1,1}h_{k-1} + w^{1,2}x_k) \\ y_k = \tilde{f}(\beta^2 + w^2 h_k) \end{cases} \tag{1}$$

the point is that h_{k-1} is fed into h_k once more and the activation functions $f, \tilde{f}$ are different. RNN structure helps reduce the parameter space and has the cascade effect.

LSTM is a special case of RNN consisting of units. Each unit has a cell and three gates. The cell keeps track of the information received so far, the input gate captures to which extent new information will flow into the cell and the forget gate captures to which extent existing information remains in the cell, the output gate controls to which extent the information in the cell will be used to compute the output of the unit. The output of the unit can be understood as "given all words seen so far, what's the meaning of the sentence". The gates are formed as

$$\begin{cases} f_k = \rho_s(w_f x_k + u_f h_{k-1} + b_f) \\ i_k = \rho_s(w_i x_k + u_i h_{k-1} + b_i) \\ o_k = \rho_s(w_o x_k + u_o h_{k-1} + b_o) \end{cases} \tag{2}$$

and the cell is formed as $C_k = f_k \odot C_{k-1} + i_k \odot \rho(w_k x_k + u_k h_{k-1} + b_k)$ and the output of the k-th unit is denoted

$h_k = O_k \odot C_k$ where $\odot$ denotes the Hadamard product.

Training

Our objective is to let the NN learn the parameters to form estimates $\hat{w}, \hat{b}$ for the optimal w^*, b^*. One need a cost function to do this, the MSE cost function is formed as

$$C(w,b) = \frac{1}{2n}\sum_x ||Z - Z(x)||^2 \tag{3}$$

where Z is the output of the NN, $Z(x)$ is the true information given in the training set and n is the size of training inputs. Note that we will not choose things like $\frac{1}{2n}\sum_x \mathbb{I}_{Z(x)\neq Z}$ as the cost function since we hope to see that small changes in w, b will result in small changes in loss/cost.

In hand recognition, for example, $Z, Z(x) \in \mathbb{R}^{10}$ since there are altogether 10 possible predictions. $Z(x)$ is always a one-hot vector, having entry 1 at the location of the true label and 0 elsewhere. Entries of $Z(x)$ can stand for the probability of the label appearing given data x, but there are also other formations. Hand recognition is a supervised learning problem since true labels are given in the training set, while for unsupervised learning problems true labels are not known and one has to discover the pattern of the data without any prior information.

Now to find $\hat{w}, \hat{b}$ such that C is as small as possible, apply gradient descent for C. If $C(v)$ depends on parameter v, then $v' = v - \eta\nabla C(v)$ gives the update with learning rate $\eta > 0$. The gradient descent ensures that $C(v') \leq C(v)$ if the Armijo condition is satisfies. Note that the learning rate can't be too large since it needs to make Taylor expansion work. It can't be too small since that will result in slow convergence so in practice the learning rate changes with time.

For the example, we have updates

$$\begin{cases} w_k \to w_k' = w_k - \eta\frac{\partial C}{\partial w_k} \\ b_k \to b_k' = b_k - \eta\frac{\partial C}{\partial b_k} \\ C = \frac{1}{n}\sum_x C_x, C_x = ||Z - Z(x)||^2 \end{cases} \tag{4}$$

the problem is that computing C is too costly in a single update when we have a large training set. Using all information in the training set might not be a good choice. This leads to the appearance of stochastic gradient descent where one only just a part of the data to compute the gradients and to update parameters.

The thought of **stochastic gradient descent (SGD)** is to estimate ∇C by computing ∇C_x for a small set of samples $X_1, ..., X_M$ randomly chosen. M is the **mini-batch size**, typically taken as a power of 2 to speed up. So ∇C is estimated by $\frac{1}{M}\sum_{i=1}^M \nabla C(x_i)$. **Epoch** means the time we have exhausted the training inputs, if there are 60000 inputs and the batch size is 1000, then in each epoch the parameters are updated for 60 times (samples randomly chosen without replacement).

The **backpropagation** procedure is adopted to compute the gradients numerically (details ommitted, refer to Pytorch for realization in practice).

Problems & Improvements

Now under the settings described above, $C = \frac{1}{n}\sum_x ||y-a||^2$ where y is the truth, a is the output of NN where $a = \sigma(z), z = wx+b$. Then $\frac{\partial C}{\partial w} = (a-y)\sigma'(z)x$, $\frac{\partial C}{\partial b} = (a-y)\sigma'(z)$. If σ is not chosen carefully, σ' can be pretty small, leading to **vanishing gradient issue**. That's why ResNet is appearing to fix this issue.

Cross entropy loss is always chosen especially in binary categorization problems.

$$C = -\frac{1}{n}\sum_x y\log(a) + (1-y)\log(1-a) \ (y, a \in [0,1]) \tag{5}$$

where $y = y(x)$ is the true data and $a = \sigma(z), z = wx+b$ is the output of NN. When C is close to 0 and $y = 0$, then a is close to 0, the same property holds when $y = 1$. So it's a good metric measuring the distance between y and a. Notice that

$$\frac{\partial C}{\partial w_j} = \frac{1}{n}\sum_x \frac{\sigma'(z)x_i}{\sigma(z)(1-\sigma(z))}(\sigma(z)-y) = \frac{1}{n}\sum_x x_i(\sigma(z)-y) \tag{6}$$

so the size of the gradient is actually proportional to the error $\sigma(z)-y$, and there's no vanishing gradient issue.

Paper Summary: Universal Approximation Results

We refer to *Multilayer Feedforward Networks are Universal Approximators by Kur Hornik, Maxwell Stinchcombe and Halber White* **and** *Recurrent Neural Networks Are Universal Approximators by Anton Schafer and Hans Zimmermann* **for the following materials**.

The universal approximation results of NN focus on the theoretical functionality of NN that for any error tolerance level $\forall \varepsilon > 0$, NN can approximate any Borel measurable function well enough regardless of the activation function and the input space environment. We first introduce the sketch of the proof that a 3-layer NN is a universal approximator and then use the result to prove that RNN is also a universal approximator.

Settings and Definitions

Denote $\mathscr{A}^r$ as the set of all affine functions $A(x) = w^T x + b$ from $\mathbb{R}^r$ to $\mathbb{R}$. For a given activation function G, a 3-layer NN first maps x to $A_j(x)$ for $A_j \in \mathscr{A}^r$ (w, b are the weight and bias between input layer and hidden layer) and then compose the activation function to get $G(A_j(x))$ as the input of the j-th neuron in the hidden layer. The output of such NN is then formed as $\sum_{j=1}^q \beta_j G(A_j(x))$, so β denotes the weight between hidden layer and output layer. As a result, we denote

$$\Sigma^r(G) = \left\{ g : \mathbb{R}^r \to \mathbb{R} : g(x) = \sum_{j=1}^q \beta_j G(A_j(x)) \right\} \tag{7}$$

as the collection of all mappings from x, the input of NN, to $g(x)$, the output of NN with activation function taken as G. For the activation function, we add some mild conditions that G is a **sigmoid function** if it's increasing and bounded, $G(\lambda) \to 1$ $(\lambda \to +\infty), G(\lambda) \to 0$ $(\lambda \to -\infty)$ (extracted from the property of the true sigmoid function).

Now to measure whether the approximation is good enough, let's define some density arguments in different sense. For metric ρ, **ρ-dense** is the same as that defined in topology, i.e. $S \subset X$ is ρ-dense in $T \subset X$ if and only if

$$\forall \varepsilon > 0, \forall t \in T, \exists s \in S, \rho(s,t) < \varepsilon \tag{8}$$

so all elements in T can be approximated arbitrarily well by some element in S. Set C^r to be the space of all continuous functions on $\mathbb{R}^r$ and M^r to be the space of all measurable functions on $\mathbb{R}^r$. Being **uniformly dense on compact sets in** C^r means that for any compact subset $K \subset C^r$, the set is dense under the metric $\rho_K(f,g) = \sup_{x\in K} |f(x) - g(x)|$ induced by the uniform norm on K. Moreover, if there is a probability measure μ on $(\mathbb{R}^r, \mathscr{B}^r)$ (Borel measurable space), the metric ρ_μ is defined as

$$\rho_\mu(f,g) = \inf \{\varepsilon > 0 : \mu(|f - g| > \varepsilon) < \varepsilon\} \tag{9}$$

one might notice that convergence under this metric is equivalent to convergence in measure μ, i.e. $\rho_\mu(f_n, f) \to 0$ $(n \to \infty)$ iff $f_n \xrightarrow{\mu} f$ $(n \to \infty)$.

Although we want to prove that $\sum^r(G)$ is generally a universal approximator, we do have to start the proof on a space with more mappings, i.e.

$$\Sigma\Pi^r(G) = \left\{ g : \mathbb{R}^r \to \mathbb{R} : g(x) = \sum_{j=1}^{q} \beta_j \prod_{k=1}^{l_j} G(A_{jk}(x)) \right\} \tag{10}$$

so the main thought is to first prove the universal approximation identity to continuous functions for good enough activation functions and a larger space $\Sigma\Pi^r(G)$, and then weaken the approximation to continuous functions into the approximation to measurable functions, weaken the condition of activation functions and shrink the space to $\Sigma^r(G)$.

Sketch of Proof for Feedforward NN

The main theorem used here is the Stone-Weierstrass theorem, telling us that when A, as a set of continuous functions, behaves well enough on a compact set K, its uniform closure includes all continuous functions on K.

Theorem 1. ***(Stone-Weierstrass Theorem)*** *A is an algebra (closed under addition, multiplication, scalar multiplication) of continuous functions on compact set K. If A separates points on K ($\forall x, y \in K, x \neq y, \exists f \in A, f(x) \neq f(y)$) and vanishes at no point of K ($\forall x \in K, \exists f \in A, f(x) \neq 0$), then the uniform closure of A contains all real continuous functions on K, i.e. A is dense under ρ_K.*

We start from the case where G is continuous and prove that it's enough for $\Sigma\Pi^r(G)$ to be uniformly dense on compact set in C^r.

Theorem 2. ***(Approx of Continuous Functions)*** *$\forall G : \mathbb{R} \to \mathbb{R}$ continuous nonconstant, $\Sigma\Pi^r(G)$ is uniformly dense on compact set in C^r.*

Proof. Take $K \subset C^r$ as any compact set, $\Sigma\Pi^r(G)$ is always an algebra on K regardless of G and it separates points on K. To see why it separates points, $\forall x, y \in K, x \neq y$, let's pick $a, b \in \mathbb{R}, a \neq b, G(a) \neq G(b)$ (G nonconstant) and pick $A \in \mathscr{A}^r$ such that it maps x to a, y to b (A affine, always possible), so $G(A(x)) = G(a) \neq G(b) = G(A(y))$.

Let's also verify that $\Sigma\Pi^r(G)$ vanishes at no point of K. Pick $b, G(b) \neq 0$ and set $w = 0$ so $A(x) = b, G(A(x)) = G(b) \neq 0$ for $\forall x \in K$.

By Stone-Weierstrass theorem, $\Sigma\Pi^r(G)$ is ρ_K-dense, so proved. □

Now let's keep G to be continuous but approximate measurable functions instead of continuous ones. Note that the definition of measurable function always depend on the sigma algerba specified (on $(\Omega, \mathscr{F}, \mathbb{P})$ a measurable function f is such that $\forall A \in \mathscr{F}, f^{-1}(A) \in \mathscr{B}_{\mathbb{R}}$). However, here we introduce a probability measure and prove that the approximation is universally good for every probability measure selected.

Theorem 3. ***(Approx of Measurable Functions, Continuous Activation Function)*** *$\forall G : \mathbb{R} \to \mathbb{R}$ continuous nonconstant, $\forall r$ (which is the dimension of input), $\forall \mu$ as probability measure on $(\mathbb{R}^r, \mathscr{B}^r)$, $\Sigma\Pi^r(G)$ is ρ_μ-dense in M^r.*

Proof. The proof is easy since it's just the application of facts in measure theory. By the theorem above, $\Sigma\Pi^r(G)$ is uniformly dense on compact set in C^r, so it must be ρ_μ-dense for any probability measure μ (uniform convergence on compact sets implies convergence in measure). To change the approximation to C^r into M^r, notice that when the measure is finite, C^r is always ρ_μ dense in M^r so we can always approximate C^r first and use continuous functions in C^r to approximate measurable functions in M^r, proved.

□

Let's then weaken the condition of the theorem by not requiring the activation function G to be continuous.

Theorem 4. ***(Approx of Measurable Functions, Sigmoid Activation Function)*** *$\forall G : \mathbb{R} \to \mathbb{R}$ sigmoid, $\forall r$ (which is the dimension of input), $\forall \mu$ as probability measure on $(\mathbb{R}^r, \mathscr{B}^r)$, $\Sigma\Pi^r(G)$ is ρ_μ-dense in M^r and uniformly dense on compact set in C^r.*

Proof. By the last theorem, we just have to prove that there exists F as continuous sigmoid activation function such that $\Sigma\Pi^r(G)$ is uniformly dense on compact set in $\Sigma\Pi^r(F)$. To prove this, we just have to prove that for any F as continuous sigmoid activation function, $\prod_{k=1}^l F(A_k(x))$ can always be uniformly approximated by functions in $\Sigma\Pi^r(G)$. This is because functions in $\Sigma\Pi^r(F)$ are just linear combinations of functions of the form $\prod_{k=1}^l F(A_k(x))$.

Here to approximate continuous sigmoid function by general sigmoid functions, we have to use a lemma that for F continuous sigmoid function fixed and any sigmoid function G, $\forall \varepsilon > 0, \exists H_\varepsilon \in \Sigma^1(G)$ such that $\sup_{\lambda\in\mathbb{R}} |F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$.

Now $\forall \varepsilon > 0$, we can find $H_\delta(x) = \sum_t \beta_t G(A_t^1(x))$ such that $\sup_{\lambda\in\mathbb{R}} |F(\lambda) - H_\delta(\lambda)| < \delta$ is small enough such that $\sup_{x\in\mathbb{R}^r} |\prod_{k=1}^l F(A_k(x)) - \prod_{k=1}^l H_\delta(A_k(x))| < \varepsilon$ (here we are using the uniform continuity of $\prod_{k=1}^l x_k$ on compact set, that's where the boundedness of sigmoid function comes into play and we plug in λ as affine function $A_k(x)$). Now observe that $\prod_{k=1}^l H_\delta(A_k(x)) \in \Sigma\Pi^r(G)$ concludes the proof (note that $H_\delta(A_k(x)) = \sum_t \beta_t G(A_t^1(A_k(x)))$, and the composition of two affine functions is still affine).

□

Finally, we can now go back from $\Sigma\Pi^r(G)$ to $\Sigma^r(G)$ and prove that a 3-layer feedforward NN is already a universal approximator.

Theorem 5. ***(Universal Approximation Theorem for 3-layer Feedforward NN)*** *$\forall G : \mathbb{R} \to \mathbb{R}$ sigmoid, $\forall r$ (which is the dimension of input), $\forall \mu$ as probability measure on $(\mathbb{R}^r, \mathscr{B}^r)$, $\Sigma^r(G)$ is ρ_μ-dense in M^r and uniformly dense on compact set in C^r.*

Proof. We only need to prove that $\Sigma^r(G)$ is uniformly dense on compact set in C^r. That's because this implies that $\Sigma^r(G)$ is ρ_μ-dense in C^r and C^r is ρ_μ-dense in M^r so the theorem is proved. (the same procedure as we have done above).

The proof of the first fact is based on the observation that by taking G as the cosine activation function in the theorem above. Since it's too technical, we omit the details here. □

Remark. *The universality comes from the fact that* ***the density argument holds for any sigmoid activation function, any dimension of input and any input space environment*** *μ. In other words, an NN with simple structure can approximate a measurable function arbitrarily well (under the metric of convergence in measure μ) on a compact set regardless of the selection of activation function, the dimension of NN and the probability measure selected on the input space.*

Remark. *Actually, **the same argument holds for the 3-layer feedforward NN that has multiple outputs**, i.e. the output is a vector in $\mathbb{R}^s$. Just replace ρ_μ with its multi-dimensional version $\rho^s_\mu(f,g) = \sum_{i=1}^s \rho_\mu(f_i, g_i)$ ($f,g : \mathbb{R}^r \to \mathbb{R}^s$) and approximate each component respectively and reduce the error tolerance level such that the sum of error tolerance level adds up to ε.*

Result for RNN

RNN has a different structure from feedforward NN that each neuron will feed the state output back to itself as the state input in the next round. Despite the structural difference, RNN also has universal approximation property directly coming from the universal approximation property of feedforward NN.

Now we have a dynamical system in discrete time

$$\begin{cases} s_{t+1} = g(s_t, u_t) \\ y_t = h(s_t) \end{cases} \tag{11}$$

where $s_t \in \mathbb{R}^J$ is the state at time t, $u_t \in \mathbb{R}^I$ is the external input at time t and $y_t \in \mathbb{R}^n$ is the output at time t. So the system has some dynamics for state evolution that would be influenced by external inputs, and different outputs are produced for different current state. Here we propose the RNN as

$$\begin{cases} s_{t+1} = f(As_t + Bu_t - \theta) \\ y_t = Cs_t \end{cases} \tag{12}$$

with all actions to be linear except the activation function f. Now the input u_t has dimension I so we naturally consider $\Sigma^{I,n}(f)$ which is the set of functions that maps the input of feedforward NN to its output (the output has dimension n). We denote its element as $NN(x)$ such that

$$NN(x) = Vf(Wx - \theta) \tag{13}$$

where V is the weight matrix between hidden layer and output layer, W, θ are the weight matrix and bias between input layer and hidden layer, and the action of f is componentwise. We denote $RNN^{I,n}(f)$ as the set of functions

$$\begin{cases} s_{t+1} = f(As_t + Bu_t - \theta) \\ y_t = Cs_t \end{cases} \tag{14}$$

in RNN for fixed activation function $f : \mathbb{R}^J \to \mathbb{R}^J$ (which is a series of states and outputs as a dynamical system). Of course we hope that RNN can replicate any dynamical system

$$\begin{cases} s_{t+1} = g(s_t, u_t) \\ y_t = h(s_t) \end{cases} \tag{15}$$

well enough in the universal sense. The result is given below.

Theorem 6. ***(Universal Approximation Theorem for RNN)*** *Assume $g : \mathbb{R}^J \times \mathbb{R}^I \to \mathbb{R}^J$ is measurable and $h : \mathbb{R}^J \to \mathbb{R}^n$ is continuous in the real dynamical system with finite time horizon $t = 1, 2, ..., T$. Then any dynamical system of the form*

$$\begin{cases} s_{t+1} = g(s_t, u_t) \\ y_t = h(s_t) \end{cases} \tag{16}$$

can be approximated by an element of $RNN^{I,n}(f)$ arbitrarily well for some continuous sigmoid activation function f.

Proof. The proof is mainly repeated use of the universal approximation theorem for feedforward NN. First approximate the state dynamics $s_{t+1} = g(s_t, u_t)$ by an NN to get the estimation of state as $\overline{s}_t$ at time t. Next approximate $h(\overline{s}_t)$ by feedforward NN again and merge the results in two parts to conclude the proof.

□

Other Results

Hornik also shows that a 3-layer feedforward NN is enough to do universal approximations to any measurable function and its derivative at the same time, which is quite important for solving stochastic control problems since in HJB equation we always have to deal with approximating derivatives. Most work on shallow networks are done.

Stochastic Control Problem