A mathematical understanding of deep learning

Jiang J.
Data Engineer, Data Scientist
ENSEEIHT Computer Science Engineering Degree, INP Toulouse Dual MSc Research Degree in AI / Big Data / Ops
France

1 – Introduction

Frameworks such as TensorFlow or PyTorch make deep learning developments easy. They have made this field wide spread for every enthusiast. Implementations only needs an instinctive understanding of deep learning. The proper math aspect is little by little forgotten.

The objective is to do a summary of the important propositions. These propositions will be mathematically proven. The subject tackled is a multi-class classification problem with – dense layers, ReLU and Softmax activation layers, Categorical cross-entropy loss, Stochastic gradient descent optimizer. However, all the elements below can be re-used or easily modified to cover a regression problem.

2 – Notation and Nomenclature

Definition – Let I a non-empty interval of $\Re$, and $f : I \to \Re$. Then f continuous function is equivalent to

$$\forall a \in I \ ,$$

$$\forall \epsilon > 0 \ , \ \exists \eta > 0 \ , \ \forall x \in I \ , \ |x-a| \leq \eta \Rightarrow |f(x)-f(a)| \leq \epsilon$$

The notation ζ^0 means the set of continuous functions from $\Re$ to $\Re$.

Definition – Let I a non-empty interval of $\Re$, and $f : I \to \Re$. Then f derivable is equivalent to

$$\forall a \in I \ , \ \exists f'(a) \in \Re \ ,$$

$$\forall \epsilon > 0 \ , \ \exists \eta > 0 \ , \ \forall x \in I \ , \ |x-a| \leq \eta \Rightarrow \left| \frac{f(x)-f(a)}{x-a} - f'(a) \right| \leq \epsilon$$

The notation $D^0(I,\Re)$ means the set of derivable functions from I to $\Re$.

Let I a non-empty interval of $\Re$, and $f: I \to \Re$. Then f piece-wise derivable is equivalent to

$\forall [a,b] \in I$, $\exists a = a_0 < a_1 < \cdots < a_n = b$, $\forall i \in \{0,\cdots,n-1\}$, $\exists f_i \in D^0([a_i, a_{i+1}],\Re)$ with f_i = f on]a_i,a_{i+1}[and f_i in D^0([a_i, a_{i+1}]

Definition – Let $a_{i,j} \in \Re$, $i \in \{1,2,\ldots,n\}$, $j \in \{1,2,\ldots,m\}$. Then the ordered rectangular array

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}$$

is a real matrix of dimension $n \times m$.

The following notations are considered

$$\forall i \in \{1,2,\ldots,n\} \ , \ \forall j \in \{1,2,\ldots,m\} \ , \ [A]_{i,j} = a_{i,j}$$

$$\forall j \in \{1,2,\ldots,m\} \ , \ [A]_{:,j} = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{n,j} \end{bmatrix}$$

$$\forall i \in \{1,2,\ldots,n\} \ , \ \forall j \in \{1,2,\ldots,m\} \ , \ [A]_{i,:} = \begin{bmatrix} a_{i,1} & a_{i,2} & \cdots & a_{i,n} \end{bmatrix}$$

The notation $M_{n,m}$ means the real matrix set of dimension $n \times m$.

Convention – A vector is a matrix with only one row. Thus, the real vector set $\Re^n$ is equivalent to $M_{1,n}$.

Notation – The matrix transpose operation will be noted as A^T .

Definition – Let $A \in M_{n,m}$, and $B \in M_{m,p}$, and let the product noted $A \times B$ or AB be

$$C = A \times B = AB$$

where C is a $m \times p$ matrix with

$$\forall i \in \{1,2,\ldots,n\} \ , \ \forall j \in \{1,2,\ldots,m\} \ ,$$

$$[C]_{i,j} = \sum_{k=1}^{m} [A]_{i,k} * [B]_{k,j}$$

3 – Activation functions

Definition – Let $z \in M_{n,m}$, and $f \in \xi^0 \cap D^0$. Then the element wise application

$$\forall i \in \{1,2,\dots,n\} \quad , \quad \forall j \in \{1,2,\dots,m\} \quad ,$$

$$f : \begin{cases} \Re \to \Re \\ [z]_{i,j} \mapsto f([z]_{i,j}) \end{cases}$$

is an activation function.

Proposition – Let $z \in M_{n,m}$. Then the following element wise applications are activation functions:

$$\forall i \in \{1,2,\dots,n\} \quad , \quad \forall j \in \{1,2,\dots,m\} \quad ,$$

$$ReLU : \begin{cases} \Re \to \Re \\ [z]_{i,j} \mapsto max(0, [z]_{i,j}) \end{cases}$$

with max the maximum between two scalars.

$$\forall i \in \{1,2,\dots,n\} \quad , \quad \forall j \in \{1,2,\dots,m\} \quad ,$$

$$Softmax : \begin{cases} \Re^m \to \Re^m \\ [z]_{i,j} \mapsto e^{[z]_{i,j}} \Big/ \sum_{j'=1}^{m} e^{[z]_{i,j'}} \end{cases}$$

with e the exponential operation.

The derivatives are the following

$$D(ReLU) : \begin{cases} \Re \to \Re \\ [z]_{i,:} \mapsto \begin{bmatrix} 1_{\Re_{\setminus\{0\}+}}([z]_{i,1}) & 0 & \cdots & 0 \\ 0 & 1_{\Re_{\setminus\{0\}+}}([z]_{i,2}) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1_{\Re_{\setminus\{0\}+}}([z]_{i,m}) \end{bmatrix} \end{cases} \quad , \quad i \in \{1,2,\dots,n\}$$

with $1_{\Re_{\setminus\{0\}}+}$ the $\Re_{\setminus\{0\}}+$ indicator function.

4 – Dense layers

Definition – Let $y \in M_{1,m}$ the input vector, and $W \in M_{1,m}$ the weight matrix. Then the simple perceptron is defined as the operation

5 – Neural Network

Definition – Suppose a data set with n samples. Each sample have m features and a corresponding one-hot encoded label among l possible labels.

Let $X \in M_{n,m}$, and $Y \in M_{n,l}$ the matrices defining the features and the labels respectively for each sample.

Suppose a neural network with k layers.

6 – References