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. All the elements below are defined for classifications but can be re-used or easily re-defined to cover regressions.

2 – Notation and Nomenclature

Definition – Let Ω a non-empty open subset of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, ${}_{m'}\|\cdot\|$ a norm on $\Re^{m'}$, and $f:\Omega\to\Re^{m'}$. Then f continuous function is equivalent to

$$\forall a\in\Omega \quad ,$$

$$\forall \epsilon>0 \quad , \quad \exists \eta>0 \quad , \quad \forall x\in\Omega \quad , \quad {}_m\|x-a\|\leq\eta \Rightarrow {}_{m'}\|f(x)-f(a)\|\leq\epsilon$$

The notation $\zeta(\Omega,\Re^{m'})$ means the set of continuous functions from Ω to $\Re^{m'}$.

The notation $\zeta(\Re^m)$ means the set of continuous functions from $\Re^m$ to $\Re^m$.

Definition – Let Ω a non-empty open subset of $\Re^m$, ${}_m\|\cdot\|$ a norm on $\Re^m$, ${}_{m'}\|\cdot\|$ a norm on $\Re^{m'}$, and $f:\Omega\to\Re^{m'}$. Then f derivable is equivalent to

$$\forall a\in\Omega \quad , \quad \exists f'(a)\in\Re^{m'} \quad ,$$

$$\forall \epsilon>0 \quad , \quad \exists \eta>0 \quad , \quad \forall x\in\Omega \quad , \quad {}_m\|x-a\|\leq\eta \Rightarrow \|\frac{f(x)-f(a)}{x-a}-f'(a)\|_{m'}\leq\epsilon$$

The notation $D(\Omega,\Re^{m'})$ means the set of derivable functions from Ω to $\Re^{m'}$.

The notation $D(\Re^{m})$ means the set of derivable functions from $\Re^{m}$ to $\Re^{m}$.

Definition – Let Ω a non-empty open subset of $\Re^{m}$, ${}_{m}\|\cdot\|$ a norm on $\Re^{m}$, ${}_{m'}\|\cdot\|$ a norm on $\Re^{m'}$, and $f:\Omega\to\Re^{m'}$. Then f piece-wise derivable is equivalent to

$$\forall K\subset\Omega \text{ such as } K \text{ compact and bounded,}$$

$$\exists(K_i)_{i\in[\![0,n]\!]} \text{ non-empty open subsets such as } \bigcup_{i=0}^{n}\overline{K_i}=K \text{ and } \forall(i,i')\in[\![0,n]\!]^2,$$
$$i\neq i'\Rightarrow K_i\cap K_{i'}=\emptyset,$$

$$\forall i\in[\![0,n]\!],\ \exists f_i\in D^0(\overline{K_i},\Re^{m'}),\ \forall x\in K_i,\ f_i(x)=f(x)$$

The notation $D_{pw}(\Omega,\Re^{m'})$ means the set of piece-wise derivable functions from Ω to $\Re^{m'}$.

The notation $D_{pw}(\Re^{m})$ means the set of piece-wise derivable functions from $\Re^{m}$ to $\Re^{m}$.

Theorem – $D(\Re^{m})\subset D_{pw}(\Re^{m})\subset\zeta(\Re^{m})$

Proof: TO DO.

Definition – Let $a_{i,j}\in\Re$ for $i\in[\![1,n]\!]$ and $j\in[\![1,m]\!]$. 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

$$\forall i\in[\![1,n]\!],\ \forall j\in[\![1,m]\!],\ A_{i,j}=a_{i,j}$$

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

$$\forall i\in[\![1,n]\!],\ A_{i,:}=\begin{bmatrix}a_{i,1} & a_{i,2} & \cdots & a_{i,n}\end{bmatrix}$$

The notation $M_{n,m}$ means the matrix set of dimension $n \times m$ with coefficients in $\Re$.

The notation $M_{n,m}(E)$ means the matrix set of dimension $n \times m$ with coefficients in $E \subseteq \Re$.

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

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

$$\forall i \in [\![1,n]\!] \ , \ \forall j \in [\![1,p]\!] \ , \ C_{i,j} = \sum_{k=1}^{m} A_{i,k} \times B_{k,j}$$

Definition – Let $a \in \Re$ and $B \in M_{n,m}$. Let the scalar wise product noted as $a \times B$ be

$$C = a \times B = B \times a$$

where C is in $M_{n,m}$ with

$$\forall i \in [\![1,n]\!] \ , \ \forall j \in [\![1,m]\!] \ , \ C_{i,j} = a \times B_{i,j}$$

Theorem – Let U and V non-empty open subsets of $\Re^n$ and $\Re^m$. Let $f : \begin{cases} U \to V \\ x \mapsto f(x) \end{cases}$ and $g : \begin{cases} V \to \Re^p \\ y \mapsto g(y) \end{cases}$ such as $f \in D(U,V)$ and $g \in D(V, \Re^p)$. Then $g \circ f : \begin{cases} U \to \Re^p \\ x \mapsto g(f(x)) \end{cases}$ is in $D(U, \Re^p)$ and its Jacobian is

$$\frac{d(g \circ f)}{dx} : \begin{cases} U \to M_{p,n} \\ x \mapsto \frac{dg}{dy}(f(x)) \times \frac{df}{dx}(x) \end{cases}$$

Proof: TO DO.

3 – Activation functions

Definition – Let $F_{act} \in D(\Re^m)$. Then the vector wise application

$$F_{act} : \begin{cases} \Re^m \to \Re^m \\ z \mapsto f(z) \end{cases}$$

is an activation function.

Definition – $ReLU$ is the following vector wise application

$$ReLU : \begin{cases} \Re^m \to \Re^m \\ z \mapsto max(0_{\Re^m}, z) \end{cases}$$

with max the element-wise maximum operation between two vectors.

Hypothesis – The notation $ReLU_j$ means the application corresponding to the coefficient j of the function $ReLU$. Let $z \in \Re^m$ then

$$\forall j \in [\![1, m]\!] \ , \ ReLU_j(z_j) = max(0, z_j) = ReLU(z)_j$$

$ReLU$ is supposed derivable on every coefficients at 0

$$\forall j \in [\![1, m]\!] \ , \ ReLU_j'(0) = 0$$

Proposition – $ReLU$ is an activation function. Its Jacobian matrix is

$$\frac{d\,ReLU}{dz} : \left\{ z \mapsto \begin{bmatrix} 1_{\Re^+_{\setminus\{0\}}}(z_1) & 0 & \cdots & 0 \\ 0 & 1_{\Re^+_{\setminus\{0\}}}(z_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1_{\Re^+_{\setminus\{0\}}}(z_m) \end{bmatrix} \right. \quad \Re^m \to M_{m,m}$$

with $1_{\Re^+_{\setminus\{0\}}}$ the $\Re^+_{\setminus\{0\}}$ indicator function on $\Re$.

Proof: TO DO.

Proposition – The following vector wise application is an activation function

$$SoftMax:\begin{cases}\Re^m \to]0,1[^m \\ z \mapsto e^{z_j} \Big/ \sum_{j'=1}^{m} e^{z_{j'}}\end{cases}$$

with e the element-wise exponential operation.

The $SoftMax$ function will be denoted as S for simplicity.

Its Jacobian matrix is

$$\frac{dS}{dz}:\begin{cases}\Re^m \to M_{m,m} \\ z \mapsto \frac{dS}{dz}(z)\end{cases}$$

where $\forall z \in \Re^m$, $\forall (j,j') \in \{1,2,\dots,m\}^2$,

$$\frac{dS}{dz}(z)_{j,j'} = S(z)_j \times (\delta_{j,j'} - S(z)_{j'})$$

with $\delta_{j,j'}$ the Kronecker delta.

Proof: TO DO.

3 – Loss

Definition – Let $\hat{\Omega} \in M_{n,m}$ and $\Omega \subseteq M_{n,m}$ non empty subsets. Let $\hat{y} \in \hat{\Omega}$ and $F_{loss} \in D(\Omega, \Re)$. Then F_{loss} is a loss function is equivalent to the application

$$F_{loss} \circ g:\begin{cases}E \to \Re \\ \epsilon \mapsto (F_{loss} \circ g)(\epsilon) = F_{loss}(\hat{y} + \epsilon)\end{cases}$$

is an increasing function according each coefficient with $E \subseteq M_{n,m}$ such as $F_{loss} \circ g$ is always defined.

The $\hat{y}$ matrix is named the ground truth.

Proposition – Let $\hat{y} \in \{0,1\}^m$ a ground truth matrix. Then the application

$$\xi:\begin{cases}]0,1[^m \to \Re \\ y \mapsto -\sum_{j=1}^{m} \hat{y}_j \log(y_j) \end{cases}$$

is a loss function. The application is named Categorical cross-entropy loss.

Its Gradient matrix is

$$\frac{d\xi}{dy}:\begin{cases}]0,1[^m \to \Re^m \\ y \mapsto -\begin{bmatrix} \frac{\hat{y}_1}{y_1} & \dots & \frac{\hat{y}_m}{y_m} \end{bmatrix} \end{cases}$$

Proof: TO DO.

Proposition – Let $\hat{y} \in \{0,1\}^m$ a ground truth matrix. Let $S: \Re^m \to]0,1[^m$ and $\xi:]0,1[^m \to \Re$ the *SoftMax* activation and Categorical cross-entropy loss functions. Then $\xi \circ S: \Re^m \to \Re$ is derivable on $\Re^m$ and its Gradient matrix is

$$\frac{d(\xi \circ S)}{dz}:\begin{cases} \Re^m \to \Re^m \\ z \mapsto S(z) - \hat{y} \end{cases}$$

Proof: TO DO.

4 – Dense layers

Definition – The application

$$L:\begin{cases} \Re^m \times M_{m',m} \times \Re^{m'} \to \Re^{m'} \\ (y, W, b) \mapsto y \times W^T + b \end{cases}$$

defines a dense layer with y named the input vector, W named the weight matrix and b named the bias matrix.

The notation L_j means the application corresponding to the coefficient j of the dense layer L. Let $y \in \Re^m$ an input vector, $W \in M_{m',m}$ a weight matrix and $b \in \Re^{m'}$ a bias matrix then

$$\forall\, j\in[\![1,m']\!]\ ,\ \ L_j(W_{j,:})=y\times(W_{j,:})^T+b_j=L(y,W,b)_j$$

Proposition – Let $L:\Re^m\times M_{m',m}\times\Re^{m'}\to\Re^{m'}$ the dense layer function. Then L is derivable according the first and third variables on $\Re^m$ and $\Re^{m'}$ respectively.

Let $y\in\Re^m$ an input vector and $b\in\Re^{m'}$ a bias matrix. Then $L_{j'}:\Re^m\to\Re$ is also derivable for all coefficient $j'\in[\![1,m']\!]$.

Its Jacobian matrices are

$$\frac{\partial L}{\partial y}:\begin{cases}\Re^m\to M_{m',m}\\ y\mapsto W\end{cases}$$

$$\forall\, j'\in[\![1,m']\!]\ ,\ \ \frac{dL_{j'}}{dw}:\begin{cases}\Re^m\to\Re^m\\ w\mapsto y\end{cases}$$

$$\frac{\partial L}{\partial b}:\begin{cases}\Re^{m'}\to M_{m',m'}\\ b\mapsto I_{m'}\end{cases}$$

with $I_{m'}$ the identity matrix of size $m'\times m'$.

Proof: TO DO.

Proposition – Let $L:\Re^m\times M_{m',m}\times\Re^{m'}\to\Re^{m'}$ and $ReLU:\Re^{m'}\to\Re^{m'}$ the dense layer and $ReLU$ activation functions. Let $F^{upstream}:\begin{cases}\Re^{m'}\to\Re\\ y'\mapsto F^{upstream}(y')\end{cases}$ such as $F^{upstream}\in D(\Re^{m'},\Re)$. Then $F_{upstream}\circ ReLU\circ L:\Re^m\times M_{m',m}\times\Re^{m'}\to\Re$ is derivable according the first and third variables on $\Re^m$ and $\Re^{m'}$ respectively.

The notation $F_{j'}^{upstream}$ means the application corresponding to the coefficient j' of $F^{upstream}$. Let $y'\in\Re^{m'}$ then

$$\forall\, j'\in[\![1,m']\!]\ ,\ \ F_{j'}^{upstream}(y'_j)=F^{upstream}(y')_{j'}$$

Let $y\in\Re^m$ an input vector and $b\in\Re^{m'}$ a bias matrix. Then $F_{j'}^{upstream}\circ ReLU_{j'}\circ L_{j'}:\Re\to\Re$ is

also derivable for all coefficient $j' \in [\![1, m']\!]$.

Its Gradient matrices are

$$\frac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial y} : \begin{cases} \Re^m \to \Re^m \\ y \mapsto \dfrac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial y}(y) \end{cases}$$

where $\forall y \in \Re^m$, $\forall j \in [\![1, m]\!]$,

$$\frac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial y}(y)_j = \sum_{j'=1}^{m'} \frac{dF^{upstream}}{dy'}(ReLU(L(y, W, b)))_{j'} \times 1_{\Re^+_{\setminus\{0\}}}(L(y, W, b)_{j'}) \times W_{j',j}$$

with $W \in M_{m',m}$ a weight matrix, $b \in \Re^{m'}$ a bias matrix and $1_{\Re^+_{\setminus\{0\}}}$ the $\Re^+_{\setminus\{0\}}$ indicator function on $\Re$.

$$\forall j' \in [\![1, m']\!] \ , \ \frac{d(F^{upstream}_{j'} \circ ReLU_{j'} \circ L_{j'})}{dw} : \begin{cases} \Re^m \to \Re^m \\ w \mapsto \dfrac{d(F^{upstream}_{j'} \circ ReLU_{j'} \circ L_{j'})}{dw}(w) \end{cases}$$

where $\forall w \in \Re^m$, $\forall j \in [\![1, m]\!]$,

$$\frac{d(F^{upstream}_{j'} \circ ReLU_{j'} \circ L_{j'})}{dw}(w)_j = F^{upstream}_{j'}{}'(ReLU_{j'}(L_{j'}(w))) \times 1_{\Re^+_{\setminus\{0\}}}(L_{j'}(w)) \times y_j$$

with $y \in \Re^m$ an input matrix, $b \in \Re^{m'}$ a bias matrix and $1_{\Re^+_{\setminus\{0\}}}$ the $\Re^+_{\setminus\{0\}}$ indicator function on $\Re$.

$$\frac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial b} : \begin{cases} \Re^{m'} \to \Re^{m'} \\ b \mapsto \dfrac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial b}(b) \end{cases}$$

where $\forall b \in \Re^{m'}$, $\forall j' \in [\![1, m']\!]$,

$$\frac{\partial(F^{upstream} \circ ReLU \circ L)}{\partial b}(b)_{j'} = \frac{dF^{upstream}}{dy'}(ReLU(L(y, W, b)))_{j'} \times 1_{\Re^+_{\setminus\{0\}}}(L(y, W, b)_{j'})$$

with $y \in \Re^m$ a weight matrix and $W \in M_{m',m}$ a weight matrix.

Proof: TO DO.

5 – Neural Network

Definition – A training data set is defined as couples of vectors $(X_i, Y_i) \in \Re^m \times \Re^l$ for $i \in [\![1, n]\!]$. The X_i are named input or feature matrices and the Y_i target or label matrices.

Definition – Let p dense layers with activation functions $F_{act}^k \circ L^k(\cdot, W_k, b_k) : \Re^{m_k} \to \Re^{m_{k+1}}$ for $k \in [\![1, p]\!]$ with $W_k \in M_{m_{k+1}, m_k}$ and $b_k \in \Re^{m_{k+1}}$ the L_k weight and bias matrices. Let a training data set $(X_i, Y_i) \in \Re^{m_1} \times \Re^{m_p}$ for $i \in [\![1, n]\!]$. Let $F_{loss} : \Re^{m_p} \to \Re$ a loss function with $(Y_i)_{i \in [\![1, n]\!]}$ as ground truth matrices.

Then a neural network is defined as the application $N : \Re^{m_1} \to \Re^{m_p}$ such as $N = F_{act}^1 \circ L^1(\cdot, W_1, b_1) \circ \ldots \circ F_{act}^p \circ L^p(\cdot, W_p, b_p)$. The optimization problem is $\min_{(W_k, b_k)_{k \in [\![1, p]\!]}} \sum_{i=1}^{n} F_{loss}(N(X_i))$.

Proposition – Let a neural network

6 – Gradient descent

7 – References