Théorie mathématique pour les forwardpropagation et backpropagation

Alexandre AZOR

18 juillet 2022

1 Notation

1.1 Entiers

- p : Nombre de couches (Autres que la couche d'entrée)

- k, l : Indices de couche, $k \in [\![0, p]\!]$, $l \in [\![0, p]\!]$
  Couche d'entrée : $k = 0$
  Couche de sortie : $k = p$

- s_k : Nombre de neurones de la couche k

1.2 Réel

$\alpha \in \mathbb{R}^{+*}$: Learning rate

1.3 Matrices

- A^T : La transposée de la matrice A

- $0_{n,m}$: La matrice $\mathbb{R}^{n \times m}$ (n lignes et m colonnes) ne contenant que des 0

- $E_{i,j}^{n\times m}$: La matrice de $\mathbb{R}^{n\times m}$ dont le coefficient ligne i colonne j et un 1, les autres étant des 0

- $w^{(k)} \in \mathbb{R}^{s_{k+1}\times s_k}, k \in [\![0, p-1]\!]$: Poids pour passer de la couche k à la couche $k+1$
  $w_{i,j}^{(k)}$: Coefficient ligne i colonne j

  On note AB le produit matriciel usuel entre 2 matrices ou vecteurs A et B de tailles compatibles

- $[a_{i,j}]_{\substack{i\in[\![1,n]\!] \\ j\in[\![1,m]\!]}}$ désigne la matrice de n lignes et m colonnes dont le coefficient ligne i colonne j est $a_{i,j}$

1.4 Vecteurs

- 0_n : Le vecteur de $\mathbb{R}^n$ n'ayant que des 0

- E_i^n : Le vecteur de $\mathbb{R}^n$ ayant un 1 sur la i-ième ligne et des 0 partout ailleurs

- $x \in \mathbb{R}^{s_0}$: Valeur d'entrée

- $y \in \mathbb{R}^{s_p}$: Valeur attendue

- $y_{\text{pred}} \in \mathbb{R}^{s_p}$: Valeur de sortie

- $b^{(k)} \in \mathbb{R}^{s_{k+1}}, k \in [\![0, p-1]\!]$: Biais pour passer de la couche k à la couche $k+1$
  $b_i^{(k)}$: Coefficient ligne i

- $z^{(k)}, k \in [\![1, p]\!]$: Valeurs des différents neurones de la couche k avant activation
  $z_i^{(k)}$: Coefficient ligne i, Valeur du i-ième neurone de la couche k avant activation
  $z^{(0)}$ n'est pas défini
  $z^{(p)} = y_{\text{pred}}$

— $a^{(k)}, k \in [\![0, p-1]\!]$: Valeurs des différents neurones de la couche k après activation
$a_i^{(k)}$: Coefficient ligne i, Valeur du i-ième neurone de la couche k après activation
$a^{(0)} = x$
$a^{(p)}$ n'est pas défini

1.5 Fonctions

— *diag* : Fonction qui à un vecteur v de $\mathbb{R}^n$ associe la matrice de $\mathbb{R}^{n \times n}$ dont le coefficient ligne i colonne i est v_i pour $i \in [\![1, n]\!]$ et tous les autres sont nuls

— $g : \mathbb{R} \to \mathbb{R}$: Fonction d'activation

On désigne également par g l'application qui à un vecteur u de $\mathbb{R}^n$ associe $[g(u_i)]_{i \in [\![1,n]\!]}$
De même, $g'(u) = [g'(u_i)]_{i \in [\![1,n]\!]}$

—
$$h_k : \left\{ \begin{array}{ccc} \mathbb{R}^{s_k} \times \mathbb{R}^{s_{k+1} \times s_k} \times \mathbb{R}^{s_k} & \longrightarrow & \mathbb{R}^{s_{k+1}} \\ (b^{(k)}, w^{(k)}, a^{(k)}) & \longmapsto & b^{(k)} + w^{(k)} a^{(k)} \end{array} \right.$$
$(k \in [\![0, p-1]\!])$
Fonction permettant de passer de $a^{(k)}$ à $z^{(k+1)}$

— $L : \mathbb{R}^{s_p} \to (\mathbb{R}^{s_p} \to \mathbb{R}^+)$: Fonction coût ayant pour paramètres la valeur attendue et la valeur prédite
$L(y)$ est la fonction mesurant l'écart avec la valeur y

— $\tilde{L}_k (k \in [\![0, p]\!])$: Fonction coût ayant pour paramètres :
◇ y : la valeur attendue

◇ $\left(b_i^{(l)}\right)_{\substack{l \in [\![k, p-1]\!] \\ i \in [\![l, s_{k+1}]\!]}}$

◇ $\left(w_{i,j}^{(l)}\right)_{\substack{l\in[\![k,p-1]\!]\\ i\in[\![1,s_{k+1}]\!]\\ j\in[\![1,s_k]\!]}}$

◇ $z^{(k)}$ si $k>0$, x sinon

Pour alléger les notations, ces paramètres seront simplement désignés par (...)

$$\tilde{L}_p(y, y_{\text{pred}}) = L\,(y)(y_{\text{pred}})$$

$$\tilde{L}_{p-1}(...) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g\left(z^{(p-1)}\right)\right)\right)$$

$$\tilde{L}_k(...) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g\circ h_{p-2}\left(b^{(p-2)}, w^{(p-2)}, g\circ h_{p-3}\left(...b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\right)\right)\right)$$

$$\tilde{L}_0(...) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g\circ h_{p-2}\left(b^{(p-2)}, w^{(p-2)}, g\circ h_{p-3}\left(...b^{(0)}, w^{(0)}, x\right)\right)\right)\right)$$

2 Forwardpropagation

On note $x\in\mathbb{R}^{s_0}$ le vecteur donné en entrée du réseau

$$a^{(0)} \leftarrow x$$
$$z^{(1)} \leftarrow w^{(0)}a^{(0)} + b^{(0)}$$

Pour k allant de 1 à $p-1$ inclus :

$$a^{(k)} \leftarrow g\left(z^{(k)}\right)$$

$$z^{(k+1)} \leftarrow w^{(k)}a^{(k)} + b^{(k)}$$

On renvoie alors $z^{(p)}$

3 Backpropagation

3.1 Objectif et principe

On veut minimiser la valeur moyenne de la fonction coût $\tilde{L}_0$ appliquée avec les couples (x, y) de la base d'apprentissage.

On utilise une méthode de descente de gradient : Pour chaque couple (x, y) de la base d'apprentissage :
— On calcule y_{pred} en enregistrant les valeurs des $a^{(k)}$
— $\forall k \in [\![0, p-1]\!]$

$$\forall (i,j) \in [\![1, s_{k+1}]\!] \times [\![1, s_k]\!] \quad w_{i,j}^{(k)} \leftarrow w_{i,j}^{(k)} - \alpha \frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}}$$

$$\forall i \in [\![1, s_{k+1}]\!] \quad b_i^{(k)} \leftarrow b_i^{(k)} - \alpha \frac{\partial \tilde{L}_0}{\partial b_i^{(k)}}$$

En posant

$$dW^{(k)} \triangleq \left[\frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}} \right]_{(i,j) \in [\![1, s_{k+1}]\!] \times [\![1, s_k]\!]}$$

$$dB^{(k)} \triangleq \left[\frac{\partial \tilde{L}_0}{\partial b_i^{(k)}} \right]_{i \in [\![1, s_k]\!]}$$

les formules précédents se réécrivent :

$$w^{(k)} \leftarrow w^{(k)} - \alpha dW$$
$$b^{(k)} \leftarrow b^{(k)} - \alpha dB$$

3.2 $dL^{(k)}$

On pose $dL^{(k)}$, le vecteur de $\mathbb{R}^{s_k}$ tel que :

$$D\tilde{L}_k(...)(0, ..., 0, u) = dL^{(k)\ T} u \Leftrightarrow \frac{\partial \tilde{L}_k}{\partial u} = dL^{(k)\ T} u$$

On construit la suite $\left(dL^{(k)}\left(z^{(k)}\right)\right)_{k \in [\![1, p]\!]}$ par ordre décroissant des indices.

— $k = p$
$dL^{(p)}$ est l'unique vecteur vérifiant pour tout $u \in \mathbb{R}^{s_p}$:

$$D\tilde{L}_p\left(y, z^{(p)}\right)(0, u) = dL^{(p)\ T} u$$

$$\tilde{L}_p(y, u) = L(y)(u)$$

$$\Rightarrow D\tilde{L}_p\left(y, z^{(p)}\right)(0, u) = D\left(L(y)\right)\left(z^{(p)}\right)(u) = \vec{\boldsymbol{\nabla}}\left(L(y)\right)\left(z^{(p)}\right)^T u$$

$$\Rightarrow \boxed{dL^{(p)} = \vec{\boldsymbol{\nabla}} L(y)\left(z^{(p)}\right)}$$

— $0 < k < p$
On suppose $dL^{(k+1)}$ connu avec $z^{(k+1)} = h_k\left(b^{(k)}, w^{(k)}, g(z^{(k)})\right)$.

$$\tilde{L}_k\left(..., b^{(k)}, w^{(k)}, z^{(k)}\right) = \tilde{L}_{k+1}\left(..., h\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\right)$$

Avec la règle de la chaîne, on obtient pour $u \in \mathbb{R}^{s_k}$:

$$D\tilde{L}_k\left(..., b^{(k)}, w^{(k)}, z^{(k)}\right)(..., 0_{s_k}, 0_{s_{k+1}\times s_k}, u)$$

$$= D\tilde{L}_{k+1}(...)\left(..., Dh\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\left(0_{s_k}, 0_{s_{k+1}\times s_k}, Dg\left(z^{(k)}(u)\right)\right)\right)$$

$$Dg(v)(u) = diag\left(g'(v)\right) u$$

$$Dh\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)(0_{s_k}, 0_{s_{k+1}\times s_k}, q) = w^{(k)} q$$

$$D\tilde{L}_{k+1}\left(..., h\left(b^{(k)}, w^{(k)}, z^{(k)}\right)\right)(..., r) = dL^{(k+1)\ T} r$$

$$\Rightarrow \frac{\partial \tilde{L}_k}{\partial u}(...) = dL^{(k+1)\ T} w^{(k)} diag\left(g'(u)\right) u$$

$$\Rightarrow \boxed{dL^{(k)}(z^{(k)}) = diag\left(g'(z^{(k)})\right) w^{(k)\ T} dL^{(k+1)}}$$

3.3 $dB^{(k)}$

Pour k allant de $p-1$ à 0

$$\tilde{L}_0\left(..., b^{(k)}, w^{(k)}, ...\right) = \tilde{L}_k\left(..., b^{(k)}, w^{(k)}, z^{(k)}\right)$$

$$\text{Avec } z^{(k)} = h_{k-1}\left(b^{(k-1)}, w^{(k-1)}, g \circ h_{k-2}\left(...\right)\right)$$

$$\tilde{L}_k(..., b^{(k)}, w^{(k)}, z^{(k)}) = \tilde{L}_{k+1}\left(..., h\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\right)$$

$$\Rightarrow \frac{\partial \tilde{L}_0}{\partial b_i^{(k)}} = \frac{\partial \tilde{L}_k}{\partial b_i^{(k)}} =$$

$$= D\tilde{L}_{k+1}\left(...\right)\left(..., Dh\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\left(E_i^{s_{k+1}}, 0_{s_{k+1}\times s_k}, 0_{s_k}\right)\right)$$

$$= dL^{(k+1)\ T} E_i^{s_{k+1}}$$

$$= \left(dL^{(k+1)}\right)_i$$

$$\Rightarrow \boxed{dB^{(k)} = dL^{(k+1)}}$$

3.4 $dW^{(k)}$

— Pour k allant de $p-1$ à 1

$$\tilde{L}_0\left(..., b^{(k)}, w^{(k)}, ...\right) = \tilde{L}_k\left(..., b^{(k)}, w^{(k)}, z^{(k)}\right)$$

$$\text{Avec } z^{(k)} = h_{k-1}\left(b^{(k-1)}, w^{(k-1)}, g \circ h_{k-2}\left(...\right)\right)$$

$$\tilde{L}_k(..., b^{(k)}, w^{(k)}, z^{(k)}) = \tilde{L}_{k+1}\left(..., h\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\right)$$

$$\Rightarrow \frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}} = \frac{\partial \tilde{L}_k}{\partial w_{i,j}^{(k)}} =$$

$$= D\tilde{L}_{k+1}\left(...\right)\left(..., Dh\left(b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\left(0_{s_{k+1}}, E_{i,j}^{s_{k+1}\times s_k}, 0_{s_k}\right)\right)$$

$$= dL^{(k+1)\ T} g\left(z_j^{(k)}\right) E_i^{s_{k+1}}$$

$$= g\left(z_j^{(k)}\right)\left(dL^{(k+1)}\right)_i$$

$$= \left[dL^{(k+1)} g\left(z^{(k)}\right)^T\right]_{i,j}$$

$$\Rightarrow \boxed{dW^{(k)} = dL^{(k+1)} g\left(z^{(k)}\right)^T}$$

— Pour $k = 0$

$$\Rightarrow \boxed{dW^{(0)} = dL^{(k+1)} x^T}$$