TD 7 - Réseaux de Neurones

B Deporte

May 15, 2024

1 Exercice 1 : un réseau dans le détail

Ici, on va calculer à la main la forward propagation et la back propagation sur un réseau simple. L'objectif du réseau est d'approximer une fonction $f : \mathbb{R}^3 \to \mathbb{R}$. Il s'agit d'une tâche de régression. Le réseau est un Multi Layer Perceptron, avec une couche cachée de deux neurones.

1. Dessiner le réseau en utilisant les notations du cours pour les poids, les calculs intermédiaires s et les sorties o des neurones.
2. Calculer les équations de forward propagation.
3. Regrouper les équations de forward propagation sous forme matricielle.
4. On choisit la norme 2 comme loss. Calculer les équations de back propagagtion.
5. Regrouper les équations de back propagation sous forme matricielle et vérifier la première formule (8) du 2.6.3 du cours.
6. On choisit la fonction *softplus* $\varphi(x) = \ln(1 + e^x)$ comme fonction d'activation. Etudier et tracer la fonction.
7. Dans les faits et pour réduire les temps de calcul, on utilise souvent la fonction *ReLU* (Rectified Linear Unit) $\varphi(x) = \max(0, x)$ au lieu de la softplus. Pourquoi ? Quel serait le problème potentiel de cette fonction d'activation par rapport aux hypothèses générales de la backpropagation ? Pourquoi est-ce que ce n'est finalement pas un problème ?

1/ Architecture du réseau

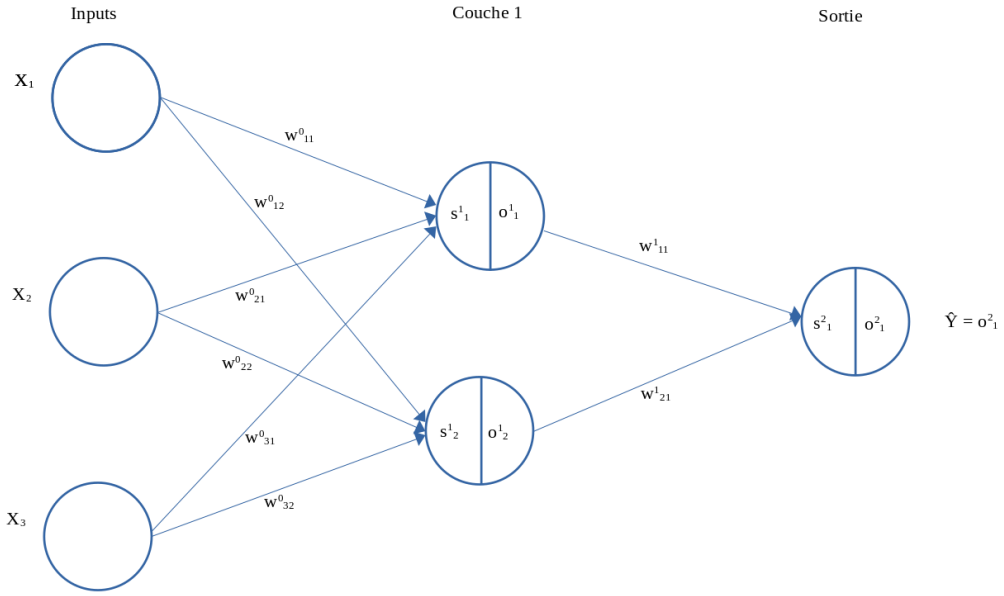

Figure 1: Architecture Reseau

2/ Equations de forward propagation
Rappel des notations:

- s_k^l : valeur du neurone k de la couche l, avant l'activation
- o_k^l : valeur du neurone k de la couche l, après l'activation
- w_{pk}^l : poids depuis le neurone p de la couche l vers le neurone k de la couche suivante
- b_k^l : biais du neurone k de la couche l

Sans oublier les biais, on a pour la couche 2 de sortie:

$$s_1^2 = w_{11}^1 o_1^1 + w_{21}^1 o_2^1 + b_1^2 \tag{1}$$
$$o_1^2 = \varphi(s_1^2) \tag{2}$$
$$\hat{y} = o_1^2 \tag{3}$$

Et pour la couche 1:

$$s_1^1 = w_{11}^0 x_1 + w_{21}^0 x_2 + w_{31}^0 x_3 + b_1^1 \tag{4}$$
$$s_2^1 = w_{12}^0 x_1 + w_{22}^0 x_2 + w_{32}^0 x_3 + b_2^1 \tag{5}$$
$$o_1^1 = \varphi(s_1^1) \tag{6}$$
$$o_2^1 = \varphi(s_2^1) \tag{7}$$
$$\tag{8}$$

3/ Forward propagation sous forme matricielle
Couche 2:

$$s_1^2 = \left(w_{11}^1, w_{21}^1, b_1^2\right) \begin{pmatrix} o_1^1 \\ o_2^1 \\ 1 \end{pmatrix} \tag{9}$$

Couche 1:

$$\begin{pmatrix} s_1^1 \\ s_2^2 \end{pmatrix} = \begin{pmatrix} w_{11}^0 & w_{21}^0 & w_{31}^0 & b_1^1 \\ w_{12}^0 & w_{22}^0 & w_{32}^0 & b_2^1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} \tag{10}$$

$$\begin{pmatrix} o_1^1 \\ o_2^2 \end{pmatrix} = \varphi \begin{pmatrix} s_1^1 \\ s_2^2 \end{pmatrix} \tag{11}$$

4/ Equations de backpropagation

On note $\hat{y}$ la prédiction du réseau, y la valeur ground truth.
La loss L2 s'écrit:

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

Pour calculer la backpropagation, on commence par la couche de sortie, et on remonte:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = \frac{\partial \mathcal{L}}{\partial o_1^2} \tag{12}$$

$$= \hat{y} - y \tag{13}$$

Pour la couche 2:

$$\frac{\partial \mathcal{L}}{\partial w_{11}^1} = \frac{\partial \mathcal{L}}{\partial o_1^2} \frac{\partial o_1^2}{\partial s_1^2} \frac{\partial s_1^2}{\partial w_{11}^1} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \varphi'(s_1^2) o_1^1 \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial w_{21}^1} = \frac{\partial \mathcal{L}}{\partial o_1^2} \frac{\partial o_1^2}{\partial s_1^2} \frac{\partial s_1^2}{\partial w_{21}^1} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \varphi'(s_1^2) o_2^1 \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial b_1^2} = \frac{\partial \mathcal{L}}{\partial o_1^2} \frac{\partial o_1^2}{\partial s_1^2} \frac{\partial s_1^2}{\partial b_1^2} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \varphi'(s_1^2) \tag{16}$$

Pour la couche 1:

$$\frac{\partial \mathcal{L}}{\partial w_{11}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \frac{\partial o_1^1}{\partial s_1^1} \frac{\partial s_1^1}{\partial w_{11}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \varphi'(s_1^1) x_1 \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial w_{21}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \frac{\partial o_1^1}{\partial s_1^1} \frac{\partial s_1^1}{\partial w_{21}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \varphi'(s_1^1) x_2 \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial w_{31}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \frac{\partial o_1^1}{\partial s_1^1} \frac{\partial s_1^1}{\partial w_{31}^0} = \frac{\partial \mathcal{L}}{\partial o_1^1} \varphi'(s_1^1) x_3 \tag{19}$$

$$\frac{\partial \mathcal{L}}{\partial b_1^1} = \frac{\partial \mathcal{L}}{\partial o_1^1} \frac{\partial o_1^1}{\partial s_1^1} \frac{\partial s_1^1}{\partial b_1^1} = \frac{\partial \mathcal{L}}{\partial o_1^1} \varphi'(s_1^1) \tag{20}$$

$$\frac{\partial \mathcal{L}}{\partial w_{12}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \frac{\partial o_2^1}{\partial s_2^1} \frac{\partial s_2^1}{\partial w_{12}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \varphi'(s_2^1) x_1 \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial w_{22}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \frac{\partial o_2^1}{\partial s_2^1} \frac{\partial s_2^1}{\partial w_{22}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \varphi'(s_2^1) x_2 \tag{22}$$

$$\frac{\partial \mathcal{L}}{\partial w_{32}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \frac{\partial o_2^1}{\partial s_2^1} \frac{\partial s_2^1}{\partial w_{32}^0} = \frac{\partial \mathcal{L}}{\partial o_2^1} \varphi'(s_2^1) x_3 \tag{23}$$

$$\frac{\partial \mathcal{L}}{\partial b_2^1} = \frac{\partial \mathcal{L}}{\partial o_2^1} \frac{\partial o_1^1}{\partial s_2^1} \frac{\partial s_2^1}{\partial b_2^1} = \frac{\partial \mathcal{L}}{\partial o_2^1} \varphi'(s_2^1) \tag{24}$$

Avec;

$$\frac{\partial \mathcal{L}}{\partial o_1^1} = \frac{\partial \mathcal{L}}{\partial o_1^2}\frac{\partial o_1^2}{\partial s_1^2}\frac{\partial s_1^2}{\partial o_1^1} \tag{25}$$

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}}\varphi'(s_1^2)w_{11}^1 \tag{26}$$

$$= (\hat{y} - y)\varphi'(s_1^2)w_{11}^1 \tag{27}$$

$$\frac{\partial \mathcal{L}}{\partial o_2^1} = \frac{\partial \mathcal{L}}{\partial o_1^2}\frac{\partial o_1^2}{\partial s_1^2}\frac{\partial s_1^2}{\partial o_2^1} \tag{28}$$

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}}\varphi'(s_1^2)w_{21}^1 \tag{29}$$

$$= (\hat{y} - y)\varphi'(s_1^2)w_{21}^1 \tag{30}$$

Une fois ces calculs faits, on met à jour les poids de façon itérative:

$$w_{pk}^{l^{(it+1)}} = w_{pk}^{l^{(it)}} - \eta\frac{\partial \mathcal{L}}{\partial w_{pk}^l}$$

5/ Backpropagation matricielle

On a vu pour une couche de sortie Y, d'entrée X, de matrice de poids W:

$$\frac{\partial \mathcal{L}}{\partial W} = \left(\frac{\partial \mathcal{L}}{\partial Y} \odot \varphi'(WX)\right) X^T \tag{31}$$

On a ici pour la couche 2:

$$\frac{\partial \mathcal{L}}{\partial Y} = \frac{\partial \mathcal{L}}{\partial o_1^2} \tag{32}$$

$$\varphi'(WX) = \varphi'(s_1^2) \tag{33}$$

$$X^T = \begin{pmatrix} o_1^1 o_2^1 1 \end{pmatrix} \tag{34}$$

et on retrouve les équations ligne à ligne du 4/

Pour la couche 1:

$$\frac{\partial \mathcal{L}}{\partial Y} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial o_1^1} \\ \frac{\partial \mathcal{L}}{\partial o_2^1} \end{pmatrix} \tag{35}$$

$$W = \begin{pmatrix} w_{11}^0 & w_{21}^0 & w_{31}^0 & b_1^1 \\ w_{12}^0 & w_{22}^0 & w_{32}^0 & b_2^1 \end{pmatrix} \tag{36}$$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} \tag{37}$$

$$WX = \begin{pmatrix} s_1^1 \\ s_2^1 \end{pmatrix} \tag{38}$$

Et on vérifie :

$$\frac{\partial \mathcal{L}}{\partial W} = \left(\begin{pmatrix} \frac{\partial \mathcal{L}}{\partial o_1^1} \\ \frac{\partial \mathcal{L}}{\partial o_2^1} \end{pmatrix} \odot \begin{pmatrix} \varphi'(s_1^1) \\ \varphi'(s_2^1) \end{pmatrix}\right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}^T \tag{39}$$

6/ Etude softplus : RAS

7/ Softplus vs ReLU

- La fonction ReLU $\varphi(x) = \max(0, \mathrm{x})$ est beaucoup moins coûteuse à calculer que la fonction softplus $\varphi(x) = \ln(1 + \mathrm{e}^{\mathrm{x}})$, et a un comportement similaire (mêmes asymptotes).
- Toutefois, ReLU n'est pas différentiable en 0, ce qui empêche théoriquement la descente de gradient (nécessité de fonctions d'activation différentiables)
- Néanmoins, 0 n'est quasiment jamais atteint numériquement.

2 Exercice 2 : Vanishing Gradients - "and suddenly, he was gone"

Dans certains types de réseau, le gradient peut tendre rapidement vers 0, ce qui pose des problèmes d'entraînement du réseau. L'objectif de l'exercice est d'explorer un exemple.

On considère le réseau suivant :

Figure 2: Reseau toy pour vanishing gradient

1. Retour sur la chain rule. Soit f différentiable sur $I \in \mathbb{R}$, g différentiable sur J (avec $f(I) \subset J$). Soit $h = g \circ f$. Ecrire la dérivée de h. En posant $y = f(x)$ et $z = h(x) = g(y)$, écrire $\frac{\partial z}{\partial x}$ et retrouver la chain rule.
2. Ecrire la forward propagation du réseau. On note g la fonction d'activation (identique pour les trois neurones), où a sont les sorties des neurones, z les entrées.
3. Ecrire le gradient de la loss par rapport à $w^{(1)}$. Isoler les termes $\frac{\partial a}{\partial z}$
4. On suppose que les fonctions d'activation sont la sigmoïde. Quel est le comportement de $\sigma(z)$ loin de 0 ? Montrer le problème du vanishing gradient.

1. On a la formule classique $h' = g'(f) \times f'$. Soit $h'(x) = g'(f(x)) \times f'(x)$, ou:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

2. La forward propagation du réseau est:

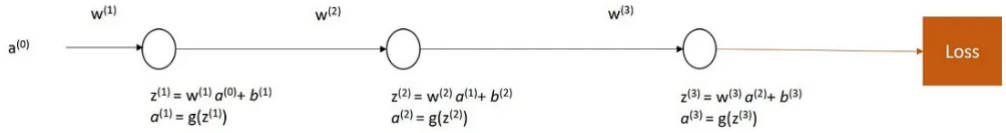

Figure 3: Exo 2 forward propagation

3. Le gradient par rapport au premier poids s'écrit:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w^{(1)}} &= \frac{\partial L}{\partial a^{(3)}}\frac{\partial a^{(3)}}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial a^{(2)}}\frac{\partial a^{(2)}}{\partial z^{(2)}}\frac{\partial z^{(2)}}{\partial a^{(1)}}\frac{\partial a^{(1)}}{\partial z^{(1)}}\frac{\partial z^{(1)}}{\partial w^{(1)}} \\
&= \frac{\partial L}{\partial a^{(3)}}\frac{\partial z^{(3)}}{\partial a^{(2)}}\frac{\partial z^{(2)}}{\partial a^{(1)}}\frac{\partial z^{(1)}}{\partial w^{(1)}}\prod_{i=1}^{3}\frac{\partial a^{(i)}}{\partial z^{(i)}} \\
&= \frac{\partial L}{\partial a^{(3)}}\frac{\partial z^{(3)}}{\partial a^{(2)}}\frac{\partial z^{(2)}}{\partial a^{(1)}}\frac{\partial z^{(1)}}{\partial w^{(1)}}\prod_{i=1}^{3} g'(z^{(i)})
\end{aligned}$$

4. Loin de 0, on a $g'(z) = \sigma'(z)$ 0. On voit alors que $\frac{\partial \mathcal{L}}{\partial w^{(1)}} \sim 0$. Le gradient tend vers 0, $w^{(1)}$ n'est plus mis à jour.

3 Exercice 3 : Entropie, Cross Entropie et KL

On rappelle les définitions, pour $\mathbb{P}$ et $\mathbb{Q}$ lois de probabilité de densités $p(x)$ et $q(x)$:

- **Entropie** : l'entropie de $\mathbb{P}$:

$$\mathcal{H}(\mathbb{P}) = -\int p(x)\ln \mathrm{p(x)dx}$$

- **Cross Entropie** : la cross-entropie de $\mathbb{Q}$ par rapport à $\mathbb{P}$ est:

$$\mathcal{H}(\mathbb{P},\mathbb{Q}) = -\int p(x)\ln \mathrm{q(x)dx}$$

- La **divergence de Kullback Leibler** de $\mathbb{P}$ par rapport à $\mathbb{Q}$ est:

$$\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) = +\int p(x)\ln\left(\frac{\mathrm{p(x)}}{\mathrm{q(x)}}\right)\mathrm{dx}$$

1. Quelles sont les conditions de support sur $p(x)$ et $q(x)$ pour que $\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) < +\infty$?
2. On considère P et Q discrètes, définies par:

$$P(0) = \frac{9}{25}$$
$$P(1) = \frac{12}{25}$$
$$P(2) = \frac{4}{25}$$
$$Q(0) = \frac{1}{3}$$
$$Q(1) = \frac{1}{3}$$
$$Q(2) = \frac{1}{3}$$

   Calculer $\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q})$ et $\mathbb{D}_{\mathbf{KL}}(\mathbb{Q},\mathbb{P})$
3. Montrer $\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) = \mathcal{H}(\mathbb{P},\mathbb{Q}) - \mathcal{H}(\mathbb{P})$
4. **Inégalité de Jensen** - Soit f convexe sur $I = [a,b] \subset [-\infty,+\infty]$ et X variable aléatoire à valeurs dans $]a,b[$. En considérant la position de f par rapport à la tangente à f en $\mathbb{E}(X)$, montrer que $f(\mathbb{E}(X) \leq \mathbb{E}(f(X))$
5. En utilisant Jensen et *log* concave, montrer que $\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) \geq 0$

1. On doit avoir mêmes supports $Supp(p) = Supp(q)$.
2. $$\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) = \frac{1}{25}\left(32\ln 2 + 55\ln 3 - 50\ln 5\right) \sim 0.0852$$
$$\mathbb{D}_{\mathbf{KL}}(\mathbb{Q},\mathbb{P}) = \frac{1}{3}\left(-4\ln 2 - 6\ln 3 + 6\ln 5\right) \sim 0.0974$$
3. $$\begin{aligned}\mathbb{D}_{\mathbf{KL}}(\mathbb{P},\mathbb{Q}) &= \int p(x)\ln\left(\frac{\mathrm{p(x)}}{\mathrm{q(x)}}\right)\mathrm{dx} \\ &= \int p(x)\ln \mathrm{p(x)dx} - \int \mathrm{p(x)}\ln \mathrm{q(x)dx} \\ &= \mathcal{H}(\mathbb{P},\mathbb{Q}) - \mathcal{H}(\mathbb{P})\end{aligned}$$

4. La tangente à f au point d'abscisse $\mathbb{E}(X)$ a pour équation $\alpha X + \beta$, avec $f(\mathbb{E}(X)) = \alpha\mathbb{E}(X) + \beta$. f étant convexe, on a pour toute valeur de X : $f(X) \geq \alpha X + \beta$. En prenant l'espérance:

$$\begin{aligned} f(X) &\geq \alpha X + \beta \\ \mathbb{E}(f(X)) &\geq \mathbb{E}(\alpha X + \beta) = \alpha\mathbb{E}(X) + \beta = f(\mathbb{E}(X)) \end{aligned}$$

5. On a, en prenant l'espérance par rapport à $\mathbb{P}$, et en utilisant $-\ln$ convexe:

$$\begin{aligned} \mathbb{D}_{\mathbf{KL}}(\mathbb{P}, \mathbb{Q}) &= \int p(x) \ln\left(\frac{\mathrm{p(x)}}{\mathrm{q(x)}}\right) \mathrm{dx} && (40)\\ &= \mathbb{E}\left(\ln\left(\frac{\mathrm{p}}{\mathrm{q}}\right)\right) && (41)\\ &= -\mathbb{E}\left(\ln\left(\frac{\mathrm{q}}{\mathrm{p}}\right)\right) && (42)\\ &= \mathbb{E}\left(-\ln\left(\frac{\mathrm{q}}{\mathrm{p}}\right)\right) && (43)\\ &\geq -\ln \mathbb{E}\left(\frac{\mathrm{q}}{\mathrm{p}}\right) && (44)\\ &= -\ln\left(\int \mathrm{p(x)}\frac{\mathrm{q(x)}}{\mathrm{p(x)}}\mathrm{dx}\right) && (45)\\ &= -\ln\left(\int \mathrm{q(x)dx}\right) && (46)\\ &= 0 && (47) \end{aligned}$$

On montre aussi que $\mathbb{D}_{\mathbf{KL}}(\mathbb{P}, \mathbb{Q}) = 0$ ssi $\mathbb{P} = \mathbb{Q}$