Backpropagation
Processus d'Apprentissage d'un Réseau de Neurones
rivo.link@gmail.com

# 1 Backpropagation - Théorie

Ceci est mon approche de la backpropagation, un processus d'apprentissage d'un réseau de neurones multicouches.

Soit un réseau de neurones à une couche cachée noté NN(i,j,k) où i est le nombre des entrées, j le nombre de neurones dans la couche cachée et k le nombre de neurones de sorties.

On notera $(x_i)_i$ les entrées présentées au neurone j de la couche cachée, $(w_{ij})_i$ les poids associés et $b_j$ le biais pour ce neurone. Le neurone j aura donc la pré-activation $z_j$ selon la formule:

$$z_j = \sum_i x_i w_{ij} + b_j \tag{1}$$

L'activation du neurone j de la couche cachée sera la sigmoïde de sa pré-activation, il aura donc comme sortie:

$$f(z_j) = \frac{1}{1 + e^{-z_j}} = f(\sum_i x_i w_{ij} + b_j) \tag{2}$$

Comme les sorties des j neurones de la couche cachée sont les entrée des k neurones de la couche de sortie, en notant $(x_j)_j$ ces entrées, $(w_{jk})_j$ les poids et $b_k$ le biais, on aura pour le neurone k de la couche de sortie, la pré-activation $z_k$:

$$\begin{aligned} x_j &= f(z_j) \\ z_k &= \sum_j x_j w_{jk} + b_k \end{aligned} \tag{3}$$

L'activation du neurone k des sorties sera la sigmoïde de sa pré-activation, il aura donc comme sortie:

$$f(z_k) = \frac{1}{1 + e^{-z_k}} = f(\sum_j x_j w_{jk} + b_k) \tag{4}$$

Finalement, en notant $\hat{y}_k$ la sortie du neurone k de la couche de sortie, $(\hat{y}_k)_k$ sera la sortie du réseau de neurones NN(i,j,k), on aura les k-équations:

$$\begin{aligned} \hat{y}_k &= f(z_k) \\ \hat{y}_k &= f(\sum_j f(\sum_i x_i w_{ij} + b_j) w_{jk} + b_k) \end{aligned} \tag{5}$$

# 2 Backpropagation - Apprentissage

On note $(y_k)_k$ la sortie attendu, associée aux entrées $(x_i)_i$ pour le réseau de neurones NN(i,j,k). On définit l'erreur de prédiction $E(\hat{y})$ par:

$$E(\hat{y}) = \sum_k \frac{1}{2}(y_k - \hat{y}_k)^2 \tag{6}$$

## 2.1 Apprentissage - Couche de sortie

La mise à jour du j-ème poids $w_{jk}$ du neurone k de la couche de sortie se fera selon la formule:

$$w_{jk} := w_{jk} - \alpha \frac{\partial E(\hat{y})}{\partial w_{jk}} \tag{7}$$

$$\frac{\partial E(\hat{y})}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \sum_k \frac{1}{2}(y_k - \hat{y}_k)^2 \tag{8}$$

Comme, seul $\hat{y}_k$ dépend de $w_{jk}$, et les elements de la somme dont les indices se diffèrent de k ont une dérivée partielle nulle alors:

$$\begin{aligned}
\frac{\partial E(\hat{y})}{\partial w_{jk}} &= -\sum_k (y_k - \hat{y}_k) \frac{\partial \hat{y}_k}{\partial w_{jk}} \\
&= -(y_k - \hat{y}_k) \frac{\partial \hat{y}_k}{\partial w_{jk}} \\
&= -(y_k - \hat{y}_k) \frac{\partial f(z_k)}{\partial w_{jk}} \\
&= -(y_k - \hat{y}_k) \frac{\partial f(z_k)}{\partial z_k} \frac{\partial z_k}{\partial w_{jk}} \\
\frac{\partial E(\hat{y})}{\partial w_{jk}} &= -(y_k - \hat{y}_k) f'(z_k) \frac{\partial}{\partial w_{jk}} (\sum_j x_j w_{jk} + b_k)
\end{aligned} \tag{9}$$

Finalement, comme "la fonction f est la fonction sigmoid", il en résulte l'équation de variation de l'erreur $E(\hat{y})$ par rapport au poid $w_{jk}$, et par rapport au biais $b_k$:

$$\frac{\partial E(\hat{y})}{\partial w_{jk}} = -(y_k - \hat{y}_k) f(z_k)(1 - f(z_k)) x_j$$

$$\frac{\partial E(\hat{y})}{\partial b_k} = -(y_k - \hat{y}_k) f(z_k)(1 - f(z_k))$$

En d'autres termes, comme $\hat{y}_k = f(z_k)$, on a les nouvelles équations:

$$\frac{\partial E(\hat{y})}{\partial w_{jk}} = -\hat{y_k}(1 - \hat{y_k})(y_k - \hat{y}_k) x_j$$

$$\frac{\partial E(\hat{y})}{\partial b_k} = -\hat{y_k}(1 - \hat{y_k})(y_k - \hat{y}_k)$$

## 2.2 Apprentissage - Couche cachée

La mise à jour du i-ème poids $w_{ij}$ du neurone j de la couche cachée se fera selon la formule:

$$w_{ij} := w_{ij} - \alpha \frac{\partial E(\hat{y})}{\partial w_{ij}} \tag{10}$$

$$\frac{\partial E(\hat{y})}{\partial w_{ik}} = \frac{\partial}{\partial w_{ij}} \sum_k \frac{1}{2}(y_k - \hat{y}_k)^2 \tag{11}$$

Il faut noter qu'ici, la somme ne disparaît pas puisqu'elle ne depend pas de $w_{ij}$. Et comme seul $\hat{y}_k$ dépend de $w_{ij}$, alors on a:

$$\begin{aligned}
\frac{\partial E(\hat{y})}{\partial w_{ij}} &= -\sum_k (y_k - \hat{y}_k)\frac{\partial \hat{y}_k}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)\frac{\partial f(z_k)}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)\frac{\partial f(z_k)}{\partial z_k}\frac{\partial z_k}{\partial x_j}\frac{\partial x_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)(\frac{\partial}{\partial x_j}\sum_j x_j w_{jk} + b_k)\frac{\partial x_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)(w_{jk})\frac{\partial x_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)(w_{jk})\frac{\partial f(z_j)}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)(w_{jk})f'(z_j)\frac{\partial z_j}{\partial w_{ij}} \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)(w_{jk})f'(z_j)(\frac{\partial}{\partial w_{ij}}\sum_i x_i w_{ij} + b_j) \\
&= -\sum_k (y_k - \hat{y}_k)f'(z_k)w_{jk}f'(z_j)x_i
\end{aligned} \tag{12}$$

Finalement, il en résulte l'équation de variation de l'erreur $E(\hat{y})$ par rapport au poid $w_{ij}$. Et on en déduit sa variation par rapport au biais $b_j$:

$$\frac{\partial E(\hat{y})}{\partial w_{ij}} = -\sum_k (y_k - \hat{y}_k)f(z_k)(1 - f(z_k))w_{jk}f(z_j)(1 - f(z_j))x_i$$

$$\frac{\partial E(\hat{y})}{\partial b_j} = -\sum_k (y_k - \hat{y}_k)f(z_k)(1 - f(z_k))w_{jk}f(z_j)(1 - f(z_j))$$