Neural network from scratch

Maxime Daigle

2018-11-09

1. Relations et dérivées de quelques fonction de base

1. sigmoid$(x) = \frac{1}{2}(tanh(\frac{1}{2}x) + 1)$

$sigmoid(x) = \frac{1}{2}(tanh(\frac{1}{2}x) + 1) \iff 2sigmoid(x) - 1 = tanh(\frac{x}{2})$

$2sigmoid(x) - 1 = \frac{2-1-exp(-x)}{1+exp(-x)} = \frac{1-exp(-x)}{1+exp(x)} = \frac{exp(\frac{x}{2})}{exp(\frac{x}{2})}(\frac{1-exp(-x)}{1+exp(-x)})$

$= \frac{exp(\frac{x}{2})-exp(\frac{x}{2})exp(-x)}{exp(\frac{x}{2})+exp(\frac{x}{2})exp(-x)} = \frac{exp(\frac{x}{2})-exp(\frac{-x}{2})}{exp(\frac{x}{2})+exp(\frac{-x}{2})} = tanh(\frac{x}{2})$

2. $\ln sigmoid(x) = -softplus(-x)$

$\ln sigmoid(x) = \ln \frac{1}{1+exp(-x)} = \ln(1) - \ln(1+exp(-x)) = 0 - \ln(1+exp(-x)) = -softplus(-x)$

3. $\frac{d\ sigmoid}{dx}(x) = sigmoid(x)(1 - sigmoid(x))$

$\frac{d\ sigmoid(x)}{dx} = \frac{d((1+exp(-x))^{-1})}{dx} = \frac{-1}{(1+e^{-x})^2}(-e^{-x}) = (\frac{1}{1+e^{-x}})(\frac{e^{-x}}{1+e^{-x}})$

$= sigmoid(x)(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}) = sigmoid(x)(1 - sigmoid(x))$

4. Dérivée de tanh : $tanh'(x) = 1 - tanh^2(x)$

$\frac{d}{dx}(\frac{e^x-e^{-x}}{e^x+e^{-x}}) = \frac{(e^x-e^{-x})'(e^x+e^{-x})-(e^x-e^{-x})(e^x+e^{-x})'}{(e^x+e^{-x})^2} = \frac{(e^x+e^{-x})^2-(e^x-e^{-x})^2}{(e^x+e^{-x})^2}$

$= \frac{(e^x+e^{-x})^2}{(e^x+e^{-x})^2} - \frac{(e^x-e^{-x})^2}{(e^x+e^{-x})^2} = 1 - tanh^2(x)$

5. Fonction sign en utilisant des fonctions indicatrices

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \implies sign(x) = \mathbb{1}_{\{x>0\}}(x) - \mathbb{1}_{\{x<0\}}(x)$$

6. Dérivée de la fonction valeur absolue $abs(x) = |x|$

$\forall x \in \mathbb{R}$, $|x| = \sqrt{x^2} \implies \frac{d|x|}{dx} = \frac{d(x2)^{\frac{1}{2}}}{dx} = \frac{1}{2}(x^2)^{-\frac{1}{2}}2x = \frac{x}{\sqrt{x^2}} = \frac{x}{|x|}$

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases} \implies |x| = x * sign(x)$$

$abs'(x) = \frac{x}{x*sign(x)} = \frac{1}{sign(x)}$ mais on veut que abs'(0) = 0.
Alors, on écrit abs'(x) = sign(x)

7. Dérivée de la fonction rect

$$rect(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Alors,

$$rect'(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \implies rect'(x) = \mathbb{1}_{\{x>0\}}(x)$$

8. Soit le carré de la norme L_2 d'un vecteur : $||x||_2^2 = \sum_i x_i^2$. Le vecteur de gradient est : $\frac{\partial ||x||_2^2}{\partial x} =$

$$\frac{\partial \sum_i x_i^2}{\partial x} = \begin{bmatrix} \frac{\partial(x_1^2+x_2^2+\cdots+x_n^2)}{\partial x_1} \\ \vdots \\ \frac{\partial(x_1^2+x_2^2+\cdots+x_n^2)}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 2x$$

9. Soit la norme L_1 d'un vecteur :$||x||_1 = \sum_i |x_i|$. Le vecteur de gradient est : $\frac{\partial ||x||_1}{\partial x} =$

$$\frac{\partial \sum_i |x_i|}{\partial x} = \begin{bmatrix} \frac{\partial(|x_1|+|x_2|+\cdots+|x_n|)}{\partial x_1} \\ \vdots \\ \frac{\partial(|x_1|+|x_2|+\cdots+|x_n|)}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} sign(x_1) \\ sign(x_2) \\ \vdots \\ sign(x_n) \end{bmatrix} = sign(x)$$

2. Calcul du gradient pour l'optimisation des paramètres d'un réseau de neurones pour la classification multiclasse

1. Vecteur des sorties des neurones de la couche cachée h^s en fonction de h^a.

$b^{(1)} \in \mathbb{R}^{d_h}$
$h^a = b^{(1)} + W^{(1)}x$

$$h^a = \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_{d_h}^{(1)} \end{pmatrix} + \begin{pmatrix} w_{11}^{(1)} & w_{12}^{(1)} & \dots & w_{1d}^{(1)} \\ \vdots & \dots & \dots & \vdots \\ w_{d_h1}^{(1)} & w_{d_h2}^{(1)} & \dots & w_{d_hd}^{(1)} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_1 \end{pmatrix}$$

$$= \begin{pmatrix} b_1^{(1)} + w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + \dots + w_{1d}^{(1)} x_d \\ \vdots \\ b_{d_h}^{(1)} + w_{d_h 1}^{(1)} x_1 + w_{d_h 2}^{(1)} x_2 + \dots + w_{d_h d}^{(1)} x_d \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_1 \end{pmatrix}$$

$\Rightarrow h_j^a = b_j^{(1)} + \sum_{i=1}^{d} w_{ji}^{(1)} x_i$
$h^s = rect(h^a)$

2. Vecteur d'activations des neurones de la couche de sortie o^a à partir de leurs entrées h^s.

$W^{(2)}$ est m x d_h et $b^{(2)} \in \mathbb{R}^m$
$o^a = b^{(2)} + W^{(2)} h^s$

$$o^a = \begin{pmatrix} b_1^{(2)} \\ \vdots \\ b_m^{(2)} \end{pmatrix} + \begin{pmatrix} w_{11}^2 & w_{12}^2 & \dots & w_{1d_h}^2 \\ \vdots & \dots & \dots & \vdots \\ w_{m1}^2 & w_{m2}^2 & \dots & w_{md_h}^2 \end{pmatrix} \begin{pmatrix} h_1^s \\ \vdots \\ h_{d_h}^s \end{pmatrix}$$

$\implies o_k^a = b_k^{(2)} + \sum_{i=1}^{d_h} w_{ki}^{(2)} h_i^s$

3. Les o_k^s sont positifs et somment à 1.

$o_k^s = softmax(o^a)_k = \frac{exp(o_k^a)}{\sum_{i=1}^m exp(o_i^a)}$
Les o_k^s sont positifs par définitions de exp(x) (i.e $\forall x \in \mathbb{R}, exp(x) > 0$)

$\sum_{k=1}^m o_k^s = \sum_{k=1}^m \frac{exp(o_k^a)}{\sum_{i=1}^m exp(o_i^a)} = \frac{\sum_{k=1}^m exp(o_k^a)}{\sum_{i=1}^m exp(o_i^a)} = 1$
Il est important que les o_k^s soient positif et qu'ils somment à 1, car cela permet d'interpréter o_k^s commpe P(Y = k|X=x) (c'est-à-dire qu'on interprète o_k^s comme étant la probabilité que l'entrée x soit de la classe k)

4. $L(x, y) = -log(o_y^s(x))$ en fonction de o^a

$L(x, y) = -log(\frac{exp(o_y^a)}{\sum_{i=1}^m exp(o_i^a)}) = log(\sum_{i=1}^m exp(o_i^a)) - log(exp(o_y^a))$
$= log(\sum_{i=1}^m exp(o_i^a)) - o_y^a$

5. Risque empire : $\hat{R}$. L'ensemble θ des paramètre du réseau. Le nombre de paramètres scalaires n_θ.

$\hat{R} = \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}(log(\sum_{j=1}^{m} exp(o_j^a(x^{(i)})) - o_{y^{(i)}}^a(x^{(i)}))$

$\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$

$n_\theta = d_h \text{ x } d + d_h + m \text{ x } d_h + m$

Le problème d'optimisation qui correspond à l'entrainement du réseau permettant de trouver une valeur optimale des paramètres est $\arg\min_\theta \hat{R}(\theta, D_{train})$

6. Pseudo-code la descente de gradient pour ce problème

Initialize θ
for N iteration :
$\theta \leftarrow \theta - \eta(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta}(log(\sum_{j=1}^{m} exp(o_j^a(x^{(i)})) - o_{y^{(i)}}^a(x^{(i)})))$

7. $\frac{\partial L}{\partial o^a} = o^s - onehot_m(y)$

Pour k $\neq$ y,

$\frac{\partial L(x,y)}{\partial o_k^a} = \frac{\partial(log(\sum_{j=1}^{m} exp(o_j^a)) - o_y^a)}{\partial o_k^a} = \frac{\partial log(\sum_{j=1}^{m} exp(o_j^a))}{\partial o_k^a}$
$= \frac{1}{\sum_{j-1}^{m} exp(o_j^a)} * \frac{\partial \sum_{j-1}^{m} exp(o_j^a)}{\partial o_k^a} = \frac{exp(o_k^a)}{\sum_{j=1}^{m} exp(o_j^a)} = o_k^s$

$\frac{\partial L(x,y)}{\partial o_y^a} = \frac{\partial(log(\sum_{j=1}^{m} exp(o_j^a)) - o_y^a)}{\partial o_y^a} = \frac{exp(o_y^a)}{\sum_{j=1}^{m} exp(o_j^a)} - 1 = o_y^s - 1$

Alors,

$$\frac{\partial L(x,y)}{\partial o^a} = \begin{pmatrix} \frac{\partial L}{\partial o_1^a} \\ \dots \\ \frac{\partial L}{\partial o_m^a} \end{pmatrix}$$

$$= \begin{pmatrix} o_1^s \\ \dots \\ o_y^s \\ \dots \\ o_m^s \end{pmatrix} - \begin{pmatrix} 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{pmatrix}$$

$= o^s - onehot_m(y)$

8. L'expression correspondante en numpy

$grad\_oa = os - np.eye(m)[y-1]$
$y \in \{1, \dots, m\}$ et le vecteur $onehot_m$ à des index de 0 à m-1

9. $\frac{\partial L}{\partial W^{(2)}}$ et $\frac{\partial L}{\partial b^{(2)}}$

pour $k \neq y$,
$\frac{\partial L(x,y)}{\partial W_{kj}^{(2)}} = o_k^s * \frac{\partial o_k^a}{\partial W_{kj}^{(2)}} = o_k^s * \frac{\partial (b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s)}{\partial W k_{kj}^{(2)}} = o_k^s * h_j^s$

pour $k = y$, $\frac{\partial L(x,y)}{\partial W_{kj}^{(2)}} = (o_y^s - 1) * \frac{\partial o_y^a}{\partial W_{yj}^{(2)}} = (o_y^s - 1) * h_j^s = o_y^s h_j^s - h_j^s$

pour $k \neq y$, $\frac{\partial L(x,y)}{\partial b_k^{(2)}} = o_k^s * \frac{\partial (b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s)}{\partial b_k^{(2)}} = o_k^s$
pour $k = y$, $\frac{\partial L(x,y)}{\partial b_k^{(2)}} = o_k^s - 1$
Alors,

$$\frac{\partial L}{\partial W^{(2)}} = \begin{pmatrix} o_1^s h_1^s & o_1^s h_2^s & \dots & o_1^s h_{d_h}^s \\ \vdots & \dots & \dots & \vdots \\ o_m^s h_1^s & o_m^s h_2^s & \dots & o_m^s h_{d_h}^s \end{pmatrix} - \begin{array}{cc} \begin{pmatrix} 0 & \dots & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ h_1^s & h_2^s & \dots & h_{d_h}^s \\ \vdots & \dots & \dots & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix} & \begin{array}{c} \\ \\ \leftarrow \text{ ligne y} \\ \\ \\ \end{array} \end{array}$$

$$\frac{\partial L}{\partial b^{(2)}} = \begin{pmatrix} o_1^s \\ \vdots \\ o_y^s \\ \vdots \\ o_m^s \end{pmatrix} - \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = o^s - onehot_m(y)$$

10. Les expression correspondantes en numpy

$grad\_b2 = os - np.eye(m)[y-1]$

grad_b2 est m x 1
grad_W2 est m x d_h

car $\frac{\partial L}{\partial b^{(2)}} = o^s - onehot_m(y)$ et
$\frac{\partial L}{\partial W^{(2)}}$

$$= o^s h^{s^T} - \begin{pmatrix} 0 & \dots & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ h_1^s & h_2^s & \dots & h_{d_h}^s \\ \vdots & \dots & \dots & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

où o^s et $onehot_m(y)$ sont m x 1, h^s est d_h x 1 et la matrice contenant que des zéros et les éléments de h^s est m x d_h

11. $\frac{\partial L}{\partial h^s}$

$\frac{\partial L}{\partial h_j^s} = \sum_{k=1}^m \frac{\partial L}{\partial o_k^a} \frac{\partial o_k^a}{\partial h_j^s} = \sum_{k=1}^m \frac{\partial L}{\partial o_k^a} \frac{\partial (b^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s)}{\partial h_j^s}$
$= \sum_{k=1}^m \frac{\partial L}{\partial o_k^a} W_{kj}^{(2)} = o_1^s W_{1j}^{(2)} + o_2^s W_{2j}^{(2)} + \dots + (o_y^s - 1) W_{yj}^{(2)} + \dots + o_m^s W_{mj}^{(2)}$

$= \sum_{k=1}^m o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}$
Alors,

$$\frac{\partial L}{\partial h^s} = \begin{pmatrix} \sum_{k=1}^m o_k^s W_{k1}^{(2)} - W_{y1}^{(2)} \\ \sum_{k=1}^m o_k^s W_{k2}^{(2)} - W_{y2}^{(2)} \\ \dots\dots \\ \sum_{k=1}^m o_k^s W_{kd_h}^{(2)} - W_{yd_h}^{(2)} \end{pmatrix}$$

12. L'expression correspondante en numpy

$\frac{\partial L}{\partial h^s} = W^{(2)^T} o^s - W^{(2)}[y,:]^T$

où $W^{(2)^T}$ est d_h x m, o^s est m x 1 et $w^{(2)}[y,:]^T$ est d_h x 1

$grad\_hs = np.dot(W2.T, os) - W2[y-1,:].reshape((d_h, 1))$

13. $\frac{\partial L}{\partial h^a}$

$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} * \frac{\partial h_j^s}{\partial h_j^a} = (\sum_{k=1}^m o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}) \frac{\partial (rect(h_j^a))}{\partial h_j^a} = (\sum_{k=1}^m o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)$

Alors,

$$\frac{\partial L}{\partial h^a} = \begin{pmatrix} (\sum_{k=1}^m o_k^s W_{k1}^{(2)} - W_{y1}^{(2)}) \mathbb{1}_{\{h_1^a > 0\}}(h_1^a) \\ \dots\dots \\ (\sum_{k=1}^m o_k^s W_{kd_h}^{(2)} - W_{yd_h}^{(2)}) \mathbb{1}_{\{h_{d_h}^a > 0\}}(h_{d_h}^a) \end{pmatrix}$$

14. L'expression correspondante en numpy

$$\frac{\partial L}{\partial h^a} = \left(\frac{\partial L}{\partial h^s} \right) \odot \begin{pmatrix} \mathbb{1}_{\{h_1^a > 0\}}(h_1^a) \\ \dots \\ \mathbb{1}_{\{h_{d_h}^a > 0\}}(h_{d_h}^a) \end{pmatrix}$$

où $\frac{\partial L}{\partial h^a}$, $\frac{\partial L}{\partial h^s}$ et le vecteur contenant les fonctions indicatrices sont d_h x 1

vector_indicator = np.array([1 if e> 0 else 0 for e in hs])
grad_ha = np.multiply(grad_hs, vector_indicator)

15. $\frac{\partial L}{\partial W^{(1)}}$ et $\frac{\partial L}{\partial b^{(1)}}$

$\frac{\partial L}{\partial W_{jl}^{(1)}} = \frac{\partial L}{\partial h_j^a} \frac{\partial h_j^a}{\partial W_{jl}^{(1)}} = (\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a) * \frac{\partial (b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i)}{\partial W_{jl}^{(1)}}$

$= (\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a) x_l = \frac{\partial L}{\partial h_j^a} * x_l$

$\frac{\partial L}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a} * \frac{\partial h_j^a}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a} * 1 = \frac{\partial L}{\partial h_j^a}$

$$\frac{\partial L}{\partial W^{(1)}} = \begin{pmatrix} \frac{\partial L}{\partial h_1^a} x_1 & \frac{\partial L}{\partial h_1^a} x_2 & \dots & \frac{\partial L}{\partial h_1^a} x_d \\ \frac{\partial L}{\partial h_2^a} x_1 & \frac{\partial L}{\partial h_2^a} x_2 & \dots & \frac{\partial L}{\partial h_2^a} x_d \\ \vdots & \dots & \dots & \vdots \\ \frac{\partial L}{\partial h_{d_h}^a} x_1 & \frac{\partial L}{\partial h_{d_h}^a} x_2 & \dots & \frac{\partial L}{\partial h_{d_h}^a} x_d \end{pmatrix}$$

$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial h^a}$

16. Sous forme matricielle et l'expression équivalente en numpy

$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial h^a}$ est d_h x 1

$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial h^a} * x^T$ est d_h x d car $\frac{\partial L}{\partial h^a}$ est d_h x 1 et x^T est 1 x d

$grad\_b1 = grad\_ha$
$grad\_W1 = np.outer(grad\_ha, x)$

17. $\frac{\partial L}{\partial x}$

$\frac{\partial L}{\partial x_l} = \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \frac{\partial h_j^a}{\partial x_l} = \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \frac{\partial (b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i)}{\partial x_l}$

$= \sum_{j=1}^{d_h} (\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a) W_{jl}^{(1)}$

Alors,

$$\frac{\partial L}{\partial x} = \begin{pmatrix} \sum_{j=1}^{d_h}(\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)})\mathbb{1}_{\{h_j^a>0\}}(h_j^a)W_{j1}^{(1)} \\ \ldots\ldots \\ \sum_{j=1}^{d_h}(\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)})\mathbb{1}_{\{h_j^a>0\}}(h_j^a)W_{jd}^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} * W_{j1}^{(1)} \\ \ldots\ldots \\ \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} * W_{jd}^{(1)} \end{pmatrix}$$

18. Le changement du gradient par rapport aux différents paramètre si on minimise $\tilde{R} = \hat{R} + \lambda_{11}(\sum_{i,j}|W_{ij}^{(1)}|) + \lambda_{12}(\sum_{i,j}(W_{ij}^{(1)})^2) + \lambda_{21}(\sum_{i,j}|W_{ij}^{(2)}|) + \lambda_{22}(\sum_{i,j}(W_{ij}^{(2)})^2)$ au lieu de $\hat{R}$

Il n'y a pas de différence pour $b^{(1)}$ et $b^{(2)}$ car $\frac{\partial \mathcal{L}(\theta)}{\partial b^{(1)}} = \frac{\partial \mathcal{L}(\theta)}{\partial b^{(2)}} = 0$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W^{(1)}} = \lambda_{11} \begin{pmatrix} sign(W_{11}^{(1)}) & sign(W_{12}^{(1)}) & \ldots & sign(W_{1d}^{(1)}) \\ sign(W_{21}^{(1)}) & \ldots & \ldots & \vdots \\ \vdots & \ldots & \ldots & \vdots \\ sign(W_{d_h1}^{(1)}) & sign(W_{d_h2}^{(1)}) & \ldots & sign(W_{d_hd}^{(1)}) \end{pmatrix} + \lambda_{12} \begin{pmatrix} 2W_{11}^{(1)} & 2W_{12}^{(1)} & \ldots & 2W_{1d}^{(1)} \\ \vdots & \ldots & \ldots & \vdots \\ 2W_{d_h1}^{(1)} & 2W_{d_h2}^{(1)} & \ldots & 2W_{d_hd}^{(1)} \end{pmatrix}$$

$= \lambda_{11} sign(W^{(1)}) + 2\lambda_{12}W^{(1)}$

Alors,
$\frac{\partial \tilde{R}}{\partial W^{(1)}} = \frac{\partial \hat{R}}{\partial W^{(1)}} + \lambda_{11} sign(W^{(1)}) + 2\lambda_{12}W^{(1)}$

et de la même façon, $\frac{\partial \tilde{R}}{\partial W^{(2)}} = \frac{\partial \hat{R}}{\partial W^{(2)}} + \lambda_{21} sign(W^{(2)}) + 2\lambda_{22}W^{(2)}$