IFT-3395 Devoir 3

Olivier St-Laurent, Maxime Daigle

2018-11-09

Question 1 Relations et dérivées de quelques fonction de base

1. Montrez que 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. Montez que $\ln sigmoid(x) = -softplus(-x)$

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

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

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

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

4. Montrez que la dérivée de tanh est : $tanh'(x) = 1 - tanh^2(x)$

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

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

5. exprimez la fonction sign en utilisant des fonctions indicatrices : sign(x) = ...

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

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

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

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

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

7. Écrivez la dérivée de la fonction rect.

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

Alors,

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

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

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

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

8. Soit la norme L_1 d<un vecteur :$||x||_1 = \sum_i |x_i|$. Écrivez le vecteur de gradient : $\frac{\partial |x|_1}{\partial x} = \dots$

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

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

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

1. Exprimez le vecteur des sorties des neurones de la couche cachée h^s en fonction de h^a.

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

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

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

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

2.Donnez la formule de calcul du vecteur d'activations des neurones de la couche de sortie o^a à partir de leurs entrées h^s sous la forme d'une expression de calcul matriciel, puis détaillez le calcul de o_a^k.

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

$$o^a = \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_m^{(2)} \end{pmatrix} + \begin{pmatrix} w_{11}^2 & w_{12}^2 & \dots & w_{1d_h}^2 \\ \vdots & \dots & \dots & \vdots \\ w_{m1}^2 & w_{m2}^2 & \dots & w_{md_h}^2 \end{pmatrix} \begin{pmatrix} h_1^s \\ \vdots \\ h_{d1}^s \end{pmatrix}$$

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

3. Démontrez que les o_k^s sont positifs et somment à 1. Pourquoi est-ce important ?

$O_k^s = softmax(O^a)_k = \frac{exp(O_k^a)}{\sum_{i=1}^{m} exp(O_i^a)}$
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. $L(x, y) = -log(O_y^s(x)) PrciserLenfonctiondeO^a$

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

5. Formuler $\hat{R}$. Indiquer précisement l'ensemble θ des paramètre du réseau. Indiquer à combien de paramètres scalaires n_θ cela correspond. Formuler le problème d'optimisation qui correspond à l ;entrainement du réseau pour trouver une valeur optimal des paramètres.

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

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

$n_\theta = d_h$ x $d + d_h$ x $d_h + m$

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

6. Exprimer avec un bref pseudo-code la descente de gradient pour ce problème

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

7. Montrez que $\frac{dL}{dO^a} = O^s - onehot_m(y)$

Pour k $\neq$ y,

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

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

Alors,

$$\frac{\partial L(x,y)}{\partial O^a} = \begin{pmatrix} \frac{\partial L}{\partial O_1^a} \\ \dots \\ \frac{\partial L}{\partial O_m^a} \end{pmatrix}$$

$$= \begin{pmatrix} O_1^s \\ \dots \\ O_y^s \\ \dots \\ O_m^s \end{pmatrix} - \begin{pmatrix} O \\ \dots \\ 1 \\ \dots \\ O \end{pmatrix}$$

$= O^s - onehot_m(y)$

8 Donner l'expression correspondante en numpy

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

9 calculer $\frac{\partial L}{\partial W^{(2)}}$ 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,

$$\frac{\partial L}{\partial W^{(2)}} = \begin{pmatrix} O_1^s h_1^s & O_1^s h_2^s & \dots & O_1^s h_{d_h}^s \\ \vdots & \dots & \dots & \vdots \\ O_m^s h_1^s & O_m^s h_2^s & \dots & O_m^s h_{d_h}^s \end{pmatrix} - \begin{pmatrix} O & \dots & \dots & O \\ \vdots & \dots & \dots & \vdots \\ h_1^s & h_2^s & \dots & h_{d_h}^s \\ \vdots & \dots & \dots & \vdots \\ O & \dots & \dots & O \end{pmatrix}$$

$$\frac{\partial L}{\partial b^{(2)}} = \begin{pmatrix} O_1^s \\ \vdots \\ O_y^s \\ \vdots \\ O_m^s \end{pmatrix} - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} = o^s - onehot_m(y)$$

10 Donner les expression correspondantes en numpy

$grad\_b2 = os - np.eye(m)[y-1]$
$grad\_W2 = np.outer(os, hs) - np.concatenate((np.zeroes((y-1, dh)), hs.reshape(1, dh), np.zeroes((m-y, dh))))$

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 $frac\partial L\partial W^{(2)}$

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

où O^s et $onehot_m(y)$ sont m x 1, h^s est $d_h x1$ et