UNIVERSITÉ DE MONTRÉAL

Rapport : Devoir2

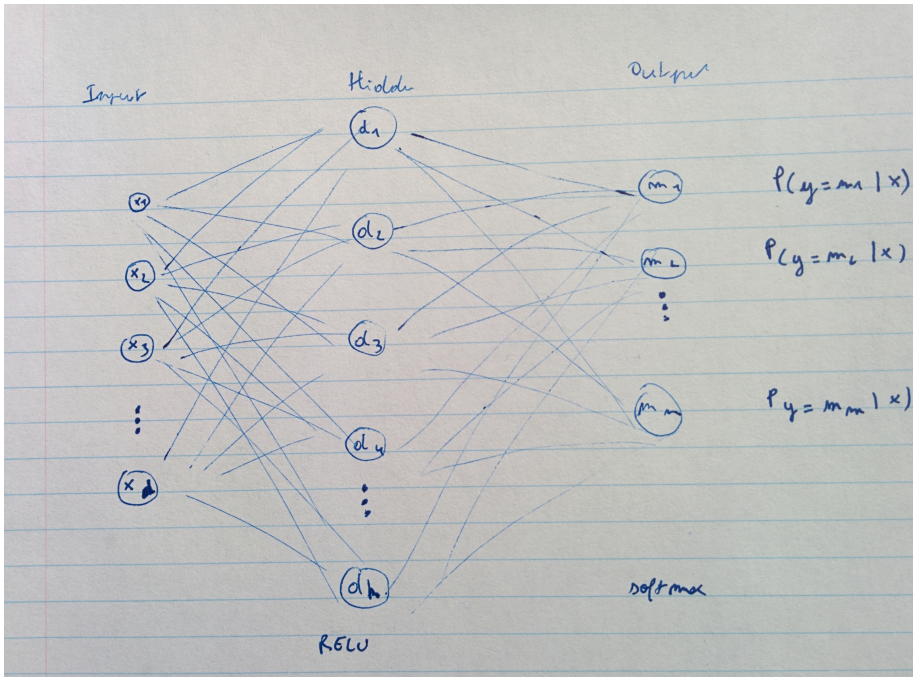
Pierre Gérard
Mathieu Bouchard

IFT3395-6390 Fondements de l'apprentissage machine
Pascal Vincent, Alexandre de Brébisson et César Laurent

Année académique 2015 - 2016

1 Partie théorique : Calcul du gradient pour l'optimisation des paramètres d'un réseau de neurones

Commençons par dessiner un rapide schéma du réseau de neurones étudié.



1.1 Exercice a)

b est de dimension d_h
Le vecteur d'activation est : $h_a = W^{(1)}x + b$
Avec $h_{a_i} = W^{(1)}_{i1}x_1 W^{(1)}_{i2}x_2 + ... + W^{(1)}_{id}x_d b_i$
Et $h_{s_i} = h_{a_i} * I_{\{h_{a_i}>0\}} = max(h_{a_i}, 0)$

1.2 Exercice b)

$W^{(2)}$ est de dimension $m \times d_h$
$b^{(2)}$ est de dimension m
Le vecteur d'activation est : $o^a = W^{(2)}h^s + b^{(2)}$
Avec $o^a_k = W^{(2)}_{k1}h^s_1 + W^{(2)}_{k2}h^s_2 + ... + W^{(2)}_{kn}h^s_n + b^{(2)}_k$

1.3 Exercice c)

$o^s = softmax(o^a) = \frac{1}{\sum_{i=1}^m e^{o^a_i}}(e^{o^a_1}, e^{o^a_2}, ..., e^{o^a_n})$

Donc $o^s_k = \frac{e^{o^a_k}}{\sum_{i=1}^m e^{o^a_i}}$

$e^x : \mathbb{R} \to \mathbb{R}^+$ donc la somme au numérateur de la fonction ci-dessus sera positive et la somme au dénominateur aussi. Une fraction de deux nombres positifs sera toujours positif donc o^s_k est toujours positif.

$\sum_{i=1}^m o^s_i = \sum_{i=1}^m \frac{e^{o^a_i}}{\sum_{j=1}^m e^{o^a_j}}$

$= \frac{1}{\sum_{j=1}^m e^{o^a_j}} \sum_{i=1}^m e^{o^a_i}$

$= \frac{\sum_{i=1}^{m} e^{o_i^a}}{\sum_{j=1}^{m} e^{o_j^a}} = 1$

C'est important car cela signifie que les sorties sont les probabilité pour l'entrée d'être d'une certaine classe et ces classes sont mutuellement exclusives.

1.4 Exercice d)

$L(x, y) = -log(o_y^s(x))$

$= -log \frac{e^{o_y^a(x)}}{\sum_{i=1}^{m} e^{o_i^a(x)}}$

$= -log(e^{o_y^a(x)}) + log(\sum_{i=1}^{m} e^{o_i^a(x)})$

$= -o_y^a(x) + log(\sum_{i=1}^{m} e^{o_i^a(x)})$

1.5 Exercice e)

L'erreur empirique vaut :

$\widehat{R}(f(x, y), D_n) = \sum_{i=1}^{n} L(x, y)$

Les paramètres sont :

$\theta = \{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$ avec $W^{(1)}$ et $b^{(1)}$ représentant les connexions synaptiques entre l'entrée et la couche cachée et $W^{(2)}$ et $b^{(2)}$ représentant les connexions synaptiques entre la couche cachée et la sortie.

Le problèmes d'optimisation revient donc à l'équation suivante :

$\theta^* = argmin_\theta \widehat{R}(f, D)$

En additionnant les dimensions de chacun des élémentes de θ, on trouve que n_θ correspond alors à $d_h \times d + d_h + m \times d_h + m$ paramètres scalaires.

1.6 Exercice f)

```
def gradient(ensemble_donne):
. somme = 0
. for x in ensemble_donne:
. . sum += derivate_L(x)
. return somme

theta = initialisation des params de maniere random
epsilon = small_value
learningRate =
while learningRate*gradient() < epsilon :  # attention aux boucles infini
. theta = theta + learningRate*gradient()
```

1.7 Exercice g)

Pour $k \neq y$:

$\frac{\partial L}{\partial O_k^a} = \frac{\partial}{\partial O_k^a}(-O_y^a(x) + log \sum_{i=1}^{m} e^{O_i^a(x)})$

$= 0 + \frac{\partial}{\partial O_k^a} log \sum_{i=1}^{m} e^{O_i^a(x)}$

$= \frac{\frac{\partial}{\partial O_k^a} \sum_{i=1}^{m} e^{O_i^a(x)}}{\sum_{i=1}^{m} e^{O_i^a(x)}}$

$= \frac{\frac{\partial}{\partial O_k^a} e^{O_k^a(x)}}{\sum_{i=1}^{m} e^{O_i^a(x)}}$

$= \frac{e^{O_k^a(x)}}{\sum_{i=1}^{m} e^{O_i^a(x)}}$

Pour $k = y$:

$\frac{\partial L}{\partial O_y^a} = \frac{\partial}{\partial O_y^a}(-O_y^a(x) + log \sum_{i=1}^m e^{O_i^a(x)})$

$= \frac{-\partial O_y^a}{\partial O_y^a} + \frac{\partial}{\partial O_y^a} log \sum_{i=1}^m e^{O_i^a(x)}$

$= -1 + \frac{e^{O_y^a(x)}}{\sum_{i=1}^m e^{O_i^a(x)}}$

Le premier terme du résultat vaut donc -1 seulement lorsque $k = y$ et 0 sinon. On obtient alors :

$\frac{\partial L}{\partial O_k^a} = \frac{1}{\sum_{i=1}^m e^{O_i^a(x)}}(e^{O_1^a(x), e^{O_2^a(x)}, \ldots, e^{O_m^a(x)}} - onehot_m(y)) = O^s - onehot_m(y)$

1.8 Exercice h)

```
import numpy as np
grad_oa = os   # avec os les sorties
grad_oa[y] = grad_oa[y] - 1  #onehot
```

1.9 Exercice i)

Réponse entière donnée

1.10 Exercice j)

La dimension de :
- $\frac{\partial L}{\partial b^{(2)}}$ est m
- $\frac{\partial L}{\partial W^{(2)}}$ est $m \times d_h$
- $\frac{\partial L}{\partial o^a}$ est $m \times 1$
- ${h^s}^T$ est $1 \times d_h$

```
grad_b2 = grad_oa
grad_w2 = grad_oa * np.transpose(h_s)
```

1.11 Exercice k)

Réponse entière donnée

1.12 Exercice l)

La dimension de :
- $\frac{\partial L}{\partial h^s}$ est d_h
- ${W^{(2)}}^T$ est $d_h \times m$
- $\frac{\partial L}{\partial o^a}$ est $m \times 1$

```
grad_hs = np.transpose(w_2) * grad_oa
```

1.13 Exercice m)

$\frac{\partial rect(z)}{\partial z} = \frac{\partial max(0,z)}{\partial z} = \begin{cases} 0 \text{ si } z \leq 0 \\ 1 \text{ sinon} \end{cases}$

Pour $h_j^a \leq 0$:

$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s}\frac{\partial h_j^s}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s}\frac{\partial rect(h_j^a)}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} * 0 = 0$

Pour $h_j^a > 0$:

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

Au final :
$\frac{\partial L}{\partial h_j^a} = \begin{cases} 0 \text{ si } h_j^a \leq 0 \\ \frac{\partial L}{\partial h_j^s} \text{ sinon} \end{cases}$
$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} * I_{\{h_j^a > 0\}}$

1.14 Exercice n)

$\frac{\partial L}{\partial h^a} = \begin{bmatrix} \frac{\partial L}{\partial h_1^s} * I_{\{h_1^a > 0\}} \\ \frac{\partial L}{\partial h_2^s} * I_{\{h_2^a > 0\}} \\ \vdots \\ \frac{\partial L}{\partial h_{d_h}^s} * I_{\{h_{d_h}^a > 0\}} \end{bmatrix}$ qui est un vecteur colonne de taill d_h

```
grad_ha = grad_hs * np.where(ha > 0, 1, 0)
```

1.15 Exercice o)

Pour $b^{(1)}$:
$\frac{\partial L}{\partial b_k^{(1)}} = \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial b_k^{(1)}}$
$= \frac{\partial L}{\partial h_k^a} \frac{\partial \sum_{j'} W_{kj'}^{(1)} x_{j'} + b_k^{(1)}}{\partial b_k^{(1)}}$
$= \frac{\partial L}{\partial h_k^a}$
Pour $W^{(1)}$:
$\frac{\partial L}{\partial W_{kj}^{(1)}} = \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial W_{kj}^{(1)}}$
$= \frac{\partial L}{\partial h_k^a} \frac{\partial \sum_{j'} W_{kj'}^{(1)} x_{j'} + b_k^{(1)}}{\partial W_{kj}^{(1)}}$
$= \frac{\partial L}{\partial h_k^a} x_j$

1.16 Exercice p)

Expressions matricielles :
$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial h^a}$
$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial h^a} (x)^T$
La dimension de :
- $\frac{\partial L}{\partial b^{(1)}}$ est d_h
- $\frac{\partial L}{\partial W^{(1)}}$ est $d_h \times n$ car $\frac{\partial L}{\partial h^{(a)}}$ est $d_h \times 1$ et x est $1 \times n$

```
grad_b1 = grad_ha
grad_w1 = grad_ha * np.transpose(x)
```

1.17 Exercice q)

$\frac{\partial L}{\partial x_j} = \sum_{k=1}^{n} \frac{\partial L}{\partial h_k^a} \frac{\partial h_a^k}{\partial x_j}$
$= \sum_{k=1}^{n} \frac{\partial L}{\partial h_k^a} (\frac{\partial}{\partial x_j} \sum_{j'} W_{kj'}^{(1)} x_{j'} + b_k^{(1)})$
$= \sum_{k=1}^{n} \frac{\partial L}{\partial h_k^a} W_{kj}^{(1)}$
Sous forme matricielle, on obtient :
$\frac{\partial L}{\partial x} = {W^{(1)}}^T \frac{\partial L}{\partial h^a}$

1.18 Exercice r)

Pour voir l'effet qu'a le terme de régularisation sur le gradient, on applique de nouveau la technique de rétropropagation mais en utilisant cette fois-ci le risque empirique régularisé $\tilde{\mathrm{R}}$. Dans la plupart des étapes, puisqu'on nous demandait de ne pas substituer les termes des expressions des dérivées partielles déjà calculées, l'ajout d'un terme de régularisation qui ne dépend que de $W^{(1)}$ et $W^{(2)}$ n'affecte pas directement l'expression de la dérivée partielle calculée.

Le premier endroit où l'on peut appercevoir un changement sur la valeur du gradiant est lors du calcul des gradients par rapport aux paramètres $W^{(2)}$ et b^2 ; soit au point i). Pour la dérivée partielle par rapport à b^2, il n'y aura pas de changement puisque $\mathcal{L}(\theta)$ n'a aucune influence sur ce paramètre. Pour $W^{(2)}$ par contre, on remarque une différence :

$$
\begin{aligned}
&\frac{\partial(L+\lambda\mathcal{L})}{\partial W^{(2)}_{kj}} = \frac{\partial L}{\partial W^{(2)}_{kj}} + \frac{\partial \lambda\mathcal{L}}{\partial W^{(2)}_{kj}} \\
&= \frac{\partial L}{\partial o^a_k} h^s_j + \frac{\partial \lambda\mathcal{L}}{\partial W^{(2)}_{kj}} \text{ (calculez en i)} \\
&= \frac{\partial L}{\partial o^a_k} h^s_j + \frac{\lambda\partial \sum_{i,j}(W^{(1)}_{i,j})^2 + \sum_{i,j}(W^{(2)}_{i,j})^2}{\partial W^{(2)}_{kj}} \\
&= \frac{\partial L}{\partial o^a_k} h^s_j + \frac{\lambda\partial \sum_{i,j}(W^{(1)}_{i,j})^2}{\partial W^{(2)}_{kj}} + \frac{\lambda\partial \sum_{i,j}(W^{(2)}_{i,j})^2}{\partial W^{(2)}_{kj}} \\
&= \frac{\partial L}{\partial o^a_k} h^s_j + 0 + \frac{\lambda\partial (W^{(2)}_{k,j})^2}{\partial W^{(2)}_{kj}} \\
&= \frac{\partial L}{\partial o^a_k} h^s_j + 2\lambda W^{(2)}_{k,j}
\end{aligned}
$$

On obtient alors une différence de $2\lambda W^{(2)}_{k,j}$ en utilisant $\tilde{\mathrm{R}}$ plutôt que $\hat{\mathrm{R}}$.

L'autre endroit où l'on note une différence se trouve au moment de calculer les gradients par rapport aux éléments des paramètres $W^{(1)}$ et b^1 de la couche cachée. Tout comme pour l'étape précédante, pour la dérivée partielle par rapport à b^1, il n'y aura pas de changement puisque $\mathcal{L}(\theta)$ n'a aucune influence sur ce paramètre. Pour $W^{(1)}$ par contre, on remarque une différence :

$$
\begin{aligned}
&\frac{\partial(L+\lambda\mathcal{L})}{\partial W^{(1)}_{kj}} = \frac{\partial L}{\partial W^{(1)}_{kj}} + \frac{\partial \lambda\mathcal{L}}{\partial W^{(1)}_{kj}} \\
&= \frac{\partial L}{\partial h^a_k} x_j + \frac{\partial \lambda\mathcal{L}}{\partial W^{(1)}_{kj}} \text{ (calculez en o)} \\
&= \frac{\partial L}{\partial h^a_k} x_j + \frac{\lambda\partial \sum_{i,j}(W^{(1)}_{i,j})^2 + \sum_{i,j}(W^{(2)}_{i,j})^2}{\partial W^{(1)}_{kj}} \\
&= \frac{\partial L}{\partial h^a_k} x_j + \frac{\lambda\partial \sum_{i,j}(W^{(1)}_{i,j})^2}{\partial W^{(2)}_{kj}} + \frac{\lambda\partial \sum_{i,j}(W^{(2)}_{i,j})^2}{\partial W^{(1)}_{kj}} \\
&= \frac{\partial L}{\partial h^a_k} x_j + \frac{\lambda\partial (W^{(1)}_{k,j})^2}{\partial W^{(1)}_{kj}} + 0 \\
&= \frac{\partial L}{\partial h^a_k} x_j + 2\lambda W^{(1)}_{k,j}
\end{aligned}
$$

On obtient alors une différence de $2\lambda W^{(1)}_{k,j}$ en utilisant $\tilde{\mathrm{R}}$ plutôt que $\hat{\mathrm{R}}$.