UNIVERSITÉ DE MONTRÉAL

Rapport : Devoir2

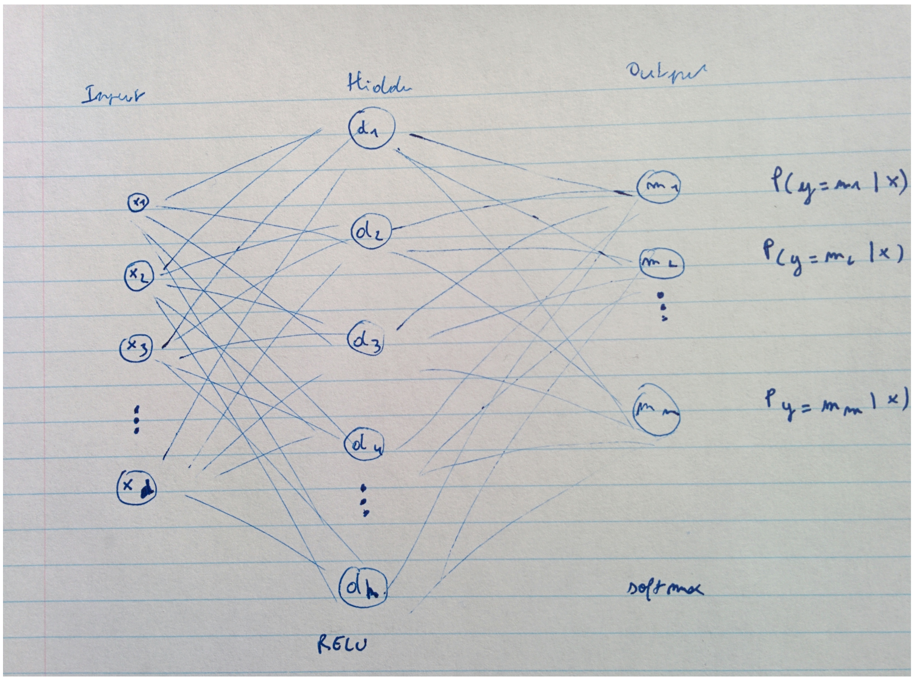
Pierre Gérard
Mathieu Bouchard

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 $h_{a_i} = W_{i1}^{(1)} x_1 W_{i2}^{(1)} x_2 + ... + W_{id}^{(1)} x_d b_i$

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 $o_k^a = W_{k1}^{(2)} h_{s_1} + W_{k2}^{(2)} h_{s_2} + ... + W_{kn}^{(2)} h_{s_n} + b_k^{(2)}$

1.3 Exercice c)

$o^s = softmax(o^a) = \frac{1}{\sum_{i=1}^m e^{o_i^a}} (e^{o_1^a}, e^{o_2^a}, ..., e^{o_n^a})$

Donc $o_k^s = \frac{e^{o_k^a}}{\sum_{i=1}^m e^{o_i^a}}$

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

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

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

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

$= 1$

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)

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

1.5 Exercice e)

L'erreur empirique vaut :
$\widehat{R}(f,D) = \sum_{i=1}^n L(x,y)$

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.

demander en tp des précisions sur la notation ci-dessus

Le problèmes d'optimisation revient donc à l'équation suivante :
$\theta^* = argmin_\theta \widehat{R}(f,D)$
Les dimensions sont :
n_θ est de dimension $d_h \times d + d_h + m \times d_h + m$

1.6 Exercice f)

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

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

1.7 Exercice g)

Pour $k \neq y$:
$\frac{\partial L}{\partial O_k^a} = \frac{\partial}{\partial O_k^a}(-O_y^a(x) + log\sum_{i=1}^m e^{O_i^a(x)})$
$= 0 + \frac{\partial}{\partial O_k^a} log\sum_{i=1}^m e^{O_i^a(x)}$
$= \frac{\frac{\partial}{\partial O_k^a}\sum_{i=1}^m e^{O_i^a(x)}}{\sum_{i=1}^m e^{O_i^a(x)}}$
$= \frac{\frac{\partial}{\partial O_k^a} e^{O_k^a(x)}}{\sum_{i=1}^m e^{O_i^a(x)}}$
$= \frac{e^{O_k^a(x)}}{\sum_{i=1}^m e^{O_i^a(x)}}$

Comment ajouter de l'espace entre les lignes pour que ce soit plus lisible ?

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 = 1/np.sum(e_oa) * np.array(e_oa)
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)

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

formatter acollade

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 :

$\frac{\partial L}{\partial h_j^a} = \{0 si h_j^a \leq 0 \frac{\partial L}{\partial h_j^s} sinon$

formatter acollade

$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} * I_{\{h_j^a > 0\}}$

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 une vecteur colonne de taill d_h

Possible de simplifier l'expression de ce vecteur ?

```
(a + np.abs(a))/2  # astuce pour garder que les nombre positif
```

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 $1 \times n$

Expression numpy ?

1.17 Exercice q)

À valider!

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

1.18 Exercice r)