FACULTÉ DES SCIENCES
DÉPARTEMENT D'INFORMATIQUE

IFT712 - Techniques d'apprentissage
TP4

préparé par
Agathe Le Bouler (leba3207)
Philippe Spino (spip2401)
Gabriel Gibeau Sanchez (gibg2501)

présenté à
Pierre-Marc Jodoin

9 avril 2020

1 Question 1

On dispose de l'expression de l'opération Softmax tel que :

$$y_{w_i}(\overrightarrow{x}) = \frac{e^{a_i}}{\sum_{c=1}^{K} e^{a_c}} \tag{1}$$

Ainsi que de l'expression de la fonction de perte entropie croisée :

$$E_D(W) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} ln(y_{w_i}(\overrightarrow{x_n})) \tag{2}$$

On calcule le gradient de la fonction de perte par rapport à a_i tel que :

$$\begin{aligned}
\frac{\partial E_D(W)}{\partial a_i} &= \frac{\partial}{\partial a_i}\left(-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} ln(y_{w_i}(\overrightarrow{x_n}))\right) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \frac{\partial}{\partial a_i}\left(ln(y_{w_i}(\overrightarrow{x_n}))\right) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \frac{\partial (ln(y_{w_k}(\overrightarrow{x_n}))}{\partial y_{w_k}\overrightarrow{x_n}} \cdot \frac{\partial y_{w_k}(\overrightarrow{x_n})}{\partial a_i} \\
\frac{\partial E_D(W)}{\partial a_i} &= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \frac{1}{y_{w_k}(\overrightarrow{x_n})} \cdot \frac{\partial y_{w_k}(\overrightarrow{x_n})}{\partial a_i}
\end{aligned} \tag{3}$$

1.0.1 Dérivée du Softmax

On s'intéresse à la dérivée de l'expression du Softmax :

$$\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} = \frac{\partial}{\partial a_j}\left(\frac{e^{a_i}}{\sum_{c=1}^{K} e^{a_c}}\right) \tag{4}$$

On note :

$$\begin{aligned}
g(a) &= e^{a_i} \\
h(a) &= \sum_{c=1}^{K} e^{a_c}
\end{aligned} \tag{5}$$

La dérivée de

$$\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} = \frac{g^{'}(a)h(a) - g(a)h^{'}(a)}{(h(a))^2} \tag{6}$$

Calcul de $\frac{\partial g(a)}{\partial a_j}$:

$$\begin{aligned}
\frac{\partial g(a)}{\partial a_j} &= \frac{\partial e^{a_i}}{\partial a_j} \\
&= \frac{\partial e^{a_i}}{\partial a_i} \cdot \frac{\partial e^{a_i}}{\partial a_j} \\
&= e^{a_i} \cdot \frac{\partial e^{a_i}}{\partial a_j} \\
\frac{\partial g(a)}{\partial a_j} &= \begin{cases} e^{a_i} \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases}
\end{aligned} \tag{7}$$

Calcul de $\frac{\partial h(a)}{\partial a_j}$:

$$\begin{aligned}
\frac{\partial h(a)}{\partial a_j} &= \frac{\partial}{\partial a_j}(\sum_{c=1}^{K} e^{a_c}) \\
&= \frac{\partial}{\partial a_j}\left(\sum_{c=1, c\neq j}^{K} e^{a_c} + e^{a_j}\right) \\
&= \frac{\partial}{\partial a_j}\left(\sum_{c=1, c\neq j}^{K} e^{a_c} + e^{a_j}\right) + \frac{\partial}{\partial a_j} e^{a_j} \\
&= 0 + e^{a_j} \\
\frac{\partial h(a)}{\partial a_j} &= e^{a_j}
\end{aligned} \tag{8}$$

On a pu déterminer $g(a)$, sa dérivée, $h(a)$ et sa dérivée, nous pouvons donc maintenant calculer $\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j}$ en considérant deux cas :

Pour i=j :

$$\begin{aligned}
\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} &= \frac{\frac{\partial g(a)}{\partial a_j} \cdot h(a) - g(a) \cdot \frac{\partial h(a)}{\partial a_j}}{h(a)^2} \\
&= \frac{(e^{a_i}) \cdot (\sum_{c=1}^{K} e^{a_c}) - (e^{a_i}) \cdot (e^{a_j})}{(\sum_{c=1}^{K} e^{a_c})^2}
\end{aligned} \tag{9}$$

Étant donné que i=j, on peut réécrire :

$$
\begin{aligned}
&= \frac{(e^{a_i}) \cdot (\sum_{c=1}^{K} e^{a_c}) - (e^{a_i}) \cdot (e^{a_i})}{(\sum_{c=1}^{K} e^{a_c})^2} \\
&= \frac{e^{a_i} \left(\sum_{c=1}^{K} e^{a_c} - e^{a_i} \right)}{(\sum_{c=1}^{K} e^{a_c})^2} \\
&= \frac{e^{a_i}}{\sum_{c=1}^{K} e^{a_c}} \cdot \left(\frac{\sum_{c=1}^{K} e^{a_c}}{\sum_{c=1}^{K} e^{a_c}} - \frac{e^{a_i}}{\sum_{c=1}^{K} e^{a_c}} \right)
\end{aligned} \tag{10}
$$

On reconnaît l'expression de $y_{\overrightarrow{w_i}}(\overrightarrow{x})$,

$$
\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} = y_{\overrightarrow{w_i}}(\overrightarrow{x}) \cdot (1 - y_{\overrightarrow{w_i}}(\overrightarrow{x})) \tag{11}
$$

pour i = j.

Pour i≠ j :

$$
\begin{aligned}
\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} &= \frac{\frac{\partial g(a)}{\partial a_j} \cdot h(a) - g(a) \cdot \frac{\partial h(a)}{\partial a_j}}{h(a)^2} \\
&= \frac{0 \cdot (\sum_{c=1}^{K} e^{a_c}) - (e^{a_i}) \cdot (e^{a_j})}{(\sum_{c=1}^{K} e^{a_c})^2} \\
&= -\frac{e^{a_i} \cdot e^{a_j}}{(\sum_{c=1}^{K} e^{a_c})^2} \\
&= -\frac{e^{a_i}}{\sum_{c=1}^{K} e^{a_c}} \cdot \frac{e^{a_j}}{\sum_{c=1}^{K} e^{a_c}}
\end{aligned} \tag{12}
$$

On reconnaît les expressions de $y_{\overrightarrow{w_i}}(\overrightarrow{x})$ et $y_{\overrightarrow{w_j}}(\overrightarrow{x})$.

$$
\frac{\partial y_{w_i}(\overrightarrow{x})}{\partial a_j} = -y_{\overrightarrow{w_i}}(\overrightarrow{x}) \cdot y_{\overrightarrow{w_j}}(\overrightarrow{x}) \tag{13}
$$

pour $i \neq j$.

1.0.2 Suite de la dérivée de l'entropie croisée

Nous avions trouvé l'expression suivante :

$$
\frac{\partial E_D(W)}{\partial a_i} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{kn}}{y_{w_k}(\overrightarrow{x_n})} \cdot \frac{\partial y_{w_k}(\overrightarrow{x_n})}{\partial a_i} \tag{14}
$$

On considère les deux cas, tel que $i = j$ et $i \neq j$ et on remplace les dérivées des softmax précédemment trouvées :

$$\begin{aligned}
\frac{\partial E_D(W)}{\partial a_i} &= -\sum_{n=1}^{N} \left[\frac{t_{in}}{\cancel{y_{\overrightarrow{w_i}}(\overrightarrow{x})}} \cdot \left(\cancel{y_{\overrightarrow{w_i}}(\overrightarrow{x_n})}(1 - y_{\overrightarrow{w_i}}(\overrightarrow{x_n})) + \sum_{k=1,k\neq i} \frac{t_{kn}}{\cancel{y_{\overrightarrow{w_k}}(\overrightarrow{x_n})}} \cdot \left(-y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) \cdot \cancel{y_{\overrightarrow{w_k}}(\overrightarrow{x_n})} \right) \right) \right] \\
&= -\sum_{n=1}^{N} \left[t_{in} \left(1 - y_{\overrightarrow{w_i}}(\overrightarrow{x_n})\right) - \sum_{k=1,k\neq i} t_{kn} \cdot y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) \right] \\
&= -\sum_{n=1}^{N} \left[t_{in} - t_{in} \cdot y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) - \sum_{k=1,k\neq i} t_{kn} \cdot y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) \right] \\
&= -\sum_{n=1}^{N} \left[t_{in} - y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) \left(t_{in} + \sum_{k=1,k\neq i} t_{kn} \right) \right] \\
&= -\sum_{n=1}^{N} \left[t_{in} - y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) \cdot \left(\sum_{k=1}^{K} t_{kn} \right) \right]
\end{aligned} \tag{15}$$

Or on sait que $\sum_{k=1}^{K} t_{kn} = 1$, on peut donc écrire :

$$\frac{\partial E_D(W)}{\partial a_i} = \sum_{n=1}^{N} (y_{\overrightarrow{w_i}}(\overrightarrow{x_n}) - t_{in}) \tag{16}$$

2 Question 2

On dispose des données telles que :

$$\overrightarrow{x_i} = [i, ii, iii, iv, v]$$

$$t_i = \{1, 2, 3, 4, 5\}$$

où $\overrightarrow{x_i}$ sont les attributs des données d'entraînement $\overrightarrow{x}$, et t_i les différentes valeurs des cibles associées à une donnée $\overrightarrow{x_i}$.

2.1 Distribution de vraisemblance

La distribution de vraisemblance se définit telle que :

$$P(\overrightarrow{x}, \overrightarrow{t} | \overrightarrow{w})$$

où $\overrightarrow{x}$ sont les données d'entraînement connus, $\overrightarrow{t}$ sont les cibles connues associés aux données d'entraînement et $\overrightarrow{w}$ sont les poids du réseau inconnus que le réseau doit déterminer durant la phase d'entraînement.
On peut également écrire l'expression telle que :

$$P((\overrightarrow{x_1}, \overrightarrow{x_2}, ..., \overrightarrow{x_N}), (\overrightarrow{t_1}, \overrightarrow{t_2}, ..., \overrightarrow{t_N}) | \overrightarrow{w_1}, \overrightarrow{w_2}, ..., \overrightarrow{w_N})$$

Afin de déterminer les valeurs des paramètres $\overrightarrow{w}$ du réseau, on peut une méthode d'estimation de maximum de vraisemblance. Les valeurs des paramètres sont trouvées telles que maximisant la probabilité de vraisemblance permettant de produire les données qui sont effectivement observées en entraînement.
Cela s'exprime tel que :

$$\overrightarrow{w} = argmax_{\overrightarrow{w}} P(\overrightarrow{x}, \overrightarrow{t} | \overrightarrow{w})$$

2.2 Distribution à priori

L'expression de la distribution à priori se définit telle que :

$$P(\overrightarrow{t})$$

que l'on peut également exprimer tel que :

$$P(\overrightarrow{t}) = \sum_{\overrightarrow{x}} P(\overrightarrow{x}, \overrightarrow{t})$$

En applicant la règle du produit des probabilités, on peut écrire :

$$\sum_{\overrightarrow{x}} P(\overrightarrow{x}, \overrightarrow{t}) = \sum_{\overrightarrow{x}} P(\overrightarrow{x} | \overrightarrow{t}) \cdot P(\overrightarrow{t})$$

Étant donné que les variables $\overrightarrow{x}$ et $\overrightarrow{t}$ sont des données connues, leur distribution est alors connue.

2.3 Hypothèse de distribution gaussienne

On considère le problème suivant que l'on définit tel que la distribution de la consommation d'essence des véhicules en sachant que ceux sont des voitures de sport.

$$P(\overrightarrow{x_i} = v | \overrightarrow{t_i} = 1)$$

De manière générale, nous sommes au fait que la consommation d'essence d'un véhicule dépend de son moteur, de l'essence utilisée et de la conduite adoptée par le chauffeur, entre autres. Les facteurs sont nombreux et il existe donc obligatoirement des variations de consommation d'essence entre les véhicules. Considérant spécifiquement les voitures de sport, leur consommation d'essence n'échappe à ces variations dépendemment des modèles de véhicules. Nous pouvons cependant imaginer sans soucis que leur consommation d'essence est importante de manière globale. Une moyenne élevée de la consommation d'essence pour les voitures de sport est donc tout à fait considérable et une hypothèse gaussienne est effectivement envisageable.

3 Question 3

On note les expressions de la descente de gradient par momentum telles que :

$$\begin{aligned} v_{t+1} &= \rho \cdot v_t + \nabla E_{\overrightarrow{x_n}}(w_t) \\ w_{t+1} &= w_t + \eta \cdot v_{t+1} \end{aligned} \tag{17}$$

où v est le terme de vélocité déterminant la direction et la vitesse à laquelle le paramètre w doit être modifié.
ρ peut être défini comme un coefficient de friction influençant directement sur la vitesse. Cet hyper-paramètre décroissant amortit la vitesse et réduit l'énergie cinétique du système dans le but de permettre la détection des minima locaux.
La vitesse du système a, quant à elle, un effet sur la position. Comme v est initialisée à 0 et est un paramètre croissant, l'accélération du système y est alors représenté.