Examen formatif

GRO-720
Réseaux de neurones artificiels

Génie robotique
Faculté de génie
Université de Sherbrooke

Hiver 2021

Copyright ©2021, Faculté de génie
Université de Sherbrooke

(Cette page est laissée vide intentionnellement)

GRO-720 #1	GRO-720 #2	GRO-720 #3
/25		

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 1

Répondez aux sous-questions suivantes par VRAI ou FAUX.

1. Lorsqu'un réseau de neurones est entraîné, il existe une seule solution optimale. [C1 – 5 pts]

2. Il est préférable d'utiliser une fonction de coût d'erreur quadratique moyenne pour entraîner un réseau de neurones dans le cadre d'une tâche de classification. [C1 – 5 pts]

3. L'augmentation des données permet de réduire le phénomène de sur apprentissage. [C1 – 5 pts]

4. L'ajout d'une couche de normalisation de lot permet d'introduire une non-linéarité dans le réseau de neurones. [C1 – 5 pts]

5. Augmenter le seuil de décision pour améliorer la précision d'un classificateur fait généralement diminuer son rappel. [C1 – 5 pts]

(*Espace supplémentaire si nécessaire pour la question 1*)

GRO-720 #1	GRO-720 #2	GRO-720 #3
/50		

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 2

Dérivez l'équation pour générer le gradient d'entrée $\frac{\partial L}{\partial \mathbf{x}}$ à partir du gradient de sortie $\frac{\partial L}{\partial \mathbf{y}}$ (*backward*) pour une couche de softmax. Vous devez démontrer les étapes mathématiques qui vous permettent d'arriver à ce résultat, et exprimer le résultat en fonction des éléments qui constituent le vecteur de sortie $\mathbf{y}$.

(*Espace supplémentaire si nécessaire pour la question 2*)

GRO-720 #1	GRO-720 #2	GRO-720 #3
	/50	

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 3

Vous devez concevoir un réseau de neurones qui déterminera si un segment audio contient ou non un signal de parole humaine. Le vecteur d'entrée est composé de 20 trames audio concatenées, chacune contenant 39 éléments qui représentent les coefficients MFCC (Mel-Frequency Cepstral Coefficients), pour un vecteur d'une dimension de 780 éléments. Le réseau sera composé de 3 couches linéaires et de fonctions d'activation (aucune couche de normalisation de lot). Ce dernier doit générer une sortie qui indique si de la parole humaine est présente (1) ou absente (0).

1. Déterminer l'architecture du réseau de neurones sous forme de schéma-bloc. Indiquez les dimensions des matrices/vecteurs qui contiennent les paramètres de chaque couche (libre à vous de choisir le nombre de neurones dans chaque couche lorsque cet hyperparamètre peut prendre plusieurs valeurs différentes). Indiquez également les dimensions des vecteurs (entrée, sortie, intermédiaires). [C2 – 20 pts]

2. Combien y a-t-il de paramètres à apprendre au total? [C2 – 10 pts]

3. Identifiez la fonction de coût utilisée pour entraîner le réseau et donnez l'équation qui permet de calculer le coût entre la prédiction et la cible. [C2 – 10 pts]

4. Lorsque tous les paramètres sont initialisés de manière aléatoire, nous pouvons supposer que la sortie $\hat{y}$ est modélisée par une variable aléatoire dont la moyenne correspond à 0.5. Quelle devrait être la valeur moyenne de l'erreur de la fonction de coût avant de débuter l'entraînement? [C2 – 10 pts]

(*Espace supplémentaire si nécessaire pour la question 3*)

GRO-720 #1	GRO-720 #2	GRO-720 #3
	/10	/70

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 4

Nous avons un ensemble de points $(x, y) \in \mathbb{R}^2$ que nous aimerions modéliser à l'aide du polynôme suivant:

$\hat{y} = a_3x^3 + a_2x^2 + a_1x + a_0$

1. Quelle fonction coût devrait être minimisée pour accomplir cette tâche? Donnez également l'équation de la fonction de coût. [C2 – 10 pts]
2. Si l'optimisation se fait en utilisant la méthode de descente de gradient, à quoi correspond le gradient pour chacun des paramètres a_0, a_1, a_2 et a_3 (on cherche donc $\frac{\partial L}{\partial a_0}$, $\frac{\partial L}{\partial a_1}$, $\frac{\partial L}{\partial a_2}$ et $\frac{\partial L}{\partial a_3}$)? [C3 – 20 pts]
3. Si le taux d'apprentissage est de $\mu = 0.01$, que les paramètres $a_3 = 1$, $a_2 = -2$, $a_1 = 0$ et $a_0 = 1$, quelles seront les nouvelles valeurs des paramètres si on applique une descente de gradient stochastique pour le point $x = 2$ et $y = 0$? [C3 – 30 pts]
4. Démontrez que la mise à jour des paramètres améliore la prédiction du modèle pour le point $x = 2$ et $y = 0$. [C3 – 10 pts]
5. En appliquant la descente de gradient pour l'ensemble des données, l'erreur diminue mais converge vers une valeur non-nulle assez importante. Quelle serait une cause possible de ce phénomène? [C3 – 10 pts]

(*Espace supplémentaire si nécessaire pour la question 4*)

GRO-720 #1	GRO-720 #2	GRO-720 #3
/25		

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 5

Un réseau de neurones est entraîné pour classifier des extraits audio. Le réseau doit déterminer si chaque extrait contient de la musique (classe positive) ou pas (classe négative). On utilise ce classificateur sur un ensemble de données de test, et on obtient les résultats suivants:

- 3642 segments audio contenant de la musique ont été identifiés comme étant de la musique
- 102 segments audio contenant de la musique ont été identifiés comme n'étant pas de la musique
- 3720 segments audio ne contenant pas de la musique ont été identifiés comme n'étant pas de la musique
- 93 segments audio contenant ne contenant pas de la musique ont été identifiés comme étant de la musique

Calculez la précision, le rappel et la justesse pour cette expérience.

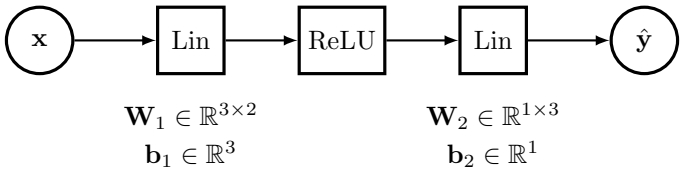
(*Espace supplémentaire si nécessaire pour la question 5*)

GRO-720 #1	GRO-720 #2	GRO-720 #3
	/40	/30

S.V.P. Ne rien inscrire dans les grilles réservées à la compilation de l'évaluation

Question 6

Soit le réseau de neurones suivant avec deux couches linéaires et un rectificateur:



Les paramètres des deux couches linéaires sont les suivants:

$$\mathbf{W}_1 = \begin{bmatrix} +1 & -2 \\ +0 & +2 \\ -1 & +1 \end{bmatrix} \mathbf{b}_1 = \begin{bmatrix} +0 \\ -2 \\ +1 \end{bmatrix} \mathbf{W}_2 = \begin{bmatrix} -1 & +1 & +1 \end{bmatrix} \mathbf{b}_2 = [+1]$$

Le vecteur d'entrée et sa cible sont:

$$\mathbf{x} = \begin{bmatrix} +3 \\ +2 \end{bmatrix}, \; y = 2$$

1. Calculez la prédiction $\hat{y}$ pour le vecteur $\mathbf{x}$. [C1 - 30 pts]
2. Calculez l'erreur pour la fonction de coût d'erreur quadratique moyenne. [C1 – 10 pts]
3. Calculez la première mise à jour de $\mathbf{W}_1$ par descente de gradient en utilisant un taux d'apprentissage de $\mu = 0.01$. [C3 – 30 pts]

(*Espace supplémentaire si nécessaire pour la question 6*)

Aide-mémoire

Dérivées

$$\frac{d}{dx}cx = c \qquad \frac{d}{dx}x^c = cx^{c-1} \qquad \frac{d}{dx}\log(x) = \frac{1}{x} \qquad \frac{d}{dx}\exp(x) = \exp(x)$$

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

Fonctions non-linéaires

ReLU: $y = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$ Sigmoïde: $y = \dfrac{1}{1+\exp(-x)}$ Tan hyperbolique: $y = \tanh(x) = \dfrac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$

Fonctions de coûts

EQM: $L = \sum_{j=1}^{J}(\hat{y} - y)^2$ Entropie croisée: $L = -\sum_{j=1}^{J} y_j \log \hat{y}_j$ DKL: $L = -\sum_{j=1}^{J} y_j \log\left(\dfrac{y_j}{\hat{y}_j}\right)$

Couches

Linéaire: $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$ $\dfrac{\partial L}{\partial \mathbf{x}} = \mathbf{W}^T\dfrac{\partial L}{\partial \mathbf{y}}$ $\dfrac{\partial L}{\partial \mathbf{W}} = \dfrac{\partial L}{\partial \mathbf{y}}\mathbf{x}^T$ $\dfrac{\partial L}{\partial \mathbf{b}} = \dfrac{\partial L}{\partial \mathbf{y}}$

Normalisation de lot:

$$\boldsymbol{\mu}_B = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i \qquad \boldsymbol{\sigma}_B^2 = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \boldsymbol{\mu}_B)^{\circ 2} \qquad \hat{\mathbf{x}} = (\mathbf{x} - \boldsymbol{\mu}_B) \oslash \sqrt{(\boldsymbol{\sigma}_B^2 + \epsilon)} \qquad \mathbf{y} = \boldsymbol{\gamma} \circ \hat{\mathbf{x}} + \boldsymbol{\beta} \qquad \frac{\partial L}{\partial \hat{\mathbf{x}}_i} = \frac{\partial L}{\partial \mathbf{y}_i} \circ \boldsymbol{\gamma}$$

$$\frac{\partial L}{\partial \boldsymbol{\sigma}_B^2} = \sum_{i=1}^{M}\frac{\partial L}{\partial \hat{\mathbf{x}}_i} \circ (\mathbf{x}_i - \boldsymbol{\mu}_B) \circ -\frac{1}{2}(\boldsymbol{\sigma}_B^2 + \epsilon)^{\circ -3/2} \qquad \frac{\partial L}{\partial \boldsymbol{\mu}_B} = \left(-\sum_{i=1}^{M}\frac{\partial L}{\partial \hat{\mathbf{x}}_i} \oslash \sqrt{(\boldsymbol{\sigma}_B^2 + \epsilon)}\right) + \frac{-2}{M}\frac{\partial L}{\partial \boldsymbol{\sigma}_B^2} \circ \sum_{i=1}^{M}(\mathbf{x}_i - \boldsymbol{\mu}_B)$$

$$\frac{\partial L}{\partial \mathbf{x}_i} = \frac{\partial L}{\partial \hat{\mathbf{x}}_i} \oslash \sqrt{(\boldsymbol{\sigma}_B^2 + \epsilon)} + \frac{2}{M}\frac{\partial L}{\partial \boldsymbol{\sigma}_B^2}(\mathbf{x}_i - \boldsymbol{\mu}_B) + \frac{1}{M}\frac{\partial L}{\partial \boldsymbol{\mu}_B} \qquad \frac{\partial L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{M}\frac{\partial L}{\partial \mathbf{y}_i} \circ \frac{\partial L}{\partial \hat{\mathbf{x}}_i} \qquad \frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{M}\frac{\partial L}{\partial \mathbf{y}_i}$$

Softmax: $\mathbf{y} = \dfrac{\exp(\mathbf{x})}{\|\exp(\mathbf{x})\|_1}$ $y_j = \dfrac{\exp(x_j)}{\sum_{i=1}^{I}\exp(x_i)}$ $\dfrac{\partial L}{\partial \mathbf{x}} = \mathbf{D}\dfrac{\partial L}{\partial \mathbf{y}}$ $d_{i,j} = \begin{cases} y_j^2 & i \neq j \\ y_j(1 - y_j) & i = j \end{cases}$

Mesures de performance

Précision: $\dfrac{VP}{VP + FP}$ Rappel: $\dfrac{VP}{VP + FN}$ Justesse: $\dfrac{VP + VN}{VP + VN + FP + FN}$