A rendre le : 16 février 2019

Instructions

- *Montrez votre démarche pour toutes les questions !*
- *Utilisez un logiciel de traitement de texte comme LaTeX.*
- *Vous devez soumettre toutes vos réponses sur la page Gradescope du cours.*

Question 1. En utilisant les définitions de la dérivée et de la fonction de *Heaviside* (fonction marche d'escalier) suivantes :

$$\frac{d}{dx}f(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \qquad H(x) = \begin{cases} 1 & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

1. Montrez que la dérivée de la fonction d'activation ReLU (Unité de Rectification Linéaire) $g(x) = \max\{0, x\}$, **partout où elle existe** est égale a la fonction de Heaviside.
2. Donnez deux définitions alternatives de $g(x)$ en utilisant $H(x)$.
3. Montrez qu'on peut bien approximer $H(x)$ en utilisant la fonction logistique (la sigmoïde) $\sigma(x) = \frac{1}{1+e^{-kx}}$ asymptotiquement (c.-à-d. pour des valeurs de k large), k étant un paramètre.

*4. Même si la fonction de Heaviside est non-dérivable, on peut définir sa **dérivée distributionelle**. Pour une fonction F, considérez la fonctionnelle $F[\phi] = \int_{\mathbb{R}} H(x)\phi(x)dx$, ϕ étant une fonction lisse (indéfiniment dérivable, c.-à-d. dans $\mathcal{C}^\infty$) à support compact ($\phi(x) = 0$ quand $|x| \geq A$, pour un certain $A > 0$).
Montrez que si F est dérivable, alors $F'[\phi] = -\int_{\mathbb{R}} F(x)\phi'(x)dx$. En utilisant cette formule comme une définition de la dérivée distributionelle dans le cas des fonctions non-dérivables, montrez que $H'[\phi] = \phi(0)$. ($\delta[\phi] \doteq \phi(0)$ est la fonction delta de Dirac (ou distribution de Dirac))

Answer 1.

Question 2. Soit $x \in \mathbb{R}^n$ un vecteur. On rappelle les définitions de la fonction softmax (fonction exponentielle normalisée) : $S : \boldsymbol{x} \in \mathbb{R}^n \mapsto S(\boldsymbol{x}) \in \mathbb{R}^n$ tel que $S(\boldsymbol{x})_i = \frac{e^{\boldsymbol{x}_i}}{\sum_j e^{\boldsymbol{x}_j}}$ et de la fonction diagonale : $\text{diag}(\boldsymbol{x})_{ij} = \boldsymbol{x}_i$ si $i = j$ et $\text{diag}(\boldsymbol{x})_{ij} = 0$ si $i \neq j$; et du symbole de Kronecker : $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

1. Montrez que la dérivée de la fonction softmax est $\frac{dS(\boldsymbol{x})_i}{d\boldsymbol{x}_j} = S(\boldsymbol{x})_i \left(\delta_{ij} - S(\boldsymbol{x})_j\right)$.
2. Exprimez la matrice Jacobienne $\frac{\partial S(\boldsymbol{x})}{\partial \boldsymbol{x}}$ en utilisant la notation matricielle/vectorielle. Utilisez $\text{diag}(\cdot)$.
3. Calculez la matrice Jacobienne de la fonction logistique $\sigma(\boldsymbol{x}) = 1/(1 + e^{-\boldsymbol{x}})$.
4. Soient $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^n$ des vecteurs, tels que $\boldsymbol{y} = f(\boldsymbol{x})$. Soit L une fonction de coût, differentiable. D'après le théorème de dérivation des fonctions composées (règle de dérivation en chaîne), $\nabla_{\boldsymbol{x}} L = (\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})^\top \nabla_{\boldsymbol{y}} L$, dont le calcul a une complexité en temps de $\mathcal{O}(n^2)$, en général. Montrer que si $f(\boldsymbol{x}) = \sigma(\boldsymbol{x})$ ou $f(\boldsymbol{x}) = S(\boldsymbol{x})$, alors la multiplication matrice-vecteur précédente peut être simplifiée, et évaluée en utilisant $\mathcal{O}(n)$ opérations.

Answer 2.

Question 3. On rappelle la définition de la fonction softmax : $S(\boldsymbol{x})_i = e^{\boldsymbol{x}_i}/\sum_j e^{\boldsymbol{x}_j}$.

1. Montrez que la fonction softmax est invariante aux translations, c'est-à-dire : $S(\boldsymbol{x} + c) = S(\boldsymbol{x})$, où c est une constante.
2. Montrez que la fonction softmax n'est pas invariante aux multiplications scalaires. On définit $S_c(\boldsymbol{x}) = S(c\boldsymbol{x})$ où $c \geq 0$. Quels seraient les effets si on choisissait $c = 0$ ou $c \to \infty$.
3. Soit $\boldsymbol{x} \in \mathbb{R}^2$ un vecteur. On peut représenter une probabilité catégorique sur deux classes en utilisant la fonction softmax. Montrez que $S(x)$ peut être reparamétrisée en utilisant la fonction sigmoïde, c'est-à-dire : $S(\boldsymbol{x}) = [\sigma(z), 1 - \sigma(z)]^\top$ où z est un scalaire, qu'il faut exprimer en fonction de $\boldsymbol{x}$.
4. Soit $\boldsymbol{x} \in \mathbb{R}^K$ un vecteur ($K \geq 2$). Montrez que $S(\boldsymbol{x})$ peut être représentée avec $K-1$ paramètres, c.-à-d. $S(\boldsymbol{x}) = S([0, y_1, y_2, ..., y_{K-1}]^\top)$ où y_i sont des scalaires, à exprimer en fonction de $\boldsymbol{x}$ pour $i \in \{1, ..., K-1\}$.

Answer 3.

Question 4. On considère un réseau de neurones à deux couches $y : \mathbb{R}^D \to \mathbb{R}^K$ de la forme suivante :

$$y(x, \Theta, \sigma)_k = \sum_{j=1}^{M} \omega_{kj}^{(2)} \sigma \left(\sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)} \right) + \omega_{k0}^{(2)}$$

pour $1 \leq k \leq K$. Les paramètres du réseau sont $\Theta = (\omega^{(1)}, \omega^{(2)})$. La fonction d'activation utilisée est σ, la fonction logistique. Montrez qu'il existe un réseau équivalent de la même forme, avec des paramètres $\Theta' = (\tilde{\omega}^{(1)}, \tilde{\omega}^{(2)})$, avec la fonction d'activation tanh, tel que $y(x, \Theta', \tanh) = y(x, \Theta, \sigma)$ pour tout $x \in \mathbb{R}^D$. Exprimez Θ' en fonction de Θ.

Answer 4.

Question 5. Soit N un entier strictement positif. On veut montrer que pour toute fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ et pour tout échantillon $\mathcal{S} \subset \mathbb{R}^n$ de taille N, il existe un ensemble de paramètres pour un réseau de neurones à deux couches, tel que la sortie $y(\boldsymbol{x})$ correspond à $f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in \mathcal{S}$. En d'autres termes, on veut interpoler la fonction f avec le réseau de neurones y pour tout ensemble fini $\mathcal{S}$.

1. Écrivez la forme générique de la fonction $y : \mathbb{R}^n \to \mathbb{R}^m$ qui définit un réseau de neurones à deux couches avec $N-1$ neurones dans la couche cachée (*hidden units*), avec une fonction d'activation ϕ, et une fonction linéaire à la dernière couche (output linéaire), en fonction des poids et biais $(\boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)})$ et $(\boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)})$.
2. Dans le reste de cet exercice, on se restreint au cas où $\boldsymbol{W}^{(1)} = [\boldsymbol{w}, \cdots, \boldsymbol{w}]^T$ pour un certain $\boldsymbol{w} \in \mathbb{R}^n$ (c'est-à-dire que les lignes de la matrice $\boldsymbol{W}^{(1)}$ sont toutes pareilles). Montrez que le problème d'interpolation sur l'ensemble $\mathcal{S} = \{\boldsymbol{x}^{(1)}, \cdots \boldsymbol{x}^{(N)}\} \subset \mathbb{R}^n$ peut être réduit à la résolution d'une équation matricielle $\boldsymbol{M}\tilde{\boldsymbol{W}}^{(2)} = \boldsymbol{F}$, $\tilde{\boldsymbol{W}}^{(2)}$ et $\boldsymbol{F}$ étant deux matrices de taille $N \times m$ définies par
$$\tilde{\boldsymbol{W}}^{(2)} = [\boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}]^\top \qquad \boldsymbol{F} = [f(\boldsymbol{x}^{(1)}), \cdots, f(\boldsymbol{x}^{(N)})]^\top$$
Exprimez la matrice $\boldsymbol{M}$ de taille $N \times N$ en fonction de $\boldsymbol{w}$, $\boldsymbol{b}^{(1)}$, ϕ et $\boldsymbol{x}^{(i)}$.
*3. **Preuve avec la fonction d'activation ReLU.** Supposons que les $\boldsymbol{x}^{(i)}$ sont tous distincts. On choisit $\boldsymbol{w}$ tel que $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ sont aussi tous distincts (Essayez de montrer l'existence d'un tel $\boldsymbol{w}$,

mais ce n'est pas requis pour le devoir - voir Assignment 0). On définit $\boldsymbol{b}_j^{(1)} = -\boldsymbol{w}^\top \boldsymbol{x}^{(j)} + \epsilon$, où $\epsilon > 0$. Trouvez une valeur de ϵ telle que $\boldsymbol{M}$ est une matrice triangulaire à éléments diagonaux non-nuls. Conclure. (Indice : Définir un ordre sur les $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$)

*4. **Preuve avec des fonctions d'activation similaires à la sigmoïde.** Supposons que ϕ est continue, bornée, $\phi(-\infty) = 0$ et $\phi(0) > 0$. On écrit $\boldsymbol{w}$ comme $\boldsymbol{w} = \lambda \boldsymbol{u}$. On définit $\boldsymbol{b}_j^{(1)} = -\lambda \boldsymbol{u}^\top \boldsymbol{x}^{(j)}$. En laissant $\boldsymbol{u}$ fixe, montrez que $\lim_{\lambda \to +\infty} \boldsymbol{M}$ est une matrice triangulaire à éléments diagonaux non-nuls. Conclure. (A noter que cela préserve le fait que les $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ sont distincts.)

Answer 5.

Question 6. Calculez la convolution complète (*full* convolution), valide (*valid* convolution), et similaire (*same* convolution), avec retournement de noyau (*kernel flipping*) pour les matrices unidimensionnelles suivantes : $\big[1, 2, 3, 4\big] * \big[1, 0, 2\big]$

Answer 6.

Question 7. On considère un réseau de neurones à convolution. On suppose que l'entrée (*input*) est une image en couleurs de taille 256×256 dans la représentation Rouge Vert Bleu (*RGB*). La première couche convolue 64 noyaux 8×8 avec l'entrée, en utilisant un pas (*stride*) de 2, et une marge (*padding*) nulle de zéro. La deuxième couche sous-échantillonne (*downsampling*) la sortie (*output*) de la première couche avec un *max-pool* 5×5 sans chevauchement (*no overlapping*). La troisième couche convolue 128 noyaux 4×4 avec un pas de 1, et une marge de 1 de chaque côté.

1. Quelle est la dimension de la sortie à la dernière couche ?
2. Sans compter les biais, combien de paramètres sont requis pour la dernière couche ?

Answer 7.

Question 8. Supposons qu'on a des données de taille $3 \times 64 \times 64$. Dans ce qui suit, donnez la configuration d'une couche d'un réseau neuronal convolutif qui satisfait les hypothèses spécifiées. Répondre avec la taille du noyau (k), le pas (s), la marge (p), et la dilatation (*dilation* d, en utilisant la convention $d = 0$ pour une convolution sans dilatation). Utilisez des fenêtres carrées seulement (par exemple, même valeur de k pour la hauteur et la largeur).

1. La taille de la sortie de la première couche est $(64, 32, 32)$.
   (a) Supposons que $k = 8$ sans dilatation.
   (b) Supposons que $d = 6$, et que $s = 2y$.
2. La taille de la sortie de la deuxième couche est $(64, 8, 8)$. Supposons que $p = 0$ et que $d = 0$.
   (a) Spécifier k et s pour une couche POOL sans chevauchement.
   (b) Quel serait la taille de la sortie si on avait $k = 8$ et $s = 4$ plutôt ?
3. La taille de la sortie de la dernière couche est $(128, 4, 4)$.
   (a) Supposons qu'on n'utilise ni marge ni dilatation.
   (b) Supposons que $d = 1$, et que $p = 2$.
   (c) Supposons que $p = 1$, et que $d = 0$.

Answer 8.