Date de remise : 4 février (11pm), 2020

Instructions

- *Montrez vos traces pour toutes les questions !*
- *Utilisez le modèle LaTeX que nous vous fournissons pour écrire vos réponses. Vous pouvez réutiliser les racourcis pour la notation, les équations et/ou les tables. Voir la politique sur les travaux pratique sur le site du cours pour plus de détails.*
- *Remettez vos questions électroniquement par Gradescope.*
- ***TAs for this assignment are Jie Fu, Sai Rajeswar, and Akilesh B***

Question 1 (4-4-4)**.** 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 à 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.

Answer 1.

1.

Question 2 (3-3-3-3)**.** 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 les y_i sont des scalaires à exprimer en fonction de $\boldsymbol{x}$ pour $i \in \{1, ..., K-1\}$.

Answer 2.

Question 3 (16)**.** On considère un réseau de neurones à deux couches $y : \mathbb{R}^D \to \mathbb{R}^K$ de la forme suivante :

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

TABLE 1 – Exemple de DA vers l'avant avec $y = f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ à $(x_1, x_2) = (2, 5)$ et fixant $\dot{x}_1 = 1$ pour calculer $\partial y / \partial x_1$.

	Trace d'évaluation vers l'avant		
	v_{-1}	$= x_1$	$= 2$
	v_0	$= x_2$	$= 5$
	v_1	$= \ln(v_1)$	$= \ln(2)$
	v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
$\Downarrow$	v_3	$= \sin(v_0)$	$\sin(5)$
	v_4	$= v_1 + v_2$	$= 0.6931 + 10$
	v_5	$= v_4 - v_3$	$= 10.6931 + 0.9589$
	y	$= v_5$	$= 11.6521$

	Trace de dérivée vers l'avant		
	$= \dot{v}_{-1}$	$\dot{x}_1$	$= 1$
	$= \dot{v}_0$	$\dot{x}_2$	$= 0$
	$\dot{v}_1$	$= \dot{v}_{-1} / v_{-1}$	$= 1/2$
	$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + v_{-1} \times \dot{v}_0$	$= 1 \times 5 + 2 \times 0$
$\Downarrow$	$\dot{v}_3$	$= \cos v_0 \times \dot{v}_0$	$= \cos(5) \times 0$
	$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
	$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
	$= \dot{y}$	$\dot{v}_5$	$= 5.5$

TABLE 2 – Exemple de DA vers l'arrière avec $y = f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ à $(x_1, x_2) = (2, 5)$. Fixant $\bar{y} = 1$, $\partial y / \partial x_1$ et $\partial y / \partial x_2$ calculés en un seul balayage inverse.

	Trace d'évaluation vers l'avant		
	v_{-1}	$= x_1$	$= 2$
	v_0	$= x_2$	$= 5$
	v_1	$= \ln(v_1)$	$= \ln(2)$
	v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
$\Downarrow$	v_3	$= \sin(v_0)$	$= \sin(5)$
	v_4	$= v_1 + v_2$	$= 0.6931 + 10$
	v_5	$= v_4 - v_3$	$= 10.6931 + 0.9589$
	y	$= v_5$	$= 11.6521$

	Trave de l'adjointe inverse		
	$\bar{x}_1$	$= \bar{v}_{-1}$	$= 5.5$
	$\bar{x}_2$	$= \bar{v}_0$	$= 1.7163$
	$\bar{v}_{-1}$	$= \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= 5.5$
	$\bar{v}_0$	$= \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= 1.7163$
$\Uparrow$	$\bar{v}_{-1}$	$= \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= 5$
	$\bar{v}_0$	$= \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= -0.2837$
	$\bar{v}_2$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= 1$
	$\bar{v}_1$	$= \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= 1$
	$\bar{v}_3$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= -1$
	$\bar{v}_4$	$= \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= 1$
	$\bar{v}_5$	$= \bar{y}$	$= 1$

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 3.

Question 4 (5-5)**.** Fondamentalement, la rétro-propagation est simplement un cas spécial de la différenciation automatique (DA) en mode inverse, appliqué à un réseau de neurone. En se basant sur la notation en "trois parties" présenté à la Table 1 et 2, représenté la trace d'évaluation et la trace de dérivée (adjointe) des exemples suivants. Dans les dernières colonnes de votre solution, évaluez numéricalement la valeur just qu'à 4 décimales près.

1. DA vers l'avant avec $y = f(x_1, x_2) = 1/(x_1 + x_2) + {x_2}^2 + \cos(x_1)$ à $(x_1, x_2) = (3, 6)$ et fixant $\dot{x}_1 = 1$ pour calculer $\partial y / \partial x_1$.
2. DA inverse avec $y = f(x_1, x_2) = 1/(x_1 + x_2) + {x_2}^2 + \cos(x_1)$ à $(x_1, x_2) = (3, 6)$. Fixant $\bar{y} = 1$, $\partial y / \partial x_1$ et $\partial y / \partial x_2$ peuvent être calculés ensemble.

Answer 4.

1. Some answers.

Question 5 (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 5. Full : $\big[,\big]$; Valid : $\big[,\big]$; Same : $\big[,\big]$.

Question 6 (5-5)**.** 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 6.

1.

Question 7 (4-4-6)**.** Supposons qu'on a des données de taille $3 \times 64 \times 64$. Dans ce qui suit, donnez une 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 $d = 1$ 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 $d = 7$, et que $s = 2y$.
2. La taille de la sortie de la deuxième couche est $(64, 8, 8)$. Supposons que $p = 0$ et que $d = 1$.
   (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 $d = 2$, et que $p = 2$.
   (c) Supposons que $p = 1$, et que $d = 1$.

Pour faciliter la correction, répondez en remplissant ce tableau :

Answer 7.

		i	p	d	k	s	o
1.	(a)	64		1	8		32
	(b)	64		7		2	32
2.	(a)	32	0	1			8
	(b)	32	0	1	8	4	
3.	(a)	8	0	1			4
	(b)	8	2	2			4
	(c)	8	1	1			4