IFT714: Traitement automatique des langues naturelles — Dû, le 17 février 2023, 23h59 heure de l'est

Devoir #3

Hiver 2023

Enseignant: Amine Trabelsi

- Ce devoir contient 3 problèmes. Les réponses aux problèmes doivent inclure les étapes de calcul intermédiaires qui vous ont amené à arriver à la réponse suggérée. Aucun crédit ne sera accordé aux réponses qui ne répondent pas à ce critère.
- L'utilisation de LaTeX est préférable mais pas obligatoire. Vous pouvez consulter un tutoriel LaTeX et ce cheatsheet. Dans la plupart des cas, si vous voulez écrire quelque chose en LaTeX, vous pouvez simplement utiliser le moteur de recherche Google "how to do {X} in latex" et les premiers liens devraient fournir la syntaxe que vous recherchez.
- Incluez votre nom et CIP avec votre soumission.
- Toutes les soumissions doivent être au format PDF suivant ce format: "prenom_nom_Devoir3_CIP".
- Si vous souhaitez soumettre des réponses manuscrites, vous pouvez les numériser et les soumettre au format PDF.
- Le devoir doit être remis sur le site du cours sur Moodle avant 23h59 heure de l'est à la date d'échéance.
- Selon la politique de jours de retard énoncée sur le plan de cours, un devoir soumis 24 heures après la date limite sera pénalisé de 3%. Un devoir soumis deux jours (24 à 48 heures) après la date limite sera pénalisé de 10%, et un devoir soumis trois jours sera pénalisé de 20%. Soumettre une livrable 72 heures (3 jours) après la date limite ne sera pas accepté.
- Les soumissions incomplètes ou en retard seront évaluées en tant que telles et aucun accommodement ou réévaluation ne seront faits.

---

1. (2 + 1 = 3 Points)

Considérez l'ensemble de données bidimensionnel suivant pour la classification binaire.

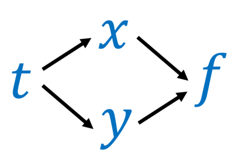
x_1	x_2	y
0	0	1
1	1	1
1	0	0

(a) Supposons que nous utilisions le perceptron pour la tâche de classification. Ici $z = \boldsymbol{w}^T\boldsymbol{x} + b$; $\boldsymbol{w} = (w_1, w_2)^T$;$\boldsymbol{x} = (x_1, x_2)^T$, et $y = \mathbb{1}\{z > 0\}$. Chaque exemple $(\boldsymbol{x}^{(i)}, y^{(i)})$, $i = 1, ..., 3$ donne une contrainte (inégalité) sur les poids w_1, w_2 et le biais b. Écrivez toutes les contraintes.

(b) Trouvez un ensemble de poids et un biais qui satisfassent strictement toutes les contraintes.

2. (1 + 2 + 2 = 5 Points)

Supposons une fonction $f(x, y)$ où $x = x(t)$ et $y = y(t)$, et f, x et $y \in \mathbb{R}$.

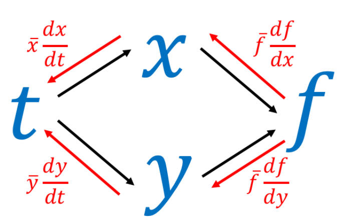


La règle de la chaîne multivariée pour calculer la dérivée de f par rapport à t est :

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t), y(t)) = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} = \bar{x}\frac{\mathrm{d}x}{\mathrm{d}t} + \bar{y}\frac{\mathrm{d}y}{\mathrm{d}t}$$

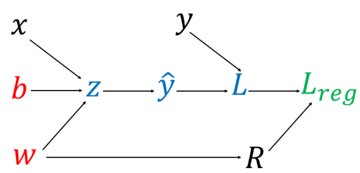
Le symbole de barre "-" indique la dérivée de la sortie f par rapport à chaque entrée utilisée pour la calculer. Une façon simple de visualiser l'opération est de schématiser le calcul à l'aide d'un graphe de calcul (computationnel) et de travailler à rebours (backward) en utilisant la notation en barres. La propagation complète est la suivante :

Passage en avant pour calculer f	Passage en arrière pour calculer $\frac{\mathrm{d}f}{\mathrm{d}t}$
t	$\bar{f} = \frac{\mathrm{d}f}{\mathrm{d}f} = 1$
$x = x(t)$	$\bar{x} = \bar{f}.\frac{\partial f}{\partial x}$
$y = y(t)$	$\bar{y} = \bar{f}.\frac{\partial f}{\partial y}$
$f = f(x, y)$	$\bar{t} = \bar{x}\frac{\mathrm{d}x}{\mathrm{d}t} + \bar{y}\frac{\mathrm{d}y}{\mathrm{d}t}$



(a) Soit $f(x, y) = x^2 + y^2, x(t) = t, y(t) = \mathrm{e}^t$. Calculez le $\frac{\mathrm{d}f}{\mathrm{d}t}$ en utilisant la passe arrière ci-dessus.

(b) Considérons la régression logistique de moindres carrés univariée régularisée en L_2 et le graphe de calcul correspondant ci-dessous.



Les nœuds représentent toutes les quantités d'entrée et de sortie et les arêtes représentent quels nœuds sont calculés directement en fonction de quels autres nœuds.

$y \in \{0, 1\}, y \sim Ber(\hat{y})$, où $\hat{y} = \sigma(z), z = wx + b$.

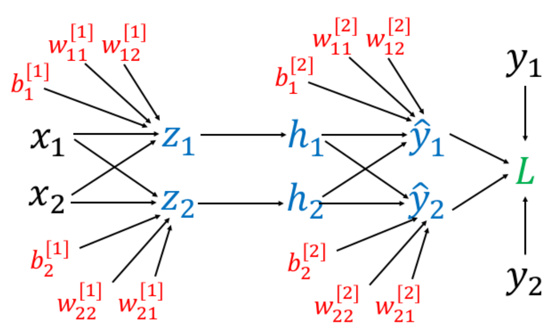
Pour un seul exemple, la perte non régularisée est $L = L(y, \hat{y}) = \frac{1}{2}(\hat{y} - y)^2$, et la perte régularisée est $L_{reg} = L + \lambda R$, où $R = \frac{1}{2}w^2$.

Écrivez les passes avant et arrière. Votre réponse devrait ressembler à ce qui suit :

Passage en avant pour calculer L_{reg}	Passage en arrière pour calculer le gradient de L_{reg}
$z =$	$\bar{L}_{reg} =$
$\hat{y} =$	$\bar{L} =$
$L =$	$\bar{R} =$
$R =$	$\bar{\hat{y}} =$
$L_{reg} =$	$\bar{z} =$
	$\bar{w} =$
	$\bar{b} =$

Notation : $\bar{u} = \frac{\mathrm{d}L_{reg}}{\mathrm{d}u}$, où $u \in \{L_{reg}, L, \hat{y}, z, w, b\}$.

(c) Considérez le perceptron multicouche (MLP) suivant avec des sorties multiples et des fonctions d'activation sigmoïdes.



Pour un seul exemple, la perte par moindres carrés est
$L = L(y,\hat{y}) = \frac{1}{2}\sum_{k=1}^{2}(\hat{y}_k - y_k)^2$;
$\hat{y}_k = \sum_{i=1}^{2} w_{ki}^{[2]} h_i + b_k^{[2]}$ pour $k = 1, 2$; $h_i = \sigma(z_i)$ pour $i = 1, 2$;
et $z_i = \sum_{j=1}^{2} w_{ij}^{[1]} x_j + b_i^{[1]}$ for $i = 1, 2$. Ecrivez la passe arrière. Votre réponse devrait ressembler à ce qui suit :

Passage en arrière pour calculer le gradient de L
$\bar{L} =$
$\bar{\hat{y}}_k =$
$\bar{w}_{ki}^{[2]} =$
$\bar{b}_k^{[2]} =$
$\bar{h}_i =$
$\bar{z}_i =$
$\bar{w}_{ij}^{[1]} =$
$\bar{b}_i^{[1]} =$

3. (1 + 2 + 3 + 1 = 7 Points)

Considérez le modèle de régression Softmax (régression logistique multi-classes). Ici, pour un vecteur d'attributs d'entrée donné, $\boldsymbol{x} = (x_1, ..., x_d)^T \in \mathbb{R}^d$, la tâche consiste à le classer dans l'une des K classes $C_1, ... C_K$. Soit y la variable de sortie dont les valeurs $y = 1, ..., K$ désignent les K étiquettes mutuellement exclusives, c'est-à-dire qu'il n'y a qu'une seule vraie étiquette.
Rappelez-vous (Devoir#2) que la probabilité conditionnelle de y, compte tenu de l'entrée x, peut être exprimée directement à l'aide de la fonction Softmax comme généralisation de la fonction sigmoïde.

$y \sim Multinoulli(\boldsymbol{s}(\boldsymbol{z}))$; $\boldsymbol{s}(\boldsymbol{z}) = (s_1(\boldsymbol{z}), ..., s_K(\boldsymbol{z}))^T$ où $s_k(\boldsymbol{z}) = p(y = k|\boldsymbol{z})$ avec $s_k(\boldsymbol{z}) = \frac{e^{z_k}}{\sum_{l=1}^{K} e^{z_l}}, k = 1, ..., K$.

$z_k = \boldsymbol{w}_k^T \boldsymbol{x} + b_k, k = 1, ..., K$ et $\boldsymbol{w}_k = (w_{k1}, ..., w_{kd})^T \in \mathbb{R}^d$, et $b_k \in \mathbb{R}$.

Rappelez-vous (Devoir#2) que pour un seul exemple $(\boldsymbol{x}, y)$ la vraisemblance est :

$$p(y|W, \boldsymbol{b}, \boldsymbol{x}) = \prod_{k=1}^{K} s_k(\boldsymbol{z})^{\mathbb{1}\{y=k\}}$$

,

où $\boldsymbol{z} = W\boldsymbol{x} + \boldsymbol{b}$, $W = \begin{pmatrix} \boldsymbol{w}_1^T \\ . \\ . \\ \boldsymbol{w}_K^T \end{pmatrix} = \begin{pmatrix} w_{11} & ... & w_{1d} \\ . & & . \\ . & & . \\ w_{K1} & ... & w_{Kd} \end{pmatrix} \in \mathbb{R}^{K \times d}$, et $\boldsymbol{b} = (b_1, ..., b_K)^T$.

(a) Au lieu de représenter la cible (y valeurs sous forme d'entiers (1,...,K), pour plus de commodité, nous utilisons un vecteur "one-hot", également appelé encodage 1 parmi K, $\boldsymbol{y} = (y_1, ..., y_K)^T$ où $y_k = \mathbb{1}\{y = k\}$, pour $k = 1, ..., K$. L'hypothèse d'exclusivité mutuelle implique qu'un seul des y_ks prend la valeur de 1, par exemple, $\boldsymbol{y} = (0, 0, 0, ..., 0, 0, 1)^T$. Par commodité, nous notons également la probabilité conditionnelle de classe prédite $s_k(\boldsymbol{z})$ par $\hat{y}_k$, et le vecteur de probabilité prédit pour toutes les classes par $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_K)^T$. Montrez que pour un exemple $(\boldsymbol{x}, y)$ la perte d'entropie croisée Softmax est :

$$L_{SCE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{k=1}^{K} y_k \log(\hat{y}_k) = -\boldsymbol{y}^T \log \hat{\boldsymbol{y}}$$

où *log* représente l'opération logarithmique par éléments.

(b) Montrez que le gradient de perte d'entropie croisée Softmax par rapport au vecteur d'entrée Softmax $\boldsymbol{z}$ est :

$$\nabla_{\boldsymbol{z}} L_{SCE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}} - \boldsymbol{y}$$

Astuce : vous pouvez supposer que seul le $k$ième élément de $\boldsymbol{y}$ est 1, ou de manière équivalente que k est la vraie classe (correcte).

(c) Pour un lot (batch) $D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$ de m exemples, montrez que le gradient de la perte d'entropie Softmax par rapport aux poids Softmax est donné par :

$$\frac{\partial L_{SCE}}{\partial w_{lj}} = \sum_{i=1}^{m} (\hat{y}_l^{(i)} - y_l^{(i)}) x_j^{(i)}$$

pour $l = 1, ..., K$ et $j = 1, ..., d$.

(d) Comparez le résultat précédent avec la forme du gradient obtenu sur la régression logistique binaire.