Introduction aux réseaux de neurones – exercices

Question 1. Considérons un problème de classification binaire où l'ensemble d'apprentissage $S \in \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ contient des couples $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$. Nous exprimons ainsi la fonction objectif à minimiser pour résoudre la régression logistique :

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} F(\mathbf{w}, b, \mathbf{x}_i, y_i)\,,$$

avec

$$\begin{aligned} F(\mathbf{w}, b, \mathbf{x}, y) &= -y(\mathbf{w} \cdot \mathbf{x} + b) + \log(1 + e^{\mathbf{w}\cdot\mathbf{x}+b}) + \frac{\rho}{2}\|\mathbf{w}\|^2 \\ &= L_{\text{nlv}}(\sigma(\mathbf{w} \cdot \mathbf{x} + b), y) + \frac{\rho}{2}\|\mathbf{w}\|^2 \end{aligned}$$

(a) Calculer $\dfrac{\partial L_{\text{nlv}}(\hat{y}, y)}{\partial \hat{y}}$, où $L_{\text{nlv}}(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$.

(b) Calculer $\dfrac{\partial \sigma(a)}{\partial a}$ où $\sigma(a) = \dfrac{1}{1 + e^{-a}}$.

(c) Pour $k \in \{1, \ldots, d\}$, calculer $\dfrac{\partial(\mathbf{w} \cdot \mathbf{x} + b)}{\partial w_k}$ où $\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$.

(d) Pour $k \in \{1, \ldots, d\}$, calculer $\dfrac{\partial \|\mathbf{w}\|^2}{\partial w_k}$.

Rappel : $\|\mathbf{w}\|^2 = \mathbf{w} \cdot \mathbf{w}$

(e) À partir des réponses aux questions précédentes, calculer $\dfrac{\partial F(\mathbf{w}, b, \mathbf{x}, y)}{\partial w_k}$.

Rappel de la règle de la dérivation en chaîne : $\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$.

(f) Déduire de la réponse précédente l'expression du gradient $\nabla_{\mathbf{w}} F(\mathbf{w}, b, \mathbf{x}, y)$.

(g) Calculez $\dfrac{\partial F(\mathbf{w}, b, \mathbf{x}, y)}{\partial b}$, c'est-à-dire le gradient du biais.

Question 2. Imaginons que nous optimisons la descente en gradient en initialisant tous les paramètres à zéro.

(a) Est-ce que cela peut poser problème dans le cas de la régression logistique ? Rappelons que la régression logistique peut être exprimée comme un réseau n'ayant aucune couche cachée.

(b) Est-ce que cela peut poser problème dans le cas d'un réseau de neurones à une couche cachée ?

(c) Dans le cas d'un réseau de neurones à une couche cachée, est-ce qu'utiliser le «dropout» sur la couche cachée peut résoudre un éventuel problème dû à l'initialisation des paramètres.

Question 3. Les fonctions d'activations utilisées pour les couches cachées d'un réseau de neurones sont rarement des fonctions linéaires $f(a) = a$. Illustrez la raison à l'aide d'un exemple.