Examen M2 AMIS
Réseaux de neurones

Jérémie Cabessa, 15 décembre 2022

Durée: 2h. Aucune documentation autorisée. Barème indicatif.

Exercice 1 (1 pt)

Expliquer brièvement le phénomène d'**overfitting**.

Exercice 2 (1 pt)

Décrire brièvement ce que représente le **dilemme biais-variance**.

Exercice 3 (1 pt)

Décrire brièvement ce qui se passe lorsqu'on entraîne un réseau de neurones par **descente de gradient**.

Exercice 4 (2 pts)

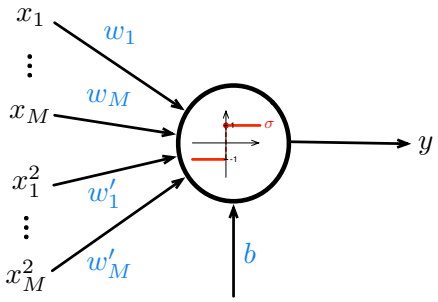
La solution d'une régression linéaire générale est donnée par: $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{arg\,min} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

1. Donner le développement qui permet de retrouver cette solution.
2. Calculer la solution $\hat{\boldsymbol{\beta}}$ associée au dataset suivant: $\left\{\left(\begin{bmatrix}0\\0\end{bmatrix}, 0\right), \left(\begin{bmatrix}0\\1\end{bmatrix}, 1\right), \left(\begin{bmatrix}1\\0\end{bmatrix}, 1\right), \left(\begin{bmatrix}1\\1\end{bmatrix}, 0\right)\right\}$.

Exercice 5 (2 pts)

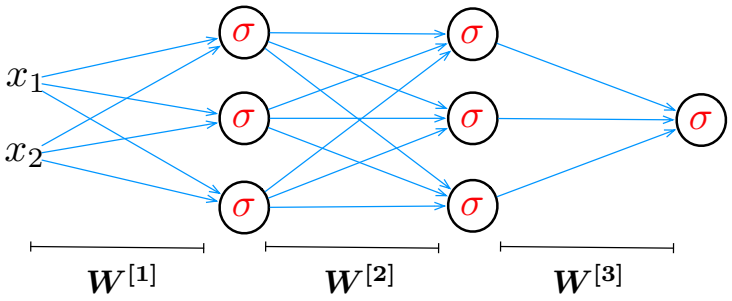
On considère un perceptron modifié comme représenté ci-dessous. Contrairement au perceptron classique, celui-ci prend en compte les carrés des inputs dans son calcul.

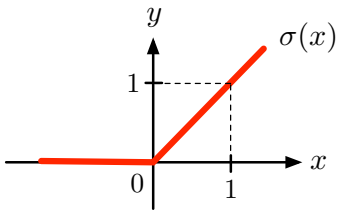
1. Donner une équation qui décrit la dynamique de ce perceptron.
2. **(Peut être difficile, ne perdez pas trop de temps...)** Donner des valeurs de w_1, w_2, w'_1, w'_2 et b qui classifient correctement les data du dataset suivant: $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix}, -1\right), \left(\begin{bmatrix}+1\\+1\end{bmatrix}, -1\right), \left(\begin{bmatrix}+1\\-1\end{bmatrix}, +1\right)\right\}$.



Exercice 6 (9 pts, bonus 2 pts)

On considère un réseau de neurones $\mathcal{N}$ dont l'architecture est illustrée ci-dessous. Les poids synaptiques (weights) sont désignés par $\mathbf{W}^{[1]}$, $\mathbf{W}^{[2]}$ et $\mathbf{W}^{[3]}$ et on suppose qu'il n'y a pas de biais (i.e., pas de vecteurs $\mathbf{b}^{[1]}, \mathbf{b}^{[2]}, \mathbf{b}^{[3]}$). La fonction d'activation σ de chaque neurone est la fonction ReLU définie par $\sigma(x) = \max\{0, x\}$, et également illustrée ci-dessous.





1. Donner les équations qui décrivent la dynamique ou "forward pass" de $\mathcal{N}$.

2. Soient $\mathbf{a}^{[0]} := \mathbf{x} \in \mathbb{R}^2$ un vecteur d'input et $\mathbf{a}^{[3]} \in \mathbb{R}$ l'output de $\mathcal{N}$ associée à $\mathbf{x}$. Donner une seule équation qui exprime $\mathbf{a}^{[3]}$ en fonction de $\mathbf{x}$, $\mathbf{W}^{[1]}$, $\mathbf{W}^{[2]}$ et $\mathbf{W}^{[3]}$

3. Donner le graphe computationnel associé au réseau $\mathcal{N}$, à savoir, une représentation graphique des opérations qui part de l'input $\mathbf{a}^{[0]} := \mathbf{x}$ pour aboutir à l'output $\mathbf{a}^{[3]}$ (comme dans le cours).

   Les variables $\mathbf{z}^{[k]}$, $\mathbf{a}^{[k]}$ et $\mathbf{W}^{[k]}$ seront respectivement représentées par des carrés $\boxed{\mathbf{z}^{[k]}}$, $\boxed{\mathbf{a}^{[k]}}$ et $\boxed{\mathbf{W}^{[k]}}$, les fonctions du type "$\mathbf{W}^{[k]}\mathbf{a}^{[k-1]}$" par un cercle $\textcircled{\times}$ et la fonction σ par un cercle $\textcircled{\sigma}$.

4. Supposons que les matrices de poids $\mathbf{W}^{[k]}$ soient uniquement constituées de valeurs k, pour $k = 1, 2, 3$ (par exemple, la matrice $\mathbf{W}^{[2]}$ est uniquement constituée de 2). Calculer l'output du réseau associé à l'input $\mathbf{x} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$

5. Soit la fonction de coût (loss function) définie par $\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, où $\hat{y} = \mathbf{a}^{[3]}$ est l'output du réseau et y est la réponse associée à l'input x, respectivement. Supposons également que les composantes de $\mathbf{z}^{[3]}$ sont toutes positives. Calculer le gradient de $\mathcal{L}$ par rapport aux poids de la troisième couche, i.e., $\nabla_{\mathbf{W}^{[3]}}\mathcal{L}(\hat{y}, y)$. Votre expression finale dépendra uniquement de $\mathbf{a}^{[3]}$, $\mathbf{a}^{[2]}$ et y.

6. Soient $\mathbf{x} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$, $y = 107$, $\hat{y} = \mathbf{a}^{[3]} = 108$, $\mathbf{a}^{[2]} = \left(\begin{smallmatrix}12\\12\\12\end{smallmatrix}\right)$, $\mathbf{W}^{[3]}$ uniquement constituée de valeurs 3 et $\lambda = 0.001$. Calculer la matrice des poids $\mathbf{W}^{[3]}$ obtenue après une itération de descente de gradient de learning rate λ sur la data $(\mathbf{x}, y)$.

   Si vous n'avez pas réussi la question précédente, supposez que $\nabla_{\mathbf{W}^{[3]}}\mathcal{L}(\hat{y}, y) = \begin{pmatrix}1296 & 1296 & 1296\end{pmatrix}$.

7. **(Bonus)** Calculer le gradient de $\mathcal{L}$ par rapport aux poids de la deuxième couche, i.e., $\nabla_{\mathbf{W}^{[2]}}\mathcal{L}(\hat{y}, y)$.