

TD 3

Exercice 1 – Apprentissage d’un réseau XOR

Le problème du XOR (“ou exclusif”) est le suivant : les points $(-1, -1), (1, 1)$ sont considérés négatifs, les points $(-1, 1), (1, -1)$ positifs.

Q 1.1 Dessiner un réseau de neurone à 2 neurones cachés pour ces données. Enumérer quelques fonctions d’activations possibles. Lesquelles sont les plus judicieuses dans ce cas précis pour les différentes couches ?

Q 1.2 Proposer des valeurs pour les poids du réseau. La solution est-elle unique ?

Q 1.3 Même question pour un échiquier à 8 cases.

Exercice 2 – Caractérisation de la solution apprise par un réseau de neurone

Considérons un réseau à une couche cachée paramétré par le vecteur \mathbf{w} . On note $f_{\mathbf{w}}(\mathbf{x})$ la sortie pour une entrée \mathbf{x} .

Nous utiliserons les notations suivantes :

- un échantillon $\mathbf{x}^i = \{x_j^i\}_{j=1,\dots,d}$, son étiquette y^i , un ensemble d’apprentissage $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1,\dots,N}$;
- les poids vers la couche cachée sont les $\mathbf{w}^1 = \{w_{jh}^1\}_{j=1,\dots,d, h=1,\dots,H}$, les poids vers la couche de sortie sont les $\mathbf{w}^s = \{w_{hk}^s\}_{h=1,\dots,H, k=1,\dots,K}$.
- les fonctions d’activation g^1, g^s des deux couches.

Q 2.1 Combien de neurones cachés compte le réseau ? De sorties ? Dessiner le réseau. A quoi correspond un nombre de sorties supérieur à un ?

Q 2.2 Exprimer la sortie $f_{\mathbf{w}}(\mathbf{x})$ en fonction des composantes de \mathbf{x} et \mathbf{w} .

Q 2.3 Donner l’expression du coût (moindres carrés) en fonction de la base d’apprentissage \mathcal{D} . Quelle est sa formulation théorique (en utilisant l’espérance d’une quantité) ?

Q 2.4 Montrer qu’en chaque \mathbf{x} , la solution optimale correspond à $f^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$. A quoi correspond ce résultat ?

Q 2.5 Pour la classification multiclass, la sortie utilisée est un vecteur : $\mathbf{y} = [\dots, 1, \dots]$ avec un 1 en k -ième position si la classe de \mathbf{x} est k . De quoi $f_k^*(\mathbf{x})$ est elle l’approximation dans ce cas là ?

Q 2.6 Dans le cas de la régression, à quoi correspond $f^*(\mathbf{x})$? Donner un exemple graphique de régression 1D bruité dans lequel plusieurs valeurs de y correspondent à un \mathbf{x} .

Q 2.7 Décomposition et interprétation du coût.

- Récrire le critère de coût en un point x pour faire intervenir les termes $y - f^*(\mathbf{x})$ et $f^*(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})$, puis $E_{y|\mathbf{x}} [\|y - f^*(\mathbf{x})\|^2]$.
- Donner une interprétation de la signification de ce terme ainsi que des autres termes que vous avez fait apparaître. Pourquoi l’apprentissage ne permet pas toujours d’obtenir un coût nul ?

Q 2.8 La solution obtenue par descente de gradient est-elle unique ? Pourquoi ? De quoi dépend elle ?

Exercice 3 (8 points) – Highway to gradient

L’architecture Highway Network a été proposé en 2015 spécifiquement pour les réseaux très profonds dédiés au traitement de l’image. Une couche de ce type de réseau est très semblable à celle d’un réseau

classique mais réalise un mélange des entrées de la couche avec les sorties de cette couche.

Prenons par exemple une couche fully-connected non linéaire d'un réseau classique définie par la fonction $H(\mathbf{x}, \mathbf{W}_H) = \sigma(\mathbf{W}_H^t \mathbf{x} + b_H)$. Le Highway Network utilise une transformation $T(\mathbf{x}, \mathbf{W}_T) = \sigma(\mathbf{W}_T^t \mathbf{x} + b_T)$ afin de mélanger l'entrée \mathbf{x} et la sortie usuelle de la couche $H(\mathbf{x}, \mathbf{W}_H)$: la sortie \mathbf{y} de la couche est (avec \odot l'opérateur de produit terme à terme)

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$$

Q 3.1 Supposons que l'entrée \mathbf{x} soit de dimension d : $\mathbf{x} \in \mathbb{R}^d$. D'après la définition, quelles sont les dimensions de \mathbf{W}_H , \mathbf{W}_T , \mathbf{y} ? (on supposera les biais b_H et b_T scalaires dans \mathbb{R}).

Q 3.2 Calculez la dérivée de la fonction sigmoïde. Dans la suite, vous pouvez la noter $\sigma'(x)$ sans développer. Rappel : $\sigma(x) = \frac{1}{1+e^{-x}}$

Q 3.3 On suppose un réseau constitué en premier d'une d'une couche Highway Network dont on notera la sortie \mathbf{z} , puis d'une couche linéaire M avec une fonction d'activation sigmoïde, et un coût aux moindres carrés :

- $\mathbf{z} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$
- $\hat{\mathbf{y}} = \sigma(\mathbf{W}_M^t \mathbf{z} + b_M)$
- $L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2$

Q 3.3.1 On suppose que $\mathbf{y} \in \mathbb{R}^p$. Quelles doivent être les dimensions de \mathbf{W}_M et $\hat{\mathbf{y}}$? (on suppose que le biais b_M est scalaire dans \mathbb{R}).

Q 3.3.2 Calculez $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}_i}$ la dérivée du coût par rapport à la i -ème sortie du réseau.

Q 3.3.3 Calculez $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^M}$ pour un poids $w_{i,j}^M$ de \mathbf{W}_M .

Q 3.3.4 Calculez $\delta_i = \frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial z_i}$ pour la i -ème sortie \mathbf{z} de la couche Highway.

Q 3.3.5 Calculez pour un poids $w_{i,j}^T$ $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^T}$

Q 3.3.6 Calculez pour un poids $w_{i,j}^H$ $\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial w_{i,j}^H}$

Q 3.3.7 Donnez l'algorithme d'optimisation du réseau.

Q 3.4 (bonus) À votre avis, quel(s) problème(s) permet de résoudre un réseau Highway et pourquoi ?