TD 6

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 $\mathbf{x}^i = \{x^i_j\}_{j=1,\ldots,d}$, son étiquette y^i, un ensemble d'apprentissage $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1,\ldots,N}$;
- les poids vers la couche cachée sont les $\mathbf{w}^1 = \{w^1_{jh}\}_{j=1,\ldots,d,\ h=1,\ldots,H}$, les poids vers la couche de sortie sont les $\mathbf{w}^s = \{w^s_{hk}\}_{h=1,\ldots,H,\ k=1,\ldots,K}$.
- 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 $f^\star(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$. A quoi correspond ce résultat ?

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

Q 2.6 Dans le cas de la régression, à quoi correspond $f^\star(\mathbf{x})$? 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 $y - f^\star(\mathbf{x})$ et $f^\star(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})$, puis $E_{y|\mathbf{x}}\left[\|y - f^\star(\mathbf{x})\|^2\right]$.
- 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 – Rétro-propagation

On se place dans le cadre général d'un réseau de neurones multi-couches (sans fixer le nombre de couches). Soit $E(\mathbf{w})$ l'erreur aux moindres carrés pour un réseau avec les poids $\mathbf{w}$.
On considère les notations suivantes : l'activation du neurone a^k_i, la sortie au neurone i de la couche k : $z^k_i = g(a^k_i)$, et le poids d'un neurone i de la couche $k-1$ vers un neurone j de la couche k : $w^k_{i,j}$.

On notera δ_i^k l'erreur associée au neurone correspondant, $g(x)$ la fonction d'activation.

Q 3.1 Dessiner un bout du réseau.

Q 3.2 Rappeler l'expression de la i-ème sortie du PMC $f_{\mathbf{w}}(\mathbf{x})_i$ en fonction des entrées (les composantes de $\mathbf{x}$) dans le cas d'un réseau à une couche cachée.

Q 3.3 En déduire, dans le cas d'un réseau général, l'expression d'un z_i^h en fonction des z_j^{h-1} (sorties de la couche précédente).

Q 3.4 Backward sur la couche de sortie
Pour appliquer l'algorithme du gradient, nous avons besoin de calculer le gradient du critère par rapport à tous les poids w_{ij}^k du réseau. Le calcul s'effectue par couche, de la dernière vers la première en répercutant les erreurs des différentes couches. Dans un premier temps, nous allons étudier $\frac{\partial E(\mathbf{w})}{\partial w_{ij}^s}$, les dérivées par rapport aux poids de la dernière couche cachée.

Q 3.4.1 Montrer que ces dérivées peuvent s'exprimer à l'aide de $\delta_j^s = \frac{\partial E(\mathbf{w})}{\partial a_j^s}$ et $\frac{\partial a_j^s}{\partial w_{ij}^s}$.

Q 3.4.2 Calculer les δ_j^s et les dérivées partielles de a_j^s par rapport à w_{ij}^s.

Q 3.4.3 En déduire la dérivée du coût par rapport à un poids w_{ij}^s.

Q 3.5 On considère maintenant le cas de cellules sur une couche cachée h. On note δ_i^h la dérivée du critère par rapport à l'activation d'une cellule d'une couche cachée a_i^h. Exprimer cette dérivée en fonction des δ_j^{h+1} correspondant aux cellules de la couche suivante et en fonction de quantités $\frac{\partial a_j^{h+1}}{\partial a_i^h}$.

Q 3.6 Que vaut $\frac{\partial a_j^{h+1}}{\partial a_i^h}$?

Q 3.7 En déduire la dérivée du coût par rapport à un poids w_{ji}^h d'une connexion d'un neurone j de la couche $h-1$ vers un neurone i de la couche h.

Q 3.8 Résumer l'algorithme de backward propagation.