UE 4I702, TD2, Réseaux de neurones “feed-forward” et l’apprentissage supervisé

EXERCICE I. PERCEPTRON (RÉSEAU FEED-FORWARD À UNE COUCHE DE POIDS)

Considérons un réseau de neurones qui consiste à un neurone de sortie avec son taux de décharge y, et D neurones d’entrée avec ses taux de décharge $\vec{x} = \left(x_1, \ldots, x_D\right)$. Le neurone d’entrée avec l’index i est connecté au neurone de sortie avec un poids synaptique w_i. Tous les poids forment le vecteur des poids $\vec{w} = (w_1, \ldots, w_D)$. La fonction d’activation est sigmoïde, $g(a) = (1 + e^{-a})^{-1}$, où $a = w_0 + \sum_{i=1}^{N} w_i x_i = w_0 + \vec{w} \cdot \vec{x}$ est l’activation du neurone de sortie. Ici, le paramètre w_0 correspond au poids synaptique d’une entrée fixe $x_0 = 1$ (appelé “le biais”).

On interprète ce réseau de neurones comme une machine de classification de données:

- considérons un jeu de données dans D dimensions, c’est-à-dire un ensemble de vecteurs (souvent appelés “motifs”) $\left\{\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(N)}\right\}$, où chacun des motifs $\vec{x}^{(p)} = (x_1^{(p)}, x_2^{(p)}, \ldots, x_D^{(p)})$, $p = 1, 2, \ldots N$. Géométriquement, un motif correspond à un point dans D dimensions.
- on choisie un motif $\vec{x}^{(p)}$, on l’introduit au réseau et on calcule la valeur de sortie $y(\vec{x}^{(p)})$
- si $y \leq 0.5$, on classifie le motif comme appartenant à la classe A. Dans le cas contraire on assigne le motif à la classe B.
- application du réseau à tous les motifs donne lieu à une séparation de l’ensemble des données en deux classes.

A. Réseau dans une dimension ($D = 1$).
   1. Exprimer y en fonction de x_1 (avec w_0 et w_1 comme paramètres de cette fonction).
   2. Trouver la valeur de x_1 qui correspond à la frontière entre deux classes (point d’inflexion de la sigmoïde).
   3. Trouver la pente de $y(x_1)$ au point d’inflexion.
   4. Pour deux motifs arbitraires (par ex. $x_1^{(1)} = 2$ et $x_1^{(2)} = 6$) proposer les poids synaptiques w_0 et w_1 pour que le réseau classifie $x_1^{(1)}$ à la classe A et $x_1^{(2)}$ à la classe B.
   5. Tracer la courbe $y(x_1)$ correspondant à w_0 et w_1 de l’étape précédente.

B. Réseau dans deux dimensions ($D = 2$).
   1. Exprimer y en fonction de $\vec{x}$ et $\vec{w}$.
   2. La frontière entre les classes A et B vérifie la condition $y = 0.5$ (Expliquer pourquoi). Trouver l’équation de cette frontière en termes de x_1 et x_2. Conclusion ?
   3. Démontrer que le vecteur des poids $\vec{w}$ est perpendiculaire à la frontière
   4. Démontrer que la distance perpendiculaire entre la frontière et l’origine est égale à $-\frac{w_0}{\|\vec{w}\|}$
   5. On choisie $w_0 = -1$ et $\vec{w} = (1,1)$. Sur le plan (x_1, x_2) représenter le vecteur de poids $\vec{w}$ et la frontière entre les classes. Que se passe-t-il si l’on change la direction de $\vec{w}$, en gardant la norme constant? Que se passe-t-il si l’on change w_0, en gardant $\vec{w}$ fixe ? Quelle est la forme de la fonction $y(x_1, x_2)$?
   6. Comment les résultats si-dessus généralisent-ils à la dimension $D > 2$?

EXERCICE II. APPRENTISSAGE : RÈGLE DELTA

On considère le réseau de neurones de l'exercice I. L'erreur totale de classification pour l'ensemble des données $\{\vec{x}^{(1)},\vec{x}^{(2)},\ldots,\vec{x}^{(N)}\}$ peut être calculée selon la formule suivante : $E(\vec{w})=\sum_{p=1}^{N}E^{(p)}=\sum_{p=1}^{N}\frac{1}{2}(y^{(p)}-t^{(p)})^2$ où :

- $E^{(p)}$ est l'erreur de classification d'un seul motif $\vec{x}^{(p)}$.
- $t^{(p)}$ est l'étiquette du motif $\vec{x}^{(p)}$; $t^{(p)}=0$ si $\vec{x}^{(p)}$ appartient à la classe A, et $t^{(p)}=1$ si $\vec{x}^{(p)}$ appartient à la classe B.

L'erreur totale est une fonction des poids $\vec{w}=(w_1,\ldots,w_D)$. Classification correcte de tous les motifs correspond au minimum de l'erreur ($E(\vec{w})=0$), i.e. quand la réponse du réseau $y^{(p)}$ est égale à l'étiquette de la classe $t^{(p)}$ pour tous les motifs.

Apprentissage du réseau correspond à la minimisation de la fonction d'erreur dans l'espace des poids $(w_1,\ldots,w_D)$ par la méthode de descente de gradient. Selon cette méthode, le changement d'un poids synaptique w_i après la présentation d'un motif $\vec{x}^{(p)}$, est proportionnel à la dérivée négative de E:

$$\Delta w_i=-\eta\frac{\partial E^{(p)}}{\partial w_i}$$

A. Démontrer que

$\Delta w_i=-\eta\delta x_i^{(p)}$ ("delta-rule")

où

$\delta=g'(a)(y^{(p)}-t^{(p)})$ ("delta-error")

B. Démontrer que

$$g'(a)=g(a)(1-g(a))$$

EXERCICE III. APPRENTISSAGE : ALGORITHME DE RÉTROPROPAGATION DU GRADIENT

On considère un réseau avec un neurone de sortie et deux couches de poids, dans D dimensions (voir les diapos du cours). La sortie du réseau :

$$y=\sum_{j=0}^{H}w_j^{(2)}z_j$$

où $w_j^{(2)}$ est le poids synaptique entre le neurone de sortie et le j ème neurone cachés, dont le taux de décharge est égal à (sauf le neurone biais pour lequel $z_0=1$):

$$z_j=g\left(\sum_{i=1}^{D}w_{ji}^{(1)}x_i\right)$$

où $w_{ji}^{(1)}$ est le poids entre j ème neurone caché et i ème neurone d'entrée.

Démontrer que la méthode de descente de gradient de l'exercice précédent correspond aux formules suivantes pour la mise à jour des poids :

$$\Delta w_j^{(2)} = -\eta \delta z_j \quad \text{où} \quad \delta = y^{(p)} - t^{(p)}$$

et

$$\Delta w_{ji}^{(1)} = -\eta \delta_j x_i \quad \text{où} \quad \delta_j = g'(a_j) w_j^{(2)} \delta$$

EXERCICE IV. PLUSIEURS NEURONES DE SORTIE

On ajoute au réseau de l'exercice I.B deux neurones de sortie, on a alors

$$y_k = g(a_k),\ k = 0,1,2$$

$$a_k = w_{k0} + \sum_{i=1}^{D} w_{ki} x_i\ ,\ D = 2\,.$$

Dans ce réseau, chaque neurone de sortie encode l'appartenance à une classe : y_0 code pour la classe A, y_1 code pour la classe B, et y_2 code pour la classe C (on a toujours autant de classes que de neurones de sortie).

Apprentissage : on introduit un motif $\vec{x}^{(p)}$, calcule le vecteur de sortie $\vec{y} = (y_0, y_1, y_2)$ et applique la méthode de descente de gradient avec les étiquettes de la forme suivante: $\vec{t} = (1,0,0)$ si $\vec{x}^{(p)}$ appartient à la classe A, $\vec{t} = (0,1,0)$ si $\vec{x}^{(p)} \in B$, et $\vec{t} = (0,0,1)$ si $\vec{x}^{(p)} \in C$.

Si l'on considère chaque neurone de sortie comme un sous-réseau (identique au réseau de l'exercice I.B), on voit que k ème neurone de sortie est entraîné pour distinguer les motifs de sa classe ($y_k = 1$) de toute les autres motifs ($y_k = 0$).

Classification : pour classifier un motif $\vec{x}$, on l'introduit au réseau, on calcule $\vec{y} = (y_0, y_1, y_2)$. On dit que le motif appartient à la classe avec l'activité maximale, par ex. à la classe B si $y_1 > y_0$ et $y_1 > y_2$.

A. Exprimer y_0, y_1 et y_2 en fonction de x_1 et x_2.

B. Quelle condition est vérifiée sur la frontière entre les classes A et B ? Quelle est l'équation de cette frontière en termes de x_1 et x_2 ?

C. La même question pour la frontière entre B et C et entre A et C.

D. Comment les résultats si-dessus généralisent-ils à la dimension $D > 2$?

E. Démontrer que la règle d'apprentissage a la forme suivante :

$$\Delta w_{ki} = -\eta \delta_k x_i^{(p)} \qquad \text{("delta-rule")}$$
$$\delta_k = g'(a_k)(y_k^{(p)} - t_k^{(p)}) \qquad \text{("delta-error")}$$