Introduction à l'apprentissage machine (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2018

EXAMEN PARTIEL

Instructions : – Une feuille aide-mémoire recto verso manuscrite est permise ;
– Durée de l'examen : 2 h 50.

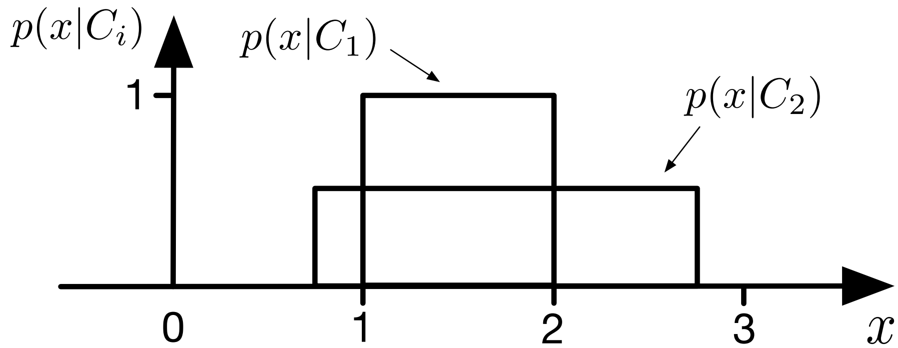
Pondération : Cet examen compte pour 30% de la note finale.

Question 1 (15 points sur 100)

Soit un problème de classement à deux classes et en une dimension, dont les vraisemblances de classe sont décrites par les densités de probabilité suivantes :

$$p(x|C_1) = \begin{cases} 1 & x \in [1, 2[\\ 0 & \text{autrement} \end{cases}, \qquad p(x|C_2) = \begin{cases} 0{,}5 & x \in [0{,}75, 2{,}75[\\ 0 & \text{autrement} \end{cases},$$

qui sont représentées dans ce qui suit.



(5) (a) Supposons que les probabilités a priori sont égales ($P(C_1) = P(C_2) = 0{,}5$), donnez la règle de décision à utiliser selon cette modélisation pour assigner une donnée x à la classe C_1 ou C_2.

(5) (b) Toujours en supposant des probabilités a priori égales, donnez les équations des probabilités a posteriori ($P(C_i|x)$) pour les deux classes.

(5) (c) Supposons maintenant que les probabilités a priori sont différentes, avec $P(C_1) = 0{,}25$ et $P(C_2) = 0{,}75$. Donnez les probabilités a posteriori pour les deux classes et la règle de décision associée.

Question 2 (25 points sur 100)

Soit un réseau de neurones de type RBF pour deux classes, composé d'une couche cachée de R neurones de type gaussien, suivi d'une couche de sortie d'un neurone avec fonction d'activation linéaire. La valeur de la sortie pour un tel réseau de neurones pour une valeur d'entrée $\mathbf{x}$ est donnée par l'équation suivante,

$$\mathrm{h}(\mathbf{x}) = \sum_{i=1}^{R} w_i \phi_i(\mathbf{x}) + w_0 = \sum_{i=1}^{R} w_i \exp\left[-\frac{\|\mathbf{x}-\mathbf{m}_i\|^2}{2s_i^2}\right] + w_0,$$

où :

— $\mathbf{m}_i$ est la valeur du centre du i-ème neurone gaussien de la couche cachée ;
— s_i est l'étalement du i-ème neurone gaussien ;
— w_i est le poids connectant le i-ème neurone gaussien de la couche cachée au neurone de sortie ;
— w_0 est le biais du neurone de sortie.

Supposons que l'on fixe les étalements s_i à des valeurs prédéterminées et que l'on veut apprendre les valeurs w_i, w_0 et $\mathbf{m}_i$ par descente du gradient, en utilisant comme critère l'erreur quadratique moyenne,

$$E = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} [r^t - \mathrm{h}(\mathbf{x}^t)]^2,$$

où :

— $r^t \in \mathbb{R}$ est la valeur désirée pour le neurone de sortie du réseau ;
— $\mathcal{X}$ est l'ensemble des N données d'entraînement.

(13) (a) Développez les équations permettant de **mettre à jour les poids** w_i et w_0 du neurone de sortie par descente du gradient, avec un taux d'apprentissage η, en utilisant le critère de l'erreur quadratique moyenne.

(12) (b) Développez les équations permettant de **mettre à jour les valeurs des centres** $\mathbf{m}_i$ des neurones gaussiens de la couche cachée par descente du gradient, en utilisant le critère de l'erreur quadratique moyenne.

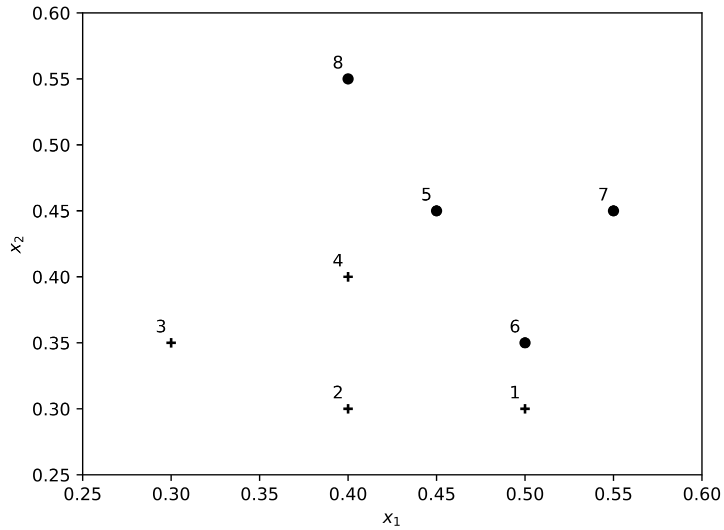
Question 3 (32 points sur 100)

Soit le jeu de données suivant, en deux dimensions :

$$\begin{array}{llll}
\mathbf{x}^1 = [0{,}5\ 0{,}3]^\top, & \mathbf{x}^2 = [0{,}4\ 0{,}3]^\top, & \mathbf{x}^3 = [0{,}3\ 0{,}35]^\top, & \mathbf{x}^4 = [0{,}4\ 0{,}4]^\top, \\
\mathbf{x}^5 = [0{,}45\ 0{,}45]^\top, & \mathbf{x}^6 = [0{,}5\ 0{,}35]^\top, & \mathbf{x}^7 = [0{,}55\ 0{,}45]^\top, & \mathbf{x}^8 = [0{,}4\ 0{,}55]^\top.
\end{array}$$

Les étiquettes de ces données sont $r^1 = r^2 = r^3 = r^4 = -1$ et $r^5 = r^6 = r^7 = r^8 = 1$.

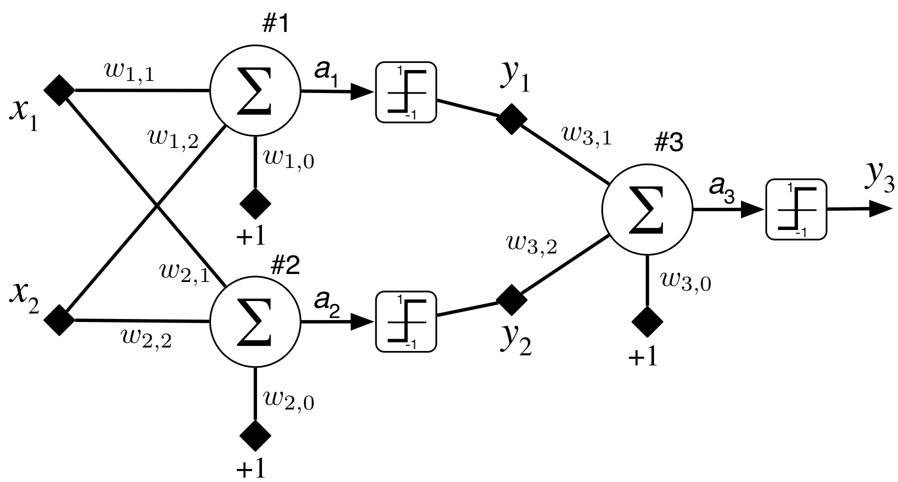
Le graphique ici-bas présente le tracé de ces données.



(10) (a) Soit un réseau de neurones utilisant la fonction signe comme fonction d'activation :

$$\mathrm{f}_{\mathrm{sgn}}(a) = \mathrm{sgn}(a) = \begin{cases} 1 & a \geq 0 \\ -1 & a < 0 \end{cases}.$$

Supposons que l'on veut entraîner le réseau suivant à trois neurones avec cette fonction d'activation signe, pour classer les données présentées ci-haut, selon le schéma suivant.



Supposons les valeurs de poids $w_{3,1} = 1$ et $w_{3,2} = 1$, et du biais $w_{3,0} = -1{,}5$. Déterminez les poids et biais des neurones 1 et 2 afin d'obtenir un taux de classement correct de $100\,\%$ sur les données.

(10) (b) Tracez les régions de décision selon les données d'entraînement présentée en préambule de la question pour un classifieur de type plus proche voisin (avec un seul voisin, $k = 1$) utilisant une distance euclidienne. Tracez le tout dans la **feuille de réponse fournie** (et non dans l'énoncé courant). Donnez également le taux de classement selon une méthodologie *leave-one-out* avec cette configuration, sur ces données.

(12) (c) Nous obtenons le résultat suivant en effectuant l'entraînement d'un SVM linéaire à **marge douce** avec ces données, avec comme valeur de paramètre de régularisation $C = 200$:

$\alpha^1 = 180, \quad \alpha^2 = 0, \quad \alpha^3 = 0, \quad \alpha^4 = 200, \quad \alpha^5 = 180, \quad \alpha^6 = 200, \quad \alpha^7 = 0, \quad \alpha^8 = 0,$
$w_0 = -11{,}6.$

En utilisant la feuille de réponse fournie, tracez les éléments suivants :
— Frontière de décision du SVM (droite continue, —-) ;
— Frontières de la marge (droites en traits pointillés, - - - -) ;
— Vecteurs de support (encerclez les points, ◯) ;
— Données dans la marge (encadrez les points, □).

Question 4 (28 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(4) (a) Expliquez en quoi consiste la régularisation dans un contexte d'apprentissage supervisé.

(4) (b) Dans un contexte de classement paramétrique s'appuyant sur des modèles de densité de probabilité des données selon les classes, la règle de Bayes est souvent utilisée de la façon suivante :

$$\underbrace{P(C|\mathbf{x})}_{\text{a posteriori}} = \frac{\overbrace{P(C)}^{\text{a priori}}\,\overbrace{p(\mathbf{x}|C)}^{\text{vraisemblance}}}{\underbrace{p(\mathbf{x})}_{\text{évidence}}}.$$

Expliquez pourquoi l'évidence $p(\mathbf{x})$ de cette formulation est généralement ignorée lorsqu'il vient le moment de définir la règle de décision (fonction $\mathrm{h}_i(\mathbf{x})$).

(4) (c) Supposons un classement paramétrique basé sur des lois normales multivariées, où l'estimation de la matrice de covariance $\mathbf{S}$ est partagée (la même) entre toutes les classes. Indiquez l'équation pour calculer cette estimation de la matrice de covariance partagée à partir d'un jeu de données étiquetées pour le classement.

(4) (d) Dans un contexte de classement par les k-plus proches voisins, expliquez l'effet du nombre de voisins k sur le classement et les frontières de décision.

(4) (e) Expliquez le lien principal que l'on peut établir entre la régression logistique et le classement paramétrique.

(4) (f) Dans un contexte de classement avec SVM à marges douces, expliquez comment on interprète les variables *slacks* ξ^t.

(4) (g) Expliquez la différence entre un apprentissage par lots (*batch*) et un apprentissage en ligne dans un réseau de neurones.