Introduction à l'apprentissage automatique (GIF-7005)
Département de génie électrique et de génie informatique
Automne 2022

UNIVERSITÉ LAVAL

EXAMEN

Instructions : – Identifiez-vous bien sur la page titre ;
– Répondez directement dans le questionnaire fourni ;
– Une feuille aide-mémoire recto verso manuscrite est permise ;
– Durée de l'examen : 1 h 50.

Pondération : Cet examen compte pour 20% de la note finale.

Prénom : _____

Nom : _____

NI : _____

Signature : _____

GIF-7005

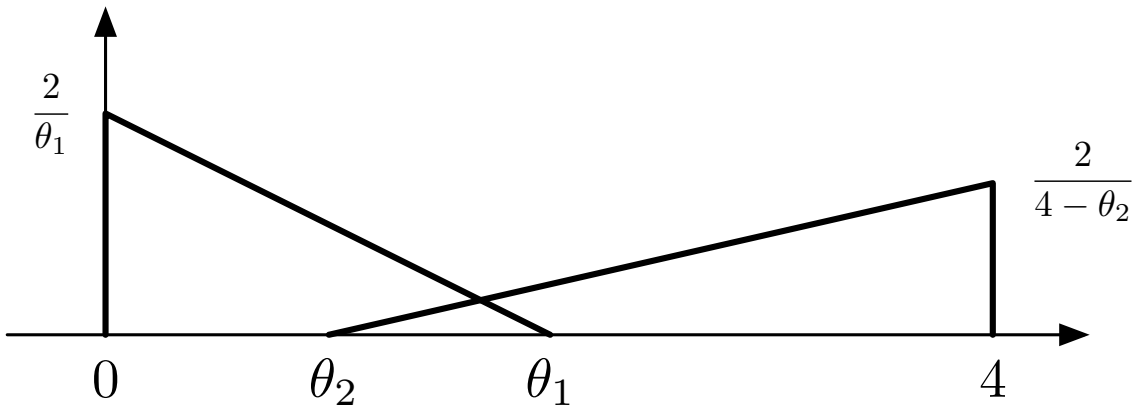
Question 1 (24 points sur 100)

Soit un système de classement paramétrique à deux classes et comportant une variable en entrée. La modélisation des distributions pour chaque classe est donnée par les équations suivantes :

$$p(x|C_1) = \begin{cases} \frac{-2\,(x-\theta_1)}{(\theta_1)^2} & \text{si } x \in [0, \theta_1] \\ 0 & \text{autrement} \end{cases},$$

$$p(x|C_2) = \begin{cases} \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} & \text{si } x \in [\theta_2, 4] \\ 0 & \text{autrement} \end{cases}.$$

Ainsi, la paramétrisation de la distribution de la classe C_1 est donnée par θ_1, alors que celle de la classe C_2 est donnée par θ_2. On fait également l'hypothèse que $0 \leq \theta_2 \leq \theta_1 \leq 4$. La figure suivante présente le tracé de ces distributions de classes.



(8 pts) (a) Supposons que $\theta_1 = 3$ et $\theta_2 = 2$, donnez la fonction $\mathrm{h}(x)$ correspondant à la prise de décision pour le classement de données selon la valeur de $x \in [0, 4]$. Supposez que les probabilités *a priori* des classes sont égales, soit $P(C_1) = P(C_2) = 0{,}5$. Supposez également une perte égale pour les différents types d'erreurs. Donnez les développements menant à votre fonction de décision.

Solution: La décision se prend selon la valeur maximale des probabilités *a posteriori* de classement, soit :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1, C_2\}}{\operatorname{argmax}} P(C_i|x).$$

Comme les évidences $p(x)$ et les probabilités *a priori* sont les mêmes pour les deux classes, la décision peut se prendre directement à partir des vraisemblances de classe, soit :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1, C_2\}}{\operatorname{argmax}} p(x|C_i).$$

Comme les vraisemblances de classe sont des fonctions linéaires, il suffit de déterminer le point où les deux distributions sont égales dans l'intervalle $[\theta_2, \theta_1]$:

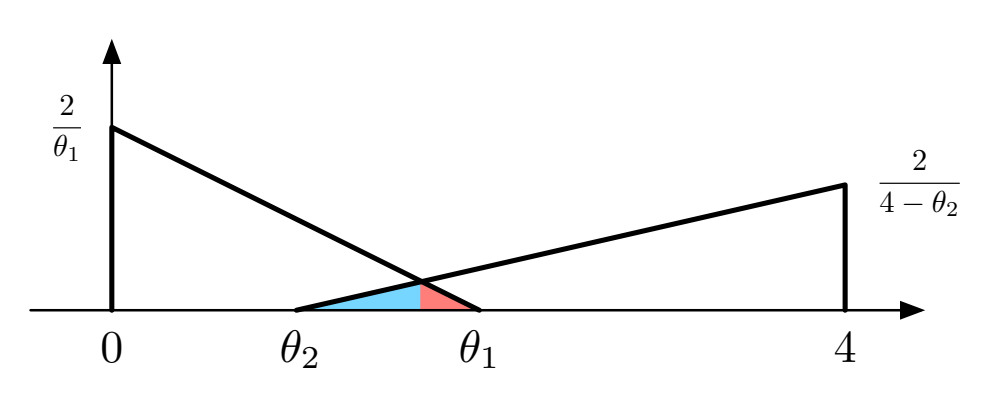
$$\begin{aligned}
p(x|C_1) &= p(x|C_2), \\
\frac{-2\,(x-\theta_1)}{(\theta_1)^2} &= \frac{2\,(x-\theta_2)}{(4-\theta_2)^2}, \\
\frac{-2\,(x-3)}{(3)^2} &= \frac{2\,(x-2)}{(4-2)^2}, \\
\frac{-x+3}{9} &= \frac{x-2}{4}, \\
(0{,}111 + 0{,}25)x &= (0{,}333 + 0{,}5), \\
x &= 2{,}3077.
\end{aligned}$$

Donc, en se basant sur la figure de l'énoncé de la question, on obtient comme fonction de prise de décision ceci :

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 2{,}3077] \\ C_2 & \text{pour } x \in [2{,}3077, 4] \end{cases}.$$

(8 pts) (b) Calculez le taux d'erreur bayésien optimal que l'on obtient avec le classifieur calculé au point précédent. Le taux d'erreur bayésien optimal correspond au taux d'erreur obtenu lorsque les données classées suivent parfaitement les distributions estimées pour le classement.

Solution: Des erreurs surviennent lorsqu'une donnée de la classe C_2 a une valeur $x < 2{,}3077$ ou qu'une donnée de la classe C_1 a une valeur $x > 2{,}3077$. La figure suivante présente les distributions données selon les classes, avec en rouge et en bleu les régions des distributions où un classement selon ces distributions résulte en une erreur.



Donc, pour estimer l'erreur de classement dans ce cas, il faut calculer l'aire des distributions où une donnée sera mal classée. Dans le cas de la classe C_1, l'erreur correspond au triangle rouge. Il faut d'abord calculer la hauteur de ce triangle :

$$H_1 = p(x = 2{,}3077|C_1) = \frac{-2\,(x-\theta_1)}{(\theta_1)^2} = \frac{-2\,(2{,}3077-3)}{3^2} = 0{,}15384.$$

Ensuite, la longueur du triangle est calculée comme étant

$$L_1 = \theta_1 - 2{,}3077 = 3 - 2{,}3077 = 0{,}6923.$$

L'aire d'un triangle rectangle se calcule ensuite comme étant le produit de la longueur et de la hauteur du triangle divisé par deux :

$$A_1 = \frac{H_1 \times L_1}{2} = \frac{0{,}15384 \times 0{,}6923}{2} = 0{,}053252.$$

Similairement, pour la classe C_2 l'erreur correspond au triangle bleu de la figure et est calculée comme suit :

$$\begin{aligned} H_2 &= p(x = 2{,}3077|C_2) = \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} = \frac{2\,(2{,}3077-2)}{(4-2)^2} = 0{,}15384, \\ L_2 &= 2{,}3077 - \theta_2 = 2{,}3077 - 2 = 0{,}3077, \\ A_2 &= \frac{H_2 \times L_2}{2} = \frac{0{,}15384 \times 0{,}3077}{2} = 0{,}023668. \end{aligned}$$

Donc, l'erreur totale est égale à la somme des deux aires multiplées par leurs probabilités a priori respectives, soit :

$$E = P(C_1)\,A_1 + P(C_2)\,A_2 = 0{,}5 \times 0{,}053252 + 0{,}5 \times 0{,}023668 = 0{,}03846.$$

Le taux d'erreur bayésien optimal est donc de 3,85 %.

(8 pts) (c) Supposons maintenant que la fonction de perte est variable selon le type d'erreur que fait notre classifieur. Plus précisément, si une donnée est classée comme étant dans la classe C_2 mais appartient en fait à la classe C_1, la perte est de $\mathcal{L}(\alpha_2,C_1) = 1$, alors que la perte pour une donnée classée comme étant de la classe C_1, mais appartenant en fait à la classe C_2 est de $\mathcal{L}(\alpha_1,C_2) = 0{,}5$. Calculez la nouvelle fonction $\mathrm{h}(x)$ correspondant à la prise de décision pour le classement de données selon cette fonction de perte dans le domaine $x \in [0, 4]$. Supposez que les autres paramètres sont les mêmes qu'aux points précédents, soit que $\theta_1 = 3$, $\theta_2 = 2$ et $P(C_1) = P(C_2) = 0{,}5$. Donnez les développements menant à votre fonction de décision.

Solution: Avec une fonction de perte, la prise de décision se base sur la minimisation du risque de classement :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1, C_2\}}{\operatorname{argmin}} \; R(C_i|x),$$

où :

$$R(C_i|x) = \sum_{C_j \in \{C_1, C_2\}} \mathcal{L}(C_i, C_j)\, P(C_j|x).$$

Étant donné que $p(x)$ ne change pas selon la classe et que $P(C_1) = P(C_2)$, on peut simplifier le risque par :

$$R(C_i|x) = \sum_{C_j \in \{C_1, C_2\}} \mathcal{L}(C_i, C_j)\, p(x|C_j).$$

Dans le cas présent, les risques pour les classes C_1 et C_2 sont donc :

$$\begin{aligned} R(C_1|x) &= \mathcal{L}(\alpha_1, C_2)\, p(x|C_2) = 0{,}5\, p(x|C_2), \\ R(C_2|x) &= \mathcal{L}(\alpha_2, C_1)\, p(x|C_1) = p(x|C_1). \end{aligned}$$

Comme les fonctions $p(x|C_1)$ et $p(x|C_2)$ sont des équations linéaires, ce qui implique que $R(C_1|x)$ et $R(C_2|x)$ le sont aussi, il suffit de déterminer le point où les deux droites $R(C_1|x)$ et $R(C_2|x)$ se croisent :

$$\begin{aligned} R(C_1|x) &= R(C_2|x), \\ 0{,}5\, p(x|C_2) &= p(x|C_1), \\ 0{,}5 \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} &= \frac{-2\,(x-\theta_1)}{(\theta_1)^2}, \\ 0{,}5 \frac{2x-4}{(4-2)^2} &= \frac{-2x+6}{(3)^2}, \\ \frac{x}{4} + \frac{2x}{9} &= \frac{1}{2} + \frac{6}{9}, \\ 0{,}47222\, x &= 1{,}16667, \\ x &= 2{,}4706. \end{aligned}$$

Donc, en se basant sur la figure de l'énoncé de la question, on obtient comme fonction de prise de décision ceci :

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 2{,}4706] \\ C_2 & \text{pour } x \in [2{,}4706, 4] \end{cases}.$$

Question 2 (32 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 de transfert 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 poids-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 est la valeur désirée pour le neurone de sortie du réseau ;
— $\mathcal{X}$ est l'ensemble des N données d'entraînement.

(16 pts) (a) Développez les équations permettant de mettre à jour les poids w_i et w_0 du neurone de sortie par descente du gradient, en utilisant le critère de l'erreur quadratique moyenne.

Solution:

$$
\begin{aligned}
e^t &= r^t - \mathrm{h}(\mathbf{x}^t) = r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right] \\
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i}\frac{1}{2N}\sum_{\mathbf{x}^t\in\mathcal{X}}(e^t)^2 = \frac{1}{2N}\sum_{\mathbf{x}^t\in\mathcal{X}} 2e^t \frac{\partial}{\partial w_i}\left(r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right]\right) \\
&= -\frac{1}{N}\sum_{\mathbf{x}^t\in\mathcal{X}} e^t \phi_i(\mathbf{x}^t) \\
\Delta w_i &= -\eta\frac{\partial E}{\partial w_i} = \frac{\eta}{N}\sum_{\mathbf{x}^t\in\mathcal{X}} e^t \phi_i(\mathbf{x}^t) \\
\frac{\partial E}{\partial w_0} &= \frac{\partial}{\partial w_0}\frac{1}{2N}\sum_{\mathbf{x}^t\in\mathcal{X}}(e^t)^2 = \frac{1}{2N}\sum_{\mathbf{x}^t\in\mathcal{X}} 2e^t \frac{\partial}{\partial w_0}\left(r^t - \left[\sum_{j=1}^{R} w_j \phi_j(\mathbf{x}^t) + w_0\right]\right) \\
&= -\frac{1}{N}\sum_{\mathbf{x}^t\in\mathcal{X}} e^t \\
\Delta w_0 &= -\eta\frac{\partial E}{\partial w_0} = \frac{\eta}{N}\sum_{\mathbf{x}^t\in\mathcal{X}} e^t \\
w_i &= w_i + \Delta w_i, \quad i = 0,\ldots,R
\end{aligned}
$$

(16 pts) (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.

Solution:

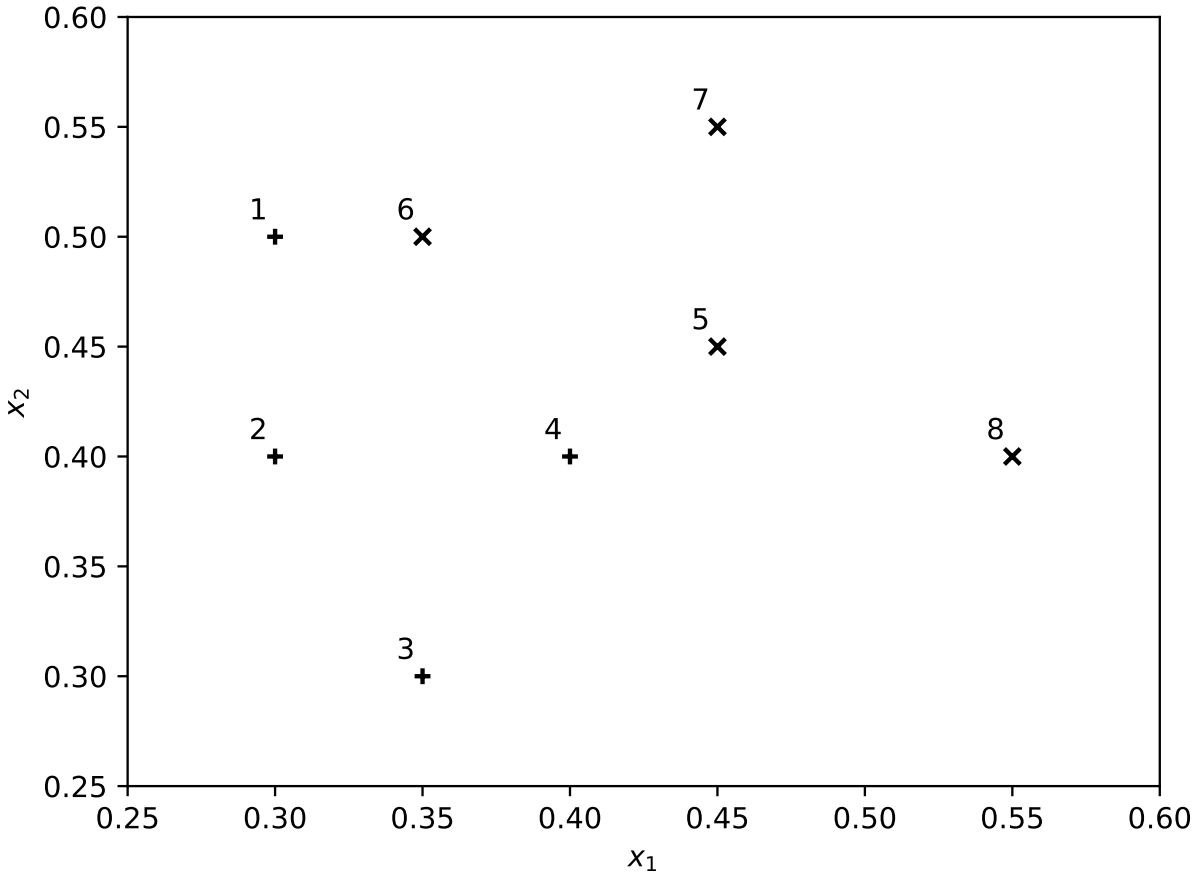
$$
\begin{aligned}
\frac{\partial \phi_i(\mathbf{x}^t)}{\partial m_{i,j}} &= \frac{\partial}{\partial m_{i,j}} \exp\left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] \\
&= \exp\left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] \frac{\partial}{\partial m_{i,j}} \left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] \\
&= \frac{(x_j^t - m_{i,j})}{s_i^2} \exp\left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] = \frac{x_j^t - m_{i,j}}{s_i^2} \phi_i(\mathbf{x}^t) \\
\frac{\partial E}{\partial m_{i,j}} &= \frac{\partial}{\partial m_{i,j}} \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} (e^t)^2 = \frac{1}{2N} \sum_{\mathbf{x}^t \in \mathcal{X}} 2e^t \frac{\partial}{\partial m_{i,j}} \left(r^t - \left[\sum_{l=1}^{R} w_l \phi_l(\mathbf{x}^t) + w_0\right]\right) \\
&= \frac{1}{N} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (-1) w_i \frac{\partial \phi_i(\mathbf{x}^t)}{\partial m_{i,j}} = -\frac{w_i}{N s_i^2} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (x_j^t - m_{i,j}) \phi_i(\mathbf{x}^t) \\
\Delta m_{i,j} &= -\eta \frac{\partial E}{\partial m_{i,j}} = \frac{\eta w_i}{N s_i^2} \sum_{\mathbf{x}^t \in \mathcal{X}} e^t (x_j^t - m_{i,j}) \phi_i(\mathbf{x}^t) \\
m_{i,j} &= m_{i,j} + \Delta m_{i,j}, \quad i = 1, \ldots, R, \quad j = 1, \ldots, D
\end{aligned}
$$

Question 3 (24 points sur 100)

Soit le jeu de données suivant, en deux dimensions :

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

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.



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

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

(8 pts) (a) Calculez les valeurs du vecteur $\mathbf{w}$ de l'hyperplan séparateur de ce classifieur.

Solution: Les valeurs du vecteur $\mathbf{w}$ sont calculées selon l'équation suivante :

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{x}^t.$$

Dans le cas présent, les valeurs du vecteur sont $\mathbf{w} = [17\ 11]^\top$.

(8 pts) (b) Déterminez les données qui sont des vecteurs de support ainsi que les données qui sont dans la marge ou mal classées.

Solution: Les données $\mathbf{x}^1$, $\mathbf{x}^4$, $\mathbf{x}^5$ et $\mathbf{x}^6$ représentent les vecteurs de support du classifieur, comme leur α^t respectif est non nul.

Les données dans la marge ou mal classées ont une valeur de α^t correspondant au paramètre de régularisation C. Donc, les données $\mathbf{x}^4$ et $\mathbf{x}^6$ sont dans la marge ou mal classées, avec $\alpha^4 = \alpha^6 = C = 200$.

(8 pts) (c) Supposons maintenant que l'on veut traiter une donnée $\mathbf{x} = [0{,}37\ 0{,}45]^\top$ avec ce SVM. Calculez la valeur $\mathrm{h}(\mathbf{x})$ correspondante (valeur réelle avant seuillage de la sortie).

Solution: On calcule la valeur de $\mathrm{h}(\mathbf{x})$ selon l'équation suivante :

$$\mathrm{h}(\mathbf{x}) = \sum_t \alpha^t r^t (\mathbf{x}^t)^\top \mathbf{x} + w_0.$$

Dans le cas présent, avec $\mathbf{x} = [0{,}37\ 0{,}45]^\top$, la sortie correspondante du classifieur est $\mathrm{h}(\mathbf{x}) = -0{,}36$. Donc, la donnée est assignée à la classe des données négatives ($r = -1$).

Question 4 (20 points sur 100)

En utilisant les données de la question précédente (question 3), en deux dimensions, répondez aux questions suivantes.

(10 pts) (a) Calculez le taux d'erreur de classement selon une approche *leave-one-out* avec un classifieur de type k-plus proches voisins, en employant $k = 1$ voisins et la distance D_∞. Explicitez la démarche menant au calcul du taux d'erreur.

Solution:

Donnée	1-PPV-LOO	Erreur ?
$\mathbf{x}^1$	$\mathbf{x}^6$	Oui
$\mathbf{x}^2$	$\{\mathbf{x}^1, \mathbf{x}^3, \mathbf{x}^4\}$	Non
$\mathbf{x}^3$	$\{\mathbf{x}^2, \mathbf{x}^4\}$	Non
$\mathbf{x}^4$	$\mathbf{x}^5$	Oui
$\mathbf{x}^5$	$\mathbf{x}^4$	Oui
$\mathbf{x}^6$	$\mathbf{x}^1$	Oui
$\mathbf{x}^7$	$\mathbf{x}^5$	Non
$\mathbf{x}^8$	$\mathbf{x}^5$	Non

Donc, il y a quatre erreurs d'effectuées sur les huit instances, pour un taux d'erreur de classement de $50\,\%$.

(10 pts) (b) Effectuez une édition de Wilson de ce jeu de données, en utilisant un voisin ($k = 1$) et une distance euclidienne. Traitez les données dans leur ordre d'indice, c'est-à-dire dans l'ordre x^1,x^2,x^3,...,x^8. Explicitez votre démarche et rapportez les données formant l'ensemble des prototypes après l'édition.

Solution:

Donnée	Prototypes	PPV-LOO	Erreur
$\mathbf{x}^1$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^6$	Oui
$\mathbf{x}^2$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^4$	Non
$\mathbf{x}^3$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\{\mathbf{x}^3, \mathbf{x}^4\}$	Non
$\mathbf{x}^4$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	Oui
$\mathbf{x}^5$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^7$	Non
$\mathbf{x}^6$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^2$	Oui
$\mathbf{x}^7$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	Non
$\mathbf{x}^8$	$\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	Non

Donc, l'ensemble de prototypes résultant de l'édition de Wilson est $\{\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$.