Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Hiver 2014

UNIVERSITÉ LAVAL

EXAMEN FINAL

Instructions : – Une feuille aide-mémoire recto verso manuscrite est permise ;
– Durée de l'examen : 2 h 50.
Pondération : – Cet examen compte pour $35\,\%$ de la note finale ;
– La note est saturée à $100\,\%$ si le total des points avec bonus excède cette valeur.

Question 1 (20 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.

(10) (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:

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

(10) (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:

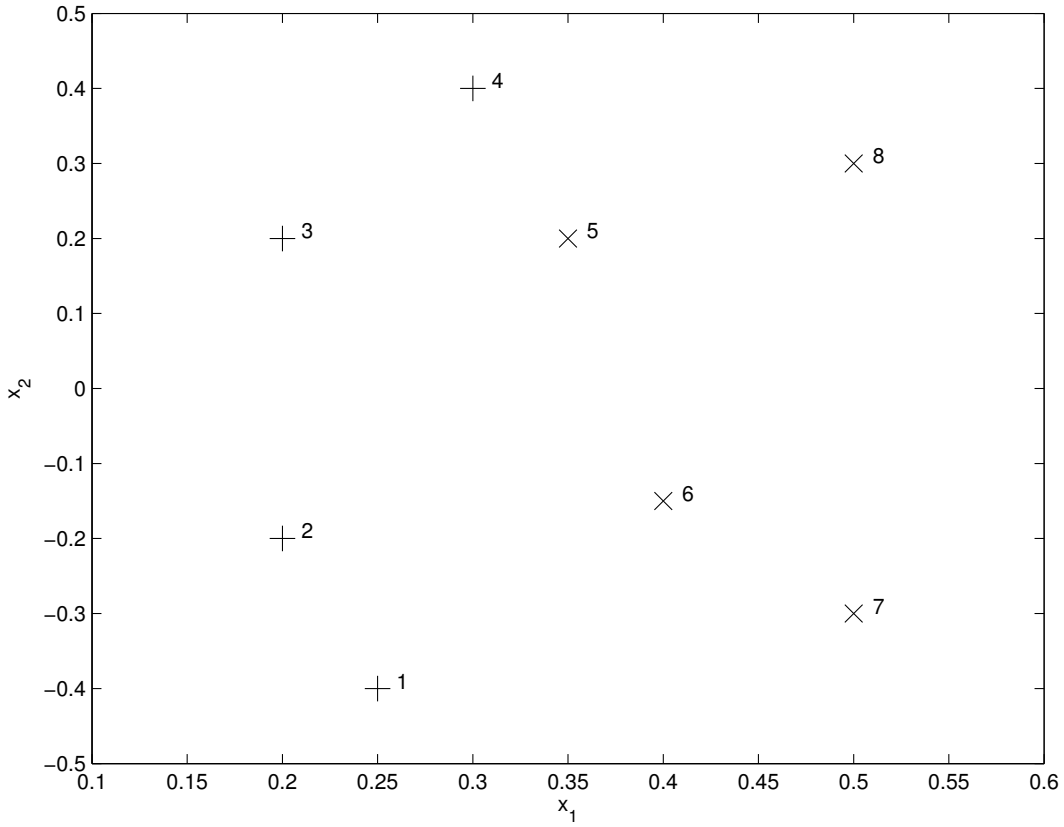
$$
\begin{aligned}
\frac{\partial \phi_i(\mathbf{x}^t)}{\partial m_{i,j}} &= \frac{\partial}{\partial m_{i,j}} \exp\left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] \\
&= \exp\left[-\frac{\|\mathbf{x}^t - \mathbf{m}_i\|^2}{2s_i^2}\right] \frac{\partial}{\partial m_{i,j}}\left[-\frac{\|\mathbf{x}^t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

Question 2 (22 points sur 100)

Soit le jeu de données suivant, en deux dimensions :

$$\mathbf{x}^1 = [0{,}25 \;\; -0{,}4]^T, \quad \mathbf{x}^2 = [0{,}2 \;\; -0{,}2]^T, \quad \mathbf{x}^3 = [0{,}2 \;\; 0{,}2]^T, \quad \mathbf{x}^4 = [0{,}3 \;\; 0{,}4]^T,$$
$$\mathbf{x}^5 = [0{,}35 \;\; 0{,}2]^T, \quad \mathbf{x}^6 = [0{,}4 \;\; -0{,}15]^T, \quad \mathbf{x}^7 = [0{,}5 \;\; -0{,}3]^T, \quad \mathbf{x}^8 = [0{,}5 \;\; 0{,}3]^T.$$

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 $C = 300$:

$\alpha^1 = 73{,}541, \quad \alpha^2 = 0, \quad \alpha^3 = 0, \quad \alpha^4 = 226{,}459, \quad \alpha^5 = 300, \quad \alpha^6 = 0, \quad \alpha^7 = 0, \quad \alpha^8 = 0,$
$w_0 = -6{,}136.$

(5) (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 :

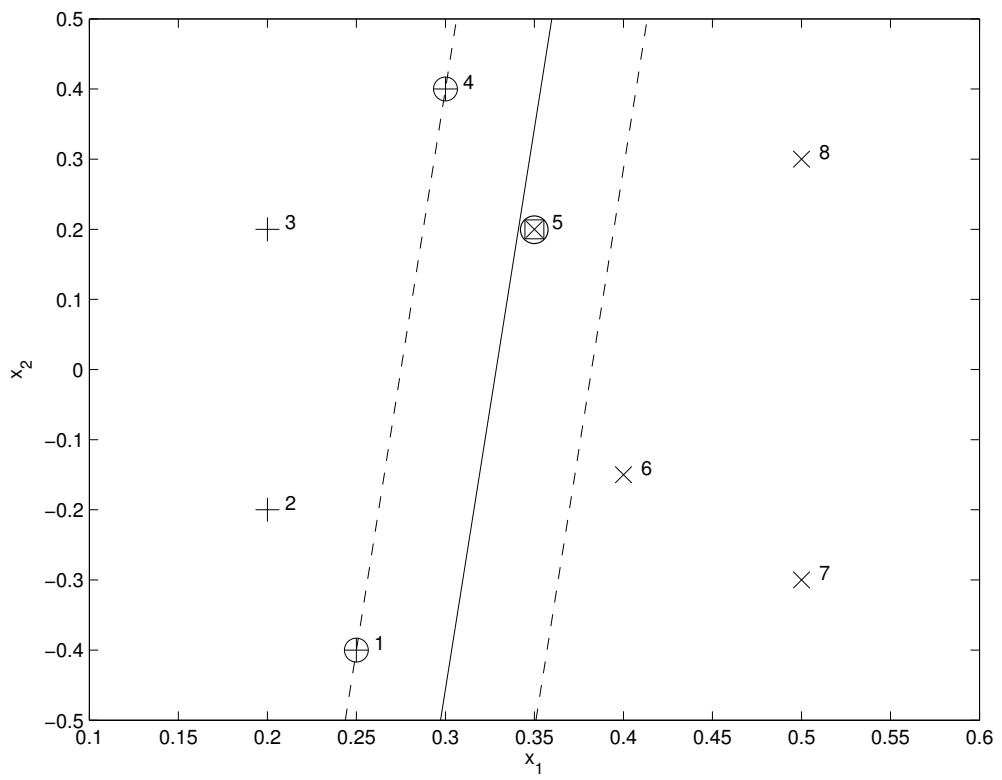
$$\mathbf{w} = \sum_t \alpha^t \, r^t \, \mathbf{x}^t.$$

Dans le cas présent, les valeurs du vecteur sont $\mathbf{w} = [18{,}677 \;\; -1{,}167]^T$.

(12) (b) Déterminez les données qui sont des vecteurs de support (mais pas dans la marge) ainsi que les données qui sont dans la marge ou mal classées. Tracez ensuite un graphique représentant toutes les données du jeu, en encerclant les vecteurs de support et en encadrant les données dans la marge ou mal classées. Tracez également la droite représentant l'hyperplan séparateur ainsi que deux droites pointillées représentant les limites de la marge. **N'utilisez pas** le graphique du préambule de l'énoncé de la question pour donner votre réponse, tracez vous-même un nouveau graphique dans votre cahier de réponse.

Solution: Les données dans la marge ou mal classées ont une valeur de α^t correspondant au paramètre de régularisation C. Donc, la donnée $\mathbf{x}^5$ est la seule donnée dans la marge ou mal classée du jeu, comme $\alpha^5 = C = 300$. Les données $\mathbf{x}^1$ et $\mathbf{x}^4$ représentent les vecteurs de support du classifieur, comme leur α^t respectif est non nul.

Le graphique demandé correspond à ce qui suit.



(5) (c) Supposons maintenant que l'on veut classer une nouvelle donnée $\mathbf{x} = [0{,}35 \quad -0{,}1]^T$ avec ce SVM. Calculez la valeur $\mathrm{h}(\mathbf{x})$ correspondante (valeur réelle avant seuillage de la sortie).

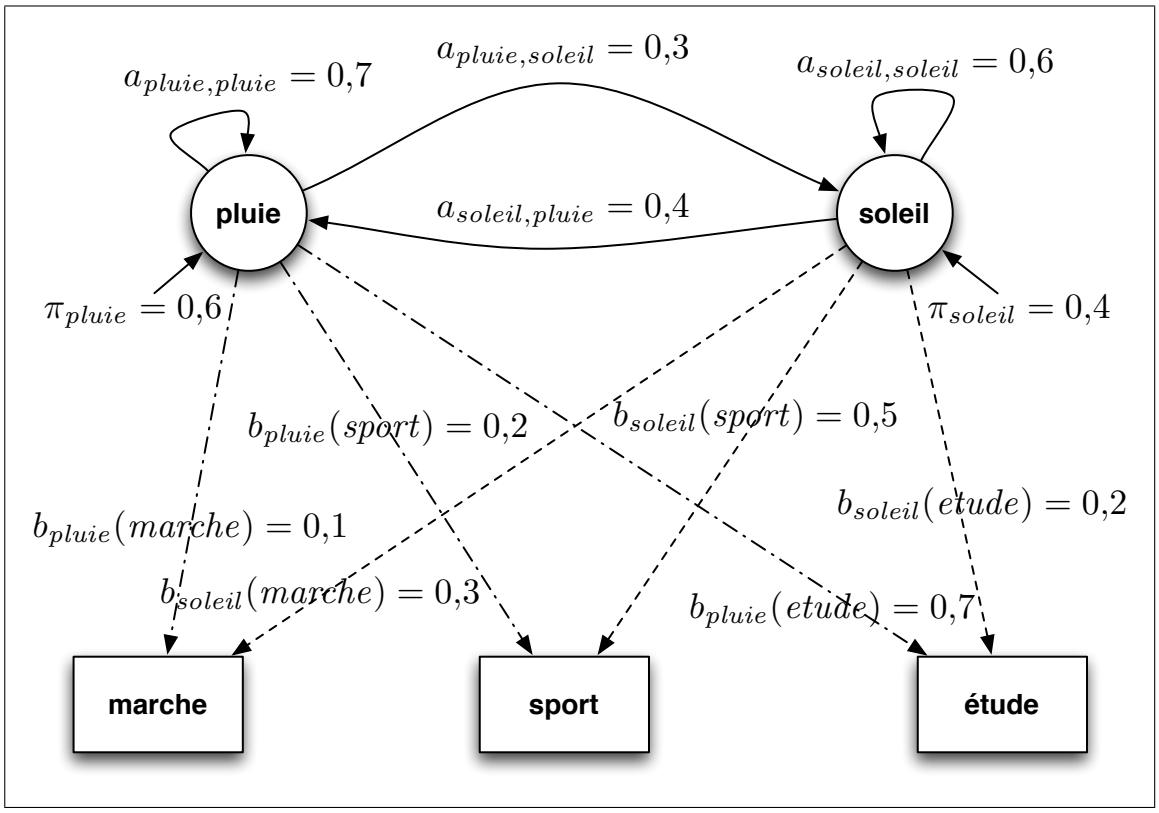
Solution: On calcule la valeur de $\mathrm{h}(\mathbf{x})$ selon l'équation suivante :

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

Dans le cas présent, avec $\mathbf{x} = [0{,}35 \quad 0{,}1]^T$, la sortie correspondante du classifieur est $\mathrm{h}(\mathbf{x}) = 0{,}5175$. Donc, la donnée est assignée à la classe des données positives.

Question 3 (16 points sur 100)

Soit le modèle de Markov caché (MMC) suivant, correspondant à l'exemple présenté en classe.



(4) (a) Supposons la séquence d'états $S = \{pluie, pluie, soleil, pluie\}$. Calculez la probabilité d'obtenir cette séquence d'états avec le MMC présenté ici haut.

Solution: La probabilité d'obtenir la séquence S est donnée par :

$$\begin{aligned} P(S|\lambda) &= \pi_{s_1} \prod_{t=1}^{T-1} a_{s_t,s_{t+1}} \\ &= \pi_{pluie}\, a_{pluie,pluie}\, a_{pluie,soleil}\, a_{soleil,pluie} \\ &= 0{,}6 \cdot 0{,}7 \cdot 0{,}3 \cdot 0{,}4 \\ &= 0{,}0504. \end{aligned}$$

Il y a donc $5{,}04\,\%$ de chance de produire la séquence d'états avec le MMC.

(8) (b) Supposons maintenant une séquence d'observations $O = \{$*étude*,*marche*,*étude*$\}$. Calculez la probabilité d'obtenir cette séquence d'observations avec le MMC présenté ici haut.

Solution: Deux approches sont possibles pour calculer la probabilité $P(O|\lambda)$ d'obtenir l'observation avec le MMC. L'approche préconisée consiste à utiliser la procédure avant, en calculant les $\alpha_t(i) = P(\{o_1, \ldots, o_t\}, s_t = S_i|\lambda)$ selon la formulation récursive

suivante :

$$\begin{aligned}\alpha_1(i) &= P(o_1|s_1 = S_i,\lambda)P(s_1 = S_i|\lambda) = \pi_i\, b_i(o_1), \quad i = 1,\ldots,N,\\ \alpha_{t+1}(j) &= P(\{o_1,\ldots,o_t\},s_t = S_j|\lambda) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i)\, a_{i,j}, \quad j = 1,\ldots,N.\end{aligned}$$

Dans le cas présent, pour la séquence d'observation à évaluer, les valeurs de $\alpha_t(i)$ sont :

t	o_t	$\alpha_t(\textit{pluie})$	$\alpha_t(\textit{soleil})$
1	*étude*	$0{,}6 \cdot 0{,}7 = 0{,}42$	$0{,}4 \cdot 0{,}2 = 0{,}08$
2	*marche*	$0{,}1 \cdot (0{,}42 \cdot 0{,}7 + 0{,}08 \cdot 0{,}3) = 0{,}0318$	$0{,}3 \cdot (0{,}42 \cdot 0{,}3 + 0{,}08 \cdot 0{,}6) = 0{,}0522$
3	*étude*	$0{,}7 \cdot (0{,}0318 \cdot 0{,}7 + 0{,}0522 \cdot 0{,}4) = 0{,}030198$	$0{,}2 \cdot (0{,}0318 \cdot 0{,}3 + 0{,}0522 \cdot 0{,}6) = 0{,}008172$

À partir de ces résultats, on peut évaluer la probabilité d'observer la séquence comme étant :

$$\begin{aligned}P(O|\lambda) &= \sum_{i=1}^{N} P(O,s_T = S_i|\lambda) = \sum_{i=1}^{N} \alpha_T(i)\\ &= \alpha_3(\textit{pluie}) + \alpha_3(\textit{soleil}) = 0{,}030198 + 0{,}008172 = 0{,}03837\end{aligned}$$

Donc, il y a $3{,}837\,\%$ de chance d'observer la séquence avec ce MMC.
Une approche alternative pour faire le calcul est d'utiliser la formulation directe de la probabilité d'observation :

$$P(O|\lambda) = \sum_{\forall S} P(O,S|\lambda).$$

Cette approche est possible dans le cas présent, car le nombre de séquences d'états possibles est réduit.

$$\begin{aligned}P(O|\lambda) &= P(O,\{\textit{pluie},\textit{pluie},\textit{pluie}\}|\lambda) + P(O,\{\textit{pluie},\textit{pluie},\textit{soleil}\}|\lambda) +\\ &\quad P(O,\{\textit{pluie},\textit{soleil},\textit{pluie}\}|\lambda) + P(O,\{\textit{pluie},\textit{soleil},\textit{soleil}\}|\lambda) +\\ &\quad P(O,\{\textit{soleil},\textit{pluie},\textit{pluie}\}|\lambda) + P(O,\{\textit{soleil},\textit{pluie},\textit{soleil}\}|\lambda) +\\ &\quad P(O,\{\textit{soleil},\textit{soleil},\textit{pluie}\}|\lambda) + P(O,\{\textit{soleil},\textit{soleil},\textit{soleil}\}|\lambda)\\ &= \pi_{\textit{pluie}}\, b_{\textit{étude}}(\textit{pluie})\, a_{\textit{pluie,pluie}}\, b_{\textit{marche}}(\textit{pluie})\, a_{\textit{pluie,pluie}}\, b_{\textit{étude}}(\textit{pluie}) +\\ &\quad \pi_{\textit{pluie}}\, b_{\textit{étude}}(\textit{pluie})\, a_{\textit{pluie,pluie}}\, b_{\textit{marche}}(\textit{pluie})\, a_{\textit{pluie,soleil}}\, b_{\textit{étude}}(\textit{soleil}) +\\ &\quad \pi_{\textit{pluie}}\, b_{\textit{étude}}(\textit{pluie})\, a_{\textit{pluie,soleil}}\, b_{\textit{marche}}(\textit{soleil})\, a_{\textit{soleil,pluie}}\, b_{\textit{étude}}(\textit{pluie}) +\\ &\quad \pi_{\textit{pluie}}\, b_{\textit{étude}}(\textit{pluie})\, a_{\textit{pluie,soleil}}\, b_{\textit{marche}}(\textit{soleil})\, a_{\textit{soleil,soleil}}\, b_{\textit{étude}}(\textit{soleil}) +\\ &\quad \pi_{\textit{soleil}}\, b_{\textit{étude}}(\textit{soleil})\, a_{\textit{soleil,pluie}}\, b_{\textit{marche}}(\textit{pluie})\, a_{\textit{pluie,pluie}}\, b_{\textit{étude}}(\textit{pluie}) +\\ &\quad \pi_{\textit{soleil}}\, b_{\textit{étude}}(\textit{soleil})\, a_{\textit{soleil,pluie}}\, b_{\textit{marche}}(\textit{pluie})\, a_{\textit{pluie,soleil}}\, b_{\textit{étude}}(\textit{soleil}) +\\ &\quad \pi_{\textit{soleil}}\, b_{\textit{étude}}(\textit{soleil})\, a_{\textit{soleil,soleil}}\, b_{\textit{marche}}(\textit{soleil})\, a_{\textit{soleil,pluie}}\, b_{\textit{étude}}(\textit{pluie}) +\\ &\quad \pi_{\textit{soleil}}\, b_{\textit{étude}}(\textit{soleil})\, a_{\textit{soleil,soleil}}\, b_{\textit{marche}}(\textit{soleil})\, a_{\textit{soleil,soleil}}\, b_{\textit{étude}}(\textit{soleil})\\ &= 0{,}03837\end{aligned}$$

Dans le cas général, cette approche n'est pas tractable étant donné l'explosion du nombre de séquences d'états possibles selon la longueur des séquences.

(4) (c) Expliquez en quoi consistent les valeurs $\gamma_t^k(i)$ et $\xi_t^k(i,j)$ utilisées par l'algorithme Baum-Welch.

Solution: La valeur $\gamma_t^k(i)$ correspond à la probabilité d'effectuer d'être dans l'état S_i au pas de temps t dans la séquence d'observations O^k. Cette valeur est calculée selon l'équation suivante :

$$\gamma_t^k(i) = \frac{\alpha_t^k(i)\,\beta_t^k(i)}{\sum_{j=1}^{N} \alpha_t^k(j)\,\beta_t^k(j)}.$$

La valeur $\xi_t^k(i,j)$ correspond à la probabilité d'effectuer une transition de l'état S_i au temps t à l'état S_j au temps $t+1$ dans le traitement de la séquence d'observations O^k. Cette valeur est calculée selon l'équation suivante :

$$\xi_t^k(i,j) = \frac{\alpha_t^k(i)\,a_{i,j}\,b_j(o_{t+1})\,\beta_{t+1}^k(j)}{\sum_u \sum_v \alpha_t^k(u)\,a_{u,v}\,b_v(o_{t+1})\,\beta_{t+1}^k(v)}.$$

Ces valeurs sont évaluées à l'étape E de l'algorithme Baum-Welch sur le jeu de données de séquences d'observations $\mathcal{O} = \{O^k\}_{k=1}^K$. Elles sont ensuite utilisées à l'étape M de l'algorithme pour estimer les paramètres $\lambda = \{\mathbf{A},\mathbf{B},\boldsymbol{\pi}\}$ du MMC.

Question 4 (42 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Expliquez en quoi le classement logistique vu en classe fait un pont, conceptuellement parlant, entre les méthodes génératives et les méthodes discriminatives de classement.

Solution: Le classement logistique effectue un apprentissage basé sur une descente du gradient appliquée à un discriminant linéaire, ce qui est une approche discriminative classique. La valeur obtenue du discriminant linéaire est ensuite utilisée comme entrée d'une fonction sigmoïde, dont la valeur peut être interprétée comme la probabilité a posteriori d'un classement selon des lois normales multivariées avec matrice de covariance partagée pour les deux classes. Le classement de données selon les probabilités a posteriori des différentes classes pour une donnée $\mathbf{x}$, soit $P(C_i|\mathbf{x})$, est caractéristique des méthodes génératives.

(3) (b) Expliquez pourquoi il est possible de traiter avec succès des données non linéairement séparables avec un discriminant linéaire en utilisant des fonctions de base.

Solution: Avec des fonctions de base, on peut projeter les données non linéairement séparables dans un nouvel espace de plus grande dimension, décrit pas les fonctions de base, où les données seront linéairement séparables, ou à tout le moins plus faciles à traiter qu'avec un discriminant linéaire travaillant dans l'espace d'entrée.

(3) (c) Expliquez pourquoi il est intéressant d'effectuer plusieurs exécutions d'un entraînement avec perceptron multicouches sur un même jeu de données, en utilisant les mêmes hyperparamètres.

Solution: Les poids initiaux du perceptron multicouches sont déterminés aléatoirement et l'algorithme de rétropropagation est reconnu comme étant relativement instable, produisant des classifieurs significativement différents d'un entraînement à l'autre. Plusieurs exécutions de l'entraînement permettent donc d'éviter d'être malchanceux dans un entraînement particulier, avec un classifieur offrant de faibles performances, en effectuant plusieurs répétitions et conservant le meilleur classifieur obtenu sur ces répétitions.

(3) (d) Expliquez la relation entre la taille du jeu de données d'entraînement et la complexité algorithmique de méthodes à noyau telles que les séparateurs à vaste marge (SVM).

Solution: Les méthodes à noyau se caractérisent par une fonction noyau $K(\mathbf{x},\mathbf{y})$ souvent appliquée à toutes les paires de données du jeu d'entraînement. Ceci implique donc une complexité algorithmique des méthodes qui est au moins quadratique selon le nombre de données du jeu d'entraînement. Dans le cas des SVM, une implémentation naïve donne une complexité algorithmique cubique selon nombre de données du jeu d'entraînement, et entre quadratique et cubique pour des implémentations plus optimisées.

(3) (e) Donnez le principal avantage et le principal désavantage d'un apprentissage en ligne comparativement à un apprentissage par lots avec une optimisation de classifieurs basée sur la descente du gradient (incluant la rétropropagation des erreurs).

Solution: Un apprentissage en ligne permet d'obtenir une convergence plus rapide de l'optimisation comparativement à un apprentissage par lots, au risque d'une plus grande instabilité relativement à la convergence vers une bonne solution.

(3) (f) Expliquez ce que l'on entend par la diversité des membres dans un contexte d'ensembles de classifieurs et pourquoi cette diversité est souhaitable.

Solution: Par diversité des membres, on entend des classifieurs différents les uns des autres, qui ne font pas les mêmes erreurs. Un exemple exagéré d'un ensemble sans diversité est le cas où un ensemble est formé de M copies identiques d'un même classifieur entraîné. La création d'un tel ensemble est superflue, comme tous les membres vont faire les mêmes erreurs sur les mêmes données. Avec un ensemble divers, on peut espérer que lorsqu'un classifieur donné fait une erreur, la majorité des autres classifieurs de l'ensemble classe la donnée correctement, de sorte que la décision de l'ensemble sera bonne.

(3) (g) Complétez la matrice de décision suivante, basée sur un code à correction d'erreurs pour un ensemble de $L = 5$ classifieurs, pour une décision d'ensemble à $K = 3$ classes, afin de maximiser la robustesse de l'ensemble relativement à une erreur des membres :

$$\begin{bmatrix} -1 & -1 & -1 & +1 & ? \\ -1 & +1 & +1 & -1 & ? \\ +1 & -1 & +1 & -1 & ? \end{bmatrix}.$$

Solution: Pour évaluer la robustesse de la matrice de décision, il faut calculer la distance de Hamming entre chaque paire de lignes :

— Distance entre les lignes 1 et 2 : 3 ;
— Distance entre les lignes 1 et 3 : 3 ;
— Distance entre les lignes 2 et 3 : 2.

Donc, il faudrait augmenter la distance entre les lignes 2 et 3 afin de maximiser la distance de Hamming. Une solution serait la matrice de décision suivante :

$$\begin{bmatrix} -1 & -1 & -1 & +1 & +1 \\ -1 & +1 & +1 & -1 & +1 \\ +1 & -1 & +1 & -1 & -1 \end{bmatrix}.$$

Ce qui ferait que la distance de Hamming serait de trois pour toutes les paires de lignes, ainsi qu'une tolérance égale à $\lfloor (3-1)/2 \rfloor = 1$ erreur de la part d'un classifieur de base de l'ensemble. D'autres variantes de la matrice de décision sont également valides.

(3) (h) Dans l'algorithme AdaBoost présenté en classe, il est indiqué que l'on interrompt l'algorithme lorsque le taux d'erreur ϵ_j est supérieur à $0{,}5$. Expliquez en quoi consiste ce taux d'erreur, en précisant l'équation utilisée pour son calcul.

Solution: Le taux d'erreur ϵ_j correspond au taux d'erreur sur les données d'entraînement pondérées à l'itération j de l'algorithme AdaBoost, selon les probabilités d'échantillonnage p_j^t des données d'entraînement, soit :

$$\epsilon_j = \sum_{t=1}^{N} p_j^t I(r^t, \mathrm{h}(\mathbf{x})),$$

avec $I(a,b)$ comme fonction de perte 0-1, retournant 1 si $a = b$ et 0 autrement.

(3) (i) Expliquez la différence principale entre une mixture d'experts et un vote pondéré dans un contexte de méthodes par ensemble.

Solution: Dans un vote pondéré, le poids de chacun des classifieurs de base formant l'ensemble est déterminé préalablement, alors qu'avec une mixture d'experts, cette pondération est déterminée dynamiquement, selon la valeur de la donnée à classer $\mathbf{x}$ et l'expertise associée de chacun des classifieurs formant l'ensemble.

(3) (j) Expliquez en quoi consiste l'hypothèse H_0 (hypothèse nulle) évaluée par le test statistique ANOVA et de quelle façon on procède pour la tester.

Solution: L'hypothèse nulle du test ANOVA consiste à déterminer si les moyennes μ_i des résultats obtenus par différentes méthodes sont égales : $H_0 : \mu_1 = \mu_2 = \cdots = \mu_L$. On teste cette hypothèse en supposant deux estimateurs de la variance σ^2 des moyennes μ_i, un premier estimateur qui suppose que l'hypothèse H_0 est valide et un deuxième estimateur qui ne n'assume rien sur la validité de ces hypothèses. Ainsi, si la différence

entre les deux estimateurs de la variance σ^2 des moyennes μ_i est inférieure à un certain seuil, on peut affirmer que l'hypothèse nulle est valide, sinon, on ne peut pas faire cette affirmation.

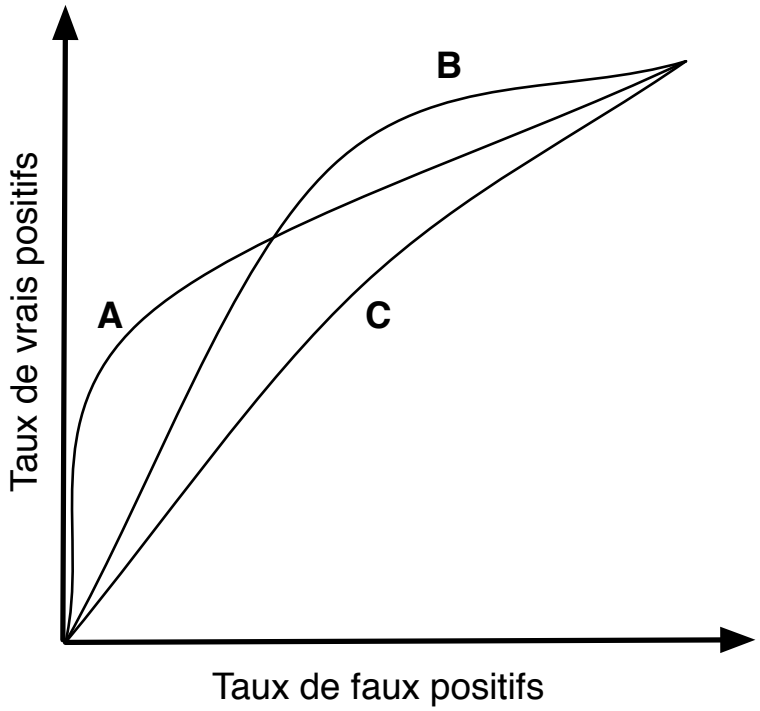
(3) (k) D'un point de vue des processus d'expérimentations, si l'on fait l'entraînement de perceptrons multicouches sur un jeu de données particulier, donnez des exemples de facteurs contrôlables que l'on voudrait étudier.

Solution: Les facteurs contrôlables lorsque l'on procède à l'entraînement d'un classifieur sont typiquement les hyperparamètres de l'algorithme de classement utilisé. Dans le cas du perceptron multicouche, on parle de la topologie du réseau de neurones (nombre de couches cachées, nombre de neurones par couches), taux d'apprentissage, variante de l'algorithme de rétropropagation utilisé (1er ordre ou 2e ordre), fonctions de transfert des différents neurones, etc.

(3) (l) Expliquez en quoi consiste une validation croisée de type 5×2.

Solution: Une validation croisée de type 5×2 consiste à effectuer 5 partitionnements aléatoires distincts, 50/50, du jeu de données. Pour chaque partitionnement aléatoire, on fait deux exécutions d'entraînement et de test, soit une première exécution en utilisant la première partition du jeu pour l'entraînement et la deuxième partition pour le test, et une deuxième exécution en utilisant la deuxième partition du jeu pour l'entraînement et la première partition pour le test. Ensuite, on fait une moyenne des 10 résultats obtenus comme résultat de l'entraînement, avec lesquels on peut estimer les distributions des résultats et effectuer des tests statistiques.

(3) (m) Soit la courbe ROC suivante, présentant les performances de trois classifieurs (A, B et C) opérant selon deux classes (données positives et négatives).



Expliquez en termes clairs et généraux quels classifieurs seraient les plus intéressants à utiliser selon les circonstances rencontrées.

Solution: Le classifieur A offre un meilleur taux de reconnaissance lorsque l'on veut conserver le taux de fausses alarmes bas, quitte à manquer des données positives. À l'opposé, le classifieur B offre les meilleures performances lorsque l'on veut avoir un taux de décisions correctes (taux de vrais positifs) élevé, quitte à ce qu'il y ait régulièrement des fausses alarmes. Le classifieur C offre des performances inférieures en toute circonstance comparativement aux classifieurs A et B, et n'est donc pas intéressant à utiliser dans le cas présent.

(3) (n) Expliquez pourquoi le classement paramétrique avec lois de probabilité multivariées peut performer relativement bien sur de petits jeux de données, mais performe généralement moins bien que d'autres méthodes sur de grands jeux de données, comparativement à des approches telles que le classement par les k-plus proches voisins et le perceptron multicouche.

Solution: Le classement paramétrique avec lois de probabilité multivariées est un modèle de classement relativement rigide, qui fonctionne bien avec peu de données lorsque la loi de probabilité utilisée correspond bien à la nature des données. C'est un classifieur dont le biais est généralement élevé, mais dont la variance est faible. Ceci est approprié pour de petits jeux de données, car l'apprentissage avec de tels jeux fait en sorte qu'une donnée différente des autres peut avoir un effet important sur le classifieur inféré lorsque l'algorithme d'apprentissage est instable ou comporte une grande variance. À l'opposé,

le classement par les k-plus proches voisins et le perceptron multicouche sont des méthodes ayant un biais faible, mais une variance élevée, de sorte que de plus gros jeux de données sont nécessaires pour minimiser l'effet de la variance du modèle.

Question 5 (15 points bonus)

Supposons que l'on veuille traiter des données provenant selon un flot continu, et qu'il n'est pas possible de stocker ces données de façon permanente. Ces données forment un jeu $\mathcal{X} = \{(\mathbf{x}^1,r^1),(\mathbf{x}^2,r^2),\ldots)\}$, où la paire $(\mathbf{x}^t,r^t)$ reçue au temps t comporte la mesure $\mathbf{x}^t$ et la classe correspondante r^t. On suppose que les données sont indépendantes et identiquement distribuées (iid), donc qu'elles sont reçues dans un ordre aléatoire relativement aux mesures $\mathbf{x}^t$ et classes r^t. Ceci implique, par exemple, que les données de chacune des classes sont bien réparties dans le temps relativement au flot de données.

On veut utiliser des méthodes de classement capables de faire un apprentissage à la volée de ces données. On considère utiliser la procédure suivante pour évaluer les performances des différentes méthodes :

1. On reçoit une paire $(\mathbf{x}^t,r^t)$ correspondant à la donnée au temps t ;
2. On évalue la performance de notre classifieur $\mathrm{h}(\cdot|\theta^{t-1})$ de l'itération précédente sur la donnée $(\mathbf{x}^t,r^t)$;
3. On entraîne le classifieur sur cette donnée, ce qui nous donne $\mathrm{h}(\cdot|\theta^t)$;
4. Tant que l'on reçoit de nouvelles données du flot, on retourne à l'étape 1.

À l'étape 2, on calcule le taux d'erreur de classement en ligne selon l'équation suivante :

$$E^t = t - \sum_{j=1}^{t} I(r^j,\mathrm{h}(\mathbf{x}^j|\theta^{j-1})),$$

où :
— $I(a,b)$ est une fonction de perte 0-1 retournant 1 lorsque $a = b$, sinon 0 ;
— $\mathrm{h}(\cdot|\theta^t)$ est le classifieur obtenu après entraînement sur la donnée de l'itération t.
D'après vous, est-ce que cette approche suit une méthodologie qui est valide ? Justifiez clairement et de façon convaincante votre réponse, sans verbiage inutile.

Solution: Oui, cette méthodologie est valide pour entraîner et évaluer la performance des classifieurs entraînés sur des flots de données. En effet, la performance du classifieur mesurée à l'étape 2 provient d'une mesure sur une donnée qui sera que par la suite utilisée pour l'entraînement (étape 3). Donc, tant que chaque donnée d'entraînement n'est traitée qu'une fois (entraînement en une passe), la mesure sera valide comme elle n'a pas été utilisée pour entraîner le classifieur. Une comparaison des différentes méthodes sera également possible selon l'erreur de classement en ligne, gardant toutefois en tête que cette mesure d'erreur de classement sera généralement inférieure à l'erreur de classement obtenue après l'entraînement sur de nouvelles données, comme les mesures ont été prises durant tout le processus d'entraînement, avec très probablement des classifieurs peu entraînés en début de processus, et donc médiocres.