Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Hiver 2013

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)

Le classement logistique, tel que présenté en classe, s'effectue selon l'équation suivante :

$$\mathrm{h}(\mathbf{x}) = \mathrm{f}_{sig}(\mathbf{w}^T\mathbf{x} + w_0) = \frac{1}{1+\exp[-(\mathbf{w}^T\mathbf{x}+w_0)]}.$$

L'entraînement d'un tel classifieur est fait en minimisant l'entropie croisée, définie comme :

$$E_{entr}(\mathbf{w},w_0|\mathcal{X}) = \sum_t E^t_{entr} = -\sum_t \left[r^t \log \mathrm{h}(\mathbf{x}^t) + (1-r^t)\log(1-\mathrm{h}(\mathbf{x}^t))\right].$$

Donnez les développements complets permettant d'obtenir la règle d'apprentissage par descente du gradient de ce classifieur.
Indice : Faites usage la règle de chaînage des dérivées pour effectuer vos développements, en tirant avantage des substitutions suivantes :

$$\begin{aligned} a^t &= \mathbf{w}^T\mathbf{x}^t + w_0 = \sum_i w_i x^t_i + w_0, \\ y^t &= \mathrm{h}(\mathbf{x}^t) = \frac{1}{1+\exp(-a^t)}. \end{aligned}$$

Solution: La descente du gradient requiert de calculer les dérivées partielles de la fonction d'erreur, ici l'entropie croisée, selon les paramètres optimisés, ici $\mathbf{w}$ et w_0. Calculons d'abord la dérivée partielle de la fonction d'entropie croisée selon y^t :

$$\begin{aligned} \frac{\partial E^t_{entr}}{\partial y^t} &= \frac{\partial}{\partial y^t} - \left[r^t \log y^t + (1-r^t)\log(1-y^t)\right] \\ &= -\left(\frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}\right) = -\frac{r^t(1-y^t)-(1-r^t)y^t}{y^t(1-y^t)} \\ &= -\frac{r^t - r^t y^t - y^t + r^t y^t}{y^t(1-y^t)} = -\frac{r^t - y^t}{y^t(1-y^t)}. \end{aligned}$$

Par la suite, nous calculons la dérivée de la fonction sigmoïde, soit la dérivée de y^t selon a^t :

$$\begin{aligned}\frac{\partial y^t}{\partial a^t} &= \frac{\partial}{a^t}\frac{1}{1+\exp(-a^t)} \\ &= \frac{\exp(-a^t)}{[1+\exp(-a^t)]^2} = \frac{1}{1+\exp(-a^t)}\frac{\exp(-a^t)+1-1}{1+\exp(-a^t)} \\ &= \frac{1}{1+\exp(-a^t)}\left(1-\frac{1}{1+\exp(-a^t)}\right) = y^t(1-y^t).\end{aligned}$$

Nous calculons également les dérivées partielles de l'équation du discriminant linéaire a^t selon $\mathbf{w}$ et w_0 :

$$\begin{aligned}\frac{\partial a^t}{\partial w_i} &= \frac{\partial}{\partial w_i^t}\sum_j w_j^t x_j^t + w_0 = x_i^t, \\ \frac{\partial a^t}{\partial w_0} &= \frac{\partial}{\partial w_0}\sum_j w_j^t x_j^t + w_0 = 1.\end{aligned}$$

Et finalement, on applique la règle de chaînage des dérivées pour déterminer les dérivées partielles de la fonction d'erreur selon les paramètres $\mathbf{w}$ et w_0 :

$$\begin{aligned}\frac{\partial E_{entr}}{w_i} &= \sum_t \frac{\partial E_{entr}^t}{w_i} = \sum_t \frac{\partial E_{entr}^t}{y^t}\frac{\partial y^t}{\partial a^t}\frac{\partial a^t}{\partial w_i} \\ &= \sum_t \left(-\frac{r^t-y^t}{y^t(1-y^t)}\right)(y^t(1-y^t))x_i^t = -\sum_t (r^t-y^t)x_i^t,\end{aligned}$$

$$\begin{aligned}\frac{\partial E_{entr}}{w_0} &= \sum_t \frac{\partial E_{entr}^t}{w_0} = \sum_t \frac{\partial E_{entr}^t}{y^t}\frac{\partial y^t}{\partial a^t}\frac{\partial a^t}{\partial w_0} \\ &= \sum_t \left(-\frac{r^t-y^t}{y^t(1-y^t)}\right)(y^t(1-y^t))1 = -\sum_t (r^t-y^t).\end{aligned}$$

La règle d'apprentissage selon la descente du gradient consiste donc à appliquer itérativement les mises à jour suivantes :

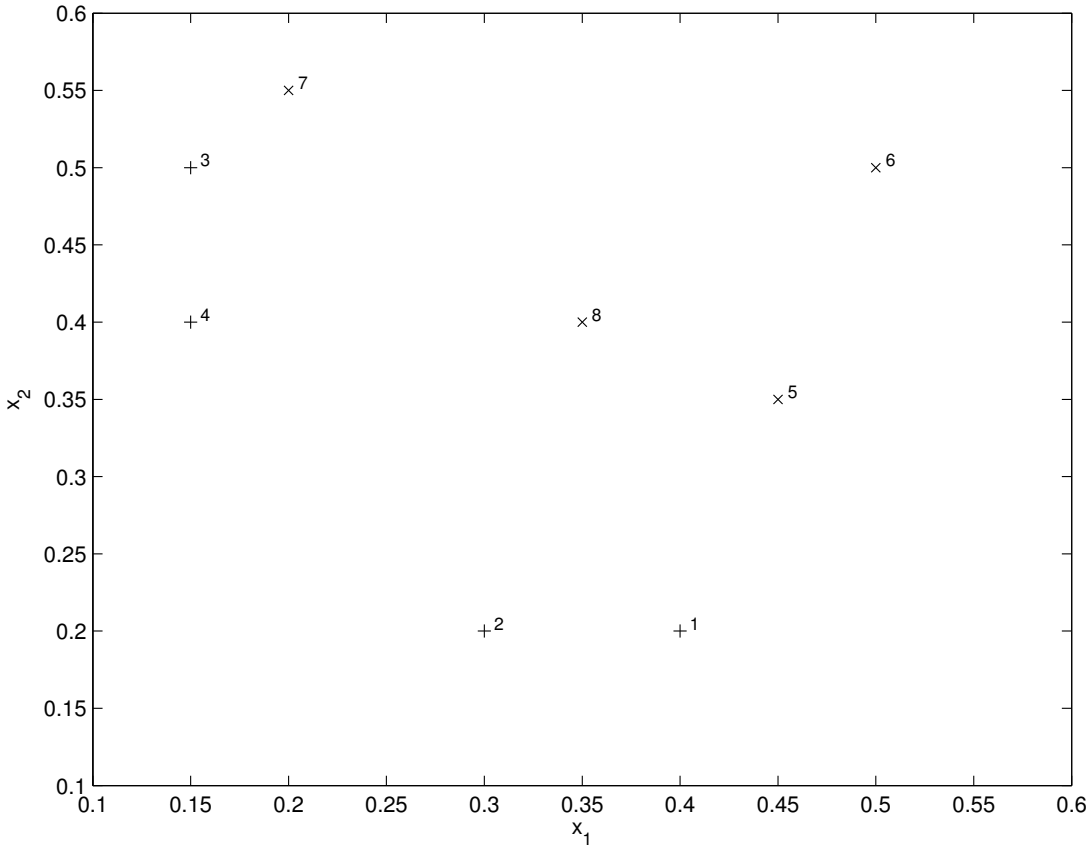
$$\begin{aligned}\Delta w_i &= -\eta\frac{\partial E_{entr}}{\partial w_i} = \eta\sum_t (r^t - \mathrm{h}(\mathbf{x}^t))\, x_i^t, \quad i = 1,\ldots,D, \\ \Delta w_0 &= -\eta\frac{\partial E_{entr}}{\partial w_0} = \eta\sum_t (r^t - \mathrm{h}(\mathbf{x}^t)).\end{aligned}$$

Question 2 (22 points sur 100)

Soit le jeu de données suivant, en deux dimensions :

$$\mathbf{x}^1 = [0{,}4\ 0{,}2]^T, \quad \mathbf{x}^2 = [0{,}3\ 0{,}2]^T, \quad \mathbf{x}^3 = [0{,}15\ 0{,}5]^T, \quad \mathbf{x}^4 = [0{,}15\ 0{,}4]^T,$$
$$\mathbf{x}^5 = [0{,}45\ 0{,}35]^T, \quad \mathbf{x}^6 = [0{,}5\ 0{,}5]^T, \quad \mathbf{x}^7 = [0{,}2\ 0{,}55]^T, \quad \mathbf{x}^8 = [0{,}35\ 0{,}4]^T.$$

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 $C = 200$:

$\alpha^1 = 44{,}44, \quad \alpha^2 = 0, \quad \alpha^3 = 200, \quad \alpha^4 = 0, \quad \alpha^5 = 0, \quad \alpha^6 = 0, \quad \alpha^7 = 162{,}96, \quad \alpha^8 = 81{,}48,$
$w_0 = -9.$

(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 :

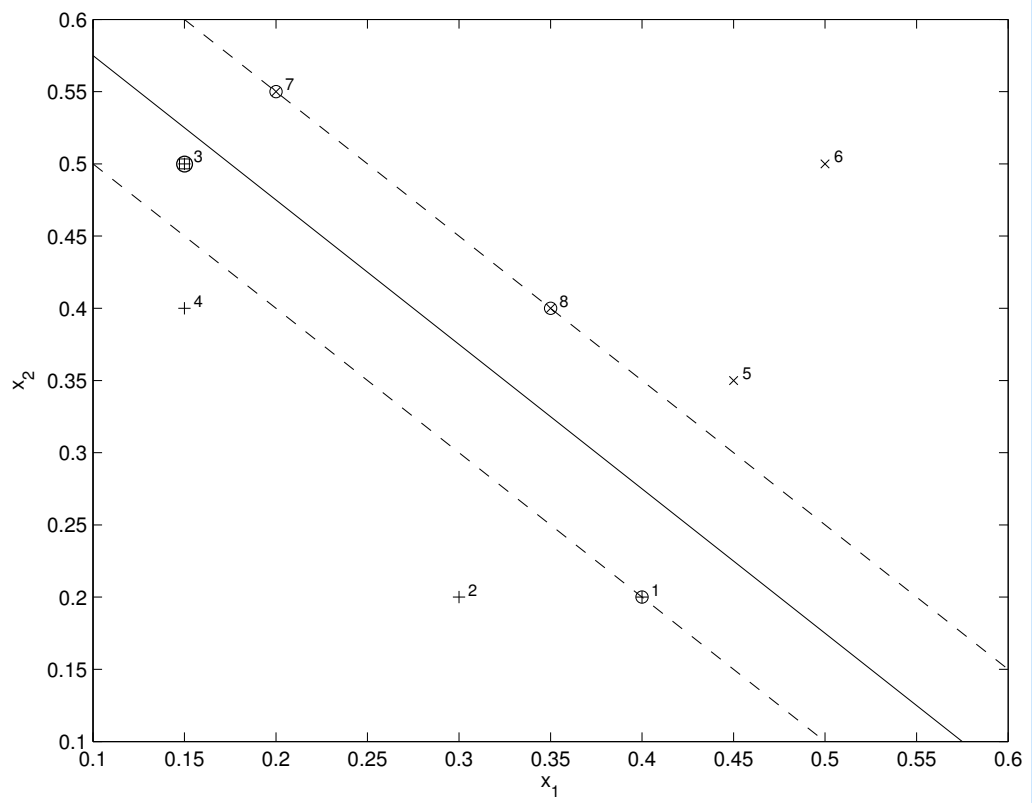
$$\mathbf{w} = \sum_t \alpha^t\, r^t\, \mathbf{x}^t.$$

Dans le cas présent, les valeurs du vecteur sont $\mathbf{w} = [13{,}33\ 13{,}33]^T$.

(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 $\mathbf{x}^3$ est la seule donnée dans la marge ou mal classée du jeu, comme $\alpha^3 = C = 200$. Les données $\mathbf{x}^1$, $\mathbf{x}^7$ et $\mathbf{x}^8$ 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 $\mathbf{x} = [0{,}37\ 0{,}35]^T$ 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 $\mathbf{x} = [0{,}37\ 0{,}35]^T$, la sortie correspondante du classifieur est $\mathrm{h}(\mathbf{x}) = 0{,}6$. Donc, la donnée est assignée à la classe des données positives.

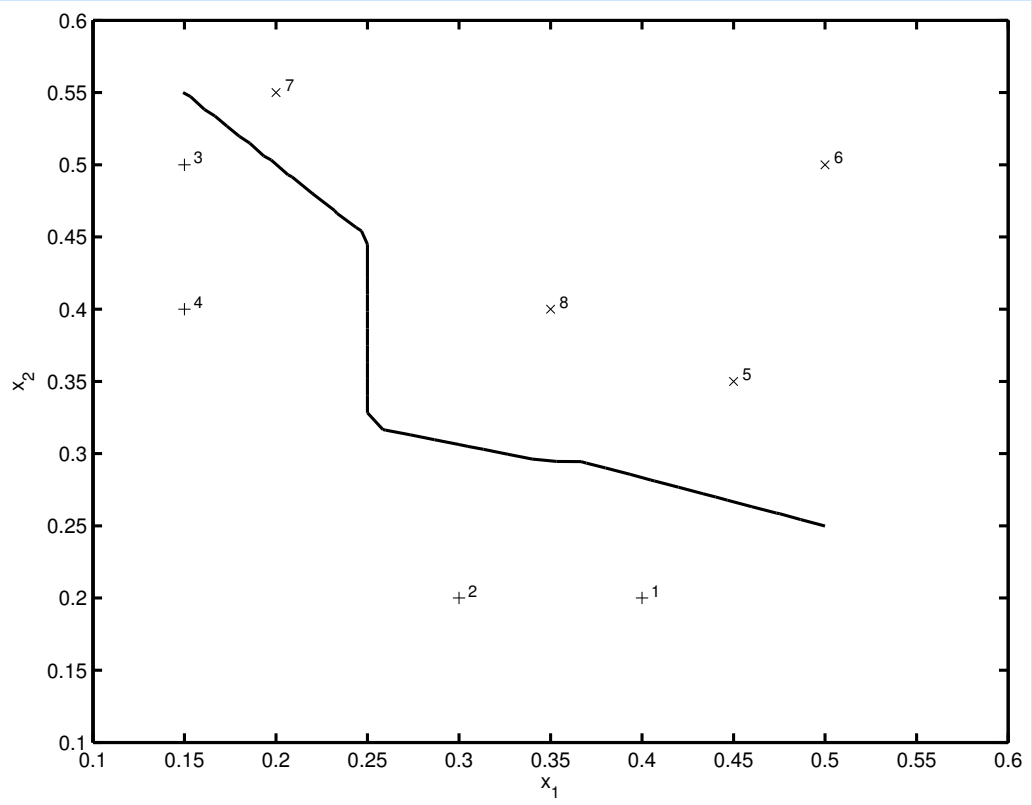
Question 3 (18 points sur 100)

Supposons que l'on veut utiliser la méthode des k-plus proches voisins (k-PPV) s'appuyant sur la distance euclidienne et un seul voisin ($k = 1$) pour classer les données présentées à la question précédente.

(6) (a) Tracer dans un graphique les points du jeu de données ainsi que les frontières de décision correspondant à ce classifieur k-PPV.

N'utilisez pas le graphique de la question précédente de l'énoncé pour donner votre réponse, tracez vous-même un nouveau graphique dans votre cahier de réponse.

Solution:



(6) (b) Calculez le taux d'erreur avec ce classifieur k-PPV selon la méthodologie *leave-one-out*.

Solution: Pour calculer le taux d'erreur avec le classifieur k-PPV, nous allons d'abord calculer la matrice des distances entre les données, qui s'avère être la suivante :

$$D = \begin{bmatrix} 0 & 0{,}1 & 0{,}3905 & 0{,}3202 & 0{,}1581 & 0{,}3162 & 0{,}4031 & 0{,}2062 \\ 0{,}1 & 0 & 0{,}3354 & 0{,}25 & 0{,}2121 & 0{,}3606 & 0{,}364 & 0{,}2062 \\ 0{,}3905 & 0{,}3354 & 0 & 0{,}11 & 0{,}3354 & 0{,}35 & 0{,}0707 & 0{,}2236 \\ 0{,}3202 & 0{,}25 & 0{,}1 & 0 & 0{,}3041 & 0{,}364 & 0{,}1581 & 0{,}2 \\ 0{,}1581 & 0{,}2121 & 0{,}3354 & 0{,}3041 & 0 & 0{,}1581 & 0{,}3202 & 0{,}1118 \\ 0{,}3162 & 0{,}3606 & 0{,}35 & 0{,}364 & 0{,}1581 & 0 & 0{,}3041 & 0.1803 \\ 0{,}4031 & 0{,}364 & 0{,}0707 & 0{,}1581 & 0{,}3202 & 0{,}3041 & 0 & 0{,}2121 \\ 0{,}2062 & 0{,}2062 & 0{,}2236 & 0{,}2 & 0{,}1118 & 0{,}1803 & 0{,}2121 & 0 \end{bmatrix}.$$

Le tableau ici-bas présente le résultat du classement de chaque donnée selon son plus proche voisin excluant elle-même (méthodologie *leave-one-out*).

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

Deux erreurs sur huit données classées seront donc effectuées, pour un taux d'erreur de $25\,\%$.

(6) (c) Appliquez l'algorithme d'édition de Wilson aux données en utilisant un voisin ($k = 1$) et en traitant les données dans l'ordre usuel ($\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots, \mathbf{x}^8$). Donnez l'ensemble des prototypes sélectionnés résultant de cette édition.

Solution:

Donnée	Prototypes	PPV	Classement	Issue
$\mathbf{x}^1$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^2$	-1	Donnée bien classée
$\mathbf{x}^2$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^1$	-1	Donnée bien classée
$\mathbf{x}^3$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^7$	+1	Donnée mal classée, retrait de x^3 de l'ensemble des prototypes
$\mathbf{x}^4$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^7$	+1	Donnée mal classée, retrait de x^4 de l'ensemble des prototypes
$\mathbf{x}^5$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^8$	+1	Donnée bien classée
$\mathbf{x}^6$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	+1	Donnée bien classée
$\mathbf{x}^7$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^8$	+1	Donnée bien classée
$\mathbf{x}^8$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	+1	Donnée bien classée

L'ensemble des prototypes sélectionnés par l'édition de Wilson est donc :

$$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}.$$

Question 4 (40 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(4) (a) Donnez l'effet de la valeur du paramètre de la largeur de fenêtre h sur une estimation de la densité de probabilité avec une fenêtre de Parzen.

Solution: Avec une faible largeur de fenêtre (valeur h faible), chaque donnée aura un effet local important dans l'estimation de la densité, alors qu'avec une plus grande largeur de fenêtre (valeur h élevée), l'estimation sera beaucoup plus douce, mais avec une perte d'information sur les variations plus brusques de la densité dans l'espace.

(4) (b) 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.

(4) (c) Indiquez combien de couches **cachées** de neurones sont nécessaires au minimum pour faire une bonne approximation de frontières de décision de forme convexe et combien de couches cachées sont nécessaires faire une bonne approximation de frontières de décisions de forme concave.

Solution: Une couche cachée est nécessaire pour approximer une frontière de décision de forme convexe et deux couches cachées sont nécessaires pour approximer des frontières de décision de forme concave.

(4) (d) Dans les méthodes par ensemble, il a été démontré que lorsque les classifieurs formant l'ensemble ont des sorties qui sont indépendantes et identiquement distribuées (i.i.d.), le taux d'erreur de l'ensemble tend vers le taux d'erreur bayésien optimal lorsqu'on utilise un très grand nombre de classifieurs de base. Expliquez pourquoi cette hypothèse i.i.d. est forte et difficile à respecter dans la pratique.

Solution: Obtenir des classifieurs dont la distribution des sorties est indépendante et identiquement distribuée est difficile à obtenir en pratique, car elle implique que les classifieurs ne font pas des erreurs pour les mêmes données, que le taux moyen d'erreurs pour chaque donnée est le même pour toutes les données, et que la corrélation des erreurs entre chaque donnée soit pratiquement nulle. En pratique il est impossible d'obtenir de telles réponses, car selon le biais inductif des classifieurs, des données sont plus difficiles que d'autres pour tous les classifieurs et que le classement de certaines données est fortement corrélé avec le classement d'autres données, par leur grande similarité ou la contradiction entre celles-ci.

(4) (e) Soit la matrice de décision suivante, correspond à un code à correction d'erreur pour la prise de décision d'un ensemble de dix classifieurs de base à deux classes (sortie -1 ou $+1$), traitant des données organisées selon trois classes.

$$\mathbf{W} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \end{bmatrix}$$

Évaluez le nombre maximum d'erreurs faites par les classifieurs de base que cet ensemble peut tolérer sans faire d'erreur de classement. Justifiez brièvement votre réponse.

Solution: Cet ensemble est tolérant à deux erreurs de classifieurs de base, car la distance de Hamming entre chaque paire de lignes de la matrice est au minimum de 5. Pour chaque erreur, la distance de Hamming est réduite de deux, de sorte qu'avec trois erreurs ou plus, une mauvaise décision d'ensemble sera prise.

(4) (f) Expliquez pourquoi dit-on que l'algorithme d'apprentissage par ensemble *Bagging* est passif, alors que l'algorithme *Boosting* (incluant les variantes *AdaBoost*) est actif.

Solution: L'algorithme *Bagging* est passif, car il s'appuie sur l'instabilité des classifieurs de base pour générer de la diversité dans l'ensemble, en entraînant ces classifieurs sur des ensembles d'entraînement produit entièrement aléatoirement. L'algorithme *Boosting* est actif, car les ensembles générés pour l'entraînement sont biaisés vers le choix de données mal classées par les autres classifieurs produits jusqu'à présent.

(4) (g) Dans les plans d'expérimentations, on désigne les facteurs incontrôlables comme étant des éléments dont on n'a pas le contrôle et dont on veut éliminer l'impact sur les décisions. Donnez un exemple d'un facteur incontrôlable spécifique à des plans d'expérimentations en apprentissage supervisé.

Solution: Le partitionnement aléatoire d'un jeu de données (ex. pour la validation croisée) est un facteur incontrôlable qui peut faire varier les performances. Les poids initiaux aléatoires d'un algorithme d'apprentissage stochastique, comme en retrouve dans le perceptron multicouche, est un autre exemple de facteur incontrôlable.

(4) (h) Expliquez ce que l'on vise à évaluer avec le test statistique ANOVA présenté en classe. Expliquez également l'hypothèse qui y est testée.

Solution: Le test ANOVA évalue si différentes configurations donnent des résultats équivalents ou non. L'hypothèse testée consiste à vérifier si deux façons différentes de calculer la variance des résultats sont équivalentes. La première façon consiste à calculer la variance en supposant que la moyenne des résultats est la même pour toutes les configurations. La deuxième façon consiste à calculer la variance en supposant que la moyenne des résultats peut être différente pour chaque configuration. Lorsque ces estimations de la variance ne concordent pas, on peut conclure qu'il y a des différences statistiquement significatives entre certaines configurations.

(4) (i) Expliquez pourquoi l'AUC-ROC est une mesure de performance intéressante lorsque l'on veut optimiser un classifieur pour des jeux de données dont les coûts pour chaque type d'erreur peuvent varier.

Solution: L'AUC-ROC est une mesure qui est indépendante du seuil de décision utilisé pour binariser une sortie réelle d'un classifieur fonctionnant pour deux classes. Cette mesure n'est donc pas spécifique à un compromis particulier du classement de données de chaque classe, et devrait donc bien fonctionner pour différents choix de coûts d'erreur de classement.

(4) (j) Expliquez la différence fondamentale entre les modèles génératifs de classement, présentés dans la première moitié du cours, et les modèles discriminatifs de classement, présentés dans la deuxième partie du cours.

Solution: Les modèles génératifs de classement visent à apprendre la distribution des données de chaque classe, donnant ainsi la probabilité *a posteriori* de classement $P(C_i|\mathbf{x})$ d'une donnée $\mathbf{x}$ à une classe C_i. Les modèles discriminatifs visent seulement à déterminer la classe à laquelle appartient une certaine donnée, sans devoir estimer les densités de probabilité des données de chaque classe.

Question 5 (15 points bonus)

Pour le projet final du cours, un étudiant propose l'approche suivante pour traiter les données MNIST :

— Des ensembles de classifieurs sont générés pour différentes configurations de classifieurs de base ;
— La prise de décision pour un ensemble s'effectue par un vote à majorité ;
— Les différentes configurations de classifieurs de base consistent en des algorithmes différents (ex. k-plus proches voisins, SVM, perceptron multicouche), ou en des valeurs différentes des hyperparamètres de ces algorithmes ;
— Pour un ensemble particulier, tous les classifieurs de base ont la même configuration ;
— Chaque classifieur de base est entraîné sur un ensemble de données différent, produit par un échantillonnage aléatoire avec remise du jeu d'entraînement d'origine (MNIST *train*) ;
— Pour évaluer la fiabilité statistique de chaque type d'ensemble, plusieurs entraînements sont effectués, avec des jeux de données à chaque fois différents ;
— La performance des ensembles de classifieurs est ensuite évaluée sur le jeu de test (MNIST *test*), en rapportant la moyenne des résultats des différents entraînements d'ensembles ayant la même configuration de classifieurs de base ;
— Une centaine de configurations de classifieurs de base sont testées, avec la sélection d'une solution finale correspondant à l'ensemble ayant le plus bas taux d'erreur sur le jeu de test.

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: La méthodologie n'est pas valide. En effet, la sélection de la configuration finale de classifieur se base sur l'évaluation d'une centaine de configurations différentes, évaluées

sur le jeu de données de test (MNIST *test*). Ce jeu de test devrait être réservé pour l'évaluation de la configuration finale seulement. L'évaluation de nombreuses configurations sur le jeu de test fait en sorte que l'on risque de sélectionner une configuration qui n'obtient de bonnes performances que par chance sur ce jeu, sans être meilleure que les autres configurations en terme de généralisation. Un exemple de bonne méthodologie serait plutôt d'utiliser un jeu de données de validation distinct, possiblement créé à partir du jeu d'entraînement, pour sélectionner la bonne configuration, qui sera ensuite la seule évaluée sur le jeu de test.