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 :

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

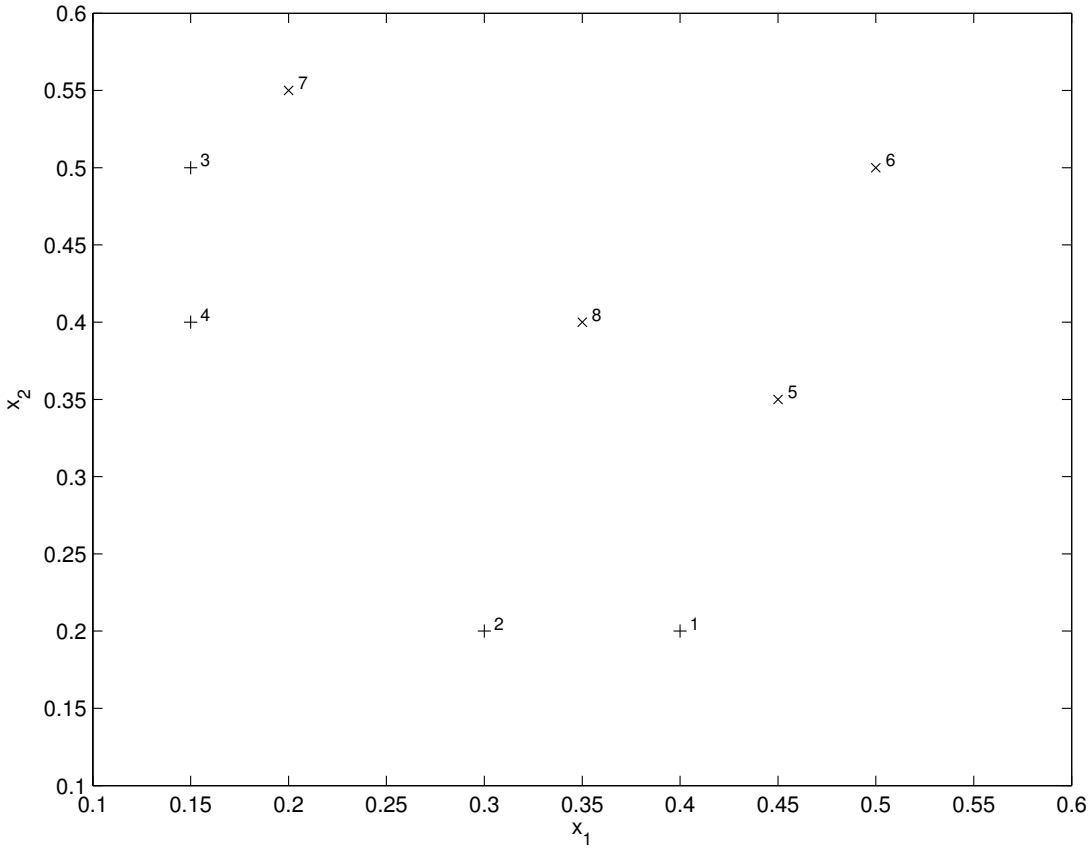
Question 2 (22 points sur 100)

Soit le jeu de données suivant, en deux dimensions :

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

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.

(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.

(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).

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.

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

(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.

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.

(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).

(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.

(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.

(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.

(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.

(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.

(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.

(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.

(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.

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.