Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2017

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 (24 points sur 100)

Supposons que l'on veut entraîner un autoencodeur à deux couches, la première couche comme encodeur et l'autre comme décodeur, toutes deux entraînées par rétropropagation des erreurs. Les poids de ces deux couches sont distincts pour les besoins de l'entraînement.

Dans ce réseau, une fonction de transfert linéaire est utilisée ($\mathrm{f}_{\mathrm{lin}}(x) = x$). La sortie d'un neurone de la couche d'encodage se modélise comme suit :

$$z_i^t = (\mathbf{w}_i^{\mathrm{enc}})^\top \mathbf{x}^t + w_{i,0}^{\mathrm{enc}}, \quad i = 1, \ldots, K.$$

La sortie de la couche de décodage se modélise comme suit, ce qui correspond à la donnée d'entrée reconstruite :

$$\hat{x}_j^t = (\mathbf{w}_j^{\mathrm{dec}})^\top \mathbf{z}^t + w_{j,0}^{\mathrm{dec}}, \quad j = 1, \ldots, D.$$

Le critère de performance utilisé est l'erreur quadratique moyenne de reconstruction :

$$e_j^t = (x_j^t - \hat{x}_j^t), \qquad E_{\mathrm{rec}}^t = \frac{1}{2}\|\mathbf{x}^t - \hat{\mathbf{x}}^t\|^2 = \frac{1}{2}\sum_{j=1}^{K}(e_j^t)^2, \qquad E_{\mathrm{rec}} = \frac{1}{N}\sum_{t=1}^{N} E_{\mathrm{rec}}^t.$$

Répondez aux questions suivantes sur l'entraînement d'un tel réseau de neurones.

(12) (a) Détaillez les équations pour mettre à jour les poids des neurones de la couche du décodeur, par descente du gradient et en utilisant l'erreur de reconstruction.

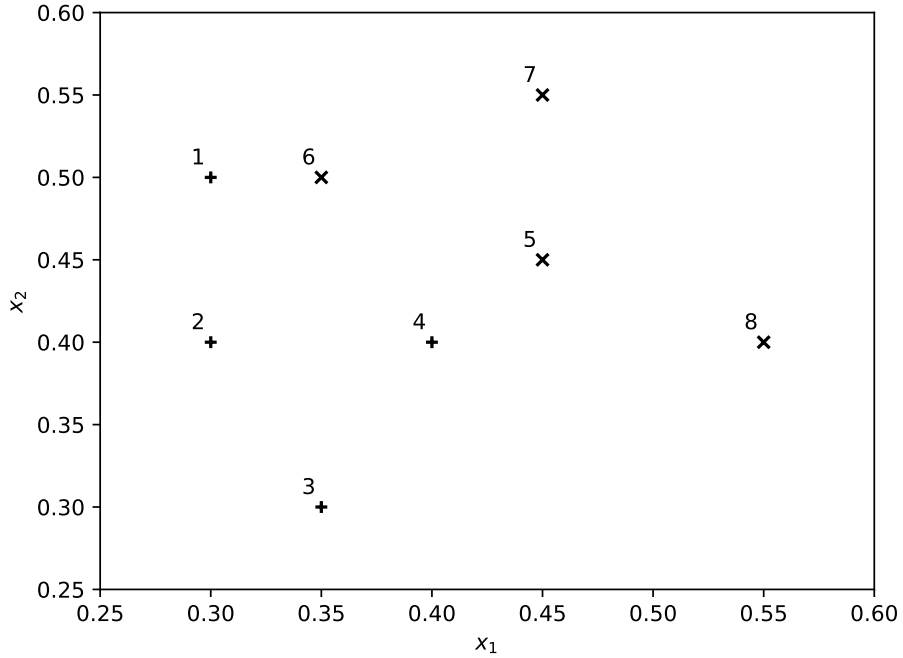
(12) (b) Détaillez maintenant les équations pour mettre à jour les poids des neurones de la couche de l'encodeur, toujours par descente du gradient et en utilisant l'erreur de reconstruction.

Question 2 (22 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 $C = 200$:

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

(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 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 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).

Question 3 (15 points sur 100)

Une matrice de décision $\mathbf{W}$, de taille $K \times L$, permet de combiner les décisions d'un ensemble de L classifieurs à deux classes, pour faire du classement de données à K classes. L'équation de décision basée sur cette matrice est la suivante :

$$\bar{\mathrm{h}}_i(\mathbf{x}) = \sum_{j=1}^{L} w_{i,j} \mathrm{h}_{j,i}(\mathbf{x}),$$

où :
— $\mathrm{h}_{j,i}(\mathbf{x})$ est le j-ème classifieur de base de l'ensemble ;
— $w_{i,j}$ est l'élément à la position (i,j) dans la matrice de décision $\mathbf{W}$;
— $\bar{\mathrm{h}}_i(\mathbf{x})$ est la décision combinée de l'ensemble pour la classe C_i.

(5) (a) Supposons que l'on veut résoudre un problème à $K = 4$ classes à l'aide d'un ensemble de classifieurs à deux classes combinés selon la méthode *un contre tous* (en anglais, *one against all*). Donnez le nombre de classifieurs à deux classes à utiliser ainsi que la matrice de décision $\mathbf{W}$ correspondant à cette configuration.

(5) (b) Supposons maintenant que l'on veut résoudre ce problème à $K = 4$ classes toujours à l'aide d'un ensemble de classifieurs à deux classes, mais cette fois en combinant les classifieurs selon la méthode de *séparation par paires* (en anglais, *pairwise separation*). Donnez le nombre de classifieurs à deux classes à utiliser ainsi que la matrice de décision $\mathbf{W}$ correspondant à cette configuration.

(5) (c) Finalement, supposons que l'on veut résoudre ce problème à $K = 4$ classes d'un ensemble redondant de $L = 9$ classifieurs, avec une matrice de décision basée sur un code à correction d'erreur (en anglais, *error code output correction*). Donnez la matrice de décision $\mathbf{W}$ correspondant à cette configuration. Déterminez également le nombre d'erreurs de classement des classifieurs de base que cette configuration de système peut tolérer sans se tromper.

Question 4 (39 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Expliquez pourquoi un SVM classique ne peut traiter que deux classes.

(3) (b) Expliquez la relation entre la fonction sigmoïde et les probabilités de classement de modèles paramétriques basés sur une loi normale.

(3) (c) Expliquez en quoi consiste le « truc du noyau » avec les SVM et autres méthodes à noyau.

(3) (d) Expliquez les principales similarités et différences entre une matrice de covariance et une matrice de Gram.

(3) (e) Expliquez pourquoi une fonction seuil ne peut pas être utilisée comme fonction de transfert de neurones dans un perceptron multicouche entraîné par rétropropagation des erreurs.

(3) (f) Expliquez l'intérêt d'effectuer de la compositionnalité de modèles dans des approches telles que l'apprentissage profond.

(3) (g) Expliquez en quoi le préentraînement non supervisé de réseau de neurones profonds était utile et nécessaire, avant l'émergence des techniques modernes d'apprentissage profond.

(3) (h) Présentez un avantage important observé avec l'apprentissage multitâches avec des réseaux de neurones profonds, outre le fait que cela permet d'apprendre des modèles capables de faire plusieurs tâches simultanément.

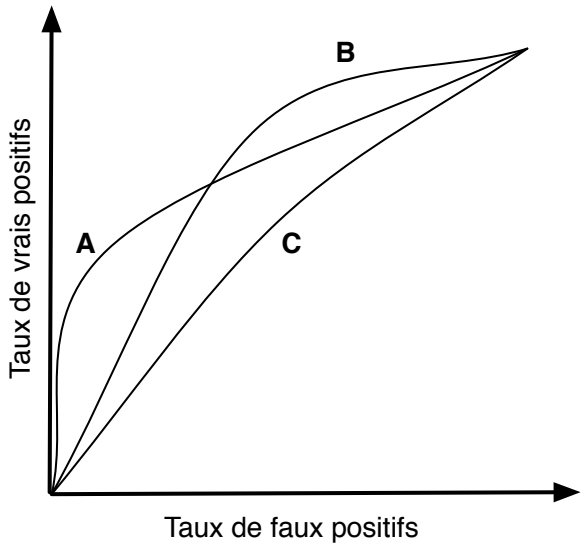
(3) (i) Présentez les avantages principaux associés à l'utilisation de graphes computationnels dans des outils d'apprentissage profonds modernes tels que TensorFlow.

(3) (j) Expliquez pourquoi un réseau de neurones profond nécessite de très grands jeux de données.

(3) (k) Expliquez pourquoi dit-on que les méthodes de *Bagging* permettent de générer de la diversité passivement, alors que les différentes approches de *Boosting* le font activement.

(3) (l) Expliquez pourquoi favorise-t-on des approches méthodologiques de type validation croisée à K-plis avec de petits jeux de données, alors qu'un partitionnement en jeux d'entraînement et de test semble suffisant avec de plus gros jeux.

(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.

Question 5 (10 points bonus)

Supposons que l'on doit choisir une méthode d'ajustement d'hyperparamètres parmi plusieurs techniques, pour un modèle d'apprentissage particulier (par exemple, paramètres C et σ d'un SVM à noyau gaussien). La comparaison expérimentale se fait sur une centaine de jeux de données standard (provenant de UCI et statlog). Présentez clairement et précisément une méthodologie expérimentale permettant de comparer les méthodes d'ajustement d'hyperparamètres et de déterminer celles offrant les meilleures performances. Prenez soin de proposer une méthode qui atténue autant que possible l'effet des facteurs de nuisance.