Introduction à l'apprentissage machine (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2018

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 30% de la note finale.

Question 1 (14 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.

(4) (a) Supposons que l'on veut résoudre un problème à $K = 5$ 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.

(4) (b) Supposons maintenant que l'on veut résoudre ce problème à $K = 5$ 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.

(6) (c) Finalement, supposons que l'on veut résoudre ce problème à $K = 5$ 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 qui maximise la robustesse des décisions d'ensemble. Déterminez également le nombre d'erreurs de classement des classifieurs de base que cette configuration de système peut tolérer sans se tromper.

Question 2 (34 points sur 100)

Supposons les données suivantes en deux dimensions :

$$\mathbf{x}^1 = \begin{bmatrix} -0{,}8 \\ 1{,}1 \end{bmatrix}, \quad \mathbf{x}^2 = \begin{bmatrix} -1{,}8 \\ 0{,}5 \end{bmatrix}, \quad \mathbf{x}^3 = \begin{bmatrix} -2{,}0 \\ 2{,}0 \end{bmatrix},$$

$$\mathbf{x}^4 = \begin{bmatrix} 1{,}5 \\ -1{,}2 \end{bmatrix}, \quad \mathbf{x}^5 = \begin{bmatrix} 0{,}5 \\ -0{,}9 \end{bmatrix}, \quad \mathbf{x}^6 = \begin{bmatrix} 2{,}0 \\ -0{,}4 \end{bmatrix}.$$

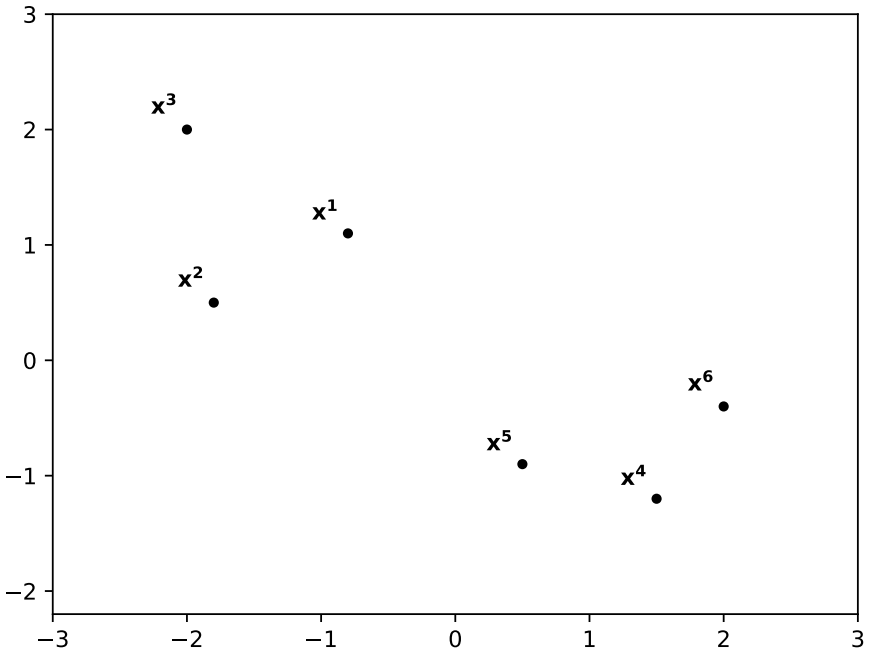
Le vecteur moyen $\mathbf{m}$ et la matrice de covariance $\mathbf{S}$ estimés de ces données sont les suivants :

$$\mathbf{m} = \begin{bmatrix} -0{,}1000 \\ 0{,}1833 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 2{,}864 & -1{,}744 \\ -1{,}744 & 1{,}534 \end{bmatrix}.$$

Les vecteurs propres et valeurs propres associés de cette matrice de covariance sont :

$$\lambda_1 = 4{,}0654, \quad \mathbf{c}_1 = \begin{bmatrix} 0{,}8235 \\ -0{,}5673 \end{bmatrix}, \quad \lambda_2 = 0{,}3323, \quad \mathbf{c}_2 = \begin{bmatrix} 0{,}5673 \\ 0{,}8235 \end{bmatrix}.$$

Finalement, les données sont tracées dans la figure suivante.



(8) (a) Donnez l'équation permettant d'effectuer une transformation blanchissante de ces données, sous la forme d'une équation linéaire ayant la formulation suivante :

$$\mathbf{z} = \mathbf{A}\,\mathbf{x} + \mathbf{b},$$

en précisant les valeurs numériques de $\mathbf{A}$ et $\mathbf{b}$.

(16) (b) Supposons un réseau de neurones de type autoencodeur avec une couche d'encodage à K neurones, une couche de décodage et une fonction de transfert linéaire ($\mathrm{f}_{\text{lin}}(x) = x$). La sortie d'un neurone i de la couche d'encodage se modélise comme suit :

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

La sortie de la couche de décodage se modélise comme suit, ce qui correspond à la donnée d'entrée reconstruite selon les D dimensions de l'espace d'origine :

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

Une analyse en composantes principales (ACP) peut être utilisée comme modèle de base pour estimer les paramètres de ce modèle. Expliquez en mots comment peut-on utiliser l'ACP pour déterminer les valeurs numériques du réseau autoencodeur.
Déterminez également les valeurs précises de l'autoencodeur ($\mathbf{w}_i^{\text{enc}}$, $w_{i,0}^{\text{enc}}$, $\mathbf{w}_j^{\text{dec}}$ et $w_{j,0}^{\text{dec}}$) devant être utilisées selon les informations données en préambule de la question, si l'encodage se fait pour une valeur de $K = 1$.

(10) (c) Exécutez l'algorithme K-means (avec $K = 2$ groupes) à partir d'une initialisation utilisant les données $\mathbf{x}^2$ et $\mathbf{x}^6$ comme centres initiaux, jusqu'à convergence de l'algorithme.

Question 3 (52 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(4) (a) Dans le cours, le fonctionnement remarquable de l'apprentissage profond est expliqué par sa capacité à faire de la composition de modèles. Expliquez en quoi consiste la « compositionnalité » dans ce contexte. Expliquez également pourquoi une méthode basée sur les distances telles que le classement par le plus proche voisin ne permet pas d'exploiter la compositionnalité.

(4) (b) Expliquez en quoi consiste la méthode d'entraînement dropout avec les réseaux profonds.

(4) (c) Un avantage certain des outils logiciels modernes pour faire des réseaux de neurones profonds est l'utilisation de gradient automatique. Expliquez en quoi ceci consiste précisément, en mentionnant l'impact pour la programmation de réseaux profonds et l'efficacité des traitements informatiques.

(4) (d) Classez en ordre chronologique d'apparition les architectures suivantes de réseaux de neurones : AlexNet, ResNet, perceptron multicouche, LeNet-5.

(4) (e) Expliquez une façon simple d'obtenir l'incertitude sur une décision faite avec une forêt aléatoire.

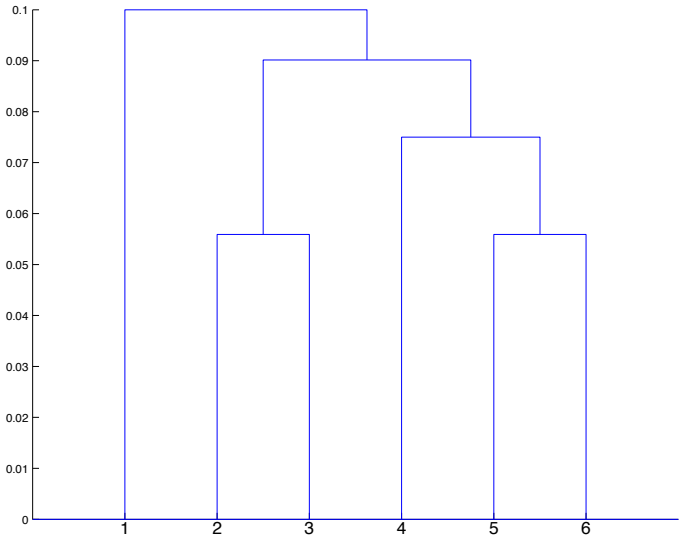
(4) (f) Expliquez pourquoi dit-on que le *bagging* induit de façon passive de la diversité entre les membres formant les ensembles, alors que AdaBoost le fait de façon active.

(4) (g) Expliquez pourquoi la standardisation des données, qui ramène la distribution des données pour chaque variable à une moyenne nulle et écart-type unitaire, n'est pas équivalente à une transformation blanchissante.

(4) (h) Dans le cours, les méthodes de sélection séquentielle avant et arrière pour la sélection de caractéristiques sont présentées comme étant des heuristiques. Expliquez ce que cela signifie en définissant le terme *heuristique* et indiquez les implications pour la sélection de caractéristiques en général.

(4) (i) Selon la présentation faite en classe de l'algorithme EM, expliquez avec précision ce à quoi consistent les appartenances h_i^t des données. Expliquez la distinction entre ces variables h_i^t et les variables cachées z_i^t de l'algorithme.

(4) (j) Soit le dendrogramme donné ici-bas.



Supposons que l'on veut former quatre groupes, donnez les indices des instances formant chacun de ces groupes.

(4) (k) Expliquez la différence entre le problème de surapprentissage et le problème de surrecherche en apprentissage machine.

(4) (l) Expliquez dans quelles circonstances une recherche aléatoire d'hyperparamètres semble mieux fonctionner qu'une recherche en grille, avec un nombre égal de configurations testées.

(4) (m) Expliquez pourquoi dit-on qu'avec un nombre suffisamment grand de données et un modèle d'une complexité suffisante, on peut obtenir des performances en généralisation se rapprochant du taux d'erreur bayésien optimal. Expliquez également pourquoi il n'est pas possible d'obtenir un meilleur taux de classement que le taux d'erreur bayésien optimal.