Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2010

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.

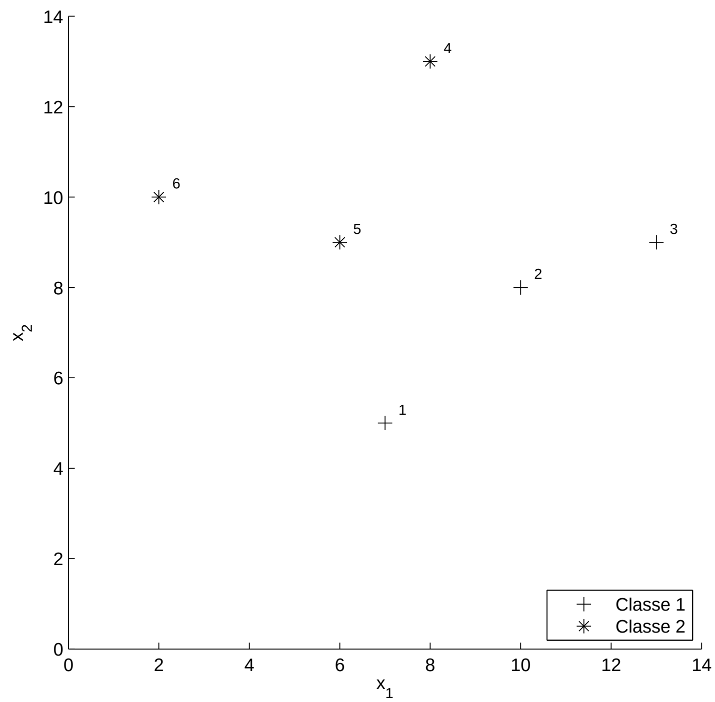
Question 1 (15 points sur 100)

Soit le jeu de données $\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^6$ présenté ci-bas.

$$\mathbf{x^1} = \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \quad r^1 = -1, \quad \mathbf{x^2} = \begin{bmatrix} 10 \\ 8 \end{bmatrix}, \quad r^2 = -1, \quad \mathbf{x^3} = \begin{bmatrix} 13 \\ 9 \end{bmatrix}, \quad r^3 = -1,$$

$$\mathbf{x^4} = \begin{bmatrix} 8 \\ 13 \end{bmatrix}, \quad r^4 = 1, \quad \mathbf{x^5} = \begin{bmatrix} 6 \\ 9 \end{bmatrix}, \quad r^5 = 1, \quad \mathbf{x^6} = \begin{bmatrix} 2 \\ 10 \end{bmatrix}, \quad r^6 = 1$$

La figure suivante trace ces points en deux dimensions.



Supposons que l'on veut classer ces données avec un classifieur de type séparateur à vastes marges (SVM) utilisant un noyau linéaire ($K(\mathbf{x},\mathbf{x}') = \langle \mathbf{x},\mathbf{x}' \rangle$), sans marge floue.

(5) (a) Dans votre **cahier de réponse**, tracez les données du jeu $\mathcal{X}$, les marges géométriques maximales obtenues avec le SVM, l'hyperplan séparateur correspondant, et encerclez les données agissant comme vecteurs de support.

(10) (b) Donnez les valeurs des poids $\mathbf{w}$ et biais w_0 correspondant au discriminant linéaire maximisant les marges géométriques tracées en a).
Indice : il n'est pas nécessaire de calculer les α^i pour répondre à la question.

Question 2 (10 points sur 100)

Soit le classifieur à noyau suivant, utilisant le critère d'erreur du perceptron.

$$\begin{aligned}
\mathrm{h}(\mathbf{x}|\boldsymbol{\alpha},w_0) &= \sum_{\mathbf{x}^s \in \mathcal{X}} \alpha^s r^s K(\mathbf{x}^s,\mathbf{x}) + w_0, \\
\alpha^t &\geq 0 \quad \forall t, \\
\text{avec :} & \\
E(\boldsymbol{\alpha},w_0|\mathcal{X}) &= -\sum_{\mathbf{x}^t \in \mathcal{Y}} r^t \mathrm{h}(\mathbf{x}^t|\boldsymbol{\alpha},w_0) + \lambda \frac{1}{2} \sum_{\alpha^t \in \boldsymbol{\alpha}} (\alpha^t)^2, \\
\mathcal{Y} &= \{\mathbf{x}^t \in \mathcal{X} | r^t \mathrm{h}(\mathbf{x}^t|\boldsymbol{\alpha},w_0) < 0\}.
\end{aligned}$$

Donnez les équations pour mettre à jour les $\boldsymbol{\alpha}$ et w_0 selon une descente du gradient.

Question 3 (15 points sur 100)

Dans le cours, il a été avancé qu'un perceptron multi-couches (PMC) avec plusieurs couches de neurones et utilisant une fonction de transfert linéaire ($\mathrm{f}_{lin}(a) = a$) pour tous les neurones peut être simplifié comme PMC à une seule couche de neurones avec fonction de transfert linéaire. Démontrez que cette affirmation est vraie à partir des équations du PMC modélisant la propagation des données de l'entrée vers la sortie.

Question 4 (15 points sur 100)

La fonction $\mathrm{h}_{j,i}$ correspond à la décision du classifieur h_j concernant la classe C_i. Ainsi, pour un problème à trois classes, les fonctions $\mathrm{h}_{j,1}$, $\mathrm{h}_{j,2}$ et $\mathrm{h}_{j,3}$ retournent les valeurs de la décision du classifieur h_j pour les classes C_1, C_2 et C_3, respectivement.

(5) (a) Supposons que l'on bâtit un ensemble de $L = 5$ classifieurs prenant des décisions sur $K = 3$ classes. La fonction $\bar{\mathrm{h}}_i$ retourne le résultat de l'ensemble pour la classe C_i. Pour une donnée $\mathbf{x}$ particulière, on obtient les résultats suivants avec cet ensemble.

	C_1	C_2	C_3
$\mathrm{h}_{1,i}(\mathbf{x})$	0,3	0,5	0,45
$\mathrm{h}_{2,i}(\mathbf{x})$	0,4	0,4	0,35
$\mathrm{h}_{3,i}(\mathbf{x})$	0,45	0,6	0,5
$\mathrm{h}_{4,i}(\mathbf{x})$	0,1	0,2	0,15
$\mathrm{h}_{5,i}(\mathbf{x})$	0,3	0,2	0,3

Pour chacune des fonctions de combinaison suivantes, utilisées pour calculer le résultat de la fonction de l'ensemble par classe $\bar{\mathrm{h}}_i$, donnez les décisions de classement pour la donnée $\mathbf{x}$:

1. somme ;
2. maximum ;
3. médiane.

(5) (b) Supposons maintenant que les décisions sont binaires, de sorte que les valeurs pour chaque classe des classifieurs de l'ensemble sont $\mathrm{h}_{j,i}(\mathbf{x}) \in \{0,1\}$, et que l'on obtienne les résultats suivants pour la donnée $\mathbf{x}$ pour un ensemble de $L = K = 3$ classifieurs.

	C_1	C_2	C_3
$\mathrm{h}_{1,i}(\mathbf{x})$	0	1	1
$\mathrm{h}_{2,i}(\mathbf{x})$	0	1	0
$\mathrm{h}_{3,i}(\mathbf{x})$	0	0	1

Si cet ensemble est entraîné selon une approche revenant à *un contre tous*, avec la matrice de décision suivante,

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

donnez la décision de classement pour cette donnée.

(5) (c) Supposons maintenant que l'on veut utiliser une approche avec code de correction d'erreur comportant huit classifieurs. On a déjà établi les sept première colonnes de la matrice de décision $\mathbf{W}$ comme suit :

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

Déterminez les valeurs sur la dernière colonnes de la matrice de décision permettant de tolérer jusqu'à deux erreurs par les classifieurs de base.

Question 5 (45 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Dans les séparateurs à vaste marge (SVM), utilisant la formulation avec marge douce, que représente une valeur de la variable *slack* ξ^t lorsqu'elle est comprise dans $0 < \xi^t \leq 1$?

(3) (b) Avec l'analyse en composantes principales à noyau, on effectue une extraction des vecteurs propres et valeurs propres de la matrice de Gram normalisée. Indiquez combien de valeurs comporte la première composante principale extraite à l'aide de cette approche.

(3) (c) Indiquez précisemment quel élément de l'algorithme de rétropropagation des erreurs du perceptron multi-couches fait en sorte qu'on le qualifie d'algorithme stochastique.

(3) (d) Dans l'algorithme de rétropropagation des erreurs du perceptron multi-couches, on utilise la règle de chaînage des dérivées pour calculer la correction à appliquer aux poids et biais du réseau. Par exemple, la correction de poids sur une couche de sortie se calcule à partir des dérivées partielles données dans l'équation suivante,

$$\frac{\partial E^t}{\partial w_{j,i}} = \frac{\partial E^t}{\partial e_j^t} \frac{\partial e_j^t}{\partial y_j^t} \frac{\partial y_j^t}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{j,i}}.$$

Quelle est la valeur de la dérivée partielle $\partial a_j^t / \partial w_{j,i}$ dans cette équation.

(3) (e) Dans les méthodes par ensemble, il est démontré que la variance des performance d'un ensemble $\bar{\mathrm{h}}$ formé de L classifieurs individuels h_j décroît selon la taille de l'ensemble selon

$$\mathrm{Var}(\bar{\mathrm{h}}) = \frac{1}{L}\mathrm{Var}(\mathrm{h}_j).$$

Cette formulation fait cependant que la réponse des classifieurs individuels h_j respecte l'hypothèse iid (identiquement et indépendamment distribués). Indiquez en vos propres mots ce que signifie cette hypothèse dans le contexte présent de classement avec ensembles.

(3) (f) Avec l'algorithme AdaBoost, on modifie une probabilité p_j^t qu'une donnée $\mathbf{x}^t$ soit échantillonnée pour entraîner un classifieur à l'itération j. Indiquez de quelle façon, d'un point de vue conceptuel, cette probabilité est modifiée à chaque itération de l'algorithme.

(3) (g) Lorsque l'on fait des expérimentations avec des algorithmes d'apprentissage supervisé pour faire du classement, on effectue souvent du partitionnement des ensembles de données avec **stratification**, où l'on respecte les proportions des données selon les différentes classes du problème (probabilités *a priori*). Indiquez pourquoi cette approche est souhaitable dans un contexte d'expérimentation et d'analyse, comparativement à un partitionnement sans stratification.

(3) (h) Dans le test de l'Analyse de variance (ANOVA), on veut comparer plusieurs algorithmes de classement, en tenant de vérifier l'hypothèse H_0 à l'effet que les moyennes des performances μ_j pour chaque classifieur sont égales. Pour vérifier cette hypothèse, on calcule deux estimateurs σ^2 de la variance des résultats pour chaque classifieur. Indiquez clairement ce que sont chacun de ces estimateurs de la variance et de quelle façon on les utilise pour déterminer si l'hypothèse H_0 est valide.

(3) (i) Indiquez précisemment les variables formant le modèle λ d'un modèle de Markov caché.

(3) (j) Selon une méthode d'évaluation des performances de type *leave-one-out*, indiquez combien de fois une données particulière $\mathbf{x}^t \in \mathcal{X}$ de l'ensemble sera utilisée pour entraîner le classifieur évalué.

(3) (k) Dans le problème d'évaluation avec un modèle de Markov caché, on veut évaluer la probabilité $P(O|\lambda)$ d'avoir un certain séquence d'observations O avec le modèle λ. Cette probabilité peut se calculer selon l'équation suivante,

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

Cependant, le calcul de cette probabilité n'est pas tractable, computationnellement parlant. Indiquez comment on doit procéder pour évaluer la probabilité $P(O|\lambda)$ selon la méthode vue en classe, qui comporte une complexité algorithmique raisonnable.

(3) (l) L'algorithme Baum-Welch permet de calculer le modèle λ d'un modèle de Markov caché à partir d'observations. Cet algorithme implique le calcul d'une probabilité $\xi_t(i,j)$. Indiquez ce que signifie précisemment cette probabilité.

(3) (m) Indiquez ce que représente précisément $Q(s,a)$ dans un contexte d'apprentissage par renforcement.

(3) (n) Selon Bellman, le valeur d'une action dans un certain état est donnée par :

$$Q^*(s,a) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q^*(s',a') \right].$$

Indiquez pourquoi il n'est pas possible d'appliquer directement cette équation dans le contexte où l'on ne possède pas de modèle satisfaisant de l'environnement.

(3) (o) Expliquez de quelle façon on détermine les actions effectuées par un agent dans un contexte d'apprentissage par renforcement, lorsque l'agent utilise une politique dite ϵ-greedy.