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

UNIVERSITÉ LAVAL

EXAMEN PARTIEL

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 (20 points sur 100)

Soit la loi exponentielle, dont la densité de probabilité est donnée par l'équation suivante.

$$p(x|\lambda) = \begin{cases} \lambda \exp[-\lambda x] & x \in [0,\infty) \\ 0 & \text{autrement} \end{cases}$$

La moyenne d'une variable aléatoire X suivant cette loi de probabilité est $\mathbb{E}[X] = \frac{1}{\lambda}$.

(7) (a) Calculez l'estimateur par un maximum de vraisemblance du paramètre λ de la loi exponentielle selon un jeu $\mathcal{X} = \{x^t\}_1^N$ à une dimension comprenant N données.

(3) (b) Déterminez si l'estimateur que vous avez développé au point précédent est biaisé. Justifiez votre réponse. Indice : $\mathbb{E}[1/y] \geq 1/\mathbb{E}[y]$ lorsque $y \in [0,\infty)$

(10) (c) Supposons maintenant que l'on veut faire du classement paramétrique avec des données organisées selon deux classes, où l'on modélise les données de chaque classe comme suivant une loi exponentielle.

$$p(x|C_1) = p(x|\lambda_1), \qquad p(x|C_2) = p(x|\lambda_2)$$

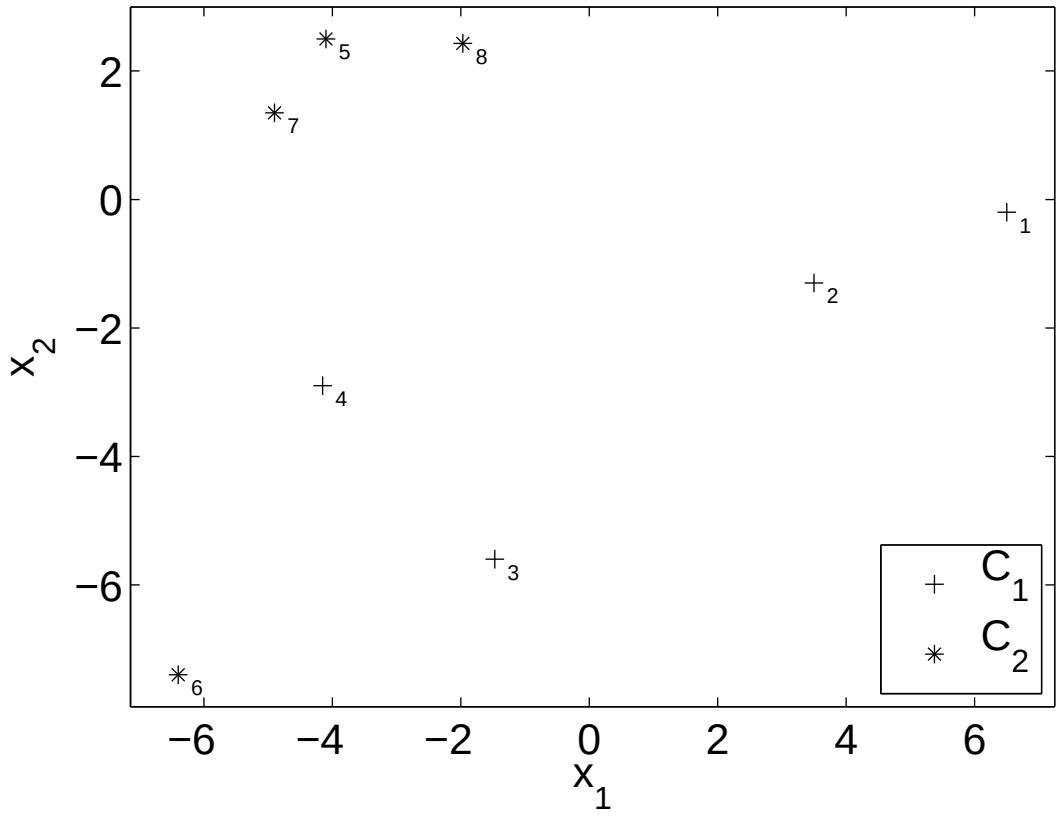
Supposons également que vous connaissez les valeurs des paramètres λ_1 et λ_2 des densités de probabilité de chaque classe. Donnez les régions de décision pour l'ensemble du domaine $x \in [0,\infty)$, c'est-à-dire les décisions de classement selon la valeur de x. Vous pouvez supposer que $\lambda_1 < \lambda_2$ et que les probabilités *a priori* sont égales, $P(C_1) = P(C_2) = 0{,}5$.

Question 2 (20 points sur 100)

Soit les données suivantes, en deux dimensions :

$$\mathbf{x}^1 = \begin{bmatrix} 6{,}5 \\ -0{,}2 \end{bmatrix}, \quad \mathbf{x}^2 = \begin{bmatrix} 3{,}5 \\ -1{,}3 \end{bmatrix}, \quad \mathbf{x}^3 = \begin{bmatrix} -1{,}47 \\ -5{,}6 \end{bmatrix}, \quad \mathbf{x}^4 = \begin{bmatrix} -4{,}15 \\ -2{,}9 \end{bmatrix},$$
$$\mathbf{x}^5 = \begin{bmatrix} -4{,}1 \\ 2{,}5 \end{bmatrix}, \quad \mathbf{x}^6 = \begin{bmatrix} -6{,}4 \\ -7{,}4 \end{bmatrix}, \quad \mathbf{x}^7 = \begin{bmatrix} -4{,}9 \\ 1{,}35 \end{bmatrix}, \quad \mathbf{x}^8 = \begin{bmatrix} -1{,}97 \\ 2{,}43 \end{bmatrix}.$$

Les données $\mathbf{x}^1$ à $\mathbf{x}^4$ appartiennent à la classe C_1, alors que les données $\mathbf{x}^5$ à $\mathbf{x}^8$ appartiennent à la classe C_2. La figure suivante présente les données.



(5) (a) Tracez la frontière de décision correspondant à un classement par la règle du plus proche voisin ($k = 1$), en utilisant une distance euclidienne.

(5) (b) Déterminez le taux d'erreur de classement sur ce jeu de données lorsque l'on fait un traitement de type *leave-one-out*, avec la règle des k plus proches voisins, en utilisant $k = 3$ voisins et une distance euclidienne.

(5) (c) Effectuez une itération de la descente du gradient basée sur critère du perceptron (mode *batch*). Utilisez les poids initiaux $\mathbf{w} = \begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix}$ et $w_0 = -1$ et un taux d'apprentissage de $\eta = 0{,}1$. Donnez les valeurs de poids $\mathbf{w}$ et w_0 résultant.

(5) (d) Tracez la frontières de décision correspondant au discriminant linéaire avec les paramètres suivants :

$$\mathrm{h}(\mathbf{x}|\mathbf{w},w_0) = \mathbf{w}^T\mathbf{x} + w_0, \quad \mathbf{w} = \begin{bmatrix} 0{,}5 \\ -0{,}5 \end{bmatrix}, \quad w_0 = 1{,}5.$$

Prenez soin d'indiquer à quelle classe appartient chaque région de décision.

Question 3 (30 points sur 100)

Supposons que l'on a un jeu de données en deux dimensions, comportant deux classes (C_1 et C_2). Les vecteurs moyens $\boldsymbol{\mu}_i$, ainsi que les valeurs propres $\lambda^{\boldsymbol{\Sigma}_i}$ et vecteurs propres $\mathbf{w}^{\boldsymbol{\Sigma}_i}$ associés à la matrice de covariance $\boldsymbol{\Sigma}_i$ de chaque classe sont les suivants :

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix}, \quad \lambda_1^{\boldsymbol{\Sigma}_1} = 2, \quad \mathbf{w}_1^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} 0{,}33 \\ 0{,}9428 \end{bmatrix}, \quad \lambda_2^{\boldsymbol{\Sigma}_1} = 0{,}8, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} 0{,}9428 \\ -0{,}33 \end{bmatrix},$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \quad \lambda_1^{\boldsymbol{\Sigma}_2} = 0{,}65, \quad \mathbf{w}_1^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} -0{,}7746 \\ 0{,}6325 \end{bmatrix}, \quad \lambda_2^{\boldsymbol{\Sigma}_2} = 4{,}5, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}6325 \\ 0{,}7746 \end{bmatrix}.$$

Les probabilités *a priori* des classes sont respectivement $P(C_1) = 0{,}6$ et $P(C_2) = 0{,}4$.

Pour rappel, l'inverse d'une matrice $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ est $\mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

(5) (a) Pour chaque classe, donnez la transformation linéaire nécessaire pour conserver 80% de la variance. Traitez les données de chaque classe indépendamment.

(5) (b) Pour chacune des deux classes, calculez la matrice de covariance associée ($\boldsymbol{\Sigma}_1$ et $\boldsymbol{\Sigma}_2$).

(5) (c) Calculez la matrice de covariance partagée (quelconque, avec valeurs hors la diagonale) par les deux classes ($\boldsymbol{\Sigma}$).

(5) (d) Pour chacune des classes, tracez les courbes de contour correspondant à une distance de Mahalanobis de 1 du vecteur moyen des densités de probabilité.

(5) (e) Donnez l'équation avec valeurs numériques des variables de la transformation linéaire correspondant à une analyse discriminante linéaire de ces données.

(5) (f) Donnez l'équation du discriminant linéaire correspondant à un classifieur à la plus proche moyenne de ces données.

Question 4 (30 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Lorsque l'on veut faire une sélection aggressive de prototypes pour le classement avec la règles des k plus proches voisins, on peut effectuer une édition de Wilson suivie d'une condensation de Hart. Expliquez pourquoi procéder dans l'ordre inverse, soit en faisant une condensation de Hart suivie d'une édition de Wilson, est une mauvaise idée et risque de donner de mauvais résultats.

(3) (b) On dit que l'algorithme K-means fait une minimisation de l'erreur de reconstruction donnée par l'équation suivant.

$$E[\{\mathbf{m}_i\}_{i=1}^K|\mathcal{X}] = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$$

Expliquez en quoi ce critère est pertinent afin d'effectuer du clustering de données.

(3) (c) Donnez la différence principale entre les tâches de régression et de classement dans un contexte d'apprentissage supervisé.

(3) (d) Pour un problème de classement donné, quel est l'effet de l'augmentation du nombre de données utilisées dans le jeu d'entraînement sur l'erreur de classement en généralisation. Justifiez brièvement votre réponse.

(3) (e) Supposons que l'on fait une validation croisée à K plis (K-*fold cross-validation*) sur un jeu comprenant N données différentes, indiquez le nombre de données moyen utilisé pour chaque entraînement de classifieur, pour chaque plis.

(3) (f) Indiquez précisemment ce que représente concrètement la vraisemblance $p(\mathbf{x}|C_i)$ dans un contexte de classement paramétrique.

(3) (g) Qu'elles sont les valeurs sur la diagonale d'une matrice de confusion, soit les valeurs de la fonction de perte $\mathcal{L}(\alpha_i,C_i)$?

(3) (h) Dans l'algorithme EM présenté en classe, on a défini que $h_i^t \equiv \mathbb{E}[z_i^t|\mathcal{X},\Phi^l]$. Expliquez ce que cela signifie en termes claires et précis.

(3) (i) Donnez les variables formant une paramétrisation Φ de l'algorithme EM pour une densité-mélange de groupes suivant des lois multinormales.

(3) (j) Pourquoi dit-on que l'on ne peut jamais faire mieux que le taux d'erreur bayésien optimal pour un certain problème, même si la valeur de ce taux d'erreur est rarement nulle.