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.

Solution:

$$\begin{aligned} L(\lambda|\mathcal{X}) &= \log \prod_{t=1}^{N} p(x|\lambda) = \sum_{t=1}^{N} \log p(x^t|\lambda) = \sum_{t=1}^{N} \log \lambda \exp[-\lambda x^t] \\ &= \sum_{t=1}^{N} \log \lambda + \sum_{t=1}^{N} (-\lambda x^t) = N \log \lambda - \lambda \sum_{t=1}^{N} x^t \\ \frac{\partial L(\lambda|\mathcal{X})}{\partial \lambda} &= \frac{\partial}{\partial \lambda} N \log \lambda - \lambda \sum_{t=1}^{N} x^t = \frac{N}{\lambda} - \sum_{t=1}^{N} x^t = 0 \\ \frac{N}{\lambda} &= \sum_{t=1}^{N} x^t \Rightarrow \lambda = \frac{N}{\sum_{t=1}^{N} x^t} \end{aligned}$$

L'estimateur selon un maximum de vraisemblance du paramètre λ est donc $\hat{\lambda} = \frac{N}{\sum_{t=1}^{N} x^t}$.

(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 $y \in [0,\infty)$

Solution: Le biais d'un estimateur est calculé par la formule suivante.

$$b_\theta(d) = \mathbb{E}_\mathcal{X}(d) - \theta$$

Le biais dans le cas présent est :

$$\begin{aligned} b_\lambda(\hat{\lambda}) &= \mathbb{E}_\mathcal{X}(\hat{\lambda}) - \lambda = \mathbb{E}_\mathcal{X}\left(\frac{N}{\sum_{t=1}^{N} x^t}\right) - \lambda = \mathbb{E}_\mathcal{X}\left(\frac{1}{\mathbb{E}[x]}\right) - \lambda \\ &\geq \frac{1}{\mathbb{E}[x]} - \lambda = \lambda - \lambda = 0 \\ b_\lambda(\hat{\lambda}) &\geq 0 \end{aligned}$$

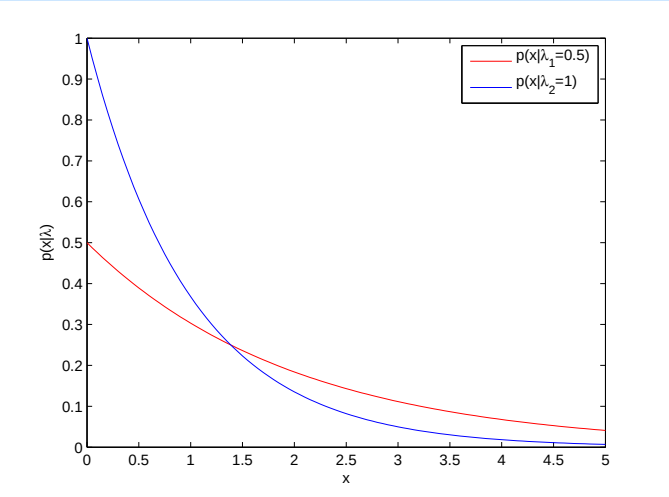
Comme le biais n'est pas nul dans le cas général, l'estimateur est biaisé.

(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, $P(C_1) = P(C_2) = 0{,}5$.

Solution: La figure suivante présente la densité de probabilité pour une valeur de $\lambda_1 < \lambda_2$.



On voit qu'il y a un point d'intersection d, de sorte que $p(x|\lambda_1) < p(x|\lambda_2)$ lorsque $x < d$ et $p(x|\lambda_1) > p(x|\lambda_2)$ lorsque $x > d$. Il suffit donc de calculer ce point d'intersection,

qui survient lorsque $p(x|\lambda_1) = p(x|\lambda_2)$.

$$\begin{aligned} p(x|\lambda_1) &= p(x|\lambda_2) \\ \lambda_1 \exp[-\lambda_1 x] &= \lambda_2 \exp[-\lambda_2 x] \\ \log \lambda_1 \exp[-\lambda_1 x] &= \log \lambda_2 \exp[-\lambda_2 x] \\ \log \lambda_1 - \lambda_1 x &= \log \lambda_2 - \lambda_2 x \\ (\lambda_1 - \lambda_2)x &= \log \lambda_1 - \log \lambda_2 = \log \frac{\lambda_1}{\lambda_2} \\ x &= \frac{\log \frac{\lambda_1}{\lambda_2}}{\lambda_1 - \lambda_2} \end{aligned}$$

Donc, les région de classement serait les suivantes :

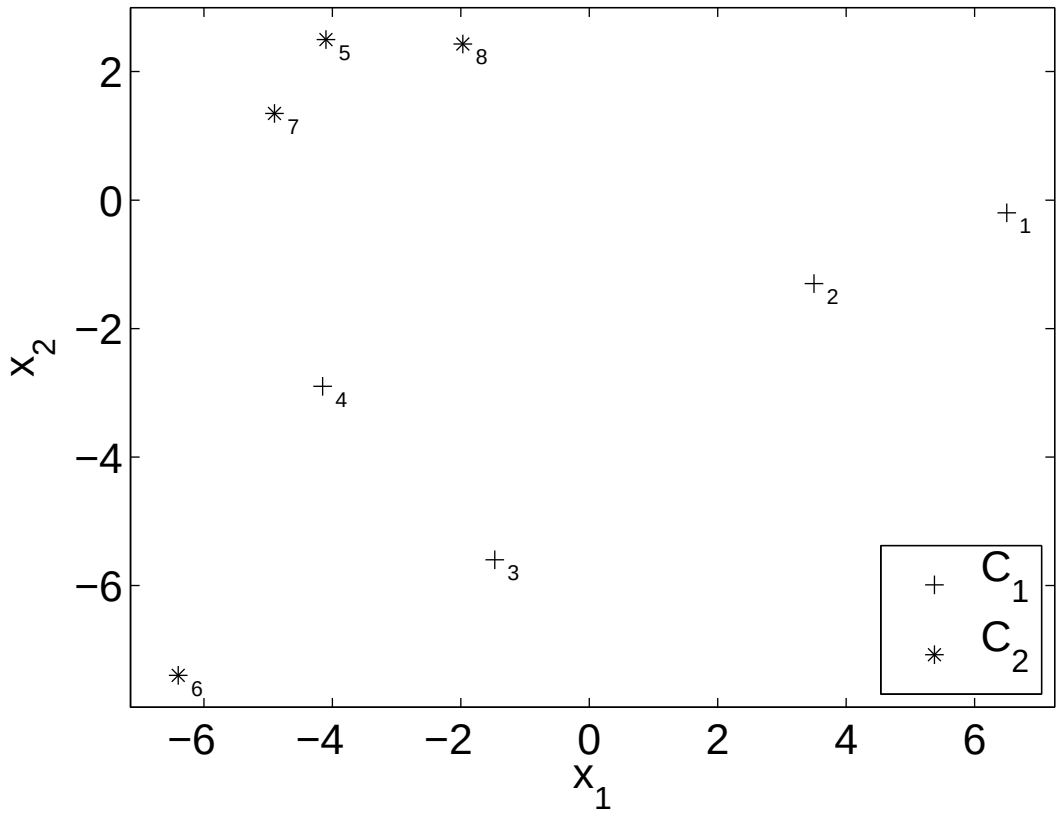
$$x \in \begin{cases} C_1 & \text{si } x > \frac{\log \frac{\lambda_1}{\lambda_2}}{\lambda_1 - \lambda_2} \\ C_2 & \text{autrement} \end{cases} .$$

Question 2 (20 points sur 100)

Soit les données suivantes, en deux dimensions :

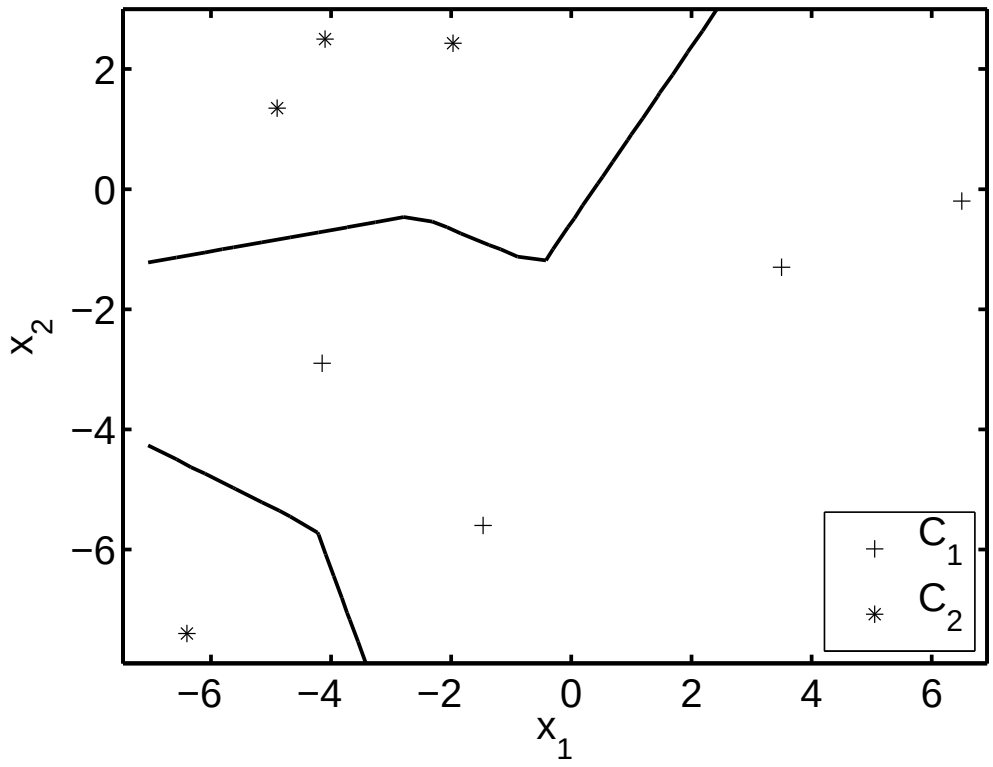
$$\mathbf{x}^1 = \begin{bmatrix} 6{,}5 \\ -0{,}2 \end{bmatrix}, \quad \mathbf{x}^2 = \begin{bmatrix} 3{,}5 \\ -1{,}3 \end{bmatrix}, \quad \mathbf{x}^3 = \begin{bmatrix} -1{,}47 \\ -5{,}6 \end{bmatrix}, \quad \mathbf{x}^4 = \begin{bmatrix} -4{,}15 \\ -2{,}9 \end{bmatrix},$$
$$\mathbf{x}^5 = \begin{bmatrix}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

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.

Solution: Les frontières de décision sont données dans la figure suivante.



(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.

Solution:
$\mathbf{x}^1$: plus proches voisins sont $\mathbf{x}^2$, $\mathbf{x}^8$ et $\mathbf{x}^3$; assigné à la classe C_1 ; bien classé.
$\mathbf{x}^2$: plus proches voisins sont $\mathbf{x}^1$, $\mathbf{x}^3$ et $\mathbf{x}^8$; assigné à la classe C_1 ; bien classé.
$\mathbf{x}^3$: plus proches voisins sont $\mathbf{x}^4$, $\mathbf{x}^6$ et $\mathbf{x}^2$; assigné à la classe C_1 ; bien classé.
$\mathbf{x}^4$: plus proches voisins sont $\mathbf{x}^3$, $\mathbf{x}^7$ et $\mathbf{x}^6$; assigné à la classe C_2 ; mal classé.
$\mathbf{x}^5$: plus proches voisins sont $\mathbf{x}^7$, $\mathbf{x}^8$ et $\mathbf{x}^4$; assigné à la classe C_2 ; bien classé.
$\mathbf{x}^6$: plus proches voisins sont $\mathbf{x}^4$, $\mathbf{x}^3$ et $\mathbf{x}^7$; assigné à la classe C_1 ; mal classé.
$\mathbf{x}^7$: plus proches voisins sont $\mathbf{x}^5$, $\mathbf{x}^8$ et $\mathbf{x}^4$; assigné à la classe C_2 ; bien classé.
$\mathbf{x}^8$: plus proches voisins sont $\mathbf{x}^5$, $\mathbf{x}^7$ et $\mathbf{x}^4$; assigné à la classe C_2 ; bien classé.
La taux d'erreur de classement est donc de $\frac{2}{8} = 0{,}25$.

(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.

Solution: La descente du gradient selon le critère du perceptron est donnée par les équations suivantes.

$$\begin{aligned}\Delta\mathbf{w} &= \sum_{\mathbf{x}^t\in\mathcal{Y}} \eta r^t \mathbf{x}^t \\ \Delta w_0 &= \sum_{\mathbf{x}^t\in\mathcal{Y}} \eta r^t\end{aligned}$$

$\mathbf{x}^1$: $\mathrm{h}(\mathbf{x}^1) = 2{,}15$; bien classé
$\mathbf{x}^2$: $\mathrm{h}(\mathbf{x}^2) = 0{,}1$; bien classé
$\mathbf{x}^3$: $\mathrm{h}(\mathbf{x}^3) = -4{,}54$; mal classé
$\mathbf{x}^4$: $\mathrm{h}(\mathbf{x}^4) = -4{,}53$; mal classé
$\mathbf{x}^5$: $\mathrm{h}(\mathbf{x}^5) = -1{,}8$; bien classé
$\mathbf{x}^6$: $\mathrm{h}(\mathbf{x}^6) = -7{,}9$; bien classé
$\mathbf{x}^7$: $\mathrm{h}(\mathbf{x}^7) = -2{,}78$; bien classé
$\mathbf{x}^8$: $\mathrm{h}(\mathbf{x}^8) = -0{,}77$; bien classé
L'ensemble des données mal classées est donc $\mathcal{Y} = \{\mathbf{x}^3,\mathbf{x}^4\}$. La correction a appliquer sera donc :

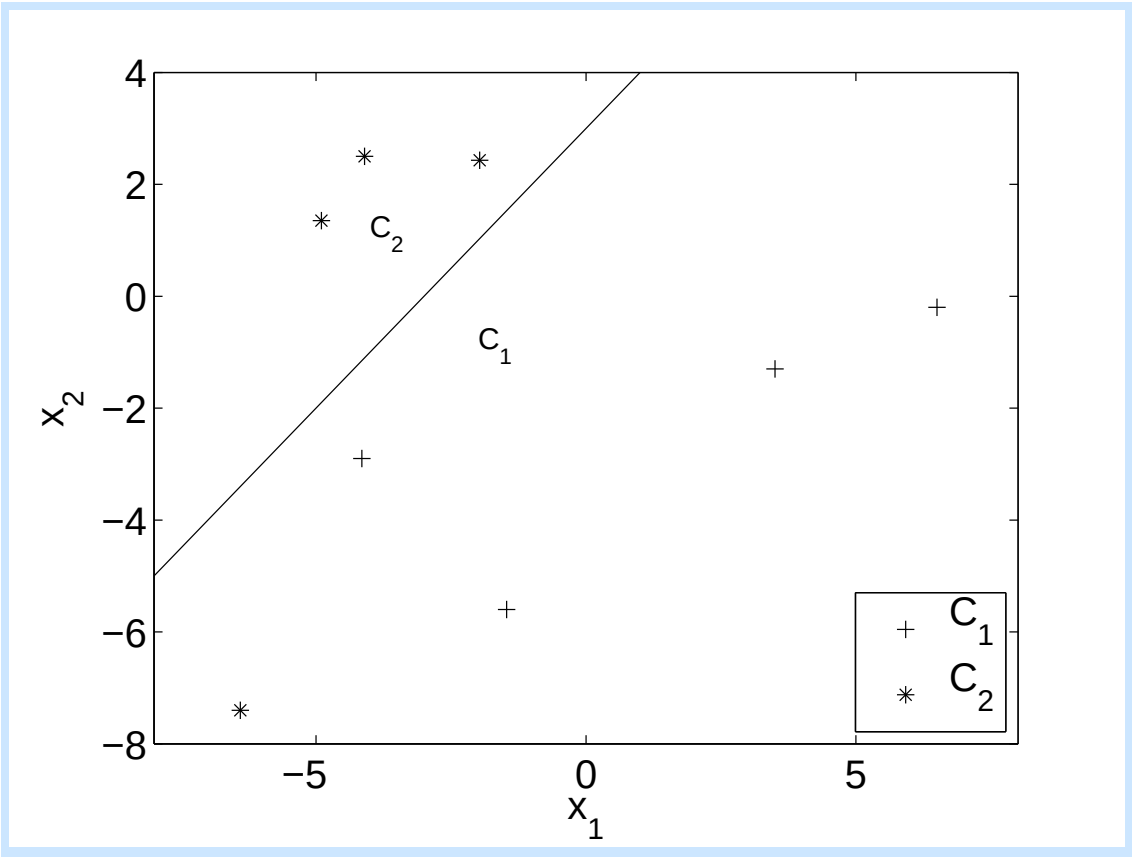
$$\begin{aligned}\Delta\mathbf{w} &= \sum_{\mathbf{x}^t\in\mathcal{Y}} \eta r^t \mathbf{x}^t \\ &= 0{,}1\left(1\begin{bmatrix}-1{,}47\\-5{,}6\end{bmatrix} + 1\begin{bmatrix}-4{,}15\\-2{,}9\end{bmatrix}\right) = \begin{bmatrix}-0.562\\-0.85\end{bmatrix} \\ \Delta w_0 &= \sum_{\mathbf{x}^t\in\mathcal{Y}} \eta r^t = 0.1(1+1) = 0.2 \\ \mathbf{w} &= \mathbf{w} + \Delta\mathbf{w} = \begin{bmatrix}0.5\\0{,}5\end{bmatrix} + \begin{bmatrix}-0.562\\-0.85\end{bmatrix} = \begin{bmatrix}-0.062\\-0.35\end{bmatrix} \\ w_0 &= w_0 + \Delta w_0 = -1 + 0.2 = -0.8\end{aligned}$$

(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.

Solution:



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.

Solution: Pour la classe C_1, on doit utiliser les deux composantes principales, car la première composante principale ne capture que 71% de la variance ($\frac{\lambda_1^{\boldsymbol{\Sigma}_1}}{\lambda_1^{\boldsymbol{\Sigma}_1} + \lambda_2^{\boldsymbol{\Sigma}_1}} = \frac{2}{2+0{,}8} =$

0,7143). La transformation linéaire correspondante sera donc la suivante :

$$\mathbf{z} = \mathbf{W}_1^T(\mathbf{x} - \boldsymbol{\mu}_1), \quad \mathbf{W}_1 = \begin{bmatrix} 0{,}33 & 0{,}9428 \\ 0{,}9428 & -0{,}33 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix}.$$

Pour la classe C_2, on peut se limiter la première composante principale, qui correspond à la deuxième paire de valeurs/vecteurs propres, comme cette composante capture 87% de la variance ($\frac{\lambda_2^{\boldsymbol{\Sigma}_2}}{\lambda_1^{\boldsymbol{\Sigma}_2}+\lambda_2^{\boldsymbol{\Sigma}_2}} = \frac{4{,}5}{0{,}65+4{,}5} = 0{,}87828$). La transformation linéaire correspondante sera donc la suivante :

$$\mathbf{z} = \mathbf{W}_2^T(\mathbf{x} - \boldsymbol{\mu}_2), \quad \mathbf{W}_2 = \mathbf{w}_2^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}6325 \\ 0{,}7746 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

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

Solution: La matrice de covariance est égale à $\boldsymbol{\Sigma} = \mathbf{W}\mathbf{D}\mathbf{W}^T$.

$$\begin{aligned}
\boldsymbol{\Sigma}_1 &= \mathbf{W}_1\mathbf{D}_1\mathbf{W}_1^T = \begin{bmatrix} 0{,}33 & 0{,}9428 \\ 0{,}9428 & -0.33 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} 0{,}33 & 0{,}9428 \\ 0{,}9428 & -0.33 \end{bmatrix} \\
&= \begin{bmatrix} 0{,}9289 & 0{,}3733 \\ 0{,}3733 & 1{,}8649 \end{bmatrix} \\
\boldsymbol{\Sigma}_2 &= \mathbf{W}_2\mathbf{D}_2\mathbf{W}_2^T = \begin{bmatrix} -0.7746 & 0{,}6325 \\ 0{,}6325 & 0{,}7746 \end{bmatrix} \begin{bmatrix} 0{,}65 & 0 \\ 0 & 4{,}5 \end{bmatrix} \begin{bmatrix} -0.7746 & 0{,}6325 \\ 0{,}6325 & 0{,}7746 \end{bmatrix} \\
&= \begin{bmatrix} 2{,}1903 & 1.8862 \\ 1.8862 & 2.9601 \end{bmatrix}
\end{aligned}$$

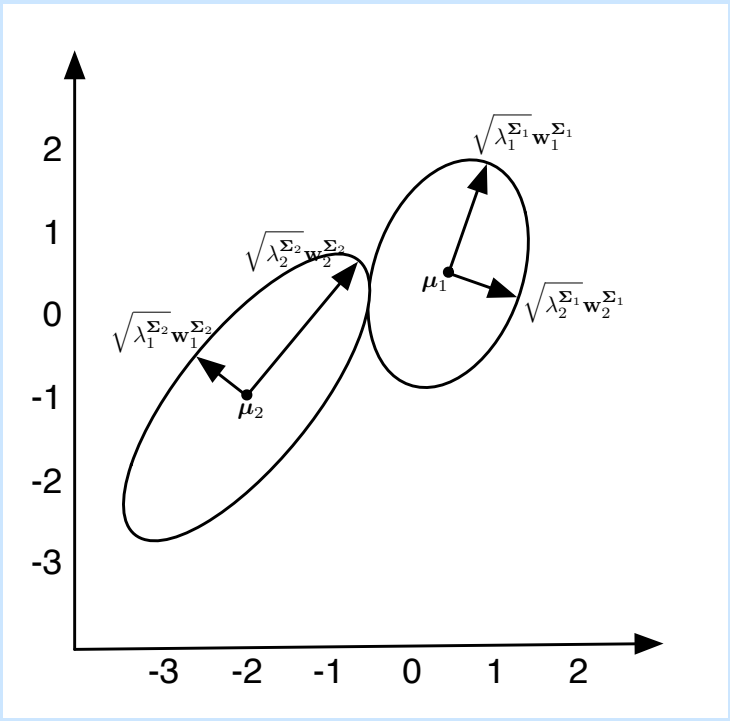
(5) (c) Calculez la matrice de covariance partagée (quelconque, avec valeurs hors la diagonale) par les deux classes ($\boldsymbol{\Sigma}$).

Solution: La matrice de covariance partagée se calcule avec la formule $\boldsymbol{\Sigma} = P(C_1)\boldsymbol{\Sigma}_1 + P(C_2)\boldsymbol{\Sigma}_2$.

$$\begin{aligned}
\boldsymbol{\Sigma} &= P(C_1)\boldsymbol{\Sigma}_1 + P(C_2)\boldsymbol{\Sigma}_2 = 0.6 \begin{bmatrix} 0{,}9289 & 0{,}3733 \\ 0{,}3733 & 1{,}8649 \end{bmatrix} + 0.4 \begin{bmatrix} 2{,}1903 & 1.8862 \\ 1.8862 & 2.9601 \end{bmatrix} \\
&= \begin{bmatrix} 1{,}4334 & 0{,}9785 \\ 0{,}9785 & 2{,}3029 \end{bmatrix}
\end{aligned}$$

(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.

Solution:



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

Solution: L'analyse discriminante linéaire de ces données se fait selon l'équation suivante :

$$z = \mathbf{w}^T \mathbf{x}.$$

La variable $\mathbf{w}$ est calculée selon l'équation $\mathbf{w} = \boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

$$\begin{aligned}
\boldsymbol{\Sigma}_W &= \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0{,}9289 & 0{,}3733 \\ 0{,}3733 & 1{,}8649 \end{bmatrix} + \begin{bmatrix} 2{,}1903 & 1.8862 \\ 1.8862 & 2.9601 \end{bmatrix} \\
&= \begin{bmatrix} 3{,}1192 & 2{,}2596 \\ 2{,}2596 & 4{,}8249 \end{bmatrix} \\
\boldsymbol{\Sigma}_W^{-1} &= \begin{bmatrix} 0{,}4852 & -0{,}2272 \\ -0{,}2272 & 0{,}3137 \end{bmatrix} \\
\mathbf{w} &= \boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{bmatrix} 0{,}4852 & -0{,}2272 \\ -0{,}2272 & 0{,}3137 \end{bmatrix} \left(\begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix} - \begin{bmatrix} -2 \\ -1 \end{bmatrix} \right) \\
&= \begin{bmatrix} 0.8722 \\ -0.0976 \end{bmatrix}
\end{aligned}$$

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

Solution: Le classifieur correspond à un discriminant linéaire dont l'hyperplan séparateur est situé à mi-chemin entre les deux moyenne de classe, et perpenticulaire à la droite connectant ces deux moyennes. La valeur du paramètre $\mathbf{w}$ peut donc être estimée par ce qui suit.

$$\begin{aligned}\begin{bmatrix} a \\ b \end{bmatrix} &= \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix} - \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} 2{,}5 \\ 1{,}5 \end{bmatrix} \\ \mathbf{w} &= \begin{bmatrix} b \\ -a \end{bmatrix} = \begin{bmatrix} -1{,}5 \\ 2{,}5 \end{bmatrix}\end{aligned}$$

Le calcul de la constante w_0 est fait de sorte que la valeur de $\mathrm{h}(\mathbf{x})$ pour $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ soit de magnitude égale, mais de signe contraire.

$$\begin{aligned}\mathrm{h}(\boldsymbol{\mu}_1) &= -\mathrm{h}(\boldsymbol{\mu}_2) \\ \mathbf{w}^T\boldsymbol{\mu}_1 + w_0 &= -\mathbf{w}^T\boldsymbol{\mu}_2 - w_0 \\ 2w_0 &= \mathbf{w}^T(-\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= \frac{1}{2}[-1{,}5 \;\; 2{,}5] \begin{bmatrix} -0.5 - -2 \\ -0{,}5 - -1 \end{bmatrix} = \frac{1}{2}[-1{,}5 \;\; 2{,}5] \begin{bmatrix} 1{,}5 \\ 0{,}5 \end{bmatrix} \\ &= \frac{1}{2}(-2{,}25 + 1.25) = -0.5\end{aligned}$$

L'équation du discriminant linéaire est donc :

$$\mathrm{h}(\mathbf{x}|\mathbf{w},w_0) = \mathbf{w}^T\mathbf{x} + w_0 = [-1{,}5 \;\; 2{,}5]\,\mathbf{x} - 0.5.$$

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.

Solution: L'édition de Wilson permet de retirer les données du jeu qui ne sont pas cohérentes avec les autres données (p. ex. données bruitées), selon un classement de type *leave-one-out*, alors que la condensation de Hart vise à retenir les données qui sont essentielle au bon classement (près des frontières de décision). Si on fait une édition de Wilson sur des données traitées par une condensation de Hart, on va travailler sur un petit ensemble de données essentielles au classement, qui risquent en bonne partie d'être

incohérentes entre elle. De cette façon, l'ensemble de prototypes résultant risque d'être très petit et d'offrir de mauvaises performances.

(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.

Solution: L'erreur de reconstruction fait la somme des différences entre les données $\mathbf{x}^t$ et les position $\mathbf{m}_i$ des centres du groupe à lequel la données est associée. Lorsque cette différence est faible, ceci veut dire que les données sont proches des centres de groupe associé, ce qui est tout à fait pertinent à la tâche de clustering, où on veut identifier des groupes de données compact.

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

Solution: Avec la tâche de régression, la valeur désirée que l'on veut apprendre est dans les réels, $r^t \in \mathbb{R}$, alors qu'en classement, la valeur désirée est discrète, p. ex. $r^t \in \{-1,1\}$ ou bien $r^t \in \{C_1,C_2,\ldots,C_K\}$.

(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.

Solution: L'augmentation de nombre de données utilisées devrait permettre de réduire l'erreur de classement en généralisation. Ceci est explicable du fait qu'un jeu de données de plus grande taille devrait capturer un plus grand éventail de valeurs de données, et donc permettre au classifieur de faire une meilleure approximation du phénomène sous-jacent.

(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.

Solution: L'entraînement pour chaque plis se fera en moyenne sur $\frac{K-1}{K} \times N$ données.

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

Solution: La vraisemblance $p(\mathbf{x}|C_i)$ représente la densité de probabilité des données provenant de la classe C_i. Autrement dit, cette fonction donne la fréquence relative qu'une certaine donnée $\mathbf{x}$ provienne de la classe C_i.

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

Solution: Les valeurs sur la diagonale de la matrice de confusion sont toutes nulles, soit $\mathcal{L}(\alpha_i,C_i) = 0, \forall i$.

(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.

Solution: La variable h_i^t représente l'espérance de la variable cachée booléenne z_i^t, associant la donnée $\mathbf{x}^t$ au groupe $\mathcal{G}_i$ ($z_i^t = 1$) ou non ($z_i^t = 1$), selon le jeu de données $\mathcal{X}$ et la paramétrisation actuelle Φ^l de la densité-mélange.

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

Solution: Avec des loi multinormales, on a une paramétrisation $\Phi = \{\pi_i,\mathbf{m}_i,\mathbf{S}_i\}_{i=1}^K$, où π_i représente la probabilité *a priori* du i-ème groupe, $\mathbf{m}_i$ le vecteur moyen du i-ème groupe et $\mathbf{S}_i$ la matrice de covariance du i-ème groupe.

(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.

Solution: Le taux d'erreur bayésien optimal correspond au taux d'erreur obtenu lorsque l'on connaît les véritables densités de probabilité des différentes classes du problème. C'est le meilleur taux d'erreur que l'on peut obtenir lorsque l'on prend la mesure sur un très vaste ensemble de données du problème. Le taux d'erreur est rarement nul car il est souvent possible d'avoir des données provenant de classes différentes de la classe pour lequel la probabilité *a posteriori* $P(C_i|\mathbf{x})$ est maximale pour un certain $\mathbf{x}$.