Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2016

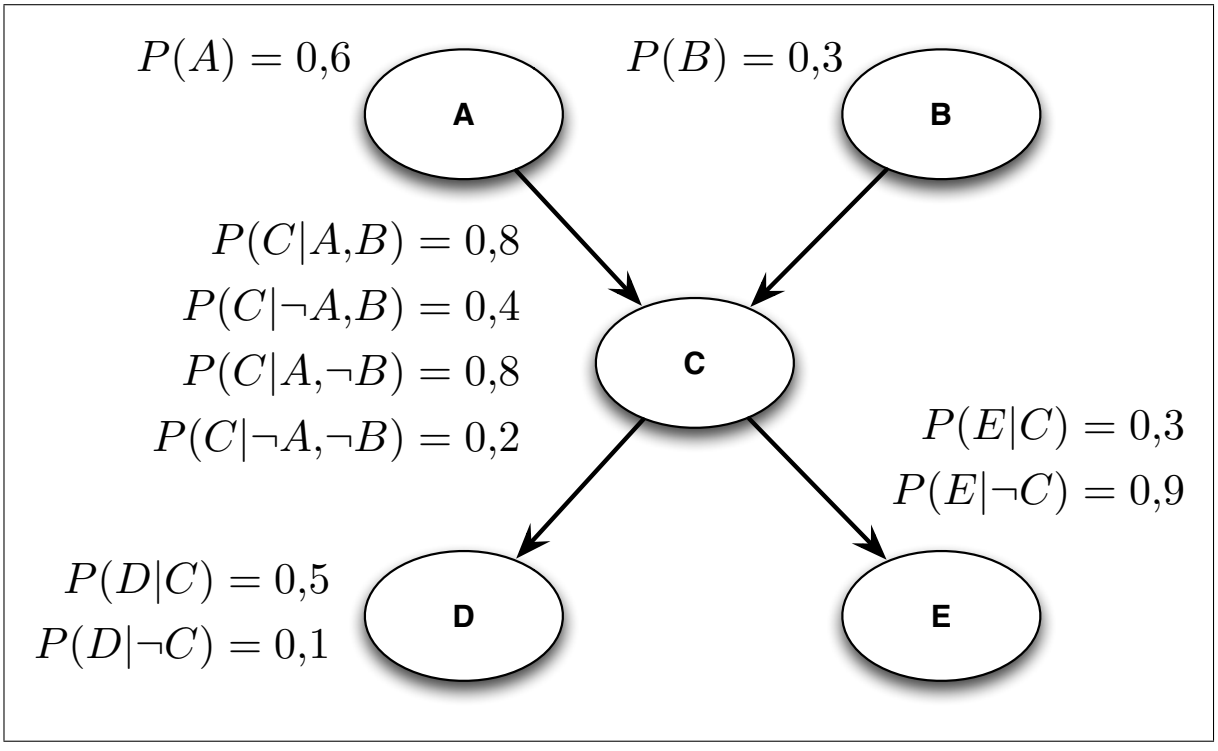
UNIVERSITÉ LAVAL

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 (10 points sur 100)

Soit le réseau bayésien suivant.



(5) (a) Selon ce réseau, calculez la valeur de la probabilité $P(D|E)$.

Solution:

$$
\begin{aligned}
P(C) &= P(C|A,B)\,P(A,B) + P(C|\neg A,B)\,P(\neg A,B) + P(C|A,\neg B)\,P(A,\neg B) + P(C|\neg A,\neg B)\,P(\neg A,\neg B) \\
&= P(C|A,B)\,P(A)\,P(B) + P(C|\neg A,B)\,P(\neg A)\,P(B) + \\
&\quad P(C|A,\neg B)\,P(A)\,P(\neg B) + P(C|\neg A,\neg B)\,P(\neg A)\,P(\neg B) \\
&= (0{,}8 \times 0{,}6 \times 0{,}3) + (0{,}4 \times 0{,}4 \times 0{,}3) + (0{,}8 \times 0{,}6 \times 0{,}7) + (0{,}2 \times 0{,}4 \times 0{,}7) \\
&= 0{,}144 + 0{,}048 + 0{,}336 + 0{,}056 = 0{,}584 \\
P(E) &= P(E|C)\,P(C) + P(E|\neg C)\,P(\neg C) \\
&= (0{,}3 \times 0{,}584) + (0{,}9 \times 0{,}416) = 0{,}1752 + 0{,}3744 = 0{,}5496 \\
P(C|E) &= \frac{P(E|C)\,P(C)}{P(E)} = \frac{0{,}3 \times 0{,}584}{0{,}5496} = 0{,}3188 \\
P(D|E) &= P(D|C)P(C|E) + P(D|\neg C)P(\neg C|E) \\
&= (0{,}5 \times 0{,}3188) + (0{,}1 \times (1 - 0.3188)) = 0{,}1594 + 0{,}06812 \\
&= 0{,}2275
\end{aligned}
$$

(5) (b) Toujours selon ce réseau, calculez la valeur de la probabilité $P(A|D)$.

Solution:

$$
\begin{aligned}
P(D) &= P(D|C)\,P(C) + P(D|\neg C)\,P(\neg C) \\
&= (0{,}5 \times 0{,}584) + (0{,}1 \times 0{,}416) = 0{,}292 + 0{,}0416 = 0{,}3336 \\
P(D|A) &= P(D|C)\,P(C|A) + P(D|\neg C)\,P(\neg C|A) \\
&= P(D|C)\,P(C|A,B)\,P(B) + P(D|C)\,P(C|A,\neg B)\,P(\neg B) + \\
&\quad P(D|\neg C)\,P(\neg C|A,B)\,P(B) + P(D|\neg C)\,P(\neg C|A,\neg B)\,P(\neg B) \\
&= (0{,}5 \times 0{,}8 \times 0{,}3) + (0{,}5 \times 0{,}8 \times 0{,}7) + (0{,}1 \times (1 - 0{,}8) \times 0{,}3) + (0{,}1 \times (1 - 0{,}8) \times 0{,}7) \\
&= 0{,}12 + 0{,}28 + 0{,}006 + 0{,}014 = 0{,}42 \\
P(A|D) &= \frac{P(D|A)\,P(A)}{P(D)} = \frac{0{,}42 \times 0{,}6}{0{,}3336} \\
&= 0{,}7554
\end{aligned}
$$

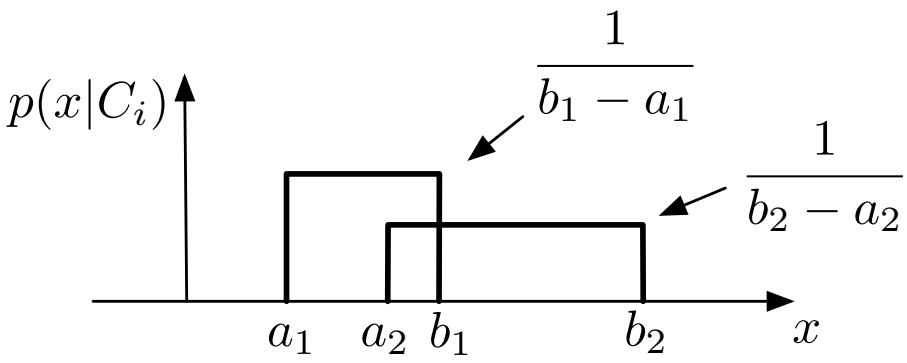
Question 2 (10 points sur 100)

Soit un système de classement paramétrique à deux classes et comportant une variable en entrée. La modélisation des distributions pour chaque classe est donnée par les équations suivantes :

$$
p(x|C_1) = \begin{cases} \frac{1}{b_1 - a_1} & \text{si } x \in [a_1, b_1] \\ 0 & \text{autrement} \end{cases},
$$

$$
p(x|C_2) = \begin{cases} \frac{1}{b_2 - a_2} & \text{si } x \in [a_2, b_2] \\ 0 & \text{autrement} \end{cases},
$$

où $a_1 < b_1$ et $a_2 < b_2$. En guise de simplification, on suppose que $a_1 \leq a_2$. La figure suivante présente le tracé de ces distributions de classes (vraisemblance).



(5) (a) En supposant que $a_1 = 0$, $b_1 = 1{,}25$, $a_2 = 1$ et $b_2 = 2$, donnez la fonction $\mathrm{h}(x)$ permettant la prise de décision pour le classement de données selon la valeur de x dans l'intervalle $[0, 2]$. Supposez que les probabilités *a priori* des classes sont égales, soit $P(C_1) = P(C_2) = 0{,}5$. Supposez également une perte égale pour les différents types d'erreurs. Donnez les développements menant à votre fonction de décision.

Solution: La décision se prend selon la valeur maximale des probabilités *a posteriori* de classement, soit :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1, C_2\}}{\operatorname{argmax}} \; P(C_i|x).$$

Comme les évidences $p(x)$ et les probabilités *a priori* sont les mêmes pour les deux classes, la décision peut se prendre directement à partir des vraisemblances de classe, soit :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1, C_2\}}{\operatorname{argmax}} \; p(x|C_i).$$

Selon les données du problème, on peut calculer que :

$$p(x|C_1) = \begin{cases} \frac{1}{b_1 - a_1} = 0{,}8 & x \in [0, 1{,}25] \\ 0 & \text{autrement} \end{cases}, \tag{1}$$

$$p(x|C_2) = \begin{cases} \frac{1}{b_2 - a_2} = 1 & x \in [1, 2] \\ 0 & \text{autrement} \end{cases}. \tag{2}$$

Comme les vraisemblances de classe sont des boîtes se chevauchant, on identifie trois intervalles d'intérêt :

— $x \in [a_1, a_2]$: associé à la classe C_1 comme $p(x|C_2) = 0$;
— $x \in [a_2, b_1]$: associé à la classe C_2 comme $p(x|C_2) > p(x|C_1)$;
— $x \in [b_1, b_2]$: associé à la classe C_2 comme $p(x|C_1) = 0$;

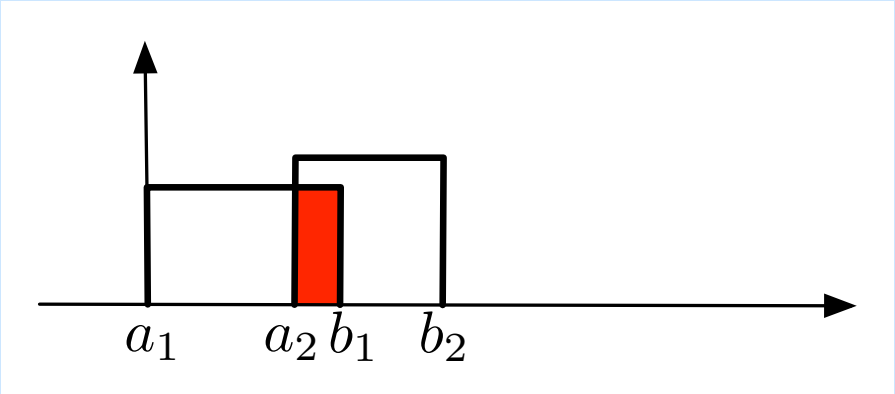
La fonction de prise de décision est donc :

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 1] \\ C_2 & \text{pour } x \in [1, 2] \end{cases}.$$

(5) (b) Calculez le taux d'erreur bayésien optimal que l'on obtient avec le classifieur calculé au

point précédent. Le taux d'erreur bayésien optimal correspond au taux d'erreur obtenu lorsque les données classées suivent parfaitement les distributions estimées pour le classement.

Solution: Des erreurs surviennent lorsqu'une donnée de la classe C_1 a une valeur $x \in [1, 1{,}25]$. La figure suivante présente les distributions données, avec en rouge la région de la distribution $p(x|C_1)$ où un classement d'une donnée de cette classe sera erroné.



Donc, pour estimer l'erreur de classement dans ce cas, il faut calculer l'aire des distributions où une donnée sera mal classée. L'erreur correspondant à l'aire du rectangle rouge dans la figure, soit :

$$\begin{aligned} \text{aire} &= \text{largeur} \times \text{hauteur} && (3)\\ &= (b_1 - a_2) \times \frac{1}{b_1 - a_1} && (4)\\ &= \frac{1{,}25 - 1}{1{,}25 - 0} = \frac{0{,}25}{1{,}25} && (5)\\ &= 0{,}2 && (6) \end{aligned}$$

Comme la probabilité a priori est de $50\,\%$ pour la première classe, le taux d'erreur bayésien optimal, sera de $10\,\%$.

Question 3 (30 points sur 100)

Supposons les données suivantes en deux dimensions :

$$\mathbf{x}^1 = \begin{bmatrix} 1{,}50 \\ -0{,}75 \end{bmatrix}, \quad r^1 = 0, \qquad \mathbf{x}^2 = \begin{bmatrix} 3{,}50 \\ -1{,}30 \end{bmatrix}, \quad r^2 = 0, \qquad \mathbf{x}^3 = \begin{bmatrix} 2{,}00 \\ -2{,}00 \end{bmatrix}, \quad r^3 = 0, \qquad \mathbf{x}^4 = \begin{bmatrix} 4{,}15 \\ -2{,}90 \end{bmatrix}, \quad r^4 = 0,$$

$$\mathbf{x}^5 = \begin{bmatrix} -0{,}10 \\ 1{,}50 \end{bmatrix}, \quad r^5 = 1, \qquad \mathbf{x}^6 = \begin{bmatrix} -2{,}00 \\ 0{,}40 \end{bmatrix}, \quad r^6 = 1, \qquad \mathbf{x}^7 = \begin{bmatrix} -1{,}20 \\ -0{,}75 \end{bmatrix}, \quad r^7 = 1.$$

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

$$\mathbf{m} = \begin{bmatrix} 1{,}1214 \\ -0{,}8286 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 5{,}3949 & -2{,}5426 \\ -2{,}5426 & 2{,}1382 \end{bmatrix}.$$

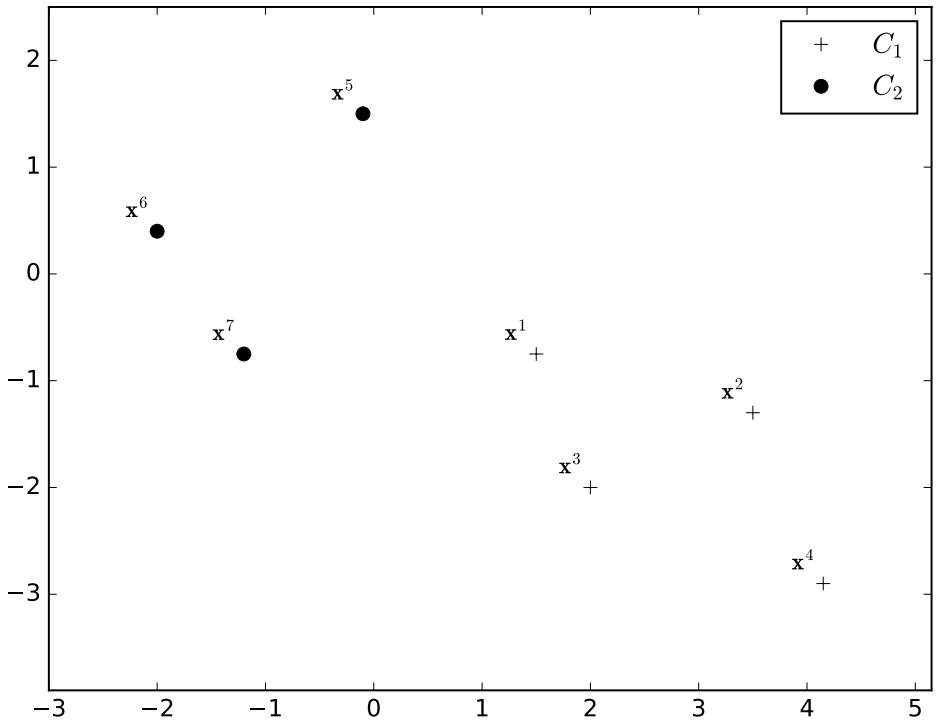
Les vecteurs propres et valeurs propres associés de cette matrice de covariance sont :

$$\lambda_1 = 6{,}7859, \quad \mathbf{c}_1 = \begin{bmatrix} 0{,}8773 \\ -0{,}4799 \end{bmatrix}, \quad \lambda_2 = 0{,}7472, \quad \mathbf{c}_2 = \begin{bmatrix} 0{,}4799 \\ 0{,}8773 \end{bmatrix}.$$

La matrice de distance euclidienne entre chaque paire de données correspond à :

$$\mathbf{D} = \begin{bmatrix} 0{,}0 & 2{,}0742 & 1{,}3463 & 3{,}4125 & 2{,}7609 & 3{,}6841 & 2{,}7 \\ 2{,}0742 & 0{,}0 & 1{,}6553 & 1{,}7270 & 4{,}5607 & 5{,}7567 & 4{,}7321 \\ 1{,}3463 & 1{,}6553 & 0{,}0 & 2{,}3308 & 4{,}0817 & 4{,}6648 & 3{,}4355 \\ 3{,}4125 & 1{,}7270 & 2{,}3308 & 0{,}0 & 6{,}1174 & 6{,}9794 & 5{,}7658 \\ 2{,}7609 & 4{,}5607 & 4{,}0817 & 6{,}1174 & 0{,}0 & 2{,}1954 & 2{,}5045 \\ 3{,}6841 & 5{,}7567 & 4{,}6648 & 6{,}9794 & 2{,}1954 & 0{,}0 & 1{,}4009 \\ 2{,}7 & 4{,}7321 & 3{,}4355 & 5{,}7658 & 2{,}5045 & 1{,}4009 & 0{,}0 \end{bmatrix}.$$

Finalement, les données sont tracées dans la figure suivante.



(6) (a) Effectuez une itération de l'algorithme K-means sur ces données, avec $K = 2$ clusters, en donnant les valeurs de b_i^t et $\mathbf{m}_i$ pour les deux groupes et toutes les données. Démarrez avec comme centres initiaux $\mathbf{m_1}(0) = \mathbf{x}^2$ et $\mathbf{m_2}(0) = \mathbf{x}^7$. Déterminez combien d'itérations sont nécessaires à l'algorithme avant qu'il y ait convergence.

Solution: Les assignations au cluster (étape E) à la première itération selon la distance sont :

$$\mathbf{b}_1 = [1\ 1\ 1\ 1\ 0\ 0\ 0]^T, \quad \mathbf{b}_2 = [0\ 0\ 0\ 0\ 1\ 1\ 1]^T,$$

le calcul des centres correspondants à ces assignations (étape M) est :

$$\mathbf{m}_1 = \begin{bmatrix} 2{,}7875 \\ -1{,}7375 \end{bmatrix}, \quad \mathbf{m}_2 = \begin{bmatrix} -1{,}1000 \\ 0{,}3833 \end{bmatrix}.$$

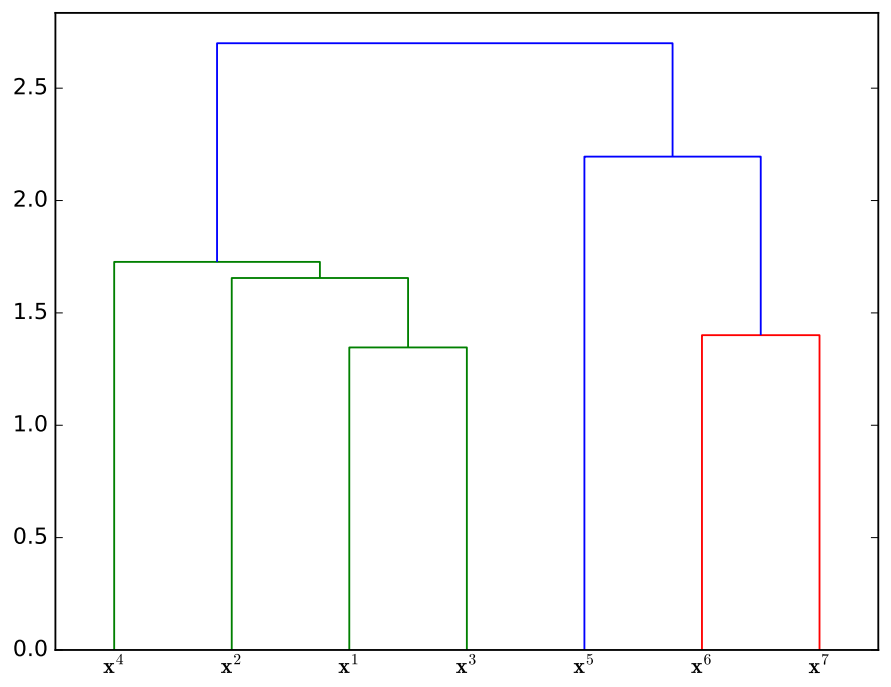
L'assignation des données aux clusters ne va pas changer à la prochaine itération, de sorte que l'algorithme K-means a convergé à la première itération.

(6) (b) Effectuez un clustering hiérarchique agglomératif de ces données. Pour ce faire, effectuez un clustering en lien simple, où la distance entre deux groupes est la distance minimale entre deux paires de données de groupes différents :

$$d(\mathcal{G}_i,\mathcal{G}_j) = \min_{\mathbf{x}^t \in \mathcal{G}_i, \mathbf{x}^s \in \mathcal{G}_j} d(\mathbf{x}^t,\mathbf{x}^s).$$

Utilisez une distance euclidienne. Tracez le résultat dans un dendrogramme. Déterminez également à quoi correspond le clustering si l'on décide de conserver $K = 3$ clusters, en indiquant quelles données forment chacun de ces clusters.

Solution: Le dendrogramme correspondant au clustering des données est le suivant.



À partir du dendrogramme, on peut établir que les trois clusters seraient composés des données $\mathcal{G}_1 = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4\}$, $\mathcal{G}_2 = \{\mathbf{x}^5\}$ et $\mathcal{G}_3 = \{\mathbf{x}^6, \mathbf{x}^7\}$.

(6) (c) Donnez l'équation permettant d'effectuer une transformation blanchissante de ces don-

nées, 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 $\mathbf{b}$.

Solution: L'équation d'une transformation blanchissante est la suivante :

$$\mathbf{z} = \boldsymbol{\Sigma}^{-0,5}(\mathbf{x} - \boldsymbol{\mu}),$$

ce qui revient à dire que $\mathbf{A} = \boldsymbol{\Sigma}^{-0,5}$ et $\mathbf{b} = -\boldsymbol{\Sigma}^{-0,5}\boldsymbol{\mu}$ selon la forme de l'équation linéaire donnée dans l'énoncé. On a donc :

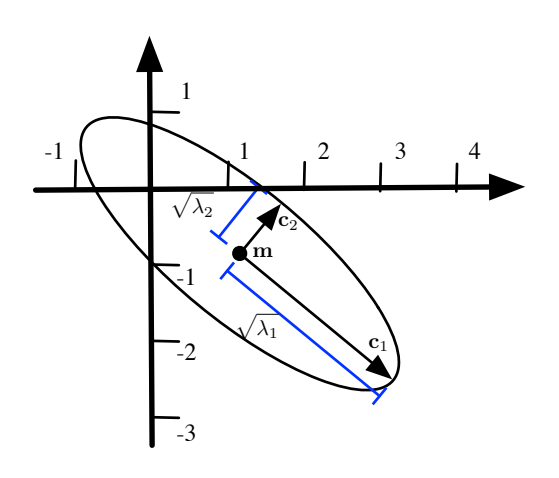
$$\mathbf{A} = \boldsymbol{\Sigma}^{-0,5} = \mathbf{C}\mathbf{D}^{-0,5}\mathbf{C}^T = \begin{bmatrix} 0{,}5619 & 0{,}3255 \\ 0{,}3255 & 0{,}9788 \end{bmatrix},$$

et :

$$\mathbf{b} = -\boldsymbol{\Sigma}^{-0,5}\boldsymbol{\mu} = \begin{bmatrix} -0{,}3605 \\ 0{,}4460 \end{bmatrix}.$$

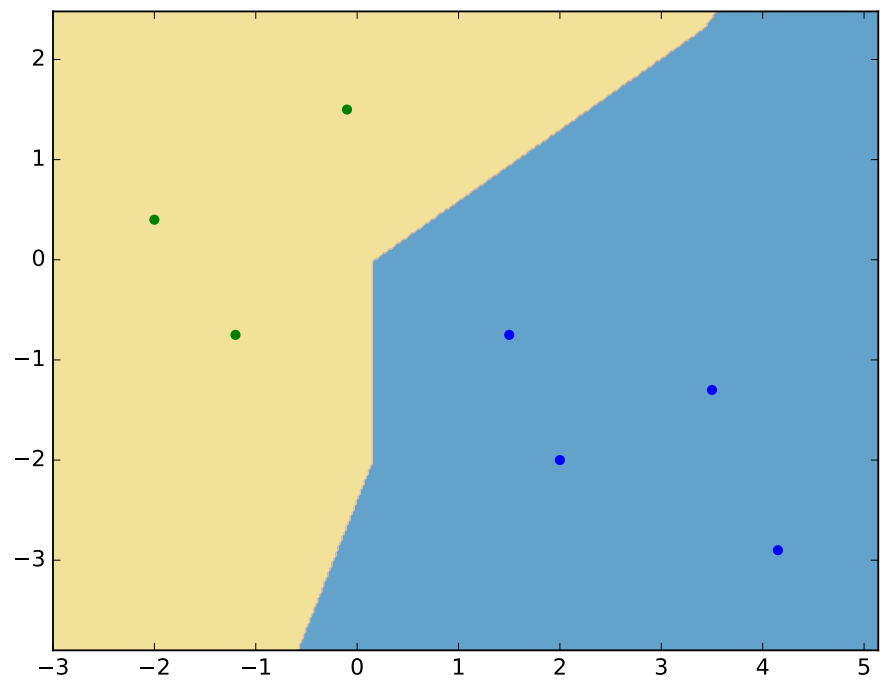
(6) (d) Faites un graphique représentant la distribution des données en deux dimensions, en y traçant la courbe de contour correspondant à une distance de Mahanalobis de 1 (ce qui est équivaut à une distance d'un écart-type en une dimension). Indiquez clairement dans le graphique l'utilisation des différentes valeurs données dans l'énoncé de la question, soit le vecteur moyen estimé $\mathbf{m}$, les valeurs propres λ_i et les vecteurs propres $\mathbf{c}_i$.

Solution:



(6) (e) Tracez les régions de décision selon les données d'entraînement pour un classifieur de type plus proche voisin (avec un seul voisin, $k = 1$). Tracez le tout dans votre **cahier bleu d'examen** (et non dans l'énoncé courant). Donnez également le taux de classement selon une méthodologie *leave-one-out* avec cette configuration, sur ces données.

Solution: Les régions de décision avec un classifieur de type plus proche voisin sont les suivantes.



Le taux de classement selon une méthodologie *leave-one-out* sur ces données est de 100 %, comme le plus proche voisin de chacune des données est une autre donnée ayant la même étiquette de classe.

Question 4 (20 points sur 100)

Supposons que l'on veut appliquer l'algorithme Espérance-Maximisation (EM) à un jeu de données à plusieurs dimensions, où chaque groupe $\mathcal{G}_i$ est décrit par une loi normale $\mathcal{N}_D(\boldsymbol{\mu}_i,\sigma_i^2\mathbf{I})$, soit :

$$p(\mathbf{x}|\boldsymbol{\mu}_i,\sigma_i^2\mathbf{I}) = \frac{1}{(2\pi)^{0,5D}\sigma_i^D} \exp\left[-\frac{\sum_j (x_j - \mu_{i,j})^2}{2\sigma_i^2}\right].$$

Selon cette paramétrisation de la loi normale multidimensionnelle, les valeurs sur la diagonale de la matrice de covariance d'un groupe sont toutes égales à σ_i, alors que les valeurs hors diagonale sont nulles. Donc, la paramétrisation du clustering par EM est donnée par $\Phi = \{\pi_i,\boldsymbol{\mu}_i,\sigma_i^2\}_{i=1}^K$. En guise de rappel, la formule de l'espérance de vraisemblance de l'algorithme EM est la suivante :

$$\mathcal{Q}(\Phi|\Phi^l) = \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i,\Phi^l),$$

où $\pi_i = P(\mathcal{G}_i|\Phi)$ est la probabilité *a priori* du groupe $\mathcal{G}_i$ et $h_i^t = P(\mathcal{G}_i|\mathbf{x},\Phi)$ est l'appartenance probabiliste de la donnée $\mathbf{x}$ au groupe $\mathcal{G}_i$.

(6) (a) Donnez le développement complet permettant de calculer les estimations π_i des probabilités *a priori* des groupes.

Solution: Comme π_i est une probabilité, on a la contrainte que $\sum_i \pi_i = 1$. On résout donc par la méthode de Lagrange :

$$\begin{aligned}\frac{\partial \mathcal{Q}(\Phi|\Phi^l)}{\partial \pi_j} &= \frac{\partial}{\partial \pi_j}\left[\sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i,\Phi^l) - \lambda\left(\sum_i \pi_i - 1\right)\right] \\ &= \sum_t \frac{h_j^t}{\pi_j} - \lambda = 0.\end{aligned}$$

Comme $\sum_i \pi_i = 1$ et $\sum_i h_i^t = 1$:

$$\begin{aligned}\pi_i \sum_t \frac{h_i^t}{\pi_i} &= \pi_i \lambda, \\ \sum_i \frac{\pi_i}{\pi_i} \sum_t h_i^t = \sum_t \sum_i h_i^t &= N = \lambda \sum_i \pi_i = \lambda, \\ \frac{1}{\pi_i}\sum_t h_i^t - N = 0, \qquad \pi_i &= \frac{\sum_t h_i^t}{N}.\end{aligned}$$

(7) (b) Donnez le développement complet permettant de calculer les estimations $\mathbf{m}_i$ des moyennes $\boldsymbol{\mu}_i$.

Solution: $\mathbf{m}_i$ s'estime directement selon le maximum de vraisemblance :

$$\begin{aligned}\frac{\partial \mathcal{Q}(\Phi|\Phi^l)}{\partial m_{u,v}} &= \frac{\partial}{\partial m_{u,v}} \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i,\Phi^l) = 0, \\ &= \sum_t h_u^t \frac{\partial}{\partial m_{u,v}} \log p(\mathbf{x}^t|\mathcal{G}_u,\Phi^l) = 0, \\ &= \sum_t h_u^t \frac{\partial}{\partial m_{u,v}} \log\left(\frac{1}{(2\pi)^{0,5D} s_u^D} \exp\left[-\frac{\sum_j (x_j^t - m_{u,j})^2}{2s_u^2}\right]\right) = 0, \\ &= \sum_t h_u^t \frac{\partial}{\partial m_{u,v}} \left(\log \frac{1}{(2\pi)^{0,5D} s_u^D} + \left[-\frac{\sum_j (x_j^t - m_{u,j})^2}{2s_u^2}\right]\right) = 0, \\ &= \sum_t h_u^t \frac{-2(-1)(x_v^t - m_{u,v})}{2s_u^2} = \sum_t h_u^t \frac{x_v^t - m_{u,v}}{s_u^2} = 0, \\ \sum_t h_u^t x_v^t &= \sum_t h_u^t m_{u,v}, \\ m_{u,v} &= \frac{\sum_t h_u^t x_v^t}{\sum_t h_u^t}.\end{aligned}$$

(7) (c) Donnez le développement complet permettant de calculer les estimations s_i^2 des σ_i^2 correspondants aux valeurs sur la diagonale des matrices de covariance.

Solution: s_i^2 s'estime directement selon le maximum de vraisemblance :

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\Phi|\Phi^l)}{\partial s_k} &= \frac{\partial}{\partial s_k}\sum_t\sum_i h_i^t \log \pi_i + \sum_t\sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i,\Phi^l) = 0,\\
&= \sum_t h_k^t \frac{\partial}{\partial s_k}\log p(\mathbf{x}^t|\mathcal{G}_k,\Phi^l) = 0,\\
&= \sum_t h_k^t \frac{\partial}{\partial s_k}\log\left(\frac{1}{(2\pi)^{0,5D}s_k^D}\exp\left[-\frac{\sum_j (x_j^t - m_{k,j})^2}{2s_k^2}\right]\right) = 0,\\
&= \sum_t h_k^t \left[\frac{\partial}{\partial s_k}\log\frac{1}{(2\pi)^{0,5D}s_k^D} + \frac{\partial}{\partial s_k}\left[-\frac{\sum_j (x_j^t - m_{k,j})^2}{2s_k^2}\right]\right] = 0,\\
&= \sum_t h_k^t \frac{(2\pi)^{0,5D}s_k^D}{(2\pi)^{0,5D}}\frac{\partial}{\partial s_k}\frac{1}{s_k^D} + \sum_t h_k^t \sum_j (x_j^t - m_{k,j})^2 \frac{\partial}{\partial s_k}\left[\frac{-1}{2s_k^2}\right] = 0,\\
&= \sum_t h_k^t \frac{s_k^D(-D)}{s_k^{D+1}} + \sum_t h_k^t \sum_j (x_j^t - m_{k,j})^2 \frac{-1(-2)}{2s_k^3} = 0,\\
&= \frac{-D}{s_k}\sum_t h_k^t + \frac{1}{s_k^3}\sum_t h_k^t \sum_j (x_j^t - m_{k,j})^2 = 0,\\
\frac{D}{s_k}\sum_t h_k^t &= \frac{1}{s_k^3}\sum_t h_k^t \sum_j (x_j^t - m_{k,j})^2,\\
s_k^2 &= \frac{\sum_t h_k^t \sum_j (x_j^t - m_{k,j})^2}{D\sum_t h_k^t}.
\end{aligned}$$

Question 5 (30 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Si l'on effectue de la validation croisée à K-plis (K-*fold cross-validation*, en anglais), indiquez combien d'entraînements de modèles seront nécessaires au minimum pour comparer la performance de deux algorithmes distincts sur un jeu de données particulier.

Solution: Pour chaque algorithme, K entraînements de modèles sera nécessaire pour évaluer la performance sur le jeu de données. Comme on compare deux algorithmes, au minimum $2 \times K$ entraînements de modèles seront nécessaires pour faire la comparaison.

(3) (b) Dans le cours, l'apprentissage automatique a été présenté comme ayant trois dimensions générales : la représentation, l'évaluation et l'optimisation. Dans le contexte de méthodes

paramétriques probabilistes présentées en classe, indiquez où l'on peut intégrer de la connaissance sur le problème relativement à l'aspect de la représentation.

Solution: Avec des méthodes paramétriques probabilistes, la représentation est un élément défini par le choix des lois de densité décrivant la distribution des données. Selon le problème auquel on s'attaque, différentes lois de densité peuvent être utilisées.

(3) (c) Dans un contexte de prise de décision probabiliste avec fonction de perte 0-1 et option de rejet, quelle est la relation entre la valeur de λ représentant le coût de rejet et les probabilités de classement des données relativement à la prise de décision.

Solution: Une décision de rejet sera effectuée lorsque les probabilités de classement sont inférieures à $1 - \lambda$ pour toutes les classes.

(3) (d) Expliquez pourquoi dit-on que les classifieurs plus complexes ont une variance plus élevée, dans une perspective du compromis biais-variance.

Solution: Les modèles qui expriment plus de variances dans leur performance sont généralement des modèles qui font plus de sur-apprentissage, apprenant bien à classer les données d'entraînement. Ainsi, ces modèles présenteront plus de variabilité d'un entraînement à l'autre apprenant des éléments spécifiques aux données d'entraînement qui peuvent être parfois utiles, mais parfois nuisibles à la généralisation.

(3) (e) Indiquez si deux variables indépendantes ont une variance nulle ou non.

Solution: La variance est une mesure intrinsèque de la distribution d'une donnée, il n'y a pas de lien avec la dépendance ou non avec d'autres variables. Sauf pour certains cas dégénérés, la variance d'un variable n'est pas nulle.
La covariance/correlation de deux variables indépendantes est toutefois nulle.

(3) (f) Dans un contexte de classement avec méthodes paramétriques utilisant des lois normales multivariées, indiquez la condition nécessaire pour que la frontière de décision entre les classes soit linéaire.

Solution: La condition nécessaire est que les matrices de covariance des lois normales multivariées représentant chaque classe soient partagées, c'est-à-dire utiliser la même matrice de covariance pour chaque classe.

(3) (g) Expliquez l'effet du paramètre h, correspondant à la largeur de la fenêtre utilisée, dans l'estimation de densités de probabilités avec une fenêtre de Parzen.

Solution: Lorsque la valeur de h est faible, la fenêtre est plus réduite, ce qui donne une estimation avec des pics plus prononcés près des données utilisées pour l'estimation, et une densité de valeur faible plus éloignée loin de ces données. Ce type d'estimation génère un résultat que l'on qualifie en général de bruité. Avec une valeur h plus élevée, l'estimation est plus douce, avec une plus grande portée des données utilisées dans l'estimation de la densité, mais une certaine perte des « hautes fréquences » de la densité.

(3) (h) On dit que les heuristiques de sélections de caractéristiques voraces, comme la sélection avant séquentielle, peuvent ne pas converger à la solution optimale. Expliquez dans quel contexte ceci peut arriver.

Solution: Les algorithmes voraces de sélection de caractéristiques se basent sur une décision locale. Il peut y avoir des cas où il y a une interaction complexe entre plus que deux variables, disons les variables x_a, x_b et x_c, qui fait en sorte que lorsque ces variables sont prises individuellement ou en paires, le gain en performance est faible. Ainsi, avec une sélection avant séquentielle, ces variables ne seront pas sélectionnées. Cependant, si on utilise une méthode moins vorace, tenant compte de l'interaction des trois variables prises ensemble, il est alors possible de détecter que ces variables offrent un gain en performance conjointement substantiel, faisant partie de la solution optimale.

(3) (i) Lorsque l'on veut faire une sélection agressive de prototypes pour le classement avec la règle 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 (ex. données bruitées), selon un classement de type *leave-one-out*, alors que la condensation de Hart vise à retenir les données qui sont essentielles 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 elles. De cette façon, l'ensemble de prototypes résultant risque d'être très petit et d'offrir de mauvaises performances.

(3) (j) Expliquez en quoi consiste la régularisation lorsque l'on fait l'inférence de modèles en apprentissage supervisé.

Solution: La régularisation consiste à intégrer une composante proportionnelle à la complexité du modèle dans la fonction de performance optimisée. Typiquement, la régularisation se fait en optimisant une fonction sous la forme :

$$J = \text{(erreur du modèle)} + \lambda \text{ (complexité du modèle)}.$$