Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Hiver 2013

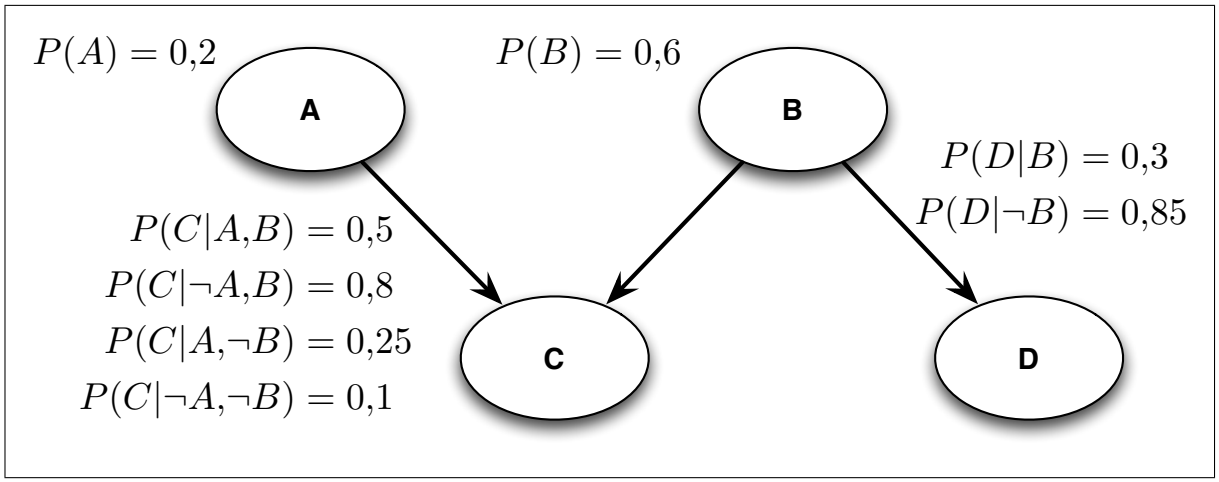
UNIVERSITÉ LAVAL

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 $P(B|C)$.

Solution:

$$
\begin{aligned}
P(C) &= P(C|A,B)\,P(A,B) + P(C|\neg A,B)\,P(\neg A,B) \\
&\quad + P(C|A,\neg B)\,P(A,\neg B) + P(C|\neg A,\neg B)\,P(\neg A,\neg B) \\
&= P(C|A,B)\,P(A)\,P(B) + P(C|\neg A,B)\,P(\neg A)\,P(B) \\
&\quad + P(C|A,\neg B)\,P(A)\,P(\neg B) + P(C|\neg A,\neg B)\,P(\neg A)\,P(\neg B) \\
&= 0{,}5 \cdot 0{,}2 \cdot 0{,}6 + 0{,}8 \cdot (1-0{,}2) \cdot 0{,}6 + 0{,}25 \cdot 0{,}2 \cdot (1-0{,}6) \\
&\quad + 0{,}1 \cdot (1-0{,}2) \cdot (1-0{,}6) \\
&= 0{,}06 + 0{,}384 + 0{,}02 + 0{,}032 = 0{,}496
\end{aligned}
$$

$$\begin{aligned} P(C|B) &= P(C|A,B)\,P(A) + P(C|\neg A,B)\,P(\neg A) \\ &= 0{,}5 \cdot 0{,}2 + 0{,}8 \cdot (1 - 0{,}2) = 0{,}1 + 0{,}64 = 0{,}74 \end{aligned}$$

$$P(B|C) = \frac{P(C|B)\,P(B)}{P(C)} = \frac{0{,}74 \cdot 0{,}6}{0{,}496} = 0{,}8952$$

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

Solution: Comme A et B sont indépendants, on peut dire que $P(B|A) = P(B)$ et $P(\neg B|A) = P(\neg B)$.

$$\begin{aligned} P(D|A) &= P(D|B)\,P(B|A) + P(D|\neg B)\,P(\neg B|A) \\ &= P(D|B)\,P(B) + P(D|\neg B)\,P(\neg B) \\ &= 0{,}3 \cdot 0{,}6 + 0{,}85 \cdot (1 - 0{,}6) = 0{,}18 + 0{,}34 = 0{,}52 \end{aligned}$$

La solution ici-haut est une révision de celle donnée en 2013, qui était erronée.

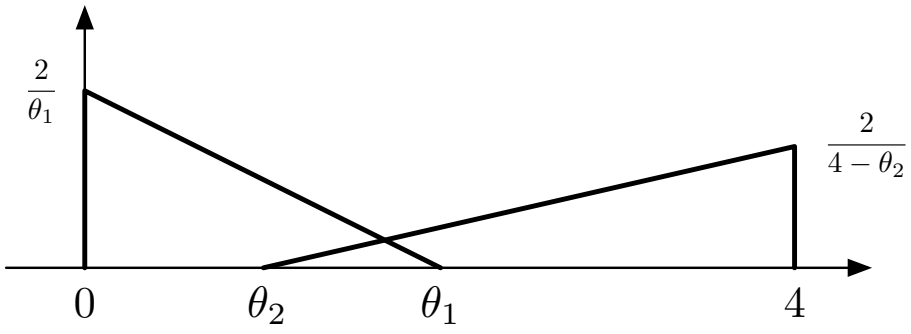
Question 2 (15 points sur 100)

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 :

$$p(x|C_1) = \begin{cases} \frac{-2\,(x-\theta_1)}{(\theta_1)^2} & \text{si } x \in [0,\theta_1] \\ 0 & \text{autrement} \end{cases},$$

$$p(x|C_2) = \begin{cases} \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} & \text{si } x \in [\theta_2,4] \\ 0 & \text{autrement} \end{cases}.$$

Ainsi, la paramétrisation de la distribution de la classe C_1 est donnée par θ_1, alors que celle de la classe C_2 est donnée par θ_2. On fait également l'hypothèse que $0 \leq \theta_2 \leq \theta_1 \leq 4$. La figure suivante présente le tracé de ces distributions de classes.



(5) (a) Supposons que $\theta_1 = 3$ et $\theta_2 = 2$, donnez la fonction $\mathrm{h}(x)$ correspondant à la prise de décision pour le classement de données selon la valeur de $x \in [0{,}4]$. 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 :

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

Comme les vraisemblances de classe sont des fonctions linéaires, il suffit de déterminer le point où les deux distributions sont égales dans l'intervalle $[\theta_2, \theta_1]$:

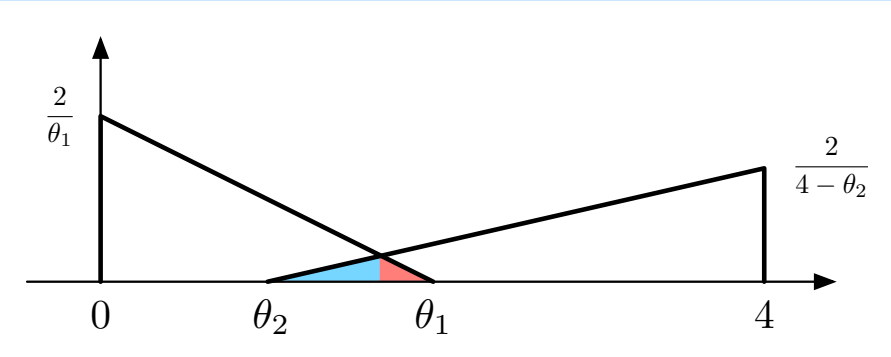
$$\begin{aligned}
p(x|C_1) &= p(x|C_2), \\
\frac{-2\,(x-\theta_1)}{(\theta_1)^2} &= \frac{2\,(x-\theta_2)}{(4-\theta_2)^2}, \\
\frac{-2\,(x-3)}{(3)^2} &= \frac{2\,(x-2)}{(4-2)^2}, \\
\frac{-x+3}{9} &= \frac{x-2}{4}, \\
(0{,}111 + 0{,}25)x &= (0{,}333 + 0{,}5), \\
x &= 2{,}3077.
\end{aligned}$$

Donc, en se basant sur la figure de l'énoncé de la question, on obtient comme fonction de prise de décision ceci :

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 2{,}3077] \\ C_2 & \text{pour } x \in [2{,}3077, 4] \end{cases}.$$

(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_2 a une valeur $x < 2{,}3077$ ou qu'une donnée de la classe C_1 a une valeur $x > 2{,}3077$. La figure suivante présente les distributions données selon les classes, avec en rouge et en bleu les régions des distributions où un classement selon ces distributions résulte en une erreur.



Donc, pour estimer l'erreur de classement dans ce cas, il faut calculer l'aire des distributions où une donnée sera mal classée. Dans le cas de la classe C_1, l'erreur correspond au triangle rouge. Il faut d'abord calculer la hauteur de ce triangle :

$$H_1 = p(x = 2{,}3077|C_1) = \frac{-2\,(x-\theta_1)}{(\theta_1)^2} = \frac{-2\,(2{,}3077-3)}{3^2} = 0{,}15384.$$

Ensuite, la longueur du triangle est calculée comme étant

$$L_1 = \theta_1 - 2{,}3077 = 3 - 2{,}3077 = 0{,}6923.$$

L'aire d'un triangle rectangle se calcule ensuite comme étant le produit de la longueur et de la hauteur du triangle divisé par deux :

$$A_1 = \frac{H_1 \times L_1}{2} = \frac{0{,}15384 \times 0{,}6923}{2} = 0{,}053252.$$

Similairement, pour la classe C_2 l'erreur correspond au triangle bleu de la figure et est calculée comme suit :

$$\begin{aligned} H_2 &= p(x = 2{,}3077|C_2) = \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} = \frac{2\,(2{,}3077-2)}{(4-2)^2} = 0{,}15384, \\ L_2 &= 2{,}3077 - \theta_2 = 2{,}3077 - 2 = 0{,}3077, \\ A_2 &= \frac{H_2 \times L_2}{2} = \frac{0{,}15384 \times 0{,}3077}{2} = 0{,}023668. \end{aligned}$$

Donc, l'erreur totale est égale à la somme des deux aires multiplées par leur probabilités a priori respective, soit :

$$E = P(C_1)\,A_1 + P(C_2)\,A_2 = 0.5 \times 0{,}053252 + 0.5 \times 0{,}023668 = 0{,}03846.$$

Le taux d'erreur bayésien optimal est donc de 3,85 %.

La solution ici-haut est une révision de celle donnée en 2013, qui était erronée.

(5) (c) Supposons maintenant que la fonction de perte est variable selon le type d'erreur que fait notre classifieur. Plus précisément, si une donnée est classée comme étant dans la

classe C_2 mais appartient en fait à la classe C_1, la perte est de $\mathcal{L}(\alpha_2,C_1) = 1$, alors que la perte pour une donnée classée comme étant de la classe C_1, mais appartenant en fait à la classe C_2 implique une perte de $\mathcal{L}(\alpha_1,C_2) = 0{,}5$. Calculez la nouvelle fonction $\mathrm{h}(x)$ correspondant à la prise de décision pour le classement de données selon cette fonction de perte dans le domaine $x \in [0,4]$. Supposez que les autres paramètres sont les mêmes qu'aux points précédents, soit que $\theta_1 = 3$, $\theta_2 = 2$ et $P(C_1) = P(C_2) = 0{,}5$. Donnez les développements menant à votre fonction de décision.

Solution: Avec une fonction de perte, la prise de décision se base sur la minimisation du risque de classement :

$$\mathrm{h}(x) = \underset{C_i \in \{C_1,C_2\}}{\operatorname{argmin}} \; R(C_i|x),$$

où :

$$R(C_i|x) = \sum_{C_j \in \{C_1,C_2\}} \mathcal{L}(C_i,C_j)\, P(C_j|x).$$

Étant donné que $p(x)$ ne change pas selon la classe et que $P(C_1) = P(C_2)$, on peut simplifier le risque par :

$$R(C_i|x) = \sum_{C_j \in \{C_1,C_2\}} \mathcal{L}(C_i,C_j)\, p(x|C_j).$$

Dans le cas présent, les risques pour les classes C_1 et C_2 sont donc :

$$\begin{aligned} R(C_1|x) &= \mathcal{L}(\alpha_1,C_2)\, p(x|C_2) = 0{,}5\, p(x|C_2), \\ R(C_2|x) &= \mathcal{L}(\alpha_2,C_1)\, p(x|C_1) = p(x|C_1). \end{aligned}$$

Comme les fonctions $p(x|C_1)$ et $p(x|C_2)$ sont des équations linéaires, ce qui implique que $R(C_1|x)$ et $R(C_2|x)$ le sont aussi, il suffit de déterminer le point où les deux droites $R(C_1|x)$ et $R(C_2|x)$ se croisent :

$$\begin{aligned} R(C_1|x) &= R(C_2|x), \\ 0{,}5\, p(x|C_2) &= p(x|C_1), \\ 0{,}5 \frac{2\,(x-\theta_2)}{(4-\theta_2)^2} &= \frac{-2\,(x-\theta_1)}{(\theta_1)^2}, \\ 0{,}5 \frac{2x-4}{(4-2)^2} &= \frac{-2x+6}{(3)^2}, \\ \frac{x}{4} + \frac{2x}{9} &= \frac{1}{2} + \frac{6}{9}, \\ 0{,}47222\, x &= 1{,}16667, \\ x &= 2{,}4706. \end{aligned}$$

Donc, en se basant sur la figure de l'énoncé de la question, on obtient comme fonction de prise de décision ceci :

$$\mathrm{h}(x) = \begin{cases} C_1 & \text{pour } x \in [0, 2{,}4706] \\ C_2 & \text{pour } x \in [2{,}4706, 4] \end{cases}.$$

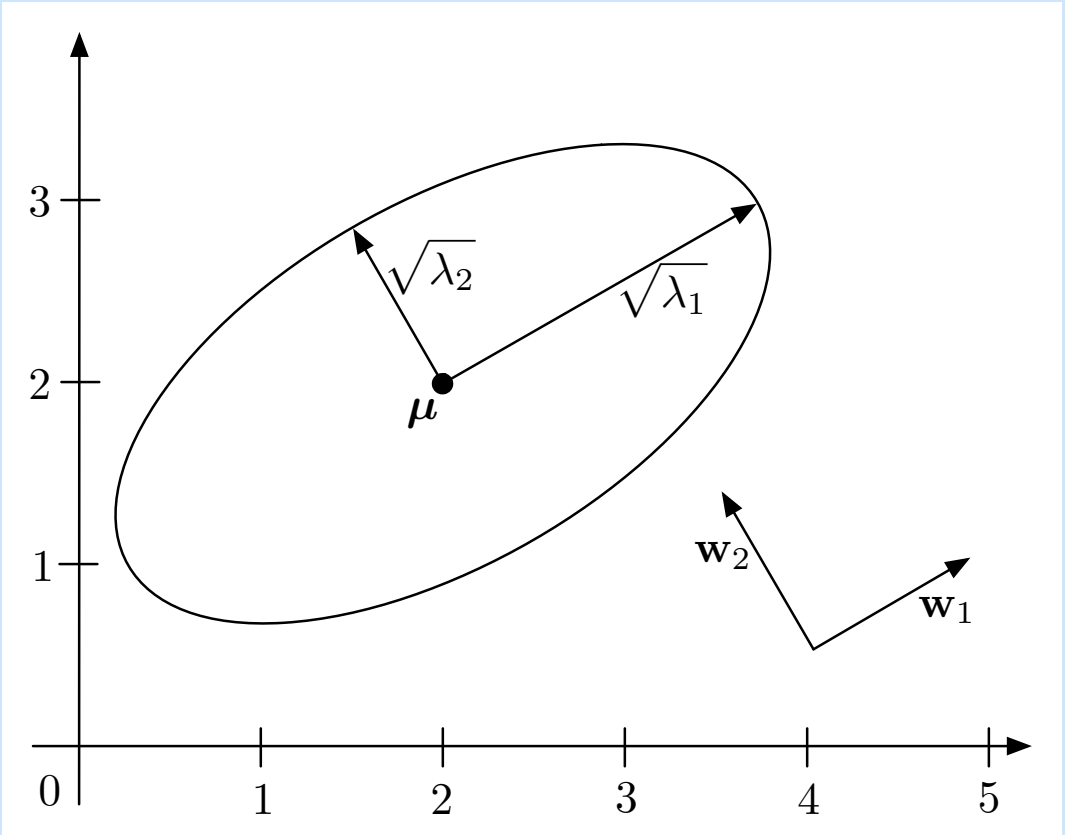
Question 3 (15 points sur 100)

Supposons que l'on a un jeu de données en deux dimensions. Le vecteur moyen $\boldsymbol{\mu}$, ainsi que les valeurs propres λ_i et vecteurs propres $\mathbf{w}_i$ associés à la matrice de covariance $\boldsymbol{\Sigma}$ des données sont les suivants :

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \lambda_1 = 4, \quad \mathbf{w}_1 = \begin{bmatrix} 0{,}866 \\ 0{,}5 \end{bmatrix}, \quad \lambda_2 = 1, \quad \mathbf{w}_2 = \begin{bmatrix} -0{,}5 \\ 0{,}866 \end{bmatrix}.$$

(10) (a) Faites un graphique représentant cette distribution en deux dimensions, en donnant 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 pour chaque distribution, soit le vecteur moyen $\boldsymbol{\mu}$, les valeurs propres λ_i et les vecteurs propres $\mathbf{w}_i$.

Solution:



(5) (b) Donnez l'équation en forme matricielle simplifiée correspondant à une transformation blanchissante des données. Détaillez les valeurs numériques des variables formant cette équation.

Solution: L'équation correspondant à une transformation blanchissante est la suivante :

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

La matrice $\mathbf{\Sigma}^{-0,5}$ est calculée comme suit :

$$\begin{aligned}
\mathbf{W} &= [\mathbf{w}_1\ \mathbf{w}_2] = \begin{bmatrix} 0{,}866 & -0{,}5 \\ 0{,}5 & 0{,}866 \end{bmatrix} \\
\mathbf{D}^{-0,5} &= \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0 \\ 0 & 1 \end{bmatrix} \\
\mathbf{\Sigma}^{-0,5} &= \mathbf{W}\,\mathbf{D}^{-0,5}\,\mathbf{W}^T \\
&= \begin{bmatrix} 0{,}866 & -0{,}5 \\ 0{,}5 & 0{,}866 \end{bmatrix} \begin{bmatrix} 0{,}5 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0{,}866 & 0{,}5 \\ -0{,}5 & 0{,}866 \end{bmatrix} \\
&= \begin{bmatrix} 0{,}866 & -0{,}5 \\ 0{,}5 & 0{,}866 \end{bmatrix} \begin{bmatrix} 0{,}433 & 0{,}25 \\ -0{,}5 & 0{,}866 \end{bmatrix} \\
&= \begin{bmatrix} 0{,}625 & -0{,}217 \\ -0{,}217 & 0{,}875 \end{bmatrix}
\end{aligned}$$

Le vecteur moyen est celui donné dans l'énoncé, soit $\boldsymbol{\mu} = [2\ 2]^T$, alors que $\mathbf{x}$ est un point dans l'espace d'origine et $\mathbf{z}$ est la position associée dans l'espace résultant de la transformation blanchissante.

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 $\mathcal{N}(\boldsymbol{\mu}_i,\sigma_i^2\mathbf{I})$, 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 tous 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 :

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

(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 $\sum_i h_i^t = 1$:

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

(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 (40 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(4) (a) Pour la sélection de caractéristiques, expliquez pourquoi on considère les approches de type « enveloppe » (*wrapper*) comme étant plus exigeantes en ressources computationnelles que les approches de type « filtre » (*filter*).

Solution: Les approches de type « enveloppe » sont plus exigeantes en calculs, car elles impliquent qu'un classifieur est entraîné et testé pour chaque sous-ensemble de caractéristiques considéré par l'algorithme. Par opposition, les approches de type « filtre » basent le choix du sous-ensemble sur des mesures de performance beaucoup moins exigeantes en traitements qu'un entraînement de classifieur, par exemple en utilisant des mesures telles que la corrélation entre les caractéristiques ou l'information mutuelle entre les caractéristiques et les étiquettes de classe.

(4) (b) Expliquez pourquoi on dit que l'analyse en composantes principales est une approche non supervisée, alors que l'analyse discriminante linéaire est une approche supervisée.

Solution: L'analyse en composantes principales est une approche non supervisée, car les étiquettes de classes des données ne sont pas du tout utilisées par la méthode. Par opposition, l'analyse discriminante linéaire est une approche supervisée, car la méthode tient compte des étiquettes de classes des données utilisées.

(4) (c) Il a été mentionné à plusieurs reprises qu'à performances égales, on doit préférer les modèles de classement les plus simples. Expliquez plus en détail pourquoi on doit procéder ainsi.

Solution: Quatre raisons principales justifient l'utilisation de modèles de classement simples, soit :

- Ce sont des modèles plus faciles à utiliser, comme leur complexité algorithmique est généralement réduite ;
- Ce sont des modèles plus faciles à entraîner, comme leur complexité en espace mémoire est également plus faible ;
- Ce sont des modèles plus facilement explicables, ce qui est intéressant lorsque l'on veut pouvoir les interpréter ;
- Ils généralisent mieux, comme les modèles plus simples sont plus plausibles (rasoir d'Ockham).

(4) (d) En général, que doit-on conclure de la complexité d'un modèle de classifieurs si l'on observe que son biais est élevé alors que sa variance est faible ?

Solution: Avec un biais élevé et une variance faible, on peut conclure que le modèle de classifieurs effectue du sous-apprentissage et qu'il est donc trop simple relativement au problème à résoudre.

(4) (e) Pour un ensemble de données particulier, si on suppose que deux variables sont indépendantes, quelle valeur de corrélation doit-on s'attendre à mesurer entre ces variables ?

Solution: Si deux variables sont indépendantes, on doit s'attendre à mesure une corrélation nulle entre celles-ci.

(4) (f) Expliquez pourquoi on dit en général qu'une bonne expertise du domaine est nécessaire pour pouvoir utiliser les réseaux bayésiens dans un contexte applicatif spécifique.

Solution: Une expertise du domaine est nécessaire pour pouvoir définir les dépendances directes entre les variables formant les nœuds du réseau. En effet, un expert du domaine peut en général les énoncer assez aisément. Il est beaucoup plus difficile d'identifier automatiquement ces dépendances à partir des données, comme des relations peuvent être identifiées entre des variables sans correspondre à un lien causal ayant rapport avec la réalité.

(4) (g) On dit qu'un algorithme de clustering tel que le K-means permet de faire de la compression de données. Expliquez plus en détail comment ceci peut être effectué et à quoi correspond alors la perte d'information associée à cette compression.

Solution: Avec un algorithme de clustering tel que le K-means, on effectue de la quantification de vecteurs, en remplaçant un vecteur dans un espace $\mathbb{R}^D$ par l'indice d'un vecteur de référence $\mathbf{m}_i$. Selon la dimensionnalité de l'espace d'origine et le nombre de vecteurs de référence utilisés, le taux de compression peut être assez important. La perte d'information associée à cette compression correspond à la différence (distance) entre le vecteur d'origine et le vecteur de référence utilisé pour le modéliser. Si la distance entre ces deux vecteurs est importante, la perte d'information est significative, alors que si la distance est faible, la perte est réduite.

(4) (h) Supposons un jeu de données de classement à plusieurs classes, où l'on veut modéliser les données de chaque classe par des distributions normales multivariées. Cependant, pour certaines de ces classes, il s'avère que les données comportent plusieurs modes, c'est-à-dire que les distributions comportent plusieurs pics significatifs à différentes positions dans l'espace des données. Expliquez de quelle façon on peut effectuer des densités-mélanges à l'aide de l'algorithme EM pour un classement paramétrique de ces données.

Solution: Pour effectuer des densités-mélanges pour chaque classe, il faut d'abord identifier le nombre de modes que comportent les données de chaque classe. Ceci peut s'effectuer par visualisation (ex. avec une Analyse en composantes principales), ou empiriquement, en appliquant un algorithme de clustering tel que K-means, où on fait varier le nombre de centres. Ensuite, on peut appliquer l'algorithme EM pour les données de chaque classe, une classe à la fois, avec le nombre de groupes identifié précédemment, afin d'estimer les paramètres des densités-mélanges associées à chacune de ces classes.

(4) (i) Expliquez en quoi la validation croisée à K plis (K-*fold cross-validation*) est intéressante pour évaluer la performance d'algorithmes d'apprentissage, lorsqu'un petit ensemble de données est disponible à cette fin.

Solution: La validation croisée à K plis permet de gérer efficacement les petits ensembles de données, en effectuant plusieurs entraînements (K répétitions) sur une portion $(K-1)/K$ des données, suivi d'un test sur la portion $1/K$ restante des données. Ainsi, pour des valeurs de K relativement élevées, la taille de l'ensemble d'entraînement à chaque répétition se rapproche de la taille de l'ensemble de données d'origine. De plus, le résultat rapporté est la moyenne des performances en test pour les K répétitions, ce qui revient au taux d'erreur sur tout l'ensemble de données. Ceci permet donc une utilisation efficace des données, avec un entraînement sur une bonne proportion des données et un test sur toutes les données. Cependant, cette approche peut être très coûteuse en traitements, comme elle implique d'effectuer K répétitions d'entraînements suivies d'un test. Mais ceci devrait être gérable avec un petit ensemble de données, comme la charge de calcul de ces opérations est généralement proportionnelle à la taille de l'ensemble de données.

(4) (j) Expliquez pourquoi en régression multivariée on se limite généralement à des équations linéaires (polynômes d'ordre 1), sauf dans le cas de données à faible dimensionnalité.

Solution: On se limite généralement à des équations linéaires en régression multivariée, car le nombre de paramètres à estimer croît significativement selon la dimensionnalité des données et l'ordre des polynômes. Avec des polynômes d'ordre relativement élevé et des données à haute dimensionnalité, le risque de sur-apprentissage serait alors trop important. Donc, avec plus d'une dizaine de variables, des équations linéaires devraient être suffisamment complexes et flexibles pour bien modéliser les données sans effectuer de sur-apprentissage important.