Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2011

UNIVERSITÉ LAVAL

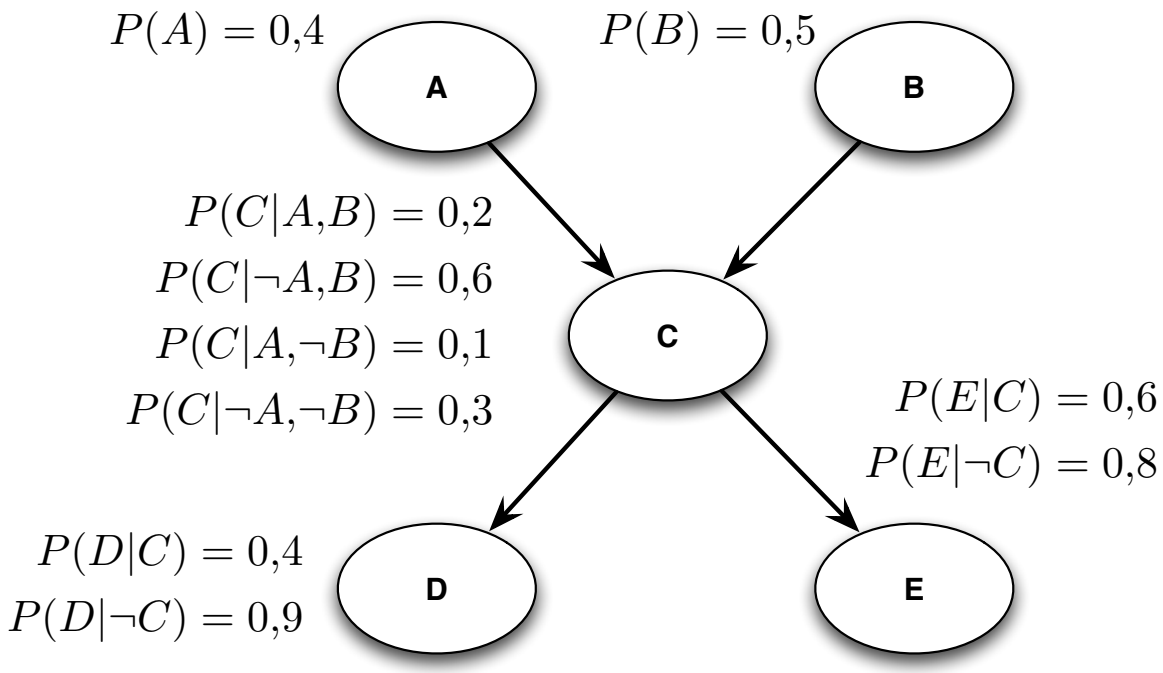
EXAMEN PARTIEL

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 (12 points sur 100)

Soit le réseau bayésien suivant.



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

Solution:

$$
\begin{aligned}
P(C) &= P(C|A,B)\,P(A,B) + P(C|\neg A,B)\,P(\neg A,B) + P(C|A,\neg B)\,P(A,\neg B) + P(C|\neg A,\neg B)\,P(\neg A,\neg B) \\
&= P(C|A,B)\,P(A)\,P(B) + P(C|\neg A,B)\,P(\neg A)\,P(B) + \\
&\quad P(C|A,\neg B)\,P(A)\,P(\neg B) + P(C|\neg A,\neg B)\,P(\neg A)\,P(\neg B) \\
&= (0{,}2 \times 0{,}4 \times 0{,}5) + (0{,}6 \times 0{,}6 \times 0{,}5) + (0{,}1 \times 0{,}4 \times 0{,}5) + (0{,}3 \times 0{,}6 \times 0{,}5) \\
&= 0{,}04 + 0{,}18 + 0{,}02 + 0{,}09 = 0{,}33
\end{aligned}
$$

$$\begin{aligned} P(E) &= P(E|C)\,P(C) + P(E|\neg C)\,P(\neg C) \\ &= (0{,}6 \times 0{,}33) + (0{,}8 \times 0{,}67) = 0{,}734 \end{aligned}$$

$$P(C|E) = \frac{P(E|C)\,P(C)}{P(E)} = \frac{0{,}6 \times 0{,}33}{0{,}734} = 0{,}2698$$

$$\begin{aligned} P(D|E) &= P(D|C)P(C|E) + P(D|\neg C)P(\neg C|E) \\ &= (0{,}4 \times 0{,}2698) + (0{,}9 \times (1 - 0.2698)) \\ &= 0{,}7651 \end{aligned}$$

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

Solution:

$$\begin{aligned} P(D) &= P(D|C)\,P(C) + P(D|\neg C)\,P(\neg C) \\ &= (0{,}4 \times 0{,}33) + (0{,}9 \times 0{,}67) = 0{,}735 \end{aligned}$$

$$\begin{aligned} P(D|A) &= P(D|C)\,P(C|A) + P(D|\neg C)\,P(\neg C|A) \\ &= P(D|C)\,P(C|A,B)\,P(B) + P(D|C)\,P(C|A,\neg B)\,P(\neg B) + \\ &\quad P(D|\neg C)\,P(\neg C|A,B)\,P(B) + P(D|\neg C)\,P(\neg C|A,\neg B)\,P(\neg B) \\ &= (0{,}4 \times 0{,}2 \times 0{,}5) + (0{,}4 \times 0{,}1 \times 0{,}5) + (0{,}9 \times (1 - 0.2) \times 0{,}5) + (0{,}9 \times (1 - 0{,}1) \times 0{,}5) \\ &= 0{,}04 + 0{,}02 + 0{,}36 + 0{,}405 = 0{,}825 \end{aligned}$$

$$\begin{aligned} P(A|D) &= \frac{P(D|A)\,P(A)}{P(D)} = \frac{0{,}825 \times 0{,}4}{0{,}735} \\ &= 0{,}449 \end{aligned}$$

Question 2 (20 points sur 100)

Supposons que l'on veut automatiser le traitement d'images de champs agricoles, afin de discriminer les mauvaises herbes des plantes de maïs. Nous avons déjà sous la main une chaîne de traitement des images qui permet de segmenter les pixels correspondants à une plante, qu'elle soit bonne ou mauvaise, des autres éléments. Nous avons également un algorithme qui extrait des pixels formant un ensemble continu selon leur voisinage, que l'on nomme *blobs*. De plus, selon la distribution des blobs dans l'image, nous sommes en mesure de différencier les régions de l'image correspondant à des rangs de maïs des régions correspondant aux espaces inter-rangs. Et finalement, nous avons une banque d'images de ce type où chacun des blobs des images a été étiqueté selon deux classes, soit les mauvaises herbes (classe C_M) et les plantes de maïs (classe C_B).

Dans des études préliminaires, il a été établi que la taille des blobs est une mesure permettant de bien discriminer les plantes de maïs des mauvaises herbes. De plus, nous faisons l'hypothèse que la distribution des mauvaises herbes est la même partout dans les images, alors que les plantes de maïs sont présentes uniquement dans les régions désignées comme étant des

rangs. Les éléments suivants peuvent donc être estimés à partir de la banque d'images étiquetées que l'on a sous la main :

— $p(x|C_M)$: densité de probabilité des mauvaises herbes selon la taille des blobs de plantes (x), cette densité est la même dans tout l'image (que l'on soit dans les rangs ou dans les inter-rangs) ;

— $p_{\text{rang}}(x)$: densité de probabilité des plantes, sans égard que ce soit des mauvaises herbes du maïs, dans les rangs selon la taille des blobs de plantes (x) ;

— $P_{\text{rang}}(C_B)$ et $P_{\text{rang}}(C_M)$: probabilités *a priori* de blobs de plantes de maïs ($P_{\text{rang}}(C_B)$) et de mauvaises herbes ($P_{\text{rang}}(C_M)$) dans les rangs ;

— $P_{\text{inter}}(C_B) = 0$ et $P_{\text{inter}}(C_M) = 1$: probabilités *a priori* de blobs de plantes de maïs (nulle, $P_{\text{inter}}(C_B) = 0$) et de mauvaises herbes ($P_{\text{inter}}(C_M) = 1$) dans les inter-rangs.

(5) (a) Donnez l'équation de la densité de probabilité de plantes (maïs ou mauvaise herbe) dans les inter-rangs, $p_{\text{inter}}(x)$, en utilisant seulement les éléments connus précédemment mentionnés.

Solution:

$$\begin{aligned} p_{\text{inter}}(x) &= P_{\text{inter}}(C_B)\, p_{\text{inter}}(x|C_B) + P_{\text{inter}}(C_M)\, p_{\text{inter}}(x|C_M) \\ &= 0 \times p_{\text{inter}}(x|C_B) + 1 \times p(x|C_M) \\ &= p(x|C_M) \end{aligned}$$

(5) (b) Donnez l'équation de la densité de probabilité qu'une plante soit du maïs dans les rangs, $p_{\text{rang}}(x|C_B)$, en utilisant seulement les éléments connus précédemment mentionnés.

Solution:

$$\begin{aligned} p_{\text{rang}}(x) &= P_{\text{rang}}(C_B)\, p_{\text{rang}}(x|C_B) + P_{\text{rang}}(C_M)\, p_{\text{rang}}(x|C_M) \\ p_{\text{rang}}(x|C_B) &= \frac{p_{\text{rang}}(x) - (P_{\text{rang}}(C_M)\, p(x|C_M))}{P_{\text{rang}}(C_B)} \end{aligned}$$

(5) (c) Donnez l'équation simplifiée de la fonction de décision $\text{h}_{\text{rang}}(x)$ donnant la classe (C_M : mauvaise herbe ; C_B : plante de maïs) d'un blob de pixels de l'image selon sa taille x, dans les rangs. À cette fin, donnez le détail du développement des probabilités $P_{\text{rang}}(C_B|x)$ et $P_{\text{rang}}(C_M|x)$.

Solution:

$$\begin{aligned} P_{\text{rang}}(C_M|x) &= \frac{P_{\text{rang}}(C_M)\, p_{\text{rang}}(x|C_M)}{p_{\text{rang}}(x)} = \frac{P_{\text{rang}}(C_M)\, p(x|C_M)}{p_{\text{rang}}(x)} \\ P_{\text{rang}}(C_B|x) &= 1 - P_{\text{rang}}(C_M,x) \\ \text{h}_{\text{rang}}(x) &= \begin{cases} C_B & \text{si } P_{\text{rang}}(C_B|x) \geq P_{\text{rang}}(C_M|x) \\ C_M & \text{autrement} \end{cases} \end{aligned}$$

(5) (d) Donnez l'équation simplifiée de la fonction de décision $\mathrm{h}_{\text{inter}}(x)$ donnant la classe (C_M : mauvaise herbe ; C_B : plante de maïs) d'un blob de pixels de l'image selon sa taille x, entre les rangs. À cette fin, donnez le détail du développement des probabilités $P_{\text{inter}}(C_B|x)$ et $P_{\text{inter}}(C_M|x)$.

Solution:

$$\begin{aligned} P_{\text{inter}}(C_M|x) &= \frac{P_{\text{inter}}(C_M)\, p_{\text{inter}}(x|C_M)}{p_{\text{inter}}(x)} = \frac{P_{\text{inter}}(C_M)\, p(x|C_M)}{p(x|C_M)} \\ &= P_{\text{inter}}(C_M) = 1 \\ P_{\text{inter}}(C_B|x) &= 1 - P_{\text{inter}}(C_M|x) = 0 \\ \mathrm{h}_{\text{inter}}(x) &= C_M, \quad \forall x \end{aligned}$$

Question 3 (15 points sur 100)

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

$$p(x|\mathcal{G}_i,\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right].$$

Donc, la paramétrisation du clustering par EM est donnée par $\Phi = \{\pi_i,\mu_i,\sigma_i\}_{i=1}^K$. En guise de rappel, la formule de l'espérance de vraisemblance de l'algorithme EM est la suivante :

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

(5) (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 :

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

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

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

(5) (b) Donnez le développement complet permettant de calculer les estimations m_i des moyennes μ_i.

Solution: Résolution par $\partial \mathcal{Q}(\Phi|\Phi^l)/\partial m_i = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\Phi|\Phi^l)}{\partial m_j} &= \frac{\partial}{\partial m_j} \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i, \Phi^l) \\
&= \sum_t \frac{\partial}{\partial m_j} \sum_i h_i^t \log \frac{1}{\sqrt{2\pi} s_i} \exp\left[-\frac{(x^t - m_i)^2}{2s_i^2}\right] \\
&= \sum_t \left(\frac{\partial}{\partial m_j} \sum_i h_i^t \log \frac{1}{\sqrt{2\pi} s_i} + \frac{\partial}{\partial m_j} \sum_i h_i^t \left[-\frac{(x^t - m_i)^2}{2s_i^2}\right] \right) \\
&= \sum_t h_j^t \left(2\frac{x^t - m_j}{2s_j^2} \right) = \sum_t h_j^t \frac{x^t - m_j}{s_j^2} \\
&= \sum_t h_j^t x^t - \sum_t h_j^t m_j = 0, \\
m_j &= \frac{\sum_t h_j^t x^t}{\sum_t h_j^t}.
\end{aligned}$$

(5) (c) Donnez le développement complet permettant de calculer les estimations s_i^2 des variances σ_i^2.

Solution: Résolution par $\partial \mathcal{Q}(\Phi|\Phi^l)/\partial s_i = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\Phi|\Phi^l)}{\partial s_j} &= \frac{\partial}{\partial s_j}\sum_t\sum_i h_i^t \log \pi_i + \sum_t\sum_i h_i^t \log p(\mathbf{x}^t|\mathcal{G}_i,\Phi^l) \\
&= \sum_t \frac{\partial}{\partial s_j}\sum_i h_i^t \log \frac{1}{\sqrt{2\pi}s_i}\exp\left[-\frac{(x^t-m_i)^2}{2s_i^2}\right] \\
&= \sum_t \left(\frac{\partial}{\partial s_j}\left[\sum_i h_i^t \log\frac{1}{\sqrt{2\pi}} - \sum_i h_i^t \log s_i\right] + \frac{\partial}{\partial s_j}\sum_i h_i^t\left[-\frac{(x^t-m_i)^2}{2s_i^2}\right]\right) \\
&= -\sum_t \frac{h_j^t}{s_j} + \sum_t h_j^t(-2)\left[-\frac{(x^t-m_j)^2}{2s_j^3}\right] \\
&= -\sum_t \frac{h_j^t}{s_j} + \sum_t h_j^t \frac{(x^t-m_j)^2}{s_j^3} = 0, \\
\sum_t \frac{h_j^t}{s_j} &= \sum_t h_j^t \frac{(x^t-m_j)^2}{s_j^3}, \\
\sum_t h_j^t &= \frac{1}{s_j^2}\sum_t h_j^t (x^t-m_j)^2, \\
s_j^2 &= \frac{\sum_t h_j^t (x^t-m_j)^2}{\sum_t h_j^t}.
\end{aligned}$$

Question 4 (20 points sur 100)

Supposons que l'on a un jeu de données en trois 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 aux matrices de covariance $\boldsymbol{\Sigma}_i$ de chaque classe sont les suivants :

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 0{,}5 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -1 \\ 3 \end{bmatrix},$$

$$\lambda_1^{\boldsymbol{\Sigma}_1} = 1{,}0201, \qquad \lambda_2^{\boldsymbol{\Sigma}_1} = 0{,}1150, \qquad \lambda_3^{\boldsymbol{\Sigma}_1} = 2{,}9288,$$

$$\mathbf{w}_1^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} 0{,}3189 \\ -0{,}8811 \\ 0{,}3493 \end{bmatrix}, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} -0{,}8167 \\ -0{,}0684 \\ 0{,}5729 \end{bmatrix}, \quad \mathbf{w}_3^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} -0{,}4809 \\ -0{,}4680 \\ -0{,}7414 \end{bmatrix},$$

$$\lambda_1^{\boldsymbol{\Sigma}_2} = 3{,}6260, \qquad \lambda_2^{\boldsymbol{\Sigma}_2} = 0{,}0162, \qquad \lambda_3^{\boldsymbol{\Sigma}_2} = 0{,}3673,$$

$$\mathbf{w}_1^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}6651 \\ 0{,}6599 \\ 0{,}3494 \end{bmatrix}, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}7165 \\ -0{,}6958 \\ -0{,}0497 \end{bmatrix}, \quad \mathbf{w}_3^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} -0{,}2103 \\ -0{,}2835 \\ 0{,}9356 \end{bmatrix}.$$

Les probabilités *a priori* des classes sont respectivement $P(C_1) = 0{,}75$ et $P(C_2) = 0{,}25$.

(5) (a) Pour chaque classe, donnez la transformation linéaire nécessaire pour conserver au moins 75 % de la variance. Traitez chaque classe indépendamment.

Solution: Pour la classe C_1, on doit utiliser les deux premières composantes principales, car la première composante principale capture 72 % de la variance ($\lambda_3^{\mathbf{\Sigma}_1}/(\lambda_3^{\mathbf{\Sigma}_1} + \lambda_1^{\mathbf{\Sigma}_1} + \lambda_2^{\mathbf{\Sigma}_1}) = 2{,}9288/(2{,}9288 + 1{,}0201 + 0{,}1150) = 0.7207$), et les deux premières capturent 97 % de la variance ($(\lambda_3^{\mathbf{\Sigma}_1} + \lambda_1^{\mathbf{\Sigma}_1})/(\lambda_3^{\mathbf{\Sigma}_1} + \lambda_1^{\mathbf{\Sigma}_1} + \lambda_2^{\mathbf{\Sigma}_1}) = (2{,}9288 + 1{,}0201)/(2{,}9288 + 1{,}0201 + 0{,}1150) = 0{,}9717$). La transformation linéaire correspondante sera donc la suivante :

$$\mathbf{z} = \mathbf{W}_1^T(\mathbf{x} - \boldsymbol{\mu}_1), \quad \mathbf{W}_1 = \begin{bmatrix} -0{,}4809 & 0{,}3189 \\ -0{,}4680 & -0{,}8811 \\ -0{,}7414 & 0{,}3493 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 0{,}5 \\ 2 \end{bmatrix}.$$

Pour la classe C_2, on peut utiliser que la première composante principale car elle capture 90 % de la variance ($\lambda_1^{\mathbf{\Sigma}_2}/(\lambda_1^{\mathbf{\Sigma}_2} + \lambda_3^{\mathbf{\Sigma}_2} + \lambda_2^{\mathbf{\Sigma}_2}) = 3{,}6260/(3{,}6260 + 0{,}3673 + 0{,}0162) = 0{,}9044$). La transformation linéaire correspondante sera donc la suivante :

$$\mathbf{z} = \mathbf{W}_2^T(\mathbf{x} - \boldsymbol{\mu}_2), \quad \mathbf{W}_2 = \begin{bmatrix} 0{,}6651 \\ 0{,}6599 \\ 0{,}3494 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -1 \\ 3 \end{bmatrix}.$$

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

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

$$\begin{aligned}
\mathbf{\Sigma}_1 &= \mathbf{W}_1\mathbf{D}_1\mathbf{W}_1^T \\
&= \begin{bmatrix} -0{,}4809 & 0{,}3189 & -0{,}8167 \\ -0{,}4680 & -0{,}8811 & -0{,}0684 \\ -0{,}7414 & 0{,}3493 & 0{,}5729 \end{bmatrix} \begin{bmatrix} 2{,}9288 & 0 & 0 \\ 0 & 1{,}0201 & 0 \\ 0 & 0 & 0{,}1150 \end{bmatrix} \begin{bmatrix} -0{,}4809 & -0{,}4680 & -0{,}7414 \\ 0{,}3189 & -0{,}8811 & 0{,}3493 \\ -0{,}8167 & -0{,}0684 & 0{,}5729 \end{bmatrix} \\
&= \begin{bmatrix} 0{,}8578 & 0{,}3790 & 1{,}1041 \\ 0{,}3790 & 1{,}4340 & 0{,}6978 \\ 1{,}1041 & 0{,}6978 & 1{,}7721 \end{bmatrix} \\
\mathbf{\Sigma}_2 &= \mathbf{W}_2\mathbf{D}_2\mathbf{W}_2^T \\
&= \begin{bmatrix} 0{,}6651 & -0{,}2103 & 0{,}7165 \\ 0{,}6599 & -0{,}2835 & -0{,}6958 \\ 0{,}3494 & 0{,}9356 & -0{,}0497 \end{bmatrix} \begin{bmatrix} 3{,}6260 & 0 & 0 \\ 0 & 0{,}3673 & 0 \\ 0 & 0 & 0{,}0162 \end{bmatrix} \begin{bmatrix} 0{,}6651 & 0{,}6599 & 0{,}3494 \\ -0{,}2103 & -0{,}2835 & 0{,}9356 \\ 0{,}7165 & -0{,}6958 & -0{,}0497 \end{bmatrix} \\
&= \begin{bmatrix} 1{,}6286 & 1{,}6053 & 0{,}7698 \\ 1{,}6053 & 1{,}6164 & 0{,}7392 \\ 0{,}7698 & 0{,}7392 & 0{,}7642 \end{bmatrix}
\end{aligned}$$

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

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

$P(C_2)\boldsymbol{\Sigma}_2$.

$$\begin{aligned}\boldsymbol{\Sigma} &= P(C_1)\boldsymbol{\Sigma}_1 + P(C_2)\boldsymbol{\Sigma}_2 \\ &= 0{,}75\begin{bmatrix} 0{,}8578 & 0{,}3790 & 1{,}1041 \\ 0{,}3790 & 1{,}4340 & 0{,}6978 \\ 1{,}1041 & 0{,}6978 & 1{,}7721 \end{bmatrix} + 0{,}25\begin{bmatrix} 1{,}6286 & 1{,}6053 & 0{,}7698 \\ 1{,}6053 & 1{,}6164 & 0{,}7392 \\ 0{,}7698 & 0{,}7392 & 0{,}7642 \end{bmatrix} \\ &= \begin{bmatrix} 1{,}0505 & 0{,}6855 & 1{,}0205 \\ 0{,}6855 & 1{,}4796 & 0{,}7081 \\ 1{,}0205 & 0{,}7081 & 1{,}5201 \end{bmatrix}\end{aligned}$$

(5) (d) Supposons que l'on fasse la projection des données sur un vecteur $\mathbf{v}$ à l'aide de la transformation linéaire suivante.

$$z = \mathbf{v}^T\mathbf{x}, \quad \mathbf{v} = \begin{bmatrix} 0{,}4082 \\ 0{,}8165 \\ 0{,}4082 \end{bmatrix}$$

Calculez le pourcentage de la variance des données qui est conservé pour chaque classe suite à cette transformation linéaire.

Solution: On peut calculer la variance de l'information conservé d'une transformation linéaire appliquée à une distribution normale de covariance $\boldsymbol{\Sigma}$ selon l'équation suivante :

$$\mathrm{Var}(\mathbf{v}^T\mathbf{x}) = \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}.$$

Dans le cas de la classe C_1, la variance conservée suite à cette transformation est :

$$\begin{aligned}\mathrm{Var}_1(\mathbf{v}^T\boldsymbol{x}) &= \mathbf{v}^T\boldsymbol{\Sigma}_1\mathbf{v} \\ &= \begin{bmatrix} 0{,}4082 & 0{,}8165 & 0{,}4082 \end{bmatrix}\begin{bmatrix} 0{,}8578 & 0{,}3790 & 1{,}1041 \\ 0{,}3790 & 1{,}4340 & 0{,}6978 \\ 1{,}1041 & 0{,}6978 & 1{,}7721 \end{bmatrix}\begin{bmatrix} 0{,}4082 \\ 0{,}8165 \\ 0{,}4082 \end{bmatrix} \\ &= 2{,}4798.\end{aligned}$$

Comme la variance totale dans les données de la classe C_1 est $(\lambda_3^{\boldsymbol{\Sigma}_1} + \lambda_1^{\boldsymbol{\Sigma}_1} + \lambda_2^{\boldsymbol{\Sigma}_1} = 2{,}9288 + 1{,}0201 + 0{,}1150 = 4{,}0638)$, on conserve donc $(2{,}4798/4{,}0638) = 61{,}0\ \%$ de la variance.
Dans le cas de la classe C_2, la variance conservée est :

$$\begin{aligned}\mathrm{Var}_2(\boldsymbol{v}^T\boldsymbol{x}) &= \mathbf{v}^T\boldsymbol{\Sigma}_2\mathbf{v} \\ &= \begin{bmatrix} 0{,}4082 & 0{,}8165 & 0{,}4082 \end{bmatrix}\begin{bmatrix} 1{,}6286 & 1{,}6053 & 0{,}7698 \\ 1{,}6053 & 1{,}6164 & 0{,}7392 \\ 0{,}7698 & 0{,}7392 & 0{,}7642 \end{bmatrix}\begin{bmatrix} 0{,}4082 \\ 0{,}8165 \\ 0{,}4082 \end{bmatrix} \\ &= 3{,}2956.\end{aligned}$$

Comme la variance totale dans les données de la classe C_2 est $(\lambda_1^{\boldsymbol{\Sigma}_2} + \lambda_3^{\boldsymbol{\Sigma}_2} + \lambda_2^{\boldsymbol{\Sigma}_2} = 3{,}6260 + 0{,}3673 + 0{,}0162 = 4{,}0095)$, on conserve donc $(3{,}2956/4{,}0095) = 82{,}2\ \%$ de la variance.

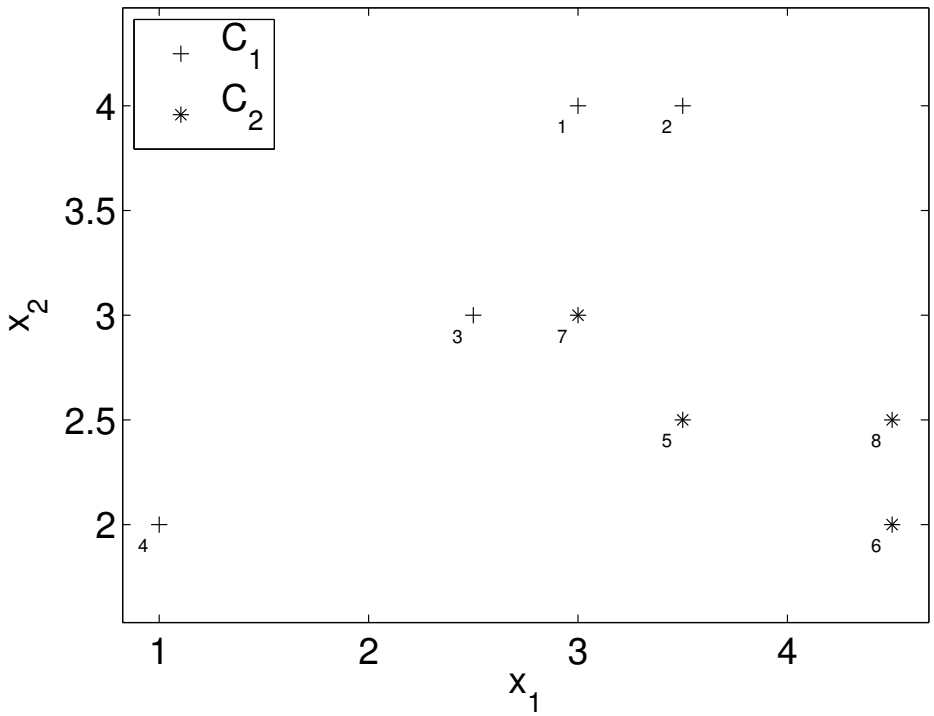
Question 5 (33 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Expliquez pourquoi l'utilisation d'un jeu de données de validation distinct du jeu d'entraînement peut servir à estimer la capacité de généralisation d'un classifieur.

Solution: En testant un classifieur sur un jeu de données distinct du jeu d'entraînement permet d'évaluer la performance du classifieur sur des données nouvelles, qui n'ont pas été vue à l'étape de détermination de la paramétrisation du classifieur. Ainsi, la performance sur le jeu de validation devrait nous donner une idée du niveau de performance sur des données inconnues du même phénomène, ce qui correspond à la notion même de la généralisation.

(3) (b) Donnez le taux de classement par la méthode du plus proche voisin ($k = 1$) avec distance euclidienne, en utilisant l'approche *leave-one-out* expliquée en classe, sur les données présentées dans la figure ici-bas.



Solution: Les données 3 et 7 sont mutuellement le plus proche voisin de l'autre, alors qu'elles appartiennent à des classes différentes. Les autres données ont comme plus proche voisin une donnée de la même classe. En conséquence, 6 données sur 8 sont bien classées par la méthode du plus proche voisin avec *leave-one-out*, pour un taux de classement de $(6/8) = 75\ \%$.

(3) (c) Expliquez pourquoi l'estimateur selon un maximum de vraisemblance de la variance σ^2

d'une loi normale unidimensionnelle $\mathcal{N}(\mu,\sigma^2)$ est biaisé.

Solution: L'espérance de l'estimateur s^2 selon un maximum de vraisemblance de la variance d'une loi normale unidimensionnelle est biaisé étant donné que sa valeur est différente de la véritable valeur σ^2 de cette variance. En effet :

$$\mathbb{E}_{\mathcal{X}}[s^2] = \frac{N-1}{N}\sigma^2 \neq \sigma^2.$$

(3) (d) Donnez la complexité algorithmique (en notation grand $O(\cdot)$, soit la borne supérieure sur le temps d'exécution) de la condensation de Hart, lorsque appliquée à un jeu de N données, en justifiant votre réponse. Vous pouvez supposer que les distances entre toutes les paires de données ont été calculées et stockées dans une structure de donnée avant l'exécution de la condensation de Hart.

Solution: À chaque itération, l'algorithme de la condensation de Hart passe au travers de toutes les données du jeu (complexité $O(N)$). Également, pour chaque données traitée à une itération, on détermine le plus proche voisin dans l'ensemble des prototypes sélectionnées jusqu'à présent (complexité $O(\log N)$). Et dans le pire des cas, la condensation de Hart va ajouter une donnée du jeu à chaque itération, jusqu'à ce que toutes les données soient ajoutées à l'ensemble de prototypes (complexité $O(N)$). Ainsi, complexité algorithmique totale de la méthode est $O(N \times \log N \times N) = O(N^2 \log N)$.

(3) (e) Expliquez pourquoi dit-on que l'analyse en composante principale (ACP) est une approche non supervisée, alors que l'analyse discriminante linéaire (ADL) est une approche supervisée.

Solution: L'ACP est dites non supervisée comme elle ne fait aucune utilisation des étiquettes de classes pour le calcul de la transformation linéaire à effectuer, se basant uniquement sur l'information contenues dans les données. En contre-partie, l'ADL est une méthode supervisée comme elle exploite les données en se basant sur leurs étiquettes de classes afin de déterminer la paramétrisation de la transformation linéaire résultant.

(3) (f) 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 prononcées 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) (g) On dit que les heuristiques de sélections de caractéristiques voraces, comme la sélection avant séquentielle, peuvent ne pas converger à la solution optimale. Présentez une situation générale illustrant bien cette possibilité.

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 ensembles, il est alors possible de détecter que ces variables offrent un gain en performance conjointement substantiel, faisant partie de la solution optimale.

(3) (h) Expliquez à quoi correspondent les arêtes formant une tesselation de Voronoï dans un contexte de classement avec la règle du plus proche voisin.

Solution: Les arêtes d'une tesselation de Voronoï correspond à une ligne exactement à mi-chemin de deux prototypes (données d'entraînement) utilisée pour le classement par la règle du plus proche voisin. En termes de classement, les arêtes à mi-chemin entre deux prototypes d'étiquettes de classe différentes correspondent aux frontières de décision.

(3) (i) On dit que l'algorithme K-means minimise l'erreur de reconstruction des données. Expliquez à quoi correspond cette erreur de reconstruction avec cet algorithme.

Solution: L'erreur de reconstruction correspond à la différence entre la valeur des données d'origines et la valeur des centres à lesquels ils sont associés. Ainsi, si un centre est très près des données appartenant à son groupe, l'erreur de reconstruction sera faible, et inversement. K-means visent à établir les centres de groupes pour lequel les distances aux données du groupe sont minimales.

(3) (j) Dans un contexte de régression polynomiale multivariée, on se limite très souvent à l'utilisation de polynômes d'ordre 1, ce qui correspond à une régression linéaire multivariée. Expliquez pourquoi il est généralement peu intéressant d'effectuer de la régression polynomiale dans un contexte multivarié, avec des polynômes d'ordre supérieur à 1.

Solution: Dans un contexte de régression polynomiale multivariée, le nombre de paramètres que l'on doit estimer augmente très rapidement selon l'ordre du polynôme utilisé. Ainsi, la complexité de la fonction de régression augmente significativement, ce qui augmente significativement le risque de sur-apprentissage des données d'entraînement. En ce sens, un polynôme d'ordre 1 est souvent amplement suffisant pour obtenir des résultats satisfaisants.

(3) (k) Expliquez ce que l'on peut conclure d'un classifieur dont on évalue que le biais est élevé, sous la perspective du compromis biais-variance.

Solution: Lorsque le biais d'un classifieur est élevé, on peut conclure que ce classifieur est trop simple ou pas assez complexe pour bien apprendre le jeu de données utilisé. Dans ce cas, il faudrait donc opter pour un classifieur ayant une capacité de modélisation plus élevée afin d'éviter de faire du sous-apprentissage.