Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Hiver 2014

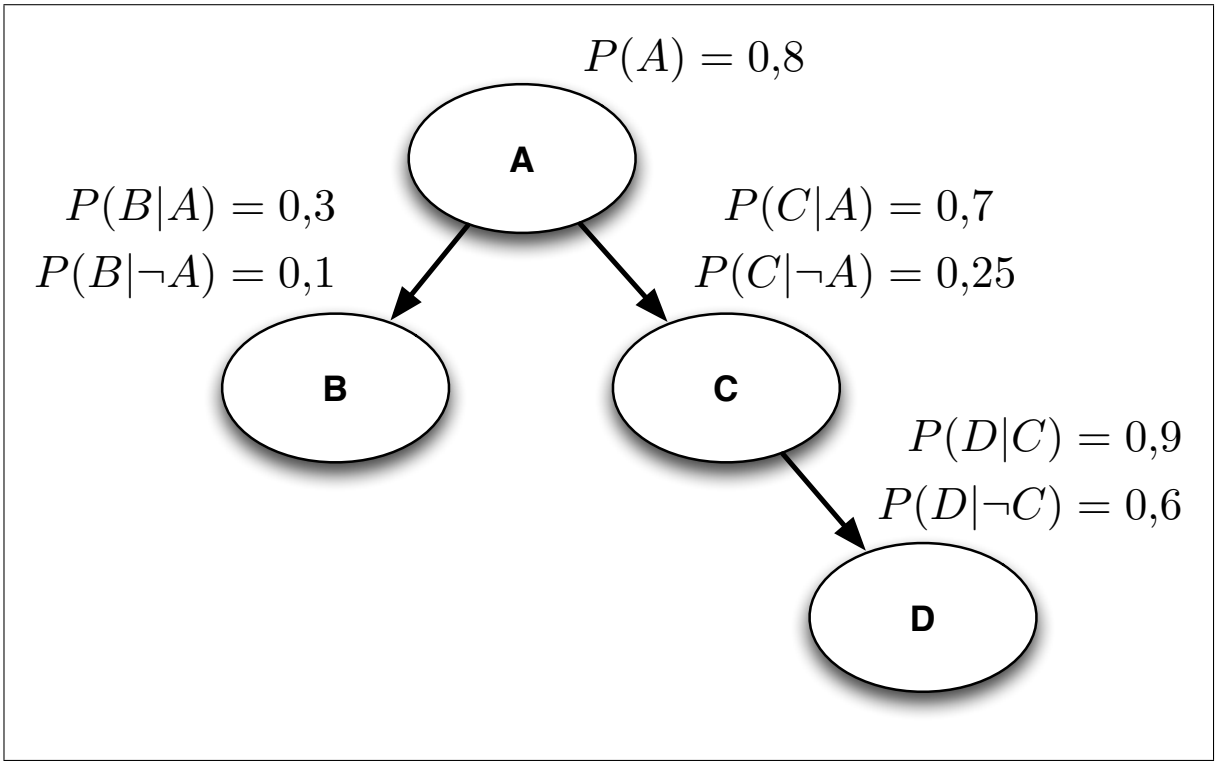
UNIVERSITÉ LAVAL

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 $P(B|\neg C)$.

Solution:

$$\begin{aligned}
P(C) &= P(C|A)\,P(A) + P(C|\neg A)\,P(\neg A) \\
&= 0{,}7 \times 0{,}8 + 0{,}25 \times (1 - 0{,}8) = 0{,}61 \\
P(A|\neg C) &= \frac{P(\neg C|A)\,P(A)}{P(\neg C)} \\
&= \frac{(1 - 0{,}7) \times 0{,}8}{(1 - 0{,}61)} = 0{,}61528 \\
P(B|\neg C) &= P(B|\neg C,A)\,P(A|\neg C) + P(B|\neg C,\neg A)\,P(\neg A|\neg C) \\
&= P(B|A)\,P(A|\neg C) + P(B|\neg A)\,P(\neg A|\neg C) \\
&= 0{,}3 \times 0{,}61528 + 0{,}1 \times (1 - 0{,}61528) \\
&= 0{,}22308
\end{aligned}$$

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

Solution:

$$\begin{aligned}
P(B) &= P(B|A)\,P(A) + P(B|\neg A)\,P(\neg A) \\
&= 0{,}3 \times 0{,}8 + 0{,}1 \times (1 - 0{,}8) = 0{,}26 \\
P(A|B) &= \frac{P(B|A)\,P(A)}{P(B)} \\
&= \frac{0{,}3 \times 0{,}8}{0{,}26} = 0{,}92308 \\
P(C|B) &= P(C|A,B)\,P(A|B) + P(C|\neg A,B)\,P(\neg A|B) \\
&= P(C|A)\,P(A|B) + P(C|\neg A)\,P(\neg A|B) \\
&= 0{,}7 \times 0{,}92308 + 0{,}25 \times (1 - 0{,}92308) = 0{,}66539 \\
P(D|B) &= P(D|B,C)\,P(C|B) + P(D|B,\neg C)\,P(\neg C|B) \\
&= P(D|C)\,P(C|B) + P(D|\neg C)\,P(\neg C|B) \\
&= 0{,}9 \times 0{,}66539 + 0{,}6 \times (1 - 0{,}66539) \\
&= 0{,}79962
\end{aligned}$$

Question 2 (18 points sur 100)

Supposons que l'on fait du classement paramétrique selon deux classes et une variable en entrée (x scalaire), en modélisant les données de chaque classe par une loi normale :

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma^2}\right], i = 1{,}2.$$

La variance de chaque classe est la même, $\sigma_1 = \sigma_2 = \sigma$, alors que les moyennes μ_i et les

probabilités a priori $P(C_i)$ sont différentes pour chaque classe. Sans perte de généralité, vous pouvez supposer que la moyenne de la classe 1 est inférieure à la classe 2, $\mu_1 < \mu_2$.

(12) (a) Avec cette modélisation, donnez l'équation analytique décrivant les frontières de décision entre les deux classes, en supposant une fonction de perte zéro-un (valeur égale pour les deux types d'erreurs). En une dimension, de telles frontières se résument à des seuils sur la valeur de x. Indiquez également de quelle façon le classement se fait dans les différentes régions séparées par les frontières.

Solution: L'équation du discriminant correspondant à la modélisation est :

$$\mathrm{h}_i(x) = \frac{p(x|C_i)\,P(C_i)}{p(x)},\ i \in \{1,2\}.$$

La frontière de décision entre les deux classes correspond au point où $\mathrm{h}_1(x) = \mathrm{h}_2(x)$:

$$\begin{aligned}
\mathrm{h}_1(x) &= \mathrm{h}_2(x) \\
\frac{p(x|C_1)\,P(C_1)}{p(x)} &= \frac{p(x|C_2)\,P(C_2)}{p(x)} \\
p(x|C_1)\,P(C_1) &= p(x|C_2)\,P(C_2) \\
\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right]P(C_1) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right]P(C_2) \\
\exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right]P(C_1) &= \exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right]P(C_2)
\end{aligned}$$

En appliquant le logarithme naturel de chaque côté de l'équation, on obtient :

$$\begin{aligned}
\ln\left[\exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right]P(C_1)\right] &= \ln\left[\exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right]P(C_2)\right] \\
\ln\exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right] + \ln(P(C_1)) &= \ln\exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right] + \ln(P(C_2)) \\
-\frac{(x-\mu_1)^2}{2\sigma^2} + \ln(P(C_1)) &= -\frac{(x-\mu_2)^2}{2\sigma^2} + \ln(P(C_2))
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2\sigma^2}\left[(x-\mu_2)^2 - (x-\mu_1)^2\right] + \ln(P(C_1)) - \ln(P(C_2)) &= 0 \\
\frac{1}{2\sigma^2}\left[x^2 - 2x\mu_2 + \mu_2^2 - (x^2 - 2x\mu_1 + \mu_1^2)\right] + \ln\left[\frac{P(C_1)}{P(C_2)}\right] &= 0 \\
\frac{1}{2\sigma^2}\left[-2(\mu_2-\mu_1)x + \mu_2^2 - \mu_1^2\right] + \ln\left[\frac{P(C_1)}{P(C_2)}\right] &= 0 \\
\frac{1}{2\sigma^2}\left[-2(\mu_2-\mu_1)x + (\mu_2-\mu_1)(\mu_2+\mu_1)\right] + \ln\left[\frac{P(C_1)}{P(C_2)}\right] &= 0 \\
\frac{\mu_2-\mu_1}{2\sigma^2}\left[-2x + \mu_1 + \mu_2\right] + \ln\left[\frac{P(C_1)}{P(C_2)}\right] &= 0
\end{aligned}$$

$$\begin{aligned} -2x &= -\mu_1 - \mu_2 - \frac{2\sigma^2}{\mu_2 - \mu_1} \ln\left[\frac{P(C_1)}{P(C_2)}\right] \\ x &= \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_2 - \mu_1} \ln\left[\frac{P(C_1)}{P(C_2)}\right] \end{aligned}$$

La frontière de décision consiste donc en un seuil θ donné par :

$$\theta = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_2 - \mu_1} \ln\left[\frac{P(C_1)}{P(C_2)}\right],$$

où les données pour lesquelles $x < \theta$ sont assignées à la classe C_1 et autrement à la classe C_2.

(6) (b) Supposons maintenant que l'on utilise une fonction avec une perte variable selon le type d'erreur, en utilisant une perte $\lambda_{1,2}$ lorsque l'on assigne une donnée de la classe C_1 alors qu'elle appartient à la classe C_2, et inversement en utilisant une perte $\lambda_{2,1}$ lorsqu'une donnée est assignée à C_2 alors qu'elle appartient à C_1. Déterminez la frontière de décision correspondant à ce cas.

Solution: Avec une fonction de perte variant le coût selon l'erreur, la frontière de décision correspondante est donnée au point où :

$$\begin{aligned} \lambda_{1,2} P(C_2|x) &= \lambda_{2,1} P(C_1|x), \\ \lambda_{1,2} \frac{p(x|C_2)\, P(C_2)}{p(x)} &= \lambda_{2,1} \frac{p(x|C_1)\, P(C_1)}{p(x)}, \\ p(x|C_2)\, \lambda_{1,2}\, P(C_2) &= p(x|C_1)\, \lambda_{2,1}\, P(C_1). \end{aligned}$$

On peut alors récupérer les développements de la question précédente, en substituant dans le résultat $\lambda_{1,2}\, P(C_2)$ pour $P(C_2)$ et $\lambda_{2,1}\, P(C_1)$ pour $P(C_1)$, ce qui donne comme θ à utiliser :

$$\theta = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_2 - \mu_1} \ln\left[\frac{\lambda_{2,1}\, P(C_1)}{\lambda_{1,2}\, P(C_2)}\right],$$

où les données pour lesquelles $x < \theta$ sont assignées à la classe C_1 et autrement à la classe C_2.

Question 3 (12 points sur 100)

Soit le jeu de données des Iris de Fisher, comprenant 150 données en 4 dimensions et organisées selon trois classes. Les estimations du vecteur moyen $\mathbf{m}$ et de la matrice de covariance $\mathbf{S}$ pour l'ensemble des données de ce jeu, en ne tenant pas compte des étiquettes de classes, sont

les suivantes :

$$\mathbf{m} = \begin{bmatrix} 0{,}6093 \\ -0{,}0727 \\ 1{,}3404 \\ 0{,}5694 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 0{,}6857 & -0{,}0393 & 1{,}2737 & 0{,}5169 \\ -0{,}0393 & 0{,}1880 & -0{,}3217 & -0{,}1180 \\ 1{,}2737 & -0{,}3217 & 3{,}1132 & 1{,}2964 \\ 0{,}5169 & -0{,}1180 & 1{,}2964 & 0{,}5824 \end{bmatrix}.$$

De plus, les vecteurs propres $\mathbf{c}_i$ et valeurs propres λ_i associées extraits de la matrice de covariance $\mathbf{S}$ sont :

$$\mathbf{c}_1 = \begin{bmatrix} 0{,}6565 \\ 0{,}7297 \\ -0{,}1758 \\ -0{,}0747 \end{bmatrix}, \quad \mathbf{c}_2 = \begin{bmatrix} 0{,}5810 \\ -0{,}5964 \\ -0{,}0725 \\ -0{,}5491 \end{bmatrix}, \quad \mathbf{c}_3 = \begin{bmatrix} 0{,}3616 \\ -0{,}0823 \\ 0{,}8566 \\ 0{,}3588 \end{bmatrix}, \quad \mathbf{c}_4 = \begin{bmatrix} -0{,}3173 \\ 0{,}3241 \\ 0{,}4797 \\ -0{,}7511 \end{bmatrix},$$
$$\lambda_1 = 0{,}2422, \qquad \lambda_2 = 0{,}0785, \qquad \lambda_3 = 4{,}2248, \qquad \lambda_4 = 0{,}0237.$$

(6) (a) Donner la transformation linéaire $\mathbf{z} = \mathrm{f}(\mathbf{x})$ permettant de projeter les données dans un espace à **deux** dimensions, tout en maximisant la variance conservée. Pour ce faire, donnez l'équation matricielle correspondant à la transformation linéaire et détaillez les valeurs et la signification de chacune des variables de cette équation.

Solution: L'équation correspondant à la transformation linéaire permettant d'effectuer la réduction de la dimensionnalité est :

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \mathbf{m}),$$

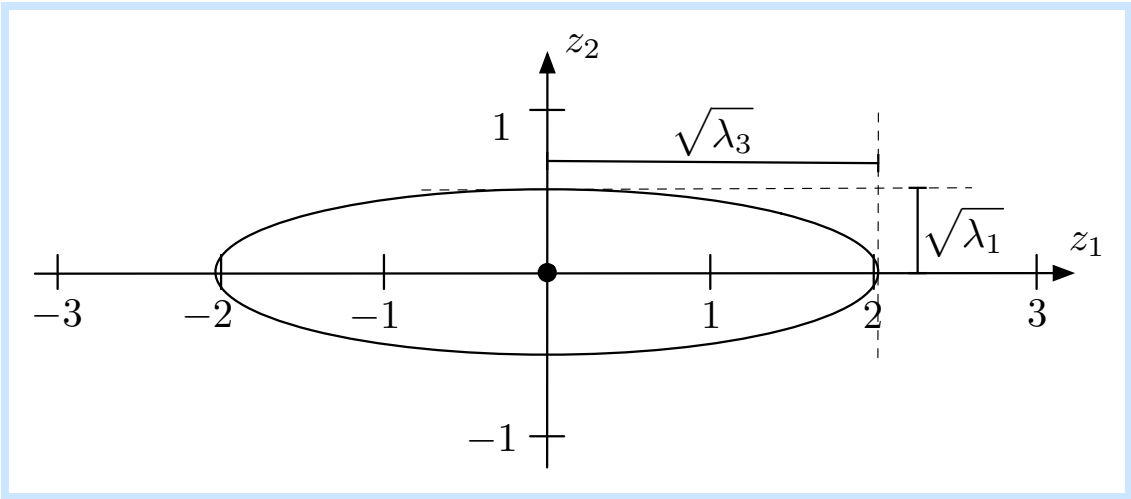
où $\mathbf{m}$ est le vecteur moyen donné dans l'énoncé de la question et la matrice $\mathbf{W}$ est construite avec les deux vecteurs propres correspondant aux valeurs propres les plus élevées, soit $\mathbf{c}_3$ (valeur propre associée de $\lambda_3 = 4{,}2248$) et $\mathbf{c}_1$ (valeur propre associée de $\lambda_1 = 0{,}2422$), $\mathbf{W} = [\mathbf{c}_3\ \mathbf{c}_1]$.

(6) (b) Tracez la courbe de contour correspondant à une distance de Mahanalobis de 1 de la distribution des données dans l'espace à deux dimensions de la sous-question précédente (l'espace des $\mathbf{z}$). Annotez votre tracé en y ajoutant toute l'information nécessaire pour préciser la position, taille et orientation de la courbe de contour.

Solution: La distribution des données dans l'espace transformé correspond à une matrice de covariance diagonale, dont les valeurs sur la diagonale correspondent aux valeurs propres utilisées pour effectuer l'analyse en composante principale :

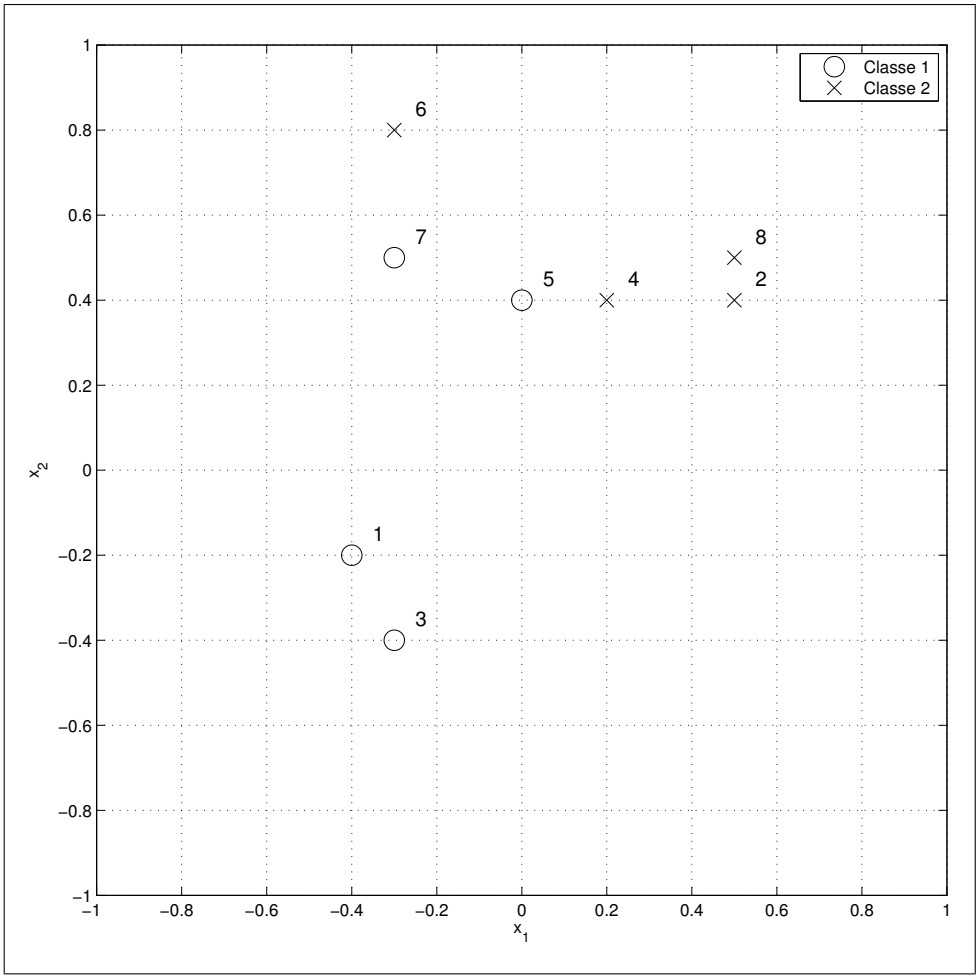
$$\mathbf{S}_{ACP} = \begin{bmatrix} \lambda_3 & 0 \\ 0 & \lambda_1 \end{bmatrix} = \begin{bmatrix} 4{,}2248 & 0 \\ 0 & 0{,}2422 \end{bmatrix}.$$

L'écart-type correspondant selon l'axe z_1 est donc de $\sqrt{\lambda_3} = 2{,}0554$ et selon l'axe z_2 de $\sqrt{\lambda_1} = 0{,}4922$. Les données sont centrées sur l'origine. Le tracé de la courbe de contour correspondant à une distance de Mahanalobis de 1 est le suivant.



Question 4 (12 points sur 100)

Soit les données suivantes, en deux dimensions.



(6) (a) Calculez le taux d'erreur de classement selon une approche *leave-one-out* avec un classifieur de type k-plus proches voisins, en employant $k = 3$ voisins et la distance de Manhattan. Explicitez la démarche menant au calcul du taux d'erreur.

Solution:

Donnée	3-PPV-LOO	Étiquette majoritaire	Erreur ?
$\mathbf{x}^1$	$\mathbf{x}^3$, $\mathbf{x}^5$, $\mathbf{x}^7$	1	Non
$\mathbf{x}^2$	$\mathbf{x}^4$, $\mathbf{x}^5$, $\mathbf{x}^8$	2	Non
$\mathbf{x}^3$	$\mathbf{x}^1$, $\mathbf{x}^5$, $\mathbf{x}^7$	1	Non
$\mathbf{x}^4$	$\mathbf{x}^2$, $\mathbf{x}^5$, $\mathbf{x}^8$	2	Non
$\mathbf{x}^5$	$\mathbf{x}^2$, $\mathbf{x}^4$, $\mathbf{x}^7$	2	Oui
$\mathbf{x}^6$	$\mathbf{x}^4$, $\mathbf{x}^5$, $\mathbf{x}^7$	1	Oui
$\mathbf{x}^7$	$\mathbf{x}^4$, $\mathbf{x}^5$, $\mathbf{x}^6$	2	Oui
$\mathbf{x}^8$	$\mathbf{x}^2$, $\mathbf{x}^4$, $\mathbf{x}^5$	2	Non

Donc, il y a trois erreurs d'effectuées sur les huit instances, pour un taux d'erreur de classement de $37{,}5\,\%$.

(6) (b) Effectuez une édition de Wilson de ce jeu de données, en utilisant un voisin ($k = 1$) et une distance euclidienne. Traitez les données dans leur ordre d'indice, c'est-à-dire dans l'ordre x^1,x^2,x^3,...,x^8. Explicitez votre démarche et rapportez les données formant l'ensemble des prototypes après l'édition.

Solution:

Donnée	Prototypes	PPV-LOO	Étiquette	Retrait ?
$\mathbf{x}^1$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^3$	1	Non
$\mathbf{x}^2$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^8$	2	Non
$\mathbf{x}^3$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^1$	1	Non
$\mathbf{x}^4$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^4,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	1	Oui
$\mathbf{x}^5$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^7$	1	Non
$\mathbf{x}^6$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^6,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^7$	1	Oui
$\mathbf{x}^7$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^5$	1	Non
$\mathbf{x}^8$	$\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$	$\mathbf{x}^2$	2	Non

Donc, l'ensemble de prototypes résultant de l'édition de Wilson est $\{\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\mathbf{x}^5,\mathbf{x}^7,\mathbf{x}^8\}$.

Question 5 (18 points sur 100)

Supposons que l'on veut appliquer l'algorithme Espérance-Maximisation (EM) à un jeu de données à D dimensions, où chaque groupe $\mathcal{G}_i$ est décrit par une loi normale $\mathcal{N}_D(\boldsymbol{\mu}_i,\mathbf{I})$, dont la covariance correspond à la matrice identité :

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

De plus, les probabilités a priori sont égales pour les K groupes, $P(\mathcal{G}_i) = 1/K$, $i = 1,\ldots,K$.

(8) (a) Donnez le développement complet de l'équation permettant l'estimation $\mathbf{m}_i$ du vecteur moyen $\boldsymbol{\mu}_i$ selon un maximum de vraisemblance.

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

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

(5) (b) Donnez les équations décrivant les calculs effectués aux étapes E et M de l'algorithme correspondant à cette modélisation.

Solution: Étape E, calcul des h_i^t :

$$\begin{aligned}
h_i^t &= P(\mathcal{G}_i|\mathbf{x}^t) = \frac{p(\mathbf{x}^t|\mathcal{G}_i)\, P(\mathcal{G}_i)}{\sum_k p(\mathbf{x}^t|\mathcal{G}_k)\, P(\mathcal{G}_k)} = \frac{p(\mathbf{x}^t|\mathcal{G}_i) 1/K}{\sum_k p(\mathbf{x}^t|\mathcal{G}_k)\, 1/K} = \frac{p(\mathbf{x}^t|\mathcal{G}_i)}{\sum_k p(\mathbf{x}^t|\mathcal{G}_k)} \\
&= \frac{\frac{1}{(2\pi)^{0,5D}} \exp \left[-\frac{\sum_j (x_j - \mu_{i,j})^2}{2} \right]}{\sum_k \frac{1}{(2\pi)^{0,5D}} \exp \left[-\frac{\sum_j (x_j - \mu_{k,j})^2}{2} \right]} = \frac{\exp \left[-\frac{\sum_j (x_j - \mu_{i,j})^2}{2} \right]}{\sum_k \exp \left[-\frac{\sum_j (x_j - \mu_{k,j})^2}{2} \right]}, \forall i, \forall t.
\end{aligned}$$

Étape M, calcul des paramètres $\mathbf{m}_i$ developpé à la sous-question précédente :

$$\mathbf{m}_i = \frac{\sum_t h_i^t \mathbf{x}^t}{\sum_t h_i^t}, \forall i.$$

(5) (c) Quelle modification supplémentaire à l'algorithme EM basé sur cette modélisation reviendrait à l'algorithme K-means présenté en classe ?

Solution: Pour que cet algorithme EM revienne à un algorithme K-means, il faudrait que les appartenances h_i^t soient binaires, en faisant en sorte qu'une donnée soit associée uniquement au groupe dont la probabilité a posteriori est la plus grande :

$$b_i^t = \begin{cases} 1 & \text{si } i = \text{argmax}_j \, h_j^t \\ 0 & \text{autrement} \end{cases} .$$

Étant donné le calcul de h_i^t donné à la sous-question précédente, ce revient à dire que :

$$b_i^t = \begin{cases} 1 & \text{si } i = \text{argmin}_j \, \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{autrement} \end{cases} .$$

En utilisant donc ce b_i^t à la place de h_i^t pour le calcul de l'appartenance, on retrouve l'algorithme K-means.

Question 6 (30 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Expliquez en quoi consiste une option de rejet lorsque l'on fait du classement bayésien.

Solution: L'option de rejet consiste à permettre une décision de rejet où le classifieur n'associe pas la donnée à une classe, mais la rejette. La décision de rejet est prise lorsque le risque de faire une erreur est élevé de sorte que le rejet a un coût plus faible que d'associer l'instance à la classe la plus probable (ou un risque moindre, dans le cas de coûts variables selon le type d'erreur).

(3) (b) Expliquez quel résultat on devrait obtenir lorsque la variance d'un estimateur d'un paramètre est élevée, selon le compromis biais-variance.

Solution: Lorsque la variance d'un estimateur est élevée, on observe que l'estimation du paramètre va varier considérablement d'un jeu d'échantillons à l'autre, même si ces échantillons sont tous récoltés à partir du même phénomène.

(3) (c) Supposons que l'on fait de la régression linéaire sur des données à D dimensions. Dites combien de paramètres (valeurs scalaires) décrivant le modèle de régression doivent être estimés.

Solution: Un modèle de régression linéaire de données à D dimensions correspond à :

$$\text{h}(\mathbf{x}|\theta) = w_D \, x_D + \ldots + w_2 \, x_2 + w_1 \, x_1 + w_0.$$

Donc, la paramétrisation $\theta = \{w_0, w_1, w_2, \ldots, w_D\}$ du modèle comporte $D + 1$ valeurs scalaires.

(3) (d) 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 :

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

(3) (e) Dans PRTools, quelle est l'utilité de la fonction `testc` ?

Solution: La fonction `testc` permet d'évaluer l'erreur de classement d'un classifieur (*mapping* entraîné) sur un certain jeu de données.

(3) (f) Lorsqu'on mesure une corrélation nulle entre deux variables, peut-on dire que ces variables sont indépendantes ? Justifiez brièvement votre réponse.

Solution: Non, lorsque la corrélation entre deux variables est nulle, on ne peut pas dire que ces variables sont indépendantes. Une relation complexe, non linéaire peut exister entre ces variables, sans être capturée par la mesure de corrélation.

(3) (g) Soit l'algorithme de sélection avant séquentielle pour la sélection de caractéristiques. Lorsque l'on utilise cet algorithme selon une approche enveloppe (*wrapper*), chaque sous-ensemble de caractéristiques considéré requiert l'entraînement d'un classifieur pour en évaluer la performance. Supposons que l'on veut sélectionner 10 caractéristiques parmi les 20 disponibles, combien d'entraînements de classifieurs seront nécessaires avec l'algorithme de sélection avant séquentielle ?

Propriété pour simplifier vos calculs : $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.

Solution: Selon l'algorithme de sélection avant séquentielle, si l'on a D variables pour démarrer, on devrait tester ces D variables pour déterminer celle offrant les meilleures performances. Ensuite, on va tester laquelle parmi les $D-1$ restantes celle qui offre la meilleure performance en conjonction de la première variable sélectionnée, et ainsi de suite. Donc, lorsque l'algorithme de sélection avant séquentielle est exécuté jusqu'au bout, soit jusqu'à la sélection des D variables du jeu de données, le nombre de sous-ensembles testés sera de $\sum_{i=1}^{D}(D-i+1)$. Si l'on s'arrête lorsque K variables sont sélectionnées, le nombre de sous-ensembles testés sera de :

$$\sum_{i=1}^{K}(D-i+1) = K\,(D+1) - \sum_{i=1}^{K} i = K\,(D+1) - \frac{K(K+1)}{2}.$$

Donc, pour $K = 10$ et $D = 20$, le nombre d'entraînements de classifieurs sera de :

$$K\,(D+1) - \frac{K(K+1)}{2} = 10 \times (20+1) - \frac{10 \times (10+1)}{2} = 210 - 55 = 155.$$

(3) (h) Quelle est l'utilité d'un multiplicateur de Lagrange lorsque l'on optimise une fonction ?

Solution: Le multiplicateur de Lagrange permet de trouver le maximum ou minimum d'une certaine fonction $\mathrm{f}(\mathbf{x})$ tout en respectant des contraintes exprimées sous forme d'égalité $\mathrm{g}(\mathbf{x}) = 0$ (l'approche est également valide pour des inégalités sous la forme $\mathrm{g}(\mathbf{x}) \geq 0$). Une équation augmentée par un multiplicateur de Lagrange à la forme :

$$L(\mathbf{x},\lambda) = \mathrm{f}(\mathbf{x}) + \lambda \mathrm{g}(\mathbf{x}),$$

avec le maximum (ou minimum) trouvé aux dérivées partielles égales à zéro :

$$\nabla L(\mathbf{x},\lambda) = 0.$$

(3) (i) Indiquez ce que représente l'équation suivante, permettant de faire du classement avec des densités-mélanges :

$$p(\mathbf{x}|C_i) = \sum_{j=1}^{k_i} p(\mathbf{x}|\mathcal{G}_{i,j}) P(\mathcal{G}_{i,j}).$$

Prenez soin d'expliquer le sens des variables formant l'équation.

Solution: Cette équation représente une densité de probabilité correspondant à la vraisemblance que la donnée $\mathbf{x}$ appartient à la classe C_i. Cette densité de probabilité est un mélange des densités des k_i groupes $\mathcal{G}_{i,j}$ associés à la classe C_i, chaque densité représentant un groupe.

(3) (j) Expliquez pourquoi une estimation non paramétrique de distribution par histogramme ne fonctionne pas bien lorsque la dimensionnalité des données est élevée.

Solution: Une estimation de distribution par histogramme exige de diviser l'espace des données en bins. Le volume des bins est important, car plus le bin est de petite taille, plus l'espace d'entrée est finement modélisé, et donc plus l'approximation de la distribution sera juste. D'autre part, avec des bins de petite taille, on doit utiliser beaucoup de bins pour modéliser un certain volume, ce qui requiert d'autant plus beaucoup de données pour avoir suffisamment de données dans chaque bin.

Avec l'augmentation de la dimensionnalité des données, la taille de l'espace augmente exponentiellement. Pour pouvoir conserver le volume des bins fixe, il faudrait donc augmenter considérablement le nombre de bins nécessaires à l'estimation de la distribution, et donc la quantité de données nécessaire pour estimer la valeur de chaque bin. Avec une croissance exponentielle de la taille de l'espace selon la dimensionnalité, et le nombre de données nécessaires pour en faire une estimation par histogramme valable devient insoutenable pour que la méthode soit utilisable en pratique.