Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de Génie Électrique et de Génie Informatique
Automne 2009

UNIVERSITÉ LAVAL

EXAMEN PARTIEL

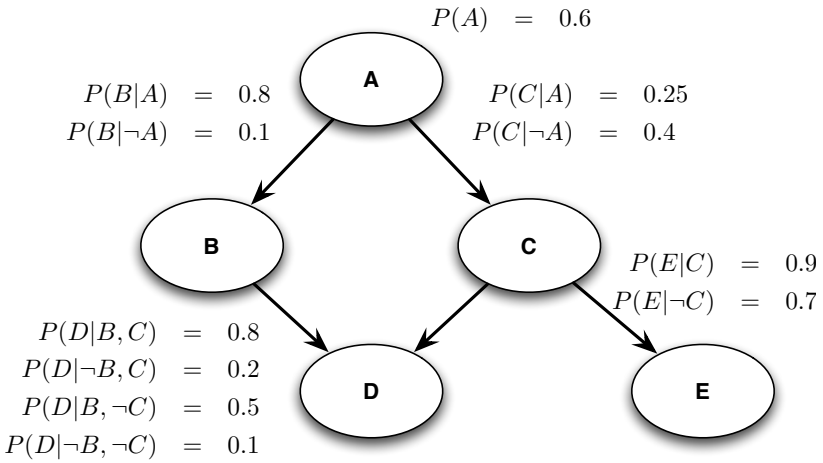
Instructions : – Aucune documentation sauf aide mémoire d'une page recto-verso manuscrit
– Ce questionnaire comporte 4 pages et 7 questions
– Durée de l'examen : 2 heure 30 minutes

Pondération : Cet examen compte pour 35% de la note finale.

---

1. Calculez les probabilités suivantes selon le réseau bayésien donné ci-bas.

   (a) (2pts) $P(C)$
   (b) (2pts) $P(D)$
   (c) (2pts) $P(B|\neg D)$
   (d) (2pts) $P(C|B)$
   (e) (2pts) $P(E|B)$



Solution :

$$\begin{aligned} P(C) &= P(C|A)P(A) + P(C|\neg A) + P(\neg A) \\ &= 0.25 \times 0.6 + 0.4 \times (1 - 0.6) = 0.15 + 0.16 = 0.31 \end{aligned}$$

$$\begin{aligned} P(B,C) &= P(B,C|A)P(A) + P(B,C|\neg A)P(\neg A) \\ &= P(B|A)P(C|A)P(A) + P(B|\neg A)P(C|\neg A)P(\neg A) \\ &= (0.8 \times 0.25 \times 0.6) + (0.1 \times 0.4 \times (1 - 0.6)) \\ &= 0.12 + 0.016 = 0.136 \\ P(\neg B,C) &= P(\neg B,C|A)P(A) + P(\neg B,C|\neg A)P(\neg A) \\ &= P(\neg B|A)P(C|A)P(A) + P(\neg B|\neg A)P(C|\neg A)P(\neg A) \\ &= ((1 - 0.8) \times 0.25 \times 0.6) + ((1 - 0.1) \times 0.4 \times (1 - 0.6)) \\ &= 0.03 + 0.144 = 0.174 \end{aligned}$$

$$
\begin{aligned}
P(B,\neg C) &= P(B,\neg C|A)P(A) + P(B,\neg C|\neg A)P(\neg A) \\
&= P(B|A)P(\neg C|A)P(A) + P(B|\neg A)P(\neg C|\neg A)P(\neg A) \\
&= (0.8 \times (1-0.25) \times 0.6) + (0.1 \times (1-0.4) \times (1-0.6)) \\
&= 0.36 + 0.024 = 0.384 \\
P(\neg B,\neg C) &= P(\neg B,\neg C|A)P(A) + P(\neg B,\neg C|\neg A)P(\neg A) \\
&= P(\neg B|A)P(\neg C|A)P(A) + P(\neg B|\neg A)P(\neg C|\neg A)P(\neg A) \\
&= ((1-0.8) \times (1-0.25) \times 0.6) + ((1-0.1) \times (1-0.4) \times (1-0.6)) \\
&= 0.09 + 0.216 = 0.306 \\
P(D) &= P(D|B,C)P(B,C) + P(D|\neg B,C)P(\neg B,C) \\
&\quad + P(D|B,\neg C)P(B,\neg C) + P(D|\neg B,\neg C)P(\neg B,\neg C) \\
&= (0.8 \times 0.136) + (0.2 \times 0.174) + (0.5 \times 0.384) + (0.1 \times 0.306) \\
&= 0.1088 + 0.0348 + 0.192 + 0.0306 = 0.3662
\end{aligned}
$$

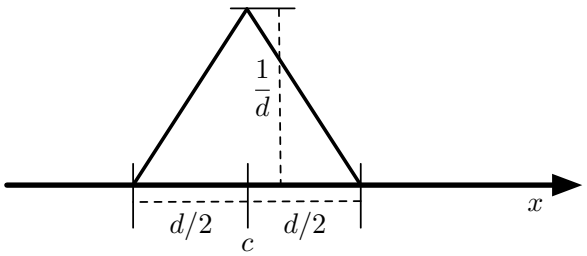
$$
\begin{aligned}
P(B) &= P(B|A)P(A) + P(B|\neg A)P(\neg A) \\
&= 0.8 \times 0.6 + 0.1 \times (1-0.6) = 0.48 + 0.04 = 0.52 \\
P(D|B) &= P(D|B,C)P(C) + P(D|B,\neg C)P(\neg C) \\
&= (0.8 \times 0.31) + (0.5 \times (1-0.31)) = 0.593 \\
P(B|\neg D) &= \frac{P(\neg D|B)P(B)}{P(\neg D)} = \frac{(1-P(D|B))P(B)}{(1-P(D))} \\
&= \frac{(1-0.593) \times 0.52}{(1-0.3662)} = 0.333922
\end{aligned}
$$

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} = \frac{0.8 \times 0.6}{0.52} = 0.923077 \\
P(C|B) &= P(C|B,A)P(A|B) + P(C|B,\neg A)P(\neg A|B) \\
&= P(C|A)P(A|B) + P(C|\neg A)P(\neg A|B) \\
&= 0.25 \times 0.923077 + 0.4 \times (1-0.923077) = 0.261538
\end{aligned}
$$

$$
\begin{aligned}
P(E|B) &= P(E|B,C)P(C|B) + P(E|B,\neg C)P(\neg C|B) \\
&= P(E|C)P(C|B) + P(E|\neg C)(1-P(C|B)) \\
&= 0.9 \times 0.261538 + 0.7 \times (1-0.261538) = 0.752308
\end{aligned}
$$

2. (10pts) Soit une densité de probabilité « triangle » en une dimension, décrite par l'équation suivante et la figure ci-bas.

$$
p(x|c,d) = \begin{cases} \frac{2x-2c+d}{d^2} & x \in [c-0.5d, c[\\ \frac{-2x+2c+d}{d^2} & x \in [c, c+0.5d] \\ 0 & \text{autrement} \end{cases}
$$



Supposons que vous avez des données pouvant appartenir à deux classes, où les données suivent une densité de probabilité « triangle » pour chaque classe. Supposons également que vous connaissez les valeurs des paramètres de ces densités, soit c_1, d_1, c_2 et d_2. Donnez les régions de décision dans tout le domaine de $x \in \mathbb{R}$ selon les différentes configurations de paramètres de densités possibles, c'est-à-dire les différents domaines de décisions de classement selon la valeur x. Vous pouvez supposer que $c_1 < c_2$ et que les probabilités *a priori* sont égales, $P(C_1) = P(C_2) = 0.5$.

Solution : Comme $P(C_1) = P(C_2)$, le classement se fait en utilisant la fonction discriminante suivante.

$$\mathrm{h}(x|c_1, d_1, c_2, d_2) = p(x|c_1, d_1) - p(x|c_2, d_2)$$

Le signe de $\mathrm{h}(x)$ change lorsque $p(x|c_1, d_1) - p(x|c_2, d_2) = 0$. On veut donc déterminer les points où $p(x|c_1, d_1) = p(x|c_2, d_2)$, ce qui correspond aux frontières de décisions.

Ces densités peuvent être égales en trois points.

– Le premier point $\theta_a < c_1$.

$$\begin{aligned}
p(x|c_1, c_2) &= p(x|c_2, d_2) \\
\frac{2x - 2c_1 + d_1}{d_1^2} &= \frac{2x - 2c_2 + d_2}{d_2^2} \\
2x - 2c_1 + d_1 &= (2x - 2c_2 + d_2)\frac{d_1^2}{d_2^2} \\
2\frac{d_2^2 - d_1^2}{d_2^2}x &= (-2c_2 + d_2)\frac{d_1^2}{d_2^2} + 2c_1 - d_1 \\
x &= \frac{(-2c_2 + d_2)d_1^2 + (2c_1 - d_1)d_2^2}{d_2^2}\frac{d_2^2}{2(d_2^2 - d_1^2)} \\
&= \frac{(-2c_2 + d_2)d_1^2 + (2c_1 - d_1)d_2^2}{2(d_2^2 - d_1^2)} = \theta_a
\end{aligned}$$

– Le deuxième point $c_1 < \theta_b < c_2$.

$$\begin{aligned}
p(x|c_1, c_2) &= p(x|c_2, d_2) \\
\frac{-2x + 2c_1 + d_1}{d_1^2} &= \frac{2x - 2c_2 + d_2}{d_2^2} \\
-2x + 2c_1 + d_1 &= (2x - 2c_2 + d_2)\frac{d_1^2}{d_2^2} \\
-2\frac{d_1^2 + d_2^2}{d_1^2}x &= (-2c_2 + d_2)\frac{d_2^2}{d_2^2} - (2c_1 + d_1) \\
x &= \left((-2c_2 + d_2)\frac{d_1^2}{d_2^2} - (2c_1 + d_1)\right)\frac{d_2^2}{-2(d_1^2 + d_2^2)} \\
&= \frac{(2c_1 + d_1)d_2^2 - (-2c_2 + d_2)d_1^2}{2(d_1^2 + d_2^2)} = \theta_b
\end{aligned}$$

– Le troisième point $\theta_c > c_2$.

$$\begin{aligned}
p(x|c_1, c_2) &= p(x|c_2, d_2) \\
\frac{-2x + 2c_1 + d_1}{d_1^2} &= \frac{-2x + 2c_2 + d_2}{d_2^2} \\
-2x + 2c_1 + d_1 &= (-2x + 2c_2 + d_2)\frac{d_1^2}{d_2^2} \\
\frac{-2d_2^2 + 2d_1^2}{d_2^2}x &= (2c_2 + d_2)\frac{d_1^2}{d_2^2} - (2c_1 + d_1) \\
x &= \frac{(2c_2 + d_2)d_1^2 - (2c_1 + d_1)d_2^2}{d_2^2}\frac{d_2^2}{-2d_2^2 + 2d_1^2} \\
&= \frac{(2c_2 + d_2)d_1^2 - (2c_1 + d_1)d_2^2}{-2(d_2^2 + d_1^2)} = \theta_c
\end{aligned}$$

Selon les valeurs de c_1, d_1, c_2 et d_2, différentes décisions peuvent être prises.

– Lorsque $(c_2 - 0.5d_2) < (c_1 - 0.5d_1)$:
  – Si $(c_2 - 0.5d_2) \leq x < \theta_a$ alors $x \in C_2$
  – Si $\theta_a \leq x < \theta_b$ alors $x \in C_1$
  – Si $\theta_b \leq x < (c_2 + 0.5d_2)$ alors $x \in C_2$
  – Autrement, rejet
– Lorsque $(c_1 + 0.5d_1) > (c_2 + 0.5d_2)$:
  – Si $(c_1 - 0.5d_1) \leq x < \theta_b$ alors $x \in C_1$
  – Si $\theta_b \leq x < \theta_c$ alors $x \in C_2$
  – Si $\theta_c \leq x < (c_1 + 0.5d_1)$ alors $x \in C_1$
  – Autrement, rejet
– Lorsque $(c_1 + 0.5d_1) < (c_2 - 0.5d_2)$:
  – Si $(c_1 - 0.5d_1) \leq x < (c_1 + 0.5d_1)$ alors $x \in C_1$
  – Si $(c_2 - 0.5d_2) \leq x < (c_2 + 0.5d_2)$ alors $x \in C_2$
  – Autrement, rejet
– Autrement :
  – Si $(c_1 - 0.5d_1) \leq x < \theta_b$ alors $x \in C_1$
  – Si $\theta_b \leq x < (c_2 + 0.5d_2)$ alors $x \in C_2$
  – Autrement, rejet

3. Soit une densité de probabilité $p(x|\sigma^2)$, décrite par l'équation suivante.

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

Cette densité de probabilité correspond à une loi normale de moyenne nulle, $p(x|\sigma^2) \sim \mathcal{N}(0, \sigma^2)$.

(a) (10pts) Déterminez s^2, l'estimateur par un maximum de vraisemblance du paramètre σ^2 de la densité de probabilité. Donnez le développement complet des équations nécessaire pour obtenir cet estimateur.

Solution : La fonction de log-vraisemblance pour cette densité de probabilité est la suivante.

$$\mathcal{L}(\sigma|\mathcal{X}) = \sum_{t=1}^{N} \log p(x|\sigma^2)$$

On calcule d'abord la dérivée de la fonction de log-vraisemblance selon σ.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma} &= \frac{\partial}{\partial \sigma} \sum_{t=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x^t)^2}{2\sigma^2}\right] \\
&= \frac{\partial}{\partial \sigma} - N \log \sqrt{2\pi} + \frac{\partial}{\partial \sigma} - N \log \sigma + \frac{\partial}{\partial \sigma} \sum_{t=1}^{N} \left[-\frac{(x^t)^2}{2\sigma^2}\right] \\
&= -\frac{N}{\sigma} + \sum_{t=1}^{N} -(-2)\frac{(x^t)^2}{2\sigma^3} \\
&= -\frac{N}{\sigma} + \sum_{t=1}^{N} \frac{(x^t)^2}{\sigma^3}
\end{aligned}$$

L'estimateur s^2 est obtenu lorsque la valeur de la dérivée de la fonction de log-vraisemblance est égale à zéro.

$$\begin{aligned}
-\frac{N}{s} + \sum_{t=1}^{N} \frac{(x^t)^2}{s^3} &= 0 \\
\sum_{t=1}^{N} \frac{(x^t)^2}{s^3} &= \frac{N}{s} \\
\sum_{t=1}^{N} (x^t)^2 &= Ns^2 \\
s^2 &= \sum_{t=1}^{N} \frac{(x^t)^2}{N}
\end{aligned}$$

(b) (5pts) Déterminez si cet estimateur s^2 du paramètre σ^2 est biaisé. Donnez le développement complet des équations vous permettant de répondre à cette question. Notez que comme la moyenne de cette densité de probabilité est nulle, $\mu = E[x^t] = 0$, alors $\sigma^2 = E[(x^t)^2] - E[x^t]^2 = E[(x^t)^2]$.

Solution : Le biais d'un estimateur est donné par l'équation suivante.

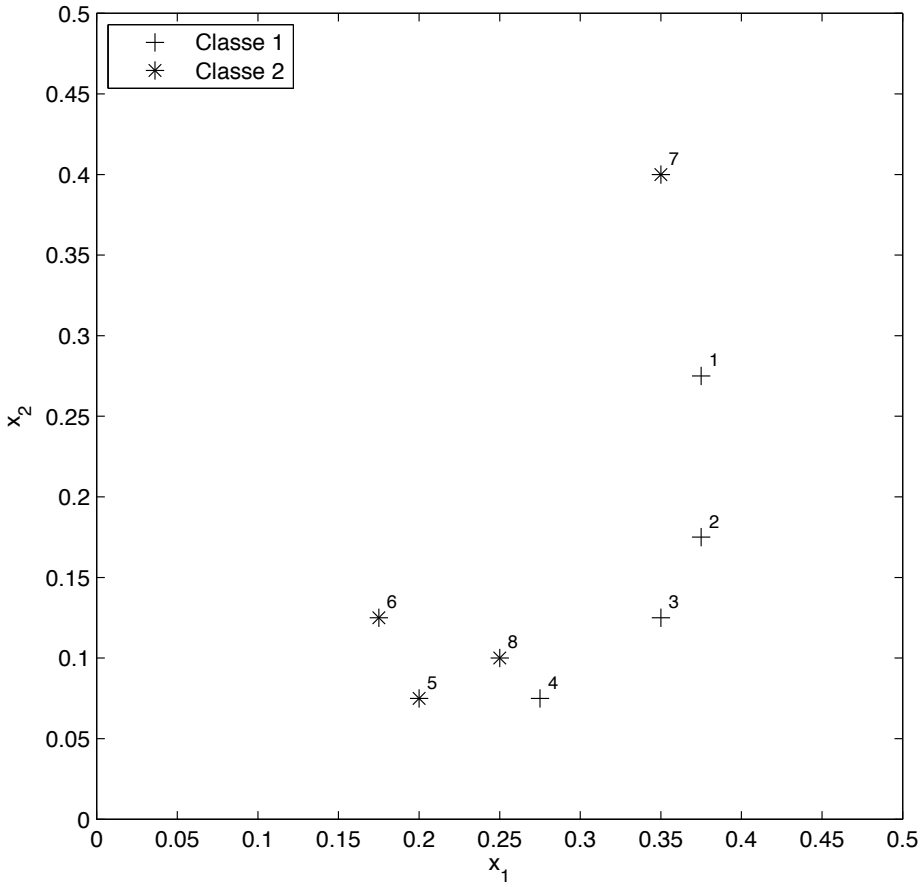
$$b_{\sigma^2(s^2)} = E[s^2(\mathcal{X})] - \sigma^2$$

En développant $E[s^2(\mathcal{X})]$:

$$\begin{aligned}
E[s^2(\mathcal{X})] &= E\left[\sum_{t=1}^{N} \frac{(x^t)^2}{N}\right] \\
&= \frac{\sum_{t=1}^{N} E\left[(x^t)^2\right]}{N} = \frac{NE\left[(x^t)^2\right]}{N} = E\left[(x^t)^2\right] \\
&= \sigma^2
\end{aligned}$$

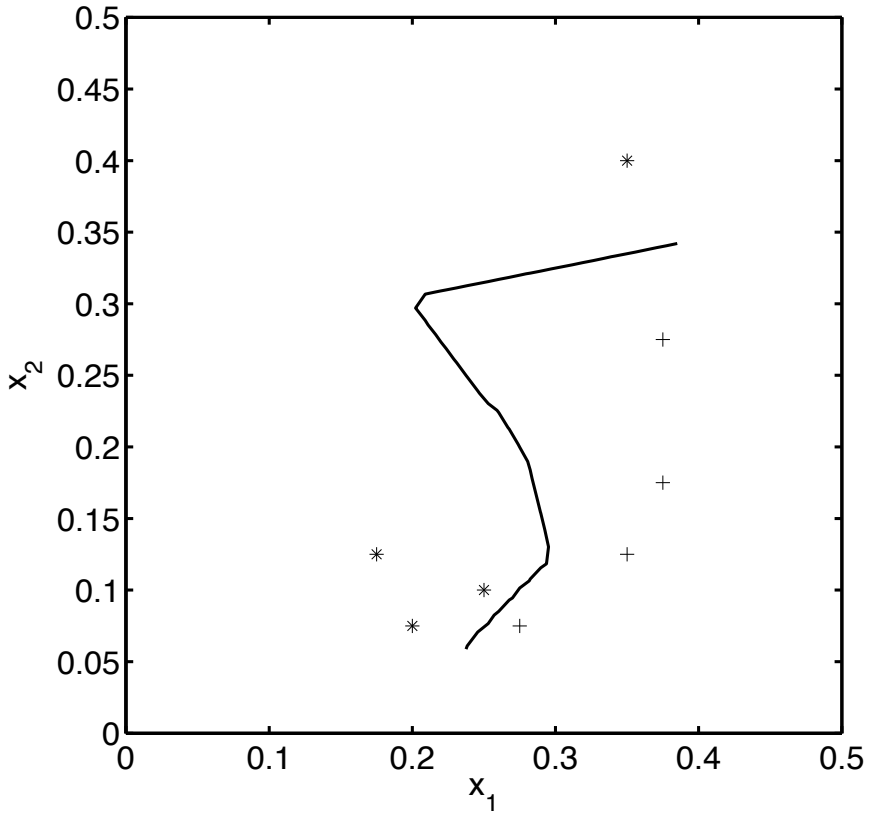
Le biais est nul car $b_{\sigma^2(s^2)} = E[s^2(\mathcal{X})] - \sigma^2 = 0$, donc cet estimateur n'est pas biaisé.

4. Soit les données en deux dimensions et deux classes, présentées dans le graphique suivant.



(a) (5pts) Donnez le taux d'erreur de classement de ces données avec un classement par le plus proche voisin ($k = 1$), selon la méthode *leave-one-out*, en utilisant une distance euclidienne.
Solution : Les données 4, 7 et 8 sont mal classées par la méthodes *leave-one-out*, avec $k = 1$ voisin. Ceci donne donc un taux d'erreur de classement de $3/8 = 37.5\%$.

(b) (5pts) Faites un tracé approximatif des frontières de décisions obtenues par un classement par le plus proche voisin ($k = 1$), en utilisant une distance euclidienne.
Solution :



(c) (5pts) Indiquez les données qui seraient retirées par une sélection de prototype basée sur une édition de Wilson, en utilisant $k = 3$ voisins. Les données sont traitées dans leur ordre croissant d'indice, c'est-à-dire $1, 2, \ldots, 8$.

Solution :
Édition de Wilson appliquée aux données :

Donnée	3-PPV	Mal classé	Prototypes sélectionnés
1	$\{2, 3, 7\}$	Non	$\{1, 2, 3, 4, 5, 6, 7, 8\}$
2	$\{1, 3, 4\}$	Non	$\{1, 2, 3, 4, 5, 6, 7, 8\}$
3	$\{2, 4, 8\}$	Non	$\{1, 2, 3, 4, 5, 6, 7, 8\}$
4	$\{3, 5, 8\}$	Oui	$\{1, 2, 3, 5, 6, 7, 8\}$
5	$\{3, 6, 8\}$	Non	$\{1, 2, 3, 5, 6, 7, 8\}$
6	$\{3, 5, 8\}$	Non	$\{1, 2, 3, 5, 6, 7, 8\}$
7	$\{1, 2, 3\}$	Oui	$\{1, 2, 3, 5, 6, 8\}$
8	$\{3, 5, 6\}$	Non	$\{1, 2, 3, 5, 6, 8\}$

L'ensemble des prototypes sélectionnés résultant est : $\{1, 2, 3, 5, 6, 8\}$.

5. Soit le jeu de données des *Iris de Fisher*, comportant quatre dimensions, où le vecteur moyen et la matrice de covariance sont estimés comme suit.

$$\mathbf{m} = \begin{bmatrix} 11.9 \\ 37.8 \\ 30.6 \\ 58.4 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 199.2 & 578.4 & 352.3 & 747.7 \\ 578.4 & 1741.7 & 1120.9 & 2335.7 \\ 352.3 & 1120.9 & 952.5 & 1781.6 \\ 747.7 & 2335.7 & 1781.6 & 3483.9 \end{bmatrix}$$

Les vecteurs propres et valeurs propres associées à la matrice de covariance $\mathbf{S}$ sont les suivants.

$$\mathbf{w}_1 = \begin{bmatrix} -0.3365 \\ -0.7107 \\ 0.5476 \\ 0.2861 \end{bmatrix}, \quad \mathbf{w}_2 = \begin{bmatrix} -0.5920 \\ -0.0269 \\ -0.6484 \\ 0.4778 \end{bmatrix}, \quad \mathbf{w}_3 = \begin{bmatrix} 0.7131 \\ -0.4782 \\ -0.3686 \\ 0.3563 \end{bmatrix}, \quad \mathbf{w}_4 = \begin{bmatrix} 0.1667 \\ 0.5153 \\ 0.3793 \\ 0.7503 \end{bmatrix}$$
$$\lambda_1 = 211.9, \qquad \lambda_2 = 7.9, \qquad \lambda_3 = 2.7, \qquad \lambda_4 = 6154.7$$

(a) (5pts) Donnez l'équation complète, avec valeurs numériques, de la transformation linéaire permettant de réduire au minimum la dimensionnalitée des données $\mathbf{x} \in \mathbb{R}^4$ tout en conservant au minimum 95% de la variance.

Solution : Les valeurs propres triées en ordre décroissant sont les suivantes.

$$\lambda_4 = 6154.7,\ \lambda_1 = 211.9,\ , \lambda_2 = 7.9,\ \lambda_3 = 2.7$$

La proportion de variance conservé en utilisant K dimensions sur D est donné par l'équation suivante.

$$\text{pdv} = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{i=1}^{D} \lambda_i}, \text{ avec } \lambda_k \geq \lambda_l,\ \forall k \leq l$$

Le tableau suivant donne la proportion de variance concervée pour les différentes valeurs de k.

K	1	2	3	4
pdv	96.51%	99.83%	99.96%	100%

Donc, une analyse en composantes principales utilisant uniquement la composante principale $\mathbf{w}_4$ associée à la valeur propre la plus grande, λ_4 serait suffisante pour capturer plus de 95% de la variance.

L'équation correspondant à la transformation linéaire avec uniquement la première composante principale est la suivante.

$$\begin{aligned} \mathbf{m} &= [11.9\ \ 37.8\ \ 30.6\ \ 58.4]^T \\ \mathbf{w}_4 &= [0.1667\ \ 0.5153\ \ 0.3793\ \ 0.7503]^T \\ z &= \mathbf{w}_4^T(\mathbf{x} - \mathbf{m}) \end{aligned}$$

(b) (5pts) Donnez l'équation complète, avec valeurs numériques, de la transformation linéaire permettant de réduire la dimensionnalitée des données de quatre dimensions ($\mathbf{x} \in \mathbb{R}^4$) à deux dimensions ($\mathbf{z} \in \mathbb{R}^2$), en conservant un maximum d'information des données, tout en normalisant les données pour obtenir une moyenne nulle, une variance unitaire et une covariance nulle, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.

Solution : Pour faire une réduction de dimensionnalité de $\mathbb{R}^4 \rightarrow \mathbb{R}^2$, en concervant un maximum de variance tout normalisant les données pour obtenir une variance unitaire, il faut utiliser les deux composantes principales, soit $\mathbf{w}_4$ et $\mathbf{w}_1$, tout en les normalisant par la racine carrée de l'inverse de leurs valeurs propres, $\lambda_i^{-0.5}$.

$$\begin{aligned} \mathbf{m} &= [11.9\ \ 37.8\ \ 30.6\ \ 58.4]^T \\ \mathbf{A} &= \begin{bmatrix} \frac{\mathbf{w}_4}{\sqrt{\lambda_4}} & \frac{\mathbf{w}_1}{\sqrt{\lambda_1}} \end{bmatrix}^T = \begin{bmatrix} \frac{0.1667}{\sqrt{6154.7}} & \frac{-0.3365}{\sqrt{211.9}} \\ \frac{0.5153}{\sqrt{6154.7}} & \frac{-0.7107}{\sqrt{211.9}} \\ \frac{0.3793}{\sqrt{6154.7}} & \frac{0.5476}{\sqrt{211.9}} \\ \frac{0.7503}{\sqrt{6154.7}} & \frac{0.2861}{\sqrt{211.9}} \end{bmatrix}^T = \begin{bmatrix} 0.0021 & -0.0231 \\ 0.0066 & -0.0488 \\ 0.0048 & 0.0376 \\ 0.0096 & 0.0197 \end{bmatrix}^T \\ \mathbf{z} &= \mathbf{A}(\mathbf{x} - \mathbf{m}) \end{aligned}$$

6. Soit les données suivantes, en deux dimensions.
$$\mathbf{x}^1 = [-0.6\ \ 0.05]^T, \quad \mathbf{x}^2 = [-0.45\ \ 0.3]^T, \quad \mathbf{x}^3 = [-0.4\ \ -0.05]^T, \quad \mathbf{x}^4 = [-0.3\ \ 0.1]^T$$

   (a) (5pts) Appliquez une itération de l'algorithme K-means à ces données, en utilisant $K = 2$ centres, avec comme position initiale des groupes $\mathbf{m}_1 = [-0.45\ \ 0.3]^T$ et $\mathbf{m}_2 = [-0.3\ \ 0.1]^T$.

   Solution :
   Étape E, calcul des b_i^t :

$\mathbf{x}^t$	$\|\mathbf{x}^t - \mathbf{m}_1\|^2$	$\|\mathbf{x}^t - \mathbf{m}_2\|^2$	b_1^t	b_2^t
$[-0.6\ \ 0.05]^T$	0.0850	0.0925	1	0
$[-0.45\ \ 0.3]^T$	0	0.0625	1	0
$[-0.4\ \ -0.05]^T$	0.1250	0.0325	0	1
$[-0.3\ \ 0.1]^T$	0.0625	0	0	1

   Étape M, calcul des $\mathbf{m}_i$:

$$\begin{aligned}
\mathbf{m}_1 &= \frac{1}{2}\left(\mathbf{x}^1 + \mathbf{x}^2\right) = \frac{1}{2}\left([-0.6\ \ 0.05]^T + [-0.45\ \ 0.3]^T\right) = [-0.525\ \ 0.175]^T \\
\mathbf{m}_2 &= \frac{1}{2}\left(\mathbf{x}^3 + \mathbf{x}^4\right) = \frac{1}{2}\left([-0.4\ \ -0.05]^T + [-0.3\ \ 0.1]^T\right) = [-0.35\ \ 0.025]^T
\end{aligned}$$

   (b) (5pts) Expliquez en quoi l'algorithme K-means se compare à un algorithme Espérance-Maximisation (EM) pour des groupes de données suivants des lois multinormales, en insistant sur les similarités et les différences entre ces deux méthodes de clustering.

   Solution : Avec un algorithme EM pour des groupes de données suivants des lois multinormales, l'étape E consiste à estimer les valeurs d'appartenances h_i^t des données $\mathbf{x}^t$ au groupes $\mathcal{G}_i$, selon l'équation suivante.

$$h_i^t = \frac{\pi_i |\mathbf{S}_i|^{-0.5} \exp\left[-0.5(\mathbf{x}^t - \mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}^t - \mathbf{m}_i)\right]}{\sum_j \pi_j |\mathbf{S}_j|^{-0.5} \exp\left[-0.5(\mathbf{x}^t - \mathbf{m}_j)^T \mathbf{S}_j^{-1}(\mathbf{x}^t - \mathbf{m}_j)\right]}$$

   h_i^t est une valeur réelle comprise dans $[0, 1]$. L'équivalent avec K-means classique est la variable b_i^t, qui donne l'association des données $\mathbf{x}^t$ au groupes $\mathcal{G}_i$.

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

   L'association b_i^t est *dure* dans le sens que la valeur de b_i^t pour une donnée $\mathbf{x}^t$ est égale à 1 pour un seul groupe, et est égale à 0 pour tous les autres groupes.

   De plus, avec l'algorithme EM, l'étape M consiste en l'évaluation de trois paramètres distincts selon les h_i^t obtenus à l'étape E, soit les probabilités *a priori* π_i des groupes, les positions des vecteurs moyens $\mathbf{m}_i$ des groupes ainsi que les matrices de covariance $\mathbf{S}_i$ des groupes. Avec l'algorithme K-means, uniquement la position du centre des groupes $\mathbf{m}_i$ est estimée à l'étape M, les autres paramètres n'étant pas considérés.

   Un autre point de vue serait de dire qu'avec l'algorithme K-means est un algorithme EM avec lois multinormales où les appartenances h_i^t sont *durcie* pour être égale à 0 ou 1, que les probabilités *a priori* π_i sont fixées comme étant égales pour chaque groupe et que les matrices de covariances sont partagées entre les groupes et d'égales valeurs pour chaque dimension, c'est-à-dire que $\mathbf{S}_i = \mathbf{S} = s\mathbf{I}$.

7. Répondez aussi brièvement et clairement que possible aux questions suivantes.

(a) (2pts) Soit un jeu de données $\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^N$, donnez l'équation permettant de calculer l'estimation de la matrice de covariance générale et partagée entre les K classes du jeu.

Solution :

$$\mathbf{S} = \sum_{C_i} P(C_i)\mathbf{S}_i = \frac{1}{N}\sum_{C_i}\sum_{\mathbf{x}^t \in C_i}(\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T$$

(b) (2pts) Décrivez à quoi consiste conceptuellement le critère suivant, utilisé par l'analyse discriminante linéaire.

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|}$$

Solution : Ce critère vise à maximiser la distance entre les classes ($|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|$) tout en minimisant la distance entre les données faisant partie d'une même classe ($|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|$).

(c) (2pts) Supposons que l'on veut sélectionner K caractéristiques dans un jeu de données de dimensionnalité D, où $K \ll D$. Indiquez quelle méthode serait normalement la plus rapide au niveau de la charge de calcul, entre une sélection séquentielle avant et une sélection séquentielle arrière, où l'évaluation des performances se fait avec un classifieur entraîné pour chaque sous-ensemble de caractéristiques candidat. Justifiez votre réponse.

Solution : La sélection séquentielle avant serait plus rapide, car les classifiers entraînés pour chaque sous-ensemble de caractéristiques candidat traiteraient des données de dimensionnalité réduite $(1, 2, \ldots, K)$ comparativement à une sélection séquentielle arrière, où les classifiers entraînés traiteraient des données de plus grande dimensionnalité ($D-1, D-2, \ldots, D-K$). Le nombre de classifieurs entraînés et testés par chaque approche serait cependant le même.

(d) (2pts) Avec la régression linéaire univariée présentée en classe, indiquez quel paramètre permet de contrôler la complexité des modèles générés.

Solution : La complexité des modèles de la régression linéaire univariée est contrôlée par l'ordre du polynôme utilisé. Plus l'ordre du polynôme est élevé, plus le modèle généré est complexe.

(e) (2pts) Indiquez ce que représente l'équation suivante, permettant de faire du classement avec des densités-mélanges.

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

Solution : Cette équation représente une densité de probabilité correspondant à la vraisemblance que la donnée $\mathbf{x}$ appartienne 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 étant associée à un groupe.

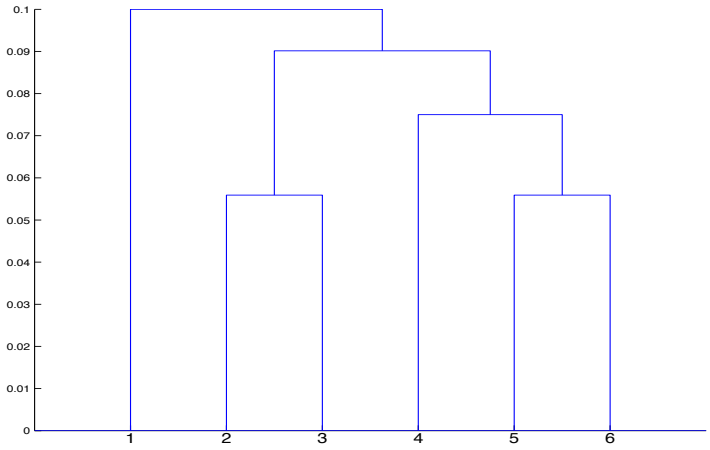
(f) (2pts) Indiquez quelle forme de fonction discriminante on obtient lorsqu'on fait du classement paramétrique basé sur des densités de classes suivants des lois multinormales, où la matrice de covariance est partagée entres toutes les classes.

Solution : On obtient une fonction discriminante linéaire.

(g) (2pts) Donnez la forme des courbes de contour de densités de probabilité suivant une loi multinormale en deux dimensions, où les valeurs de la matrice de covariance hors diagonale sont négatives, $\sigma_{i,j} < 0,\ \forall i \neq j$.

Solution : Les courbes de contour de ces densités ont une forme ellipsoïde étirées dans le sens d'une droite de pente négative (étirées selon les directions Nord-Ouest et Sud-Est).

(h) (2pts) Soit le dendrogramme présenté ci-bas, obtenu par un clustering hiérarchique agglomératif. Supposons que l'on veut former trois groupes à partir de ce dendrogramme, donnez les indices des données formant les groupes.



Solution : Groupe 1 : $\{1\}$, groupe 2 : $\{2, 3\}$ et groupe 3 : $\{4, 5, 6\}$.

(i) (2pts) Expliquez à quoi sert l'espérance de vraisemblance $\mathcal{Q}(\Phi|\Phi^l)$ à l'étape M de l'algorithme EM.

Solution : L'espérance de vraisemblance $\mathcal{Q}(\Phi|\Phi^l)$ est utilisée comme critère à optimiser pour déterminer le nouvel ensemble de paramètres Φ^l selon la paramétrisation actuelle Φ. Ces paramètres Φ représentent la paramétrisation des densité de probabilités de chacun des groupes de données, par exemple les $\mathbf{m}_i$ et $\mathbf{S}_i$ dans le cas de lois multinormales, ainsi que les probabilités *a priori* π_i des groupes.

(j) (2pts) Indiquez en quoi consiste le compromis entre le biais et la variance.

Solution : Ce compromis est lié à la complexité d'un modèle d'apprentissage. Plus le modèle est simple, plus on devrait s'attendre à ce qu'il y ait du sous-apprentissage, ce qui devrait donnez un biais fort mais une variance réduite. En contre partie, si le modèle est plus complexe, le biais devrait diminuer (éventuellement disparaître), alors que la variance du modèle devrait augmenter, pour éventuellement faire sur sur-apprentissage avec des modèles trop complexes pour le problème à résoudre.

(k) (2pts) Donnez la différence entre les modèles génératifs (méthodes paramétriques, non-paramétriques, densités-mélanges, clustering) et les modèles discriminatifs (fonctions discriminantes linéaires, non-linéaires) pour faire du classement.

Solution : Avec les modèles génératifs, on vise à estimer les densités de probabilité des données pour chaque classe, et à utiliser ces estimations de densités pour faire du classement. Avec les modèles discriminatifs, on vise plutôt à obtenir des fonctions paramétrées directement pour discriminer les données, sans égard aux densités de probabilités des données.

(l) (2pts) L'apprentissage de paramètres de classifieurs par une descente du gradient se fait à l'aide de la formule $w = w + \Delta w$, où w est le paramètre appris. Indiquez ce que représente Δw dans cette formule.

Solution : Δw représente la modification appliquée à w par l'itération actuelle de la descente du gradient. Cette modification correspond à changer la valeur de w selon la direction du gradient de la fonction d'erreur $E(w)$ par rapport w, en utilisant un pas η.

$$\Delta w = -\eta \frac{\partial E}{\partial w}$$

(m) (2pts) Soit le problème du *ou exclusif* (XOR), où on veut inférer un classifieur à partir des données suivantes.

$$\mathbf{x}^1 = [0\ \ 0]^T,\ r^1 = 0, \quad \mathbf{x}^2 = [0\ \ 1]^T,\ r^2 = 1$$
$$\mathbf{x}^3 = [1\ \ 0]^T,\ r^3 = 1, \quad \mathbf{x}^4 = [1\ \ 1]^T,\ r^4 = 0$$

Donnez une fonction de base $\phi(\mathbf{x}) = [\phi_1(\mathbf{x})\ \ \phi_2(\mathbf{x})\ \ \cdots\ \ \phi_K(\mathbf{x})]^T$ pouvant résoudre parfaitement ce problème de classement avec la fonction discriminante suivante.

$$\mathrm{h}(\mathbf{x}) = \sum_{i=1}^{K} w_i \phi_i(\mathbf{x})$$

Solution : Une solution serait d'utiliser la fonction de base suivante.

$$\phi(\mathbf{x}) = [x_1\ \ x_2\ \ (x_1 x_2)]^T$$

Une autre solution serait d'utiliser une fonction de base radiale centrées les quatre données du jeu d'entraînement :

$$\phi(\mathbf{x}) = [\phi_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}^1)\ \ \phi_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}^2)\ \ \phi_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}^3)\ \ \phi_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}^4)]^T,$$

où

$$\phi_{\mathrm{RBF}}(\mathbf{x}, \mathbf{m}) = \exp\left[-\frac{\|\mathbf{x} - \mathbf{m}\|^2}{2s_i^2}\right].$$

(n) (2pts) Expliquez comment on peut effectuer une fenêtre de Parzen à l'aide d'un classifieur de type *Radial Basis Function* (RBF), pour un jeu de données $\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$ à deux classes, où $r^t = 1$ si $\mathbf{x}^t \in C_1$ et $r^t = -1$ autrement.

Solution : Il suffit de poser une fonction gaussienne sur chacune des données de l'ensemble d'entraînement, avec N fonctions gaussiennes où $h = s$, $\mathbf{m}_i = \mathbf{x}^i$ et $w_i = r^i/(\sqrt{2\pi}Nh^D)$ pour $i = 1, \ldots, N$.

$$\mathrm{h}(\mathbf{x}) = \sum_{i=1}^{N} w_i \phi_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}Ns^D} \sum_{i=1}^{N} r^i \exp\left[-\frac{\|\mathbf{x} - \mathbf{m}\|^2}{2s^2}\right]$$

(o) (2pts) Des algorithmes tels que l'édition de Wilson, la condensation de Hart ou le K-means sont désignés comme étant des heuristiques. Indiquez en quoi consiste précisemment une heuristique.

Solution : Les heuristiques sont des méthodes simples qui permettent de donner en temps polynomial des réponses satisfaisantes certains problèmes, sans nécessairement donner des réponses optimales.

$\mathcal{FIN}$

6 Novembre 2009
CG