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

EXAMEN PARTIEL

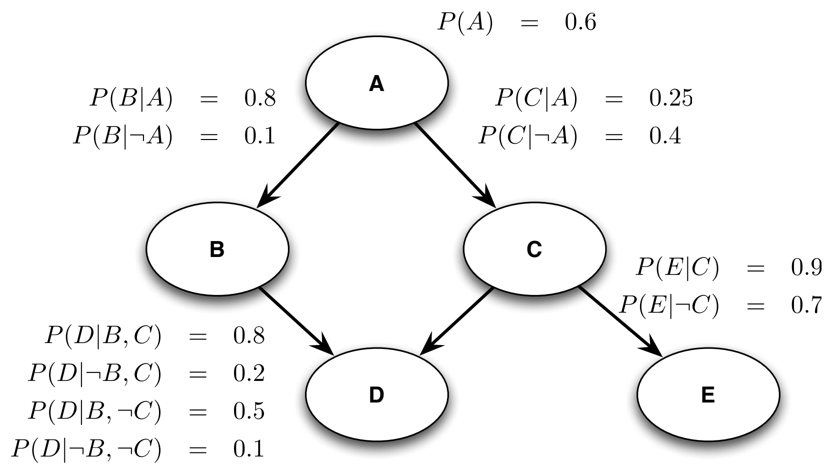
<u>Instructions</u> : – Aucune documentation sauf aide mémoire d'une page recto-verso manuscrit
– Ce questionnaire comporte 4 pages et 7 questions
– Durée de l'examen : 2 heure 30 minutes

<u>Pondération</u> : Cet examen compte pour 35% de la note finale.

---

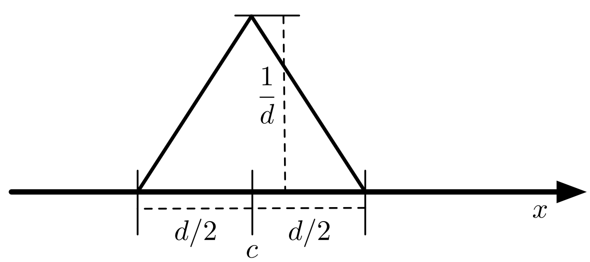
1. Calculez les probabilités suivantes selon le réseau bayésien donné ci-bas.

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



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, $P(C_1) = P(C_2) = 0.5$.

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

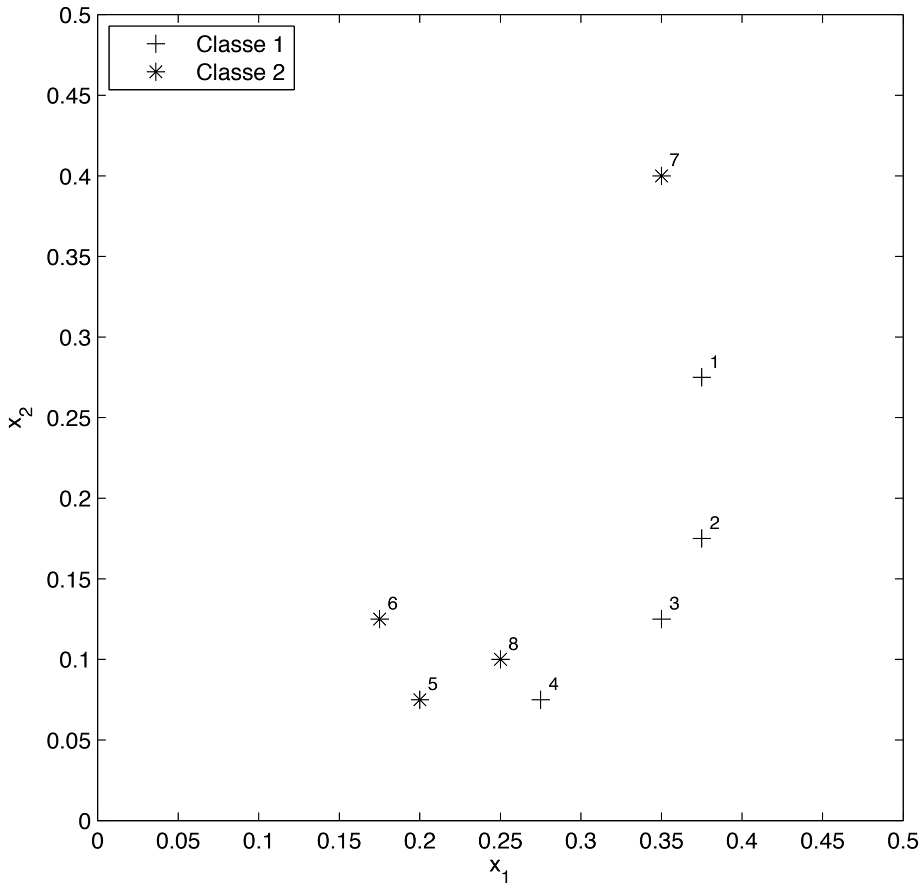
$$p(x|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

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.

(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 $\sigma^2 = E[(x^t)^2] - E[x^t]^2 = E[(x^t)^2]$.

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.

(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.

(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 $1, 2, \ldots, 8$.

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.

(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, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.

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 $\mathbf{m}_2 = [-0.3 \;\; 0.1]^T$.

(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.

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.

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

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

(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.

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

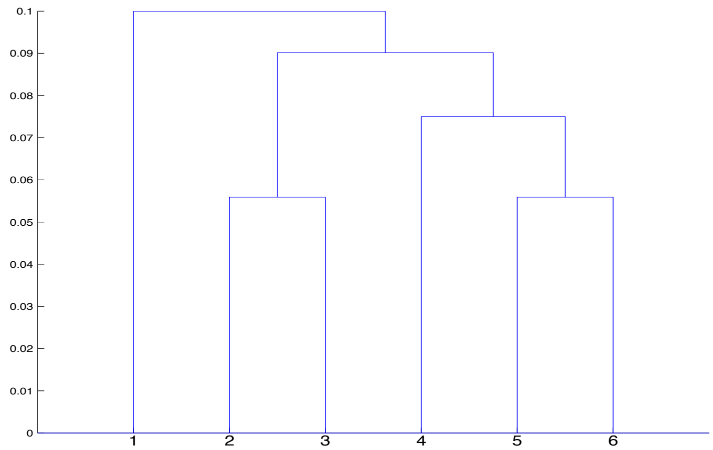
(e) (2pts) Indiquez ce que représente l'équation suivante, permettant de faire du classement avec des densités-mélanges.

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

(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.

(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, $\sigma_{i,j} < 0,\ \forall i \neq j$.

(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.



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

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

(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.

(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.

(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.

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

(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.

(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 précisément une heuristique.

$\mathcal{FIN}$

6 Novembre 2009
CG