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

UNIVERSITÉ LAVAL

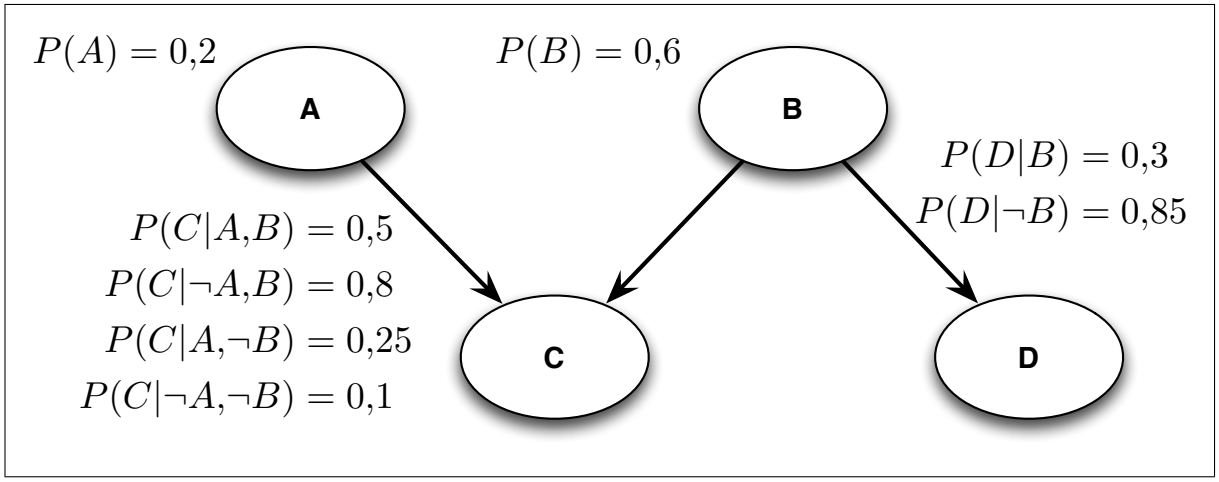
EXAMEN PARTIEL

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 $P(B|C)$.

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

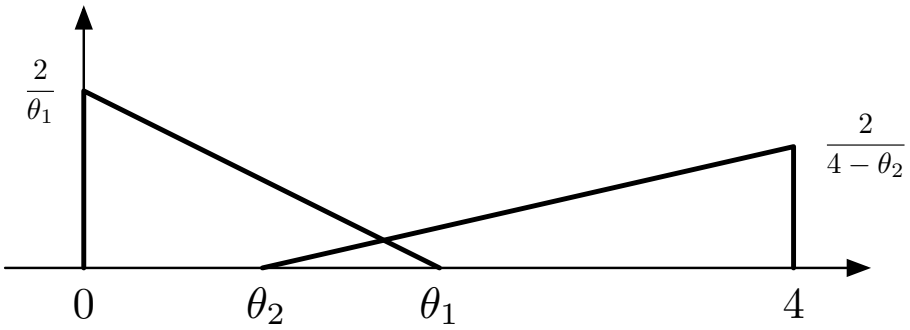
Question 2 (15 points sur 100)

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.

(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.

(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.

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 $\mathbf{w}_i$.

(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.

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.

(7) (b) Donnez le développement complet permettant de calculer les estimations $\mathbf{m}_i$ des moyennes $\boldsymbol{\mu}_i$.

(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.

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*).

(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.

(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.

(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 ?

(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 ?

(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.

(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.

(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.

(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.

(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.