Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2016

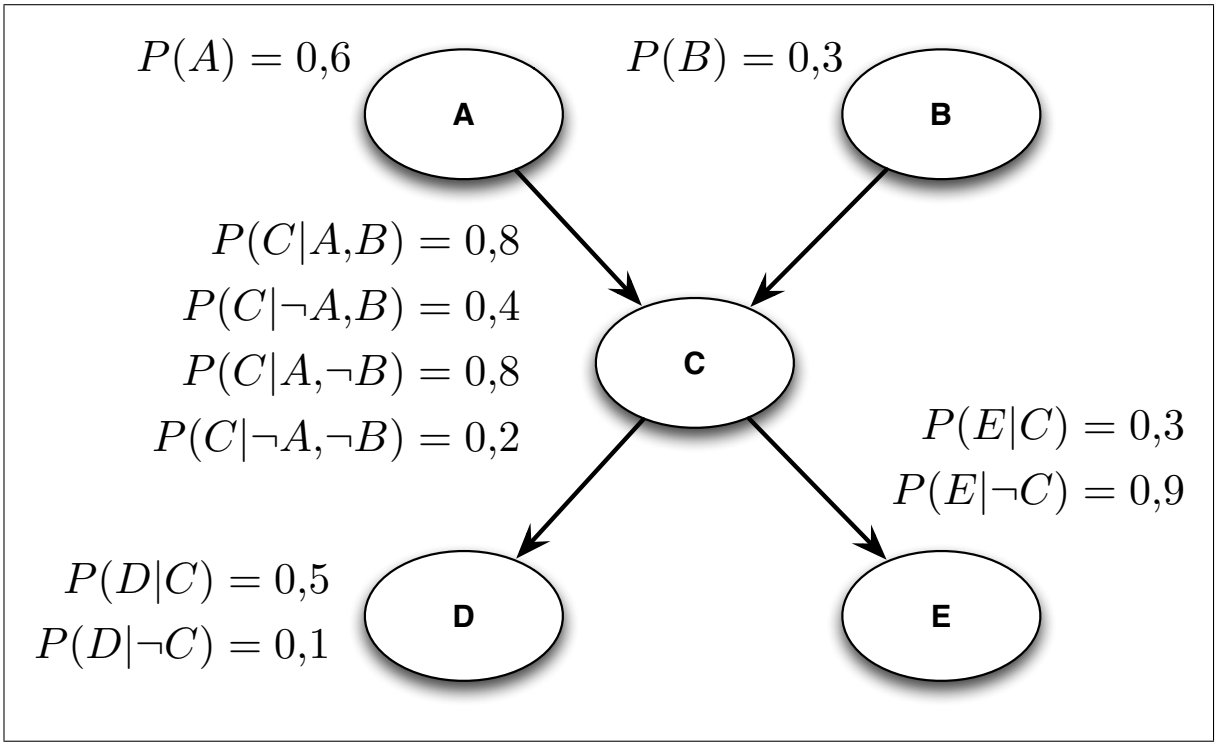
UNIVERSITÉ LAVAL

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 $P(D|E)$.

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

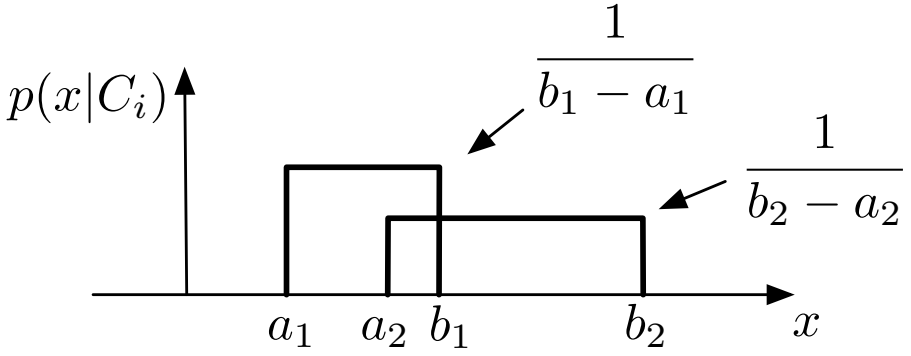
Question 2 (10 points sur 100)

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 :

$$p(x|C_1) = \begin{cases} \frac{1}{b_1-a_1} & \text{si } x \in [a_1,b_1] \\ 0 & \text{autrement} \end{cases},$$

$$p(x|C_2) = \begin{cases} \frac{1}{b_2-a_2} & \text{si } x \in [a_2,b_2] \\ 0 & \text{autrement} \end{cases},$$

où $a_1 < b_1$ et $a_2 < b_2$. En guise de simplification, on suppose que $a_1 \leq a_2$. La figure suivante présente le tracé de ces distributions de classes (vraisemblance).



(5) (a) En supposant que $a_1 = 0$, $b_1 = 1{,}25$, $a_2 = 1$ et $b_2 = 2$, donnez la fonction $\mathrm{h}(x)$ permettant la prise de décision pour le classement de données selon la valeur de x dans l'intervalle $[0, 2]$. 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.

Question 3 (30 points sur 100)

Supposons les données suivantes en deux dimensions :

$$\mathbf{x}^1 = \begin{bmatrix} 1{,}50 \\ -0{,}75 \end{bmatrix}, \quad r^1 = 0, \qquad \mathbf{x}^2 = \begin{bmatrix} 3{,}50 \\ -1{,}30 \end{bmatrix}, \quad r^2 = 0, \qquad \mathbf{x}^3 = \begin{bmatrix} 2{,}00 \\ -2{,}00 \end{bmatrix}, \quad r^3 = 0, \qquad \mathbf{x}^4 = \begin{bmatrix} 4{,}15 \\ -2{,}90 \end{bmatrix}, \quad r^4 = 0,$$

$$\mathbf{x}^5 = \begin{bmatrix} -0{,}10 \\ 1{,}50 \end{bmatrix}, \quad r^5 = 1, \qquad \mathbf{x}^6 = \begin{bmatrix} -2{,}00 \\ 0{,}40 \end{bmatrix}, \quad r^6 = 1, \qquad \mathbf{x}^7 = \begin{bmatrix} -1{,}20 \\ -0{,}75 \end{bmatrix}, \quad r^7 = 1.$$

Le vecteur moyen $\mathbf{m}$ et la matrice de covariance $\mathbf{S}$ estimés de ces données sont les suivants :

$$\mathbf{m} = \begin{bmatrix} 1{,}1214 \\ -0{,}8286 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 5{,}3949 & -2{,}5426 \\ -2{,}5426 & 2{,}1382 \end{bmatrix}.$$

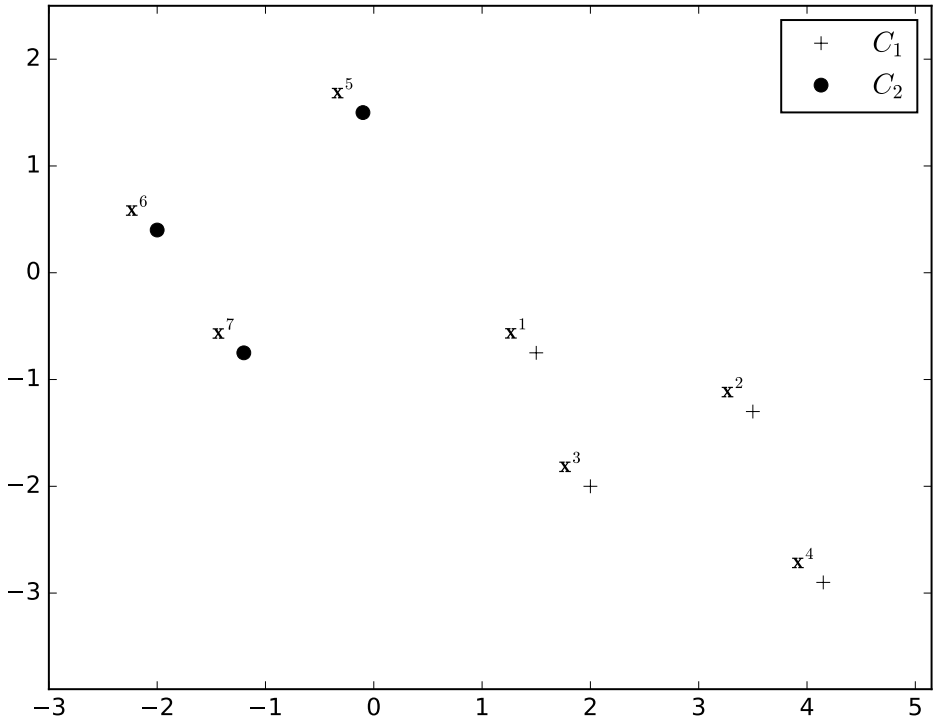
Les vecteurs propres et valeurs propres associés de cette matrice de covariance sont :

$$\lambda_1 = 6{,}7859, \quad \mathbf{c}_1 = \begin{bmatrix} 0{,}8773 \\ -0{,}4799 \end{bmatrix}, \quad \lambda_2 = 0{,}7472, \quad \mathbf{c}_2 = \begin{bmatrix} 0{,}4799 \\ 0{,}8773 \end{bmatrix}.$$

La matrice de distance euclidienne entre chaque paire de données correspond à :

$$\mathbf{D} = \begin{bmatrix} 0{,}0 & 2{,}0742 & 1{,}3463 & 3{,}4125 & 2{,}7609 & 3{,}6841 & 2{,}7 \\ 2{,}0742 & 0{,}0 & 1{,}6553 & 1{,}7270 & 4{,}5607 & 5{,}7567 & 4{,}7321 \\ 1{,}3463 & 1{,}6553 & 0{,}0 & 2{,}3308 & 4{,}0817 & 4{,}6648 & 3{,}4355 \\ 3{,}4125 & 1{,}7270 & 2{,}3308 & 0{,}0 & 6{,}1174 & 6{,}9794 & 5{,}7658 \\ 2{,}7609 & 4{,}5607 & 4{,}0817 & 6{,}1174 & 0{,}0 & 2{,}1954 & 2{,}5045 \\ 3{,}6841 & 5{,}7567 & 4{,}6648 & 6{,}9794 & 2{,}1954 & 0{,}0 & 1{,}4009 \\ 2{,}7 & 4{,}7321 & 3{,}4355 & 5{,}7658 & 2{,}5045 & 1{,}4009 & 0{,}0 \end{bmatrix}.$$

Finalement, les données sont tracées dans la figure suivante.



(6) (a) Effectuez une itération de l'algorithme K-means sur ces données, avec $K = 2$ clusters, en donnant les valeurs de b_i^t et $\mathbf{m}_i$ pour les deux groupes et toutes les données. Démarrez avec comme centres initiaux $\mathbf{m_1}(0) = \mathbf{x}^2$ et $\mathbf{m_2}(0) = \mathbf{x}^7$. Déterminez combien d'itérations sont nécessaires à l'algorithme avant qu'il y ait convergence.

(6) (b) Effectuez un clustering hiérarchique agglomératif de ces données. Pour ce faire, effectuez un clustering en lien simple, où la distance entre deux groupes est la distance minimale entre deux paires de données de groupes différents :

$$d(\mathcal{G}_i,\mathcal{G}_j) = \min_{\mathbf{x}^t\in\mathcal{G}_i,\mathbf{x}^s\in\mathcal{G}_j} d(\mathbf{x}^t,\mathbf{x}^s).$$

Utilisez une distance euclidienne. Tracez le résultat dans un dendrogramme. Déterminez également à quoi correspond le clustering si l'on décide de conserver $K = 3$ clusters, en indiquant quelles données forment chacun de ces clusters.

(6) (c) Donnez l'équation permettant d'effectuer une transformation blanchissante de ces données, sous la forme d'une équation linéaire ayant la formulation suivante :

$$\mathbf{z} = \mathbf{A}\,\mathbf{x} + \mathbf{b},$$

en précisant les valeurs numériques de $\mathbf{A}$ et $\mathbf{b}$.

(6) (d) Faites un graphique représentant la distribution des données en deux dimensions, en y traçant 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, soit le vecteur moyen estimé $\mathbf{m}$, les valeurs propres λ_i et les vecteurs propres $\mathbf{c}_i$.

(6) (e) Tracez les régions de décision selon les données d'entraînement pour un classifieur de type plus proche voisin (avec un seul voisin, $k = 1$). Tracez le tout dans votre **cahier bleu d'examen** (et non dans l'énoncé courant). Donnez également le taux de classement selon une méthodologie *leave-one-out* avec cette configuration, sur ces données.

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 $\mathcal{N}_D(\boldsymbol{\mu}_i,\sigma_i^2\mathbf{I})$, 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 toutes 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 :

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

où $\pi_i = P(\mathcal{G}_i|\Phi)$ est la probabilité *a priori* du groupe $\mathcal{G}_i$ et $h_i^t = P(\mathcal{G}_i|\mathbf{x},\Phi)$ est l'appartenance probabiliste de la donnée $\mathbf{x}$ au groupe $\mathcal{G}_i$.

(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 (30 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Si l'on effectue de la validation croisée à K-plis (K-*fold cross-validation*, en anglais), indiquez combien d'entraînements de modèles seront nécessaires au minimum pour comparer la performance de deux algorithmes distincts sur un jeu de données particulier.

(3) (b) Dans le cours, l'apprentissage automatique a été présenté comme ayant trois dimensions générales : la représentation, l'évaluation et l'optimisation. Dans le contexte de méthodes paramétriques probabilistes présentées en classe, indiquez où l'on peut intégrer de la connaissance sur le problème relativement à l'aspect de la représentation.

(3) (c) Dans un contexte de prise de décision probabiliste avec fonction de perte 0-1 et option de rejet, quelle est la relation entre la valeur de λ représentant le coût de rejet et les probabilités de classement des données relativement à la prise de décision.

(3) (d) Expliquez pourquoi dit-on que les classifieurs plus complexes ont une variance plus élevée, dans une perspective du compromis biais-variance.

(3) (e) Indiquez si deux variables indépendantes ont une variance nulle ou non.

(3) (f) Dans un contexte de classement avec méthodes paramétriques utilisant des lois normales multivariées, indiquez la condition nécessaire pour que la frontière de décision entre les classes soit linéaire.

(3) (g) 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.

(3) (h) On dit que les heuristiques de sélections de caractéristiques voraces, comme la sélection avant séquentielle, peuvent ne pas converger à la solution optimale. Expliquez dans quel contexte ceci peut arriver.

(3) (i) Lorsque l'on veut faire une sélection agressive de prototypes pour le classement avec la règle des k plus proches voisins, on peut effectuer une édition de Wilson suivie d'une condensation de Hart. Expliquez pourquoi procéder dans l'ordre inverse, soit en faisant une condensation de Hart suivie d'une édition de Wilson, est une mauvaise idée et risque de donner de mauvais résultats.

(3) (j) Expliquez en quoi consiste la régularisation lorsque l'on fait l'inférence de modèles en apprentissage supervisé.