Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2011

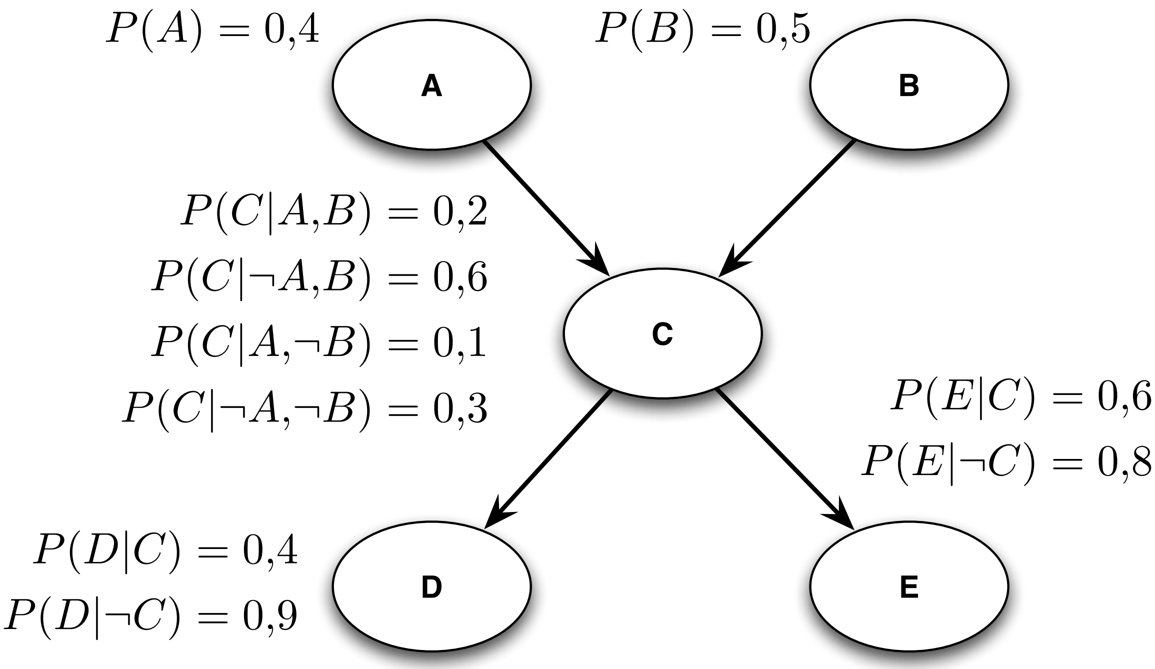
EXAMEN PARTIEL

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 (12 points sur 100)

Soit le réseau bayésien suivant.



(6) (a) Selon ce réseau, calculez la valeur de la probabilité $P(D|E)$.

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

Question 2 (20 points sur 100)

Supposons que l'on veut automatiser le traitement d'images de champs agricoles, afin de discriminer les mauvaises herbes des plantes de maïs. Nous avons déjà sous la main une chaîne de traitement des images qui permet de segmenter les pixels correspondants à une plante, qu'elle soit bonne ou mauvaise, des autres éléments. Nous avons également un algorithme qui extrait des pixels formant un ensemble continu selon leur voisinage, que l'on nomme *blobs*. De plus, selon la distribution des blobs dans l'image, nous sommes en mesure de différencier les régions de l'image correspondant à des rangs de maïs des régions correspondant aux espaces inter-rangs. Et finalement, nous avons une banque d'images de ce type où chacun des blobs des images a été étiqueté selon deux classes, soit les mauvaises herbes (classe C_M) et les plantes de maïs (classe C_B).

Dans des études préliminaires, il a été établi que la taille des blobs est une mesure permettant de bien discriminer les plantes de maïs des mauvaises herbes. De plus, nous faisons l'hypothèse que la distribution des mauvaises herbes est la même partout dans les images, alors que les plantes de maïs sont présentes uniquement dans les régions désignées comme étant des rangs. Les éléments suivants peuvent donc être estimés à partir de la banque d'images étiquetées que l'on a sous la main :

- $p(x|C_M)$: densité de probabilité des mauvaises herbes selon la taille des blobs de plantes (x), cette densité est la même dans tout l'image (que l'on soit dans les rangs ou dans les inter-rangs) ;
- $p_{\text{rang}}(x)$: densité de probabilité des plantes, sans égard que ce soit des mauvaises herbes du maïs, dans les rangs selon la taille des blobs de plantes (x) ;
- $P_{\text{rang}}(C_B)$ et $P_{\text{rang}}(C_M)$: probabilités *a priori* de blobs de plantes de maïs ($P_{\text{rang}}(C_B)$) et de mauvaises herbes ($P_{\text{rang}}(C_M)$) dans les rangs ;
- $P_{\text{inter}}(C_B) = 0$ et $P_{\text{inter}}(C_M) = 1$: probabilités *a priori* de blobs de plantes de maïs (nulle, $P_{\text{inter}}(C_B) = 0$) et de mauvaises herbes ($P_{\text{inter}}(C_M) = 1$) dans les inter-rangs.

(5) (a) Donnez l'équation de la densité de probabilité de plantes (maïs ou mauvaise herbe) dans les inter-rangs, $p_{\text{inter}}(x)$, en utilisant seulement les éléments connus précédemment mentionnés.

(5) (b) Donnez l'équation de la densité de probabilité qu'une plante soit du maïs dans les rangs, $p_{\text{rang}}(x|C_B)$, en utilisant seulement les éléments connus précédemment mentionnés.

(5) (c) Donnez l'équation simplifiée de la fonction de décision $\mathrm{h}_{\text{rang}}(x)$ donnant la classe (C_M : mauvaise herbe ; C_B : plante de maïs) d'un blob de pixels de l'image selon sa taille x, dans les rangs. À cette fin, donnez le détail du développement des probabilités $P_{\text{rang}}(C_B|x)$ et $P_{\text{rang}}(C_M|x)$.

(5) (d) Donnez l'équation simplifiée de la fonction de décision $\mathrm{h}_{\text{inter}}(x)$ donnant la classe (C_M : mauvaise herbe ; C_B : plante de maïs) d'un blob de pixels de l'image selon sa taille x, entre les rangs. À cette fin, donnez le détail du développement des probabilités $P_{\text{inter}}(C_B|x)$ et $P_{\text{inter}}(C_M|x)$.

Question 3 (15 points sur 100)

Supposons que l'on veut appliquer l'algorithme Espérance-Maximisation (EM) à un jeu de données en une dimension, où chaque groupe $\mathcal{G}_i$ est décrit par une loi normale $\mathcal{N}(\mu_i,\sigma_i^2)$, soit :

$$p(x|\mathcal{G}_i,\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right].$$

Donc, la paramétrisation du clustering par EM est donnée par $\Phi = \{\pi_i,\mu_i,\sigma_i\}_{i=1}^K$. En guise de rappel, la formule de l'espérance de vraisemblance de l'algorithme EM est la suivante :

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

(5) (a) Donnez le développement complet permettant de calculer les estimations π_i des probabilités *a priori* des groupes.

(5) (b) Donnez le développement complet permettant de calculer les estimations m_i des moyennes μ_i.

(5) (c) Donnez le développement complet permettant de calculer les estimations s_i^2 des variances σ_i^2.

Question 4 (20 points sur 100)

Supposons que l'on a un jeu de données en trois dimensions comportant deux classes, C_1 et C_2. Les vecteurs moyens $\boldsymbol{\mu}_i$, ainsi que les valeurs propres $\lambda^{\boldsymbol{\Sigma}_i}$ et vecteurs propres $\mathbf{w}^{\boldsymbol{\Sigma}_i}$ associés aux matrices de covariance $\boldsymbol{\Sigma}_i$ de chaque classe sont les suivants :

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 0{,}5 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -1 \\ 3 \end{bmatrix},$$

$$\lambda_1^{\boldsymbol{\Sigma}_1} = 1{,}0201, \qquad \lambda_2^{\boldsymbol{\Sigma}_1} = 0{,}1150, \qquad \lambda_3^{\boldsymbol{\Sigma}_1} = 2{,}9288,$$

$$\mathbf{w}_1^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} 0{,}3189 \\ -0{,}8811 \\ 0{,}3493 \end{bmatrix}, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} -0{,}8167 \\ -0{,}0684 \\ 0{,}5729 \end{bmatrix}, \quad \mathbf{w}_3^{\boldsymbol{\Sigma}_1} = \begin{bmatrix} -0{,}4809 \\ -0{,}4680 \\ -0{,}7414 \end{bmatrix},$$

$$\lambda_1^{\boldsymbol{\Sigma}_2} = 3{,}6260, \qquad \lambda_2^{\boldsymbol{\Sigma}_2} = 0{,}0162, \qquad \lambda_3^{\boldsymbol{\Sigma}_2} = 0{,}3673,$$

$$\mathbf{w}_1^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}6651 \\ 0{,}6599 \\ 0{,}3494 \end{bmatrix}, \quad \mathbf{w}_2^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} 0{,}7165 \\ -0{,}6958 \\ -0{,}0497 \end{bmatrix}, \quad \mathbf{w}_3^{\boldsymbol{\Sigma}_2} = \begin{bmatrix} -0{,}2103 \\ -0{,}2835 \\ 0{,}9356 \end{bmatrix}.$$

Les probabilités *a priori* des classes sont respectivement $P(C_1) = 0{,}75$ et $P(C_2) = 0{,}25$.

(5) (a) Pour chaque classe, donnez la transformation linéaire nécessaire pour conserver au moins $75\ \%$ de la variance. Traitez chaque classe indépendamment.

(5) (b) Pour chacune des deux classes, calculez la matrice de covariance associée ($\mathbf{\Sigma}_1$ et $\mathbf{\Sigma}_2$).

(5) (c) Calculez la matrice de covariance partagée (quelconque, avec valeurs hors la diagonale) par les deux classes ($\mathbf{\Sigma}$).

(5) (d) Supposons que l'on fasse la projection des données sur un vecteur $\mathbf{v}$ à l'aide de la transformation linéaire suivante.

$$z = \mathbf{v}^T\mathbf{x}, \quad \mathbf{v} = \begin{bmatrix} 0{,}4082 \\ 0{,}8165 \\ 0{,}4082 \end{bmatrix}$$

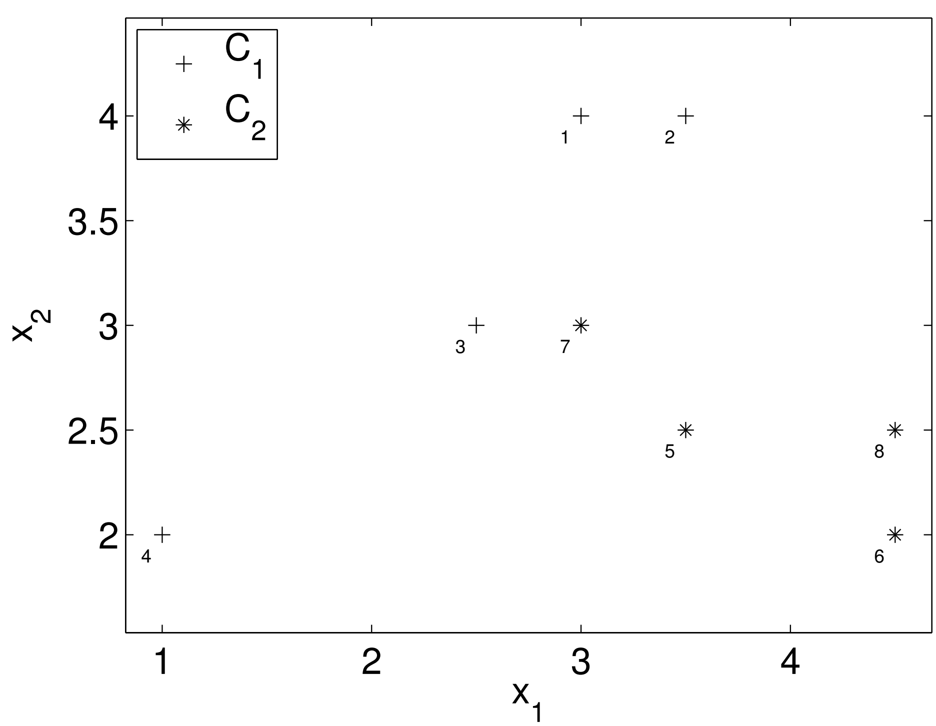
Calculez le pourcentage de la variance des données qui est conservé pour chaque classe suite à cette transformation linéaire.

Question 5 (33 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Expliquez pourquoi l'utilisation d'un jeu de données de validation distinct du jeu d'entraînement peut servir à estimer la capacité de généralisation d'un classifieur.

(3) (b) Donnez le taux de classement par la méthode du plus proche voisin ($k = 1$) avec distance euclidienne, en utilisant l'approche *leave-one-out* expliquée en classe, sur les données présentées dans la figure ici-bas.



(3) (c) Expliquez pourquoi l'estimateur selon un maximum de vraisemblance de la variance σ^2 d'une loi normale unidimensionnelle $\mathcal{N}(\mu,\sigma^2)$ est biaisé.

(3) (d) Donnez la complexité algorithmique (en notation grand $O(\cdot)$, soit la borne supérieure sur le temps d'exécution) de la condensation de Hart, lorsque appliquée à un jeu de N données, en justifiant votre réponse. Vous pouvez supposer que les distances entre toutes les paires de données ont été calculées et stockées dans une structure de donnée avant l'exécution de la condensation de Hart.

(3) (e) Expliquez pourquoi dit-on que l'analyse en composante principale (ACP) est une approche non supervisée, alors que l'analyse discriminante linéaire (ADL) est une approche supervisée.

(3) (f) 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) (g) On dit que les heuristiques de sélections de caractéristiques voraces, comme la sélection avant séquentielle, peuvent ne pas converger à la solution optimale. Présentez une situation générale illustrant bien cette possibilité.

(3) (h) Expliquez à quoi correspondent les arêtes formant une tesselation de Voronoï dans un contexte de classement avec la règle du plus proche voisin.

(3) (i) On dit que l'algorithme K-means minimise l'erreur de reconstruction des données. Expliquez à quoi correspond cette erreur de reconstruction avec cet algorithme.

(3) (j) Dans un contexte de régression polynomiale multivariée, on se limite très souvent à l'utilisation de polynômes d'ordre 1, ce qui correspond à une régression linéaire multivariée. Expliquez pourquoi il est généralement peu intéressant d'effectuer de la régression polynomiale dans un contexte multivarié, avec des polynômes d'ordre supérieur à 1.

(3) (k) Expliquez ce que l'on peut conclure d'un classifieur dont on évalue que le biais est élevé, sous la perspective du compromis biais-variance.