Apprentissage et reconnaissance (GIF-4101 / GIF-7005)
Département de génie électrique et de génie informatique
Automne 2017

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)

Supposons que l'on fait du classement paramétrique selon deux classes et une variable en entrée (x scalaire), en modélisant les données de chaque classe par une loi normale :

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

La variance de chaque classe est la même, $\sigma_1 = \sigma_2 = \sigma$, alors que les moyennes μ_i et les probabilités a priori $P(C_i)$ sont différentes pour chaque classe. Sans perte de généralité, vous pouvez supposer que la moyenne de la classe 1 est inférieure à la classe 2, $\mu_1 < \mu_2$.

Avec cette modélisation, donnez l'équation analytique décrivant les frontières de décision entre les deux classes, en supposant une fonction de perte zéro-un (valeur égale pour les deux types d'erreurs). En une dimension, de telles frontières se résument à des seuils sur la valeur de x. Indiquez également de quelle façon le classement se fait dans les différentes régions séparées par les frontières.

Question 2 (20 points sur 100)

Soit une loi de Bernoulli de paramètre p, soit la probabilité de succès (probabilité d'obtenir une valeur 1). Une loi géométrique de paramètre p modélise le nombre de répétitions de tirages selon une loi de Bernoulli de paramètre p nécessaire pour obtenir un premier succès. La loi s'exprime selon la probabilité $P(x|p)$ d'obtenir un premier succès après x tirages :

$$P(x|p) = (1-p)^{x-1}\, p.$$

La variable x est un entier positif ($x = 1{,}2, \ldots$). L'espérance d'une variable X suivant une loi géométrique de paramètre p est $\mathbb{E}(X) = \frac{1}{p}$ alors que sa variance est de $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

(10) (a) Calculez la fonction pour estimer par un maximum de vraisemblance l'espérance d'une loi géométrique sur un jeu $\mathcal{X} = \{x^t\}_{t=1}^N$ de N échantillons. Donnez les développements analytiques complets pour en arriver à votre réponse.

Remarque : On traite le développement d'estimateur selon le maximum de vraisemblance avec une loi de probabilité de la même façon qu'on le fait avec une loi de densité de probabilité.

(5) (b) Nous avons accès à K machines à sous distinctes dans un casino. Chaque machine est programmée pour fournir un montant m_i, connu des joueurs et différent pour chaque machine. Une partie avec une machine correspond à une expérience aléatoire où l'on fait le tirage d'une valeur binaire x_i suivant une loi de Bernoulli de paramètre p_i, qui est différent pour chaque machine et inconnu des joueurs. Une valeur de $x_i = 1$ signifie que le montant m_i est donné au joueur alors qu'autrement une valeur $x_i = 0$ est obtenue.

Supposons qu'une estimation $\hat{p}_i$, $i = 1, \ldots, K$, des paramètres de chaque machine est disponible. Donnez l'équation permettant de choisir la machine qui maximise les gains espérés du joueur en justifiant votre réponse.

(5) (c) Supposons maintenant qu'un joueur n'a pas accès aux estimations $\hat{p}_i$ des paramètres des machines à sous. La stratégie est de supposer des paramètres initiaux des machines $\hat{p}_i$ arbitraires et identiques pour toutes les machines. Tant qu'aucun gain n'est fait, les machines sont jouées à tour de rôle. Lors du premier gain par une machine, l'estimation $\hat{p}_i$ de la machine gagnante est mise à jour à partir du gain observé et remplace le paramètre initial. Pour les gains subséquents d'une même machine, l'estimation $\hat{p}_i$ se fait à partir des gains historiques observés. De plus, la prise de décision se base sur la maximisation du gain espéré, une fois le premier gain est observé.

Cette approche pour la prise de décision comporte des problèmes importants. Expliquez dans votre cahier les principaux éléments problématiques. Proposez également des correctifs permettant de régler le problème.

Question 3 (34 points sur 100)

Supposons les données suivantes en deux dimensions :

$$\mathbf{x}^1 = \begin{bmatrix} 2{,}50 \\ 1{,}00 \end{bmatrix}, \quad r^1 = 0, \quad \mathbf{x}^2 = \begin{bmatrix} 3{,}50 \\ 1{,}30 \end{bmatrix}, \quad r^2 = 0, \quad \mathbf{x}^3 = \begin{bmatrix} 2{,}00 \\ 2{,}00 \end{bmatrix}, \quad r^3 = 0,$$

$$\mathbf{x}^4 = \begin{bmatrix} 4{,}15 \\ 2{,}90 \end{bmatrix}, \quad r^4 = 1, \quad \mathbf{x}^5 = \begin{bmatrix} -0{,}10 \\ -1{,}50 \end{bmatrix}, \quad r^5 = 1, \quad \mathbf{x}^6 = \begin{bmatrix} -2{,}00 \\ -0{,}40 \end{bmatrix}, \quad r^6 = 1.$$

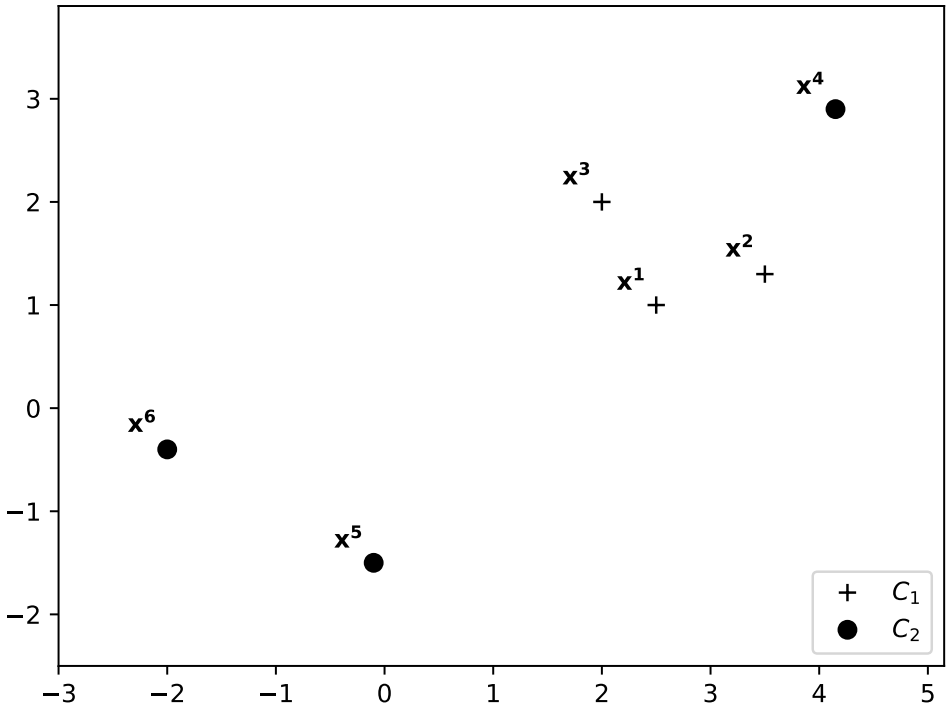
Le vecteur moyen $\mathbf{m}$ et la matrice de covariance $\mathbf{S}$ estimés de ces données sont les suivants :

$$\mathbf{m} = \begin{bmatrix} 1{,}67500 \\ 0.88333 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 5{,}37975 & 3{,}03150 \\ 3{,}03150 & 2{,}56567 \end{bmatrix}.$$

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

$$\lambda_1 = 7{,}3148, \quad \mathbf{c}_1 = \begin{bmatrix} 0{,}84291 \\ 0{,}53805 \end{bmatrix}, \quad \lambda_2 = 0{,}63059, \quad \mathbf{c}_2 = \begin{bmatrix} -0{,}53805 \\ 0{,}84291 \end{bmatrix}.$$

Finalement, les données sont tracées dans la figure suivante.



(10) (a) La procédure suivante a été brièvement présentée en classe afin d'initialiser l'algorithme K-means à partir d'une analyse en composantes principales (ACP) :

1. effectuer une ACP sur l'ensemble des données pour en extraire la composante principale (première composante) ;
2. projeter les données en une dimension selon la composante principale ;
3. partitionner les données en K groupes de taille égale dans l'espace unidimensionnel ;
4. en utilisant le partitionnement de l'étape précédente, calculer les centres des groupes dans l'espace d'origine.

Effectuez cette procédure pour déterminer les paramètres initiaux de l'algorithme K-means, avec $K = 2$ groupes.

(8) (b) Exécutez l'algorithme K-means (avec $K = 2$ groupes) à partir l'initialisation obtenue à la sous-question précédente, jusqu'à convergence de l'algorithme.

(8) (c) 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.

(8) (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 (équivalent 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 directions des vecteurs propres $\mathbf{c}_i$.

Question 4 (36 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

(3) (a) Dans la formalisation mathématique présentée en classe, un modèle d'apprentissage supervisé est représenté par une fonction $\mathrm{h}(\mathbf{x}|\theta)$. Dans ce modèle, expliquez ce que représente les variables $\mathbf{x}$ et θ.

(3) (b) Supposons que l'on fait l'évaluation de modèles de classement selon une approche de type validation croisée à K-plis (en anglais : K-*fold cross-validation*) avec un jeu de données $\mathcal{X}$. Donnez le nombre d'instances sur lequel le taux de classement a été calculé.

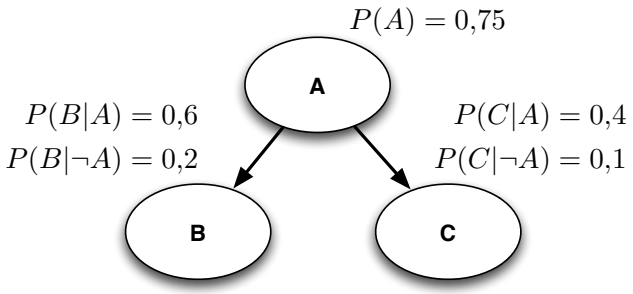
(3) (c) Dans une approche de classement probabiliste, expliquez à quoi correspond $P(C_i|\mathbf{x})$

(3) (d) D'après vous, où se situe un classifieur de type plus proche voisin (avec $k = 1$ voisin) selon le compromis biais-variance. Justifiez votre réponse.

(3) (e) On dit que deux variables indépendantes ont une corrélation nulle, mais que l'inverse n'est pas nécessairement vrai. Expliquez cette affirmation.

(3) (f) Dans un contexte de classement paramétrique avec lois normales multivariées, donnez la condition sur les modèles faisant en sorte que les frontières de décision entre les classes soient linéaires.

(3) (g) Soit le réseau bayésien suivant, calculez $P(C|\neg B)$.



(3) (h) Expliquez précisément pourquoi les algorithmes de sélection de variables avant séquentielle et arrière séquentielle, tels que présentés dans le cours, sont des heuristiques sans garantie d'optimalité.

(3) (i) Expliquez l'effet d'une transformation blanchissante sur un jeu de données quelconque.

(3) (j) Selon les explications faites en classe sur l'algorithme Espérance-Maximisation (EM), indiquez à quoi correspondent précisément les variables $\mathbf{z}^t$ ainsi que leur contenu.

(3) (k) Expliquez pourquoi dit-on qu'il est préférable de bien normaliser les données avant l'application d'un classement de type k-plus proches voisins basé sur une distance euclidienne.

(3) (l) Expliquez pourquoi il est déconseillé de faire une condensation de Hart suivie d'une édition de Wilson, alors que le contraire est possible et même suggéré dans certains cas.