Final SY09 Printemps 2009

Partie 1 - Sans document

Nom :	**Signature :**
Prénom :	

Répondre sur ce document uniquement. La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée.

Exercice 1 (1 point)

Analyse en Composantes Principales

1. Indiquer parmi les affirmations suivantes celles qui sont exactes
   - ☐ l'ACP est une méthode de discrimination
   - ☐ l'ACP est une méthode de visualisation
   - ☐ l'ACP est une méthode de classification
   - ☐ l'ACP est une méthode de réduction de variables
2. Une ACP faite en diagonalisant la matrice de variance ou la matrice de corrélation donne :
   - ☐ des axes factoriels différents, mais des inerties expliquées égales
   - ☐ des axes factoriels identiques, mais des inerties expliquées différentes
   - ☐ des axes factoriels et des inerties expliquées différents
3. La relation $X = CU^\top$
   - ☐ fait intervenir les matrices des valeurs propres, des vecteurs propres, et des individus
   - ☐ permet de recalculer les coordonnées des individus dans la base de départ
   - ☐ nécessite de connaître les composantes principales et les axes factoriels
   - ☐ calcule les contributions relatives des axes aux individus
4. Soit λ_k l'inertie expliquée par l'axe factoriel δ_k (pour $k = 1, \ldots, p$) ; si l'on décide de n'utiliser que les axes δ_i, δ_j et δ_k pour représenter les individus :
   - ☐ l'inertie expliquée par le sous-espace associé est égale à $\lambda_i + \lambda_j + \lambda_k$
   - ☐ les individus seront parfaitement représentés dans le sous-espace défini par δ_i, δ_j et δ_k
   - ☐ les individus seront parfaitement représentés à condition de bien choisir δ_i, δ_j et δ_k

Exercice 2 (2 points)

On dispose de deux jeux de données X_1 et X_2 dans $\mathbb{R}^2$, chacun comptant $K = 3$ classes. Chaque classe de X_1 compte 1000 individus générés suivant une loi normale bidimensionnelle. Dans X_2, les classes comptent trois composantes gaussiennes de 300 individus chacune. Les classes de X_1 ont même matrice de variance-covariance, de même que les composantes des classes de X_2. La figure **??** représente les individus des deux ensembles.

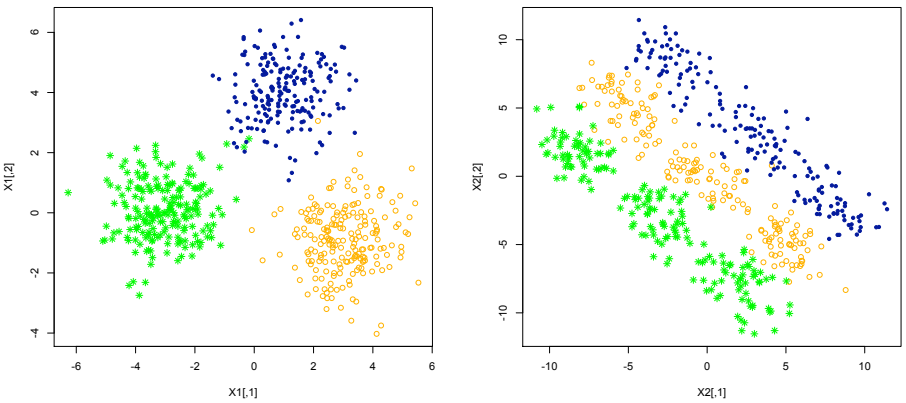

FIG. 1 – Individus des jeux de données X_1 (gauche) et X_2 (droite).

1. Peut-on s'attendre à ce que l'algorithme des K-means trouve une partition satisfaisante des données X_1, des données X_2 ?

Deux exécutions successives de l'algorithme des centres-mobiles sur le jeu de données X_2 ont donné les deux partitions représentées sur la figure **??**. La partition P_1 (figure de gauche) correspond à une somme des inerties intra-classes $I_{w1} = 9313$; la partition P_2 (droite), à une somme $I_{w2} = 8685$.

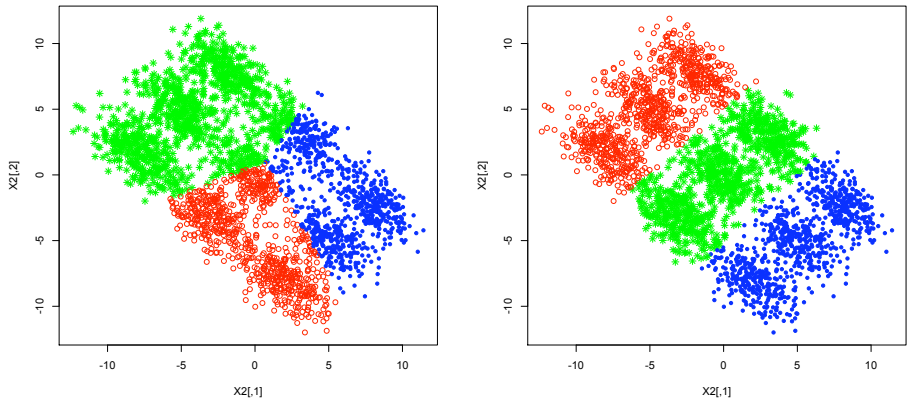

FIG. 2 – Partitions P_1 (gauche) et P_2 (droite) obtenues par exécution de l'algorithme des centres-mobiles sur le jeu de données X_2.

2. Comment expliquez-vous l'obtention de deux résultats différents sur ce même jeu de données ?

3. Expliquer en quoi on peut considérer que la partition P_2 est meilleure que P_1. Proposer une méthode pour obtenir automatiquement P_2 sans avoir visualiser la partition fournie par l'algorithme.

4. Comment serait-il possible de calculer une partition des données X_2 correspondant aux vraies classes ?

Final SY09 Printemps 2009

Partie 2 - Sans document

Nom :	**Signature :**
Prénom :	

Répondre sur ce document uniquement. La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée.

Exercice 1 (3 points)

Dans cette partie, chaque question peut avoir 0, 1 ou plusieurs réponses correctes. Répondre en ***entourant*** *la lettre V (Vrai) ou F (Faux). Une bonne réponse est comptée* $+1$*, une erreur* -1 *et une absence de réponse* 0*. Répondre de manière lisible. Toute réponse ambiguë sera comptée comme fausse.*

1. La règle de Neyman-Pearson
   - V F suppose la connaissance des probabilités a priori ;
   - V F suppose la connaissance des coûts de décision ;
   - V F s'exprime en fonction du rapport de vraisemblance $f_1(x)/f_2(x)$;
   - V F a une probabilité d'erreur qui tend vers 0 quand la taille de l'ensemble d'apprentissage tend vers l'infini.
2. La règle de Bayes
   - V F suppose la connaissance des probabilités a priori ;
   - V F suppose la connaissance des coûts de décision ;
   - V F s'exprime dans le cas de 2 classes en fonction du rapport de vraisemblance $f_1(x)/f_2(x)$;
   - V F a une probabilité d'erreur qui tend vers 0 quand la taille de l'ensemble d'apprentissage tend vers l'infini.
3. En général, plus le nombre de paramètres d'un classifieur est important
   - V F moins il commettra d'erreurs sur l'ensemble d'apprentissage ;
   - V F moins il commettra d'erreurs sur de nouvelles données.
4. Dans le cas gaussien, le classifieur de Bayes :
   - V F ne dépend que des moyennes dans chaque classe ;
   - V F est un classifieur linéaire ;
   - V F suppose l'égalité des matrices de variance conditionnellement à chaque classe.
5. Lorsque les composantes d'un vecteur aléatoire sont indépendantes et non constantes, sa matrice de variance est
   - V F scalaire ;
   - V F diagonale ;
   - V F égale à la matrice identité ;
   - V F inversible.
6. Dans la méthode d'analyse discriminante quadratique, le nombre de paramètres à estimer
   - V F augmente linéairement avec le nombre de classes ;
   - V F augmente linéairement avec le nombre de variables ;
   - V F est plus important que dans la méthode d'analyse discriminante linéaire, ce qui garantit une probabilité d'erreur plus faible.
7. Le classifieur de Bayes naïf
   - V F suppose l'indépendance des variables d'entrée ;
   - V F suppose l'indépendance des variables d'entrée conditionnellement à la classe ;
   - V F est linéaire dans le cas gaussien ;
   - V F a un nombre de paramères dépendant linéairement du nombre de variables d'entrée dans le cas gaussien.

8. Supposons que l'on dispose initialement de p variables explicatives et que l'on applique une procédure incrémentale ascendante de sélection de variables avec un certain critère de séparabilité des classes. Le nombre d'évaluations du critère augmente avec p de façon
   V F linéaire ;
   V F quadratique ;
   V F exponentielle.
9. Le bootstrap
   V F permet d'estimer la loi d'une statistique ;
   V F permet d'estimer la probabilité d'erreur d'un classifieur ;
   V F est basé sur des tirages aléatoires sans remise ;
   V F est basé sur une partition de l'ensemble d'apprentissage.
10. La règle des k plus proches voisins
   V F est une méthode de discrimination paramétrique ;
   V F tend vers la règle de Bayes quand $n \to \infty$, pour toute valeur de k ;
   V F tend vers la règle de Bayes quand $k \to \infty$, pour toute valeur de n ;
   V F a une probabilité d'erreur asymptotique inférieure à deux fois la probabilité d'erreur de Bayes.

Final SY09 Printemps 2009

Partie 3 - Avec document

Nom :	**Signature :**
Prénom :	

Répondre sur ce document uniquement. La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée.

Exercice 1 (4 points)

On dispose de données portant sur 4 individus et 3 variables dont le tableau centré et la matrice de variance associés sont respectivement :

$$\begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ -1 & 1 & -1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 0.50 & -0.25 & 0.50 \\ -0.25 & 0.50 & 0 \\ 0.50 & 0 & 1 \end{pmatrix}.$$

Ses valeurs propres sont 0.59, 1.33 et 0.08 et les vecteurs propres rangés dans le même ordre sont $\begin{pmatrix} 0.31 \\ -0.87 \\ -0.38 \end{pmatrix}$, $\begin{pmatrix} 0.54 \\ -0.16 \\ 0.82 \end{pmatrix}$ et $\begin{pmatrix} 0.78 \\ 0.46 \\ -0.42 \end{pmatrix}$.

1. Quels sont les axes factoriels $\mathbf{u}_1$, $\mathbf{u}_2$ et $\mathbf{u}_3$ et les inerties expliquées λ_1, λ_2 et λ_3 associées.

2. Calculer les deux premières composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$ et donner leur variance.

3. Tracer la représentation des individus dans le premier plan factoriel et donner le pourcentage d'inertie expliqué par ce plan.

4. Tracer la représentation des variables dans le premier plan factoriel.

Exercice 2 (3 points)

On dispose du tableau de distances suivant :

$$\begin{pmatrix} 0 & 4.2 & 1 & 5 \\ 4.2 & 0 & 3.6 & 1.1 \\ 1 & 3.6 & 0 & 4.4 \\ 5 & 1.1 & 4.4 & 0 \end{pmatrix}.$$

1. Tracer la hiérarchie indicée obtenue par la Classification hiérachique ascendante du lien minimum à partir de ce tableau de distances.

2. Proposer à partir de cette représentation une partition « raisonnable ».

3. Déterminer l'ultramétrique associée à la hiérarchie indicée obtenue dans la question 1.

Final SY09 Printemps 2009

Partie 4 - Avec document

Nom :	**Signature :**
Prénom :	

Répondre sur ce document uniquement. La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée.

Exercice 1 (7 points)

Faire d'abord les calculs au brouillon et ne reporter que les grandes lignes du raisonnement et les principaux résultats intermédiaires.
On considère un problème de discrimination à deux classes $\Omega = \{\omega_1, \omega_2\}$ et une variable $X \in \mathbb{R}$. On suppose que la variable X suit dans chaque classe une loi normale avec les espérances μ_1 et μ_2 ($\mu_2 > \mu_1$) et une variance égale à 1. Les probabilités a priori sont notées π_1 et π_2.

1. Montrer que la règle de Bayes avec coûts 0-1 revient à comparer x à un seuil s que l'on précisera.

2. Donner l'expression littérale de la probabilité d'erreur de Bayes ϵ^* (on notera ϕ la fonction de répartition de la loi normale centrée-réduite).

3. On dispose des données d'apprentissage suivantes :
– classe ω_1 : 0.1, 0.2, 0.6, 0.7 ;
– classe ω_2 : 0.5, 0.57, 1.5

3a. Estimer numériquement la probabilité d'erreur de Bayes ϵ^*, en utilisant la formule de la question 2 (on a $\phi(0.4) \approx 0.7$ et $\phi(0.9) \approx 0.8$).

3b. Estimer la probabilité d'erreur de la règle du plus proche voisin par la méthode « leave-one-out ».

3c. Soit $\widehat{\delta}$ la règle de décision obtenue en remplaçant dans l'expression de la règle de Bayes μ_1 et μ_2 par leurs estimations $\widehat{\mu}_1$ et $\widehat{\mu}_2$ et en supposant $\pi_1 = \pi_2$. Estimer la probabilité d'erreur de $\widehat{\delta}$ par la méthode « leave-one-out ». (On présentera les calculs intermédiaires dans un tableau).