Université Paris Descartes
**UFR de Mathématiques et Informatique**

# Classification Master 1 Ingénierie Mathématique
## Examen partiel du 5 avril 2018

On écrira toutes les commandes demandées dans l'éditeur en prenant soin d'enregistrer le fichier à la fin de la séance (peu importe le nom et l'emplacement du fichier). On pourra aussi enregistrer les résultats graphiques dans des fichiers pdf (Menu "Export" puis "Save as pdf").

### Exercice 1

Soit le tableau de données suivant (8 individus, une variable quantitative $X$ et une variable $Y \in \{0, 1\}$

| | $e_1$ | $e_2$ | $e_3$ | $e_4$ | $e_5$ | $e_6$ | $e_7$ | $e_8$ |
|---|---|---|---|---|---|---|---|---|
| $X$ | 9 | 16 | 6 | 7 | 18 | 15 | 19 | 4 |
| $Y$ | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |

1. Représenter graphiquement les observations sur un axe suivant la valeur de la variable $X$.
2. Donner le taux d'erreur d'apprentissage et le taux d'erreur obtenu par validation croisée "leave-one-out" pour la méthode du plus proche voisin, puis pour la méthode des trois plus proches voisins.

### Exercice 2

Le jeu données `avimedi` de la librairie `ade4` rassemble des informations sur certains sites géographiques en France et la répartition de différentes espèces d'oiseaux sur ces sites.

1. Charger la librairie et le jeu de données. Consulter l'aide sur ces données et expliquer ce que contiennent les trois tableaux de données disponibles.
2. Rappeler en quoi consiste la dissimilarité de Dice pour des données binaires. A l'aide de la fonction `dist.binary`, calculer les dissimilarités de Dice entre les "individus" (c'est-à-dire ici les sites géographiques) pour le tableau de données `fau`.
3. Effectuer des classifications ascendantes hiérarchiques (CAH) sur la matrice des dissimilarités, en comparant plusieurs méthodes d'agrégation. Afficher les différents dendrogrammes obtenus sur une même figure.

On s'intéresse à comparer les classes que l'on peut obtenir par une des méthodes CAH avec d'une part la répartition des sites en deux régions (Provence ou Corse), et d'autre part leur répartition en six niveaux de végétation.

4. En observant les dendrogrammes obtenus précédemment, choisir une méthode d'agrégation qui donne une classification en deux classes pertinente (justifier votre choix). Comparer ensuite la classification en deux classes obtenue avec la répartition des sites en deux régions. Commenter.
5. Reprendre la question précédente en choisissant cette fois une classification en six classes, et en comparant avec la répartition en six niveaux de végétation. Commenter.

On cherche à comparer les méthodes CAH pour ces données avec une autre méthode de partitionnement.

6. Expliquer pourquoi la méthode des $k$-moyennes n'est pas adaptée à ces données.

La méthode des $k$-**médoïdes** est une variante des $k$-moyennes qui consiste en l'algorithme suivant :

— Initialisation : choix de $k$ individus parmi $e_1, \cdots, e_n$ pour former les $k$ centres initiaux des classes $C_1, \ldots, C_k$.

— On itère les deux étapes suivantes :

  — Pour chaque individu $e_j$, on l'intègre à la classe $C_i$ correspondant au centre qui le plus proche de $e_j$ pour la dissimilarité considérée.

  — Pour chaque classe $C_i$ on calcule son médoïde : il s'agit de l'individu le plus central dans la classe, c'est-à-dire tel que la somme des dissimilarités aux autres individus de la classe est minimale. Les médoïdes de chaque classe deviennent les nouveaux centres.

L'algorithme est itéré jusqu'à ce que les classes n'évoluent plus.

7. Expliquer quelle différence il y a entre la méthode des $k$-moyennes et la méthode des $k$-médoïdes.

8. La fonction `pam` de la librairie `cluster` permet d'effectuer un partitionnement par une méthode proche de la méthode des $k$-médoïdes décrite ci-dessus (il s'agit en fait d'un algorithme un peu différent afin d'être plus rapide). Regarder l'aide de cette fonction puis utilisez-la pour partitionner les données en 2 et 6 classes, et comparer les résultats avec ceux obtenus avec les méthodes de classification hiérarchique.

9. Un autre avantage de la méthode des $k$-médoïdes souvent mentionné est le fait qu'elle est moins sensible aux données aberrantes (individus dont les observations sont erronées et de ce fait très éloignées des autres observations). Essayer d'imaginer et de représenter graphiquement une situation pour des données quantitatives avec deux variables où la présence de données aberrantes pourrait aboutir à un mauvais partitionnement avec la méthode des $k$-moyennes, alors qu'elle resterait cohérente avec la méthode des $m$-médoïdes.