Université Paris Descartes
UFR de Mathématiques et Informatique

Classification Master 1 Ingénierie Mathématique
Examen partiel du 5 avril 2018 - Correction

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.

Correction.

- Méthode du plus proche voisin :
  - taux d'erreur d'apprentissage : 0 car chaque individu a pour plus proche voisin lui-même, et se retrouve donc bien classé.
  - taux d'erreur par leave-one-out : On regarde le plus proche voisin de chaque individu, à l'exclusion de lui-même. On obtient une erreur de classification lorsque la classe du plus proche voisin est différente de la classe de l'individu. Ainsi on aura 5 erreurs (pour e_3, e_4, e_1, e_5 et e_7) et donc un taux d'erreur de $5/8 = 62.5\%$.
- Méthode des trois plus proches voisins :
  - taux d'erreur d'apprentissage : on peut voir rapidement que tous les individus du groupe de gauche sur le graphique (e_8, e_3, e_4, e_1) seront classés en $Y = 0$ car pour chacun d'eux les trois plus proches voisins se trouvent dans ce groupe, et il n'y a qu'un seul individu $Y = 1$ parmi eux. De même pour le groupe de droite, tous les individus seront classés en $Y = 1$. Par conséquent il y a deux erreurs et le taux d'erreur d'apprentissage est donc de $2/8 = 25\%$.

— Par leave-one-out le raisonnement précédent reste vrai en fait, car même en considérant les trois plus proches voisins hors lui-même, chacun des individus du groupe de gauche sera classé en $Y = 0$ et chacun des individus du groupe de droite sera classé en $Y = 1$ (car e_8 est plus proche de e_1 que e_6, et e_7 est plus proche de e_6 que e_1). Le taux d'erreur est donc là aussi de 25%.

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. Expliquer ce que contiennent les trois tableaux de données disponibles.

   Correction. Le tableau `fau` indique la présence ou non présence de chacune des 51 espèces d'oiseaux sur chacun des 302 sites géographiques. Le tableau `plan` donne deux informations supplémentaires sur chacun des 302 sites : situation géographique globale (Provence ou Corse) et niveau de végétation (suivant une notation de 1 à 6). Enfin `nomesp` donne les noms scientifiques des 51 espèces d'oiseaux correspondant au tableau `fau`.

2. Rappeler en quoi consiste la dissimilarité de Dice pour des données binaires.

   Correction. Pour deux individus donnés e_i et e_j, de variables $X_i = (X_i^1, \dots, X_i^p)$ et $X_j = (X_j^1, \dots, X_j^p)$, on note :
   $a_{ij} = \text{Card}\{k, 1 \leqslant k \leqslant p, X_i^k = X_j^k = 1\}$ (nombre de variables valant 1 pour e_i et pour e_j),
   $a_{i\bar{j}} = \text{Card}\{k, 1 \leqslant k \leqslant p, X_i^k = 1, X_j^k = 0\}$ (variables valant 1 pour e_i et 0 pour e_j),
   $a_{\bar{i}j} = \text{Card}\{k, 1 \leqslant k \leqslant p, X_i^k = 1, X_j^k = 0\}$ (variables valant 0 pour e_i et 1 pour e_j),
   $a_{\bar{i}\bar{j}} = \text{Card}\{k, 1 \leqslant k \leqslant p, X_i^k = 1, X_j^k = 0\}$ (variables valant 0 pour e_i et 0 pour e_j).
   La dissimilarité de Dice est alors définie par

$$\delta(e_i, e_j) = 1 - \frac{2a_{ij}}{2a_{ij} + a_{\bar{i}j} + a_{i\bar{j}}}.$$

3. Voir fichier R
4. Voir fichier R
5. Voir fichier R
6. Expliquer pourquoi la méthode des k-moyennes n'est pas adaptée à ces données.

   Correction. Les données ne sont pas quantitatives, et donc la méthode des k-myennes ne peut pas être utilisée. En effet cette méthode implique de calculer les moyennes des variables constituant chaque groupe, ce qui n'a pas de sens pour des variables qualitatives.

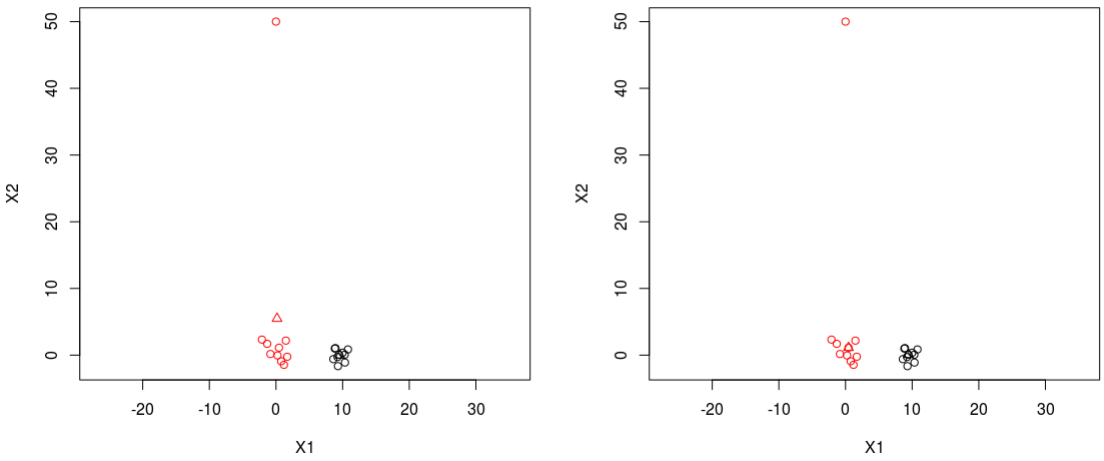
7. Expliquer quelle différence il y a entre la méthode des k-moyennes et la méthode des k-médoïdes.

   Correction. La seule différence se situe lors de la mise à jour des centres de chaque groupe. Pour la méthode des k-médoïdes ces centres correspondent aux individus les plus centraux, minimisant la somme des dissimilarités aux autres individus du groupe. Pour la méthode des k-moyennes les centres correspondent à des moyennes des variables du groupe. Par

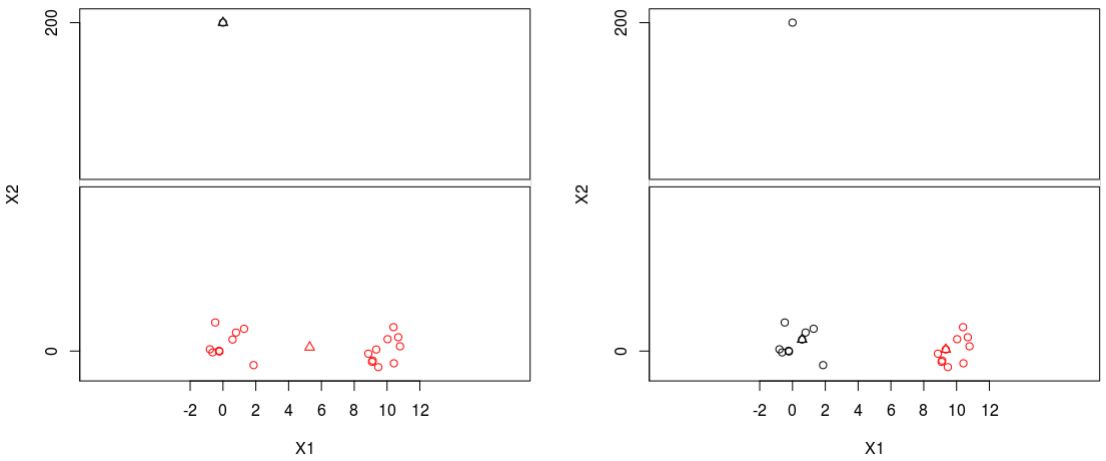
conséquent la méthode des k-médoïdes peut être utilisée pour n'importe quelles données d'entrées (qualitatives ou quantitatives) ; elle peut être appliquée en connaissant uniquement une matrice de dissimilarités entre les individus.

8. Voir fichier R
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.

   Correction. On peut imaginer une situation où les données sont nettement séparées en deux classes mais avec un individu "outlier", avec une valeur de variable erronée. A n'importe quelle étape de l'algorithme des k-moyennes (utilisé avec $k = 2$ ici), la moyenne du centre du groupe contenant l'outlier va se retrouver loin du groupe, de sorte qu'à l'étape suivante les individus de ce groupe risquent de se retrouver affectés à l'autre groupe. Pour l'algorithme des k-médoïdes le centre reste dans tous les cas bien positionné au sein de son groupe.



Situation 1 : les deux algorithmes (k-moyennes à gauche et k-médoïdes à droite) renvoient des groupes cohérents et l'outlier est simplement affecté à l'un des deux groupes. Cependant on voit que le centre du groupe rouge (représenté par un triangle) est fortement décentré pour les k-moyennes alors qu'il reste bien centré pour les k-médoïdes.



Situation 2 : Cette fois l'outlier est positionné beaucoup plus à l'écart (l'axe des ordonnées a été coupé) et la situation précédente n'est plus stable pour les k-moyennes : l'outlier va finalement former un groupe à lui tout seul, tandis que tous les autres individus se retrouvent dans un deuxième groupe. Pour les k-médoïdes on retrouve la partition cohérente en deux groupes.