TP – Clustering et classification

Andi WANG

— En utilisant R appliquez les fonctions de clustering (classification nonsupervisée) sur les données mammal.dentition qui existent dans la base R dans la librarie cluster.datasets. Il faut, naturellement, ignorer la première colonne. Représentez (listez) les résultats de chaque méthode de clustering.

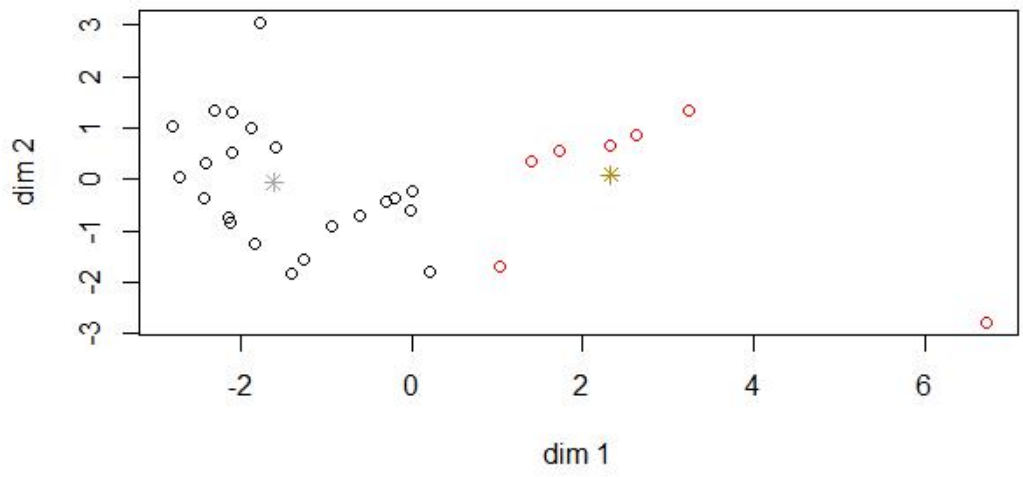
Interprétez les résultats obtenus.

Avant de classification, comme il y trop de dimensions variables dans ces données, il n'est pas possible d'afficher tous les variables en une image. Donc, pour bien afficher et pour plus facilement analyser, on a d'abord fait une analyse en composantes principales, on a obtenu 8 composantes (valeurs propres) pour ces données. Et puis on choisit les première 2 composantes qui sont les plus grandes pour analyser des questions suivantes. En utilisant les 2 vecteurs propres correspondus à ces deux valeurs propres, et puis on tranforme chaque objet à 2 dimensions (2 variables).

Et puis on utilise des méthodes : k-means méthode, k-medoids méthode, hierarchical et la méthode EM pour analyser des données.

k-means méthode :

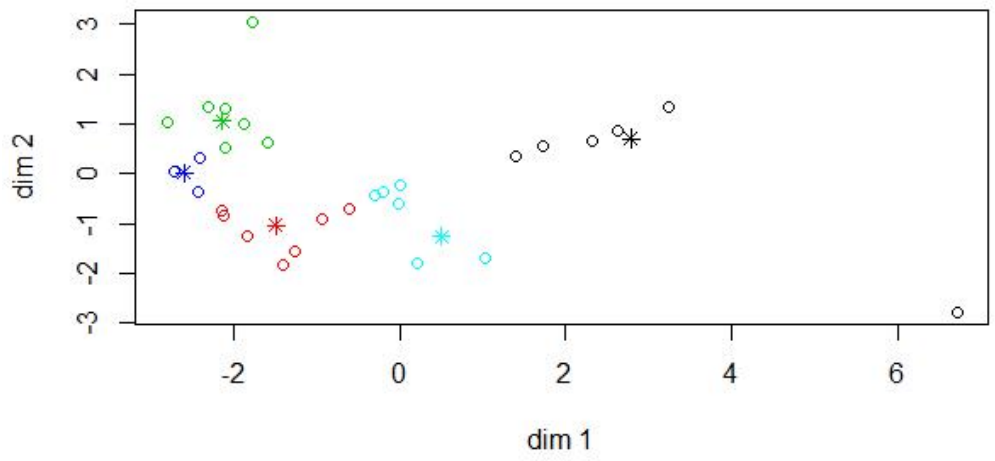
quand k=2 :



Séparation de 2 clusters (index de objets):

```
grp1:1  2  3  4  5  6  7  8  9 10 11 32 33 34 35 36 37 38 39 40 41
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 61
grp2 : 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
59 60 62 63 64 65 66
```

Quand k=5 :



Séparations de 5 clusters (index de objet):

grp1 :

12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
grp2 :

1 2 4 5 6 11 32 33 34 35 37 57
grp3 :

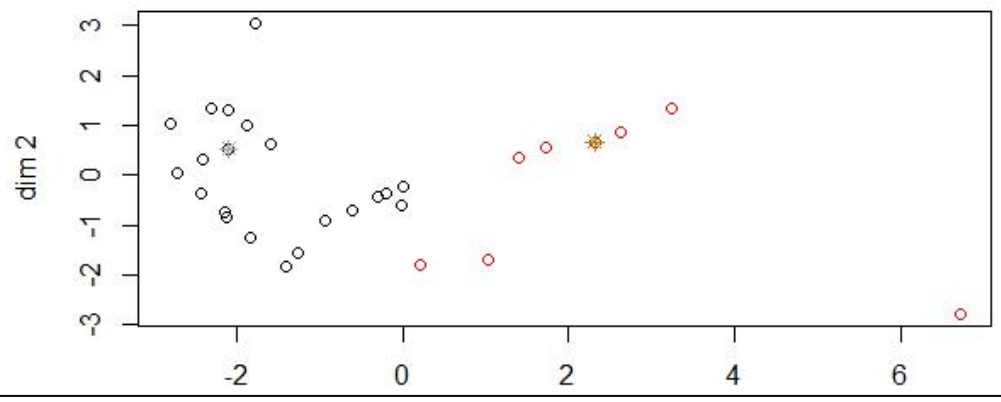
40 41 42 44 45 47 48 49 50 51 52 53 54 55 56
grp4 :

36 38 39 43 46
grp5 :

3 7 8 9 10 58 59 60 61 62 63 64 65 66

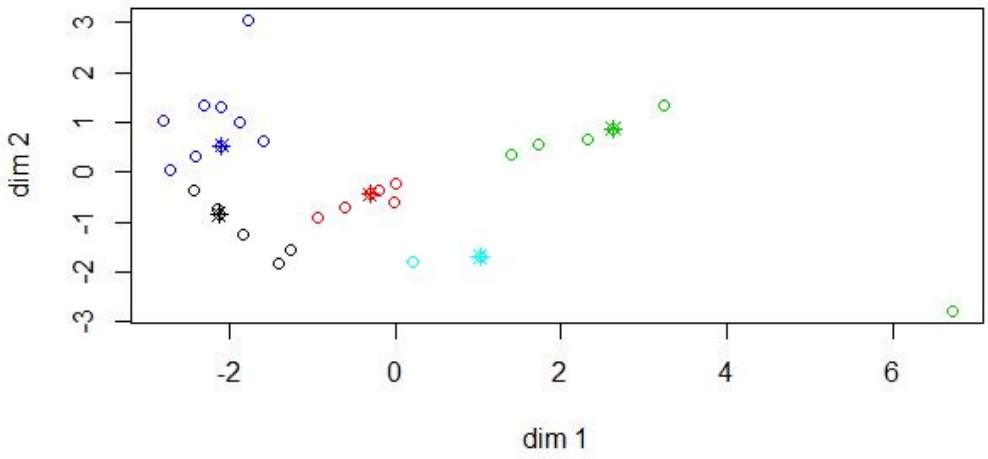
k-medoids méthodes :

Quand k=2 :



Séparation de 2 clusters(index de objet) :
grp1 : 1 2 3 4 5 6 7 8 9 10 11 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
grp2 : 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 58 59 60 61 62 63 64 65 66

Quand k=5 :



Séparation de 5 clusters (index de objets) :

grp1

1 2 4 32 33 34 35 36 37
grp2

3 5 6 7 8 9 10 11 57

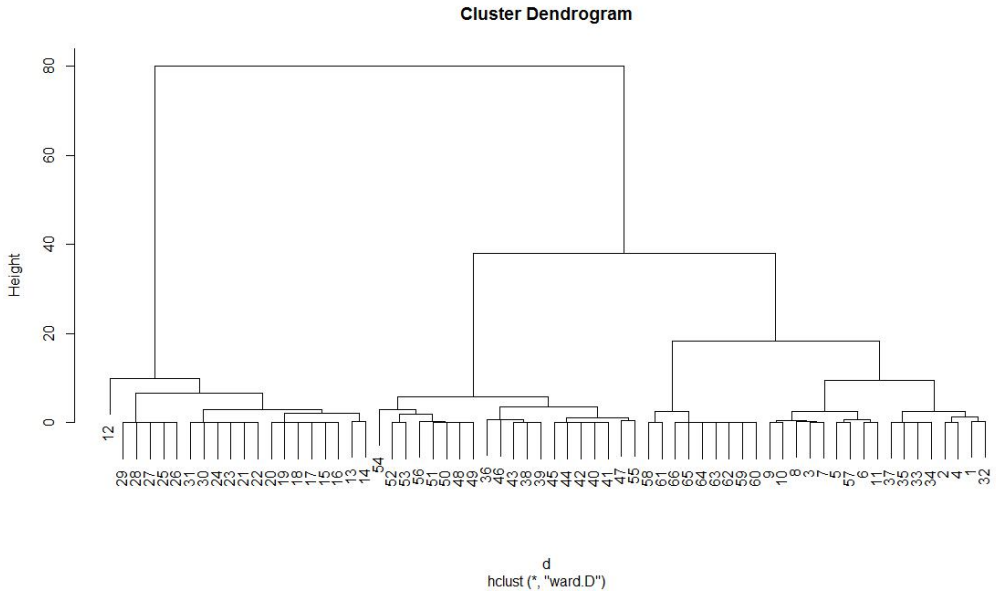
grp3

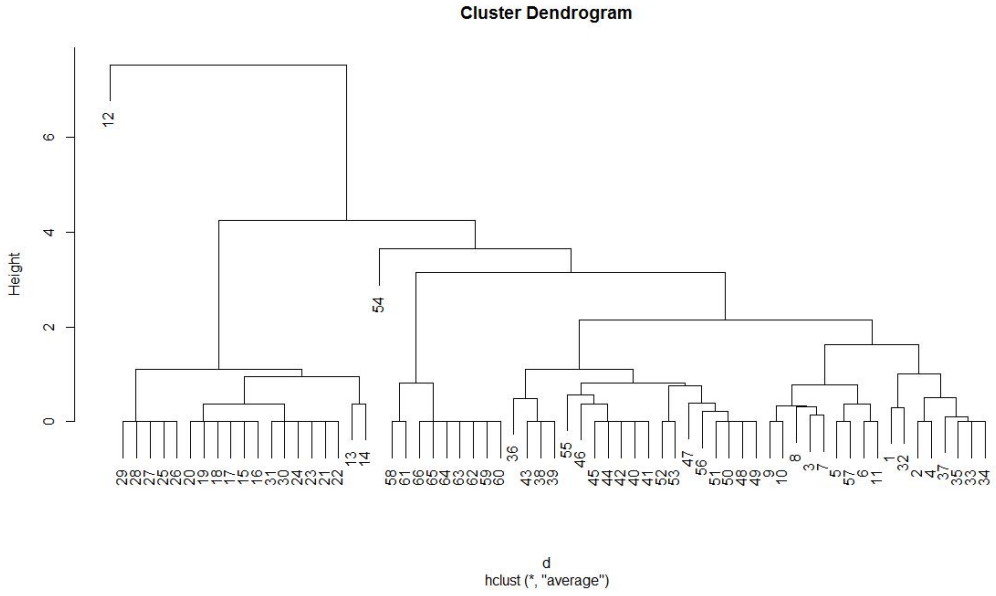
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
grp4

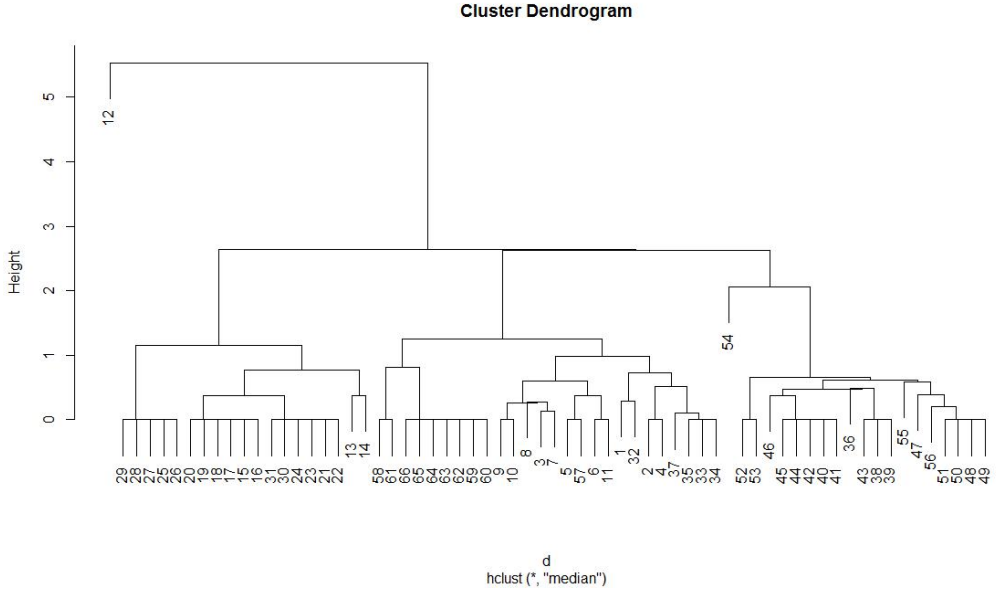
38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56
grp5

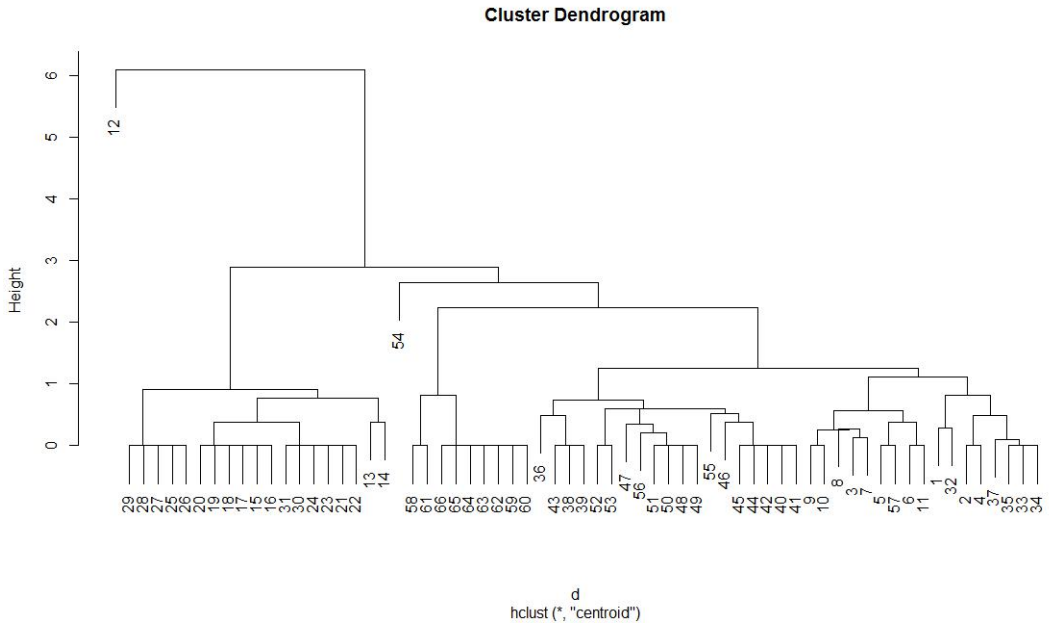
58 59 60 61 62 63 64 65 66

Hierarchical:

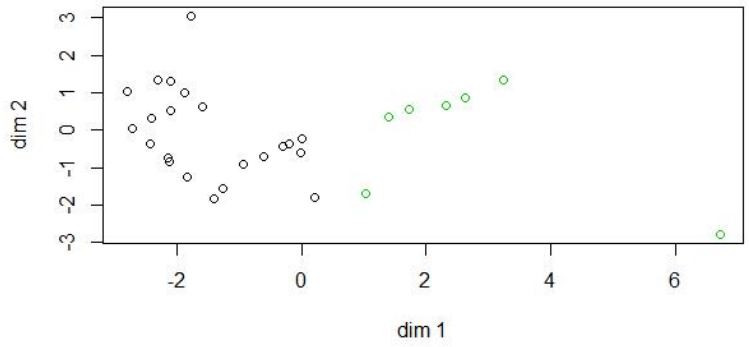


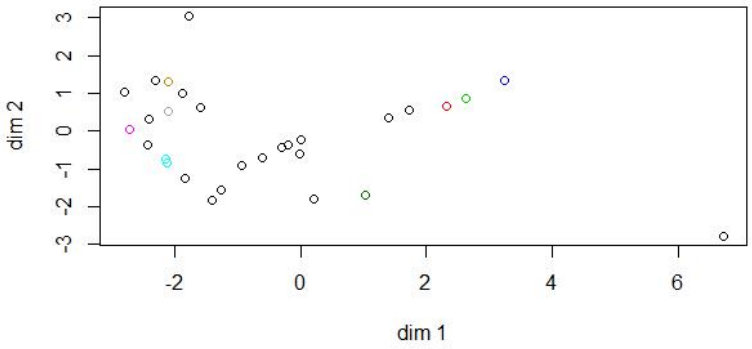






Méthode EM:





On peut voir tous les résultats obtenus par des méthodes différentes, quand k=2, on peut obtenir presque des même clusters sauf seulement 1 ou 2 points. Mais quand k=5, on peut voir que les résultats sont beaucoup différents à cause de méthodes internes. Quand on utilise des méthodes différentes, on va trouver peut-être une convergence différente pour chaque méthode. Donc on peut obtenir des résultats différents. Quand k plus grand, les différences sont de même plus grande.

— Classifiez manuellement les animaux, puis utilisez maintenant la fonction de classification par la méthode K-NN. Dressez la matrice de confusion.

On va d'abord séparer des données manuellement d'après des résultats que l'on a reçu dans les questions précédentes.

k=1:

```
> mknn1 [1] 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2
Levels: 1 2
> table(mknn1, donnees[J,3])
mknn1  1  2
    1 14  0
    2  0  8
```

k=3:

```
> mknn3 [1] 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2
Levels: 1 2
> table(mknn3, donnees[J,3])
mknn3  1  2
    1 14  0
    2  0  8
```

k=7:

```
> mknn7 [1] 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2
Levels: 1 2
> table(mknn7, donnees[J,3])
mknn7  1  2
    1 14  0
    2  0  8
```

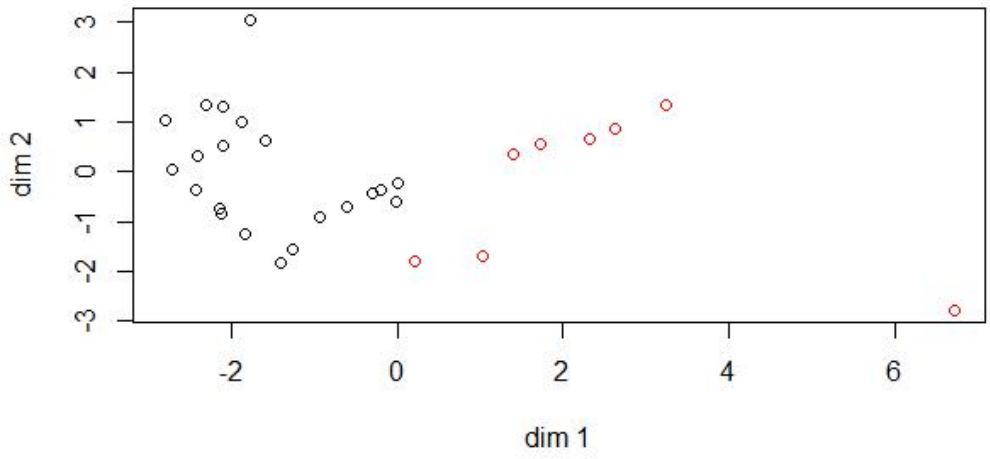
k=11:

```
> mknn11
[1] 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1
Levels: 1 2
> table(mknn11, donnees[J,3])
mknn11  1  2
     1 14  2
     2  0  6
```

Cross-Validation:

```
> model
[1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1
1 1 1 1 1 1 1 1
[41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
Levels: 1 2
> table(cl,model)
model
cl   1  2
  1 37  2
  2  0 27
```

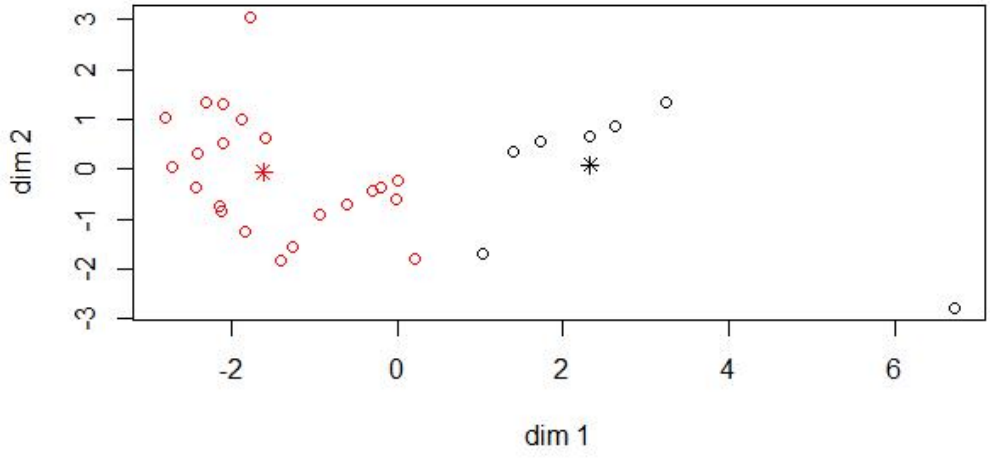
On dessine des clusters de ces résultats:



— (optionnel) Programmez dans le langage de votre choix la méthode de K-moyennes.

Compte-rendu sous la forme d'un unique fichier (code R ou archive .zip ou .rar) à rendre sur Campus.

Test du programme avec les mêmes données comme précédantes:



Résultats de clusters:

```
[1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1
1 1 1 1 1 1 1
[41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 2 2 2
```

C'est le même des résultats de question 1.