Méthodes de Partionnement et d'apprentissage non supervisé Classification Hiérarchique et Kmeans

Anne Badel et Frédéric Guyon

2019-02-20

Partitionnement et apprentissage

Partionnement = Clustering

Partionnement = Clustering

Apprentissage

Apprentissage: Séparation linéaire

Méthodes

Géométrie et distances

Partitionnement et apprentissage

Partitionnement et apprentissage

Partionnement = Clustering

Partionnement = Clustering

Partionnement = Clustering

Partionnement = Clustering

Apprentissage

Apprentissage

Apprentissage: Séparation linéaire

Apprentissage: Séparation linéaire

Méthodes

Méthodes

Géométrie et distances

Géométrie et distances

Les données

Les données

Les variables

Les variables

UVisualisation des données

Visualisation des données

Cas d'étude : TCGA Breast Invasive Cancer (BIC)

Cas d'étude : TCGA Breast Invasive Cancer (BIC)

TP : analyse de données d'expression

TP : analyse de données d'expression

Géométrie et distances

Géométrie et distances

Distances

Distances

Distances utilisées dans R

Distances utilisées dans R

Distances utilisées dans R

Distances utilisées dans R

Autres distances non géométriques (pour information)

Autres distances non géométriques (pour information)

└ Distances plus classiques en génomique

Distances plus classiques en génomique

☐ Distances entre groupes

Distances entre groupes

☐ Distances entre groupes

Distances entre groupes

Les données

Les données

```
Les données
```

```
str(mes.iris)
```

```
'data.frame': 150 obs. of 4 variables:
```

\$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 \$ Sepal.Width: num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.5

\$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 : \$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.2

summary(mes.iris)

Separ.	Lengtn	Sepal.	wiath	Petal.	Lengtn	Petal.	.Wl
Min.	:4.300	Min.	:2.000	Min.	:1.000	Min.	:0
1st Qu.	:5.100	1st Qu.	:2.800	1st Qu	:1.600	1st Qu	.:0
Median	:5.800	Median	:3.000	${\tt Median}$:4.350	Median	:1
Mean	:5.843	Mean	:3.057	Mean	:3.758	Mean	:1

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1 Max. :7.900 Max. :4.400 Max. :6.900 Max. :2

UVisualisation des données

Visualisation des données

Visualisation des données - coloration par espèces

Visualisation des données - coloration par espèces

UVisualisation des données

Visualisation des données

Nettoyage des données (1): données manquantes

Nettoyage des données (1): données manquantes

└ Nettoyage des données (2) : variables constantes

Nettoyage des données (2) : variables constantes

└ Normalisation

Normalisation

On peut visuellement regarder l'effet de la normalisation :

On peut visuellement regarder l'effet de la normalisation :

Méthodes de Partionnement et d'apprentissage non supervisé

On peut visuellement regarder l'effet de la normalisation :

par un plot des données

par un plot des données

Méthodes de Partionnement et d'apprentissage non supervisé								
On peut visuellement regarder l'effet de la normalisation :								
par un plot des données								

Méthodes de Partionnement et d'apprentissage non supervisé

On peut visuellement regarder l'effet de la normalisation :

... par une boîte à moustaches (boxplot)

... par une boîte à moustaches (boxplot)

Méthodes de Partionnement et d'apprentissage non supervisé								
On peut visuellement regarder l'effet de la normalisation :								
└ par une boîte à moustaches (boxplot)								

Méthodes de Partionnement et d'apprentissage non supervisé

On peut visuellement regarder l'effet de la normalisation :

... par une image

... par une image



Méthodes de Partionnement et d'apprentissage non supervisé

On peut visuellement regarder l'effet de la normalisation :

... par une projection sur une ACP

... par une projection sur une ACP

Méthodes de Partionnement et d'apprentissage non supervisé

La matrice de distances

La matrice de distances

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

La classification hiérarchique

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

Principe

Principe

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

Principe

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

Notion importante, cf distances

Notion importante, cf distances

Méthodes de Partionnement et d'apprentissage non supervisé	
La classification hiérarchique	
Notion importante, cf distances	

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

L'algorithme

L'algorithme

L'algorithme

étape 1:

- départ : n individus = n clusters distincts
- calcul des distances entre tous les individus
 - choix de la métrique à utiliser en fonction du type de données
- ▶ regroupement des 2 individus les plus proches => (n-1) clusters



Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

au départ

au départ



La classification hiérarchique

Lidentification des individus les plus proches

identification des individus les plus proches

Méthodes de Partionnement et d'apprentissage non supervisé	
La classification hiérarchique	
identification des individus les plus proches	

Méthodes de Partionnement et d'apprentissage non supervisé
La classification hiérarchique

Construction du dendrogramme

construction du dendrogramme

Méthodes de Partionnement et d'apprentissage non supervisé			
L	· La classification hiérarchique		
	construction du dendrogramme		

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

étape j :

étape j :

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

étape j :

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

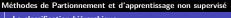
└calcul des nouveaux représentants BE et CD

calcul des nouveaux représentants BE et CD

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

calcul des nouveaux représentants BE et CD



La classification hiérarchique

<u>calcul des dis</u>tances de l'individu restant A aux points moyens

calcul des distances de l'individu restant A aux points moyens

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

calcul des distances de l'individu restant A aux points moyens

Méthodes de Partionnement et d'apprentissage non supervisé
La classification hiérarchique

A est plus proche de ...

A est plus proche de ...

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

A est plus proche de ...

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

dendrogramme

dendrogramme

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

dendrogramme

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

pour finir

pour finir

∟pour finir

ightharpoonup à l'étape (n-1), tous les individus sont regroupés dans un même cluster

Méthodes de Partionnement et d'apprentissage non supervisé
La classification hiérarchique

dendrogramme final

dendrogramme final



Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

└ Je ne fais pas attention à ce que je fais ...

Je ne fais pas attention à ce que je fais ...

Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

Je ne fais pas attention à ce que je fais ...

Méthodes de Partionnement et d'apprentissage non supervisé
La classification hiérarchique

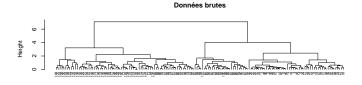
Sur données normalisées

Sur données normalisées

```
La classification hiérarchique
```

Sur données normalisées

```
par(mfrow = c(2, 1))
plot(iris.hclust, hang = -1, cex = 0.5, main = "Données bro
plot(iris.scale.hclust, hang = -1, cex = 0.5, main = "Normal")
```





Normalisées



Méthodes de Partionnement et d'apprentissage non supervisé

La classification hiérarchique

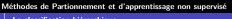
Sur données normalisées

La classification hiérarchique

└En utilisant une autre métrique

En utilisant une autre métrique

Méthodes de Partionnement et d'apprentissage non supervisé	
L	-La classification hiérarchique
	En utilisant une autre métrique



La classification hiérarchique

En utilisant un autre critère d'aggrégation

En utilisant un autre critère d'aggrégation

Les k-means

Les k-means

L'algorithme

L'algorithme

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

— étape 1 :

étape 1 :

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

_étape 1 :

L'algorithme

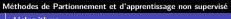
Choix des centres provisoires

Choix des centres provisoires

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Choix des centres provisoires



L'algorithme

Calcul des distances aux centres provisoires

Calcul des distances aux centres provisoires

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Calcul des distances aux centres provisoires

Affectation à un cluster

Affectation à un cluster

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

∟ Affectation à un cluster

L'algorithme

Calcul des nouveaux centres de classes

Calcul des nouveaux centres de classes

Méthodes de Partionnement et d'apprentissage non supervisé
L'algorithme
L'Etape j :

Etape j:

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Etape j:

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

⊢Fin:

Fin:

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Arrêt:

Arrêt:

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Arrêt:

L'algorithme

Un premier k-means en 5 groupes

Un premier k-means en 5 groupes

```
L'algorithme
```

Un premier k-means en 5 groupes

iris.scale.kmeans5\$cluster

table(iris.scale.kmeans5\$cluster)

Méthodes de Partionnement et d'apprentissage non supervisé
L'algorithme
Un premier k-means en 5 groupes

Visualisation des clusters

Visualisation des clusters

Méthodes de Partionnement et d'apprentissage non supervisé

L'algorithme

Visualisation des clusters

Méthodes de Partionnement et d'apprentissage non supervisé

└─L'algorithme

Combien de clusters ?

Combien de clusters ?

Méthodes de Partionnement et d'apprentissage non supervisé
L'algorithme
Combien de clusters ?

L'algorithme

Classification hiérarchique

Classification hiérarchique





└_K-means

K-means

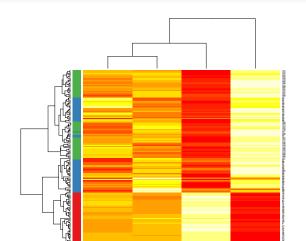




Heatmap

Heatmap

Heatmap



Comparaison de clustering: Rand Index

Comparaison de clustering: Rand Index

Méthodes de Partionnement et d'apprentissage non supervisé

Comparaison de clustering: Adjusted Rand Index

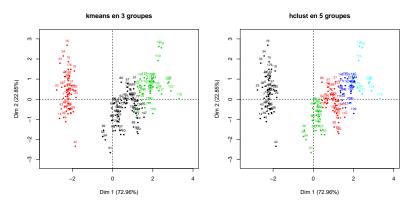
Comparaison de clustering: Adjusted Rand Index

Comparaison des résultats des deux classifications

Comparaison des résultats des deux classifications

par une visualisation

```
par(mfrow=c(1,2))
plot(iris.scaled.acp, col.ind=cluster.kmeans3, choix="ind"
plot(iris.scaled.acp, col.ind=cluster.hclust5, choix="ind"
```



par(mfrow=c(1 1))

Comparaison avec la réalité

Comparaison avec la réalité

Méthodes de Partionnement et d'apprentissage non supervisé

Comparaison avec la réalité

La réalité

La réalité





Comparer k-means avec la réalité

Comparer k-means avec la réalité

 $\begin{tabular}{ll} \bf M\acute{e}thodes \ de \ Partionnement \ et \ d'apprentissage \ non \ supervis\'e \\ \begin{tabular}{ll} $ & \\ \hline \ & \\ \ & \\ \hline \ & \\ \ & \\ \hline \ & \\ \ & \\ \hline \ & \\ \ & \\ \hline \ & \\ \$

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

Setosa vs others

Setosa vs others

```
Comparer k-means avec la réalité
Setosa vs others
```

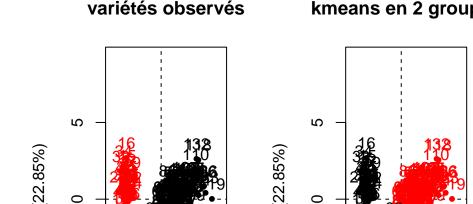
Visualisation

```
variete2 <- rep("notSetosa", 150)
variete2[variete=="setosa"] <- "setosa"
variete2 = factor(variete2)
table(variete2)</pre>
```

```
variete2
notSetosa setosa
100 50
```

Comparer k-means avec la réalité

```
par(mfrow=c(1,2))
plot(iris.scaled.acp, col.ind=variete2, title="variétés obscluster.kmeans2 <- kmeans(mes.iris.scaled, center=2)$cluster.kmeans2, title="kmeans2", title="
```



Setosa vs others

Table de confusion et calcul de performances

conf.kmeans <- table(variete2, cluster.kmeans2)
kable(conf.kmeans)</pre>

	1	2
notSetosa	3	97
setosa	50	0

```
Comparer k-means avec la réalité
```

▶ table de confusion, taux de bien prédits, spécificité, sensibilité,

```
TP \leftarrow conf.kmeans[1,1]
FP \leftarrow conf.kmeans[1,2]
FN \leftarrow conf.kmeans[2,1]
TN \leftarrow conf.kmeans[2,2]
P <- TP + FN # nb positif dans la réalité
N <- TN + FP
                       # nb négatif dans la réalité
FPrate <- FP / N # = false alarm rate
Spe <- TN / N \# = sp\acute{e}cificit\acute{e}
Sens <- recall <- TPrate <- TP / P # = hit rate ou re
PPV <- precision <- TP / (TP + FP)
accuracy \leftarrow (TP + TN) / (P + N)
F.measure <- 2 / (1/precision + 1/recall)
performance <- c(FPrate, TPrate, precision, recall, accuracy
names(performance) <- c("FPrate", "TPrate", "precision", ";</pre>
```

Comparer k-means avec la réalité

Setosa vs others

kable(performance, digits=3)

	×
FPrate	1.000
TPrate	0.057
precision	0.030
recall	0.057
accuracy	0.020
F.measure	0.039
Spe	0.000
PPV	0.030



Setosa vs others

rand index et adjusted rand index

clues::adjustedRand(as.numeric(variete2), cluster.kmeans2)

Rand HA MA FM Jaccard 0.9605369 0.9204051 0.9208432 0.9639434 0.9302767

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

Setosa vs others

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

└Versicolor vs !Versicolor

Versicolor vs !Versicolor

```
Comparer k-means avec la réalité

Versicolor vs !Versicolor
```

Visualisation

```
variete2 <- rep("notVersicolor", 150)</pre>
variete2[variete=="versicolor"] <- "versicolor"</pre>
variete2 = factor(variete2)
table(variete2)
variete2
notVersicolor
                versicolor
          100
                          50
par(mfrow=c(1,2))
plot(iris.scaled.acp, col.ind=variete2)
cluster.kmeans2 <- kmeans(mes.iris.scaled, center=2)$clust@
plot(iris.scaled.acp, col.ind=cluster.kmeans2)
```

Individuals factor map (PC Individuals factor map

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

Versicolor vs !Versicolor

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

Table de confusion et calcul de performances

Table de confusion et calcul de performances

Comparer k-means avec la réalité

☐ Table de confusion et calcul de performances

kable(performance, digits=3)

	X	
FPrate	0.943	
TPrate	0.515	
precision	0.500	
recall	0.515	
accuracy	0.353	
F.measure	0.508	
Spe	0.057	
PPV	0.500	

clues::adjustedRand(as.numeric(variete2), cluster.kmeans2)

Rand HA MA FM Jaccard 0.53995526 0.07211421 0.07722223 0.57895580 0.40737752

Méthodes de Partionnement et d'apprentissage non supervisé

Comparer k-means avec la réalité

└─Table de confusion et calcul de performances

 $Contact:\ anne.badel @univ-paris-diderot.fr$