Clustering hiérarchique et critères de (dis)similarité

Master parcours SSD - UE Apprentissage Statistique I

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

Rapppels : clustering

Clustering

Clustering = **classification** non-supervisée

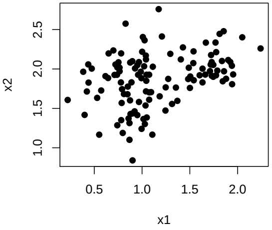
- catégoriser les observations (sous-populations)
- ...ou catégoriser les variables (corrélation / redondance)

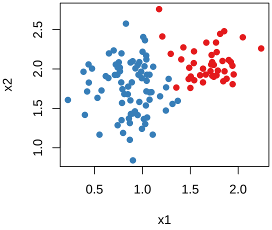
Objectifs : exploratoire

- présence de sous-groupes dans les données
- adéquation critères de similarité / données

Clustering

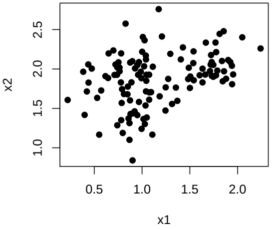
Catégoriser les observations ?

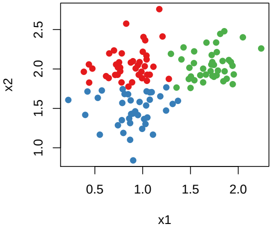




Clustering

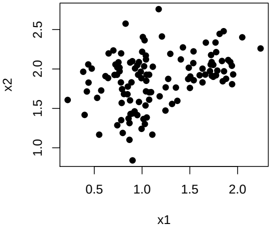
Catégoriser les observations ? Oui mais...

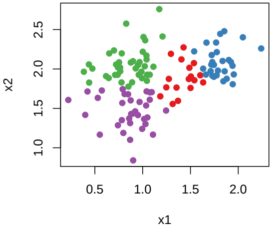




Clustering

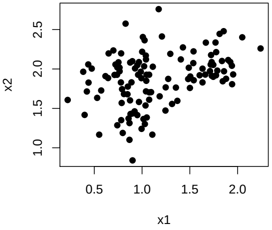
Catégoriser les observations ? Oui mais...

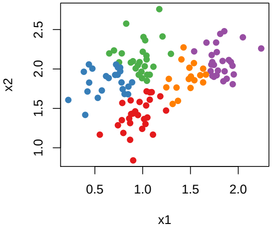




Clustering

Catégoriser les observations ? Oui mais...





Clustering - qualité

But du clustering :

- déterminer des ensembles de points proches
- qui soient distants les uns des autres

Clustering - qualité

But du clustering :

- déterminer des ensembles de points proches
- qui soient distants les uns des autres

Fonction objective (à minimiser) = dispersion "intra" cluster

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d(x_i, x_j)$$

- K = nombre de clusters
- C = clustering : $C(i) = k \leftrightarrow x_i \in$ cluster k
- $d(x, y)$ = distance/dissimilarité entre x et y

$\Rightarrow$ problème combinatoire, présence de minima locaux.

Clustering - challenges

Questions centrales :

- choisir la fonction de distance entre observations
  - dicté par nature du problème et des données
- choisir le nombre de clusters
  - pas de réponse absolue, tester plusieurs valeurs
- évaluer la stabilité du clustering

Méthodes clé :

- clustering hiérarchique
- K-means
- modèles de mélanges

Clustering hiérarchique

Clustering hiérarchique

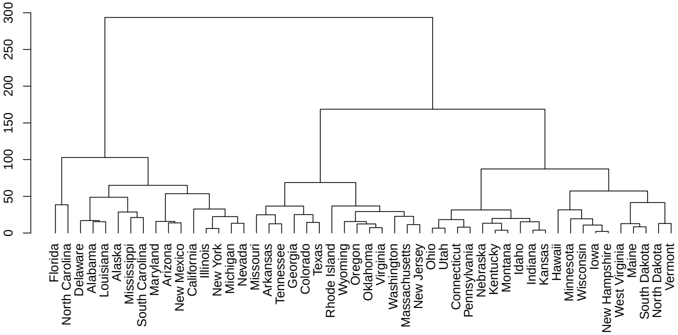


Figure: Clustering hiérarchique du jeu `USArrests`.

- procédure itérative d'agglomération (ou de division)
- s'appuie sur une mesure de distance
- le plus simple des algorithmes de clustering

Clustering hiérarchique - algorithme

Algorithme :

1. Introduire un cluster par observations
2. Calculer la similarité entre observations/clusters
3. Tant que le nombre de clusters est > 1
   3.1 fusionner les deux clusters les plus similaires
   3.2 re-calculer la similarité entre clusters

Clustering hiérarchique - algorithme

Algorithme :

1. Introduire un cluster par observations
2. Calculer la similarité entre observations/clusters
3. Tant que le nombre de clusters est > 1
   3.1 fusionner les deux clusters les plus similaires
   3.2 re-calculer la similarité entre clusters

$\Rightarrow$ définit une hiérarchie de partitions

Clustering hiérarchique - algorithme

Algorithme :

1. Introduire un cluster par observations
2. Calculer la similarité entre observations/clusters
3. Tant que le nombre de clusters est > 1
   3.1 fusionner les deux clusters les plus similaires
   3.2 re-calculer la similarité entre clusters

$\Rightarrow$ définit une hiérarchie de partitions

$\Rightarrow$ on la résume dans un arbre appelé dendrogramme

Clustering hiérarchique - algorithme

Algorithme :

1. Introduire un cluster par observations
2. Calculer la similarité entre observations/clusters
3. Tant que le nombre de clusters est > 1
   3.1 fusionner les deux clusters les plus similaires
   3.2 re-calculer la similarité entre clusters

$\Rightarrow$ définit une hiérarchie de partitions

$\Rightarrow$ on la résume dans un arbre appelé dendrogramme

$\Rightarrow$ le nombre de clusters n'est pas défini à l'avance

Clustering hiérarchique - illustration

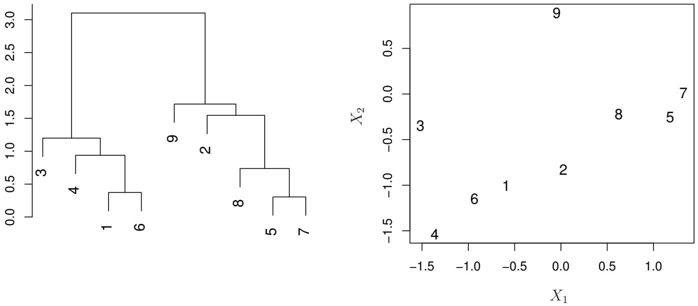


Figure: Figure tirée de James et al. (2013)

Clustering hiérarchique - en pratique

Questions ouvertes :

- la mesure de distance entre observations
- la mesure de distance entre clusters
- l'interprétation du dendrogramme
- le choix du nombre de clusters

Clustering hiérarchique - distances inter-observations

Quelle critère de distance pour quelles observations ?

Clustering hiérarchique - distances inter-observations

Quelle critère de distance pour quelles observations ?

- très dépendant du problème
- choix usuel / par défaut = distance Euclidienne
- voir 2nde partie du cours

Clustering hiérarchique - distances inter-clusters

Distance entre clusters : 3 stratégies principales

- "average" : distance moyenne entre observations :

$$d_C(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} d(x_1, x_2)$$

- "complete" : distance maximale entre observations

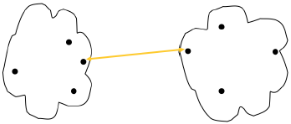
$$d_C(C_1, C_2) = \max\left\{ d(x_1, x_2) : \ \ x_1 \in C_1, \ x_2 \in C_2 \right\}$$

- "single" : distance minimale entre observations

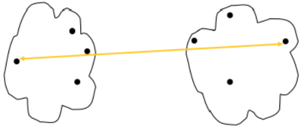
$$d_C(C_1, C_2) = \min\left\{ d(x_1, x_2) : \ \ x_1 \in C_1, \ x_2 \in C_2 \right\}$$

Clustering hiérarchique - distances inter-clusters

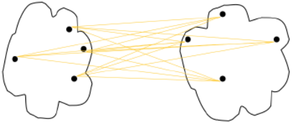
Distance min



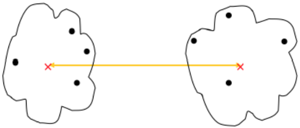
Diamètre maximum



Distance moyenne

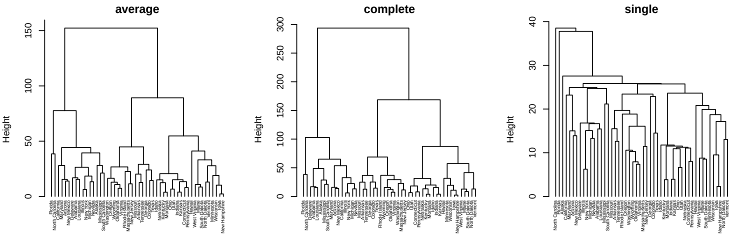


Distance des centres de gravité



Clustering hiérarchique - distances inter-clusters

Illustration :



⇒ impact fort sur les résultats !

Clustering hiérarchique - distances inter-clusters

Comment choisir ?

- pas de règle bien définie $\Rightarrow$ inspection du dendrogramme

Clustering hiérarchique - distances inter-clusters

Comment choisir ?

- pas de règle bien définie $\Rightarrow$ inspection du dendrogramme

Néanmoins (en général) :

Clustering hiérarchique - distances inter-clusters

Comment choisir ?

- pas de règle bien définie $\Rightarrow$ inspection du dendrogramme

Néanmoins (en général) :

- "single" & "complete" : sensibles aux bruit (outliers)

Clustering hiérarchique - distances inter-clusters

Comment choisir ?

- pas de règle bien définie $\Rightarrow$ inspection du dendrogramme

Néanmoins (en général) :

- "single" & "complete" : sensibles aux bruit (outliers)
- "single" : perte de compacité
  - cluster compact : toutes les observations sont proches
  - perdu par distance minimum par effet de chaînage

Clustering hiérarchique - distances inter-clusters

Comment choisir ?

- pas de règle bien définie $\Rightarrow$ inspection du dendrogramme

Néanmoins (en général) :

- "single" & "complete" : sensibles aux bruit (outliers)
- "single" : perte de compacité
  - cluster compact : toutes les observations sont proches
  - perdu par distance minimum par effet de chaînage
- "complete" : compacité, mais perte de "proximité"
  - "proximité" (closeness) : observations plus proches de celles de leur cluster que de celles des autres clusters.
  - pas garanti par l'utilisation du maximum

Clustering hiérarchique - distances inter-clusters

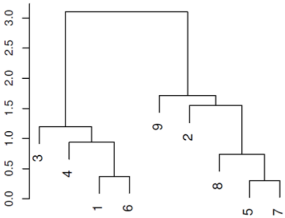
Comment choisir ?

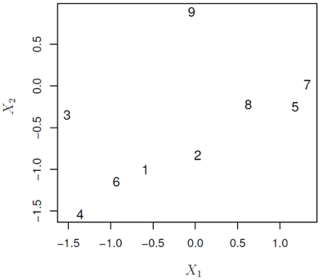
- pas de règle bien définie ⇒ inspection du dendrogramme

Néanmoins (en général) :

- "single" & "complete" : sensibles aux bruit (outliers)
- "single" : perte de compacité
  - cluster compact : toutes les observations sont proches
  - perdu par distance minimum par effet de chaînage
- "complete" : compacité, mais perte de "proximité"
  - "proximité" (closeness) : observations plus proches de celles de leur cluster que de celles des autres clusters.
  - pas garanti par l'utilisation du maximum
- "average" : bon compromis, clusters équilibrés

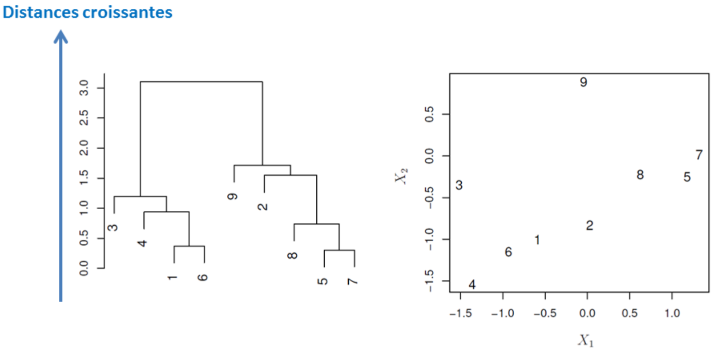
Interprétation du dendrogramme





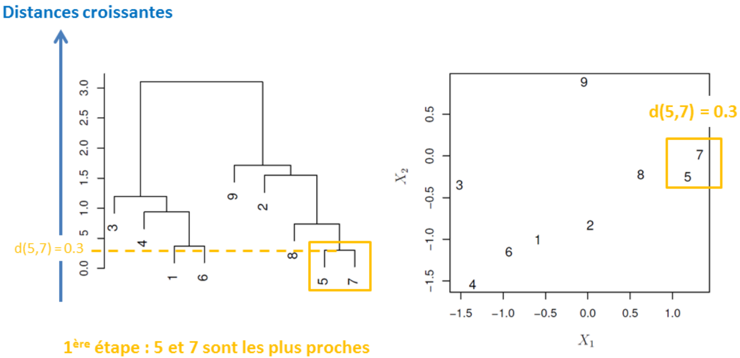
- on construit l'arbre de bas en haut

Interprétation du dendrogramme



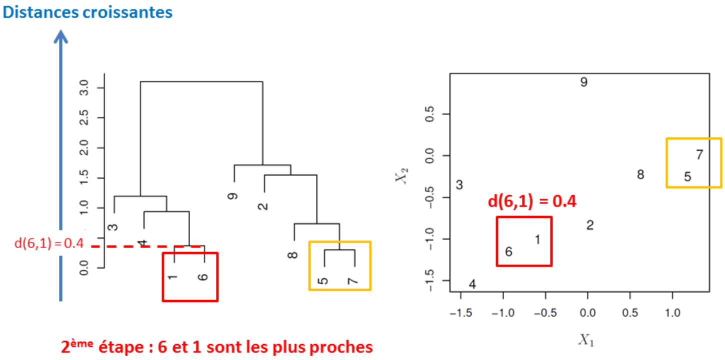
- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters

Interprétation du dendrogramme



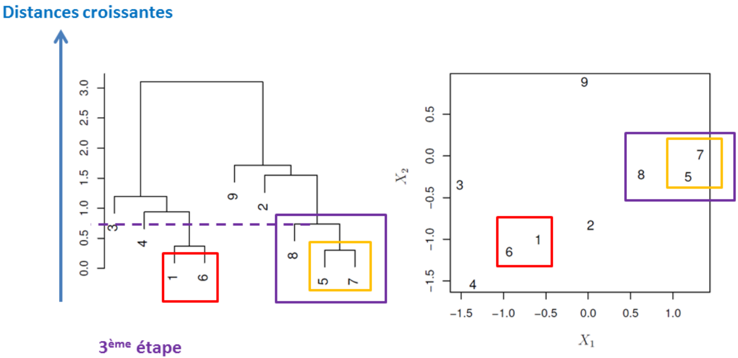
- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters
- hauteur d'un cluster = distance entre ses deux fils

Interprétation du dendrogramme



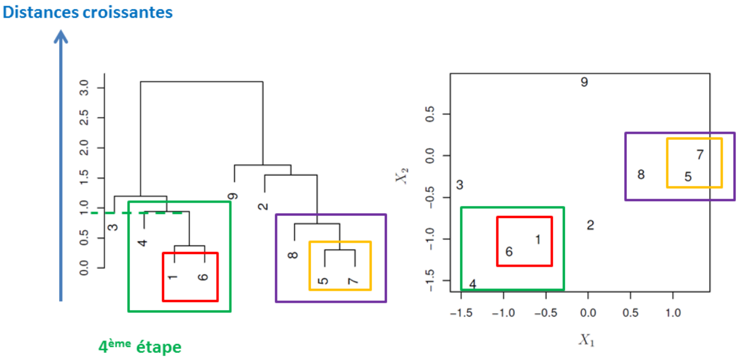
- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters
- hauteur d'un cluster = distance entre ses deux fils
- ...

Interprétation du dendrogramme



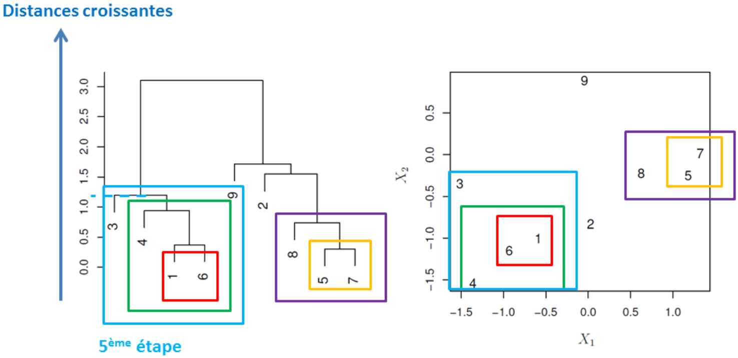
- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters
- hauteur d'un cluster = distance entre ses deux fils
- ...

Interprétation du dendrogramme



- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters
- hauteur d'un cluster = distance entre ses deux fils
- ...

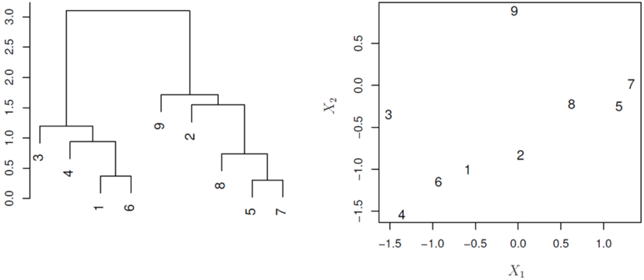
Interprétation du dendrogramme



- on construit l'arbre de bas en haut
- hauteur dans l'arbre = distance entre clusters
- hauteur d'un cluster = distance entre ses deux fils
- ...on continue jusqu'à ce qu'on ait 1 seul cluster

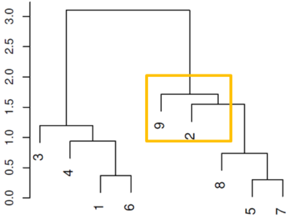
Interprétation du dendrogramme

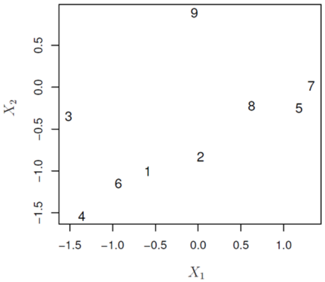
Attention :



Interprétation du dendrogramme

Attention :

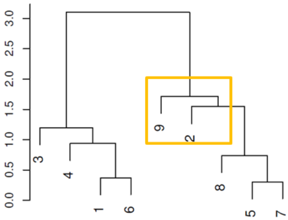


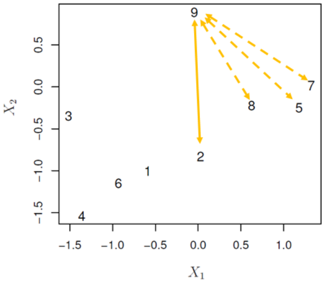


- il est tentant de conclure que 2 et 9 sont proches...

Interprétation du dendrogramme

Attention :





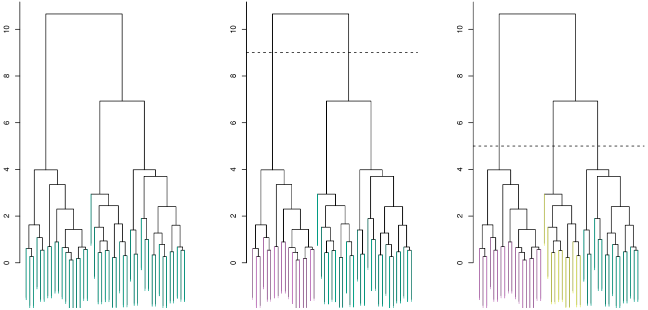
- il est tentant de conclure que 2 et 9 sont proches...
- ...mais 9 n'est pas plus proche de 2 que de 5, 7 ou 8.

$\Rightarrow$ proximité dans l'arbre $\not\Rightarrow$ distance faible

- c'est la hauteur dans l'arbre qui compte
- la topologie de l'arbre est (en partie) arbitraire

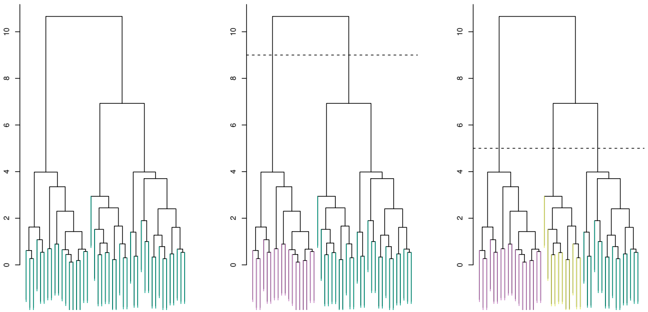
Choix du nombre de clusters

On obtient des clusters en coupant le dendrogramme :



Choix du nombre de clusters

On obtient des clusters en coupant le dendrogramme :



- choisir le "bon" nombre de clusters : difficile et subjectif
- intérêt du dendrogramme : continuum de solution
- choix essentiellement empirique - "à l'oeil"

(⇒ à suivre la semaine prochaine)

Conclusion

En conclusion :

- représentation très populaire
- fournit un ensemble de solution à différents niveaux de granularité
- très simple à mettre en oeuvre
- limite : influence assez forte du critère de "linkage"

Conclusion

En conclusion :

- représentation très populaire
- fournit un ensemble de solution à différents niveaux de granularité
- très simple à mettre en oeuvre
- limite : influence assez forte du critère de "linkage"

Extensions :

- autres critères de linkage
  - meilleure interprétation de la "coupe" du dendrogramme
- ré-échantillonnage pour améliorer la stabilité des clusters

Mise en oeuvre R

1. construction d'un objet de type `dist`
   - `d = dist(X)` si `X` = observations
   - `d = as.dist(D)` si `D` = matrice de distance

2. construction du dendrogramme : `hc = hclust(d)`
   - avec options de linkage

3. "coupe" du dendrogramme : `C = cutree(hc, ...)`
   - avec hauteur de coupe, ou nombre de clusters
   - `C` = vecteur donnant les indices des clusters

Heatmaps

Clustering hiérarchique & heatmaps

Clustering :

- catégoriser les observations (sous-populations)

Clustering hiérarchique & heatmaps

Clustering :

- catégoriser les observations (sous-populations)
- ...ou catégoriser les variables (corrélation / redondance)

Clustering hiérarchique & heatmaps

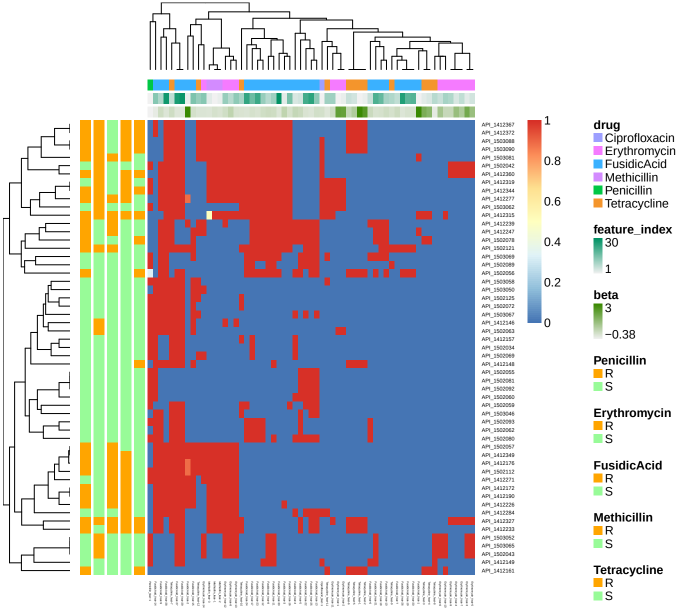
Clustering :

- catégoriser les observations (sous-populations)
- ...ou catégoriser les variables (corrélation / redondance)
- ... ou les deux !

Heatmaps :

- représentation graphique de matrice de données
- clustering hiérarchique pour ré-ordonner lignes & colonnes
- fait ressortir des structures "de bloc"
- très populaire en biologie
- très pratique quand on a beaucoup de variables

Heatmap



API_1412367
API_1412372
API_1503088
API_1503090
API_1503081
API_1502042
API_1412360
API_1412319
API_1412344
API_1412277
API_1503062
API_1412315
API_1412239
API_1412247
API_1502078
API_1502121
API_1503069
API_1502089
API_1502056
API_1503058
API_1503050
API_1502125
API_1502072
API_1503067
API_1412146
API_1502063
API_1412157
API_1502034
API_1502069
API_1412148
API_1502055
API_1502081
API_1502092
API_1502060
API_1502059
API_1503046
API_1502093
API_1502062
API_1502080
API_1502057
API_1412349
API_1412176
API_1502112
API_1412271
API_1412172
API_1412190
API_1412226
API_1412284
API_1412327
API_1412233
API_1503052
API_1503065
API_1502043
API_1412149
API_1412161
1
0.8
0.6
0.4
0.2
0
drug
Ciprofloxacin
Erythromycin
FusidicAcid
Methicillin
Penicillin
Tetracycline
feature_index
30
1
beta
3
−0.38
Penicillin
R
S
Erythromycin
R
S
FusidicAcid
R
S
Methicillin
R
S
Tetracycline
R
S

Heatmap

Principe :

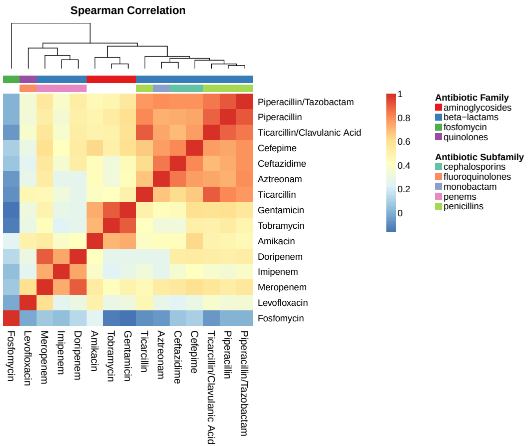
- représentation d'une matrice comme une image
  - "carte de chaleur" : couleur = intensité
- lignes & colonnes ré-ordonnées par clustering hiérarchique

Paramètres :

- fonction de distance entre lignes ou colonnes
- stratégie d'aggrégation pour le clustering hiérarchique
  - e.g., complete, single, average
- (+ spécifier si on ré-ordonne lignes & colonnes ou seulement l'un ou l'autre)

Mise en oeuvre

- fonction "native" `heatmap(X)`
- fonction `aheatmap(X)` du package NMF
  - ajout facile d'annotations sur lignes et/ou colonnes



Distances / Similarités

Distances et (dis)similarités

Point d'entrée de l'algorithme précédent :

- matrice de distance ou disimilarité entre les observations

$\Rightarrow$ quelle(s) mesure(s) utiliser ?

Distances et (dis)similarités

Point d'entrée de l'algorithme précédent :

- matrice de distance ou disimilarité entre les observations

$\Rightarrow$ quelle(s) mesure(s) utiliser ?

1. distance vs (dis)similarité
2. variables quantitatives : distances & espaces vectoriels
3. la distance Euclidienne
4. variables qualitatives : critères de similarité usuels
5. distance & données structurées

Distances & (dis)similarités

Une fonction de similarité s définie sur un espace $\mathcal{X}$:

- est une fonction de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$, et souvent dans $\mathbb{R}^+$
- qui quantifie la proximité entre couples d'instances

Distances & (dis)similarités

Une fonction de similarité s définie sur un espace $\mathcal{X}$:

- est une fonction de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$, et souvent dans $\mathbb{R}^+$
- qui quantifie la proximité entre couples d'instances

Propriétés naturelles d'un critère de similarité s :

- $s(x, y) = s(y, x)$: `symétrie`
- $s(x, x) = 1$ (ou plus généralement $s(x, x) = S > 0$)
- $s(x, y) \leq s(x, x)$

$\Rightarrow$ $s(x, y)$ grand si x et y proches/similaires

Distances & (dis)similarités

A partir d'un critère de similarité s on peut définir un critère de dissimilarité d :

$$d(x, y) = 1 - s(x, y)$$

(ou plus généralement $d(x, y) = S - s(x, y)$ si $s(x, x) = S$)

Distances & (dis)similarités

A partir d'un critère de similarité s on peut définir un critère de dissimilarité d :

$$d(x, y) = 1 - s(x, y)$$

(ou plus généralement $d(x, y) = S - s(x, y)$ si $s(x, x) = S$)

Conséquences :

- $d(x, x) = 0$
- $d(x, y) \geq 0$
- (+ reste symétrique)

$\Rightarrow$ $d(x, y)$ grand si x et y distants/différents

Distances & (dis)similarités

A partir d'un critère de similarité s on peut définir un critère de dissimilarité d :

$$d(x, y) = 1 - s(x, y)$$

(ou plus généralement $d(x, y) = S - s(x, y)$ si $s(x, x) = S$)

Conséquences :

- $d(x, x) = 0$
- $d(x, y) \geq 0$
- (+ reste symétrique)

$\Rightarrow d(x, y)$ grand si x et y distants/différents

Néanmoins, ce n'est pas suffisant pour parler de **distance**.

Distances & (dis)similarités

Une distance d est une fonction de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}^+$

- (NB : $d(x, y) \geq 0$)

qui vérifie les propriétés suivantes :

- symétrie : $d(x, y) = d(y, x)$
- séparation : $d(x, y) = 0 \Leftrightarrow x = y$
- inégalité triangulaire : $d(x, z) \leq d(x, y) + d(y, z)$

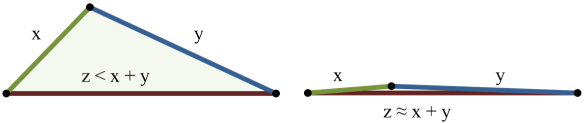


Figure: Image adaptée de Wikipedia.

Distances et (dis)similarités

1. distance vs (dis)similarité
2. variables quantitatives : distances & espaces vectoriels
3. la distance Euclidienne
4. variables qualitatives : critères de similarité usuels
5. distance & données structurées

Distances et espaces vectoriels

Distances usuelles quand $\mathcal{X} = \mathbb{R}^p$:

- distance Euclidienne : $d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

Distances et espaces vectoriels

Distances usuelles quand $\mathcal{X} = \mathbb{R}^p$:

- distance Euclidienne : $d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$
- distance de Manhattan : $d_1(x, y) = \sum_{i=1}^{p} |x_i - y_i|$

Distances et espaces vectoriels

Distances usuelles quand $\mathcal{X} = \mathbb{R}^p$:

- distance Euclidienne : $d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$
- distance de Manhattan : $d_1(x, y) = \sum_{i=1}^{p} |x_i - y_i|$
- distance de Chebyshev : $d_\infty(x, y) = \max_{i=1,\ldots,p} |x_i - y_i|$

Distances et espaces vectoriels

Distances usuelles quand $\mathcal{X} = \mathbb{R}^p$:

- distance Euclidienne : $d_2(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$
- distance de Manhattan : $d_1(x,y) = \sum_{i=1}^{p} |x_i - y_i|$
- distance de Chebyshev : $d_\infty(x,y) = \max_{i=1,\dots,p} |x_i - y_i|$

$\Rightarrow$ des cas particuliers de la distance de Minkowski :

$$d_q(x,y) = \left(\sum_{i=1}^{p} |x_i - y_i|^q \right)^{1/q}$$

(mais en pratique on prend essentiellement $q \in \{1, 2, \infty\}$)

Distances et espaces vectoriels

Distance de Manhattan ?

- ou distance de city-block, de taxi.

Distances et espaces vectoriels

Distance de Manhattan ?

- ou distance de city-block, de taxi.

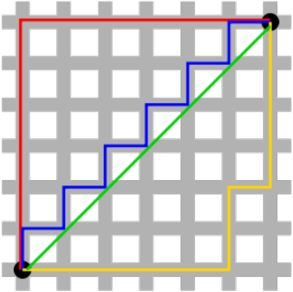


Figure: Image tirée de Wikipedia

Quizz : distance Euclidienne vs Manhattan sur cet exemple ?

Distances et (dis)similarités

1. distance vs (dis)similarité
2. variables quantitatives : distances & espaces vectoriels
3. la distance Euclidienne
4. variables qualitatives : critères de similarité usuels
5. distance & données structurées

La distance Euclidienne

Définition : $d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

La distance Euclidienne

Définition : $d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

Relation avec le produit scalaire :

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{p} x_i y_i$$

$$\langle x, x \rangle = x^T x = \sum_{i=1}^{p} x_i^2 = ||x||_2^2$$

La distance Euclidienne

Définition : $d_2(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$

Relation avec le produit scalaire :

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{p} x_i y_i$$

$$\langle x, x \rangle = x^T x = \sum_{i=1}^{p} x_i^2 = ||x||_2^2$$

On a donc :

$$\begin{aligned} d_2(x,y) &= \sqrt{\langle x - y, x - y \rangle} \\ &= ||x - y||_2. \end{aligned}$$

Distance Euclidienne & produit scalaire

On a donc :

$$d_2^{2}(x,y) = ||x-y||_2^2$$
$$= \langle x-y, x-y\rangle$$

Distance Euclidienne & produit scalaire

On a donc :

$$\begin{aligned} d_2^2(x,y) &= ||x-y||_2^2 \\ &= \langle x-y, x-y \rangle \end{aligned}$$

Par distributivité du produit scalaire, on peut écrire :

$$\begin{aligned} d_2^2(x,y) &= \langle x,x \rangle + \langle y,y \rangle - 2\langle x,y \rangle \\ &= ||x||^2 + ||y||^2 - 2\langle x,y \rangle \end{aligned}$$

Distance Euclidienne & produit scalaire

On a donc :

$$\begin{aligned} d_2^2(x, y) &= ||x - y||_2^2 \\ &= \langle x - y, x - y \rangle \end{aligned}$$

Par distributivité du produit scalaire, on peut écrire :

$$\begin{aligned} d_2^2(x, y) &= \langle x, x \rangle + \langle y, y \rangle - 2\langle x, y \rangle \\ &= ||x||^2 + ||y||^2 - 2\langle x, y \rangle \end{aligned}$$

Si nos observations sont normées, alors $||x|| = 1$ et :

$$d_2^2(x, y) = 2(1 - \langle x, y \rangle)$$

$\Rightarrow$ produit scalaire = similarité liée à la distance Euclidienne.

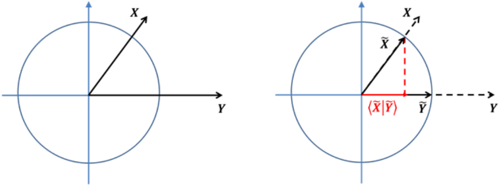
Distance Euclidienne & produit scalaire

Remarque : fonction de similarité du cosinus =

$$\cos(x, y) = \frac{\langle x, y \rangle}{||x|| \, ||y||}$$

⇒ correspond au produit scalaire entre vecteurs normalisés :

$$\cos(x, y) = \langle \tilde{x}, \tilde{y} \rangle, \quad \text{avec} \;\; \tilde{x} = \frac{x}{||x||}.$$



Distance Euclidienne généralisée

Si les variables ont des variances différentes, il peut être judicieux de le prendre en compte par :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{s_i^2}},$$

où s_i est l'écart type (empirique) de la $i$ème variable.

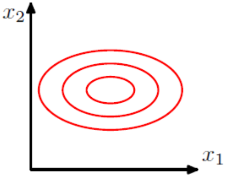
Distance Euclidienne généralisée

Si les variables ont des variances différentes, il peut être judicieux de le prendre en compte par :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{s_i^2}},$$

où s_i est l'écart type (empirique) de la $i$ème variable.

Rappel :



- une différence sur l'axe 1 comptera moins qu'une différence sur l'axe 2.

Distance Euclidienne généralisée

Distance Euclidienne normalisée :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{s_i^2}},$$

où s_i est l'écart type empirique de la $i$ème variable.

Ecriture matricielle :

$$d(x, y) = \sqrt{(x - u)^T M(x - y)},$$

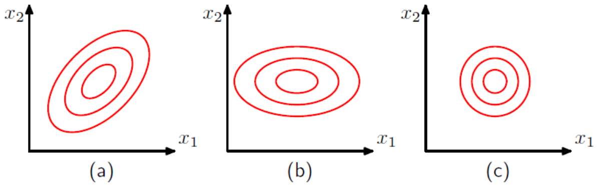
avec $M = \text{diag}(1/s_1^2, ..., 1/s_p^2)$.

Distance Euclidienne généralisée

Plus généralement, la distance de Mahalanobis :

$$d(x, y) = \sqrt{(x - y)^T M(x - y)},$$

où $M = S^{-1}$ est la matrice de covariance empirique, permet de prendre en compte les corrélations entre variables.



Distances et (dis)similarités

1. distance vs (dis)similarité
2. variables quantitatives : distances & espaces vectoriels
3. la distance Euclidienne
4. variables qualitatives : critères de similarité usuels
5. distance & données structurées

Variables qualitatives

Variable qualitative :

- prend des valeurs 0 ou 1
- code pour la présence / l'absence d'une caractéristique

Variables qualitatives

Variable qualitative :

- prend des valeurs 0 ou 1
- code pour la présence / l'absence d'une caractéristique

Pourquoi un traitement particulier ?

Variables qualitatives

Variable qualitative :

- prend des valeurs 0 ou 1
- code pour la présence / l'absence d'une caractéristique

Pourquoi un traitement particulier ?

Exemple du produit scalaire :

$$\langle x, y \rangle = \{\#\text{ de variables présentes en même temps}\}$$

Variables qualitatives

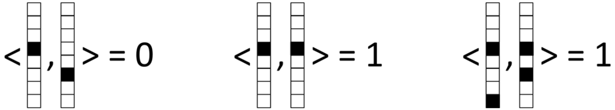
Variable qualitative :

- prend des valeurs 0 ou 1
- code pour la présence / l'absence d'une caractéristique

Pourquoi un traitement particulier ?

Exemple du produit scalaire :

$$\langle x, y \rangle = \left\{ \# \text{ de variables présentes en même temps} \right\}$$



$\Rightarrow$ quid des variables absentes en même temps ?
$\Rightarrow$ $s(x, y) = s(x, x)$ pour des x et y très différents

Variables qualitatives

Beaucoup de critères de similarité pour données qualitatives.

Motivations = prendre en compte :

- la présence conjointe de variables dans x et y
- l'absence conjointe de variables dans x et y
- le nombre de variables présentes dans x ou y

Variables qualitatives

Beaucoup de critères de similarité pour données qualitatives.

Motivations = prendre en compte :

- la présence conjointe de variables dans x et y
- l'absence conjointe de variables dans x et y
- le nombre de variables présentes dans x ou y

On définit les quantités suivantes :

	$y = 1$	$y = 0$
$x = 1$	M_{11}	M_{10}
$x = 0$	M_{01}	M_{00}

(NB : $M_{11} + M_{10} + M_{01} + M_{00} = p$)

Critères de similarité usuels

Produit scalaire :

$$s(x,y) = M_{11}$$

Simple Matching coefficient :

$$s(x,y) = \frac{M_{11}+M_{00}}{M_{11}+M_{10}+M_{01}+M_{00}} = \frac{M_{11}+M_{00}}{p}$$

Coefficient de Jaccard :

$$s(x,y) = \frac{M_{11}}{M_{11}+M_{01}+M_{10}}$$

Coefficient de Dice :

$$s(x,y) = \frac{2M_{11}}{2M_{11}+M_{01}+M_{10}}$$

Critères de similarité usuels - interprétation

Produit scalaire :

$$s(x, y) = M_{11}$$

$\Rightarrow$ nombre de variables présentes en même temps

Critères de similarité usuels - interprétation

Produit scalaire :

$$s(x, y) = M_{11}$$

$\Rightarrow$ nombre de variables présentes en même temps

Simple Matching coefficient :

$$s(x, y) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}} = \frac{M_{11} + M_{00}}{p}$$

$\Rightarrow$ proportion de variables identiques en même temps

- présentes OU absentes

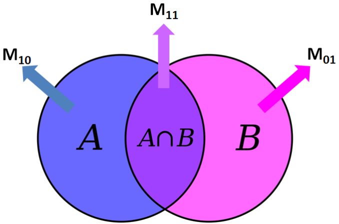
Critères de similarité usuels - interprétation

Coefficient de Jaccard :

$$s(x,y) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} = \frac{|X \cap Y|}{|X \cup Y|}$$

Coefficient de Dice :

$$s(x,y) = \frac{2M_{11}}{2M_{11} + M_{01} + M_{10}} = \frac{2|X \cap Y|}{|X| + |Y|}$$



Critères de similarité usuels - implémentation

Produit scalaire :

$$s(x, y) = \langle x, y \rangle$$

Critères de similarité usuels - implémentation

Produit scalaire :

$$s(x, y) = \langle x, y \rangle$$

Simple Matching coefficient :

$$s(x, y) = \frac{\langle x, y \rangle + \langle \neg x, \neg y \rangle}{p}$$

Critères de similarité usuels - implémentation

Produit scalaire :

$$s(x,y) = \langle x, y \rangle$$

Simple Matching coefficient :

$$s(x,y) = \frac{\langle x, y \rangle + \langle \neg x, \neg y \rangle}{p}$$

Coefficient de Jaccard :

$$s(x,y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\langle x, y \rangle}{\langle x, x \rangle + \langle y, y \rangle - \langle x, y \rangle}$$

Critères de similarité usuels - implémentation

Produit scalaire :

$$s(x,y) = \langle x, y \rangle$$

Simple Matching coefficient :

$$s(x,y) = \frac{\langle x, y \rangle + \langle \neg x, \neg y \rangle}{p}$$

Coefficient de Jaccard :

$$s(x,y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{\langle x, y \rangle}{\langle x, x \rangle + \langle y, y \rangle - \langle x, y \rangle}$$

Coefficient de Dice :

$$s(x,y) = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2\langle x, y \rangle}{\langle x, x \rangle + \langle y, y \rangle}$$

Distances et (dis)similarités

1. distance vs (dis)similarité
2. variables quantitatives : distances & espaces vectoriels
3. la distance Euclidienne
4. variables qualitatives : critères de similarité usuels
5. distance & données structurées

Distances & données structurées

Données structurées :

- pas de représentation vectorielle naturelle
- présentes dans bon nombre d'applications réelles

Distances & données structurées

Données structurées :

- pas de représentation vectorielle naturelle
- présentes dans bon nombre d'applications réelles

2 approches :

1. se ramener à une représentation vectorielle
   - "feature extraction" - quantitatif vs qualitatif (0/1)
2. définir des mesures de similarité entre objets

Distances & données structurées

Données structurées :

- pas de représentation vectorielle naturelle
- présentes dans bon nombre d'applications réelles

2 approches :

1. se ramener à une représentation vectorielle
   - "feature extraction" - quantitatif vs qualitatif (0/1)
2. définir des mesures de similarité entre objets

Illustrations :

1. distance d'édition & données de séquences
2. "dynamic time warping" & séries temporelles

Distance d'édition pour chaînes de caractères

Distance d'édition = similarité entre chaînes de caractères

3 opérations élémentaires :

1. insertion : $ac \rightarrow abc$ $(\epsilon \rightarrow b)$
2. délétion : $abc \rightarrow ac$ $(b \rightarrow \epsilon)$
3. substitution : $abd \rightarrow abc$ $(d \rightarrow c)$

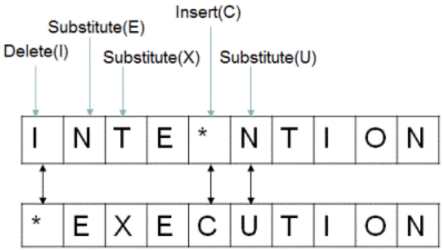
$\Rightarrow$ chaque opération = un coût

$\Rightarrow$ $d(x, y)$ = transformation de coût minimal

Très utilisée en bioinformatique et traitement du langage.

Distance d'édition pour chaines de caractères

Illustration :

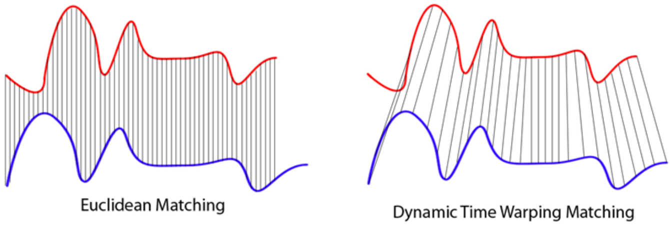


- aussi appelée distance de Levenshtein
- une vraie distance
  - sous certaines conditions (assez générales) sur les coûts.
- calcul par algorithmes de programmation dynamique

"Dynamic Time Warping" pour séries temporelles

DTW = similarité entre séries temporelles

- univariées et + généralement "séquences" (e.g., vidéos)



- appariement optimal par "déformation du temps"
- calcul par algorithmes de programmation dynamique
- pas une vraie distance
- applications : reconnaissance de la parole, vidéo

Conclusion

Conclusion

Clustering hiérarchique :

- algorithme de base de clustering
- hiérarchie de partitions : dendrogramme

Heatmaps :

- clustering hiérarchique pour la visualisation
- identification de structures de bloc

Distance/similarités :

- choix propre à l'application
- rarement anodin

TP : matrices de distance, clustering hiérarchique, heatmaps.

Exercice

Exercice

Soit la matrice de distance suivante :

$$D = \begin{bmatrix} 0 & 0.5 & 3 & 1.3 & 4 \\ 0.5 & 0 & 1 & 3.5 & 5 \\ 3 & 1 & 0 & 1.5 & 2.2 \\ 1.3 & 3.5 & 1.5 & 0 & 6 \\ 4 & 5 & 2.2 & 6 & 0 \end{bmatrix}$$

Calculer (à la main) le dendrogramme par la méthode d'agrégation de votre choix.

Exercice - solution (1/2)

	1	2	3	4	5
1	0	0.5	3	1.3	4
	2	0	1	3.5	5
		3	0	1.5	2.2
			4	0	6
				5	0

step 1 : merge 1 & 2

	1	2	3	4	5
1	0	0.5	3	1.3	4
	2	0	1	3.5	5
		3	0	1.5	2.2
			4	0	6
				5	0

SINGLE

step 2 : merge 1/2 & 3

	1\2	3	4	5
1\2	0	1	1.3	4
	3	0	1.5	2.2
		4	0	6
			5	0

step 3 : merge 1/2/3 & 4

	1\2\3	4	5
1\2\3	0	1.3	2.2
	4	0	6
		5	0

step 3 : merge 1/2/3/4 & 5

	1\2\3\4	5
1\2\3\4	0	2.2
	5	0

COMPLETE

step 2 : merge 3 & 4

	1\2	3	4	5
1\2	0	3	3.5	5
	3	0	1.5	2.2
		4	0	6
			5	0

step 3 : merge 1/2 & 3/4

	1\2	3\4	5
1\2	0	3.5	5
	3\4	0	6
		5	0

step 3 : merge 1/2/3/4 & 5

	1\2\3\4	5
1\2\3\4	0	6
	5	0

AVERAGE

step 2 : merge 3 & 4

	1\2	3	4	5
1\2	0	2	2.4	4.5
	3	0	1.5	2.2
		4	0	6
			5	0

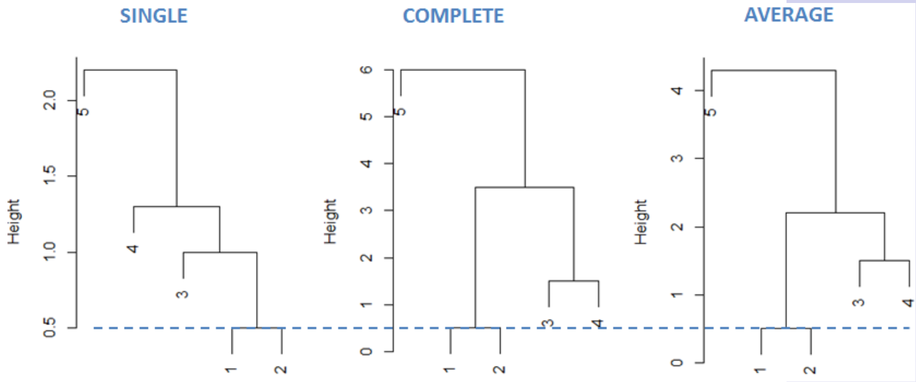
step 3 : merge 1/2 & 3/4

	1\2	3\4	5
1\2	0	2.2	4.5
	3\4	0	4.1
		5	0

step 3 : merge 1/2/3/4 & 5

	1\2\3\4	5
1\2\3\4	0	4.3
	5	0

Exercice - solution (2/2)



Références

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.