MA431 : Mathématiques appliquées à la sécurité

Clustering

D. Barcelo

Grenoble INP ESISAR

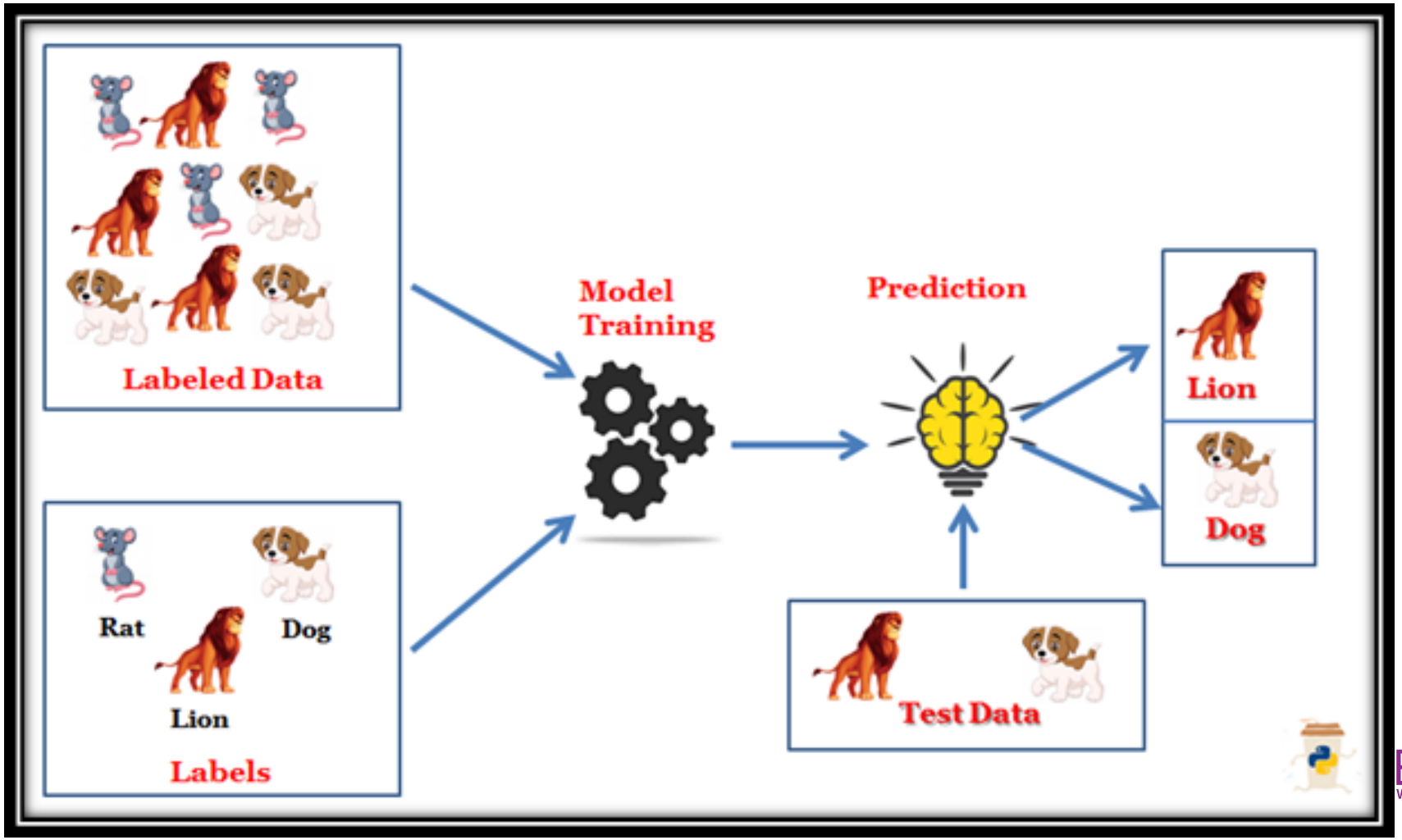
2022/2023

Classification (fr) ou Classification (eng) ?

Classrification

La classification permet d'affecter des individus à des classes existantes à priori. Il s'agit d'un apprentissage supervisé. Le terme francophone pour désigner cette technique est classement.

Apprentissage supervisé



Mathématiquement

Classification

On observe p variables sur n individus.
On note X_i le vecteur des observations des variables sur l'individu i.
Chaque individu i est affecté à une classe Y_i.
Il y a k classes possibles $(1 \leqslant k \leqslant n)$.

- variable Y_i connue,
- objectif : à partir des observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ construire une régle de classement r.

$$r : X \mapsto r(X) = Y$$

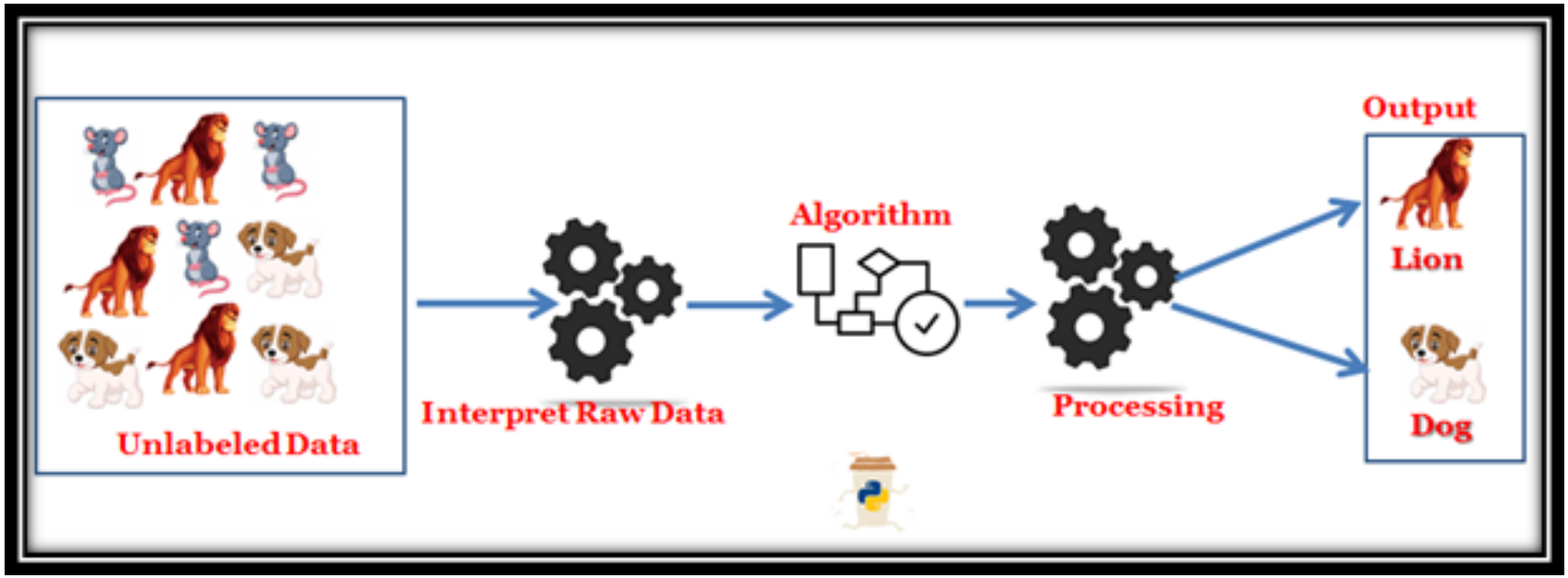
- r permet d'affecter une classe à de nouveaux individus de classe inconnue.

Classification (fr) ou Classification (eng) ?

Clustering

La classification automatique permet de regrouper des individus dans des classes (clusters) non définies à priori. Il s'agit d'un apprentissage automatique non supervisé. Le terme anglophone pour désigner cette technique est cluster analysis ou clustering (en français, on parle de classification).

Apprentissage non supervisé



Mathématiquement

Clustering

On observe p variables sur n individus.
On note X_i le vecteur des observations des variables sur l'individu i.
Chaque individu i doit être affecté à une classe Y_i.
Il y a au maximum n classes possibles.

- variable Y_i inconnue,
- objectif : à partir des observations $X_1, \ldots, X_n$ construire $Y_1, \ldots Y_n$.
- Les classes ainsi construites pourront être interprétées.

Clustering : objectifs

- Objectif : produire des groupements d'individus ou de variables afin de donner du sens à des jeux de données.
- Identifier des groupes ayant des caractères similaires.
- Techniques d'apprentissage non supervisées.
- De nombreuses techniques existent.
- Méthodes présentées ici :
  1. une technique hiérarchique :
     la **Classification Ascendante Hiérarchique** (**CAH**),
  2. une technique de centres mobiles :
     Méthodes des k-**means** (ou k-moyennes in french),
  3. une technique basée sur la densité : **DBSCAN**.

Clustering et combinatoire

Pourquoi avoir recours à des algorithmes pour partitionner un ensemble ?

Nombre de partitions

Soit $n \in \mathbb{N}$, $n > 0$.
Le nombre de partitions d'un ensemble de cardinal n est donné par le n-ième nombre de Bell. Par récurrence, on a : $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_{n-k}$.

On peut montrer que : $B_n = \frac{1}{e} \sum_{k \geqslant 0} \frac{k^n}{n!}$.

Le nombre de partitions possibles devient rapidement très grand.
Recherche d'algorithmes convergeant assez rapidement vers une solution (sans en explorer tous les cas).
Par exemple, on a : $B_{12} = 4213597$, $B_{13} = 27644437$ et $B_{30} \approx 8,47.10^{23}$.

Différents types de clustering

- **Clustering intrinsèque** : classes inconnues, analyse des données.
- **Clustering extrinsèque** : individus reliés à des catégories (étude préalable des données)
- **Clustering par agglomération** : un individu = une classe puis fusion progressive des classes (approche "bottom up", du bas vers le haut).
- **Clustering par division** : une seule classe initiale puis séparation progressive en plusieurs classes (approche "top down", du haut vers le bas).
- **Clustering hiérarchique** : classes "parents" et classes "enfants". Représentation graphique sous forme de **dendrogramme**.
- **Clustering non hiérarchique** : individus répartis dans des classes qui forment une partition de l'ensemble des individus, sans relation hiérarchique entre les classes.

Exemples étudiés

Nous étudierons trois techniques de clustering :

- **Classification Ascendante hiérarchique** : classification intrinsèque hiérarchique par agglomération
- *k*-**means** : classification intrinsèque non hiérarchique
- **DBSCAN** : classification intrinsèque non hiérarchique

Clustering et distances

- Techniques basées sur des études de proximité,
- Nécessité de définir une distance ou une mesure.
- Pour la suite du cours : distance d dont le choix ne sera pas étudié.

Clustering et distances

Par exemple : Soit $\vec{x} \in \mathbb{R}^n$ et $\vec{y} \in \mathbb{R}^n$.

- **Distance euclidienne :** $d(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$.
- **Distance Manhattan :** $d(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|$.
- **Distance de Mahalanobis :**

  $$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - y_i) Cov(\alpha_i, \alpha_j)(x_j - y_j)}$$

  ou matriciellement $d(X; Y) = \sqrt{(X - Y)^t V (X - Y)}$ avec V matrice de variance-covariance des variables observées.
- **Distance du χ^2** pour des variables nominales
- **Distance de Hamming** pour des variables binaires,
- etc. .

Critères d'un bon clustering

Critères d'un bon clustering :

- détection des structures présentes dans les données,
- obtenir le nombre optimal de clusters,
- fournir des clusters bien différenciés,
- fournir des clusters stables en cas de légères modification de données,
- pouvoir traiter efficacement de grands volumes de données.

Clustering : principes

- Les individus d'une même classe (cluster) doivent se ressembler.
- Les individus de deux classes différentes doivent se démarquer.
- Il faut donc créer des classes homogènes éloignées les unes des autres.
- Ces classes doivent aider à interpréter les données.
- Méthode complémentaire des méthodes descriptives type ACP.

Clustering : principes

- On considère un nuage de points $(x_1, \ldots, x_n)$.
- On veut le partager en k classes.
- Chaque point ne doit appartenir qu'à une seule classe.
- Chaque classe possède un centre de gravité $(g_1, \ldots, g_k)$.
- Chaque x_i est associé à un centre de gravité g_j.
- x_i appartient à la classe j de centre de gravité g_j.

Clustering : questions

- Comment déterminer k ?
- Comment déterminer les centres de gravité ?
- Comment associer un individu à un centre de gravité ?
- Comment comparer deux partitions pour déterminer la meilleure ?

Clustering : réponses ?

- Comment déterminer k ? pas si évident ...
- Comment déterminer les centres de gravité ?
  Il faut choisir les k premiers.
- Comment associer un individu à un centre de gravité ?
  En l'affectant au centre de gravité le plus proche.
- Comment comparer deux partitions pour déterminer la meilleure ?
  En calculant les inerties de chaque classe et en les comparant.

Choix de k

- spécifique à certaines méthodes (inutile avec DBSCAN),
- détermination visuelle à l'aide d'un graphique (CAH),
- faire varier k et chercher une valeur sur un graphique ($k$means),
- utiliser un indice type Calinski-Harabasz ou Davies-Bouldin ($k$means).

Centre de gravité

Centroïde :

Soit $j \in [\![1; k]\!]$. On note C_j le $j$ième cluster d'effectif n_j.
On appelle Centroïde du cluster C_j l'isobarycentre (ou centre de gravité) du cluster :

$$\vec{g_j} = \frac{1}{n_j} \sum_{\vec{x} \in C_j} \vec{x}$$

On appelle medoïde le point du cluster le plus proche du centroïde.

Homogénéité

Homogénéité

Soit $j \in [\![1; k]\!]$. On note C_j le $j$ième cluster d'effectif n_j.
On appelle homogénéité du cluster C_j la distance moyenne des individus du cluster à son centroïde.

$$H_j = \frac{1}{n_j} \sum_{x_i \in C_j} d(x_i; g_j)$$

On appelle homogénéité globale : $H = \frac{1}{k} \sum_{j=1}^{k} H_j$.

Un cluster est homogène si son homogénéité est faible.
Plus l'homogénéité globale est faible, plus les clusters sont homogènes.

Inertie

Inertie intraclasse

Soit $j \in [\![1; k]\!]$. On note C_j le $j$ième cluster d'effectif n_j.
On définit l'inertie de la classe C_j par : $I_j = \sum_{x_i \in C_j} d^2(x_i; g_j)$.

On définit l'inertie intraclasse par : $I_{intra} = \sum_{j=1}^{k} I_j$.

Un cluster est homogène si son inertie est faible.
Plus l'inertie intraclasse est faible, plus les clusters sont homogènes.

Séparabilité

Séparabilité

Soit $j \in [\![1; k]\!]$ et soit $l \in [\![1; k]\!]$. On note C_j le $j$ième cluster et C_l le $l$ième cluster .
On appelle séparabilité des clusters C_j et C_l la distance entre leurs centroïdes.

$$S_{j,l} = d(g_j; g_l)$$

On appelle séparabilité globale : $S = \frac{2}{k(k-1)} \sum_{j=1}^{k} \sum_{l=j+1}^{k} S_{j,l}$.

Plus la séparabilité globale est grande, plus les clusters sont séparés.

Inertie

Inertie interclasse

Soit $j \in [\![1; k]\!]$. On note C_j le $j$ième cluster d'effectif n_j et g l'isobarycentre des individus.

On définit l'inertie interclasse par : $I_{inter} = \frac{1}{n} \sum_{j=1}^{k} n_i d^2(g_i; g)$.

Plus l'inertie interclasse est grande, plus les clusters sont séparés.
Attention, l'inertie interclasse augmente avec le nombre de clusters.

Inertie et qualité

Inertie totale

On définit l'inertie totale par : $I_{tot} = \frac{1}{n}\sum_{j=1}^{k} d^2(x_i; g)$.

Formule de Huygens : $I_{tot} = I_{inter} + I_{intra}$.

On peut calculer la proportion de l'inertie expliquée : $R^2 = \frac{I_{inter}}{I_{tot}}$ pour évaluer la qualité du clustering mais sans chercher à la maximiser.

Inertie et qualité

Indice de Calinski-Harabasz

On définit l'indice de Calinski-Harabasz par : $S_{CH} = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$.

k représente le nombre de clusters et l'indice correspond au rapport de la variance interclasse sur la variance intraclasse.

Pour améliorer la qualité d'un clustering, on peut maximiser S_{CH} mais il est préférable de ne pas l'utiliser avec certaines méthodes.

Clustering hiérarchique

la Classification Ascendante Hiérarchique ou **CAH**

CAH

Classification Ascendante Hiérarchique

1. Objectif : obtenir une **hiérarchie** (ensemble de partitions emboitées).
2. Chaque cluster pourra se diviser en sous-clusters, jusqu'à aboutir aux individus.
3. Algorithme **ascendant** : on construit les clusters en regroupant deux à deux des éléments.
4. Ensemble des partitions possibles hiérarchisées représenté sous la forme d'un **dendrogramme**.
5. On peut obtenir entre n clusters et un unique cluster.

CAH

Classification Ascendante Hiérarchique

Pour obtenir une partition :

1. procéder à une coupure du dendrogramme.
2. Plus cette coupure sera haute, moins on aura de clusters et moins elles seront homogènes.

CAH

On considère un ensemble E de n individus caractérisés par p variables.
On suppose l'espace $\mathbb{R}^p$ muni d'une distance d appropriée.

Distance entre groupes et éléments

Règles de calcul des distances entre groupes :
On considère C_i et C_j deux clusters. On regroupe C_i et C_j si :

- deux éléments sont proches (single linkage)
  distance du saut minimal : $d(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y)$
- tous les éléments sont proches (complete linkage)
  distance du saut maximal : $d(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x, y)$
- les éléments sont proches en moyenne (average linkage)
  distance du saut moyen : $d(C_i, C_j) = \frac{1}{n_i n_k} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$
- les centroïdes sont proches (centroïd linkage)
  distance du saut centroïde : $d(C_i, C_j) = d(g_i, g_j)$

CAH

Méthode de Ward

L'inertie totale d'un nuage de point I se décompose en somme des deux inerties interclasse et intraclasse :

$$I_{tot} = I_{inter} + I_{intra}$$

Méthode de Ward : Agréger des clusters en minimisant l'inertie intraclasse et en maximisant l'inertie interclasse.

CAH

Choix de la distance

- la distance du saut minimale est sensible aux effets de chaine mais détecte bien les formes allongées,
- la distance du saut moyen est peu sensible au bruit et produit des clusters de même variance,
- la distance du saut centroïd est robuste mais moins précise,
- la méthode de Ward est la plus utilisée mais produit des clusters sphériques et de mêmes effctifs.

Algorithme

Etape 1 :

On considère un ensemble I de n individus caractérisés par p variables.

Il y a donc n éléments à classer.

La première partition, P_1, est constituée de n clusters, chacune contenant un unique individu.

Algorithme

Etape 2 :

On construit la matrice de distances entre les n éléments.

On cherche les deux éléments les plus proches et on les agrège ensemble en un nouveau cluster.

La deuxième partition, P_2, est constituée de $n-1$ clusters.

Algorithme

Etape 2 :
On construit une nouvelle matrice de distances entre les $n-1$ éléments de P_2.
On cherche les deux éléments les plus proches et on les agrège ensemble en un nouveau cluster.
La troisième partition, P_3, est constituée de $n-2$ clusters.

Algorithme

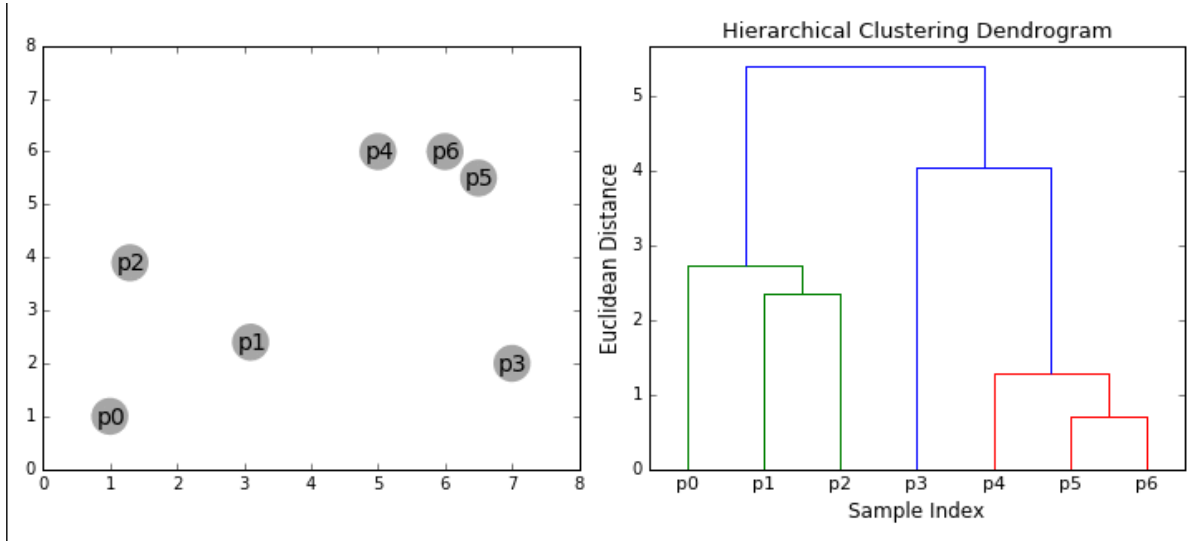
Etape q :
On construit une nouvelle matrice de distances entre les $n-(q-1)$ éléments de P_{q-1}.
On cherche les deux éléments les plus proches et on les agrège ensemble en un nouveau cluster.
La troisième partition, P_q, est constituée de $n-q$ clusters.
On réitère le processus jusqu'à n'avoir plus qu'un seul cluster regroupant tous les individus, on obtient ainsi la dernière partition.

Algorithme

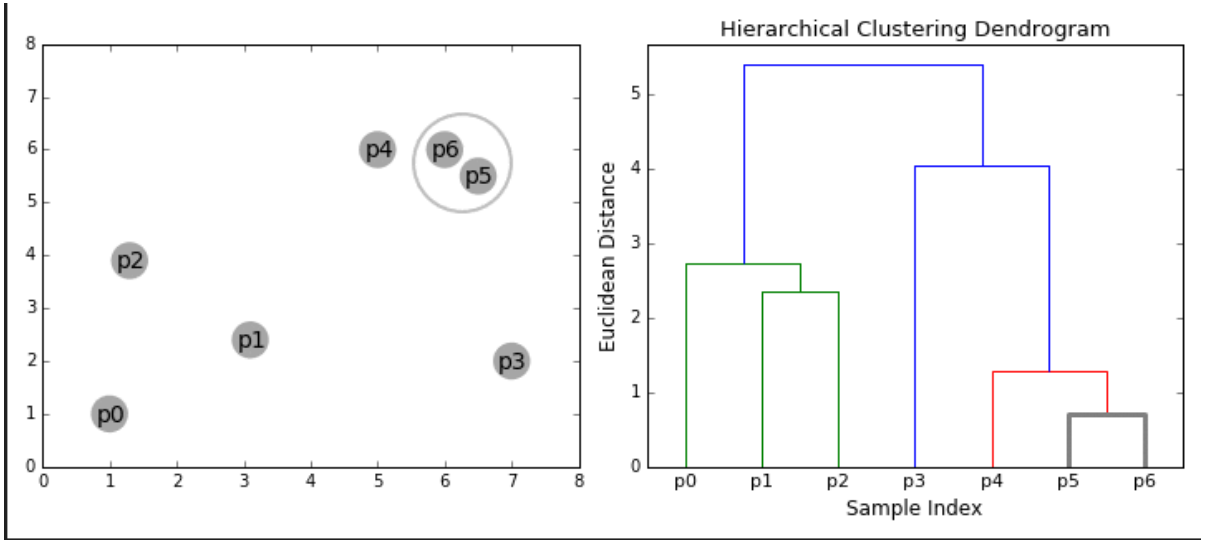
Vocabulaire :

- La famille de l'ensemble des partitions créées lors de l'éxécution de l'algorithme s'appelle une hiérarchie.
- Cette hiérarchie est représentée à l'aide d'un dendrogramme.
- Les individus sont les éléments terminaux du dendrogramme.

Dendrogramme

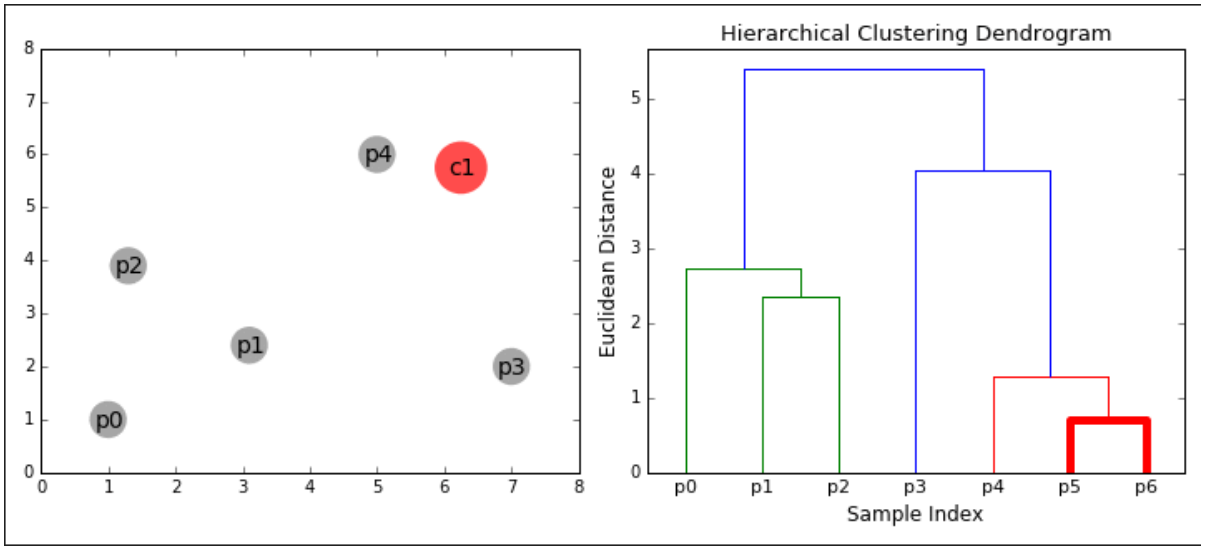


Dendrogramme



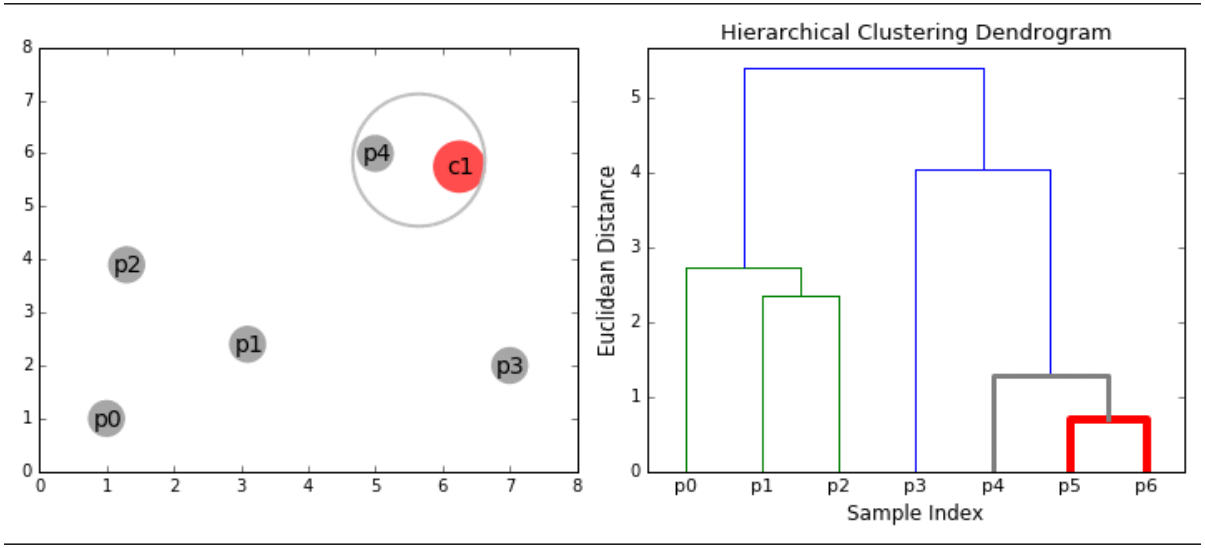
Animation : dashee87.github.io

Dendrogramme



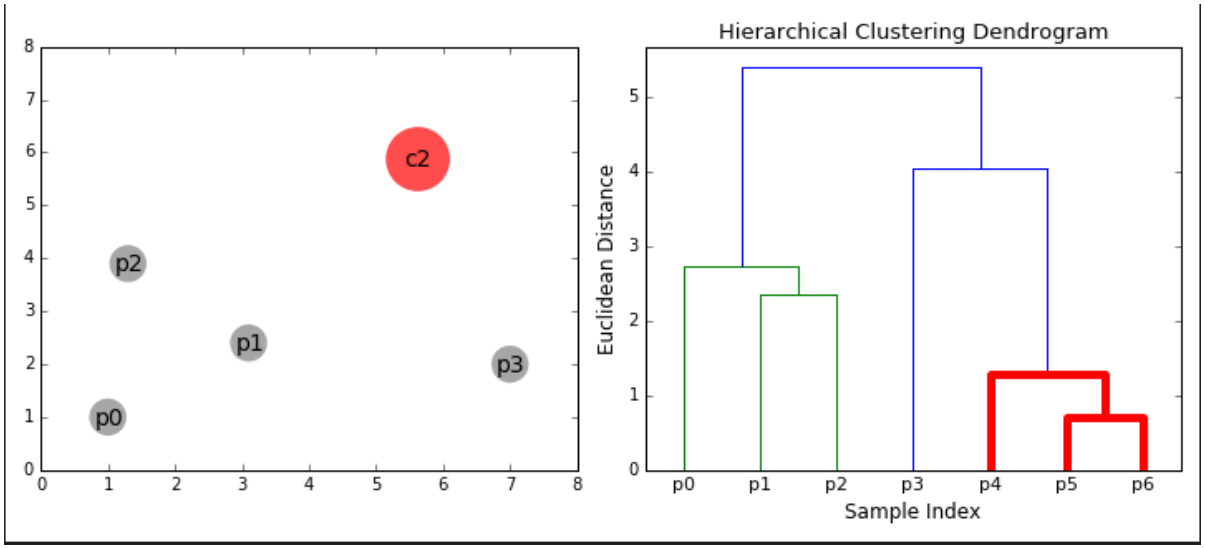
Animation : dashee87.github.io

Dendrogramme



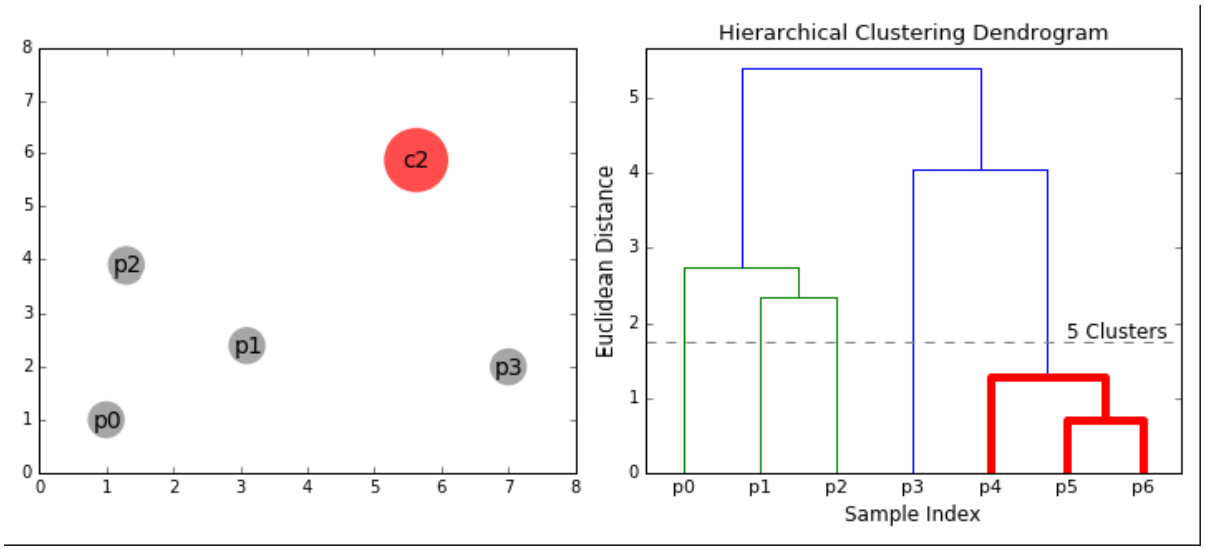
Animation : dashee87.github.io

Dendrogramme

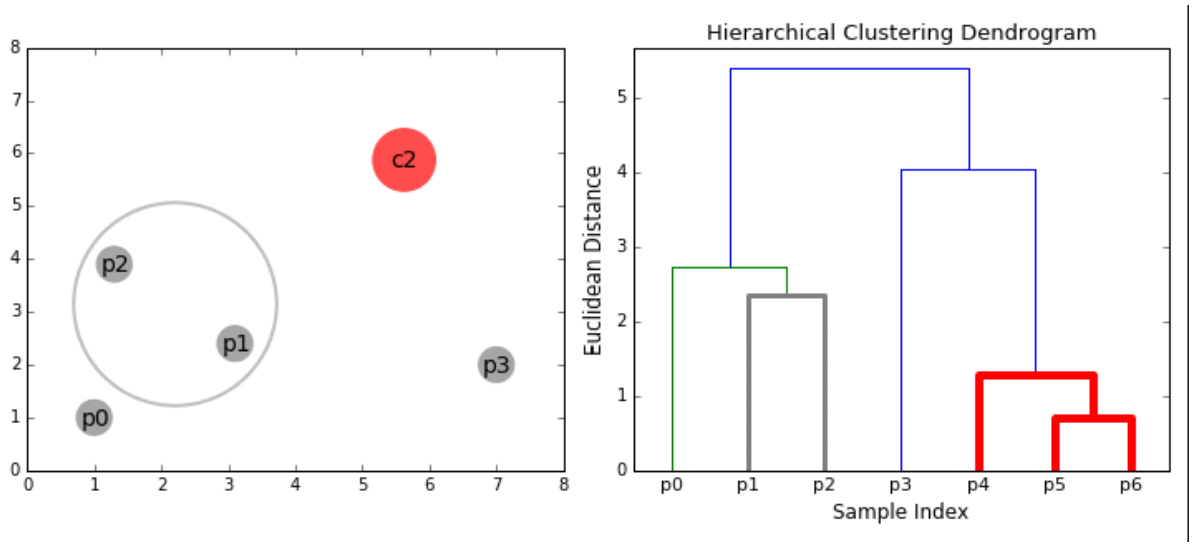


Animation : dashee87.github.io

Dendrogramme

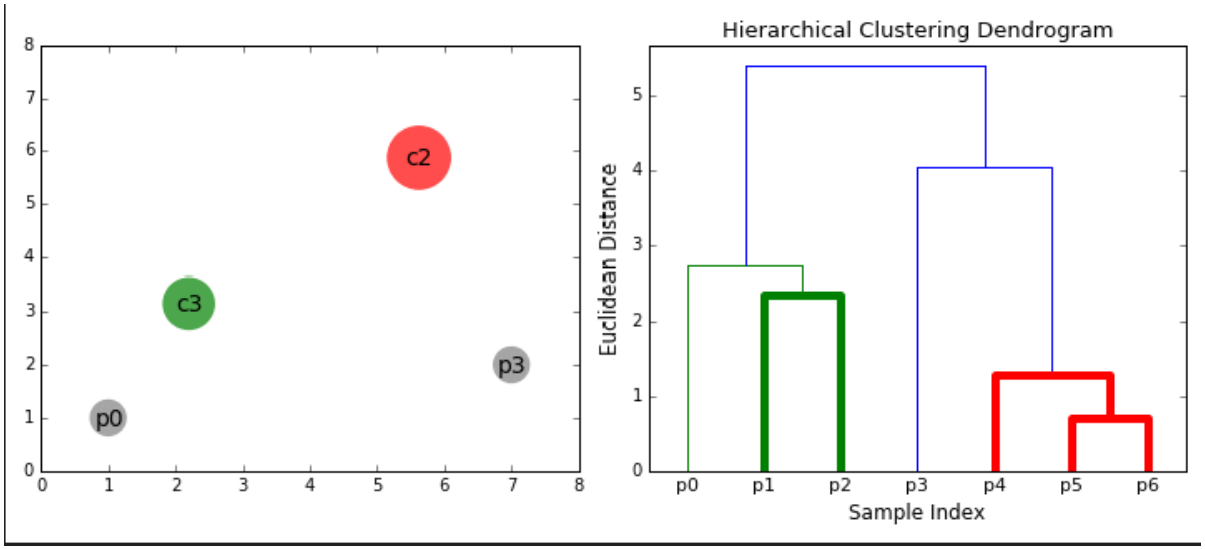


Dendrogramme



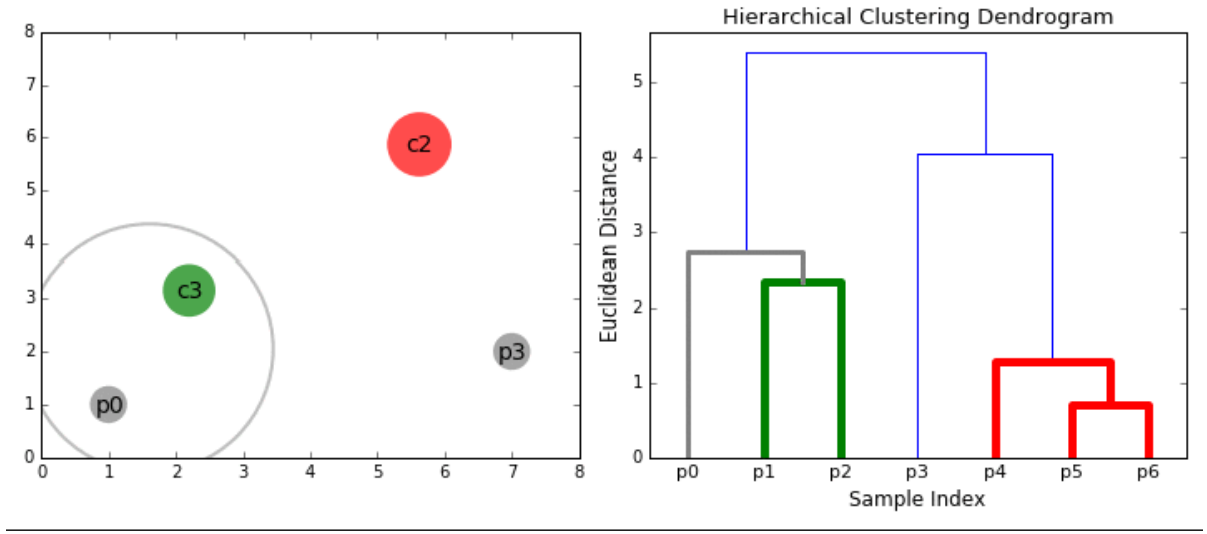
Animation : dashee87.github.io

Dendrogramme



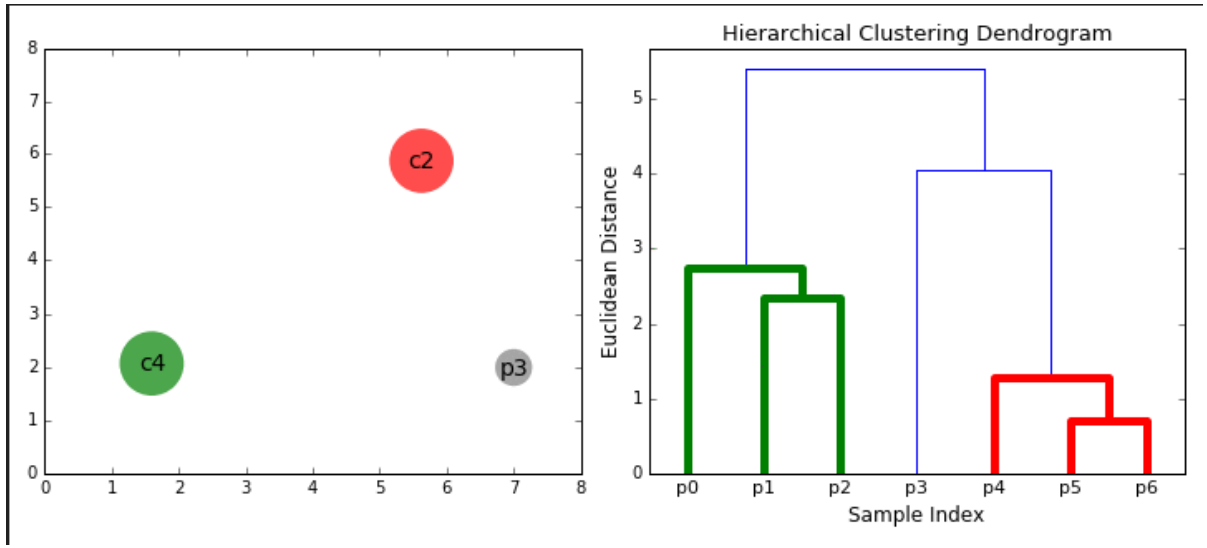
Animation : dashee87.github.io

Dendrogramme



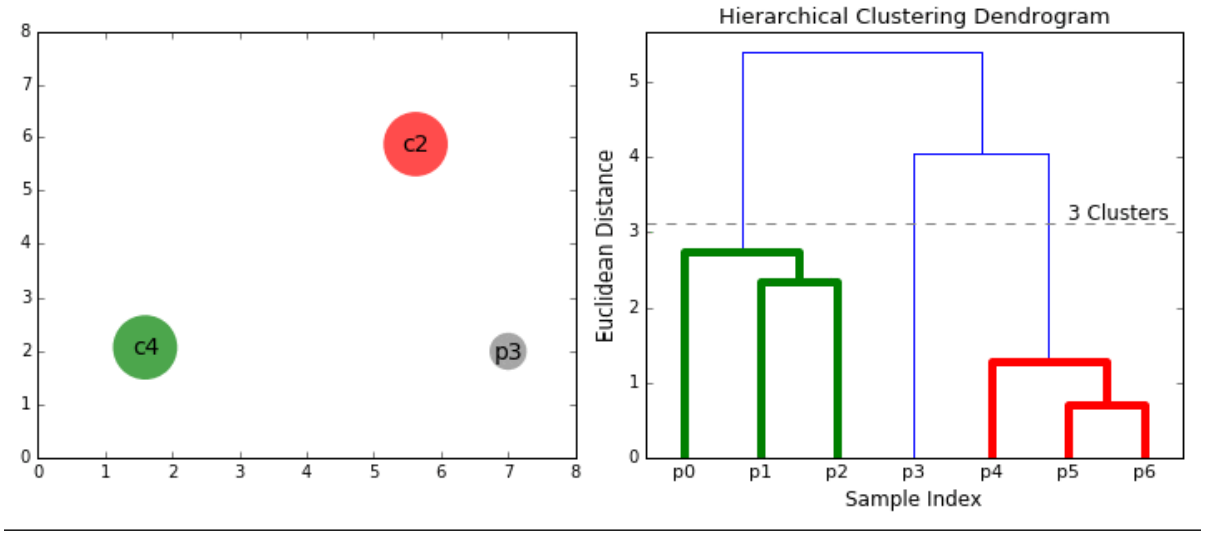
Animation : dashee87.github.io

Dendrogramme

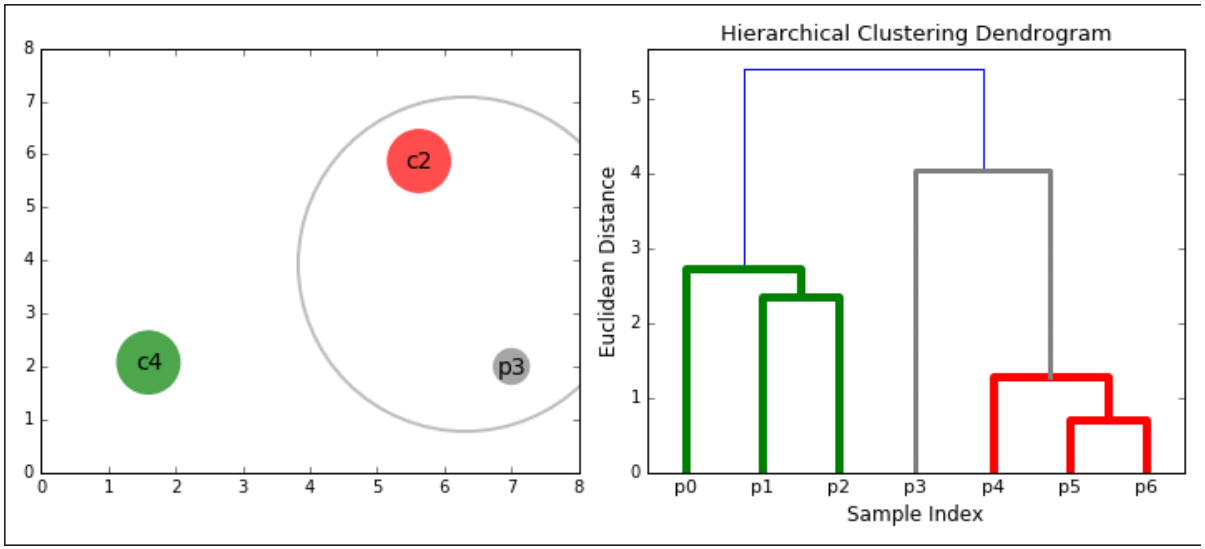


Animation : dashee87.github.io

Dendrogramme

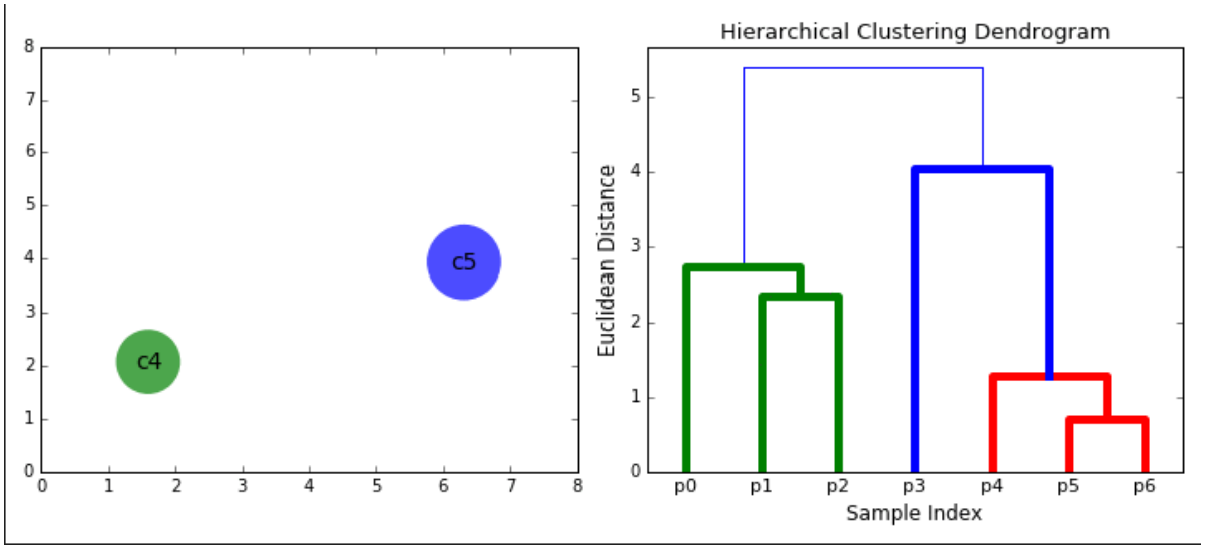


Dendrogramme



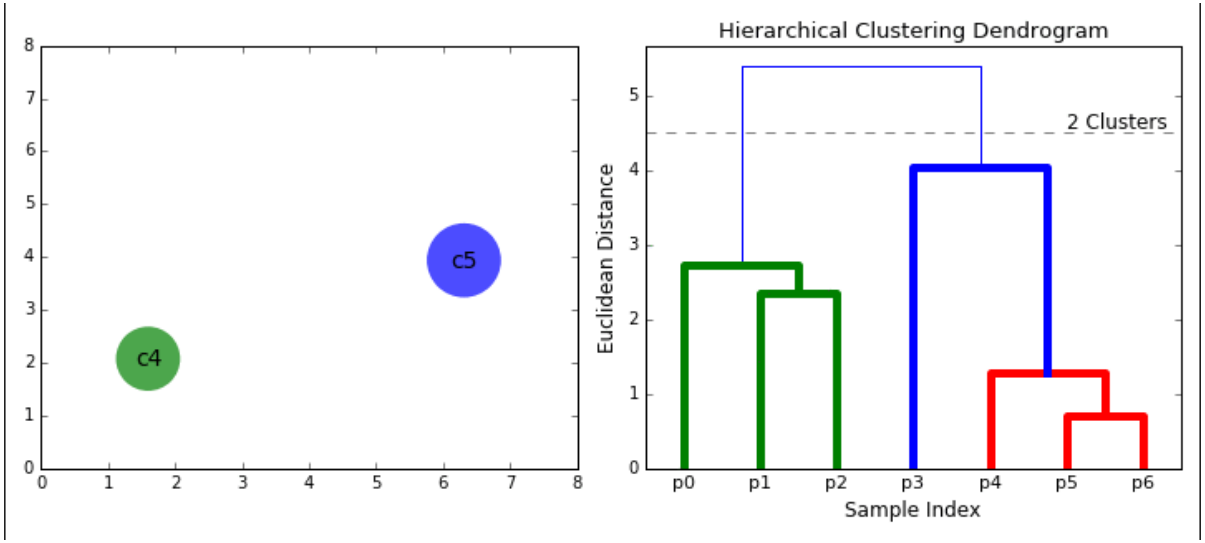
Animation : dashee87.github.io

Dendrogramme



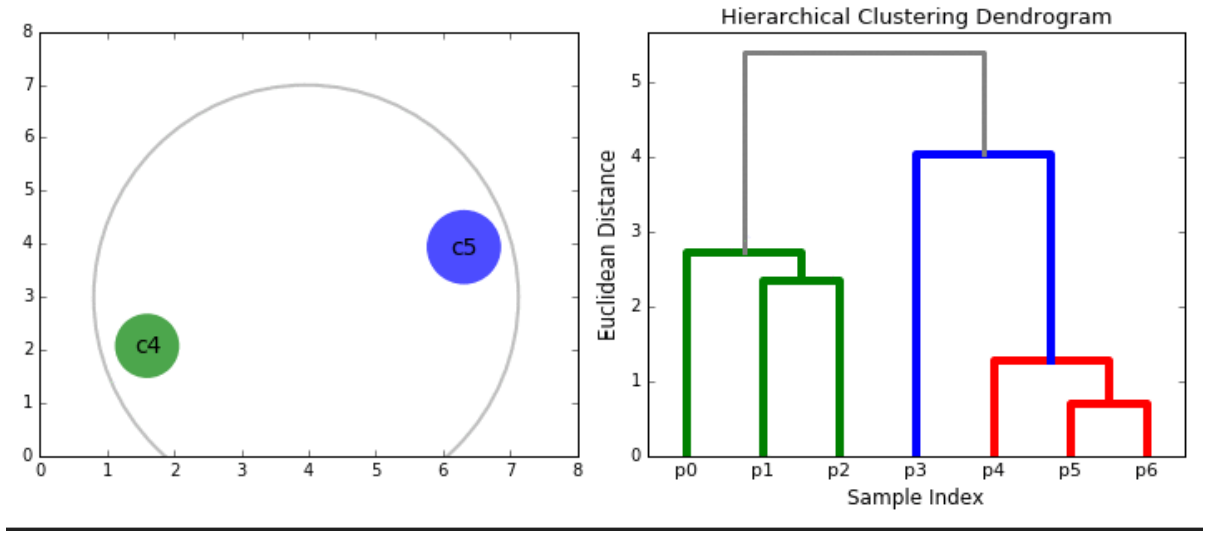
Animation : dashee87.github.io

Dendrogramme



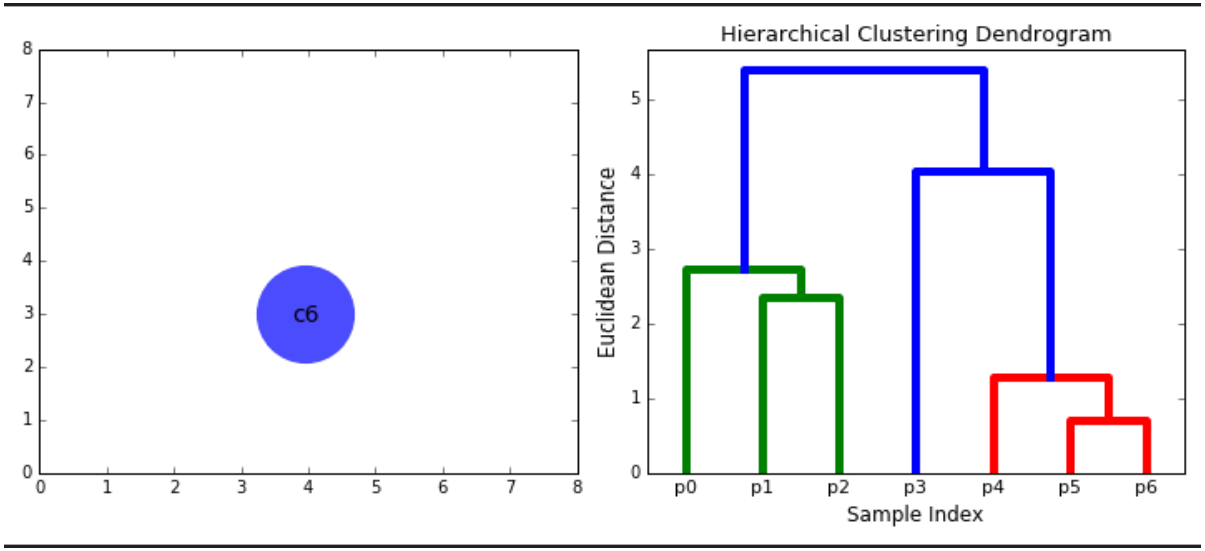
Animation : dashee87.github.io

Dendrogramme



Animation : dashee87.github.io

Dendrogramme



Animation : dashee87.github.io

Méthode

Critique :

- Avec la méthode de Ward, on agrége les classes dont l'agrégation fait perdre le moins d'inertie interclasse,
- méthode pas à pas qui tourne lentement,
- complexité algorithmique en $\mathcal{O}(pn^2)$,
- le résultat ne sera pas optimal si le nombre de pas est trop élevé,
- méthode peu robuste, si on change une distance le saut change.

Exemple sous R

Exemple sous R