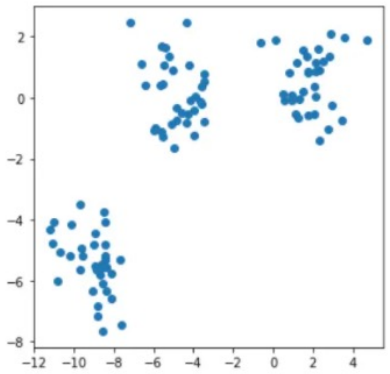
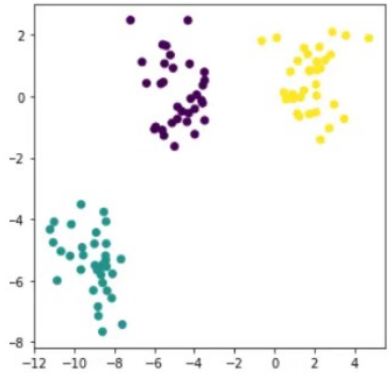
Classification Ascendante Hiérarchique

CAH

Le Principe





Sur ce **jeu de données en 2D** il apparaît clair que l'**on peut le diviser en 3 groupes**. Concrètement comment s'y prend-on ?

- L'idée de départ est de considérer que **chacun des points de votre jeu de données est un centroïde**. Cela revient à considérer qu'**à chaque point correspond une unique étiquette** (0,1,2,3, 4…).
- Ensuite **on regroupe chaque centroïde avec son centroïde voisin le plus proche**. Ce dernier prend l'étiquette du centroïde qui l'a « absorbé ».
- On **calcule** alors les **nouveaux centroïdes** qui seront les **centres de gravité des clusters nouvellement créés**.
- On **réitère l'opération** jusqu'à **obtenir un unique cluster** ou bien un nombre de clusters préalablement défini.

Dans notre exemple le nombre de clusters optimal est 3 et le résultat final se trouve sur la figure de droite.

Dans la classification ascendante hiérarchique généralement on utilise la **distance euclidienne**, soient p = (p1,….,pn) et q = (q1,….,qn)

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_i-q_i)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Elle permet d'**évaluer la distance entre les centroïdes**. A chaque étape de regroupement entre deux centroïdes on obtiendra un nouveau cluster et un **nouveau centroïde qui n'est autre que le centre de gravité du nuage de points** comme expliqué plus haut.

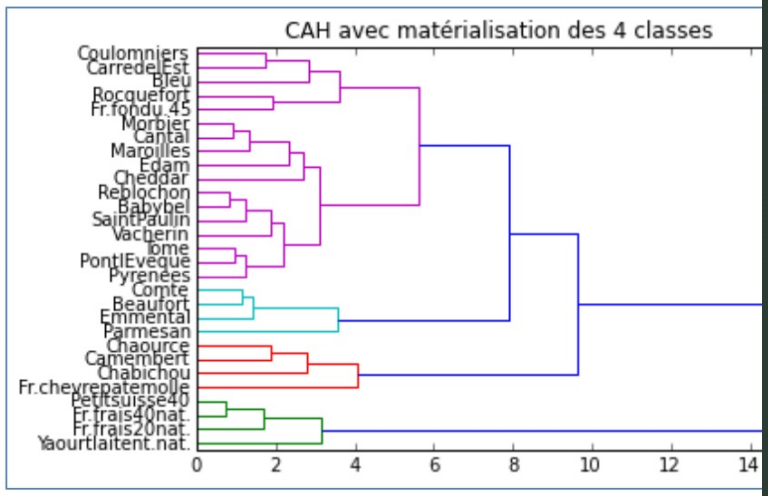
- Un dendrogramme est un diagramme fréquemment utilisé pour illustrer l'arrangement de groupes générés par un regroupement hiérarchique ou hiérarchisant

Le dendrogramme

Le dendrogramme « suggère » un découpage en 4 groupes. On note qu'une classe de fromages, les « fromages frais » (tout à gauche), se démarque fortement des autres au point qu'on aurait pu envisager aussi un découpage en 2 groupes seulement. Nous y reviendrons plus longuement lorsque nous mixerons l'analyse avec une analyse en composantes principales (ACP).

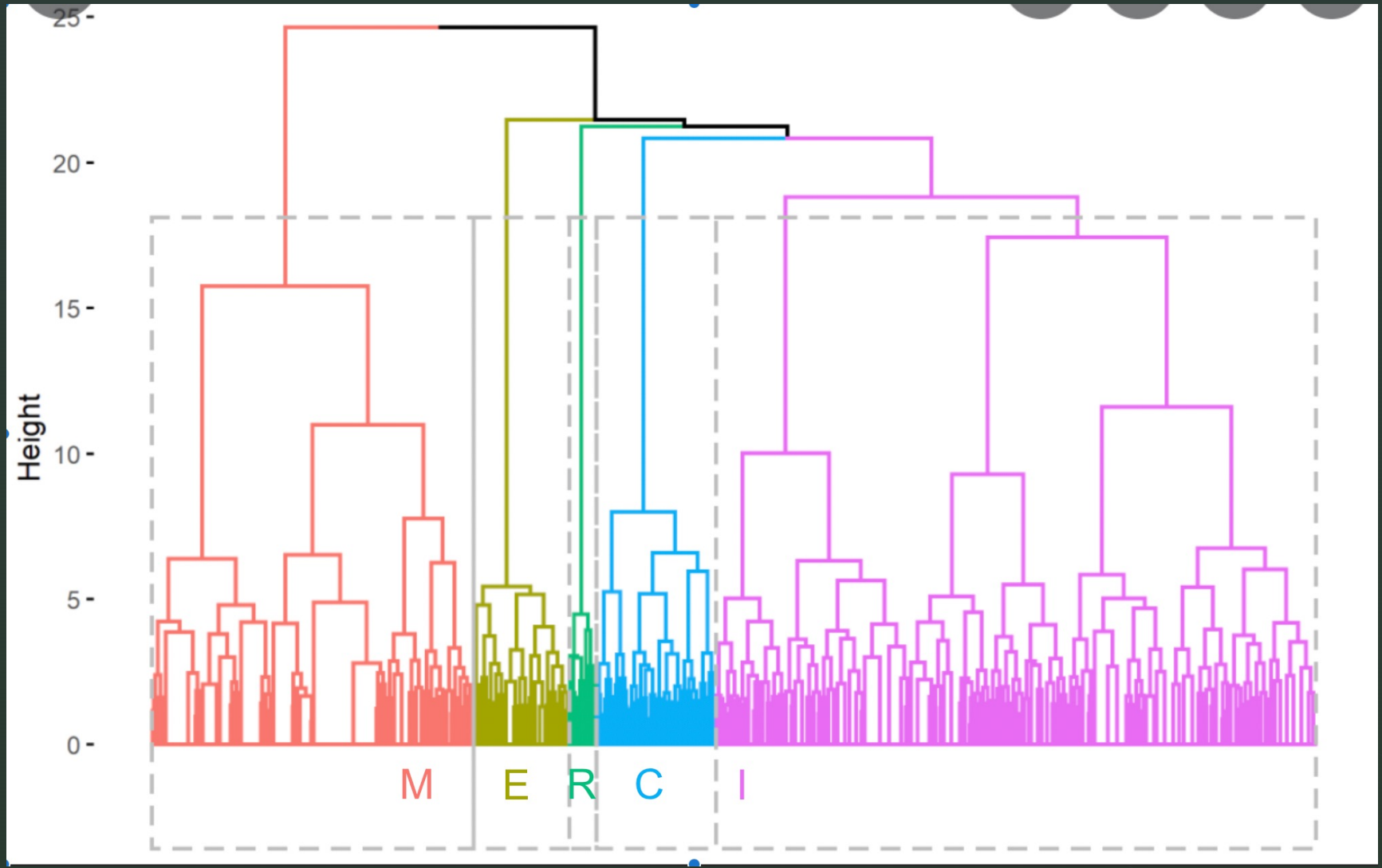
```
#affichage des observations et leurs groupes
print(pandas.DataFrame(fromage.index[idg],groupes_cah[idg]))
```

Groupe	Fromage
1	Yaourtlaitent.nat.
1	Fr.frais20nat.
1	Petitsuisse40
1	Fr.frais40nat.
2	Fr.chevrepatemolle
2	Camembert
2	Chabichou
2	Chaource
3	Emmental
3	Parmesan
3	Beaufort
3	Comte
4	Pyrenees
4	PontlEveque
4	Rocquefort
4	SaintPaulin
4	Tome
4	Reblochon
4	CarredelEst
4	Maroilles
4	Vacherin
4	Edam
4	Coulomniers
4	Cheddar
4	Cantal
4	Bleu
4	Babybel
4	Morbier



Le 1er groupe est constitué de fromages frais.
Le 2nd de fromages à pâte molle.
Le 3ème de fromages « durs ».
Le 4ème est un peu fourre-tout (de mon point de vue).

Mes compétences en fromage s'arrêtent là (merci à Wikipédia). Pour une caractérisation à l'aide des variables de l'étude, il faut passer par des techniques statistiques univariées (simples à lire) ou multivariées (tenant compte des relations entre les variables).


25
20
15
10
5
0
Height
M
E
R
C
I

Sourcing:

-http://eric.univ-lyon2.fr/~ricco/cours/didacticiels/Python/cah_kmeans_avec_python.pdf

-https://datascientest.com/machine-learning-clustering-focus-sur-algorithme-cah