Classification non supervisée

Statistiques II pour IENM2

Tiré du cours de Thomas Rieutord

Occitane Barbaux (occitane.barbaux@umr-cnrm.fr)

24 Janvier 2023

Déroulement:

- 4h de cours en deux parties: Présentation et TD
- 3h de TP informatiques en 3 groupes (sur Linux les 1, 3 et 6 février.)

Présentation:

- Diplômée de l'ENSAE (Statistiques et Machine Learning)
- En thèse au CNRM/GMGEC/CLIMSTAT depuis octobre 2022
- Sur les températures extrêmes (Théorie des valeurs extrêmes, bayésiens, etc)
- Vous pouvez me contacter à occitane.barbaux@umr-cnrm.fr

■ Introduction
► Définitions ► Notations
■ Distance
► Notion ► Distance entre caractéristiques ► Distance entre individus
■ Classification en partitions
► Algorithme des K-means ► Choix de K ► Limites du K-means ► Variantes
■ Classification Hiérarchique
► Classification hiérarchique ► Sens ► Distance cophrénétique ► Dendrogramme
► Avantages et limites ► Choix de K
■ Classification par Densité
► Algorithme DBSCAN ► Avantages et limites
■ Conclusion
■ Bibliographie

Classification automatique : un algorithme qui attribue une classe à chaque individu selon ses caractéristiques.

- Algorithme : suite d'opérations explicites permettant de résoudre un problème (ici : regrouper les individus en classes).
- Classe : groupement d'individus dont les caractéristiques sont proches.
- Individu : un élément de notre échantillon. Ils sont représentés par un vecteur de ses caractéristiques.
- Caractéristique : Propriété mesurables d'un individu. Ce sont les composantes des vecteurs qui forment les individus.

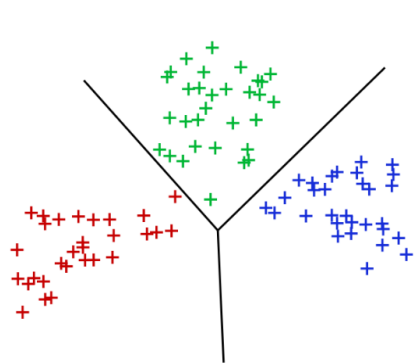
En statistiques, on cherche à expliquer une variable de sortie Y en fonction de variables d'entrée $X_1, ..., X_p$. Cela revient à trouver la fonction de décision f qui donne la meilleur approximations de

$$Y = f(X_1, ..., X_p)$$

- Lorsque Y est une variable continue, on parle de régression.
- Lorsque Y est une variable discrète, on parle de classification.

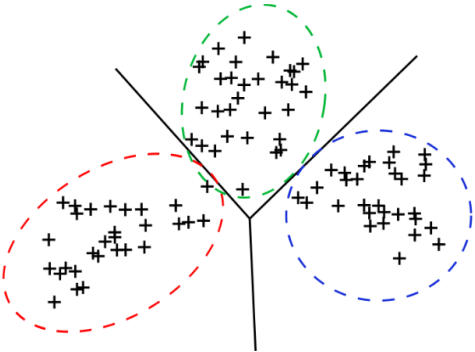
Ici, pour chaque individu i; Y_i est la meilleure classe. Mais il est possible de définir un jeu de probabilités d'appartenance à chacune des classes : "soft clustering".

Note: En machine learning, on cherche à prédire Y pour un nouvel individu.



Meilleures frontières à partir des

- caractéristiques (position des points)
- étiquettes (couleur)



Zone de fortes densités à partir des:

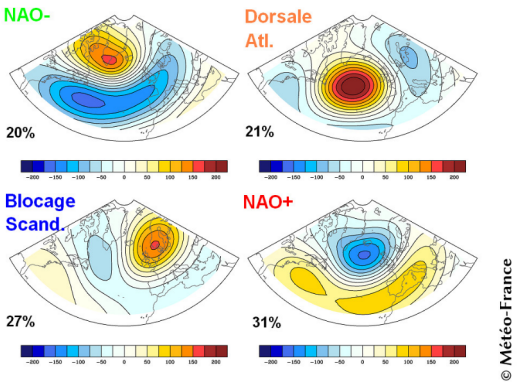
- caractéristiques UNIQUEMENT (position des points)
- pas d'étiquettes (couleur)

Idée générale: déterminer la structure sous-jacente des données.

Pour l'épidémiologie, l'économie et les science sociales : faire émerger des hypothèses explicatives. (John Snow, l'épidémie de Choléra de Londres)

En Machine Learning: Algorithmes de recommandations (permet d'identifier et cibler des comportements propres à chaque classe de consommateurs.)

D'autres approches existent: apprentissage mixte (Apprentissage d'images), apprentissage par renforcement.

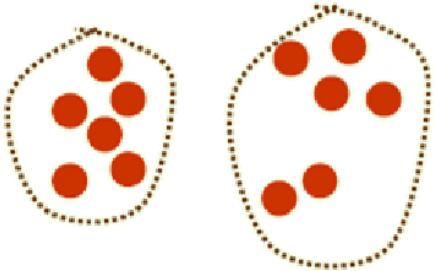


Individus: cartes mensuelles météorologiques
Caractéristiques : 10 premières composantes principales

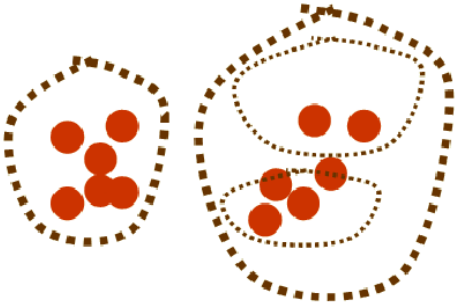
But: identification de schémas récurrents qui aident à prévoir le temps à moyenne échéance

Le résultat d'une classification peut être soit:

Une partition mathématique avec un nombre de classes fixe (ex: k-mean)



Une hiérarchie avec des classes emboîtées (ex: classification hiérarchique)



Classification automatique : un algorithme qui attribue une classe à chaque individu selon ses caractéristiques.

- Individu i : un vecteur $x^i = (x^i_1, ..., x^i_p)$
- Caractéristique j : Composante $j \in [1, p]$ des vecteurs x^i
- Algorithme : pour tout individu i, attribue une classe k : $C(i) = k$
- Classe : Ensemble tel que $C_k = \{i, C(i) = k\}$ proches.

On dispose de N individus avec k caractéristiques.

■ Introduction
▶ Définitions ▶ Notations

■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus

■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes

■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K

■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites

■ Conclusion

■ Bibliographie

Une seule hypothèse : plus deux individus sont proches, plus ils ont de chances de faire partie de la même classe.
Pour deux individus i et i', on définit la notion de distance $d(x^i, x^{i'})$

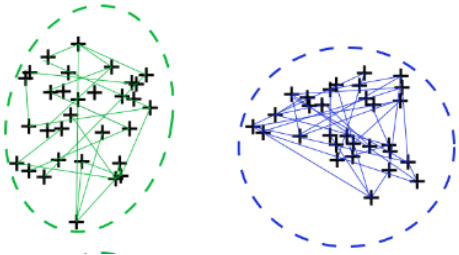
Évaluer la qualité d'une classification:

- dispersion à l'intérieur d'une classe (minimiser pour des observations homogènes)
- distance entre les classes (maximiser pour des clusters distincts)

La fonction d'attribution C doit minimiser une fonction de coût qui dépend de :

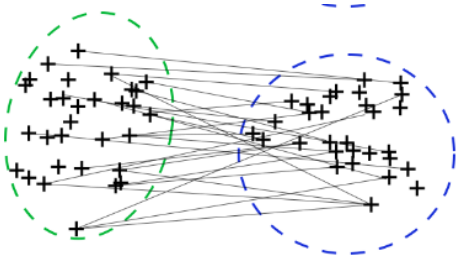
La dispersion intra-cluster (Within)

$$W(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x^i, x^{i'})$$



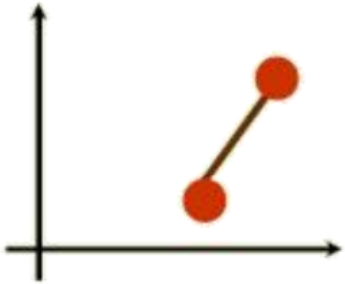
La dispersion inter-cluster (Between)

$$B(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d(x^i, x^{i'})$$

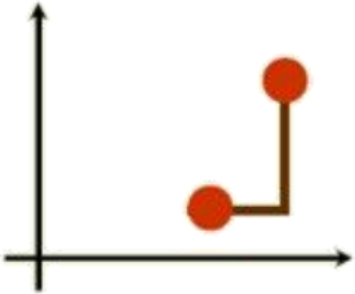


Pour une variable quantitative: Toute distance convient.

Distance Euclidienne:



Distance de Manhattan:



Pour une variable ordinale:

Exemple : $X_j \in$ très mauvais, mauvais, neutre, bien, très bien
Les valeurs prises par X_j sont numérotées de 1 à M . On remplace chaque élément de la colonne j par

$$\frac{k - 1/2}{M}$$

où $k \in [1, M]$ est le rang de X_j^i parmi les valeurs possibles .

On traite ensuite cette variable transformée comme une variable quantitative.

Pour une variable catégorielle:

Exemple : Espèce, pictogramme d'évènement météo $X_j \in$ pluie, neige, soleil, etc
Il faut définir une matrice de dissimilarité.

	I	II	III	IV
I	-	0.1	0.4	0.6
II		-	0.5	0.5
III			-	0.6
IV				-

Pour k une valeur possible, on pose généralement $d_j(x^i, x^{i'}) = 0$ si $k \neq k'$ et 1 sinon.

Distance entre individus données par:

$$d_{ii'} = d(x^i, x^{i'}) = \sum_{j=1}^{p} w_j d_j(x^i, x^{i'})$$

Problèmes :

- Choix des poids w_j entre les caractéristiques
- Problèmes des unités
- Problème des données manquantes

Choix des poids pour les caractéristiques ?

Dissimilarité totale de l'échantillon:

$$\bar{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x^i, x^{i'}) = \sum_{j=1}^{p} w_j \bar{d}_j$$

Avec $\bar{d}_j = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_j(x_j^i, x_j^{i'})$ pour une caractéristique

- La contribution de la j-ème caractéristique à la dissimilarité totale est $w_j \bar{d}_j$.
- On peut égaliser les contributions
- Mais certaines caractéristiques "séparent" peut-être mieux que d'autres.

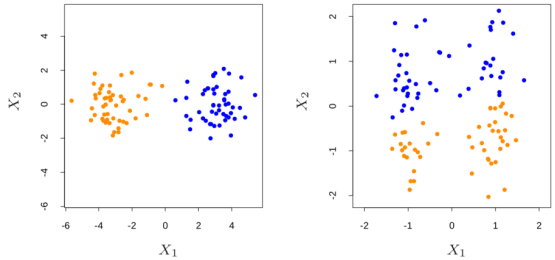


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with K=2) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights* $1/[2 \cdot \mathrm{var}(X_j)]$. *The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

©figure de Hastie, Tibshirani, Friedman (2001).

Exemple : Si X_1 est une P_{mer} en hectopascal (≈ 1000) et X_2 est une surcôte en mètres (≈ 0.1), X_1 est prépondérant dans la dissimilarité à cause de son unité

Possibilité de normaliser les caractéristiques:

- "uniforme" : $\hat{X}_j = \frac{X_j - min(X_j)}{max(X_j) - min(X_j)}$
- "normale": $\hat{X}_j = \frac{X_j - \bar{(}X_j)}{\sigma(X_j)}$
- etc.

Ce choix peut influer fortement sur les résultats. Il doit être fait par un expert.

Raison variées : (panne capteur, erreur, valeur aberrante...)

Soient deux individus x^i et x^j:

- Si x^i et x^j ont des caractéristiques communes, on omet les caractéristiques communes et on adapte les poids.
- Si x^i et x^j n'ont aucune caractéristique commune, la distance est incalculable. Possibilités:
  - On retire un individu.
  - On complète les valeurs manquantes

■ Introduction
▶ Définitions ▶ Notations
■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus
■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes
■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K
■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites
■ Conclusion
■ Bibliographie

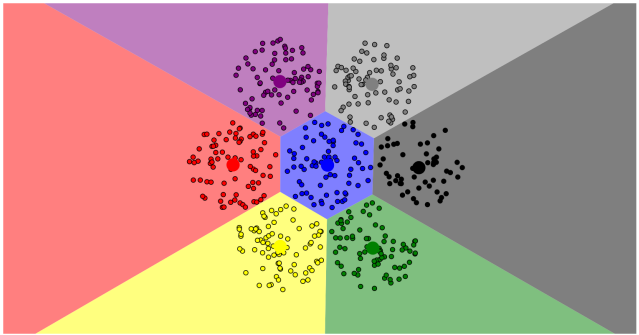
- Algorithme de partitionnement des K-means (MacQueen, 1967)
- Nombre de classes K fixé au préalable
- Les séparations sont des hyperplans (droites en 2D)

Déroulement

1. Prendre aléatoirement K points comme centroïde
2. Assigner chaque individu au centroïde le plus proche.
3. Recalculer le centroid comme le centre du cluster
4. Recommencer 2. et 3. jusqu'à convergence

Illustration

Exemple de résultats pour 7 clusters :



La convergence est :

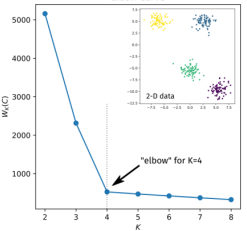
- Rapide (Quelques dizaines d'itérations)
- Seulement vers un minimum local du coût, pas nécessairement le maximum global
- De complexité faible $O(n)$ (n le nombre d'individus)

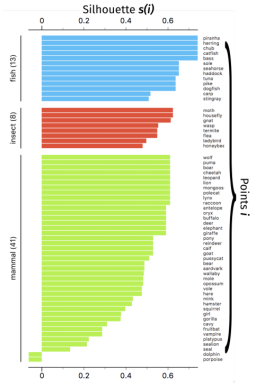
Critères de convergence

- Pas (ou peu) de réassignement de points dans un autre cluster
- Pas (ou peu) de déplacement des centroïdes
- Minimum de la diminution de la somme des erreurs au carré (Ici pour chaque point, distance au centroïde)

Choix du nombre de classes:

- Peut être imposé par le problème (ex: 10 chiffres)
- Répéter l'algorithme pour divers valeurs de K et chercher un "coude" dans la courbe de $W_K(C)$
- Optimiser un score de classification (Méthode de la silhouette)

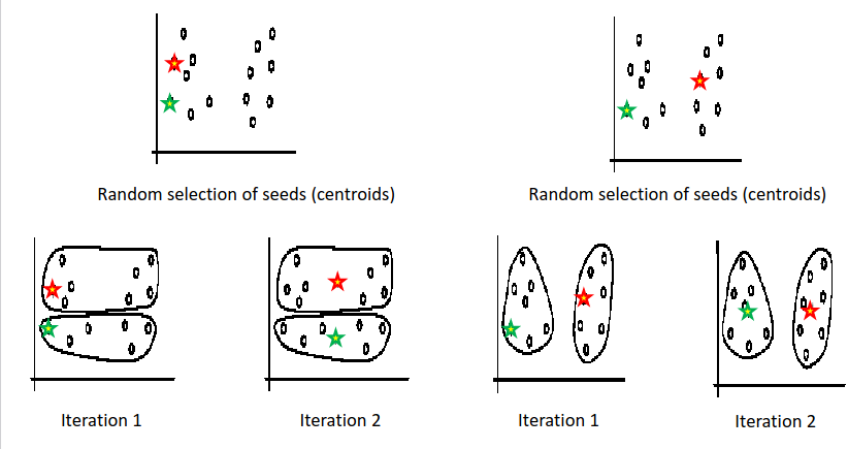




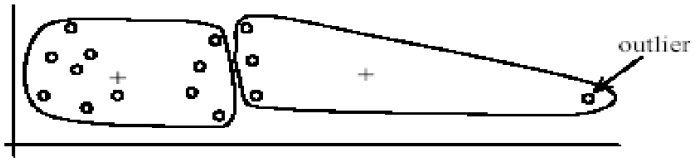
Limites du K-means

- Sensibilité aux centroïde initiaux
- Sensible à la présence d'outliers
- Inadapté selon la forme des données.

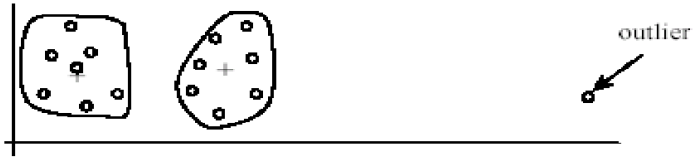
Les résultats sont très sensibles au choix des premiers centres de cluster:



La présence d'Outliers peut aussi biaiser la formation des clusters:

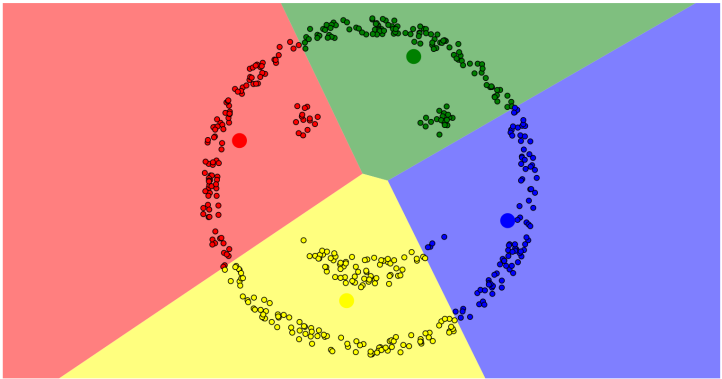


(A): Undesirable clusters



(B): Ideal clusters

L'algorithme est inadapté si les clusters à trouver ne sont pas des hyper-ellipsoïdes:



Diverses variantes existent:

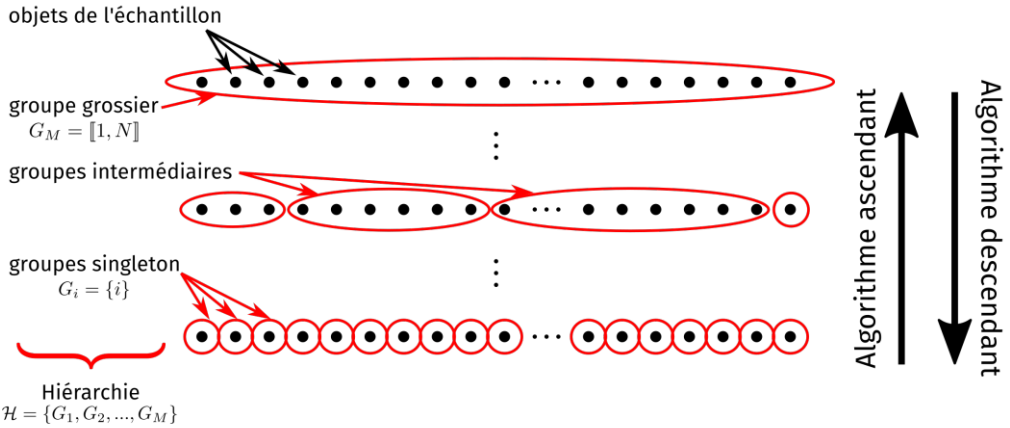
- K-medoids : Les centres de clusters sont systématiquement des points (individus). Moins sensible aux outliers et au bruit.
- K-median: Distance de Manhattan et le centre est une médiane des points du cluster. Plus adapté pour des variables discrètes ou catégorielles.
- Hierarchical k-means clustering: Pas de K fixé, car il combine k-means et classification hiérachique.

L'algorithme des k-means:

- Très populaires
- Nombre de classes K fixé à priori.
- Converge rapidement, vers un minimum local.
- Basé sur la minimisation de la distance au centre du groupe.
- Sensible à l'initialisation et aux outliers.
- Des variantes existent pour l'adapter à certaines contraintes (choix de la dissimilarité, définition des centroïdes)
- Pas adapté pour toutes les données.

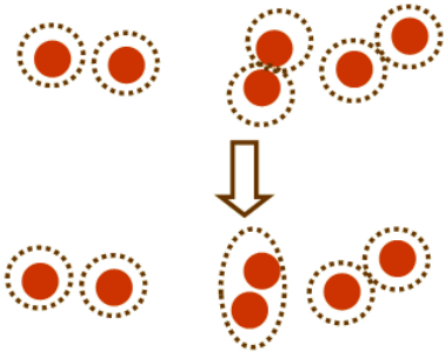
■ Introduction
▶ Définitions ▶ Notations
■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus
■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes
■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K
■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites
■ Conclusion
■ Bibliographie

Une classification hiérarchique fournit une hiérarchie de groupes (i.e. des groupes imbriqués les uns dans les autres).



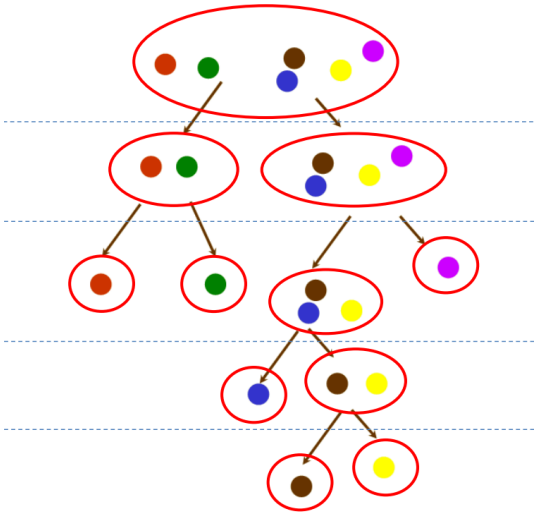
Classification Ascendante

1. Par le bas: Chaque singleton compose un groupe.
2. On réunit on maximisant la distance entre les groupes restants.
3. On réunit les deux groupes les plus proches.
4. Recommencer 2. jusqu'à ce que tous les points soient dans un seul groupe.



Classification Descendante / divisive.

1. Par le haut: Le groupe contient tous les individus.
2. On divise en maximisant la distance entre les groupes
3. A chaque étape, on prend le groupe avec la plus grande variance intra, et on le divise suivant (2.)
4. Recommencer 3. jusqu'à ce qu'il ne reste que des singletons.



Dans un algorithme ascendant, on commence par considérer tous les points comme des groupes puis on réunit les 2 groupes les plus proches en un groupe père, et ainsi de suite.

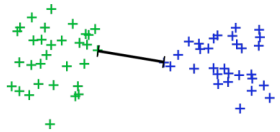
$\Longrightarrow$ Par construction, la distance entre les groupes réunis est croissante : La distance cophénétique.
$\Longrightarrow$ La distance choisie influe sur la formation des groupes.
$\Longrightarrow$ On souhaite que les classes qui en résultent soient bonnes :

- Compacité : classes de petit diamètre
- Proximité : points d'une même classe plus proche que ceux de la classe voisine

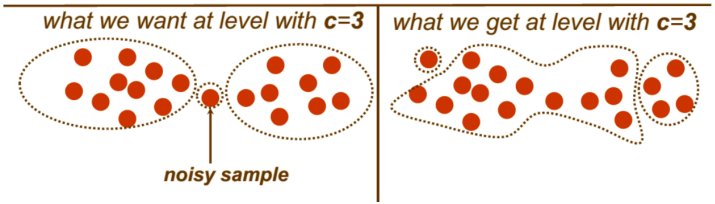
$\Longrightarrow$ Il existe un grand nombre de distances possibles.
$\Longrightarrow$ Choix ? Arbitrage avantages/inconvénients préférablement par un expert du problème à résoudre.

Soient G et H deux groupes.

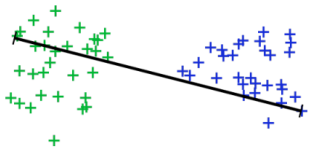
Saut minimal : $D(G,H) = \min_{i \in G, i' \in H} d_{ii'}$



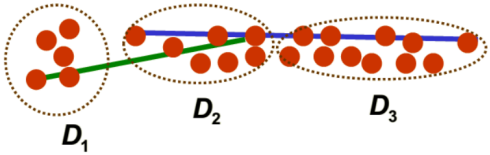
- Nommé aussi: simple linkage , plus proche voisin.
- Tendance à faire des chaines de points : les clusters résultants seront étirés, de large diamètre. Bonne détection de bandes/strates.
- Limites : Clusters peu compacts et Sensible au bruit.



Saut maximal : $D(G,H) = \max_{i \in G, i' \in H} d_{ii'}$



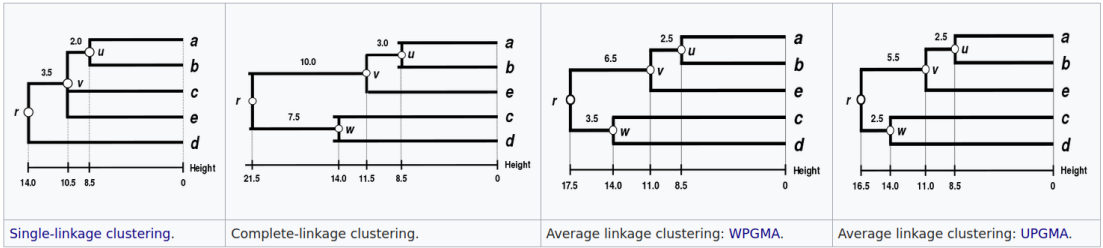
- Nommé aussi: complete linkage , voisin le plus éloigné.
- Tendance à faire des clusters compacts et de faible diamètre.
- Limites : les points au bord peuvent être plus proche d'un autre cluster que d'un autre point du même cluster. Fonctionne mal pour les clusters allongés.



De nombreuses distances existent:

- Distance moyenne : Compromis compacité/proximité
- Distance aux centroïdes: adaptée pour les clusters avec des formes assez régulières et de taille équivalente. Ne tiens pas compte de la taille ou la forme des clusters.
- Distance de Ward: maximise l'augmentation d'inertie inter-classe (= Coût pour la distance Euclidienne). Ne tiens pas compte de la forme.
- Etc.

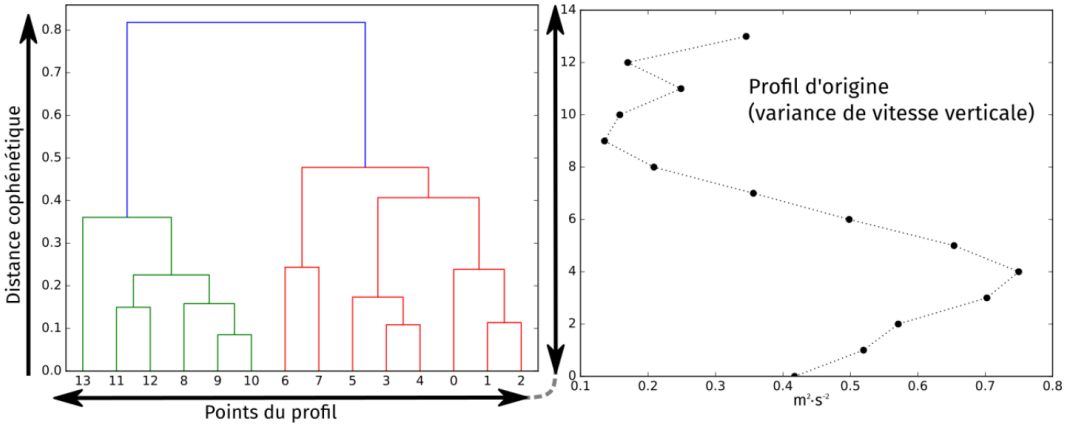
Selon la distance choisie, les clusters seront composés différemment.

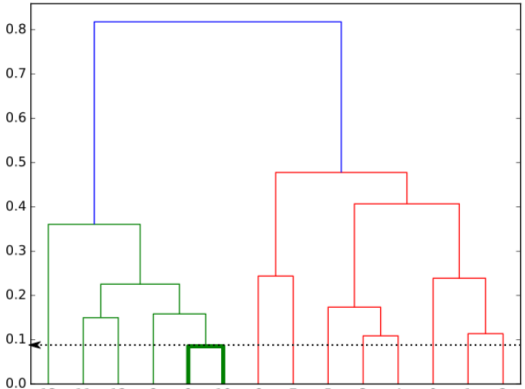


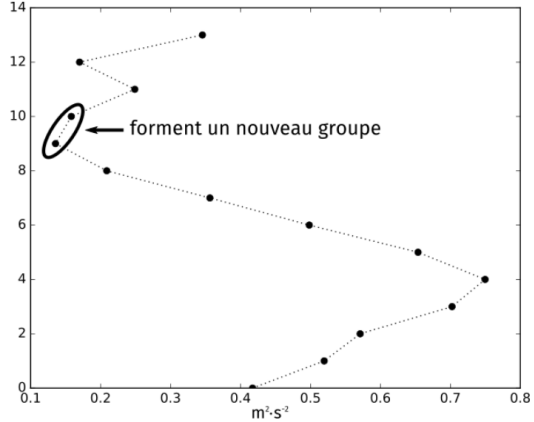
Single-linkage clustering. | Complete-linkage clustering. | Average linkage clustering: WPGMA. | Average linkage clustering: UPGMA.

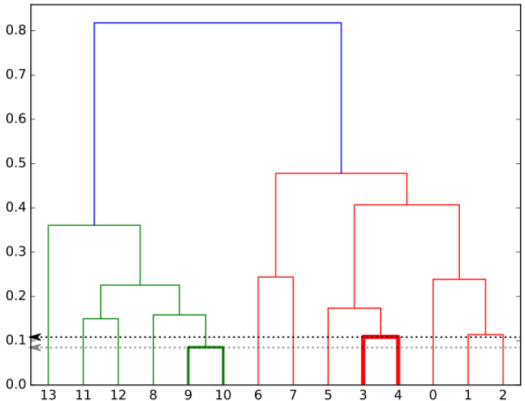
Source : Wikipédia WPGMA

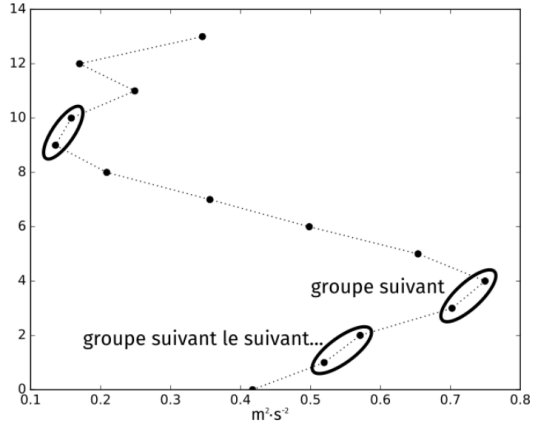
On peut représenter les réunions successives par une structure en arbre : Un dendrogramme.

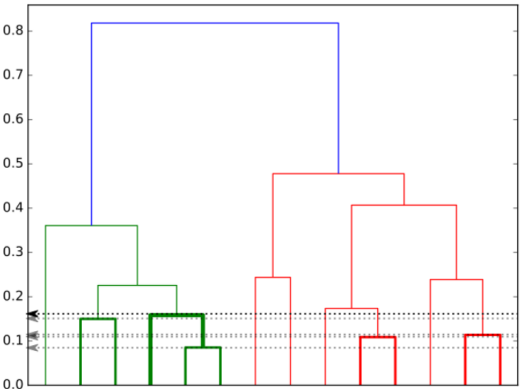

Distance cophénétique
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
13 11 12 8 9 10 6 7 5 3 4 0 1 2
Points du profil
Profil d'origine
(variance de vitesse verticale)
0
2
4
6
8
10
12
14
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m²·s⁻²

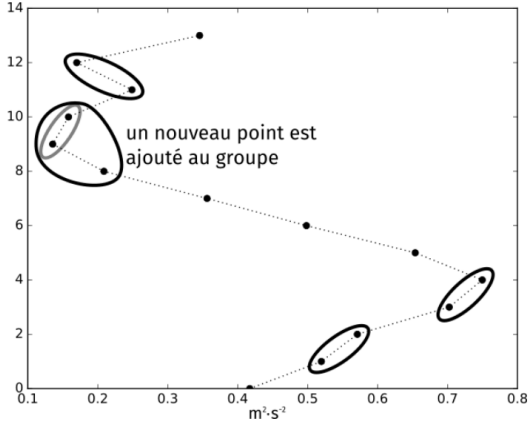

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
13 11 12 8 9 10 6 7 5 3 4 0 1 2

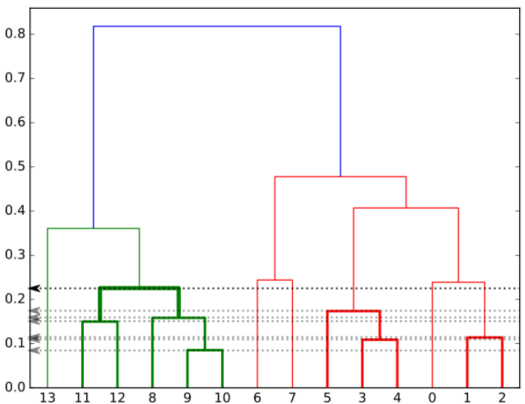

forment un nouveau groupe
14
12
10
8
6
4
2
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m².s⁻²

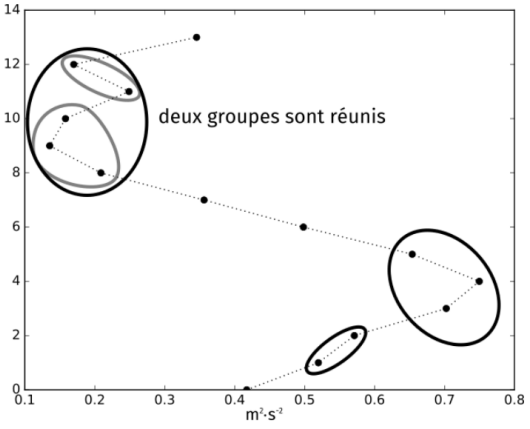

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
13 11 12 8 9 10 6 7 5 3 4 0 1 2

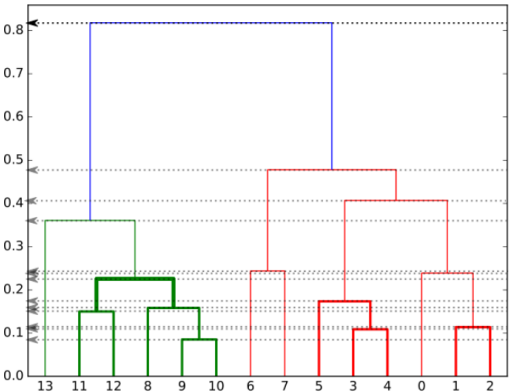

14
12
10
8
6
4
2
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m²·s⁻²
groupe suivant
groupe suivant le suivant...

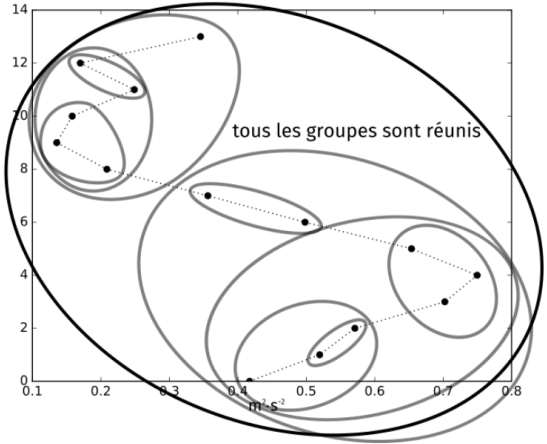

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
13 11 12 8 9 10 6 7 5 3 4 0 1 2


un nouveau point est
ajouté au groupe
14
12
10
8
6
4
2
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m²·s⁻²


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
13 11 12 8 9 10 6 7 5 3 4 0 1 2


deux groupes sont réunis
14
12
10
8
6
4
2
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m²·s⁻²


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
13 11 12 8 9 10 6 7 5 3 4 0 1 2


tous les groupes sont réunis
14
12
10
8
6
4
2
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m·s-2

Avantages:

- Résumé visuel et complet des données.
- Permet d'identifier des groupes à plusieurs échelles.
- Les groupes sont imbriqués (cohérence entre les échelles).
- Une fois le nombre de groupes fixés, les groupes sont déjà calculés. calculés.

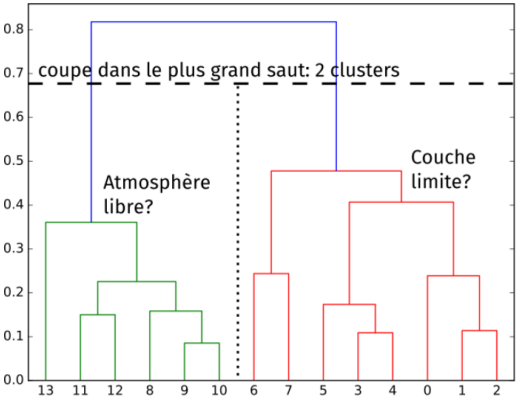
Inconvénients:

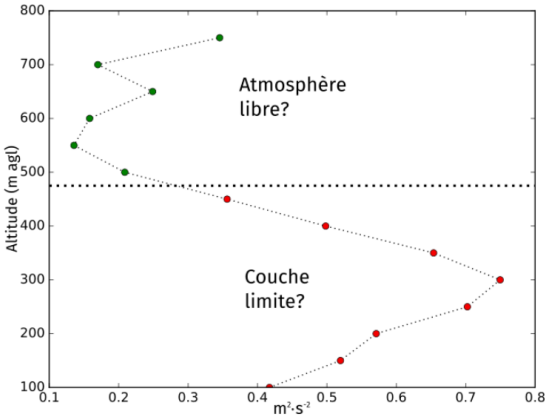
- Dépend du choix de la distance cophénétique.
- Donne toujours une structure hiérarchique, même lorsqu'il n'y en a pas.
- De petites modifications des données peuvent produire un dendrogramme assez différent.
- Pour obtenir une partition, il faut choisir où couper.

Problème : le clustering hiérarchique fourni une hiérarchie et non une partition.

Comment fixer le nombre de classes ?

- Visuellement : Chercher les noeuds qui suivent de grands sauts
- De façon automatique : Calculer les coefficients d'inconsistance.


coupe dans le plus grand saut: 2 clusters
Atmosphère libre?
Couche limite?
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
13 11 12 8 9 10 6 7 5 3 4 0 1 2

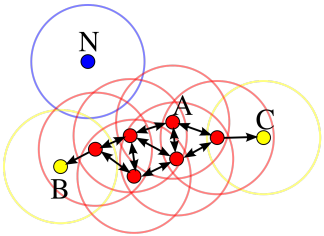

Atmosphère libre?
Couche limite?
Altitude (m agl)
100
200
300
400
500
600
700
800
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
m².s⁻²

■ Introduction
▶ Définitions ▶ Notations
■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus
■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes
■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K
■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites
■ Conclusion
■ Bibliographie

- Density-Based Spatial Clustering of Applications with Noise (Ester et al., 1996)
- On ne fixe pas K, mais m (nombre de voisins minimums) et ε (distance maximale entre voisins)
- Le partitionnement est basé sur la densité de points.

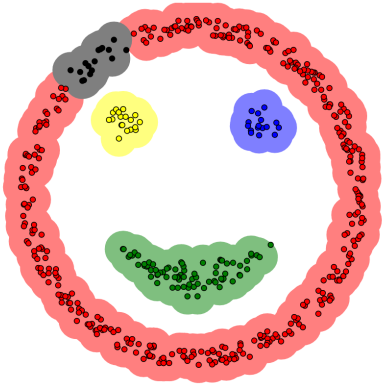
Déroulement

1. Pour tout point x^i, trouver son ε-voisinage.
2. Les points avec plus de m voisins appelé points coeur (A). Les points coeurs connecté entre eux sont mis dans le même cluster
3. Les points non-coeur sont soit au bord (B,C) s'ils sont voisins d'un point coeur, soit aberrant (N) sinon.



Illustration

Exemple de résultats :



Avantages:

- Trouve naturellement le nombre de partitions: pas besoin de fixer K au préalable.
- Les groupes peuvent être de forme quelconque (circulaire, entrelacés...).
- Robuste au bruit : capable de repérer des points aberrants.

Inconvénients:

- Il faut fixer m (nombre de voisins minimums) et ϵ (distance maximale entre voisins)
- Les points au bord reliés à plus d'un groupe peuvent changer d'affectation suivant l'ordre des points.
- Les groupes doivent être de densité comparable.
- Les résultats dépendent de la distance choisie.

■ Introduction
▶ Définitions ▶ Notations
■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus
■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes
■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K
■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites
■ Conclusion
■ Bibliographie

Conclusion:

- La classification non-supervisée est un champ de recherche actif, avec un grand nombre d'algorithmes et de variantes.
- C'est une méthode très utilisée dans de nombreux domaines
- La classification est très utile pour faire ressortir des groupes dans les données, mais difficile à évaluer.
- La méthodes et les résultats sont extrêmement dépendants des données, distances, paramètres, etc. choisis.

■ Introduction
▶ Définitions ▶ Notations
■ Distance
▶ Notion ▶ Distance entre caractéristiques ▶ Distance entre individus
■ Classification en partitions
▶ Algorithme des K-means ▶ Choix de K ▶ Limites du K-means ▶ Variantes
■ Classification Hiérarchique
▶ Classification hiérarchique ▶ Sens ▶ Distance cophrénétique ▶ Dendrogramme
▶ Avantages et limites ▶ Choix de K
■ Classification par Densité
▶ Algorithme DBSCAN ▶ Avantages et limites
■ Conclusion
■ Bibliographie

Utilisés pour ce cours:

- Thomas A Rieutord. Classification non-supervisée. École d'ingénieur. France. 2021. ⟨meteo-02465143v2⟩
- Ullman et al. Unsupervised learning Clustering. 9.54 MIT. 2014. Cours
- Wikipédia K-means DBSCAN Classification hiérarchique

Classification non supervisée

Merci d'avoir suivi ce cours
Des questions?