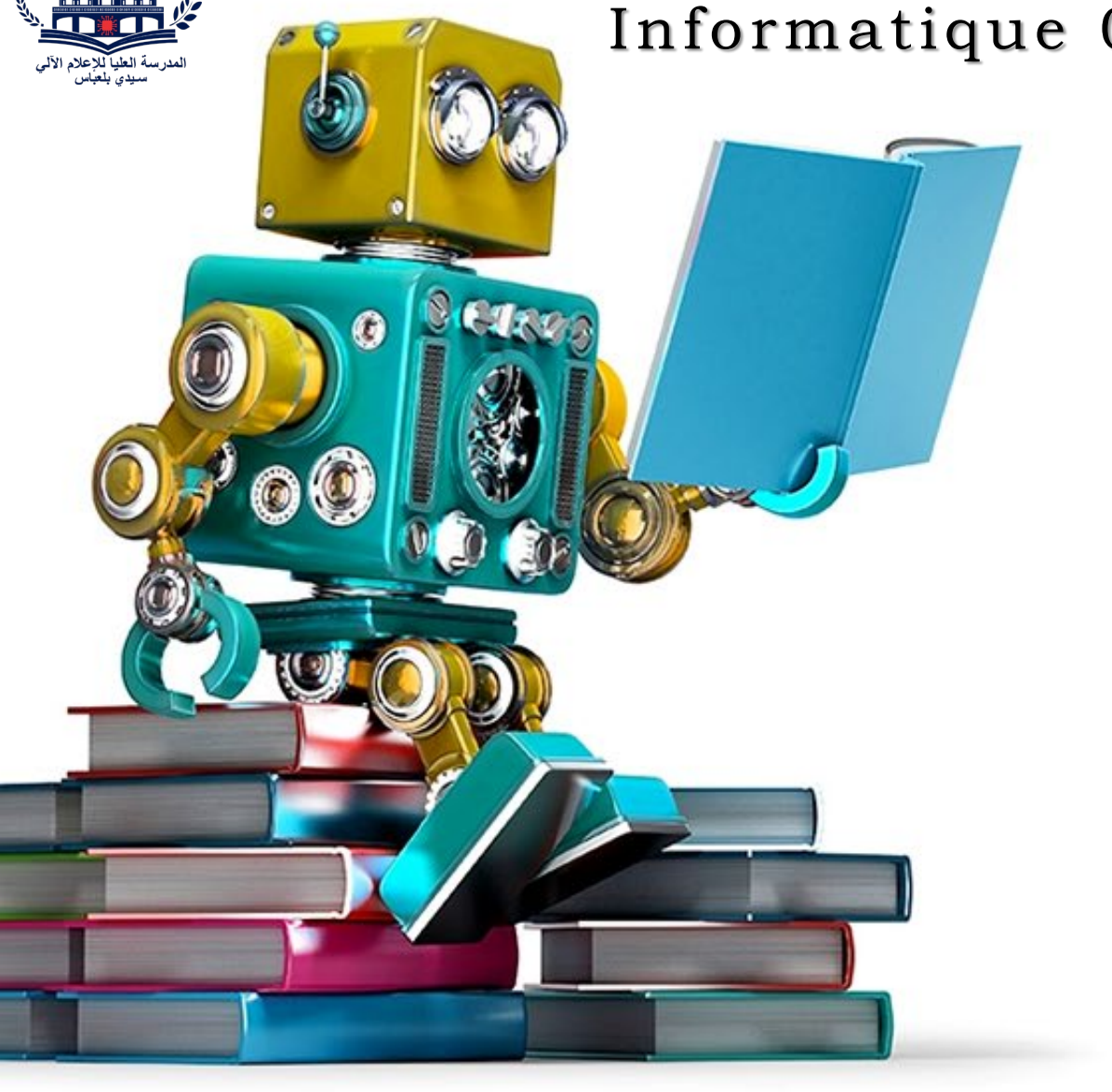
ESI-SBA - École Supérieure en Informatique 08-MAI-1945

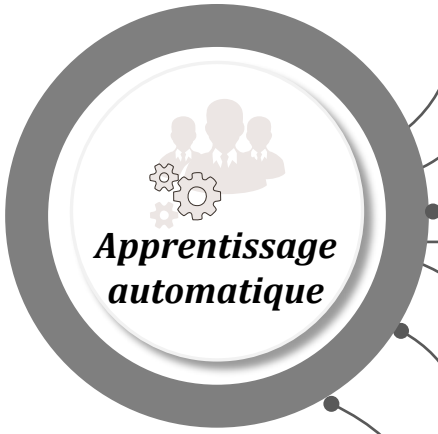


Apprentissage Automatique

Intelligence Artificielle et Sciences de Données (IASD)

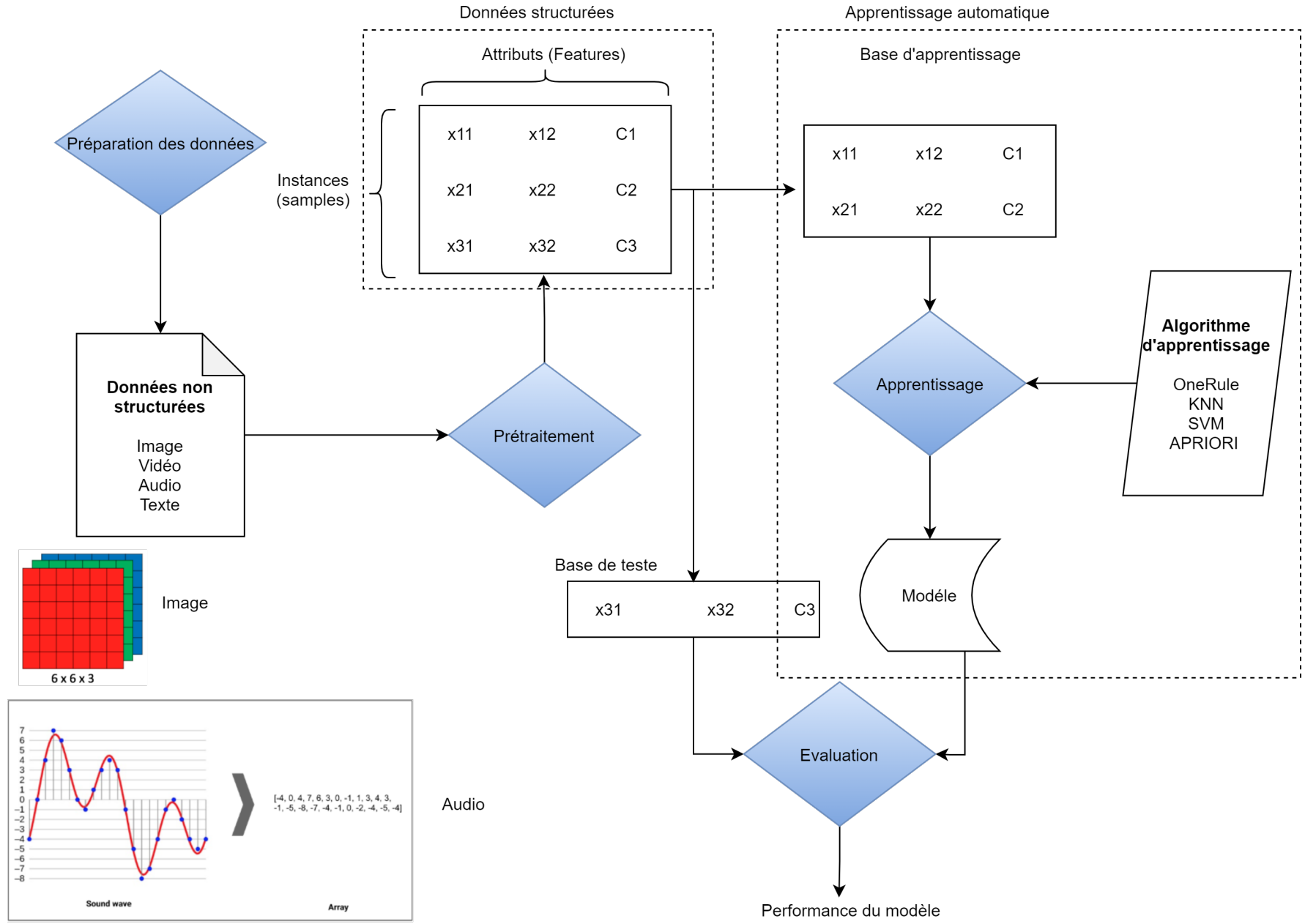
DR N. DIF

Programme



1. Introduction à l'apprentissage automatique
2. Régression
3. Évaluation des performances
4. Classification
5. **Clustering**
6. Sélection des attributs et réduction de la dimensionnalité
7. Méthodes ensemblistes

Rappel


Données structurées
Apprentissage automatique
Attributs (Features)
Base d'apprentissage
Préparation des données
Instances (samples)
x11 x12 C1
x21 x22 C2
x31 x32 C3
Données non structurées
Image
Vidéo
Audio
Texte
Prétraitement
Apprentissage
Algorithme d'apprentissage
OneRule
KNN
SVM
APRIORI
Base de teste
Modéle
Image
6 x 6 x 3
Audio
Sound wave
Array
Evaluation
Performance du modèle

1. CLUSTERING

I CAN CREATE CATEGORIES BY MY SELF

1.1. Principe

- Technique d'apprentissage non supervisée qui permet de regrouper les données non classifiées auparavant.
- Le but principal du clustering est de grouper les points de données similaires dans des clusters de telle sorte à maximiser la variance intercluster et de minimiser la variance intracluster.
- La similarité dépend d'un ensemble de paternes prédéfini dans la base d'apprentissage, comme : la forme, la taille, la couleur…

Domaines d'application :

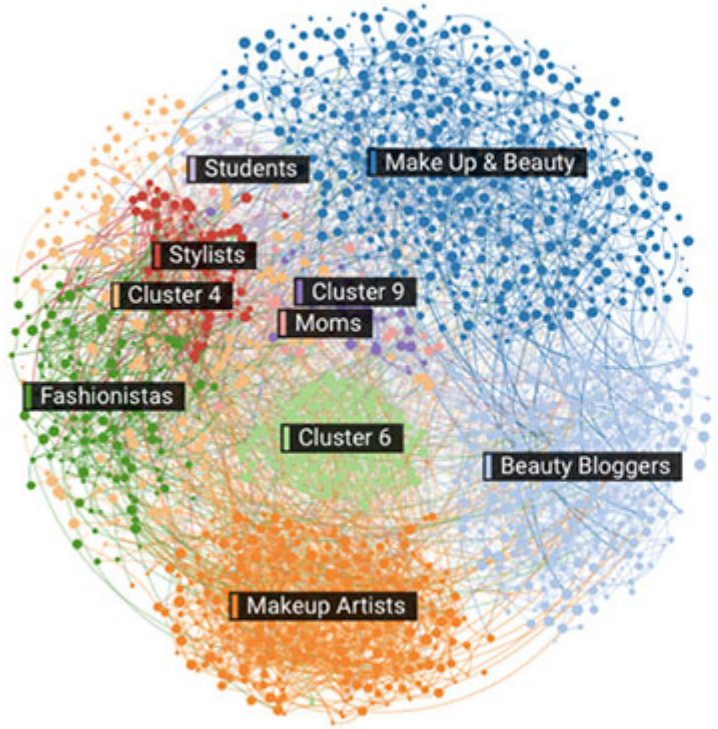
1. La segmentation du marché.
2. L'analyse des réseaux sociaux.
3. La segmentation des images.
4. Les systèmes de recommandation sur Amazon et Netflix.

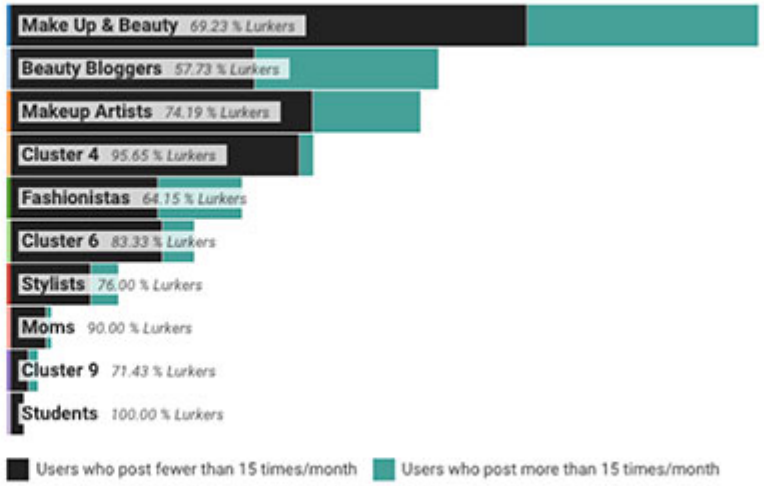
1. CLUSTERING

1.1. Principe

Domaines d'application :

L'analyse des réseaux sociaux.





From : https://www.metamaven.com/14-ways-machine-learning-can-boost-marketing/loreal_cluster_800px_web/

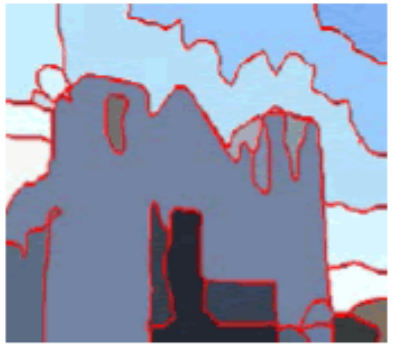
1. CLUSTERING

1.1. Principe

Domaines d'application :

La segmentation des images.

(a)



(b)

(c)

Figure 1: (a) is the original image; (b) and (c) are the segmentation results.

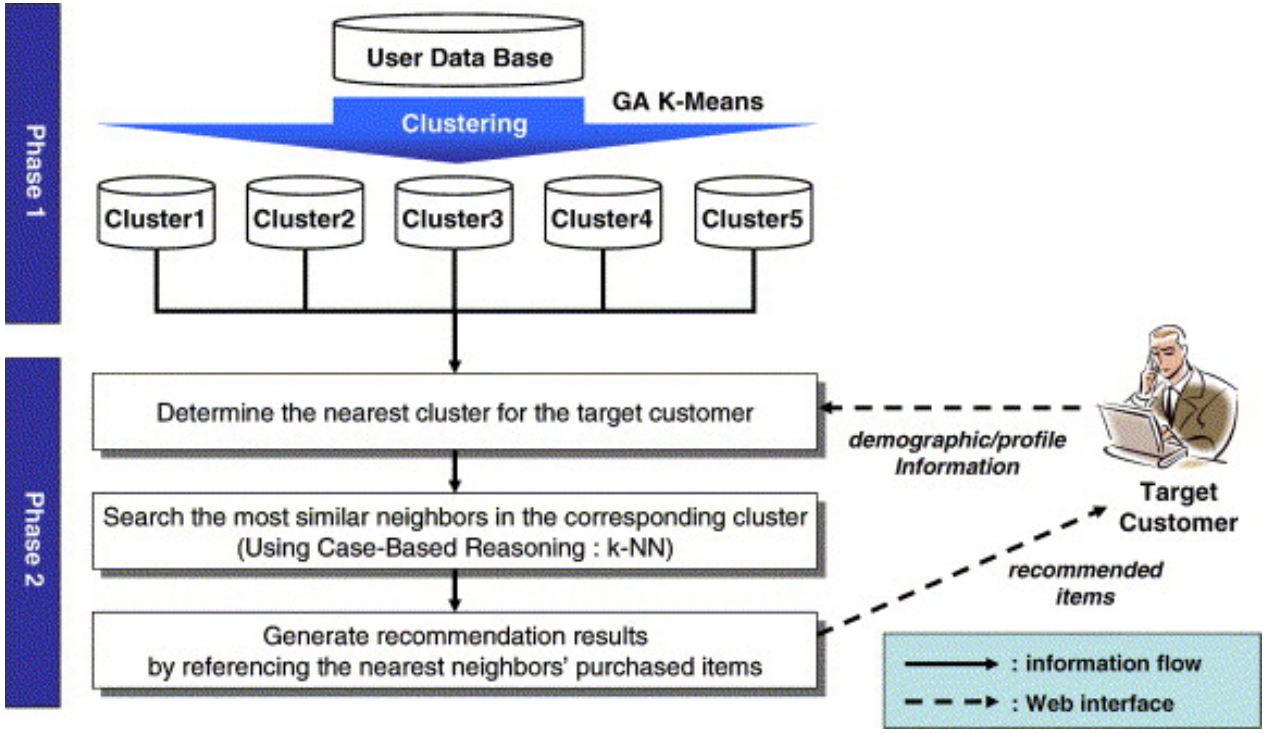
From : https://www.kdnuggets.com/2019/08/introduction-image-segmentation-k-means-clustering.html

1. CLUSTERING

1.1. Principe

Domaines d'application :

Les systèmes de recommandation.



From : Kim, K. J., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. Expert systems with applications, 34(2), 1200-1209.

1. CLUSTERING

1.1. Principe

Il existe plusieurs approches en clustering, le choix de la technique approprié dépend essentiellement des données traitées et leurs distributions. Parmi les approches les plus fréquemment utilisées en clustering, nous avons :

1. Les méthodes hiérarchiques.
2. Les méthodes de clustering à base de centroïde.
3. Les méthodes de clustering à base de distribution.
4. Les méthodes de clustering à base de densité.

1. CLUSTERING

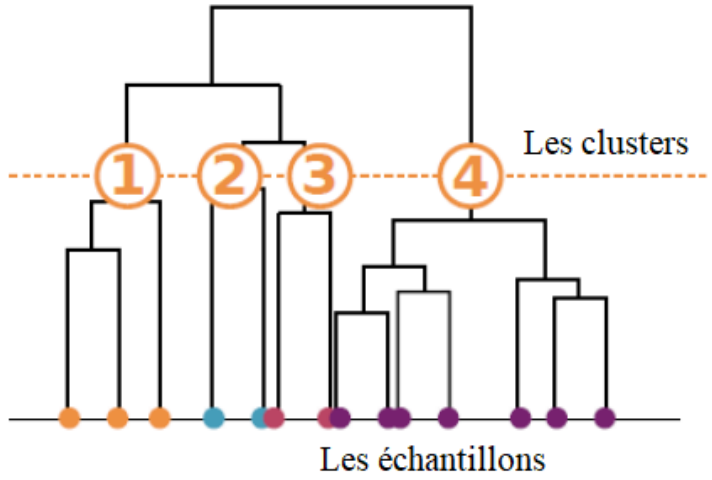
1.2. Les Méthodes hiérarchiques

- Les méthodes hiérarchiques sont de deux types :

1. Clustering ascendant (agglomératif).
2. Clustering descendant (divisif).

- Le résultat est représenté sous forme d'un dendrogramme (arbre binaire):

1. Les feuilles sont les échantillons,
2. Les nœuds sont les clusters
3. La hauteur des branches est proportionnelle à la distance entre clusters.

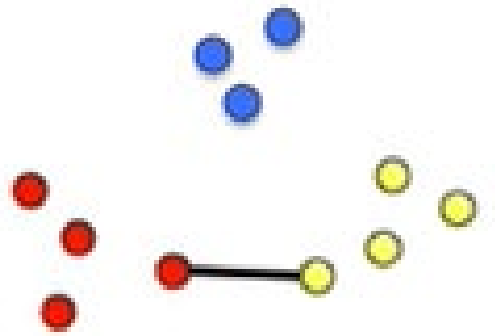


1. CLUSTERING

1.2. Les Méthodes hiérarchiques

1.2.1. Le clustering ascendant

- En premier, l'algorithme considère chaque point de données en tant que cluster unique, ensuite, ces clusters sont fusionnés itérativement jusqu'à arriver à un seul cluster.
- Les clusters proches sont fusionnés en fonction de la mesure de distance. Parmi les méthodes utilisées pour la fusion entre deux clusters :



1. **Méthode de liaison simple ou voisin le plus proche :** La proximité entre deux clusters est la proximité entre leurs deux objets les plus proches : $D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(c_1, c_2)$

N'est pas conseillé pour les clusters sphériques

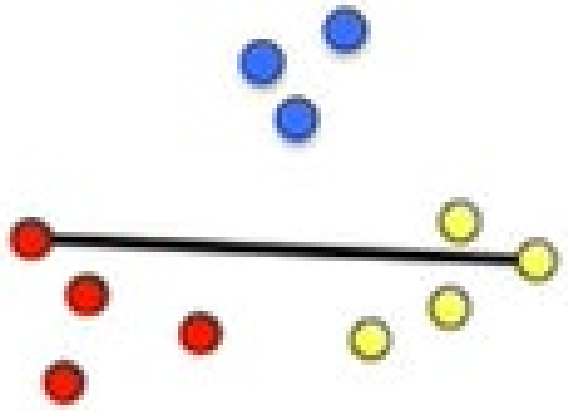
1. CLUSTERING

1.2. Les Méthodes hiérarchiques

1.2.1. Le clustering ascendant

2. **Méthode de liaison complète ou voisin le plus éloigné :** La proximité entre deux clusters est la proximité entre leurs deux objets les plus éloignés (diamètre).

$$D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(c_1, c_2)$$



- Elle permet de forcer les clusters sphériques.

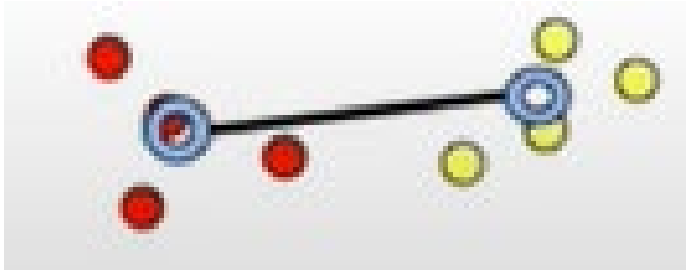
1. CLUSTERING

1.2. Les Méthodes hiérarchiques

1.2.1. Le clustering ascendant

3. **Méthode centroïde** : La proximité entre deux grappes est la proximité entre leurs centroïdes géométriques.

$$D(c_1, c_2) = \mathrm{D}(\left(\frac{1}{|c_1|}\sum_{x \in c_1} \vec{x}\right), \left(\frac{1}{|c_2|}\sum_{x \in c_2} \vec{x}\right))$$



1. CLUSTERING

1.2. Les Méthodes hiérarchiques

1.2.1. Le clustering ascendant

4. **Méthode de Ward** [1] : perte d'inertie interclasse due à cette agrégation revient au gain minimum d'inertie intraclasse. Le principe consiste à choisir les deux clusters dont l'union augmente moins la variance intraclassée.

$$D(\mathrm{A},\mathrm{B}) = \sum_{i\in A\cup B}\|\overrightarrow{x_i} - \overrightarrow{m}_{A\cup B}\|^2 - \sum_{i\in A}\|\overrightarrow{x_i} - \overrightarrow{m}_A\|^2 - \sum_{i\in B}\|\overrightarrow{x_i} - \overrightarrow{m}_B\|^2 = \frac{n_A n_B}{n_A+n_B}\|\overrightarrow{m_A} - \overrightarrow{m}_B\|^2 \quad [1]$$

1. CLUSTERING

1.2. Les Méthodes hiérarchiques

1.2.1. Le clustering descendant

- Effectue l'opération inverse du clustering agglomérative, car il s'agit d'une approche descendante.
- Ne nécessite pas la spécification du nombre de clusters.
- Initialise un cluster initial qui contient toute la dataset, ensuite, divise itérativement cette dataset par le fractionnement de clusters, jusqu'à arriver à des clusters contenant un seul point de données.
- Pour choisir le cluster à diviser, la somme des carrés de l'erreur ($SSE = \sum_{i=1}^{n}(X_i - \bar{X})$) est calculée pour chaque cluster, ensuite, le cluster qui maximise cette valeur est sélectionné.
- Parmi les techniques exploitées pour la division du cluster sélectionné est le critère de Ward pour rechercher la plus grande réduction dans la SSE.

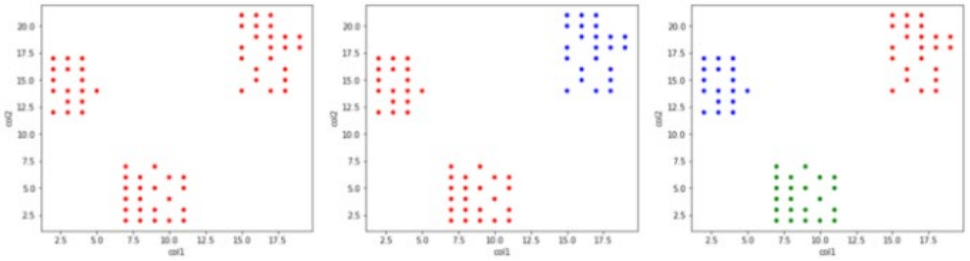


illustration d'un clustering descendant;

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

- Divise la base d'apprentissage en plusieurs groupes non hiérarchiques.
- La dataset est divisées en K groupes, où le nombre de groupes doit être prédéfini à l'avance (paramètres).
- Les clusters sont créés de telle sorte à minimiser la distance entre les points de données appartenant au même groupe.
- Ces techniques sont sensible aux conditions initiales (le paramètre K, et les centroïdes initialisés) et aux valeurs aberrantes.
- K-means est Parmi les algorithmes, les plus connu à base de centroïde.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means

- Algorithme d'apprentissage automatique non supervisé.
- Appartient aux méthodes de clustering à base de centroïde, où chaque cluster est associé à un centroïde.
- Le but principal de cet algorithme est de minimiser la distance entre les différents points de données et les centroïdes de leurs clusters.

L'algorithme K-means

1- Choisir le nombre K (des clusters).

2- Sélectionner aléatoirement K point de données en tant que centroïdes.

3- Assigner chaque point de données au centroïde plus proche.

4- Recalculer les nouveaux centroïdes des nouveaux clusters (la moyenne de chaque cluster).

5- Répéter les étapes 3 et 4.

6- Arrêter le traitement après stabilisation des clusters.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means

- La distance entre un point de données X et un centroïde C est calculé à base de la distance d'Euclide suivant :

$$D = \sqrt{\sum_{i=1}^{N}(X_i - C_i)^2}$$ N est la dimensionnalité des vecteurs X et C

- Le nouveau cluster C' est calculé suivant :

$$C' = \frac{1}{m}\sum_{j=1}^{m} X_j$$

m est le nombre de points dans le cluster, et X_j est un point de données appartenant au même cluster.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : comment choisir le nombre de clusters

- Parmi les techniques connues pour la sélection du K est la **méthode de Elbow**.
- Basée sur le concept de **Within Cluster Sum of Squares** (WCSS) qui définit la somme des variations à l'intérieur d'un cluster. WCSS est calculée suivant :

$$\sum_{P_i \in cluster1} distance\ (P_i, C_1)^2 + \sum_{P_i \in cluster2} distance\ (P_i, C_2)^2 + \sum_{P_i \in cluster3} distance\ (P_i, C_3)^2$$

$\sum_{P_i \in cluster_j} distance\ (P_i, C_j)^2$ est la somme du carré des distances (euclidienne ou autre) entre chaque point de données et son centroïde au sein d'un cluster C_j.

L'algorithme de la méthode de Elbow

1. Exécuter l'algorithme K-means pour différentes valeurs de K (de 1 à N).
2. Calculer la valeur $WCSS$ pour chaque K.
3. Tracer une courbe entre les valeurs WCSS calculées et le nombre de clusters K.
4. Le point de courbure pointu (forme de bras) est considéré comme la meilleure valeur de K.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : comment choisir le nombre de clusters

- Parmi les techniques connues pour la sélection du K est la **méthode de Elbow**.
- Basée sur le concept de **Within Cluster Sum of Squares** (WCSS) qui définit la somme des variations à l'intérieur d'un cluster. WCSS est calculée suivant :

$$\sum_{P_i \in cluster1} distance\ (P_i, C_1)^2 + \sum_{P_i \in cluster2} distance\ (P_i, C_2)^2 + \sum_{P_i \in cluster3} distance\ (P_i, C_3)^2$$

$\sum_{P_i \in cluster_j} distance\ (P_i, C_j)^2$ est la somme du carré des distances entre chaque point de données et son centroïde au sein d'un cluster C_j.

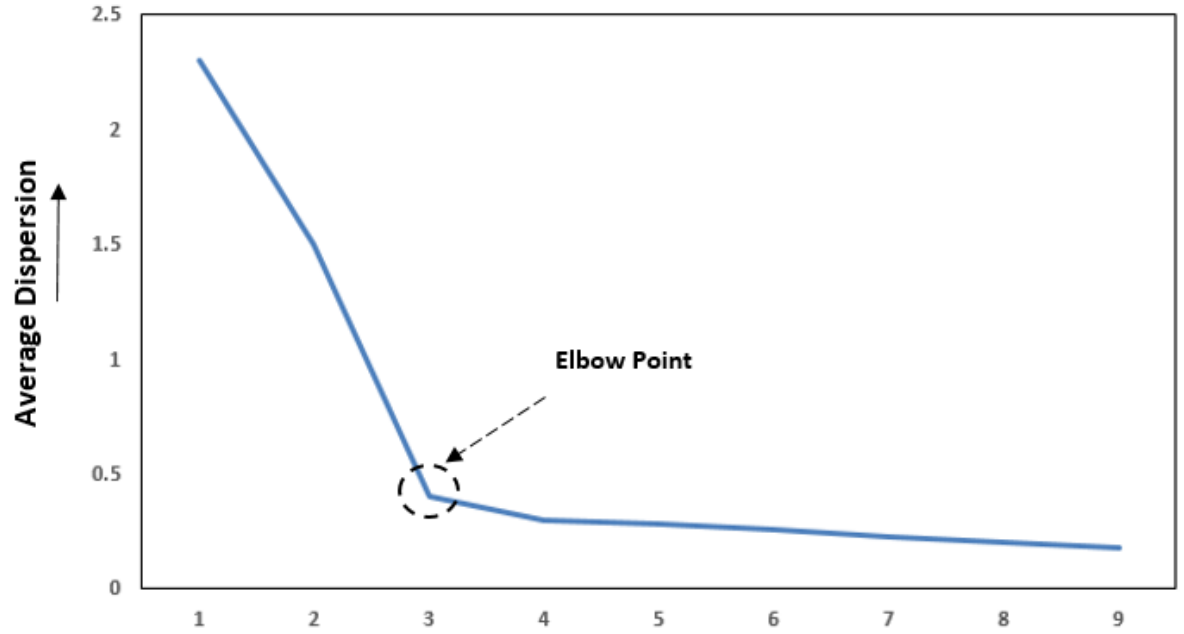
L'algorithme de la méthode de Elbow

1. Exécuter l'algorithme K-means pour différentes valeurs de K (de 1 à N).
2. Calculer la valeur $WCSS$ pour chaque K.
3. Tracer une courbe entre les valeurs WCSS calculées et le nombre de clusters K.
4. Le point de courbure pointu (forme de bras) est considéré comme la meilleure valeur de K.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : comment choisir le nombre de clusters

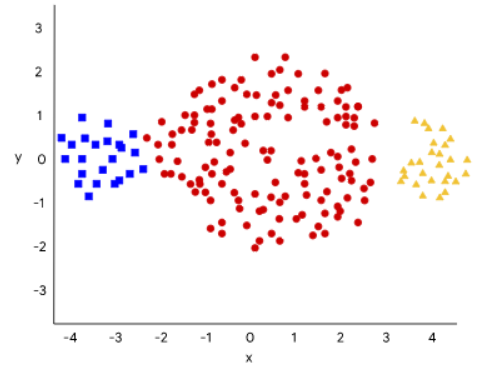


1. CLUSTERING

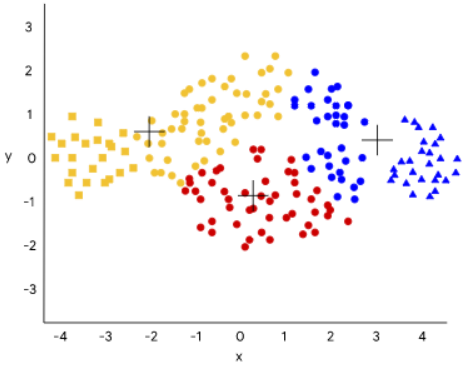
1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : inconvénients

- La sensibilité aux initialisations (K et centroïdes).
- Le choix manuel du K.
- Le problème de non-généralisation lors du regroupement des clusters caractérisés par différentes densités et tailles (solutions **: K-means Gaussian mixture models**). La figure à droite illustre les désavantages de K-means en cas de clusters de différentes densités.
- Sensible aux valeurs aberrantes (éliminer ces données avant le clustering).
- Sensible aux données caractérisées par une grande dimensionnalité (prétraitement : réduction de dimensionnalité à base de PCA).



Les clusters réels.



Les clusters prédits par K-means.

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : Exemple

Pour l'ensemble des données suivantes, déterminer les clusters avec K= 2, et les centroïdes initiaux (D2 et D4)

Points de données	W1	W2
D1	2	0
D2	1	3
D3	3	5
D4	2	2
D5	4	6

Itération 1 : la distance entre points de données D1, D3, C5 et les centroïdes D1, D2.

Centroïdes	D1	D2	D3	D4	D5
D2	3.17	0	2.83	-	4.25
D4	2	-	3.17	0	4.84

Cluster 1: (D1, D4) Cluster 2: (D2, D3, D5)
Les nouveaux centroids: C1 (2,1), C2 (2.67, 4.67).

1. CLUSTERING

1.3. Les Méthodes de clustering à base de centroides

1.3.1. K-means : Exemple

Itération 2 : la distance entre points de données D1, D2, D3, D4, D5 et les centroïdes C1, C2.

Centroïdes	D1	D2	D3	D4	D5
C1	1.0	2.24	4.13	1	5.39
C2	4.72	2.37	0.47	2.76	1.89

Cluster 1: (D1, D2, D4) Cluster 2: (D3, D5)
Les nouveaux centroids: C3 (1.67,1.67), C4 (3.5, 5.5).

Itération 3 : la distance entre points de données D1, D2, D3, D4, D5 et les centroïdes C3, C4.

Centroïdes	D1	D2	D3	D4	D5
C3 (1.67,1.67)	1.70	1.49	3.59	0.47	4.91
C4 (3.5, 5.5)	3.8	3.54	0.71	3.81	0.71

Cluster 1: (D1, D2, D4) Cluster 2: (D3, D5)

Les nouveaux centroïdes: C5 (1.67,1.67), C6 (3.5, 5.5).

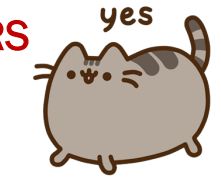
Les clusters (C3, C4) et (C5, C6) ont les mêmes valeurs, d'où la stabilisation des clusters.

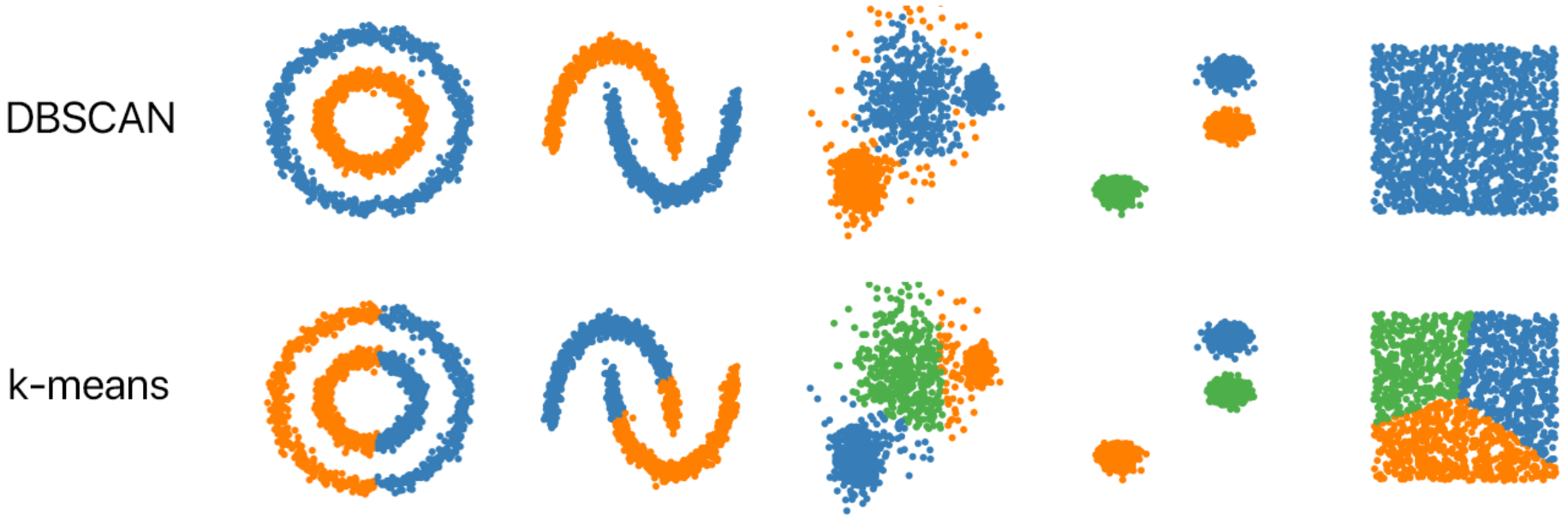
1. CLUSTERING

1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN)

YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !



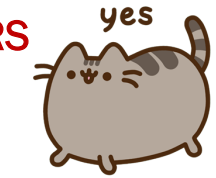


1. CLUSTERING

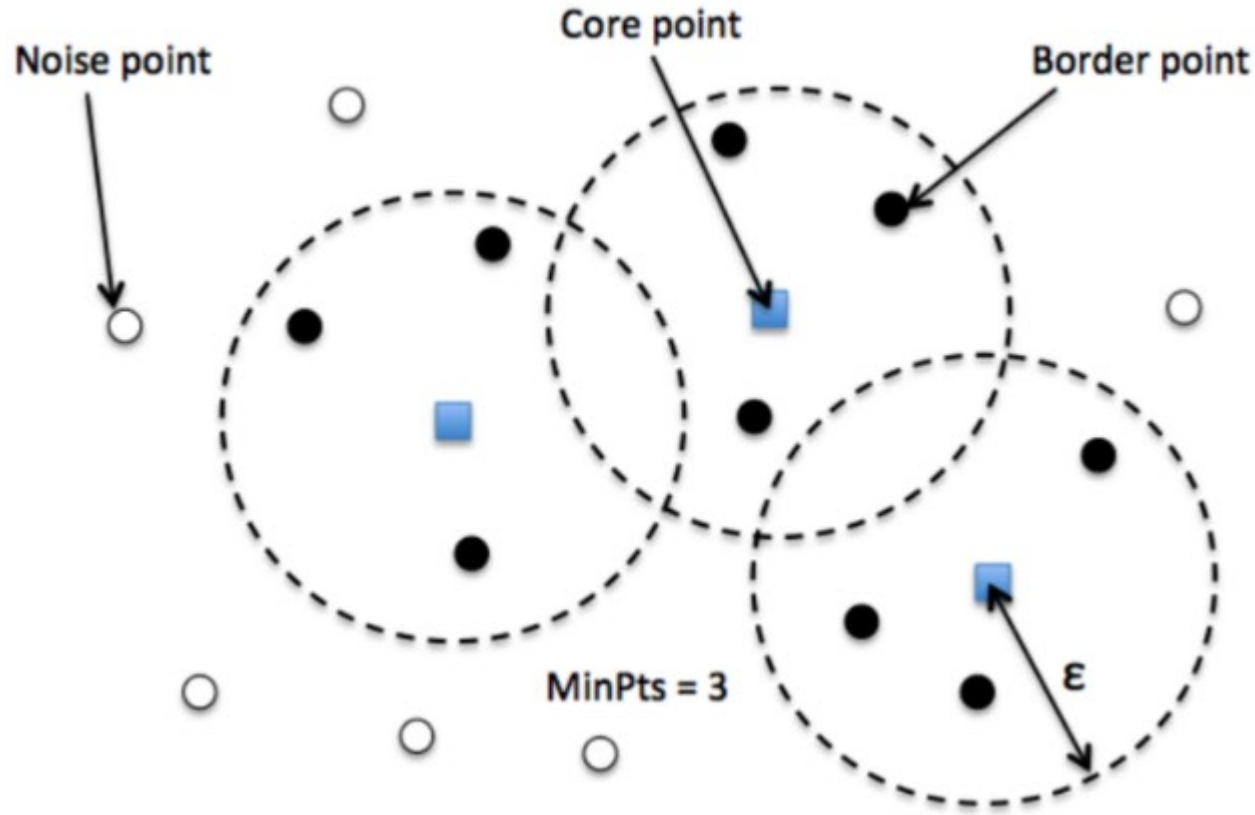
1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN)

YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !



- Algorithme d'apprentissage automatique non supervisé.
- Appartient aux méthodes de clustering à base de densité. Il s'agit d'un algorithme basé sur la densité. Il peut découvrir des clusters de différentes formes et tailles, qui contiennent à des valeurs bruité et aberrantes.
- L'algorithme DBSCAN utilise 2 paramètres : la distance ε et le nombre minimum de points « MinPts » devant se trouver dans un rayon ε.

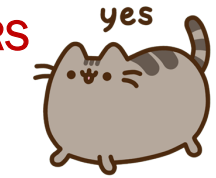


1. CLUSTERING

1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN)

YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !



- Algorithme d'apprentissage automatique non supervisé.
- Appartient aux méthodes de clustering à base de densité. Il s'agit d'un algorithme basé sur la densité. Il peut découvrir des clusters de différentes formes et tailles, qui contiennent à des valeurs bruité et aberrantes.
- L'algorithme DBSCAN utilise 2 paramètres : la distance ε et le nombre minimum de points « MinPts » devant se trouver dans un rayon ε.

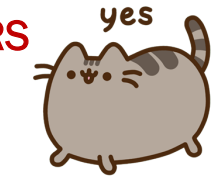
```
for each o ∈ D do
    if o is not yet classified then
        if o is a core-object then
            collect all objects density-reachable from o
            and assign them to a new cluster.
        else
            assign o to NOISE
```

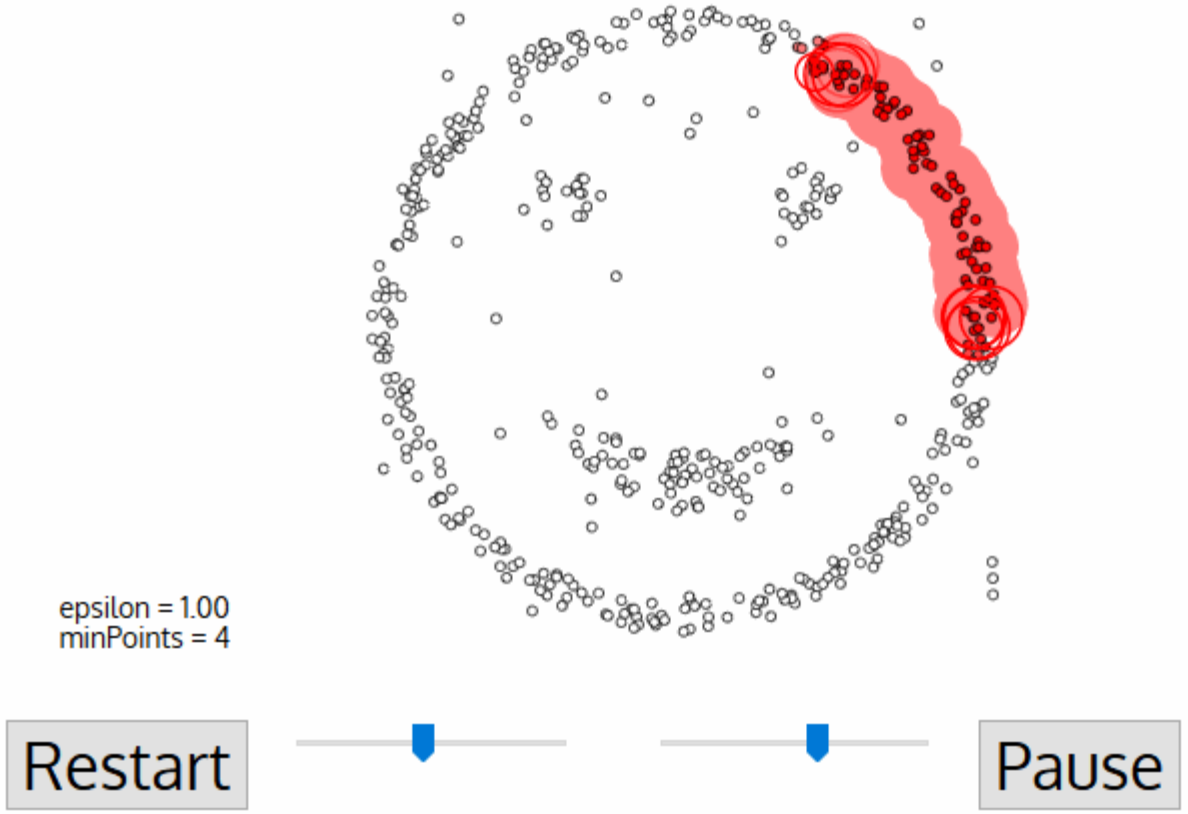
1. CLUSTERING

1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN)

YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !



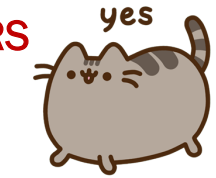


1. CLUSTERING

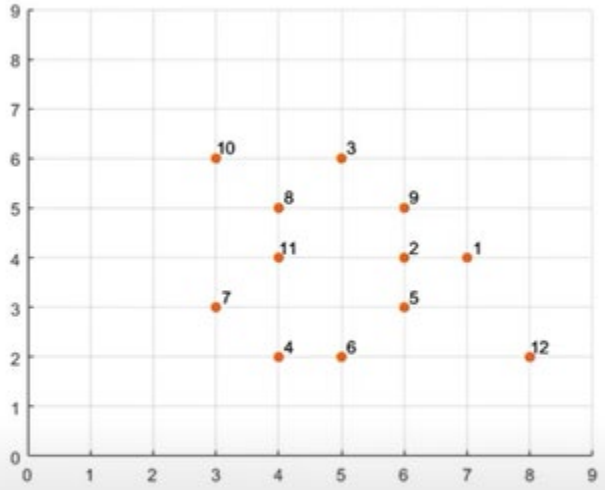
1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN) : Exemple

YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !



Point	X	Y
P1	7	4
P2	6	4
P3	5	6
P4	4	2
P5	6	3
P6	5	2
P7	3	3
P8	4	5
P9	6	5
P10	3	6
P11	4	4
P12	8	2



Eps=1.9 MinPts=4

P1	0.00											
P2	1.00	0.00										
P3	2.83	2.24	0.00									
P4	3.61	2.83	4.12	0.00								
P5	1.41	1.00	3.16	2.24	0.00							
P6	2.83	2.24	4.00	1.00	1.41	0.00						
P7	4.12	3.16	3.61	1.41	3.00	2.24	0.00					
P8	3.16	2.24	1.41	3.00	2.83	3.16	2.24	0.00				
P9	1.41	1.00	1.41	3.61	2.00	3.16	3.61	2.00	0.00			
P10	4.47	3.61	2.00	4.12	4.24	4.47	3.00	1.41	3.16	0.00		
P11	3.00	2.00	2.24	2.00	2.24	2.24	1.41	1.00	2.24	2.24	0.00	
P12	2.24	2.83	5.00	4.00	2.24	3.00	5.10	5.00	3.61	6.40	4.47	0.00
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12

Eps=1.9

MinPts=4

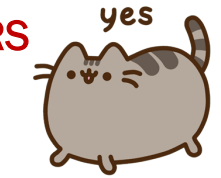
P1: P2,P5, P9	**P2:P1, P5, P9**	P3:P8, P9	P4: P6, P7
P5:P1,P2,P6	P6:P4, P5	P7: P4, P11	**P8:P3, P10, P11**
P9:P1, P2, P3	P10:P8	P11: P7, P8	P12:

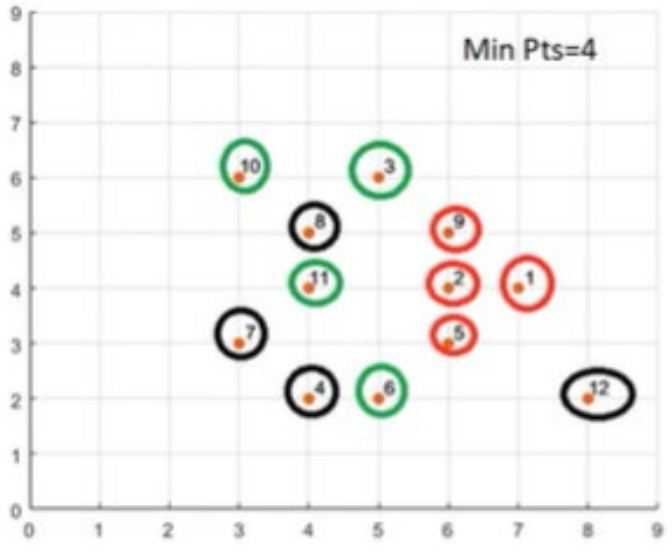
1. CLUSTERING

1.4. Les Méthodes de clustering à base de densité

1.4.4. Density-based spatial clustering (DBSCAN) : Exemple

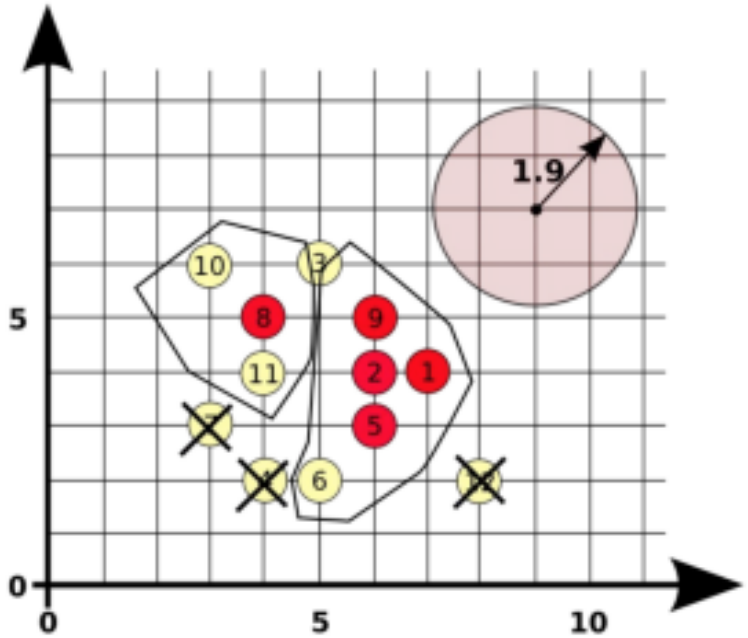
YOU DON'T NEED TO SPECIFY THE NUMBER OF CLUSTERS TRUST ME !





Point	Status	
P1	Core	
P2	Core	
P3	Noise	Border
P4	Noise	
P5	Core	
P6	Noise	Border
P7	Noise	
P8	Core	
P9	Core	
P10	Noise	Border
P11	Noise	Border
P12	Noise	

P1: P2,P5, P9	**P2:P1, P5, P9**	P3:P8, P9	P4:P6, P7
P5:P1,P2,P6	P6:P4, P5	P7:P4, P11	**P8:P3, P10, P11**
P9:P1, P2, P3	P10:P8	P11:P7, P8	P12:



3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

- Contrairement à l'apprentissage supervisé, les techniques d'apprentissage non supervisé n'exigent pas de diviser la dataset en base d'apprentissage et de test pour l'évaluation, car dans ces techniques, l'apprentissage ne nécessite pas une supervision, donc le risque de sur-apprentissage à cause des catégories de la base d'apprentissage n'est pas présent.
- Pour valider les résultats de l'apprentissage en clustering, deux techniques statistiques peuvent être exploitées : des techniques de validation interne, et des techniques de validation externe.

3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

3.1. La validation interne

- La majorité des méthodes de validation interne combinent la **cohérence** à l'intérieur du cluster et la **séparation** entre les différents clusters pour estimer le score de validation.

- Le principe consiste à calculer le score de validation de chaque cluster et les combiner d'une manière pondérée pour obtenir le score final du clustering. Étant donné un ensemble S de clusters $\{C_1 , C_2 , \dots \dots , C_n\}$ associé à un ensemble de poids $\{W_1 , W_2 , \dots \dots , W_n\}$, la mesure de validation de S est calculée suivant l'équation suivante, où W_k et le poids du cluster C_k.

$$Score_{validation(S)} = \sum_{k=1}^{n} W_k \; Score_{validation(C_k)}$$

- Parmi les mesures les plus utilisées en validation interne : **l'indice de Davies-Bouldin (à minimiser) et le coefficient de Silhouette (à maximiser).**

3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

3.1. La validation interne

- **L'indice de Davies-Bouldin :** est une mesure qui est basée sur le ratio entre les distances intra-cluster et inter-cluster. L'indice de Davies-Bouldin est calculé suivant l'équation suivante, où D_{ij} est le rapport de distance intra-et-inter cluster pour les cluster i, j, et $\bar{d}_i$ est distance moyenne entre chaque point du cluster i et son centroïde, et d_{ij} est la distance euclidienne entre les centroïdes des deux clusters.

$$\begin{cases} DB(C) = \dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \max_{i \leq k, j \neq i} D_{ij}\,, K = |C| \\ D_{ij} = \dfrac{(\bar{d}_i + \bar{d}_j)}{d_{ij}} \end{cases}$$

Exemple : étant donné 4 clusters C_0 , C_1 , C_2 , C_3 , le rapport de distance intra-et-inter cluster D est calculé entre les clusters 0 et 1, 0 et 2, 0 et 3, ensuite le maximum de ces valeurs est considéré. Ce rapport est calculé de la même manière pour les clusters 1, 2, et 3, enfin la moyenne des maximums définit l'indice de Davies-Bouldin.

3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

3.1. La validation interne

- **Le coefficient de Silhouette :** est une mesure de la similarité moyenne des objets appartenant à un cluster et de leur distance par rapport aux autres objets des autres clusters.

Premièrement, pour chaque point de données i, une distance moyenne a(i) du point i à tous les autres points appartenant au même cluster Ci est défini suivant l'équation suivante. Une grande valeur de a(i) signifie que le point de donnée i est différent de son cluster.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in c_i, i \neq j} d(i,j)$$

Ensuite, une distance moyenne b(i) du point i à tous les autres points du cluster voisin est définie suivant l'équation suivante :

$$b(i) = \min_{k \neq i} \left(\frac{1}{|C_i|} \sum_{j \in c_k} d(i,j) \right)$$

3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

3.1. La validation interne

Une grande valeur de b(i) signifie que le point de donnée i est différent de son cluster voisin. Pour calculer b(i), les distances moyennes du point i à tous les autres clusters sont calculées, ensuite la distance minimale (au cluster voisin) est prise en considération.

Enfin, le coefficient $S(i) \in [-1,1]$ de silhouette du point i et le coefficient de silhouette global S sont calculés suivant :

$$S(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$$

$$S = \frac{1}{N} \sum S(i)$$

Le coefficient de silhouette est plus informatif par rapport à l'indice de Davies-Bouldin, d'autre part, il est caractérisée par une complexité élevée.

3. APPRENTISSAGE NON SUPERVISÉ : CLUSTERING

3.2. La validation externe

- Contrairement aux méthodes de validation interne, les méthodes de validation externe exigent la connaissance préalable des vrais labels ou clusters.

- Dans ces techniques, l'ensemble S représente l'ensemble de clusters prédits {*C1 , C2 , ……, Cn*} et un autre ensemble de clusters qui représente les vrais labels est défini par P = {*P1 , P2 , ……, Pn*}. Dans cette situation, les mêmes mesures de classification peuvent être exploitées. La validation externe est déconseillée en cas de clustering, car il est impossible de savoir qu'un cluster Ck définit une classe Pk.

Références

[1] Hierarchical Clustering . 36-350, Data Mining. 17 September 2008 : https://www.stat.cmu.edu/~cshalizi/350/2008/lectures/08/lecture-08.pdf