Plan du Module

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Chapitre 1 : Introduction à l'apprentissage non supervisé
- Chapitre 2 : Mesures de similarité et de dissimilarités
- **Chapitre 3 : Clustering**
  - K-means, Partitionnement spectral
- Chapitre 4 : Réduction de dimension :
  - ACP
- Chapitre 5 : Evaluation des performances
- Chapitre 6 : Synthèse et Comparaison

Université Abdelmalek Essaadi

Faculté des sciences de Tétouan

Licence d'Excellence

Machine Learning et Intelligence Artificielle

Intitulé du Module :
Unsupervised Learning Algorithms

Chapitre 3 :
Classification non supervisée (clustering)

Pr. HARCHLI FIDAE

A.U : 2024 / 2025

Objectifs du chapitre

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Comprendre le concept de clustering, son importance en apprentissage non supervisé et ses types
- Découvrir les défis du clustering et savoir comment les surmonter
- Comprendre le fondement théorique de la méthode de k-means (centres mobiles)
- Analyser les avantages et les inconvénients de la méthode de k-means
- Découvrir les stratégies d'amélioration de la méthode de k-means ainsi que ses variantes

Plan

- Motivation
- Classification non supervisée
- Types de clustering
- Problématique
- Critères d’évaluation des partitions
- Méthode de K-means

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Motivation : Problème

- Imaginons une université qui souhaite améliorer l'accompagnement de ses étudiants en première année.

- **L'objectif** est d'identifier des groupes d'étudiants ayant des besoins similaires pour leur proposer un soutien adapté.

- Pour cela, elle collecte des données sur les étudiants : leurs résultats au test d'entrée, leur assiduité aux cours, leurs habitudes de travail (via des enquêtes ou des plateformes en ligne) et leur niveau d'engagement dans les activités parascolaires.

Motivation : Importance du clustering

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- En appliquant un algorithme de clustering sur ces données, l'université découvre trois groupes principaux :

  - **Les performants indépendants** : des étudiants avec de bons résultats, mais qui interagissent peu avec leurs camarades ou les enseignants.

  - **Les étudiants en difficulté** : des étudiants ayant de faibles résultats et une faible participation, nécessitant un suivi personnalisé.

  - **Les motivés collaboratifs** : des étudiants qui, bien que leurs résultats soient moyens, participent activement aux activités de groupe et montrent un potentiel de progression.

Motivation : Importance du clustering

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Impact :**
  - Le groupe 1 peut bénéficier d'ateliers pour les encourager à collaborer davantage et développer des compétences relationnelles.
  - Le groupe 2 peut être accompagné avec des sessions de tutorat et des ressources pédagogiques adaptées.
  - Le groupe 3 peut être encouragé à devenir mentor pour leurs pairs, renforçant leur propre apprentissage tout en aidant les autres.

Le clustering peut transformer une approche générique en une stratégie ciblée et efficace, favorisant ainsi la réussite académique et le développement personnel des étudiants.

Clustering : Définition

- La classification non supervisé (Clustering, segmentation, regroupement) est un outil très important en analyse exploratoire de données non étiquetées (aucune information antérieure).
- Elle Vise a regrouper les observations dans des clusters d'une homogénéité maximale de telle sorte que :
  - Les objets d'un même cluster sont très similaires
  - Les objets dans différents clusters sont très distincts.
- C'est un système d'analyse en clusters qui prend en entrée un ensemble de données et une mesure de similarité entre ces données et Produit en sortie un ensemble de partitions décrivant la structure générale de l'ensemble de données.

Le clustering permet de décrire de façon simple une réalité complexe via son résumé.

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Clustering : Structure des données à classer

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Soit une matrice rectangulaire dont :
  - Les lignes correspondent aux individus,
  - Les colonnes correspondent aux variables.

- Soit une matrice carrée de similarités, distances entre :
  - Individus,
  - Variables (par exemple : la matrice des corrélations).

- Ces structures permettent de classer les individus ou variables.

Clustering : Types

- Dans le clustering dur, les algorithmes attribuent une étiquette de classe $l_i \in \{1, 2, \ldots, k\}$ à chaque objet x_i pour identifier sa classe de cluster, où k est le nombre de clusters.
- En d'autres termes, dans le hard clustering, chaque objet est supposé appartenir à un et un seul cluster.
- Mathématiquement, le résultat des algorithmes de clustering dur peut être représenté par la matrice U

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k1} & u_{n2} & \cdots & u_{kn} \end{pmatrix}$$

où n désigne le nombre d'enregistrements dans l'ensemble de données, k désigne le nombre de grappes et u_{ij} satisfait

$$u_{ji} \in \{0, 1\}, \quad 1 \leq j \leq k, \; 1 \leq i \leq n,$$

$$\sum_{j=1}^{k} u_{ji} = 1, \quad 1 \leq i \leq n,$$

$$\sum_{i=1}^{n} u_{ji} > 0, \quad 1 \leq j \leq k.$$

Clustering : Types

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Dans le clustering flou, l'hypothèse est relâchée afin qu'un objet puisse appartenir à un ou plusieurs clusters avec des probabilités.
- Le résultat des algorithmes de clustering flou peut également être représenté par la matrice $U = (u_{ji})$:

$$u_{ji} \in [0, 1], \quad 1 \leq j \leq k, \; 1 \leq i \leq n,$$

$$\sum_{j=1}^{k} u_{ji} = 1, \quad 1 \leq i \leq n,$$

$$\sum_{i=1}^{n} u_{ji} > 0, \quad 1 \leq j \leq k.$$

Clustering : Structure des clusters obtenus

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Soit deux classes sont toujours disjointes (méthodes de partitionnement)
  - Généralement, le nombre de classes est défini à priori,
  - Certaines méthodes permettent de s'affranchir de cette contrainte (analyse relationnelle, méthodes paramétriques par estimation de densité).
- Soit deux classes sont disjointes ou l'une contient l'autre (méthodes hiérarchiques) :
  - Ascendantes (agglomération progressive d'éléments 2 à 2),
  - Descendantes.
- Soit deux classes peuvent avoir des objets en commun :
  - Analyse floue : chaque objet a une certaine probabilité d'appartenir à une classe donnée.

Clustering : Critères d'un bon algorithme

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Détecter les structures présentes dans les données,
- Permettre de déterminer le nombre optimal de classes,
- Fournir des classes bien différenciées,
- Fournir des classes stables vis-à-vis de légères modifications des données,
- Traiter efficacement les grands volumes de données,

Clustering : Critères de choix d'un bon modèle

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

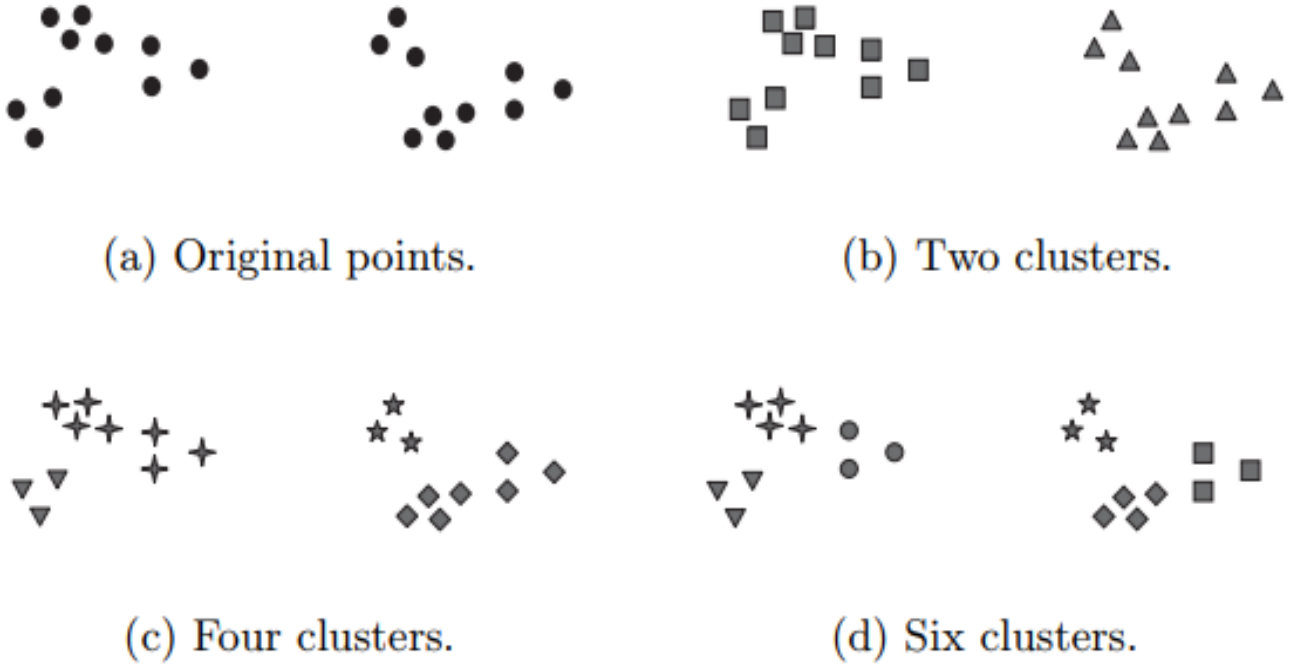
- Prise en compte de données de différents types : symbolique, numérique, matrice de similarité, données mixtes
  - Comment comparer des objets caractérisés à la fois par des attributs numériques et symboliques
- Minimisation du nombre de paramètres à fixer
  - Avec certaines méthodes, il est nécessaire de préciser le nombre de classes recherchées
- Insensibilité à l'ordre de présentation des exemples
  - Certaines méthodes ne génèrent pas les mêmes classes si l'ordre de parcours des données est modifié

Clustering : Critères de choix d'un bon modèle

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Résistance au bruit et aux anomalies
  - La recherche des points isolés est un sujet de recherche en soi
- Problème en grande dimensions
  - Certaines méthodes ne sont applicables que si les objets ou individus sont décrits sur deux ou trois dimensions
  - Construction et interprétation de clusters en grande dimension
- Passage à l'échelle
  - Certaines méthodes ne sont pas applicables sur de gros volumes de données.
- Interprétation et utilisation des résultats
  - Les utilisateurs doivent pouvoir donner un sens aux classes découvertes.
  - Comment un utilisateur peut il influencer un processus de catégorisation en introduisant des contraintes.

Problématique : Nombres de Clusters



(a) Original points.

(b) Two clusters.

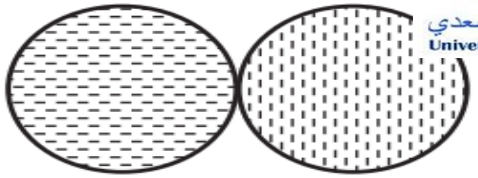
(c) Four clusters.

(d) Six clusters.

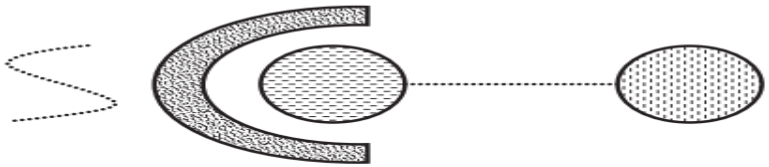
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Problématique : Types de Clusters

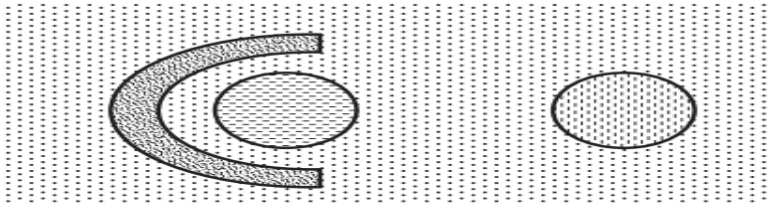
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.



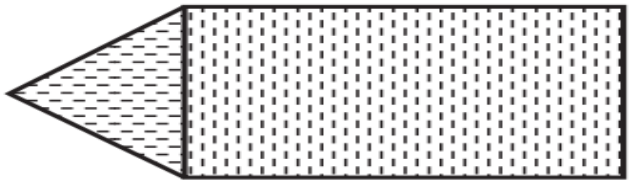
(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

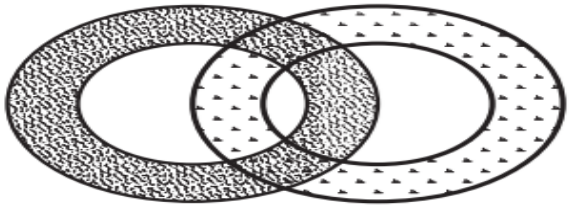


(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.





(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

Problématique : Types de données

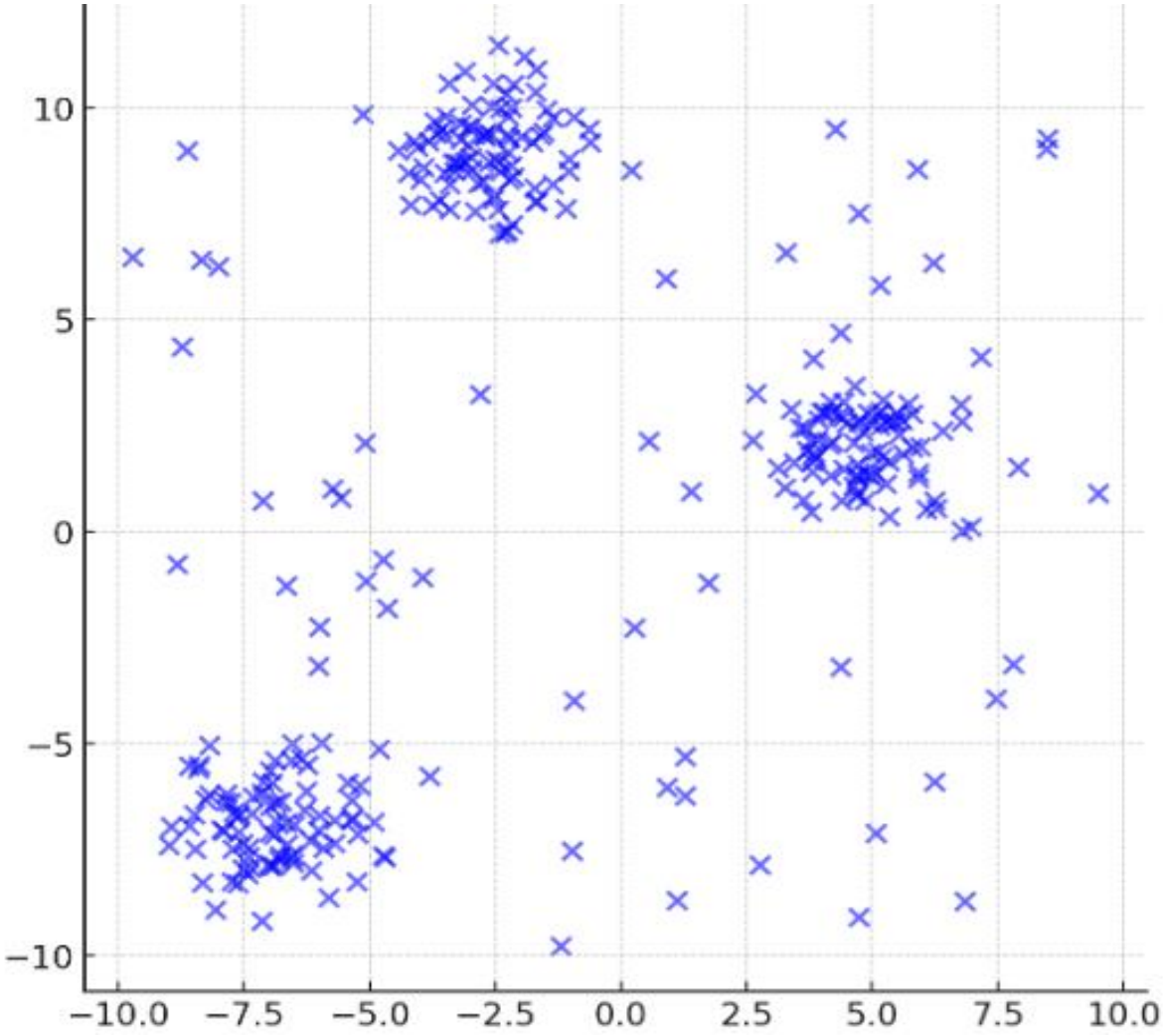
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Données bruitées

- Problème : Les valeurs aberrantes ou les données bruitées perturbent les algorithmes de clustering, car elles peuvent être attribuées à un cluster erroné ou créer des clusters inutiles. (Le point aberrant éloigné perturbe la répartition naturelle des clusters et fausse leur analyse.)
- Une correction du problème (par exemple, suppression ou traitement des aberrations) améliorerait significativement les résultats du clustering.

Client	Revenu annuel (k€)	Dépenses annuelles (k€)
Client 1	40	30
Client 2	50	35
Client 3	45	32
Client 4	60	40
Client 5	55	38
Client 6	200	5

Problématique : Types de données



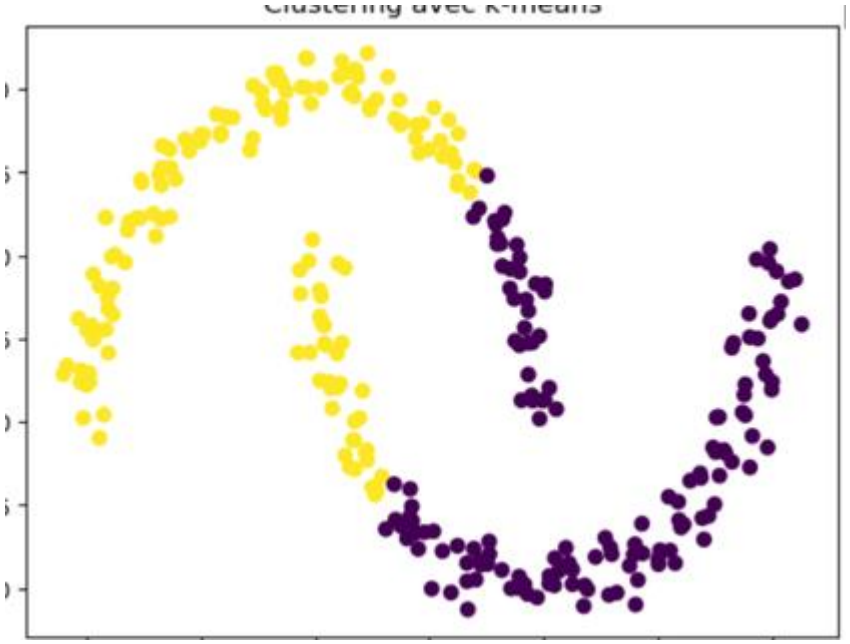
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

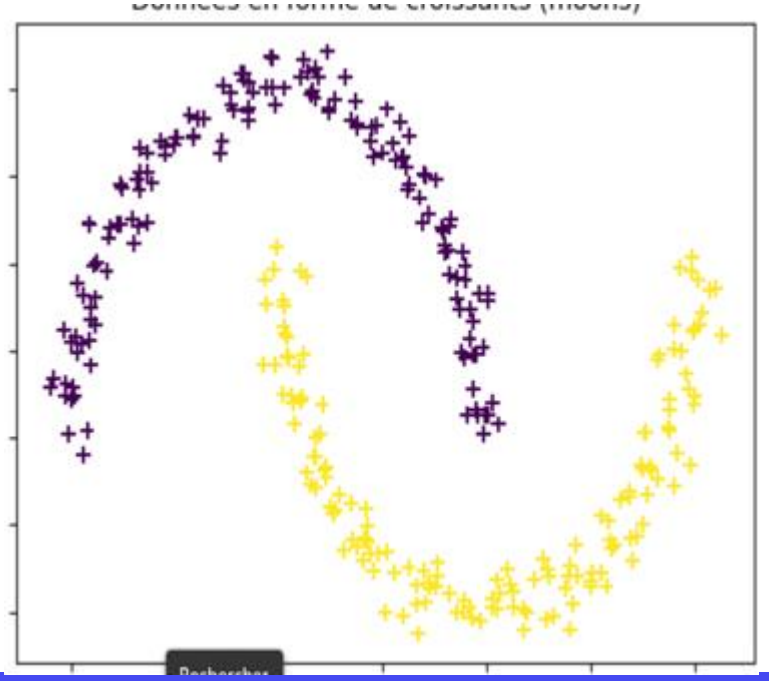
Problématique : Types de données

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

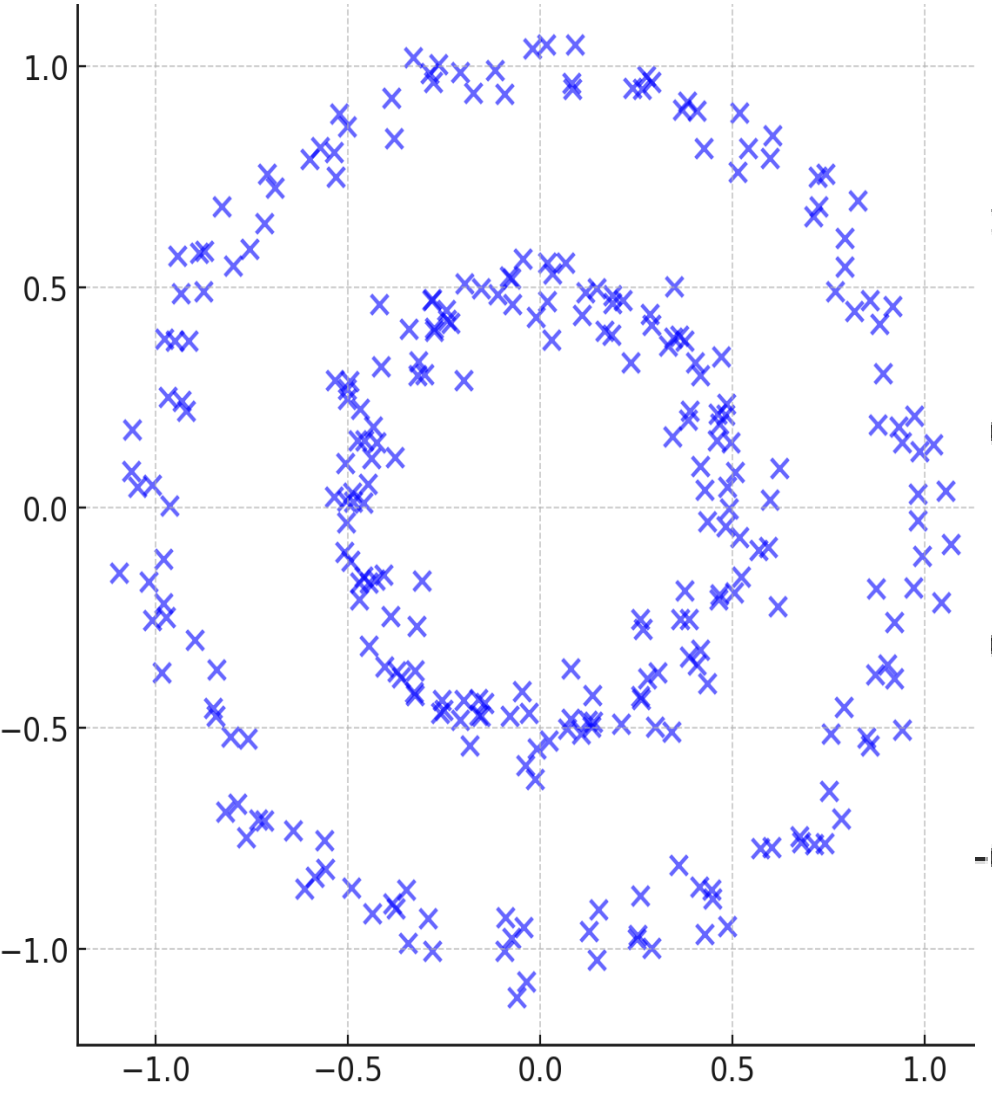
Données non linéairement séparables

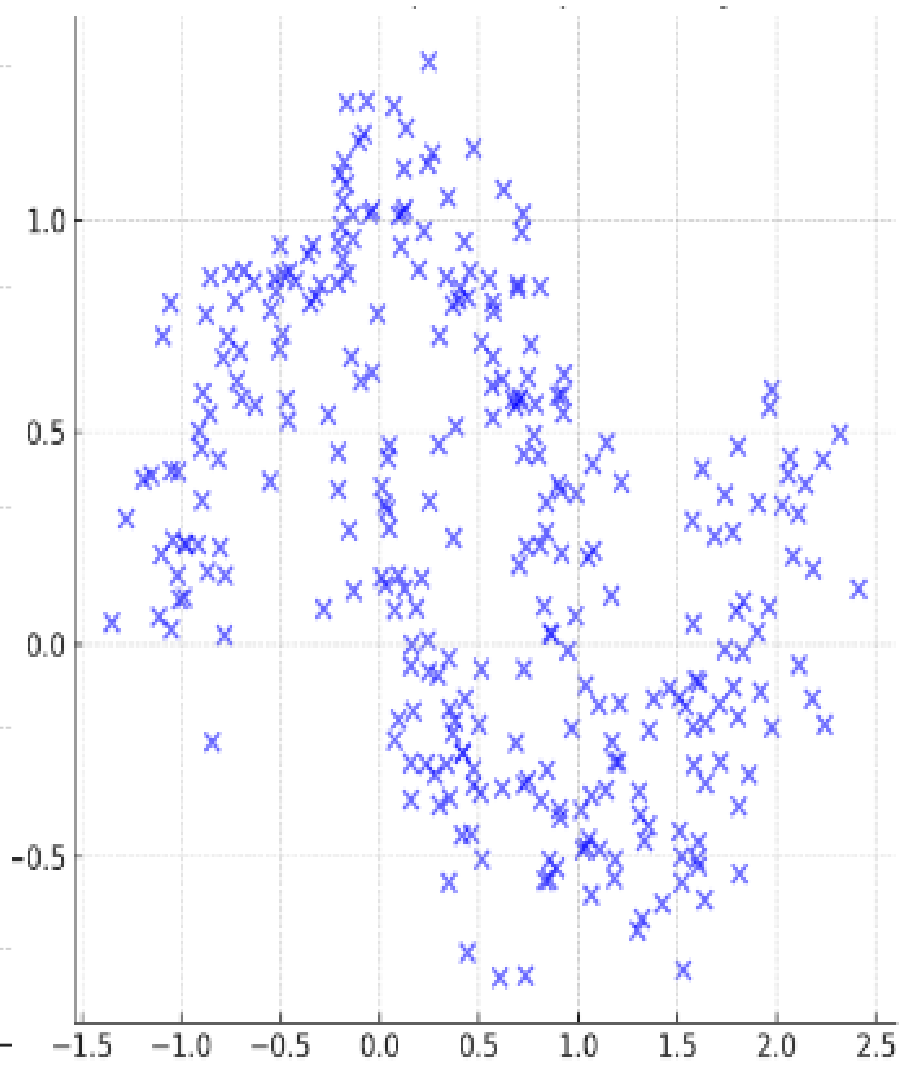
- Problème : Les données ne peuvent pas être séparées en clusters distincts par une frontière linéaire. Par exemple, des clusters en forme de cercles imbriqués ou de spirales.
- Exemple : Deux groupes d'étudiants, l'un vivant à proximité de l'université et l'autre plus éloigné, mais répartis de manière concentrique autour du campus.





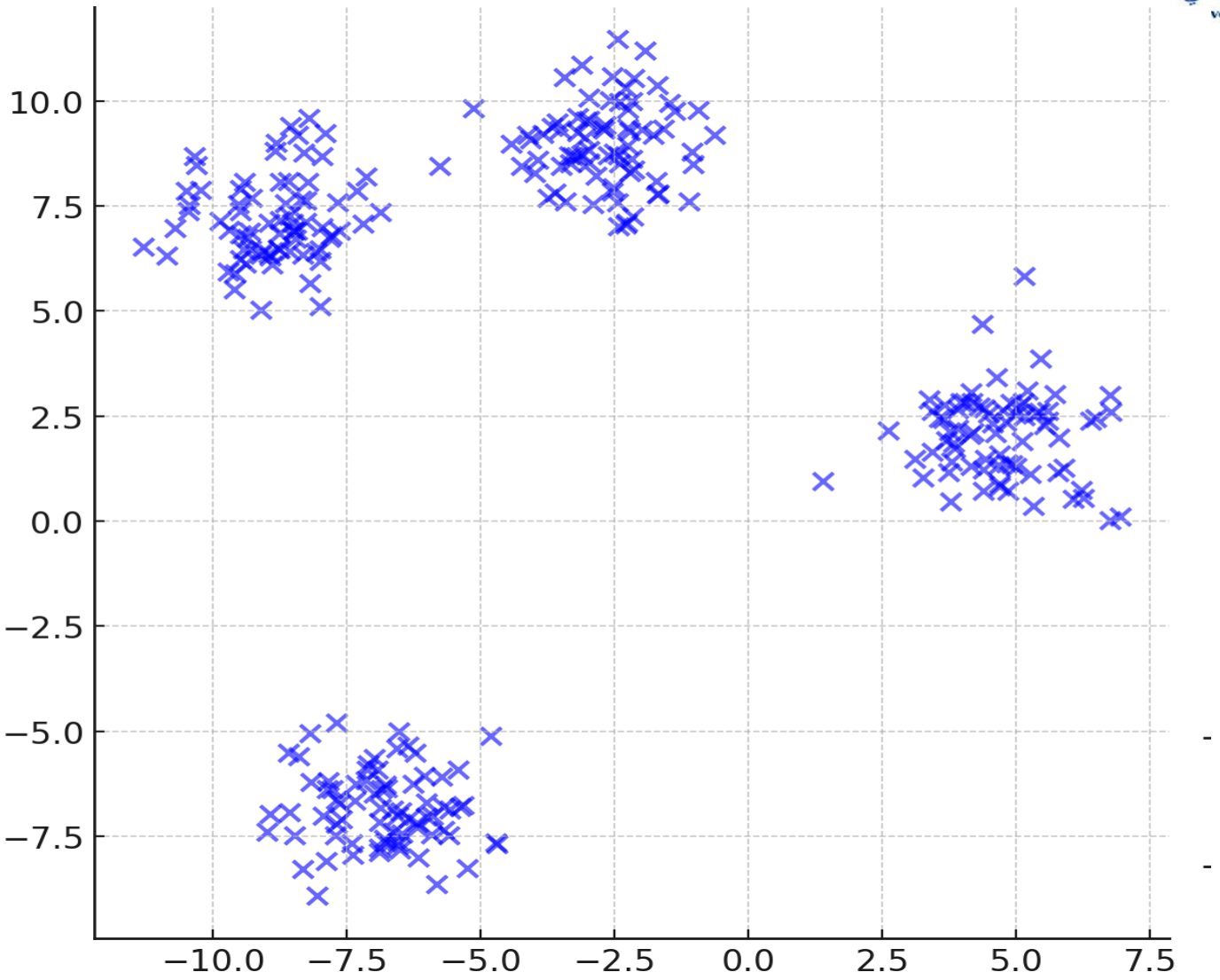
Problématique : Types de données





جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Problématique : Types de données (compact clusters)



جامعة عبد المالك السعدي
versité Abdelmalek Essaadi

Problématique : Types de données

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

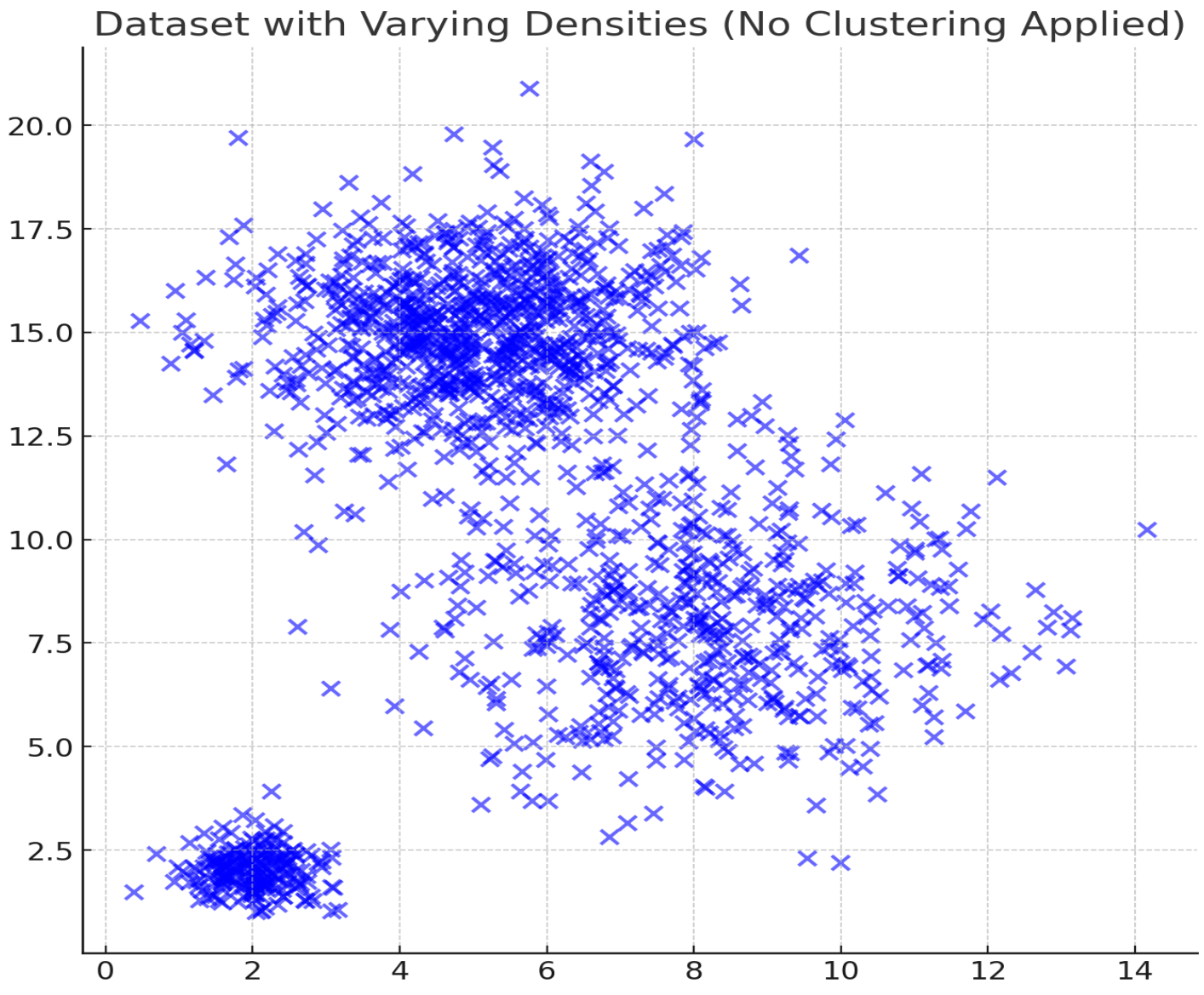
Données avec des densités variables

- Problème : Si les clusters ont des densités très différentes, les algorithmes comme K-Means ou la méthode des k-médoïdes échouent à séparer correctement les groupes.
- Exemple : Des étudiants regroupés selon leurs préférences de cours, où certains clusters contiennent des centaines d'étudiants (ex. informatique) et d'autres quelques dizaines (ex. arts).

Données à haute dimension (curse of dimensionality)

- En haute dimension, les distances entre les points deviennent moins significatives, rendant difficile la formation de clusters.
- Les clusters deviennent difficiles à distinguer et les algorithmes basés sur les distances produisent des résultats moins fiables.

Problématique : Types de données

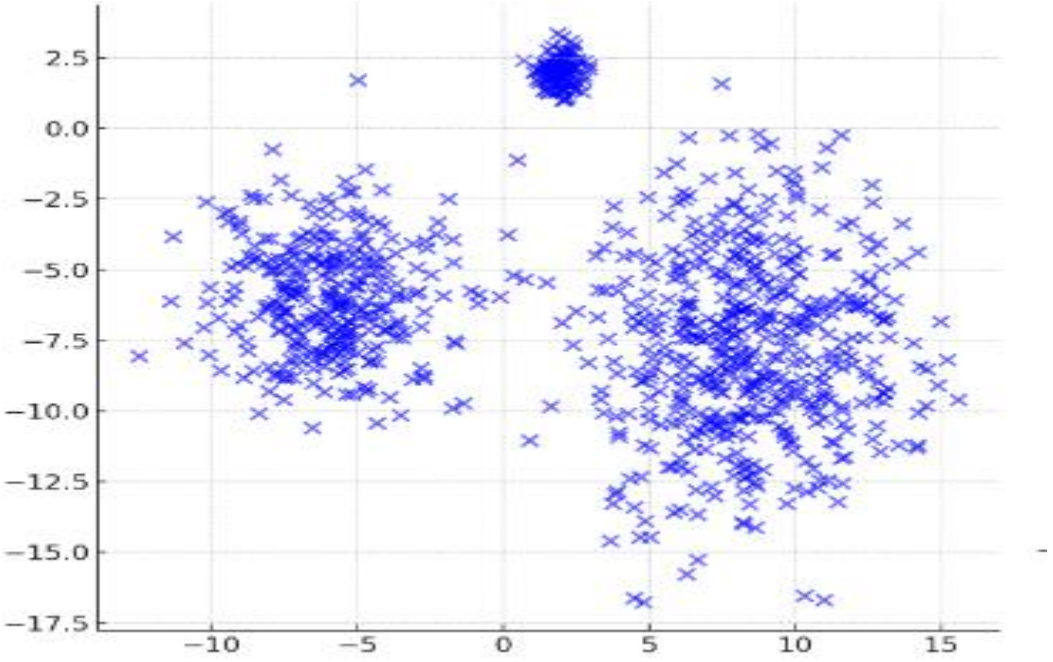


Problématique : Types de données

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Clusters de tailles inégales

- Problème : Les algorithmes comme K-Means supposent que les clusters ont des tailles similaires. Lorsque les clusters ont des tailles très différentes, les petits clusters risquent d'être ignorés.
- Exemple : Un groupe majoritaire d'étudiants dans une filière très populaire et quelques groupes minoritaires dans des filières spécialisées.



Problématique : Types de données

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Données mixtes (catégoriques et numériques)

- Problème : Les métriques standard comme la distance euclidienne ne sont pas adaptées pour les données contenant à la fois des valeurs numériques (ex. âges) et catégoriques (ex. filières).
- Exemple : Les données sur les étudiants contiennent leur âge, leur sexe, leur filière, et leur région d'origine.

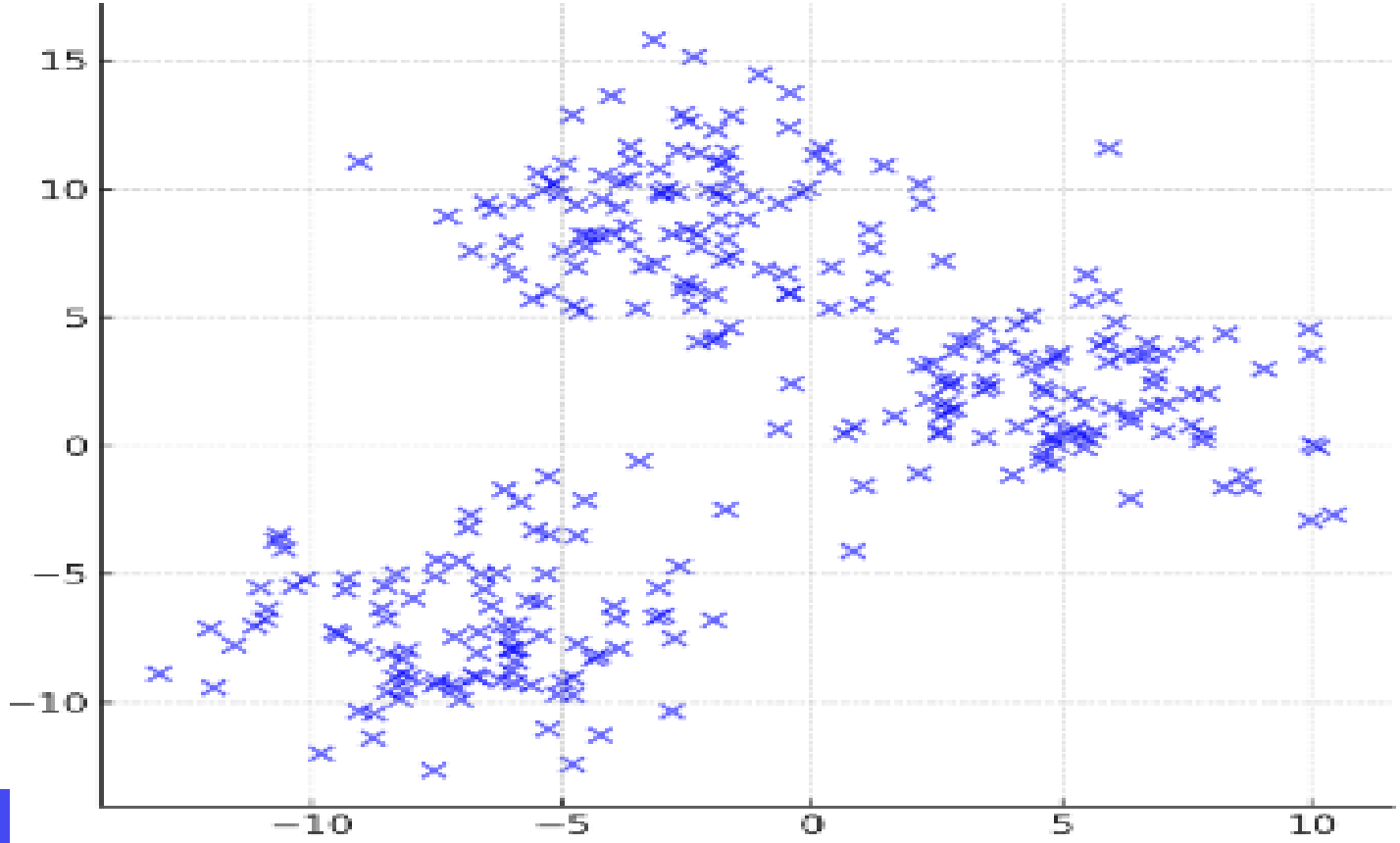
Données temporelles ou séquentielles

- Problème : Les points de données évoluent dans le temps, et le clustering doit tenir compte de cette dépendance temporelle.
- Exemple : L'évolution des résultats académiques des étudiants au fil des semestres.

Problématique : Types de données

Clusters qui se chevauchent

- Problème : Les clusters ne sont pas distincts et partagent des zones communes, ce qui rend leur séparation complexe.
- Exemple : Les étudiants intéressés par plusieurs disciplines, comme ceux qui étudient à la fois l'informatique et la biologie.



جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Problématique : Types de données (Synthèse)

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Ces types de données nécessitent souvent :

- **des prétraitements spécifiques,**
- **le choix d'un algorithme adapté et**
- **une validation rigoureuse des résultats pour s'assurer que les clusters trouvés ont une réelle signification.**

Problématique : Combien de façon de segmenter?

- **Formalisation**
  - Décomposer un ensemble X en sous-ensembles non vides tel que chaque élément $x \in X$ se retrouve dans un et un seul sous ensemble.
  - $X = \{x_1, x_2, \dots, x_N\}$ un ensemble de données
  - $p = (C_1, C_2, \dots, C_k)$ partition de X : $X = \bigcup C_i$ et souvent $\bigcap C_i = \emptyset$
- **Exemple**
  - $X = \{a, b, c\}$. Il existe 5 partitions :
  - $\{\{a\},\{b\},\{c\}\}, \{\{a,b\},\{c\}\}, \{\{a\},\{b,c\}\}, \{\{a,c\},\{b\}\}, \{\{a,b,c\}\}$
- Nombre de partitions d'un ensemble en sous-ensembles :
  - k=3, k=2, k=1

Problématique : Combien de façon de segmenter?

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Nombre de Stirling $S(n,k)$**
  - Equations de récurrence :
    - $S(n,k) = S(n-1,k-1) + k \times S(n-1,k)$
    - $S(0,0) = 1\ et\ \forall n > 0, S(n,0) = S(0,n) = 0$
  - Formulation explicite : $S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\, C_j^k\, j^n$
  - Où C_j^k est le nombre de combinaisons de j parmi k
- **Nombre total de partitions**
  - Nombre de Bell est le nombre de partitions d'un ensemble à n éléments distincts.
  - $\mathrm{B(n)} = \sum_{j=1}^{n} S(n,k)$

Problématique : Combien de façon de segmenter?

- Nombre de Bell $B(n) = \sum_{j=1}^{n} S(n,k)$

		k										
		1	2	3	4	5	6	6	8	9	10	Nb Bell
n	1	1										1
	2	1	1									2
	3	1	3	1								5
	4	1	7	6	1							15
	5	1	15	25	10	1						52
	6	1	31	90	65	15	1					203
	7	1	63	301	350	140	21	1				877
	8	1	127	966	1701	1050	266	27	1			4139
	9	1	255	3025	7770	6951	2646	428	35	1		21112
	10	1	511	9330	34105	42525	22827	5214	708	44	1	115266

Clustering : Problématiques

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Nature des observations : données binaires, textuelles, numériques, ... ?
- Présence des points aberrants et de changement d'échelle
- Notion de similarité (ou de dissimilarité entre observations)
- Définition d'un cluster
- Evaluation de la validité d'un cluster.
- Nombre de clusters pouvant être identifiés dans les données
- Quels algorithmes ?
- Comparaison de différents résultats de clustering.

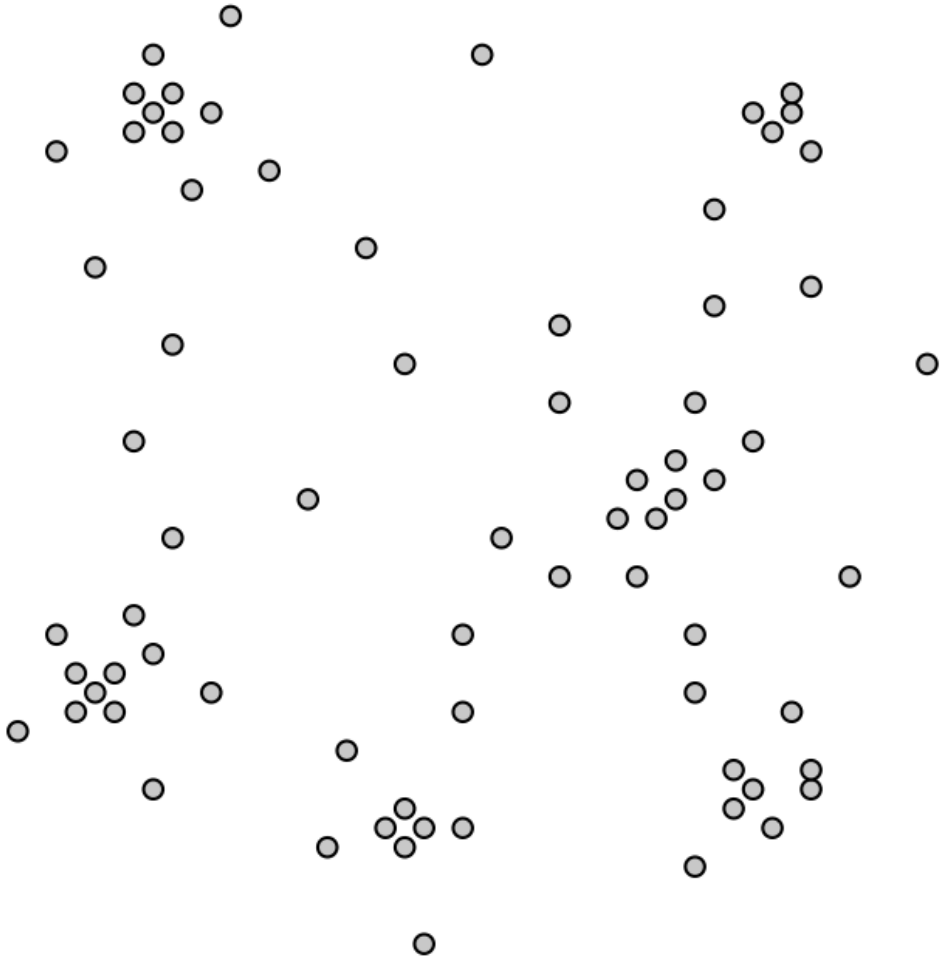
Méthode de k-means

K-means : Historique

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

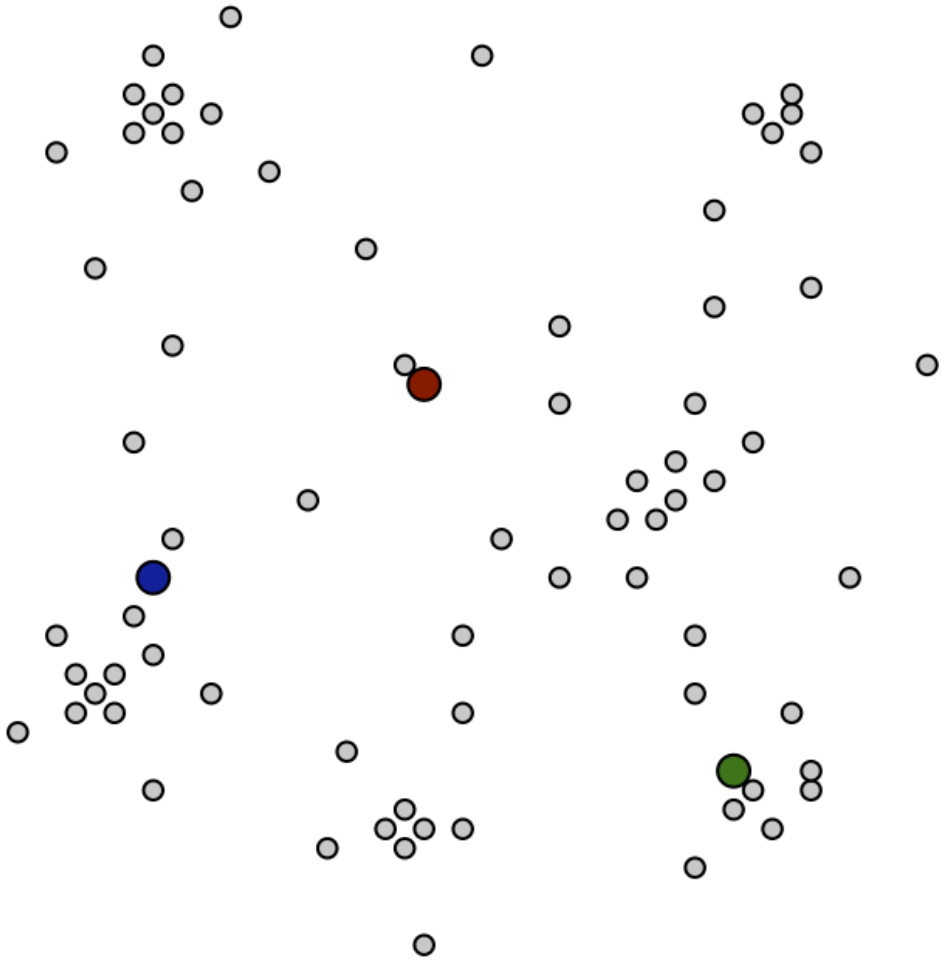
- La méthode k-moyenne est un algorithme non supervisé qui a été introduit par MacQueen en 1967.
- Il est l'un des plus simples algorithmes de classification automatique non supervisé des données qui se base sur l'erreur quadratique comme critère d'évaluation de la partition.
- L'objectif général de la méthode est d'obtenir la partition optimale qui, pour un nombre k fixé de classes, minimise l'erreur quadratique.
- Le principe de cet algorithme est de choisir un ensemble de centres bien déterminé à l'avance et de chercher **itérativement** la partition optimale.
- Chaque membre est affecté au centre le plus proche, après la réalisation des clusters, la moyenne de chaque cluster est recalculée, elle constitue les nouveaux représentants des clusters.
- L'algorithme s'arrête lorsqu'il aboutit à un état **stationnaire**, c'est-à-dire, aucun élément ne change de place.

K-means : Illustration



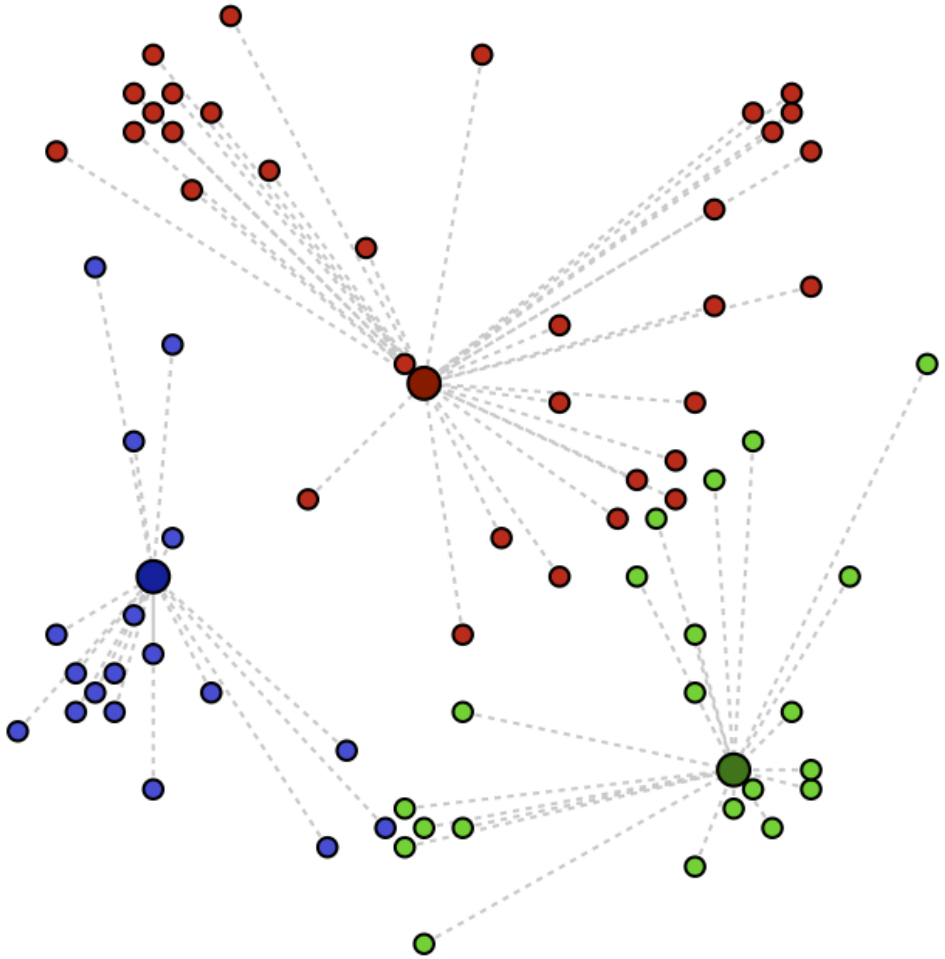
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



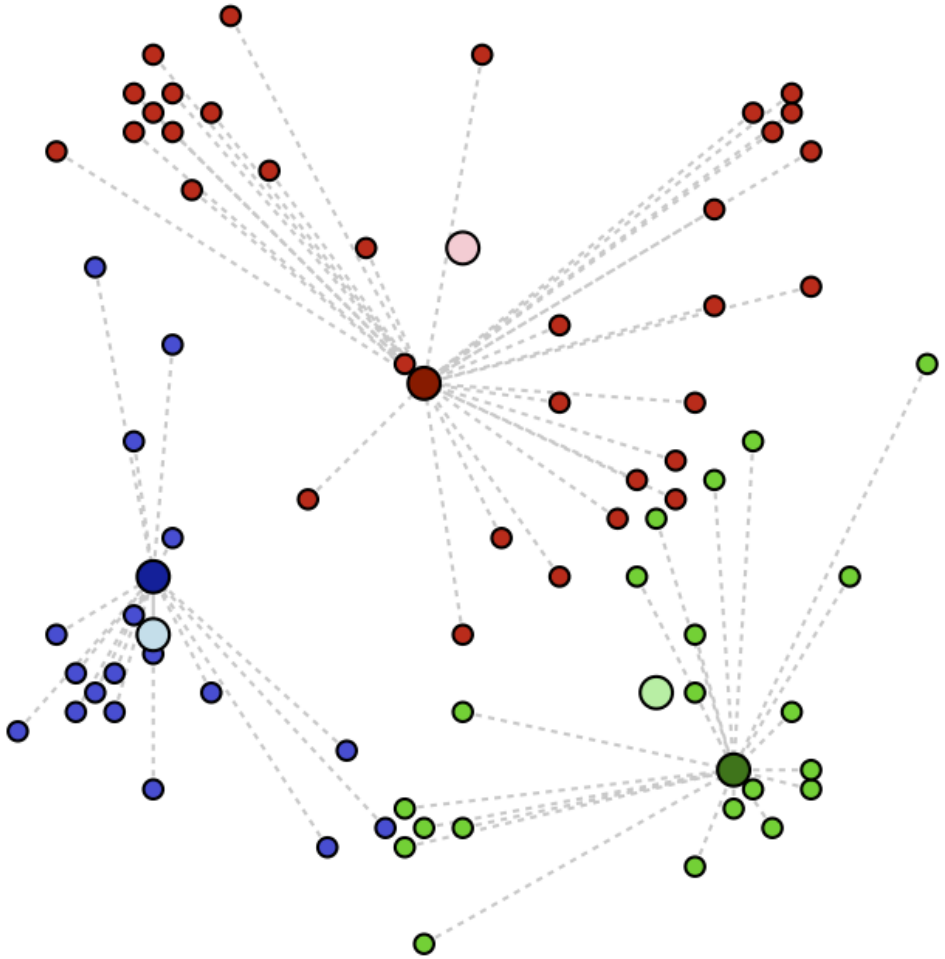
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



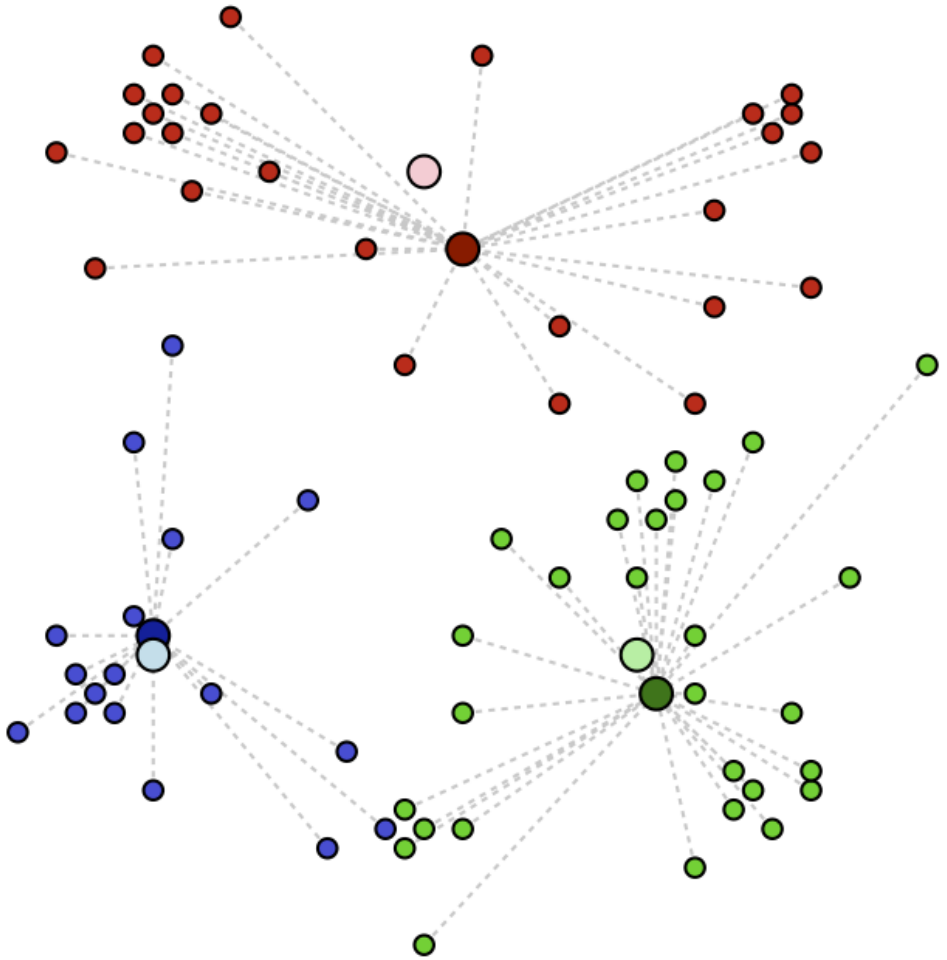
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



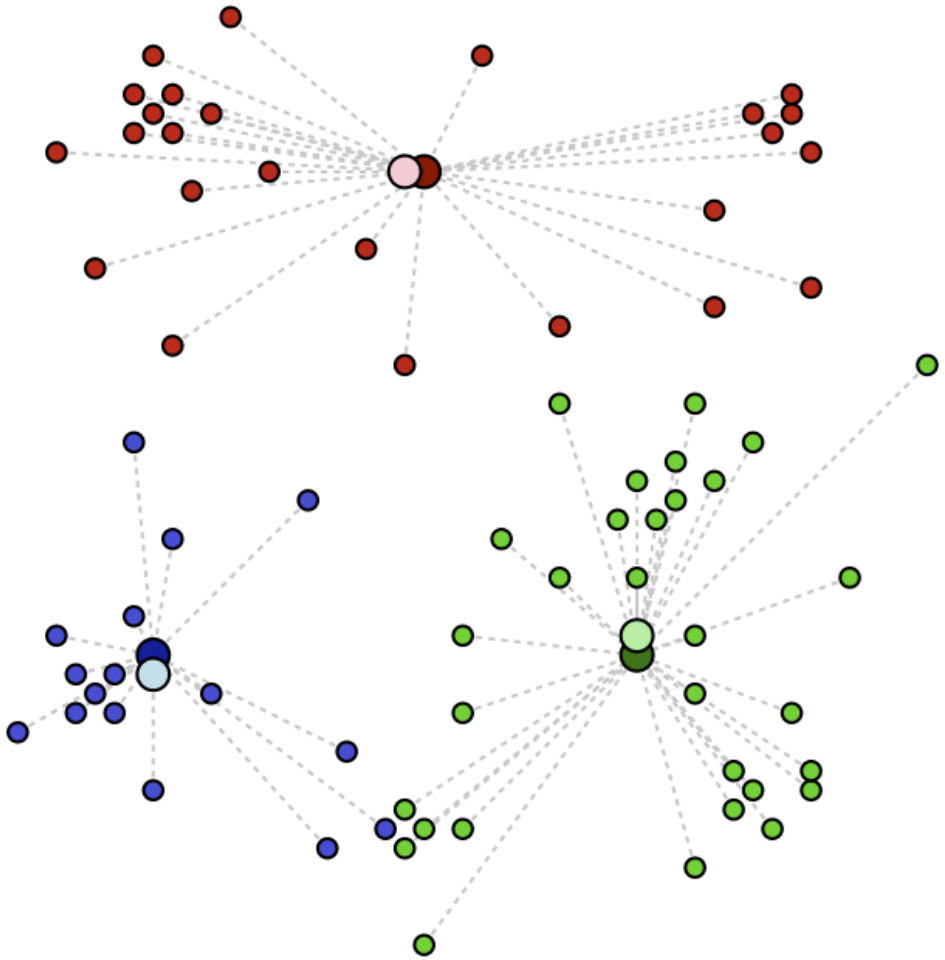
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



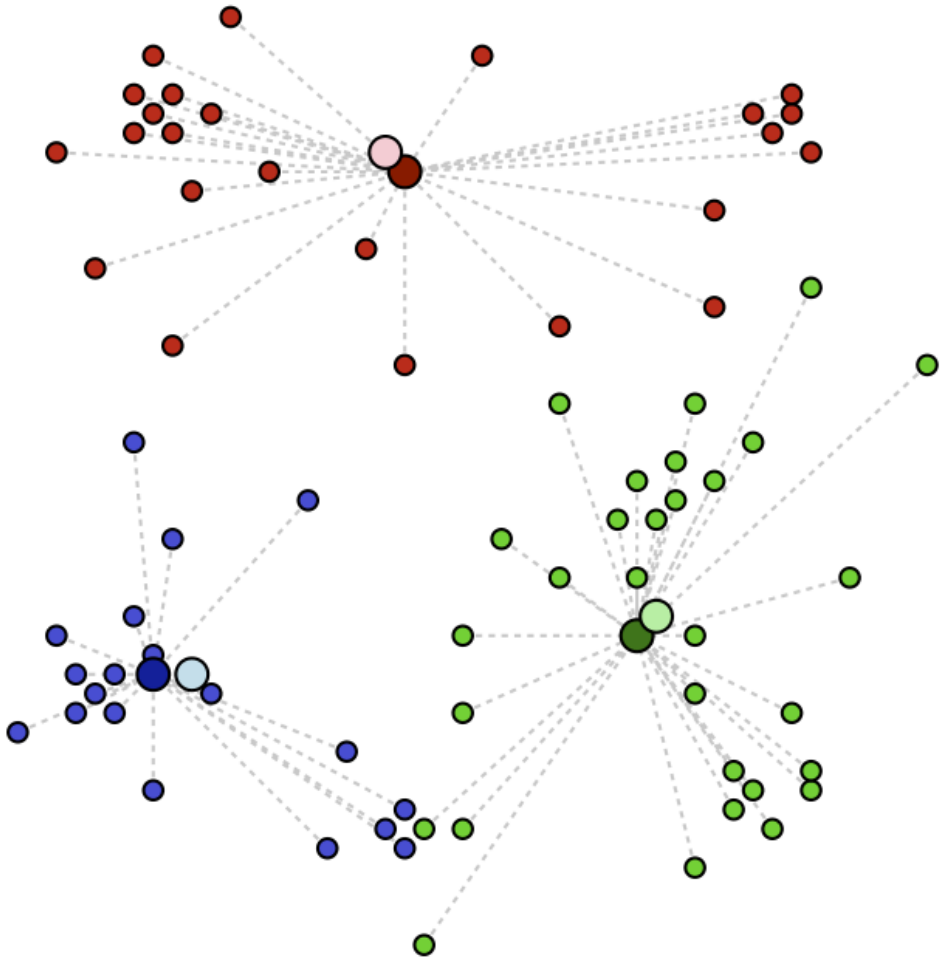
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



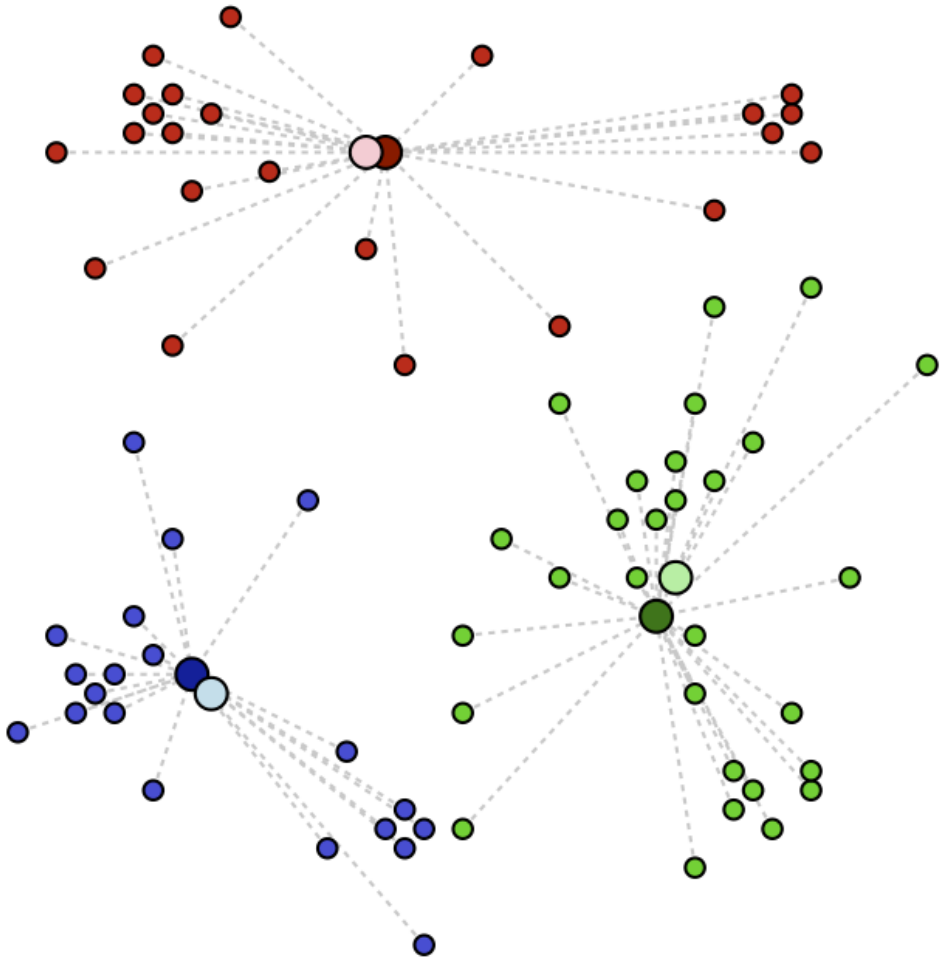
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



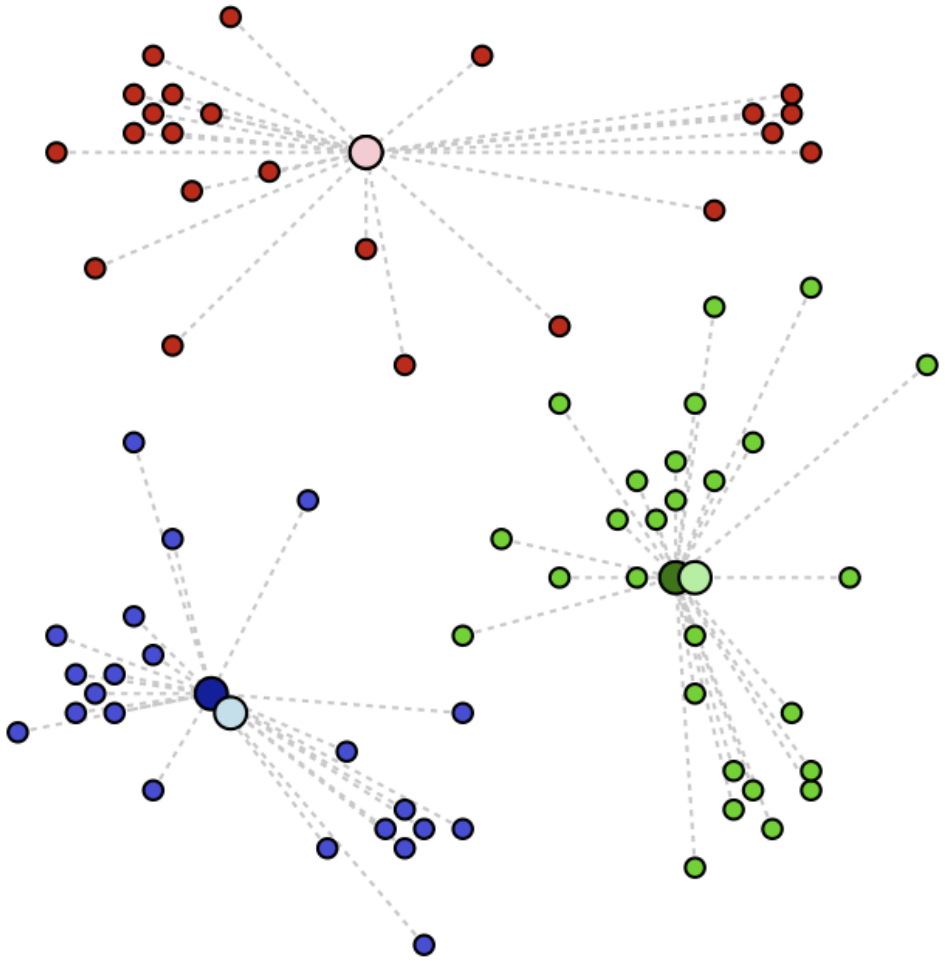
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



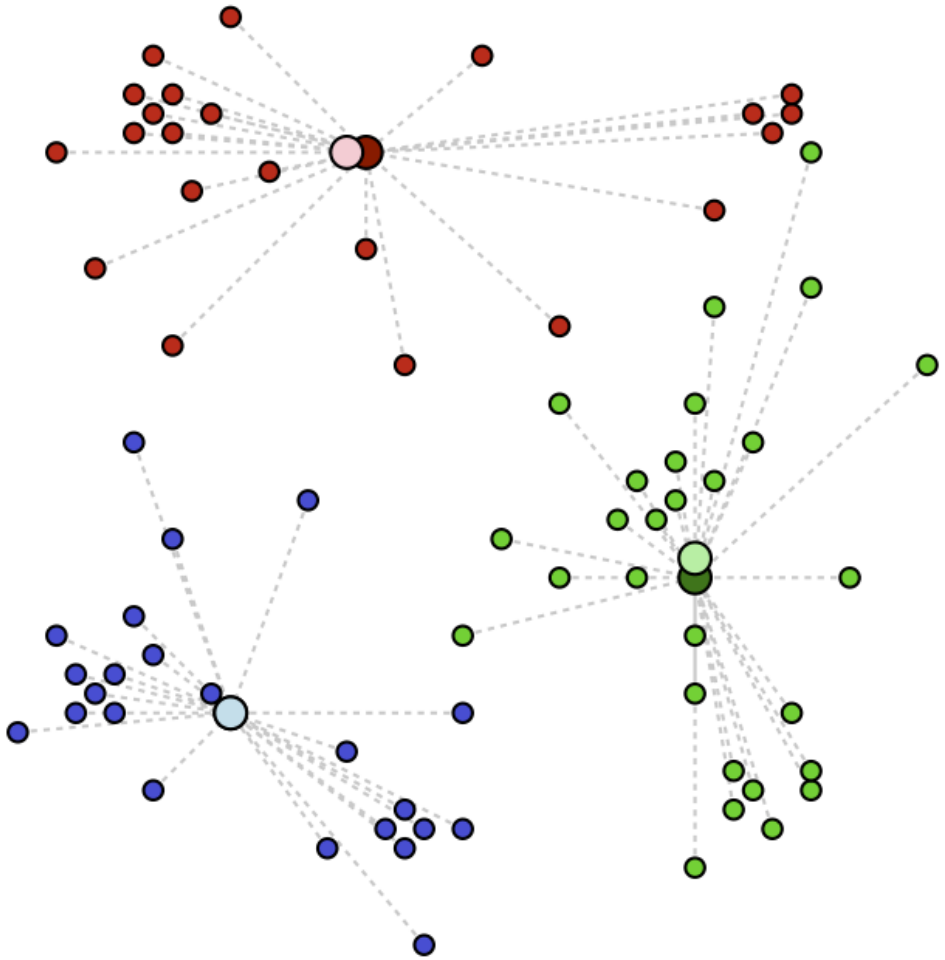
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



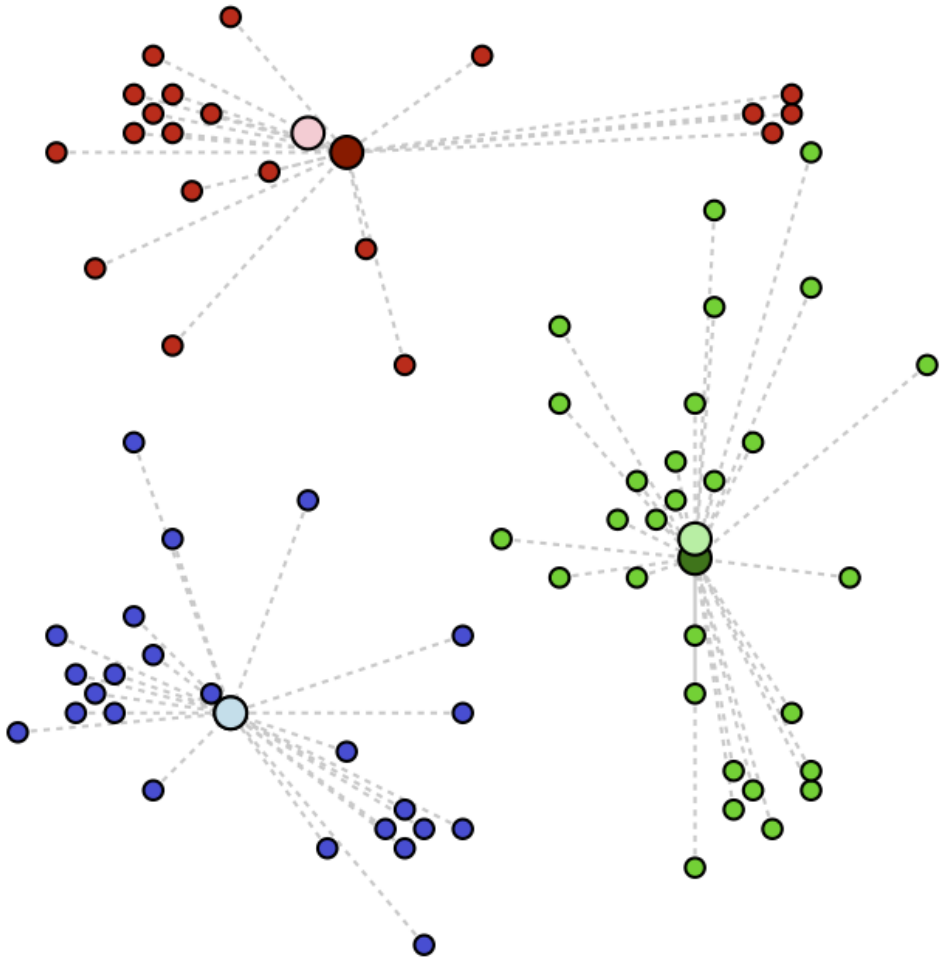
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



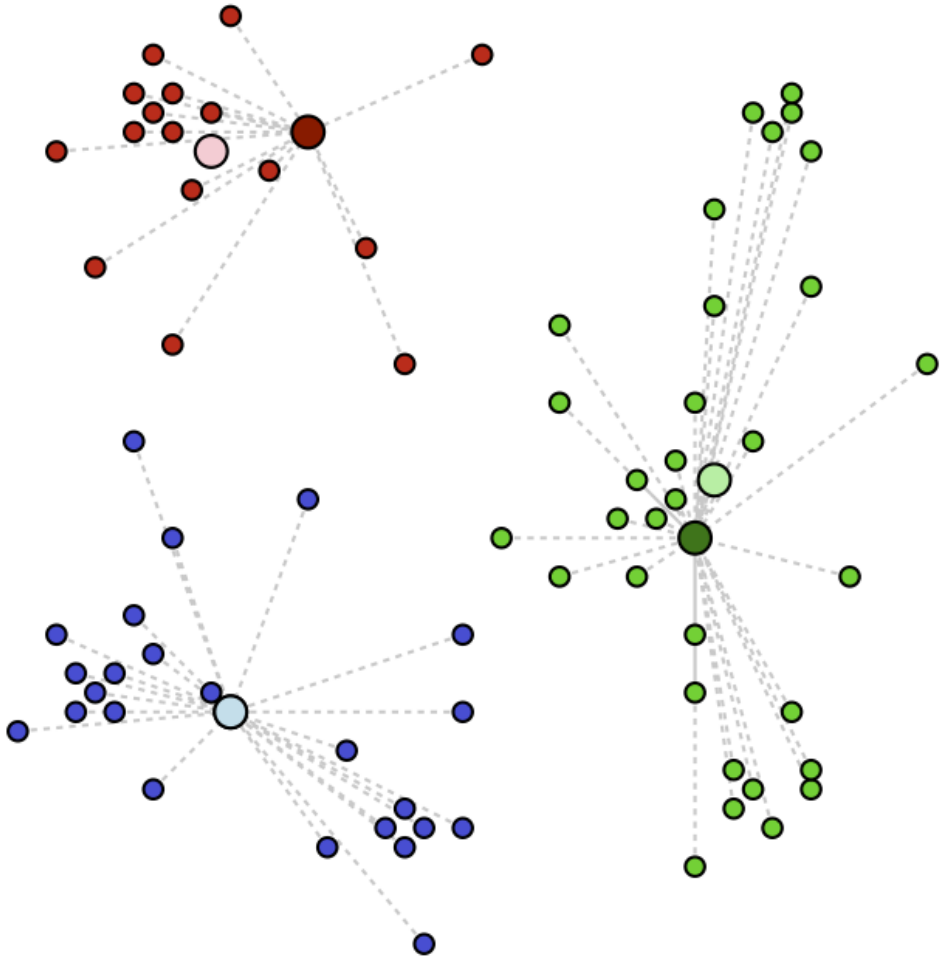
جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Illustration



جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

K-means : Modélisation du problème

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Données :**
  - A= $\{X_1, X_2, \dots, X_n\}$ l'ensemble des n points de données, où chaque point X_i appartient à un espace euclidien de dimension d ($X_i \in \mathbb{R}^d$).
  - K le nombre de cluster.
- **Variables :**
  - $W = \{W_1, \dots, W_k\}$: vecteur des centres tel que $W_i \in \mathbb{R}^d$
  - $\aleph(X_i)$: fonction d'affectation qui prend des valeurs entières :

$$\aleph : \mathbb{R}^d \to \mathbb{N}$$

$$X_i \to j \in \{1, \dots, k\}$$

  - u_{ij} : variable binaire telle que $u_{ij} = \begin{cases} 1 & \text{si } X_i \text{ est affecté au cluster j} \\ 0 & sinon \end{cases}$

K-means : Modélisation du problème

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Objectif :**

$$Min\ I(W, \aleph) = Min\left(\sum_{x_i \in A} \left\|X_i - W_{\aleph(x_i)}\right\|^2\right) = Min\left(\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}.\left\|X_i - W_j\right\|^2\right)$$

- **Contraintes :**

$$\sum_{i=1}^{n} u_{ij} > 0 \qquad \forall j = 1, \dots, k$$

$$\sum_{j=1}^{k} u_{ij} = 1 \qquad \forall i = 1, \dots, n$$

K-means : Modélisation finale

$$\begin{cases} Min\ I(W,\aleph) = Min\left(\sum_{x_i \in A} \|X_i - W_{\aleph(x_i)}\|^2\right) = Min\left(\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}.\|X_i - W_j\|^2\right) \\ \textbf{S. C.} \\ \sum_{i=1}^{n} u_{ij} > 0 \qquad \forall j = 1,\dots,k \\ \sum_{j=1}^{k} u_{ij} = 1 \qquad \forall i = 1,\dots,n \\ u_{ij} = \{0,1\} \quad \forall i = 1,\dots,n \ \forall\, j = 1,\dots,k \\ W \in \mathbb{R}^{d*k} \end{cases}$$

K-means : Résolution

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Le processus de minimisation de la fonction coût par la méthode de k-moyennes consiste à optimiser deux variables (vecteurs référents W et la fonction d'affectation $\aleph$) **alternativement**.
- L'algorithme procède d'une manière itérative et chaque **itération** comporte

  deux phases :

  - ✓ La $\mathbf{1}^{\grave{e}re}$ **phase** minimise la fonction objectif $I(W, \aleph)$ par rapport au paramètre $\aleph$ en supposant les valeurs des **centres fixées** aux valeurs calculées précédemment. Dans cette étape, on génère une **nouvelle partition** en calculant la **nouvelle fonction d'affectation** .
  - ✓ La **seconde phase** suppose que la fonction **d'affectation** est **fixée** à la valeur qui vient d'être calculée et **minimise** la fonction objectif par rapport au paramètre W.
- Cette procédure fait **décroître** la valeur de $I(W, \aleph)$ à chaque itération.

K-means : Résolution

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- La méthode de k-moyennes détermine les variables en minimisant la **fonction objectif** suivante :

$$I(W, \aleph) = \sum_{X_i \in A} \left\| X_i - W_{\aleph(X_i)} \right\|^2$$

- On a $\aleph(X_i) = c \Rightarrow X_i \in P_c$ (la classe c) alors :

$$I(W, \aleph) = \sum_{c=1}^{k} \sum_{X_i \in P_c \cap A} \|X_i - W_c\|^2$$

$$I(W, \aleph) = \sum_{c=1}^{k} \sum_{X_i \in P_c} \|X_i - W_c\|^2$$

K-means : Résolution

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- On pose : $\mathrm{I_c} = \sum_{x_i \in P_c} \|X_i - W_c\|^2$
- Alors on aura :

$$\boldsymbol{I(W, \aleph) = \sum_{c} I_c}$$

- L'expression I_c représente :
  - L'inertie locale, par rapport au référent W_c, des observations de l'ensemble d'apprentissage A qui lui sont affectées; ces observations appartiennent donc au sous ensemble P_c.
  - L'erreur de quantification obtenue quand on désire remplacer les observations de P_c par le référent W_c qui les représente.
- La quantité $I(W, \aleph)$ que l'on cherche à minimiser représente la somme des inerties locales I_c.

K-means : Phases d'apprentissage

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

Phase d'affectation :

- C'est la première phase de l'itération t qui consiste à minimiser $I\ (W, \aleph)$ par rapport à la fonction d'affectation $\aleph$, et fixer la première variable (l'ensemble des centres W) aux valeurs calculées précédemment.

- La minimisation par rapport à la partition s'obtient en affectant chaque observation X_i au référent W_c selon la nouvelle fonction d'affectation :

$$\aleph(X_i) = argmin_{c=1,\dots,k} \|X_i - W_c\|^2$$

- Cette nouvelle fonction d'affectation définit une nouvelle partition P de l'ensemble de données, où chaque observation X_i est affectée au plus proche référent W_c.

- En affectant chaque observation X_i au référent le plus proche W_c, on réduit le terme correspondant à X_i dans la fonction objectif.

K-means : Phases d'apprentissage

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- **Phase de minimisation :**
- A la seconde phase de l'itération t, la fonction coût décroit à nouveau en fonction de l'ensemble des vecteurs référents et en fixant la deuxième variable (fonction d'affectation) aux valeurs trouvées précédemment.

- La fonction coût $I\ (W, \aleph)$ est quadratique et convexe par rapport à la variable W alors elle atteint le minimum global pour:

$$\frac{\partial I}{\partial W}(W, \aleph) = 0 \Rightarrow \begin{bmatrix} \frac{\partial I}{\partial W_1} \\ \frac{\partial I}{\partial W_2} \\ \vdots \\ \frac{\partial I}{\partial W_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

K-means : Phases d'apprentissage

جامعة عبد المالك السعدي
Université Abdelmalek Essaadi

- Le calcul du vecteur gradient associé à tout référent W_c permet d'obtenir un ensemble d'équations vectorielles :

$$\forall c = 1 \dots k \qquad \sum_{X_i \in P_c} (X_i - W_c)$$

- Ces équations définissent les k nouveaux centres, et la mise à jour au cours de l'algorithme se fait selon cette formule :

$$\forall c = 1 \dots k \qquad W_c = \frac{\sum_{X_i \in P_c} X_i}{|P_c|}$$

où $|P_c|$ représente le nombre d'éléments de la classe C de la partition P.

- D'après cette formule les référents W_c sont alors les centres de gravité des observations de l'ensemble P_c.