Introduction à la fouille de données

M. Ledmi
m_ledmi@esi.dz

Département d'Informatique Khenchela

2020/2021

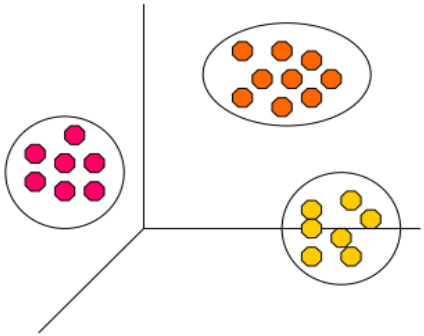
Plan

Vous êtes ici

Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

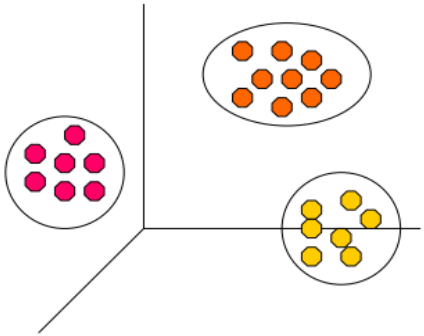
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment.
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

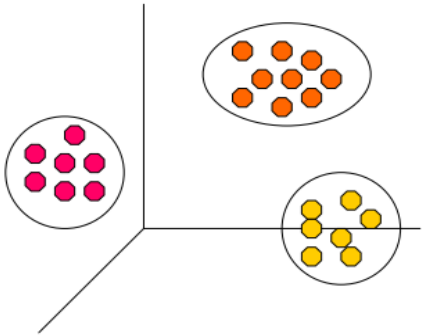
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

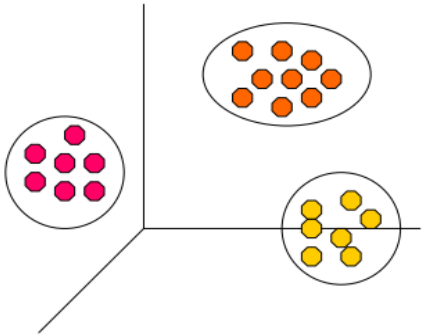
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.

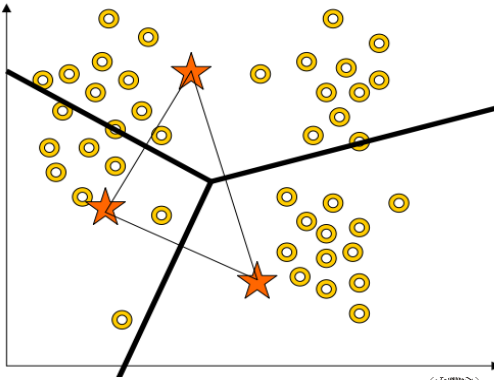


Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

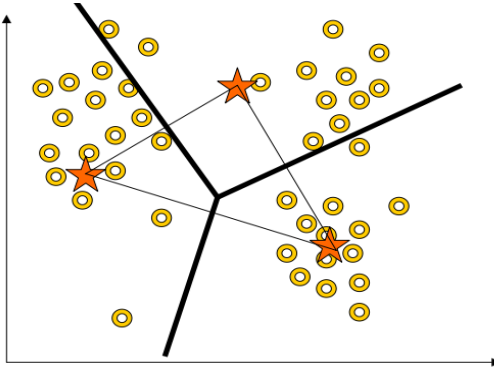
Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.



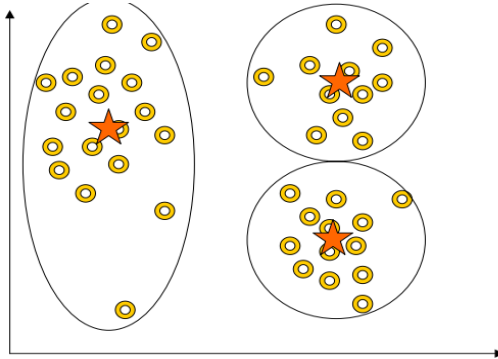
Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.



Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

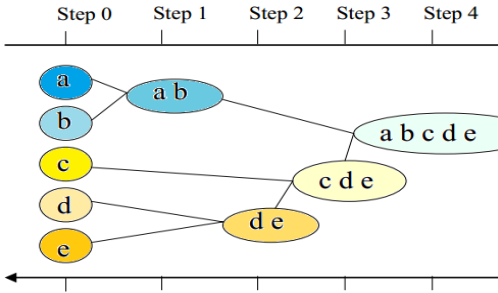


Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

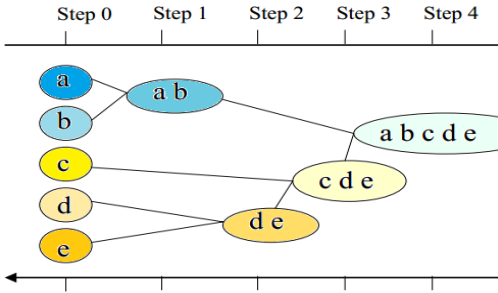
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



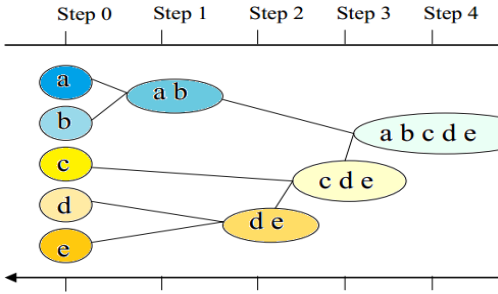
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



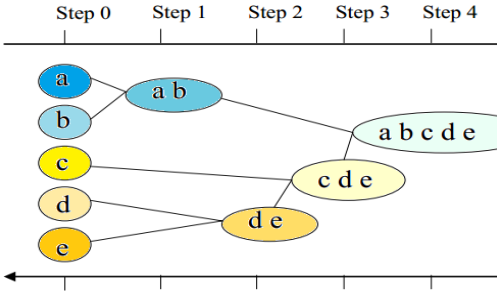
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



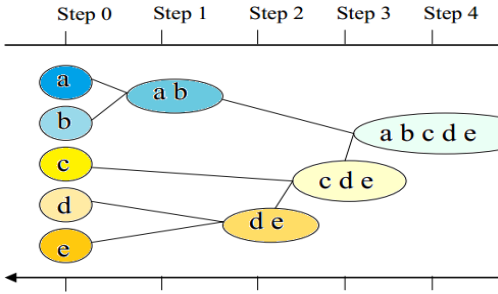
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Problèmatique

Problèmatique

Soit $\mathcal{P}$ une polpulation d'instances de données à N attributs, trouver un partitionnement en K clusters (groupes) $\{C_1, C2, \ldots C_K\}$ de $\mathcal{P}$ telque :

$$\bigcup_{k=1}^{K} C_k = \mathcal{P}$$

Où les clusters C_k soient :

1. Homogènes que possible (similaires au sein d'un même groupe).
2. Distincts que possible (dissimilaires quand ils appartiennent à des groupes différents).

Problèmatique

Problèmatique

Soit $\mathcal{P}$ une polpulation d'instances de données à N attributs, trouver un partitionnement en K clusters (groupes) $\{C_1, C2, \ldots C_K\}$ de $\mathcal{P}$ telque :

$$\bigcup_{k=1}^{K} C_k = \mathcal{P}$$

Où les clusters C_k soient :

1. Homogènes que possible (similaires au sein d'un même groupe).
2. Distincts que possible (dissimilaires quand ils appartiennent à des groupes différents).

Problèmatique

Problèmatique

Soit $\mathcal{P}$ une polpulation d'instances de données à N attributs, trouver un partitionnement en K clusters (groupes) $\{C_1, C2, \ldots C_K\}$ de $\mathcal{P}$ telque :

$$\bigcup_{k=1}^{K} C_k = \mathcal{P}$$

Où les clusters C_k soient :

1. Homogènes que possible (similaires au sein d'un même groupe).
2. Distincts que possible (dissimilaires quand ils appartiennent à des groupes différents).

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Qualité d'un clustering

- Une bonne méthode de clustering produira des clusters d'excellente qualité avec :
  - Similarité intra-classe importante.
  - Similarité inter-classe faible.
- La qualité d'un clustering dépend de :
  - La mesure de similarité utilisée.
  - L'implémentation de la mesure de similarité.
- La qualité d'une méthode de clustering est évaluée par son abilité à découvrir certains ou tous les "pattern" cachés.

Distance et Dissimilarité

Distance

On appelle distance sur un ensemble E, une application $d : E \times E \leftarrow \mathbb{R}^+$ telle que :

1. Séparation : $\forall (x, y) \in E^2 : d(x, y) = 0$ ssi $x = y$
2. Symétrie : $\forall (x, y) \in E^2 : d(x, y) = d(y, x)$
3. Inégalité triangulaire : $\forall (x, y, z) \in E^3 : d(x, z) \leq d(x, y) + d(y, z)$

Distance et Dissimilarité

Distance

On appelle distance sur un ensemble E, une application $d : E \times E \leftarrow \mathbb{R}^+$ telle que :

1. Séparation : $\forall (x, y) \in E^2 : d(x, y) = 0$ ssi $x = y$
2. Symétrie : $\forall (x, y) \in E^2 : d(x, y) = d(y, x)$
3. Inégalité triangulaire : $\forall (x, y, z) \in E^3 : d(x, z) \leq d(x, y) + d(y, z)$

> Une dissimilarité est une application qui a les propriétés de la distance sauf éventuellement l'inégalité triangulaire.

Distance et Dissimilarité

Distance

On appelle distance sur un ensemble E, une application $d : E \times E \leftarrow \mathbb{R}^+$ telle que :

1. Séparation : $\forall (x, y) \in E^2 : d(x, y) = 0$ ssi $x = y$
2. Symétrie : $\forall (x, y) \in E^2 : d(x, y) = d(y, x)$
3. Inégalité triangulaire : $\forall (x, y, z) \in E^3 : d(x, z) \leq d(x, y) + d(y, z)$

- Une dissimilarité est une application qui a les propriétés de la distance sauf éventuellement l'inégalité triangulaire.

Distance et Dissimilarité

Distance

On appelle distance sur un ensemble E, une application $d : E \times E \leftarrow \mathbb{R}^+$ telle que :

1. Séparation : $\forall (x, y) \in E^2 : d(x, y) = 0$ ssi $x = y$
2. Symétrie : $\forall (x, y) \in E^2 : d(x, y) = d(y, x)$
3. Inégalité triangulaire : $\forall (x, y, z) \in E^3 : d(x, z) \leq d(x, y) + d(y, z)$

- Une dissimilarité est une application qui a les propriétés de la distance sauf éventuellement l'inégalité triangulaire.

Distance et Dissimilarité

Distance

On appelle distance sur un ensemble E, une application $d : E \times E \leftarrow \mathbb{R}^+$ telle que :

1. Séparation : $\forall (x, y) \in E^2 : d(x, y) = 0$ ssi $x = y$
2. Symétrie : $\forall (x, y) \in E^2 : d(x, y) = d(y, x)$
3. Inégalité triangulaire : $\forall (x, y, z) \in E^3 : d(x, z) \leq d(x, y) + d(y, z)$

- Une dissimilarité est une application qui a les propriétés de la distance sauf éventuellement l'inégalité triangulaire.

Choix d'une Distance

Définir une distance sur chacun des attributs :

- Distance : $d(x, y) = |x - y|$,
- Distance normalisée : $d(x, y) = \frac{|x-y|}{d_{max}}$

Choix d'une Distance

Définir une distance sur chacun des attributs :

- Distance : $d(x, y) = |x - y|$,
- Distance normalisée : $d(x, y) = \frac{|x-y|}{d_{max}}$

Choix d'une Distance

Définir une distance sur chacun des attributs :

- Distance : $d(x, y) = |x - y|$,
- Distance normalisée : $d(x, y) = \frac{|x-y|}{d_{max}}$

Exemple : Age, taille, poids.

Choix d'une Distance

- **Attributs discrets** :
  - *Données binaires* : $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = 1$.
  - *Données énumératives* : distance nulle si les valeurs sont égales et 1 sinon.
  - *Données énumératives ordonnées* : On peut définir une distance utilisant la relation d'ordre.
- **Données de types complexes** : textes, images, données génétiques, ...etc.

Choix d'une Distance

- **Attributs discrets** :
  - *Données binaires* : $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = 1$.
  - *Données énumératives* : distance nulle si les valeurs sont égales et 1 sinon.
  - *Données énumératives ordonnées* : On peut définir une distance utilisant la relation d'ordre.
- **Données de types complexes** : textes, images, données génétiques, ...etc.

Choix d'une Distance

- **Attributs discrets** :
  - *Données binaires* : $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = 1$.
  - *Données énumératives* : distance nulle si les valeurs sont égales et 1 sinon.
  - *Données énumératives ordonnées* : On peut définir une distance utilisant la relation d'ordre.
- **Données de types complexes** : textes, images, données génétiques, . . . etc.

Choix d'une Distance

- **Attributs discrets** :
  - *Données binaires* : $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = 1$.
  - *Données énumératives* : distance nulle si les valeurs sont égales et 1 sinon.
  - *Données énumératives ordonnées* : On peut définir une distance utilisant la relation d'ordre.
- **Données de types complexes** : textes, images, données génétiques, . . . etc.

Choix d'une Distance

- **Attributs discrets** :
  - *Données binaires* : $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = 1$.
  - *Données énumératives* : distance nulle si les valeurs sont égales et 1 sinon.
  - *Données énumératives ordonnées* : On peut définir une distance utilisant la relation d'ordre.
- **Données de types complexes** : textes, images, données génétiques, . . . etc.

Distance : Données numériques

Standardiser les données

- Calculer l'écart absolu moyen :
  $S_f = \frac{1}{n}(|x_{1f} - M_f| + |x_{2f} - M_f| + \cdots + |x_{nf} - M_f|)$ où
  $M_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$
- Calculer la mesure standardisée (z-score) : $z_{if} = \frac{x_{if} - M_f}{S_f}$

Utiliser une distance : Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$

- Distance **Euclidienne** : $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distance de **Manhattan** : $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
- Distance de **Minkowski** : $d(x, y) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^q}$

Distance : Données numériques

Standardiser les données

- Calculer l'écart absolu moyen :
  $S_f = \frac{1}{n}(|x_{1f} - M_f| + |x_{2f} - M_f| + \cdots + |x_{nf} - M_f|)$ où
  $M_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$
- Calculer la mesure standardisée (z-score) : $z_{if} = \frac{x_{if} - M_f}{S_f}$

Utiliser une distance : Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$

- Distance **Euclidienne** : $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distance de **Manhattan** : $d(x, y) = \sum_{i=1}^{n}|x_i - y_i|$
- Distance de **Minkowski** : $d(x, y) = \sqrt[q]{\sum_{i=1}^{n}|x_i - y_i|^q}$

Distance : Données numériques

Standardiser les données

- Calculer l'écart absolu moyen :
  $S_f = \frac{1}{n}(|x_{1f} - M_f| + |x_{2f} - M_f| + \cdots + |x_{nf} - M_f|)$ où
  $M_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$
- Calculer la mesure standardisée (z-score) : $z_{if} = \frac{x_{if} - M_f}{S_f}$

Utiliser une distance : Soient $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$

- Distance **Euclidienne** : $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Distance de **Manhattan** : $d(x, y) = \sum_{i=1}^{n}|x_i - y_i|$
- Distance de **Minkowski** : $d(x, y) = \sqrt[q]{\sum_{i=1}^{n}|x_i - y_i|^q}$

Distance : Données numériques

Exemple :

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3

d(P1,P2)=120 d(P1,P3)=132

- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données numériques

Exemple :

- Calculer d(P1,P2), d(P1,P3) sans standardiser les données :
  Conclusion : P1 ressemble plus à P2 qu'à P3
- Calculer d(P1,P2), d(P1,P3) après avoir standardisé les données :
  Conclusion : P1 ressemble plus à P3 qu'à P2

	Age	Salaire
P1	50	11000
P2	70	11100
P3	60	11122
P4	60	11074

d(P1,P2)=120 d(P1,P3)=132

	Age	Salaire
P1	-2	-2
P2	2	0.7
P3	0	1.3
P4	0	0

d(P1,P2)=6.7 d(P1,P3)=4.3

Distance : Données binaires

- **Coefficient de correspondance simple** : (similarité invariante, si la variable binaire est *symétrique*) :

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- **Coefficient de Jaccard** : (similarité non invariante, si la variable binaire est *asymétrique*) :

$$d(i,j) = \frac{b+c}{a+b+c}$$

		Objet J		
		1	0	Somme
Objet I	1	a	b	$a+b$
	0	c	d	$c+d$
	Somme	$a+c$	$b+d$	n

TABLE – Table de dissimilarité

Distance : Données binaires

- **Coefficient de correspondance simple** : (similarité invariante, si la variable binaire est *symétrique*) :

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- **Coefficient de Jaccard** : (similarité non invariante, si la variable binaire est *asymétrique*) :

$$d(i,j) = \frac{b+c}{a+b+c}$$

		Objet J		
		1	0	Somme
Objet I	1	a	b	$a+b$
	0	c	d	$c+d$
	Somme	$a+c$	$b+d$	n

TABLE – Table de dissimilarité

Distance : Données binaires

Exemple :

Nom	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Salim	Oui	N	P	N	N	N
Karima	Oui	N	P	N	P	N
Ali	Oui	P	N	N	N	N

TABLE – Table de patients

Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(Salim, Karima) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Salim, Ali) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Karima, Ali) = \frac{2+1}{\ldots} = 0.75$$

Distance : Données binaires

Exemple :

Nom	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Salim	Oui	N	P	N	N	N
Karima	Oui	N	P	N	P	N
Ali	Oui	P	N	N	N	N

TABLE – Table de patients

Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(Salim, Karima) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Salim, Ali) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Karima, Ali) = \frac{2+1}{1+2+1} = 0.75$$

Distance : Données binaires

Exemple :

Nom	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Salim	Oui	N	P	N	N	N
Karima	Oui	N	P	N	P	N
Ali	Oui	P	N	N	N	N

TABLE – Table de patients

Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(Salim, Karima) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Salim, Ali) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Karima, Ali) = \frac{2+1}{1+2+1} = 0.75$$

Distance : Données binaires

Exemple :

Nom	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Salim	Oui	N	P	N	N	N
Karima	Oui	N	P	N	P	N
Ali	Oui	P	N	N	N	N

TABLE – Table de patients

Calculer la distance entre patients, basée sur le coefficient de Jaccard.

$$d(Salim, Karima) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Salim, Ali) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Karima, Ali) = \frac{2+1}{1+2+1} = 0.75$$

Distance : Données énumératives

- Généralisation des variables binaires, avec plus de 2 états : rouge, jaune, bleu, vert ... etc.
- *Méthode 1* : **Correpondance simple** m : # de correspondances, p : # total de variables

$$d(i,j) = \frac{p-m}{p}$$

- *Méthode 2* : Utiliser un grand nombre de variables binaires :
  - Créer une variable binaire pour chaque modalité (ex : variable rouge qui prend la valeur vrai ou faux)

Distance : Données énumératives

- Généralisation des variables binaires, avec plus de 2 états : rouge, jaune, bleu, vert . . . etc.
- *Méthode 1* : **Correpondance simple** m : # de correspondances, p : # total de variables

$$d(i,j) = \frac{p-m}{p}$$

- *Méthode 2* : Utiliser un grand nombre de variables binaires :
  - Créer une variable binaire pour chaque modalité (ex : variable rouge qui prend les valeurs vrai ou faux).

Distance : Données énumératives

- Généralisation des variables binaires, avec plus de 2 états : rouge, jaune, bleu, vert ... etc.
- *Méthode 1* : **Correpondance simple** m : # de correspondances, p : # total de variables

$$d(i,j) = \frac{p-m}{p}$$

- *Méthode 2* : Utiliser un grand nombre de variables binaires :
  - Créer une variable binaire pour chaque modalité (ex : variable rouge qui prend les valeurs vrai ou faux).

Distance : Données énumératives

- Généralisation des variables binaires, avec plus de 2 états : rouge, jaune, bleu, vert ... etc.
- *Méthode 1* : **Correpondance simple** m : # de correspondances, p : # total de variables

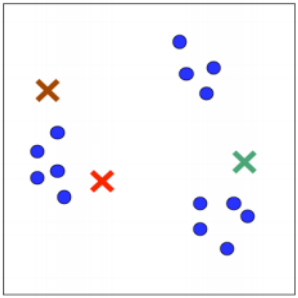
$$d(i,j) = \frac{p-m}{p}$$

- *Méthode 2* : Utiliser un grand nombre de variables binaires :
  - Créer une variable binaire pour chaque modalité (ex : variable rouge qui prend les valeurs vrai ou faux).

Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;

Algorithme **k-Means** (MacQueen'67) :

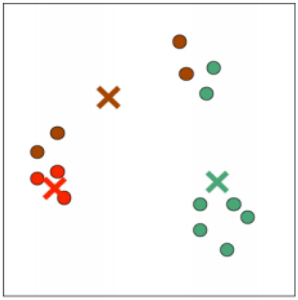
Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;

Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

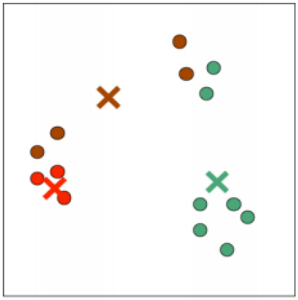
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

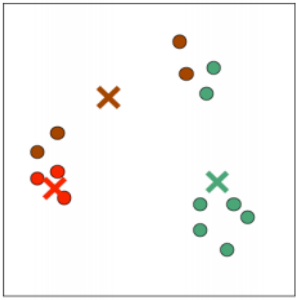
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

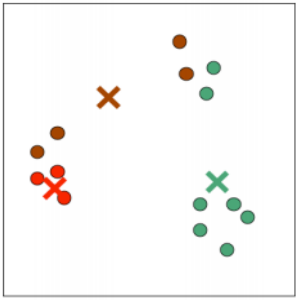
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

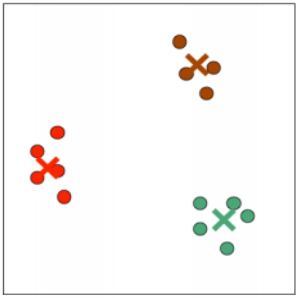
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

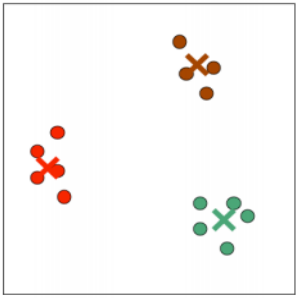
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

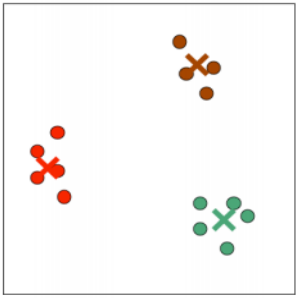
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

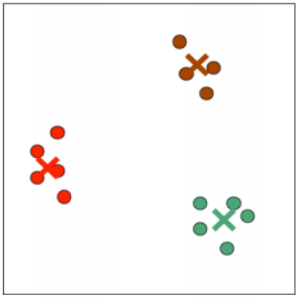
1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme **k-Means** (MacQueen'67) :

Entrées : un ensemble de m enregistrements $x_1, \ldots, x_m$

1. Choisir k centres initiaux $c_1, \ldots, c_k$;
2. Répartir chacun des m enregistrements dans le groupe i dont le centre c_i est le plus proche.;
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes;
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i.;
5. Aller en 2.;



Algorithme k-Means

Exemple :

- 8 points $A, B, \ldots, H$ de l'espace euclidien 2D. $k = 2$ (2 groupes)
- Tire aléatoirement 2 centres : B et D choisi.

Point	Centre	Centre	Centre
	B(2,2)	D(2,4)	J(7/4,12/4)
	D(2,4)	I(27/7,17/7)	K(22/4,9/4)
A(1,3)	B	D	J
B(2,2)	B	D	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Algorithme k-Means

Exemple :

- 8 points $A, B, \ldots, H$ de l'espace euclidien 2D. $k = 2$ (2 groupes)
- Tire aléatoirement 2 centres : B et D choisi.

Point	Centre	Centre	Centre
	B(2,2)	D(2,4)	J(7/4,12/4)
	D(2,4)	I(27/7,17/7)	K(22/4,9/4)
A(1,3)	B	D	J
B(2,2)	B	D	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Algorithme k-Means

Exemple :

- 8 points $A, B, \ldots, H$ de l'espace euclidien 2D. $k = 2$ (2 groupes)
- Tire aléatoirement 2 centres : B et D choisi.

Point	Centre	Centre	Centre
	B(2,2)	D(2,4)	J(7/4,12/4)
	D(2,4)	I(27/7,17/7)	K(22/4,9/4)
A(1,3)	B	D	J
B(2,2)	B	D	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Algorithme k-Means

Exemple :

- 8 points $A, B, \ldots, H$ de l'espace euclidien 2D. $k = 2$ (2 groupes)
- Tire aléatoirement 2 centres : B et D choisi.

Point	Centre	Centre	Centre
	B(2,2)	D(2,4)	J(7/4,12/4)
	D(2,4)	I(27/7,17/7)	K(22/4,9/4)
A(1,3)	B	D	J
B(2,2)	B	D	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K