
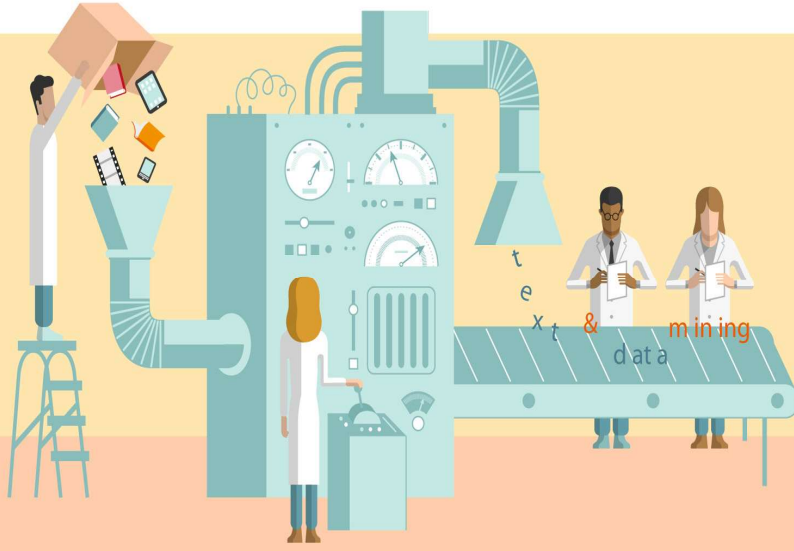


Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences
Département d'Informatique



IPSS
Intelligent Processing
Systems & Security



Data Mining & Machine Learning

Master IPS
Faculté des sciences – Rabat
Université Mohamed V

193


Chapitre 3 :

Techniques descriptives

Classification/(Clustering)

Méthodes de partitionnement

- Nuées dynamiques
- Centres-mobiles
- K-means
- Réseaux de Kohonen (SOM)

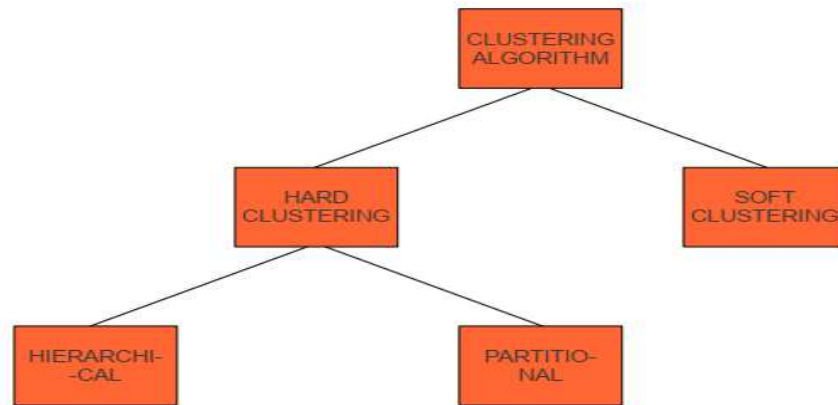


ipes.boutyour@gmail.com 194

194

Clustering : Algorithmes de Clustering

Types :



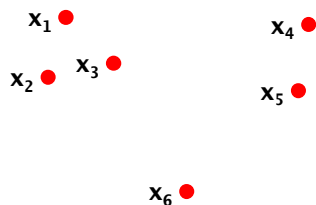
ipes.boutyour@gmail.com

195

195

Clustering : Algorithmes de Clustering

Le clustering par partitionnement



- Nous recherchons une partition plate **C** telle que les objets appartenant à un cluster sont similaires et que les objets appartenant à des clusters différents sont dissemblables.

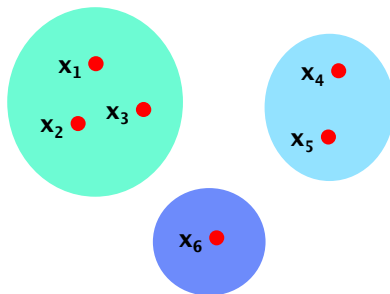
ipes.boutyour@gmail.com

196

196

Clustering : Algorithmes de Clustering

Le clustering par partitionnement



- Nous recherchons une partition plate **C** telle que les objets appartenant à un cluster sont similaires et que les objets appartenant à des clusters différents sont dissemblables.
- Rappelons qu'une partition est la même chose qu'un regroupement ou qu'une relation d'équivalence.

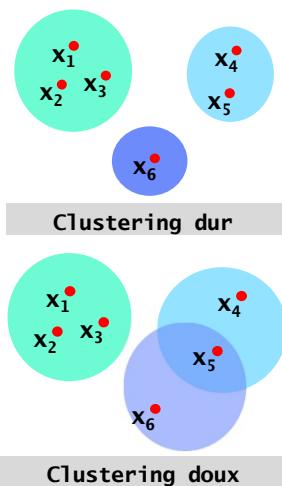
ipes.boutyour@gmail.com

197

197

Clustering : Algorithmes de Clustering

Le clustering par partitionnement



- On peut faire la distinction entre le clustering par partitionnement dur (ou crisp) et doux (ou fuzzy) :
 - Dans le **clustering dur** : un objet n'appartient qu'à un seul cluster.
 - Dans le **clustering doux**, un objet peut appartenir à plusieurs clusters et dans ce cas, il a une valeur d'appartenance non nulle avec tous les clusters auxquels il appartient.

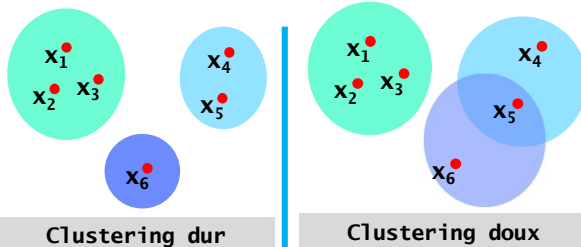
ipes.boutyour@gmail.com

198

198

Clustering : Algorithmes de Clustering

Le clustering par partitionnement



Exemple de la matrice de répartition

	C_1	C_2	C_3
x_1	1	0	0
x_2	1	0	0
x_3	1	0	0
x_4	0	1	0
x_5	0	1	0
x_6	0	0	1

$$U =$$

	C_1	C_2	C_3
x_1	0.9	0.05	0.05
x_2	0.7	0.2	0.1
x_3	0.6	0.25	0.15
x_4	0.2	0.7	0.1
x_5	0.25	0.5	0.25
x_6	0.25	0.3	0.45

$$U =$$

○ On peut faire la distinction entre le clustering par partitionnement dur (ou crisp) et doux (ou fuzzy) :

- Dans le **clustering dur** : un objet n'appartient qu'à un seul cluster.
- Dans le **clustering doux**, un objet peut appartenir à plusieurs clusters et dans ce cas, il a une valeur d'appartenance non nulle avec tous les clusters auxquels il appartient.

ipes.boutyour@gmail.com

199

199

Clustering : Méthodes de partitionnement

Principales méthodes de clustering

- Méthodes hiérarchiques :
- **Méthodes de partitionnement :**
 - Nuées dynamiques
 - Centres mobiles
 - K-means
 - Réseaux de Kohonen
- Méthodes à estimation de densité
- Méthodes mixtes

ipes.boutyour@gmail.com

200

200

Clustering : Méthodes de partitionnement

Nuées dynamiques

- Idée de base de cet algorithme dans des travaux de Hugo Steinhaus et Stuart Lloyd en 1957
- Principe général des algorithmes de partitionnement
- Noyau = sous-ensemble d'individus appartenant à la classe
 - * Noyau bien constitué → meilleure représentativité de la classe
- Puisqu'on fixe le nombre de classes, on parle de typologie ou d'analyse typologique

Clustering : Méthodes de partitionnement

Nuées dynamiques

- Algorithme :
 1. **Indiquer** le nombre de classes souhaitées **k**
 2. **Tirer** aléatoirement **k** objets qui constituent les noyaux initiaux des classes
 3. **Affecter** chaque objet restant à la classe dont le noyau est le plus proche (calcul de la distance)
 4. **Recalculer** les noyaux des **k** classes
 5. **Répéter** de 3. jusqu'à stabilisation des classes ou qu'un nombre défini d'itération soit atteint

Clustering : Méthodes de partitionnement

Principales méthodes de clustering

- Méthodes hiérarchiques :
- **Méthodes de partitionnement :**
 - Nuées dynamiques
 - **Centres mobiles**
 - K-means
 - Réseaux de Kohonen
- Méthodes à estimation de densité
- Méthodes mixtes

Clustering : Méthodes de partitionnement

Centres mobiles

- Algorithme conçu en 1965
- Remplace le noyau des classes par le barycentre
- Le barycentre n'est pas forcément un objet « réel » dans la classe

Clustering : Méthodes de partitionnement

Centres mobiles

○ Algorithme :

1. **Indiquer** le nombre de classes souhaitées k
2. **Tirer** aléatoirement k objets comme **centres initiaux** des classes à constituer
3. **Rattacher** chaque objet restant au centre le plus proche (calcul de la distance)
4. **Recalculer** les barycentres des k classes
5. **Répéter de 3.** jusqu'à stabilisation des classes ou qu'un nombre défini d'itération soit atteint

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

1. Soient $X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$, appliquer l'algorithme des centres mobiles pour créer les deux clusters.

Clustering : Méthodes de partitionnement

Centres mobiles

Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$		
1			$d^2(1, C_1) =$	$d^2(1, C_2) =$
2			$d^2(2, C_1) =$	$d^2(2, C_2) =$
9			$d^2(9, C_1) =$	$d^2(9, C_2) =$
12			$d^2(12, C_1) =$	$d^2(12, C_2) =$
20			$d^2(20, C_1) =$	$d^2(20, C_2) =$
Itération 1				

ipes.boutyour@gmail.com

207

207

Clustering : Méthodes de partitionnement

Centres mobiles

Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$		
1			$d^2(1, C_1) = 0$	$d^2(1, C_2) =$
2			$d^2(2, C_1) = (2-1)^2 = 1$	$d^2(2, C_2) =$
9			$d^2(9, C_1) = (9-1)^2 = 64$	$d^2(9, C_2) =$
12			$d^2(12, C_1) = (12-1)^2 = 121$	$d^2(12, C_2) =$
20			$d^2(20, C_1) = (20-1)^2 = 361$	$d^2(20, C_2) =$
Itération 1				

ipes.boutyour@gmail.com

208

208

Clustering : Méthodes de partitionnement

Centres mobiles

Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	0	361
2	1	324
9	64	121
12	121	64
20	361	0
Itération 1		

$$d^2(1, C_1) = 0$$

$$d^2(1, C_2) = (1-20)^2 = 361$$

$$d^2(2, C_1) = (2-1)^2 = 1$$

$$d^2(2, C_2) = (2-20)^2 = 324$$

$$d^2(9, C_1) = (9-1)^2 = 64$$

$$d^2(9, C_2) = (9-20)^2 = 121$$

$$d^2(12, C_1) = (12-1)^2 = 121$$

$$d^2(12, C_2) = (12-20)^2 = 64$$

$$d^2(20, C_1) = (20-1)^2 = 361$$

$$d^2(20, C_2) = (20-20)^2 = 0$$

ipes.boutyour@gmail.com

209

209

Clustering : Méthodes de partitionnement

Centres mobiles

Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	0	361
2	1	324
9	64	121
12	121	64
20	361	0
Itération 1		



Cluster 1	Cluster 2
1	12
2	20
9	-
Barycentres	

ipes.boutyour@gmail.com

210

210

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{1\}$ et $C_2 = \{20\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	0	361
2	1	324
9	64	121
12	121	64
20	361	0
Itération 1		



Cluster 1	Cluster 2
1	12
2	20
9	-
Barycentres	
$(1+2+9)/3 = 4$	$(12+20)/2 = 16$

ipes.boutyour@gmail.com

211

211

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{4\}$ et $C_2 = \{16\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1		
2		
9		
12		
20		
Itération 2		

$$d^2(1, C_1) =$$

$$d^2(1, C_2) =$$

$$d^2(2, C_1) =$$

$$d^2(2, C_2) =$$

$$d^2(9, C_1) =$$

$$d^2(9, C_2) =$$

$$d^2(12, C_1) =$$

$$d^2(12, C_2) =$$

$$d^2(20, C_1) =$$

$$d^2(20, C_2) =$$

ipes.boutyour@gmail.com

212

212

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{4\}$ et $C_2 = \{16\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	9	225
2	4	196
9	25	49
12	64	16
20	256	16
Itération 2		

$$d^2(1, C_1) = (1-4)^2 = 9$$

$$d^2(1, C_2) = (1-16)^2 = 225$$

$$d^2(2, C_1) = (2-4)^2 = 4$$

$$d^2(2, C_2) = (2-16)^2 = 196$$

$$d^2(9, C_1) = (9-4)^2 = 25$$

$$d^2(9, C_2) = (9-16)^2 = 49$$

$$d^2(12, C_1) = (12-4)^2 = 64$$

$$d^2(12, C_2) = (12-16)^2 = 16$$

$$d^2(20, C_1) = (20-4)^2 = 256$$

$$d^2(20, C_2) = (20-16)^2 = 16$$

ipes.boutyour@gmail.com

213

213

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{4\}$ et $C_2 = \{16\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	9	225
2	4	196
9	25	49
12	64	16
20	256	16
Itération 2		



Cluster 1	Cluster 2
1	12
2	20
9	-
Barycentres	
$(1+2+9)/3 = 4$	$(12+20)/2 = 16$

Les clusters sont stables \rightarrow convergence de l'algorithme

ipes.boutyour@gmail.com

214

214

Clustering : Méthodes de partitionnement

Centres mobiles

○ Exemples :

$X(1 \ 2 \ 9 \ 12 \ 20)$, $k=2$, $C_1 = \{4\}$ et $C_2 = \{16\}$

	$d^2(x, C_1)$	$d^2(x, C_2)$
1	9	225
2	4	196
9	25	49
12	64	16
20	256	16
Itération 2		



Cluster 1	Cluster 2
1	12
2	20
9	-
Barycentres	
$(1+2+9)/3 = 4$	$(12+20)/2 = 16$

Les clusters sont stables → convergence de l'algorithme

ipes.boutyour@gmail.com

215

215

Clustering : Méthodes de partitionnement

Principales méthodes de clustering

- Méthodes hiérarchiques :
- **Méthodes de partitionnement :**
 - Nuées dynamiques
 - Centres mobiles
 - **K-means**
 - Réseaux de Kohonen
- Méthodes à estimation de densité
- Méthodes mixtes

ipes.boutyour@gmail.com

216

216

Clustering : Méthodes de partitionnement

K-Means

- Algorithme conçu en 1967
- Variante de l'algorithme des centres mobiles
- Recalcule du barycentre de chaque classe après introduction de chaque nouvel individu
- Moins d'itérations nécessaires avant stabilisation des classes → plus grande rapidité
- Rapidité dépend de l'ordre d'introduction des individus

ipes.boutyour@gmail.com

217

217

Clustering : Méthodes de partitionnement

K-Means

- Algorithme :
 1. **Indiquer** le nombre de classes souhaitées **k**
 2. **Tirer** aléatoirement **k** objets comme **centres initiaux** des classes à constituer
 3. **Pour chaque objet**:
 - **Rattacher** l'objet au centre le plus proche (calcul de la distance)
 - **Recalculer** les barycentres des classes concernées
 4. **Répéter de 3.** jusqu'à stabilisation des classes ou qu'un nombre défini d'itération soit atteint

ipes.boutyour@gmail.com

218

218

Clustering : Méthodes de partitionnement

K-Means

○ Qualité d'une classification: R^2

- Proportion de la variance (inertie) expliquée par les classes.

$$R^2 = \frac{I_R}{I_{Total}}$$

- $0 \leq R^2 \leq 1$
- Plus c'est proche de 1, plus la classification est bonne
- Critère d'arrêt de fusion des classes: arrêter après le dernier changement de valeur important du R^2

○ Pseudo F :
$$Pseudo\ F = \frac{\frac{R^2}{nombre\ de\ classes - 1}}{\frac{1 - R^2}{nombre\ d'observations - nombre\ de\ classes}}$$

- Proportion de la variance (inertie) expliquée par les classes.

ipes.boutyour@gmail.com

219

219

Clustering : Méthodes de partitionnement

K-Means

○ Comment choisir k:

- Afin de trouver le nombre optimal de clusters pour un k-means, il est recommandé de le choisir en se basant sur :
 - Le contexte du problème traité
 - Les 4 approches suivantes :
 - La méthode du coude (**Elbow method**) (qui utilise les sommes des carrés à l'intérieur des groupes WSS).
 - Méthode de la silhouette moyenne
 - Méthode de la statistique de l'écart (**Gap**)
 - Algorithme basé sur le consensus

Plus d'info sur l'article suivant: <https://cran.rproject.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf>

ipes.boutyour@gmail.com

220

220

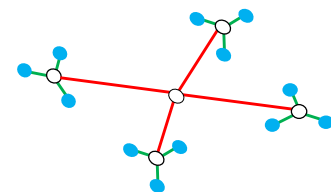
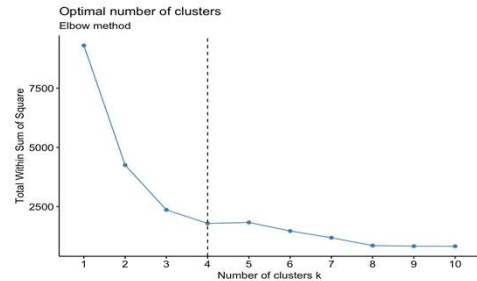
Clustering : Méthodes de partitionnement

K-Means

○ Comment choisir k:

○ Méthode du coude (Elbow method):

- Calculer l'algorithme de clustering (par exemple, le clustering k-means) pour différentes valeurs de k. Par exemple, en faisant varier k de 1 à 10 clusters.
- Pour chaque k, calculer la variance inter-classe I_A du cluster (**WSS: within-cluster sum of square**).
- Tracer la courbe de I_A en fonction du nombre de clusters k.
- L'emplacement d'un coude (genou) dans la parcelle est généralement considéré comme un indicateur du nombre approprié de clusters.



$$I_{\text{Total}} = I_R + I_A$$

(Anglais) TSS = BSS + WSS

ipes.boutyour@gmail.com

221

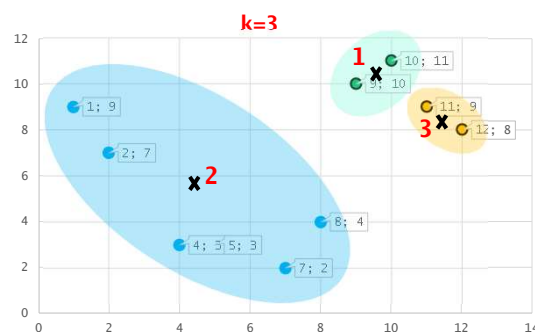
221

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



WSS: Within-cluster sum of square

BSS: Between Sum of Squares

TSS: Total Sum of Squares

ipes.boutyour@gmail.com

222

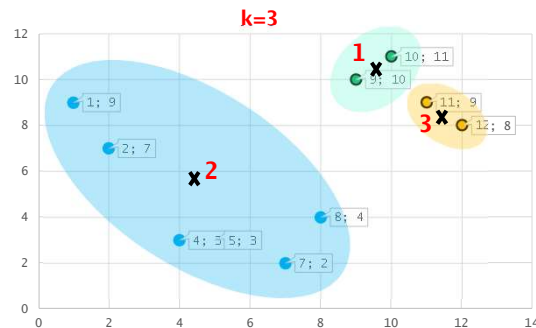
222

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Calculons WSS et BSS

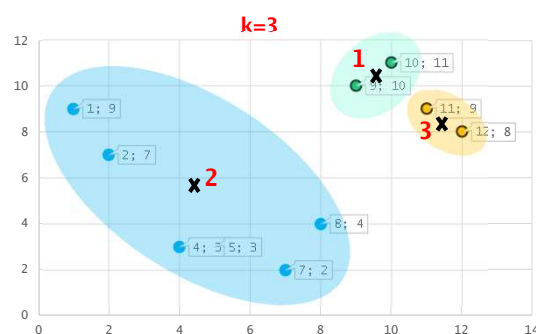
223

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$WSS = \sum_{i=1}^{n_c} \sum_{x \in I} d(x, \bar{x}_i)^2$$

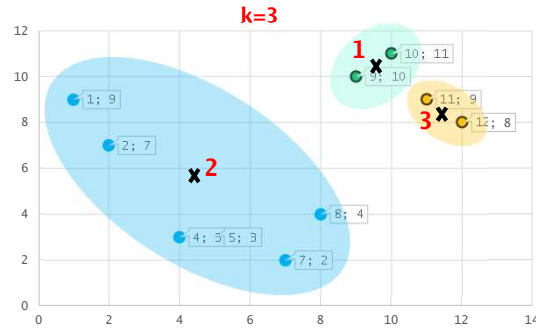
224

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$WSS = \sum_{i=1}^n \sum_{x \in i} d(x, \bar{x}_i)^2 = WSS_1 + WSS_2 + WSS_3$$

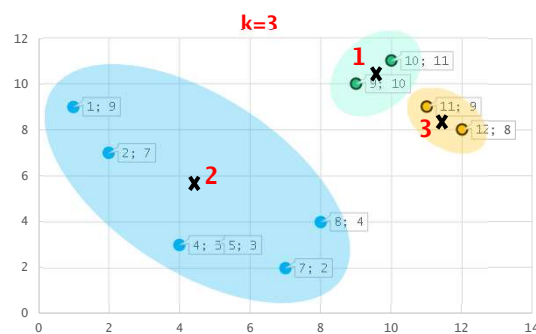
$$WSS_3 =$$

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$WSS = \sum_{i=1}^n \sum_{x \in i} d(x, \bar{x}_i)^2 = WSS_1 + WSS_2 + WSS_3$$

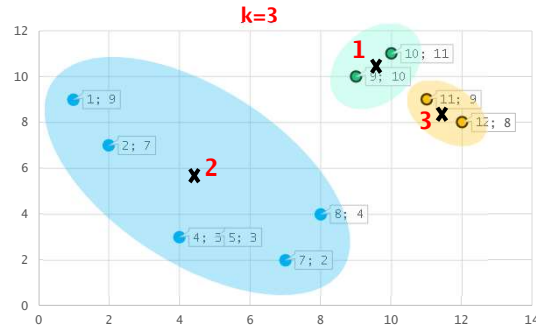
$$WSS_3 = \sum_{x \in 3} d(x, \bar{x}_3)^2 =$$

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$WSS = \sum_{i=1}^n \sum_{x \in i} d(x, \bar{x}_i)^2 = WSS_1 + WSS_2 + WSS_3$$

$$WSS_3 = \sum_{x \in 3} d(x, \bar{x}_3)^2 = (12 - 11.5)^2 + (8 - 8.5)^2 + (11 - 11.5)^2 + (9 - 8.5)^2$$

ipes.boutyour@gmail.com

227

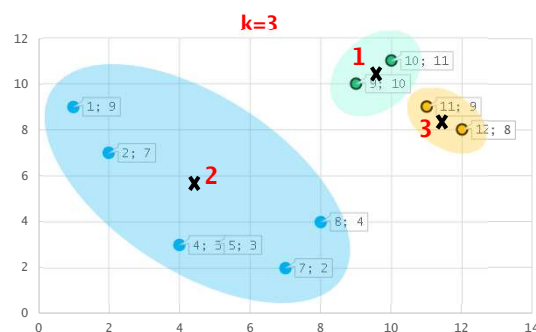
227

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$WSS = \sum_{i=1}^n \sum_{x \in i} d(x, \bar{x}_i)^2 = WSS_1 + WSS_2 + WSS_3$$

$$WSS_3 = \sum_{x \in 3} d(x, \bar{x}_3)^2 = (12 - 11.5)^2 + (8 - 8.5)^2 + (11 - 11.5)^2 + (9 - 8.5)^2$$

$$= 0.5^2 + 0.5^2 + 0.5^2 + 0.5^2 = 1$$

ipes.boutyour@gmail.com

228

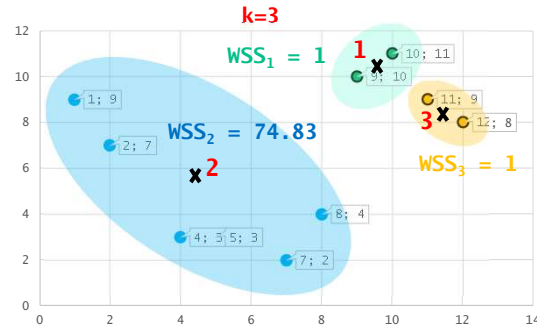
228

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



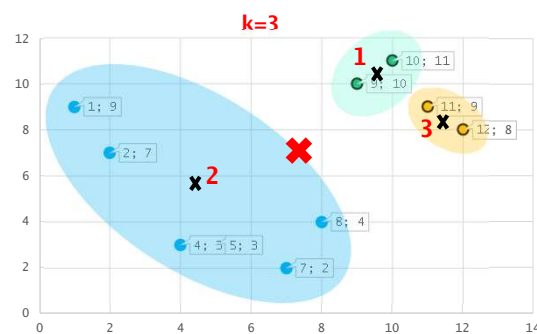
$$WSS = \sum_{i=1}^{n_c} \sum_{x \in i} d(x, \bar{x}_i)^2 = WSS_1 + WSS_2 + WSS_3 = 1 + 74.83 + 1 = 76.83$$

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



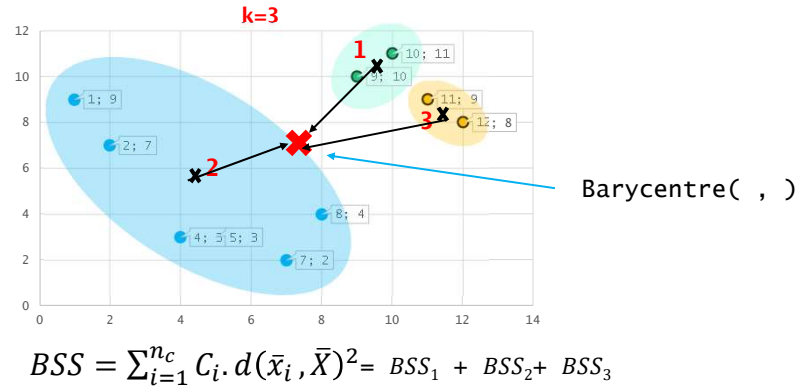
$$BSS = \sum_{i=1}^{n_c} C_i \cdot d(\bar{x}_i, \bar{X})^2 = BSS_1 + BSS_2 + BSS_3$$

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$BSS = \sum_{i=1}^{n_c} C_i \cdot d(\bar{x}_i, \bar{X})^2 = BSS_1 + BSS_2 + BSS_3$$

ipes.boutyour@gmail.com

231

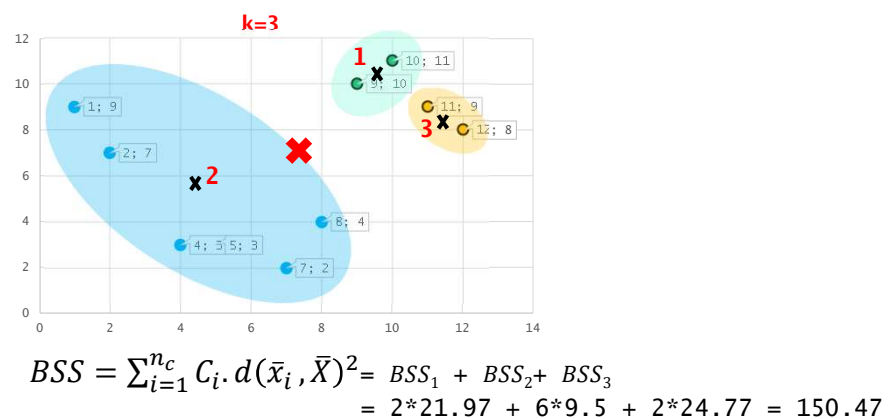
231

Clustering : Méthodes de partitionnement

K-Means

○ Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$BSS = \sum_{i=1}^{n_c} C_i \cdot d(\bar{x}_i, \bar{X})^2 = BSS_1 + BSS_2 + BSS_3$$

$$= 2 \cdot 21.97 + 6 \cdot 9.5 + 2 \cdot 24.77 = 150.47$$

ipes.boutyour@gmail.com

232

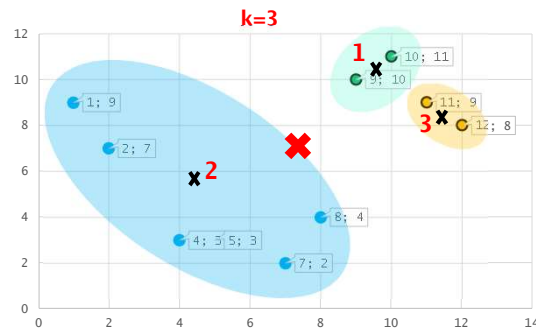
232

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



$$BSS = \sum_{i=1}^{n_c} C_i \cdot d(\bar{x}_i, \bar{X})^2 = BSS_1 + BSS_2 + BSS_3$$

$$= 2 \cdot 21.97 + 6 \cdot 9.5 + 2 \cdot 24.77 = 150.47$$

ipes.boutyour@gmail.com

233

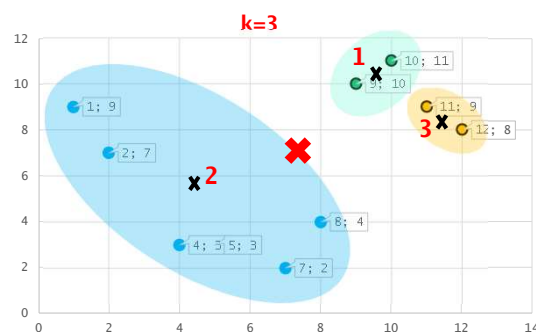
233

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Notes

Le résultat des clusters obtenus dépend des centroïdes initiaux. Pour avoir le k optimal, on change k de 2 jusqu'à N, et on calcule l'inertie inter-classe (WSS en anglais) puis on affiche le graphique du coude (Elbow graph).

ipes.boutyour@gmail.com

234

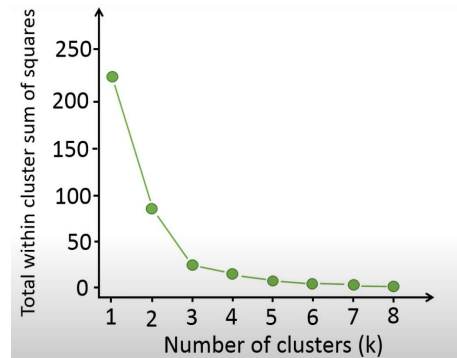
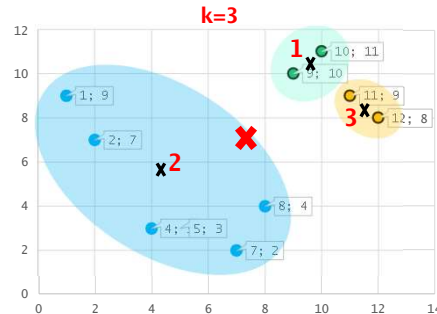
234

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Notes

Le résultat des clusters obtenus dépend des centroïdes initiaux
 Pour avoir le k optimal, on change k de 2 jusqu'à N, et on calcule l'inertie inter-classe (WSS en anglais) puis on affiche le graphique du coude (Elbow graph).

ipes.boutyour@gmail.com

235

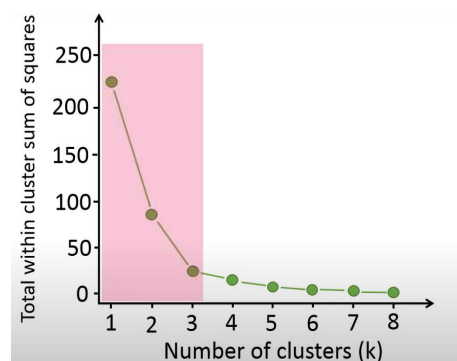
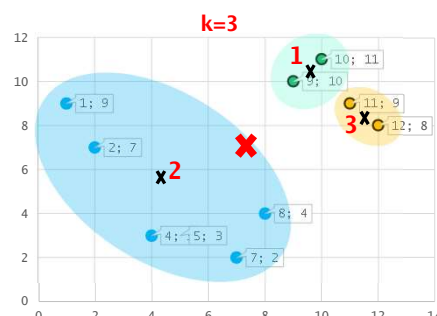
235

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Notes

Le résultat des clusters obtenus dépend des centroïdes initiaux
 Pour avoir le k optimal, on change k de 2 jusqu'à N, et on calcule l'inertie inter-classe (WSS en anglais) puis on affiche le graphique du coude (Elbow graph).

ipes.boutyour@gmail.com

236

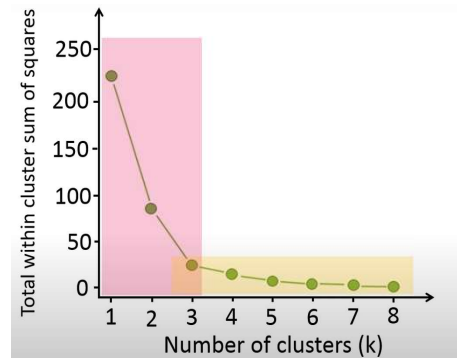
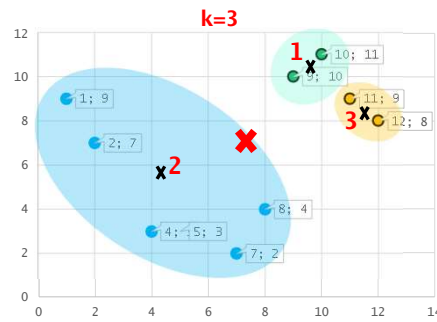
236

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Notes

Le résultat des clusters obtenus dépend des centroïdes initiaux
Pour avoir le k optimal, on change k de 2 jusqu'à N, et on calcule l'inertie inter-classe (WSS en anglais) puis on affiche le graphique du coude (Elbow graph).

ipes.boutyour@gmail.com

237

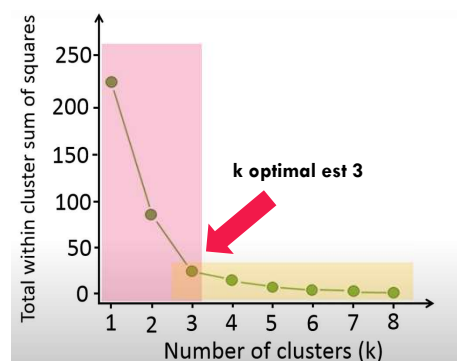
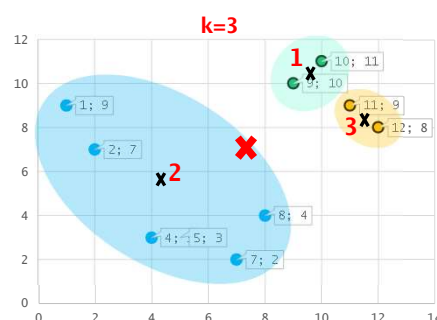
237

Clustering : Méthodes de partitionnement

K-Means

Exemple:

	X	Y
1	1	9
2	2	7
3	4	3
4	5	3
5	7	2
6	8	4
7	10	11
8	11	9
9	9	10
10	12	8



Notes

Le résultat des clusters obtenus dépend des centroïdes initiaux
Pour avoir le k optimal, on change k de 2 jusqu'à N, et on calcule l'inertie inter-classe (WSS en anglais) puis on affiche le graphique du coude (Elbow graph).

ipes.boutyour@gmail.com

238

238

Clustering : Méthodes de partitionnement

Principales méthodes de clustering

- Méthodes hiérarchiques :
- **Méthodes de partitionnement :**
 - Nuées dynamiques
 - Centres mobiles
 - K-means
 - **Réseaux de Kohonen**
- Méthodes à estimation de densité
- Méthodes mixtes

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

- La carte auto-organisatrice est un type de réseau neuronal artificiel.
- Réseaux de neurones à apprentissage non supervisé
- Son réseau est entraîné avec un algorithme d'apprentissage compétitif.
- SOM est utilisé pour les techniques de regroupement et de cartographie (ou de réduction de la dimensionnalité).
- Réduire des problèmes complexes pour en faciliter l'interprétation.
- Le SOM comporte deux couches, l'une étant la couche d'entrée et l'autre la couche de sortie.

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

○ Objectif :

- Représenter un grand échantillon d'individus avec plusieurs variables en une carte à deux dimensions où les nœuds les plus proches sont similaires.

○ Principe :

- « Assigner des centres de classe à une couche radiale en soumettant de façon itérative des formes d'apprentissage au réseau, et en ajustant les pondérations des centres des unités radiales gagnantes (plus proches), et de ses voisins »

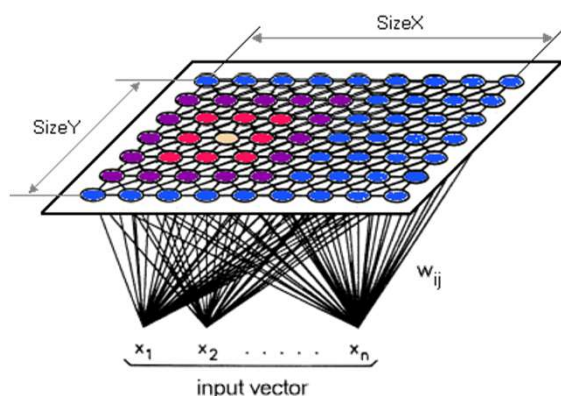
ipes.boutyour@gmail.com

241

241

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)



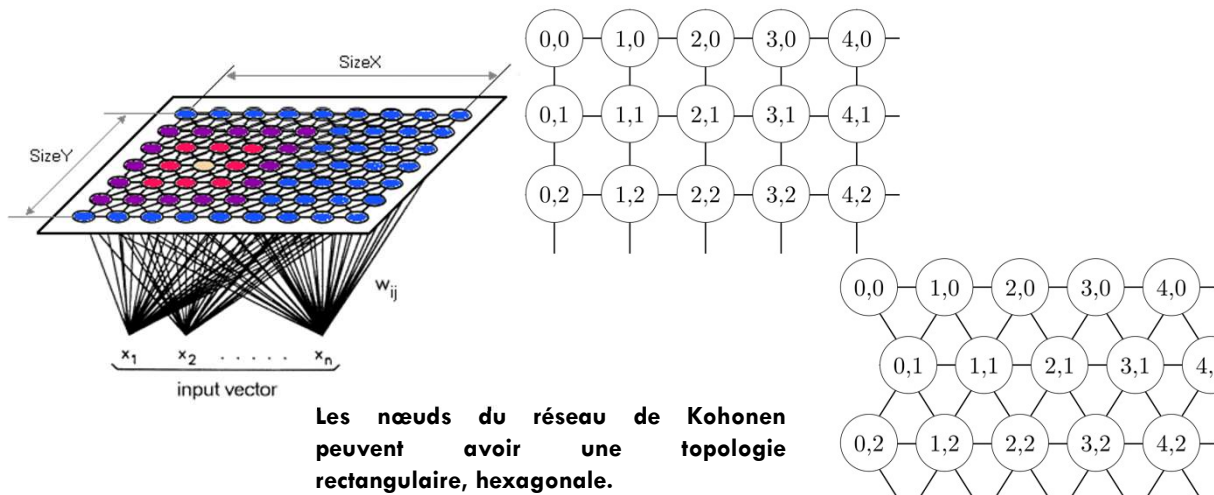
ipes.boutyour@gmail.com

242

242

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)



ipes.boutyour@gmail.com

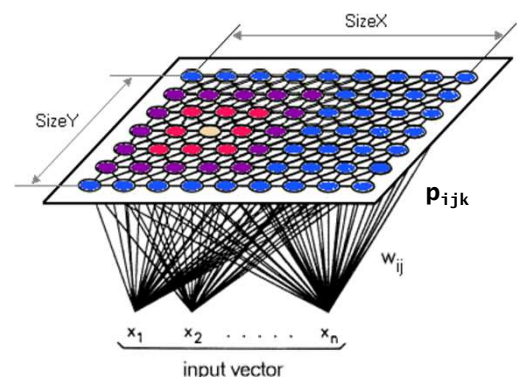
243

243

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

- La taille et la forme de la grille sont choisies par l'utilisateur et peuvent évoluer au cours de l'apprentissage.
- Chaque nœud d'entrée est connecté à tous les nœuds de sortie avec une pondération p_{ijk} qui est initialisé aléatoirement.
- La réponse d'un nœud (i,j) à un individu $x = (x_1 \dots x_N)$ est la distance : $d_{ij}(X) = \sum_{k=1}^N (x_k - p_{ijk})^2 \rightarrow$ fonction de score
- Le nœud (i,j) retenu pour représenter l'individu x est celui qui a le meilleur score = minimise $d_{ij}(X)$



ipes.boutyour@gmail.com

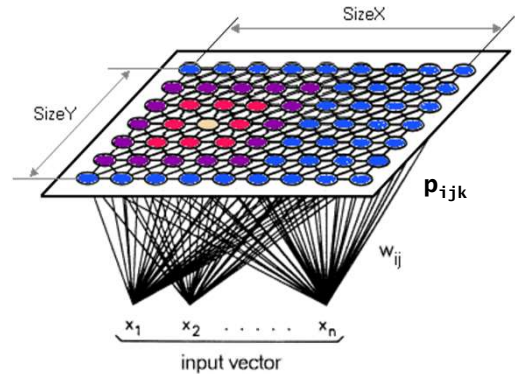
244

244

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

- Les nœuds de sortie sont en compétition entre eux pour être le gagnant → Algorithme compétitif.
- Pour chaque individu, un seul nœud de sortie est activé: **le nœud gagnant = cluster représentant de l'individu.**
- Les poids du nœud gagnant et ses voisins sont réajustés → Deux individus proches sont représentés par deux nœuds proches.
- Apprentissage achevé quand :
 - Chaque individu a été présenté au réseau
 - Tous les poids ont été ajustés



ipes.boutyour@gmail.com

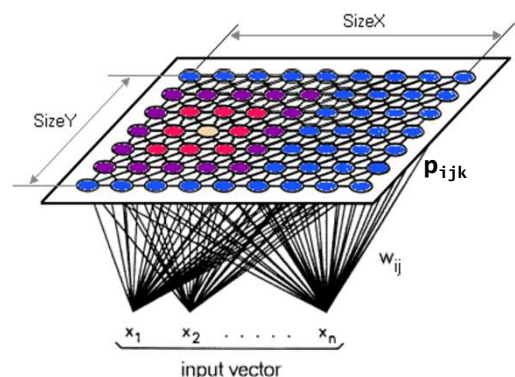
245

245

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

- **Algorithme SOM :**
 1. Initialisation des poids
 2. Pour 1 à N nombre d'époques
 3. Sélectionner l'échantillon pour l'apprentissage (Training)
 4. Calculer le vecteur gagnant
 5. Mettre à jour le vecteur gagnant
 6. Répéter 3, 4, 5 pour tous les individus de l'échantillon d'apprentissage.
 7. Procéder à la classification de l'échantillon du test



ipes.boutyour@gmail.com

246

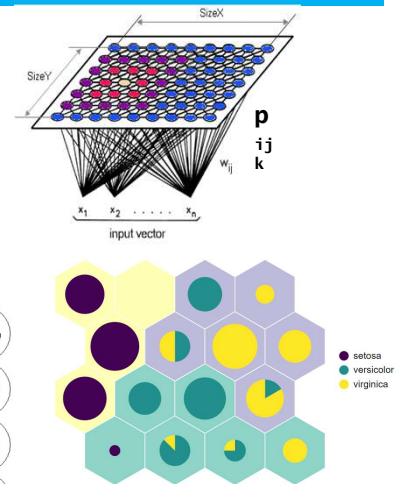
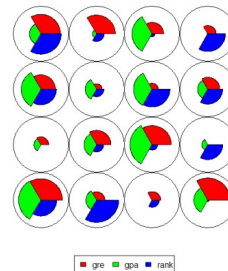
246

Clustering : Méthodes de partitionnement

Réseaux de Kohonen : SOM (Self Organizing Map)

- Sous le langage R, on utilise le package **kohonen** ou le package **aweSOM**.

```
library(kohonen)
bmi.cr <- scale(bmi)
bmi.g <- somgrid(xdim=3,ydim=3)
map <- som(bmi.cr,grid= bmi.g)
plot(map)
plot(map,type='mapping')
plot(map,type='count')
```



ipes.boutyour@gmail.com

247

247

Clustering : Méthodes de partitionnement

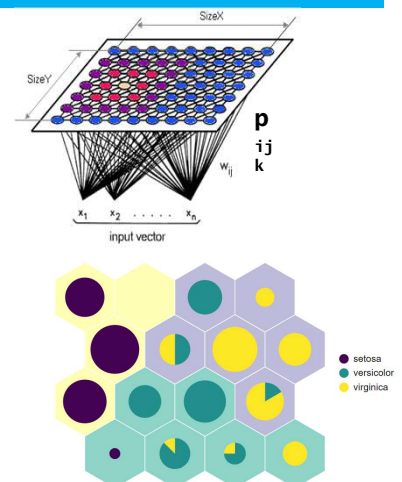
Réseaux de Kohonen : SOM (Self Organizing Map)

○ Avantages:

- Classification raffinée
- Une technique de réduction de dimension non linéaire

○ Inconvénients:

- Difficulté de lire les résultats
- Nécessité de fixer le nombre de classes à l'avance
- Nécessite de normaliser les variables
- sensibles aux valeurs extrêmes



ipes.boutyour@gmail.com

248

248

Chapitre 4: Techniques descriptives

Classification/(Clustering) Méthodes à estimation de densité

- DBSCAN
- DENCLUE

ipes.boutyour@gmail.com

249

249

Clustering

Principales méthodes de clustering

- Méthodes hiérarchiques :
- Méthodes de partitionnement :
- **Méthodes à estimation de densité**
 - **DBSCAN**
 - DENCLUE
- Méthodes mixtes

ipes.boutyour@gmail.com

250

250

Clustering : Méthodes à estimation de densité

- Principales caractéristiques :
 - Découvrir des clusters de forme arbitraire
 - Gérer les bruits
 - Un seul balayage
 - Nécessité de paramètres de densité comme condition de terminaison
- Algorithmes les plus célèbres:
 - **DBSCAN** : Ester, et al. (KDD'96) **OPTICS**: Ankerst, et al (SIGMOD'99)
 - **DENCLUE**: Hinneburg & D. Keim (KDD'98) **CLIQUE**: Agrawal, et al. (SIGMOD'98)

ipes.boutyour@gmail.com

251

251

Clustering : Méthodes à estimation de densité

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Deux paramètres :
 - **Eps**: la distance maximale entre deux points pour qu'ils soient considérés comme voisins
 - **MinPts**: le nombre minimum de points nécessaires pour former un cluster
- Trois types de points à considérer :
 1. **Point central (core point)**: il a plus que le nombre spécifié de points (MinPts) dans Eps
 2. **Point de frontière (border point)**: il a moins de MinPts dans Eps, mais se trouve dans le voisinage d'un point central.
 3. **Point bruit (noise point)**: autre point que le «point central » et le « point de frontière »

ipes.boutyour@gmail.com

252

252

Clustering : Méthodes à estimation de densité

DBSCAN

○ Choix des paramètres :

○ **Eps :**

- Si trop petite, une grande partie des données ne sera pas classifiée car sera considérée comme bruit;
- Si trop grande, les clusters risquent d'être fusionnés → grand nombre d'observations dans même cluster;
- En général, privilégier un Eps de petite valeur (pas trop)

○ **MinPts :**

- Les grandes valeurs sont préférables si données bruitées
- La valeur minimale conseillée est 3
- Plus les grand le dataset, plus grande la valeur de MinPoints;
- Astuce: $\text{MinPts} \geq \text{nombre de variables} + 1$

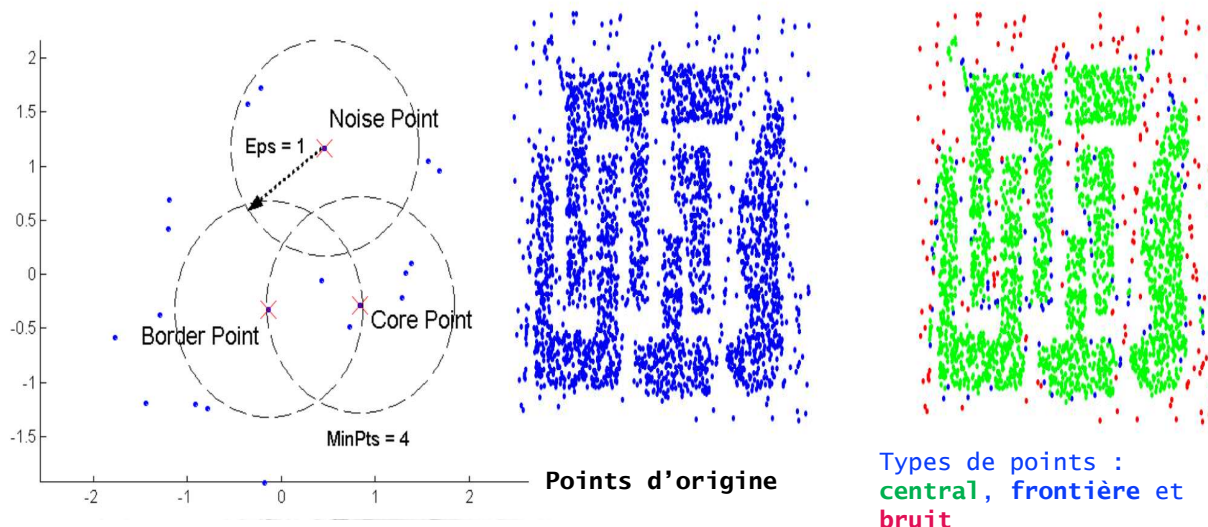
ipes.boutyour@gmail.com

253

253

Clustering : Méthodes à estimation de densité

DBSCAN



ipes.boutyour@gmail.com

254

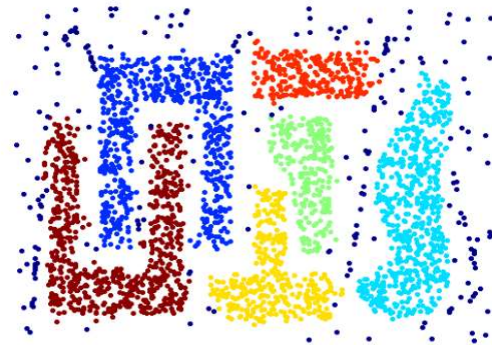
254

Clustering : Méthodes à estimation de densité

DBSCAN



Points d'origine



Clusters

ipes.boutyour@gmail.com

255

255

Clustering : Méthodes à estimation de densité

DBSCAN

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

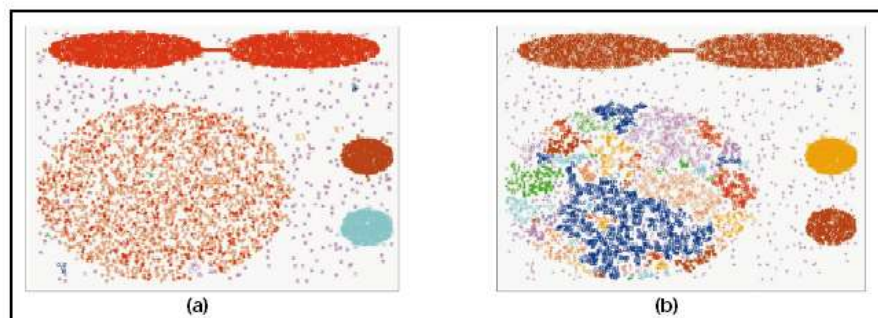
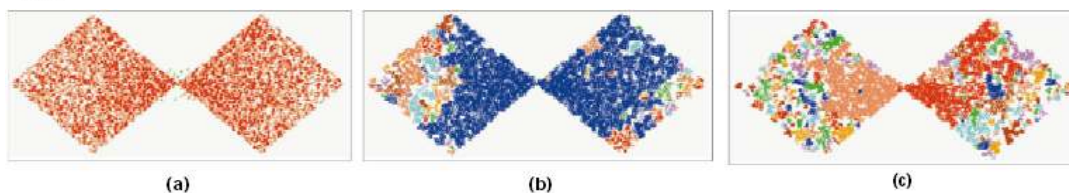


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



ipes.boutyour@gmail.com

256

256

Clustering : Méthodes à estimation de densité

DBSCAN : complexité

○ Complexité en temps :

$O(n^2)$ -pour chaque point il faut déterminer s'il s'agit d'un point central, il peut être réduit à $O(n \cdot \log(n))$ dans des espaces de dimension inférieure en utilisant des structures de données efficaces (n est le nombre d'objets à regrouper) ;

○ Complexité de l'espace :

$O(n)$

ipes.boutyour@gmail.com

257

Clustering : Méthodes à estimation de densité

DBSCAN :

○ Avantages:

- Efficace en temps de calcul
- Pas besoin de fixer le nombre de clusters à l'avance
- Permet de trouver des clusters de formes arbitraires

○ Inconvénients:

- Difficile à utiliser quand le nombre de variables est grand
- Le choix des paramètres est délicat

ipes.boutyour@gmail.com

258

Clustering

Principales méthodes de clustering

- Méthodes hiérarchiques :
- Méthodes de partitionnement :
- **Méthodes à estimation de densité**
 - DBSCAN
 - **DENCLUE**
- Méthodes mixtes

ipes.boutyour@gmail.com

259

259

Clustering : Méthodes à estimation de densité

DENCLUE

- **DENS**ity based **CLUSt**Ering
- Adapté aux données bruitées
- Opère en deux étapes:
 - 1. Construire un hyperrectangle des données à classifier
 - Chaque hyperrectangle est composé d'hypercubes dont la dimension est le nombre de variables
 - On ne considère que les hypercubes peuplés
 - 2. Déterminer les clusters à partir des hypercubes à forte densité (dont le nombre de points dépasse un seuil fixé) et les hypercubes voisins

ipes.boutyour@gmail.com

260

260

Clustering : Méthodes à estimation de densité

DENCLUE

- Basé sur le concept de calcul de l'influence entre les points → **fonction d'influence**
- Décrit l'impact (l'influence) d'un point dans son voisinage (sur ses points voisins)

$$f_{Gauss}(x, y) = \exp \frac{-d(x, y)^2}{2\sigma^2} \text{ avec } \sigma \text{ est le rayon de voisinage de } x$$

- La somme des fonctions d'influence de tous les points → **fonction de densité**

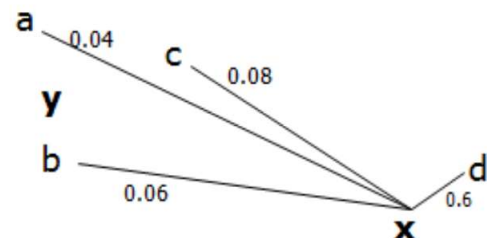
$$f_D(x) = \sum_{i=1}^N f_{Gauss}(x, x_i)$$

avec D est l'ensemble des objets à classifier et N le cardinal de D

Clustering : Méthodes à estimation de densité

DENCLUE

- Exemple:
- Soit $D = \{a, b, c, d\}$
- $f_D(x) = f_{Gauss}(x, a) + f_{Gauss}(x, b) + f_{Gauss}(x, c) + f_{Gauss}(x, d)$
 $= 0,78$



Clustering : Méthodes à estimation de densité



DENCLUE

- Les clusters sont déterminés en utilisant la méthode hill-climbing
 - Identifier les « density attractors » = maxima locaux de la fonction de densité
 - Les objets associés au même density attractor appartiennent au même cluster

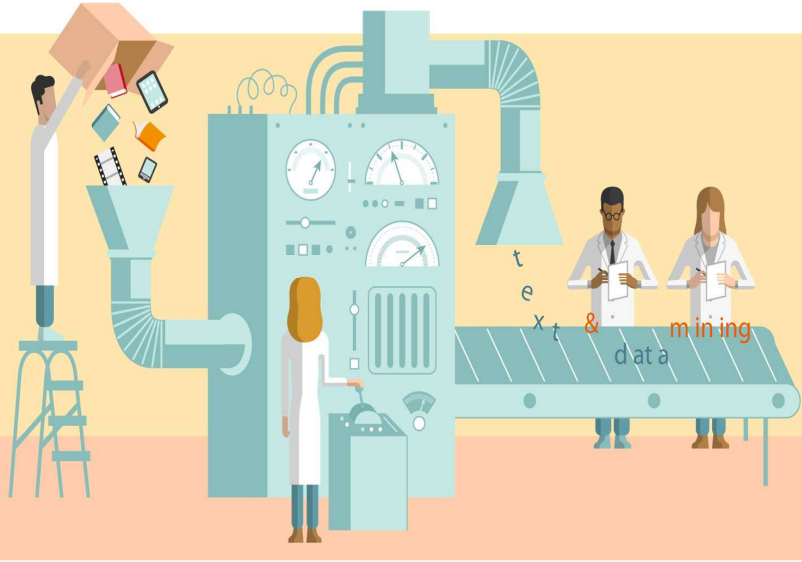
Clustering : Méthodes à estimation de densité

DENCLUE

- Avantages:
 - Il repose sur une solide base numérique et généralise plusieurs approches de regroupement, telles que les méthodes de partitionnement, hiérarchiques et basées sur la densité.
 - Il possède de bonnes propriétés de regroupement pour les ensembles de données comportant de grandes quantités de bruit.
 - Ces méthodes nécessitent une sélection minutieuse du paramètre de densité σ et du seuil de bruit ξ , car la sélection de ces paramètres peut influencer de manière significative la qualité des résultats du clustering.



Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences
Département d'Informatique



Data Mining & Machine Learning

Master IPS
Faculté des sciences – Rabat
Université Mohamed V