

• Typologie de la réduction de dimension

- méthode de base: ACP
- “groupement (clustering) des dimensions”
- extensions:
 - ACP non-linéaire (NLPCA)
 - échelonnement multidimensionnel (multidimensional scaling – MDS)
 - cartes auto-organisatrices (self-organizing maps – SOM)
 - local linear embedding (LLE)
 - ISOMAP
 - courbes principales (principal curves)

• Densités du mélange

- modèle semi-paramétrique:

$$p(\mathbf{x}|\Theta) = \sum_{\ell=1}^k p(\mathbf{x}|C_{\ell}, \Theta_{\ell})P(C_{\ell})$$

- k classes
- vecteur des paramètres: $\Theta = (\Theta_1, \dots, \Theta_k)$
- densités de composante: $p(\mathbf{x}|C_{\ell}, \Theta_{\ell})$
- probabilités a-priori (paramètres du mélange): $P(C_{\ell})$

• Objectif

- estimer Θ , ($P(C_{\ell})$) étant donné $X_n = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$

• Typologie de groupement (clustering)

- méthode de base: k-moyennes
- groupement (clustering) des points
- extensions:
 - k-moyennes flou (fuzzy k-means) \equiv SOM
 - densités du mélange \subseteq k-moyennes flou
 - groupement hiérarchique (hierarchical clustering)

• Approche de maximum de vraisemblance

- $p(X_n|\Theta) = \prod_{i=1}^n p(\mathbf{x}_i|\Theta)$

- $l = \sum_{i=1}^n \log p(\mathbf{x}_i|\Theta)$

$$\begin{aligned} \nabla_{\Theta_{\ell}} l &= \sum_{i=1}^n \frac{1}{p(\mathbf{x}_i|\Theta)} \nabla_{\Theta_{\ell}} \left[\sum_{j=1}^k p(\mathbf{x}_i|C_j, \Theta_j)P(C_j) \right] \\ &= \sum_{i=1}^n P(C_{\ell}|\mathbf{x}_i, \Theta) \nabla_{\Theta_{\ell}} \log p(\mathbf{x}_i|C_{\ell}, \Theta_{\ell}) = 0 \end{aligned}$$

- où $P(C_{\ell}|\mathbf{x}_i, \Theta) = \frac{p(\mathbf{x}_i|C_{\ell}, \Theta_{\ell})P(C_{\ell})}{p(\mathbf{x}_i|\Theta)}$

• Algorithme itératif

```

DENSITÉSDUMÉLANGE( $X_n$ )
1   $\Theta^{(0)} \leftarrow \{\Theta_1^{(0)}, \dots, \Theta_k^{(0)}\}, j \leftarrow 0$ 
2  faire
3    pour  $\ell \leftarrow 1$  à  $k$  faire
4      pour  $i \leftarrow 1$  à  $n$  faire
5         $P_{\ell,i}^{(j)} = P(C_\ell | \mathbf{x}_i, \Theta^{(j)}) \leftarrow \frac{p(\mathbf{x}_i | C_\ell, \Theta_\ell^{(j)}) P(C_\ell)}{p(\mathbf{x}_i | \Theta^{(j)})}$ 
6      pour  $\ell \leftarrow 1$  à  $k$  faire
7         $\Theta_\ell^{(j+1)} \leftarrow \text{solution} \left\{ \sum_{i=1}^n P_{\ell,i}^{(j)} \nabla_{\Theta_\ell} \log p(\mathbf{x}_i | C_\ell, \Theta_\ell) = 0 \right\}$ 
8       $j \leftarrow j + 1$ 
9  jusqu'à  $\left(1 - \frac{j^{(j)}}{j^{(j+1)}}\right) < \text{seuil}$ 
    
```

• Solution ($b > 1$)

- $\mu_\ell = \frac{\sum_{i=1}^n \mathbf{W}_{i,\ell}^b \mathbf{x}_i}{\sum_{i=1}^n \mathbf{W}_{i,\ell}^b}$
- $W_{i,\ell} = \frac{(1/d_{i\ell})^{1/(b-1)}}{\sum_{\ell=1}^k (1/d_{i\ell})^{1/(b-1)}}$, ($d_{i\ell} = \|\mathbf{x}_i - \mu_\ell\|^2$)
- algorithme itératif

• k-moyennes flou (fuzzy k-means)

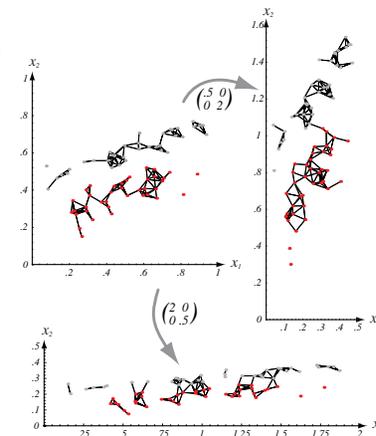
- \mathbf{x}_i appartient à V_ℓ avec un poids $\mathbf{W}_{i,\ell}$ ($\sim P(C_\ell | \mathbf{x}_i)$)
- $\mathbf{W}_{i,\ell}$ est normalisé pour tous les points \mathbf{x}_i :

$$\sum_{\ell=1}^k \mathbf{W}_{i,\ell} = 1$$

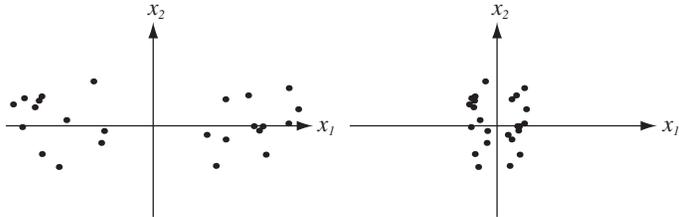
- objectif: minimiser

$$J_{\text{fuz}} = \sum_{\ell=1}^k \sum_{i=1}^n \mathbf{W}_{i,\ell}^b \|\mathbf{x}_i - \mu_\ell\|^2$$

• Normalisation



• Normalisation



• Critères différents

- métrique de Minkowski:

$$d(\mathbf{x}, \mathbf{x}') = \left(\sum_{i=1}^d |x_i - x'_i|^p \right)^{1/p}$$

- mesures de similarité:

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}' \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|}$$

- $\mathbf{x}' \mathbf{x}'$ est le nombre des attributs partagés (variables binaires)
- $\|\mathbf{x}\| \|\mathbf{x}'\|$ est la moyenne géométrique des attributs possédés par \mathbf{x} et \mathbf{x}'
- $s(\mathbf{x}, \mathbf{x}')$: possession relative des attributs

• Critères différents

- versions différentes:

- fraction des attributs partagés: $s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}' \mathbf{x}'}{d}$
- distance de Tanimoto: $s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}' \mathbf{x}'}{\mathbf{x}' \mathbf{x} + \mathbf{x}'' \mathbf{x}' - \mathbf{x}' \mathbf{x}'}$

• Critères différents

- métrique quadratique:

$$J_e = \sum_{i=1}^c \sum_{\mathbf{x} \in V_i} \|\mathbf{x}_i - \mathbf{v}_i\|^2 = \frac{1}{2} \sum_{i=1}^c n_i \bar{s}_i$$

- où $\bar{s}_i = \frac{1}{n_i^2} \sum_{\mathbf{x} \in V_i} \sum_{\mathbf{x}' \in V_i} \|\mathbf{x} - \mathbf{x}'\|^2$

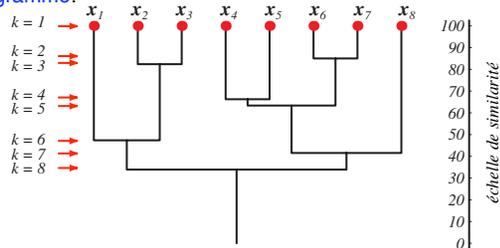
- généralisations:

$$\bar{s}_i = \frac{1}{n_i^2} \sum_{\mathbf{x} \in V_i} \sum_{\mathbf{x}' \in V_i} s(\mathbf{x}, \mathbf{x}')$$

$$\bar{s}_i = \max_{\mathbf{x}, \mathbf{x}' \in V_i} s(\mathbf{x}, \mathbf{x}')$$

• Groupement hiérarchique

• dendrogramme:



• Distances des groupes

- $d_{min}(V_i, V_j) = \min_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \|\mathbf{x} - \mathbf{x}'\|$
- $d_{max}(V_i, V_j) = \max_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \|\mathbf{x} - \mathbf{x}'\|$
- $d_{avg}(V_i, V_j) = \frac{1}{n_i n_j} \sum_{\mathbf{x} \in V_i} \sum_{\mathbf{x}' \in V_j} \|\mathbf{x} - \mathbf{x}'\|$
- $d_{mean}(V_i, V_j) = \|i - j\|$

• Groupement hiérarchique agglomératif

GROUPEMENT HIERARCHIQUE AGGLOMERATIF (X_n, c)	
1	$\hat{c} \leftarrow n$
2	pour $i \leftarrow 1$ à n faire
3	$V_i \leftarrow \{\mathbf{x}_i\}$
4	faire
5	trouver les groupes les plus proches V_i et V_j
6	fusionner V_i et V_j
7	$\hat{c} \leftarrow \hat{c} - 1$
8	jusqu'à $c = \hat{c}$

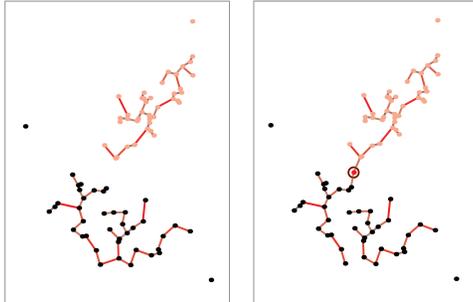
• Groupement hiérarchique – plus proche voisin

- $d_{min}(V_i, V_j) = \min_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \|\mathbf{x} - \mathbf{x}'\|$
- algorithme du lien simple (single-linkage)
- arbre couvrant minimal (Kruskal)

Apprentissage non-supervisé

17

- Groupement hiérarchique – plus proche voisin



Apprentissage non-supervisé

18

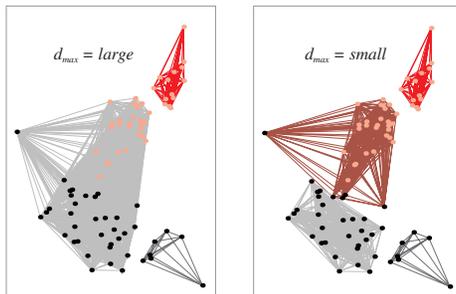
- Groupement hiérarchique – plus loin voisin

- $d_{max}(V_i, V_j) = \max_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \|\mathbf{x} - \mathbf{x}'\|$
- algorithme du **lien complet** (complete linkage)
- augmenter le diamètre le moins possible

Apprentissage non-supervisé

19

- Groupement hiérarchique – plus loin voisin



Apprentissage non-supervisé

20

- Groupement hiérarchique incrémentiel

GROUPEMENT HIÉRARCHIQUE INCREMENTIEL (X_n, c)

- 1 $\hat{c} \leftarrow n$
- 2 **pour** $i \leftarrow 1$ à n **faire**
- 3 $V_i \leftarrow \{\mathbf{x}_i\}$
- 4 **faire**
- 5 trouver V_i et V_j dont la fusion change une critère le moins
- 6 fusionner V_i et V_j
- 7 $\hat{c} \leftarrow \hat{c} - 1$
- 8 **jusqu'à** $c = \hat{c}$

- critère: $J_e = \sum_{i=1}^c \sum_{\mathbf{x} \in V_i} \|\mathbf{x} - \mathbf{x}_i\|^2$

- distance: $d_e(V_i, V_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \|i - j\|$

Apprentissage non-supervisé

21

- Groupement hiérarchique – approche de **théorie de graphe**

- Matrice (graphe) de **similarité**

$$S_{ij} = \begin{cases} 1 & \text{si } d(\mathbf{x}_i, \mathbf{x}_j) < d_0 \\ 0 & \text{sinon.} \end{cases}$$

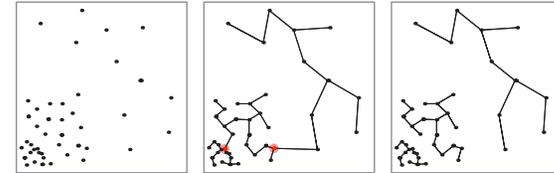
- d_{min} → **composantes connexes**
- d_{max} → **sous-graphes complets**
- Approche de **division**
- Statistique de **longueurs des arrêtes**
- **Chemin de diamètre**

Apprentissage non-supervisé

22

- Groupement hiérarchique – approche de **division**

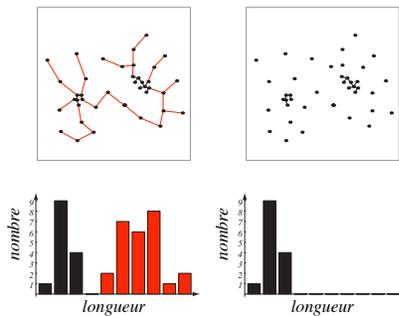
- construire un **arbre couvrant minimal**
- **couper** les arrêtes “longues”



Apprentissage non-supervisé

23

- Groupement hiérarch. – statistique de **longueurs des arrêtes**



Apprentissage non-supervisé

24

- Groupement hiérarchique – **métrique générée**

- $\delta(\mathbf{x}, \mathbf{x}')$ “**dissimilarité**” non-métrique

- non-négativité: $\delta(\mathbf{x}, \mathbf{x}') \geq 0$
- réflexivité: $\delta(\mathbf{x}, \mathbf{x}') = 0$ ssi $\mathbf{x} = \mathbf{x}'$

- “**dissimilarité**” des groupes

$$\delta_{min}(V_i, V_j) = \min_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \delta(\mathbf{x}, \mathbf{x}')$$

$$\delta_{max}(V_i, V_j) = \max_{\substack{\mathbf{x} \in V_i \\ \mathbf{x}' \in V_j}} \delta(\mathbf{x}, \mathbf{x}')$$

- $d(\mathbf{x}, \mathbf{x}')$ **métrique générée**:

- le niveau de **groupement plus bas** où \mathbf{x} et \mathbf{x}' se trouvent dans le **même groupe**
- aussi **symétrique** et satisfait l'**inégalité de triangle**

Apprentissage non-supervisé

25

- Groupement hiérarchique – dans l'espace des attributs
 - trouver les attributs les plus corrélés
 - matrice de covariance: $\mathbf{R} = [\sigma_{ij}]$
 - coefficients de corrélation: $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}}$
 - $0 \leq \rho_{ij} \leq 1$: mesure de similarité entre deux attributs