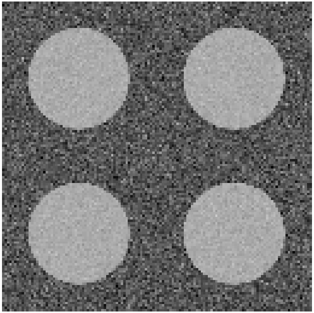
Clustering par kmeans

L. Macaire - M1 IVI - RDF - Cours 8

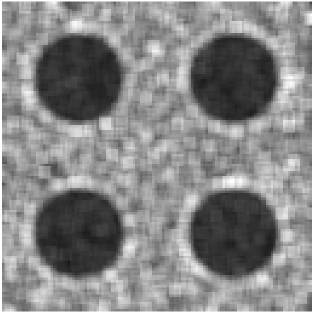
11 mars 2014

Motivations et objectifs du cours

- Classification des données sans connaissance a priori
- Mise en oeuvre de la méthode K-means
- Segmentation d'une image multi-variée



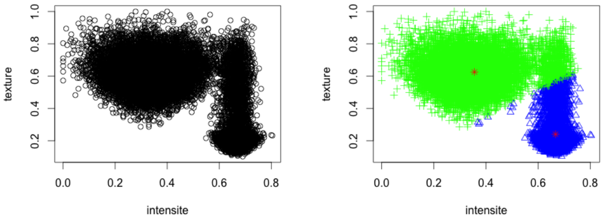
Niveaux de gris rdf-2-classes-texture-1.png



Niveaux de texture rdf-2-classes-texture-1-text.png

Motivations et objectifs du cours

- Projection des pixels dans l'espace des attributs
- Mise en oeuvre de la méthode K-means

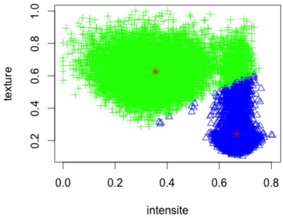


Espace d'atributs

Observations classées en 2 classes

Motivations et objectifs du cours

- Classification des pixels par comparaison des distances observations-centres de gravité



Observations classées en 2 classes

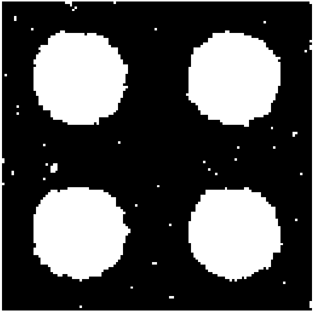


Image segmentée en 2 classes

Matrice des données discrètes $\mathbf{X}$

- Soient N observations pour chacune des D attributs $X_i, i = 1, ..., D$.
- On peut donc représenter les N observations de l'attribut x_i sous la forme d'un vecteur

$$\mathbf{x}_i = (x_{i,1}, ..., xi, j, ..., x_{i,N})^T$$

- On peut alors rassembler les D vecteurs d'attributs $\mathbf{x}_i$ dans une matrice $\mathbf{X}$ de dimension $D \times N$.
- $x_{i,j}$ est donc le $i$eme attribut de la $j$eme observation.
- x est une observation quelconque de dimension D.

Clustering par kmeans - Principes

- Soit K le nombre de classes C_k à retrouver, donné par l'utilisateur.
- L'algorithme kmeans va identifier les K centres de gravité $\hat{\boldsymbol{\mu}}_k$ des classes.
- Ils minimisent la distance entre les points assignés à chaque classe et les centre de gravité associés :

$$\frac{1}{N}\sum_{\mathbf{x}}(\mathbf{x}-\boldsymbol{\mu}_{\hat{\omega}(\mathbf{x})})^T.(\mathbf{x}-\boldsymbol{\mu}_{\hat{\omega}(\mathbf{x})})$$

- où $\hat{\omega}(\mathbf{x})$ est la classe d'assignation de la donnée $\mathbf{x}$.

Clustering par kmeans -Principes

- Minimisation de la distance entre le centre de gravité et les points assignés à chaque classe :

$$\frac{1}{N}\sum_{\mathbf{x}}(\mathbf{x}-\boldsymbol{\mu}_{\hat{\omega}(\mathbf{x})})^T.(\mathbf{x}-\boldsymbol{\mu}_{\hat{\omega}(\mathbf{x})})$$

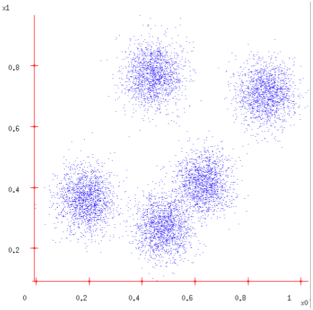
- où $\hat{\omega}(\mathbf{x})$ est la classe d'assignation de la donnée $\mathbf{x}$.
- $\hat{\omega}(\mathbf{x})$ est la classe dont le centre de gravité est le plus proche de $\mathbf{x}$:

$$\hat{\omega}(\mathbf{x}) = argmin_k(\mathbf{x}-\boldsymbol{\mu}_k)^T.(\mathbf{x}-\boldsymbol{\mu}_k)$$

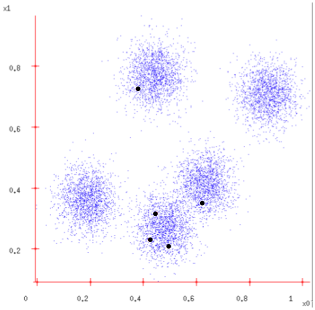
Clustering par kmeans - Algorithme

- Données d'entrée : K et $\mathbf{X}$.
- Soient les positions initiales des centres de gravité des classes $\hat{\boldsymbol{\mu}}_k^{(0)}$, $k = 1, .., K$.
- t = 1
- Tant que critère d'arrêt non satisfait
  - Assignation des points aux K classes
    $\hat{\omega}(\mathbf{x})^{(t)} = argmin_k(\mathbf{x} - \boldsymbol{\mu}_k^{(t-1)})^T.(\mathbf{x} - \boldsymbol{\mu}_k^{(t-1)})$
  - Soit $S_k^{(t)}$ l'ensemble des points assignés à la classe k
    $S_k^{(t)} = \{\mathbf{x}.tq.\hat{\omega}(\mathbf{x})^{(t)} = k\}$
  - Mise à jour des centres de gravité des K classes
    $\boldsymbol{\mu}_k(t) = \frac{1}{|S_k^{(t)}|} \sum_{\mathbf{x} \in S_k^{(t)}} \mathbf{x}$
  - $t = t + 1$

Kmeans- Exemple 1

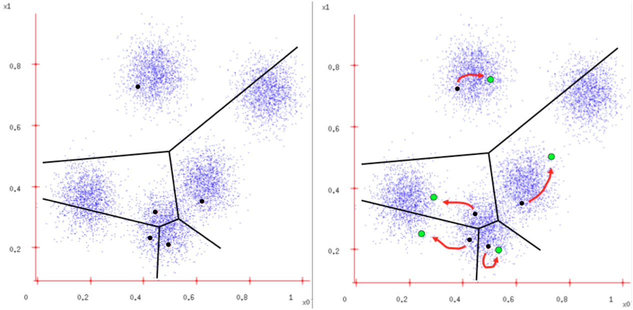


Espace d'atributs



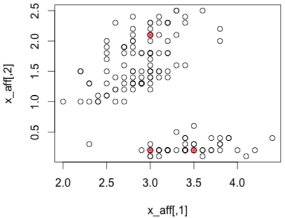
$\hat{\boldsymbol{\mu}}_k^{(0)}$

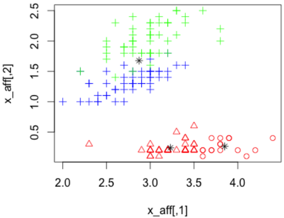
Kmeans- Exemple 1

- 

Assignation　　　Mise à jour des centres $\hat{\boldsymbol{\mu}}_k^{(1)}$

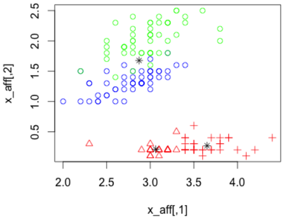
Kmeans- Exemple 2



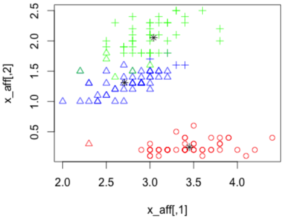


Espace d'atributs et $\hat{\boldsymbol{\mu}}_k^{(0)}$ t=1

Kmeans- Exemple 2



t=2



t=3

Limites de la méthode kmeans

- Sensible aux positions initiales des centres de gravité
- Nécessite de préciser le nombre K des classes
- Critères d'arrêt :
  - Nombre max d'itérations
  - seuil entre 2 itérations : $\sum_k (\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^{(t-1)})^T (\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^{(t-1)})$
- Adaptée aux nuages sphériques