

# 1 Prise en main des programmes

Vous devez réaliser le programme permettant d'étudier l'algorithme de coalescence vu durant le cours

```
[clas,g2]=coalescence(x,K,M,g);
```

Ce programme applique la méthode de coalescence afin de regrouper les différents individus en  $K$  classes. Ce programme ne fonctionne que pour un ensemble d'individu élément de  $R^2$  (2 mesures pour chaque individu).

Les paramètres :

- La variable **x** est une matrice qui doit contenir les différents individus de l'ensemble d'apprentissage rangés par colonne. Le nombre de ligne est 2 et le nombre de colonne est  $n$ .
- La variable **K** indique le nombre de classe que l'on désire.
- La variable **M** est la métrique que l'on désire utiliser. **M** est donc une matrice de taille  $2 \times 2$ .
- La variable **g** contient les centres de gravité initiaux rangés par colonne. **g** est donc une matrice de 2 lignes et **K** colonnes.

Le résultat :

- La variable **clas** est un vecteur qui indique le résultat de l'algorithme de coalescence. **clas(i)** indique le numéro de la classe de l'individu **x(:,i)**.
- La variable **g2** contient les centres de gravité finaux rangés par colonne.

Nous avons aussi à notre disposition la fonction (pour octave/Matlab)

```
affiche_classe(x,clas);
```

Cette fonction permet de visualiser graphiquement dans une représentation planaire le résultat de la classification.

Afin de tester le programme de coalescence, vous allez créer un ensemble d'apprentissage correspondant à l'acquisition de 3 classes d'individus chacun associé à une distribution gaussienne bidimensionnelle.

La première classe est composée de 128 individus et est associée à la fonction de densité

$$f(x) = \frac{1}{(2\pi)} \exp \left( -\frac{1}{2} \left( x - \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left( x - \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right) \right)$$

La seconde classe est composée de 128 individus et est associée à la fonction de densité

$$f(x) = \frac{1}{(2\pi)^4} \exp \left( -\frac{1}{2} \left( x - \begin{pmatrix} -4 \\ -4 \end{pmatrix} \right)^T \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}^{-1} \left( x - \begin{pmatrix} -4 \\ -4 \end{pmatrix} \right) \right)$$

La troisième classe est composée de 128 individus et est associée à la fonction de densité

$$f(x) = \frac{1}{(2\pi)} \exp \left( -\frac{1}{2} \left( x - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right)^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left( x - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) \right)$$

- Indiquer pour chaque classe le vecteur moyenne et la matrice de variance-covariance.
- Pour créer ces données de test, sous Octave/Matlab on peut faire

```
x1=[randn(1,128)+4;randn(1,128)+4];
x2=[randn(1,128)*2-4;randn(1,128)*2-4];
x3=[randn(1,128);randn(1,128)];
x=[x1 x2 x3];
```

- En utilisant la fonction **affiche\_classe** représenter graphiquement la partition optimale.
- Développer l'algorithme de coalescence.
- Appliquer l'algorithme de coalescence. Vous devez donc créer 3 centres de gravité initiaux (paramètres **g** de la fonction). Pour cela, vous allez choisir au hasard 3 individus de l'ensemble d'apprentissage qui constitueront les centres initiaux.
- Etudier graphiquement le résultat de la partition. Mesurer le taux d'erreur : pour cela, vous avez le programme :

```
nbre_erreur=erreur_classif(clas,clas2)
```

avec **clas** la classification attendue et **clas2** la classification obtenue.

- Tester plusieurs fois.

## 2 Influence du point de l'initialisation

- Charger des données `td2_d1.txt`.

Il s'agit de l'enregistrement de 256 individus, à chacun correspondant 2 mesures.

- Visualiser le nuage de points dans une représentation planaire.
- Appliquer plusieurs fois la méthode de coalescence à partir de centres initiaux choisis au hasard et visualiser le résultat de la partition. La partition vous convient-elle à chaque fois ? Expliquer le problème.
- Proposer des centres initiaux (d'une façon non aléatoire) afin d'obtenir une partition satisfaisante.

## 3 Influence de la métrique

- Charger des données `td2_d2`.

Il s'agit de l'enregistrement de 256 individus, à chacun correspondant 2 mesures.

- Visualiser le nuage de points dans une représentation planaire.
- Appliquer la méthode de coalescence à partir de centres initiaux choisis au hasard et visualiser le résultat de la partition.
- Si le résultat de la partition ne vous convient pas, recommencer (en proposant d'autres centres initiaux).

On sait que l'écart type de la mesure 1 est de 0.1 et que l'écart type de la mesure 2 est de 2.

- Proposer une solution permettant d'obtenir une partition satisfaisante.

## 4 Choix du nombre de classes

- Charger des données `td2_d3`.

Il s'agit de l'enregistrement de 256 individus, à chacun correspondant 2 mesures.

- Visualiser le nuage de points dans une représentation planaire.

Afin de déterminer le nombre de classes de l'ensemble d'apprentissage, vous allez faire plusieurs tests en utilisant différent nombre de classe.

- Appliquer la méthode de coalescence sur l'ensemble d'apprentissage avec  $K = 6, 5, \dots, 1$  classes. Pour chacune des partitions obtenues, mesurer le critère qui est optimisé dans l'algorithme de coalescence. Le représenter sous forme d'une courbe en fonction du nombre de classes.
- Conclure sur le nombre de classes de l'ensemble.