O.de Cambry **MA 412**

ESIEE **11/12**

Traitement statistique des données

TP $n^0 2$

Classification

Les fichiers mis à votre disposition se trouvent à l'adresse :
http ://www.esiee.fr/∼decambro

Il s'agit de **Pollco2.mat**, le fichier de données étudié dans le TP1, et de **classr.m**, une fonction Matlab.
Le but est d'étudier et de comparer différentes méthodes de classification sur les données contenues dans **Metaux.mat** que l'on aura pris soin de normaliser auparavant.
On peut utiliser pour cela l'instruction matlab suivante :

```
Xn=(X-repmat(mean(X),size(X,1),1))*inv(diag(std(X,1)')) ;
```

Vous trouverez les fonctions Matlab nécessaires dans la partie **"Cluster analysis"** de la **"Statistics toolbox"** de Matlab. Il est recommandé de lire attentivement les notices d'aide de ces fonctions pour comprendre leur fonctionnement.

- 1 - Algorithme des k-means

On utilise la fonction **kmeans**

- a - Etude de l'algorithme

On fixe le nombre de classes à 4 et on utilise la distance euclidienne.

1. *Calculer une classification par k-means des données et représentez les dans le plan principal à l'aide de* **classr**
2. *A l'aide des options de la fonction* **kmeans**, *'start' et 'display', déterminer le critère de la somme des inerties à la fin de la classification et vérifier sa sensibilité au choix des centres initiaux. Utiliser l'option 'replicate' pour compenser un mauvais choix de centre.*

- b - Calcul du critère

Dans la partie suivante on aura besoin de la valeur du critere de la somme des inerties. *Programmez dans Matlab une fonction* **critere**, *permettant de calculer le critère de la somme des inerties à partir d'un tableau de données X et L vecteur des étiquettes des classes. Vous pourrez vérifier votre programme facilement puisque la fonction* **kmeans**, *donne ce renseignement*

- 2 - Classification par construction ascendante hiérarchique

On utilisera en particulier les fonctions suivantes :

pdist

linkage

dendrogram

cluster

1. *Déterminer les classifications obtenues pour 4 classes par cette méthode en fonction des différents choix de critères d'aggrégation vus en cours ('centroid', 'single','average','complete' et 'ward').*
2. *Calculer la somme des inerties des classifications obtenues et comparer avec les rsultats obtenus par la fonction kmeans.*

Chaque binôme rendra un rapport comprenant les programmes, les résultats et les commentaires utiles avant le 21/10/11.