1 Function realization

The main objective is the realization of the K-mean algorithm, studied during the course.

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

This function applied the K-mean method in order to obtain a clustering of the data x with K clusters. To simplify the numerical development, we consider that each sample is associated with 2 measures $x \in NxR^2$.

The parameters of the function :

— La variable `x` a matrix that contains the observations (each column is a sample). The number of lines is 2 and the number of the columns is N.

— `K` is the number of clusters.

— `M` is the metric for the computation of the distance. The size of `M` is thus 2×2.

— `g` is the initial means of the clusters (for example the kth column is the mean of the class number k. `g` is a matrix $2 \times K$

The output :

— `clas` a vector that contains the result of the algorithm : `clas(i)` is the label of the observation `x(:,i)`.

— `g2` is the final means of the clusters.

I propose some additional functions (for octave/Matlab)

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

This function permits one to show graphically the result of the clustering process : x is the matrix of observations and *clas* contains the labels of the observations.

To test your algorithm, we propose to create a synthetic set of observations associated with two clusters. For each cluster, the distributions following by the observations are a Gaussian distribution.

The first class contains 128 samples and follows the density function such that :

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

The second class contains 128 samples and follows the density function such that :

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

— *In order to create the data, with Octave/Matlab you can execute these commands*

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

— *Display the waiting result for the clustering method with the function* `affiche_classe` .
— *I*mplement the K-mean algorithm.
— *Apply our K-mean algorithm to the test observations, created at the precedent item (for this, it is necessary to select two initial means for example with a random process).*
— *Analyze the result of the k-mean algorithm (you can use the graphical representation). Compute the error rate (for this, you know the perfect result), with the program :*

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

with `clas` the "'theorical"' clustering and `clas2` the obtained classification with the K-mean algorithm.

— *Test the method several times with different initial centers.*

2 The influence of the initial centers

— *load the data* `td2_d1.txt`.
It is a set of 256 observations, with 2 measures for each observation.
— *Display the cloud of the observations.*
— *Test the K-mean method several times with different random initial centers. Display the classification result. Is the partition satisfactory ? Explain the problem.*
— *Set initial means (without random process) in order to obtain a satisfactory partition.*

3 The influence of the definition of the distance

— *Load the data* `td2_d2.txt`.
It is a set of 256 observations, with 2 measures for each observation.
— *Display the cloud of the observations.*
— *Test the K-mean method several times with different random initial centers. Display the classification result. Is the partition satisfactory ?*
— *Set initial means (without random process). Is the partition satisfactory ?*
We known that the standard deviation of the measure 1 is equal to 2 the standard deviation of the measure 2 is equal to 0.5.
— *From this information, propose a solution in order to obtain a satisfactory clustering.*

4 The choice of the number of clusters

— *load the data* `td2_d3.txt`.

It is a set of 256 observations, with 2 measures for each observation.

— *Display the cloud of the observations.*

In order to determine the number of clusters in the data set, we propose to generate the criteria curve for the input data by running the k-means for all values of k between 1 and K, and computing the criteria (the inertia or the distortion described in the course) of the resulting clustering.

— *running the k-means for all values of k between 1 and K, and computing the criteria (the inertia) of the resulting clustering* $K = 6, 5..., 1$. *Show the criteria curve.*

— *Conclude about the number of clusters.*