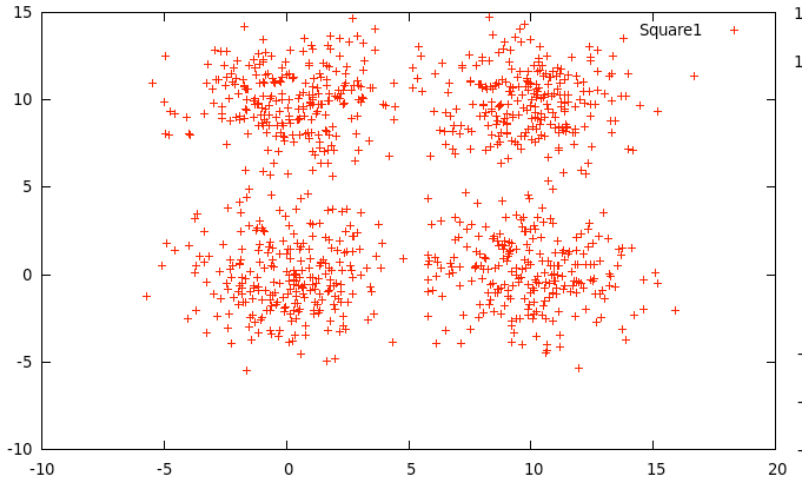
TP
Clustering avec K-Means

On se propose de réaliser un algorithme de clustering. La méthode choisie est K-Means. Pour plus de rapidité il vous est conseillé de programmer de façon relativement générique.

Segmentation de la base de données « Square 1»

On désire faire une segmentation d'un ensemble de points.



Les données sont disponibles dans square1.dat sur moodle.

Travail demandé sur Square 1

1. Détermination de la distance utilisée (Choix, justification).
2. Détermination de la notion de moyenne
3. Implémentation de l'algorithme K-Means (Langage de programmation de votre choix).
4. Utilisation sur la base de données Square 1.
5. Validation de l'algorithme
   Le programme donnera à la fin de l'exécution les différentes valeurs.
   a. Calcul de WC (minimiser)
   b. Calcul de BC (maximiser)
   c. Calcul du ratio BC/WC
6. Etude de l'importance des différents paramètres (choix de K, de la distance ...) via les métriques précédemment implémentée.
7. Utiliser Excel/Openoffice/Libreoffice ou autre (gnuplot) pour visualiser les clusters obtenus.
8. Interprétation, analyse et commentaires sur les résultats obtenus

Segmentation de la base de données « Iris »

On désire faire une segmentation d'un ensemble de points des Iris du fichiers.

Les Iris de différentes familles sont caractérisées par des longueurs et des largeurs de pétales et de sépales différentes.
La base de données contient 150 instances.
Les différents champs de la base sont :
sepal length in cm
sepal width in cm
petal length in cm
petal width in cm
class: Iris Setosa / Iris Versicolour / Iris Virginica

Travail demandé sur Iris

1. Détermination de la distance utilisée (Choix, justification).
2. Détermination de la notion de moyenne
3. Implémentation de l'algorithme K-Means (Langage de programmation de votre choix).
4. Utilisation sur la base de données IRIS.
5. Validation de l'algorithme
    a. Le programme donnera à la fin de l'exécution les différentes valeurs.
    b. Calcul de WC
    c. Calcul de BC
    d. Calcul du ratio
    e. Calcul du tableau de contingence

On affecte le label de la classe majoritaire du cluster au cluster. Ainsi si dans le cluster il y a le plus d'Iris A, on dira que le cluster contient les Iris A.

	Iris Setosa	Iris Versicolour	Iris Virginica	Total
Cluster X (IRIS Setosa)				
Cluster Y (IRIS Versicolor)				
Cluster 3 (IRISVirginica)				
Total				

L'algorithme est évalué sur son nombre d'erreur qui correspond à la partie en rouge du tableau.

6. Etude de l'importance des différents paramètres (choix de K, de la distance ...) via les métriques précédemment implémentée.
7. Utiliser Excel/Openoffice/Libreoffice ou autre (gnuplot) pour visualiser les clusters obtenus.
8. Interprétation, analyse et commentaires sur les résultats obtenus

Segmentation de la base de données « Titanic »

Il s'agit de données concernant les passagers et membres d'équipage du célèbre bateau "le Titanic" qui a fait naufrage le 15 avril 1912.
Source : http://www.amstat.org/publications/jse/datasets/titanic.txt
VARIABLE DESCRIPTIONS:

```
Column  Name      Values
 1      Iden       Anonymous descriptor
 2      Class      0 = crew,    1 = first, 2 = second, 3 = third
 3      Age        0 = child,   1 = adult
 4      Sex        0 = female, 1 = male
 5      Survived   0 = no,      1 = yes
```

Travail demandé sur Titanic

1. Détermination de la distance utilisée (Choix, justification).
2. Détermination de la notion de moyenne