Data Mining : la classification non supervisée

***Clustering* : une affaire de distance ….**

Etude préliminaire.Valeurs discrètes.

Soient les deux individus suivants correspondant à des séquences ADN :
X = AGGGTGGC et Y = AGGCGTAA

1. Dans quel espace « vivent » les points X et Y ? A quelle dimension ? Si on code A=0, G=1, C=2 et T=3, quelle est la distance euclidienne d(X,Y) ? Cela a-t-il un sens en terme de similitude entre les séquences ADN X et Y ? Expliquez notamment en comparant d(A,G) et d(A,T) ?

2. En bioinformatique, la comparaison de séquences ADN deux à deux doit permettre de trouver des homologies c'est-à-dire comment les séquences ont muté à travers les espèces durant l'évolution. Pour cela, on a regroupé les séquences ADN par famille (clustering). A l'intérieur de ces familles, on a réalisé des mesures statistiques. On s'est aperçu que les mutations trans-nucléotides sont déséquilibrées à l'intérieur de famille de séquences appariées. Des matrices de substitution sont utilisées en guise d'heuristiques à la recherche de séquences homologues. Soit par exemple, la matrice de pondération suivante (s'inspirant des matrices de substitution type BLOSUM62 utilisées en bioinformatique) :

$$S=\begin{pmatrix} & A & C & G & T \\ A & 0 & 1 & 0.01 & 1 \\ C & 1 & 0 & 1 & 0.01 \\ G & 0.01 & 1 & 0 & 1 \\ T & 1 & 0.01 & 1 & 0 \end{pmatrix}$$

   Le coefficient 0.01 à la croisée de la ligne G et de la colonne A traduit la très grande fréquence observée de ce type de substitution dans les séquences déjà appariées. A l'inverse, un coefficient 1 indique une très grande rareté observée.
   Proposez une nouvelle mesure de proximité entre deux séquences ADN et l'appliquer aux deux séquences proposées.

3. En quoi tout cela aide-t-il à la découverte de séquences semblables ?

4. Que se passe-t-il si les séquences sont de longueurs différentes ?

Google, Bing, Duchduckgo : indexation, analyse de documents et « text mining »

Essayez de faire une recherche sur un moteur de recherche avec les mots suivants : *loovres* puis *kouvres*, puis *souvres*. Quelle est la connaissance capturée par le coefficient de substitution k ↔ l ou s ↔ d ?
Essayez : *tartable* puis *tirtable* puis *tistable*. Concluez. Pensez au web sémantique.

Algorithmes séquentiels

Soit l'ensemble de vecteurs 2D suivant :
$x_1=[1,1]^T$, $x_2=[1,2]^T$, $x_3=[2,2]^T$, $x_4=[2,3]^T$, $x_5=[3,3]^T$, $x_6=[3,4]^T$, $x_7=[4,4]^T$, $x_8=[4,5]^T$, $x_9=[5,5]^T$, $x_{10}=[5,6]^T$, $x_{11}=[-4,5]^T$, $x_{12}=[-3,5]^T$,$x_{13}=[-4,4]^T$, $x_{14}=[-3,4]^T$.

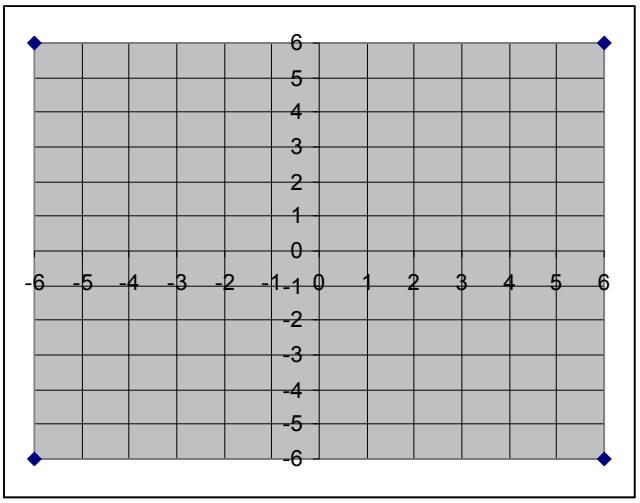
On peut le représenter comme un ensemble d'individus pour lesquels on mesure sur une échelle de –6 à +6 leur intérêt pour le sport et leur intérêt pour l'art.

Individu	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
Intérêt sportif	1	1	2	2	3	3	4	4	5	5	-4	-3	-4	-3
Intérêt artistique	1	2	2	3	3	4	4	5	5	6	5	5	4	4

On considère que la mesure de dissimilarité choisie est la distance euclidienne d. La distance d'un point à une classe est prise égale au minimum parmi les distances de ce point à tous les points de la classe.

1. Faites tourner sur papier les algorithmes séquentiels BSAS et MBSAS quand les vecteurs sont présentés dans l'ordre lexicographique indiqué. On prendra comme seuil $\Theta = \sqrt{2}$ (la diagonale d'un carré de côté 1 afin de faciliter le raisonnement géométrique)
2. Changer l'ordre de présentation à x_1, x_{10}, x_2, x_3, x_4, x_{11}, x_{12}, x_5, x_6, x_7, x_{13}, x_8, x_{14}, x_9 et refaire tourner les algorithmes.

4. Placez les points sur le graphe ci-dessous et évaluez les résultats de ces algorithmes, notamment par rapport au clustering visuel que vous feriez.



Valeurs mixtes : réelles et discrètes.

Soit le tableau suivant résumant les données caractérisant des entreprises.

Entreprise	1er budget	2ème budget	3ème budget	Activité à l'étranger	Nombre d'employés
1 (x1)	1.2	1.5	1.9	0	1
2 (x2)	0.3	0.4	0.6	0	0
3 (x3)	10	13	15	1	2
4 (x4)	6	6	7	1	1

Les trois premières caractéristiques correspondent à leur budget annuel en millions d'euros, la quatrième indique si elles ont une activité à l'internationale, et la dernière estime la taille de l'entreprise : 0 pour un petit nombre d'employés, 1 pour un nombre moyen et 2 pour un très grand nombre.

Proposez une mesure de similarité pour comparer ces entreprises.

Clustering : créer des concepts ? Langage C

Exercice 0

Récupérez le fichier *classif.tar.gz* sur mon site (http://www.math-info.univ-paris5.fr/~lomn/Cours/DM/Material/Data/). Dans ce répertoire, vous bénéficiez à présent d'un programme codé dans le fichier *visualise.c* qui lit une image du type *pgm* (voir les fichiers *nuage1.pg*m et *nuage2.pgm*), puis charge dans une matrice ***points*** l'ensemble des points noirs du nuage considéré, et les réaffiche grâce à la matrice ***mat_out*** en couleur en utilisant le format d'image *ppm*.

Modifier ce programme pour implémenter un algorithme de clustering et testez sur les nuages de points proposé pour différents nombre de classes. On utilisera la distance euclidienne classique.

On affichera les résultats sous la forme d'une image au format ppm en couleur avec une couleur différente pour chaque classe.

Rappels : générer le fichier objet ***visualise.o*** , commande Unix

```
$cc -c visualise.c
#ou gcc
```

puis l'exécutable `visualise` :

```
$cc -o visualise visualise.o
#ou gcc
```

Exécution :

```
$./visualise
```