TP Projet

Vous trouverez tout à l'URL suivante : http://www.math-info.univ-paris5.fr/~lomn/Cours/BC/

A. Cette première partie traite de problématiques de Fouille de Données et en particulier d'apprentissage non supervisé.

Algorithmes séquentiels (6 points)

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

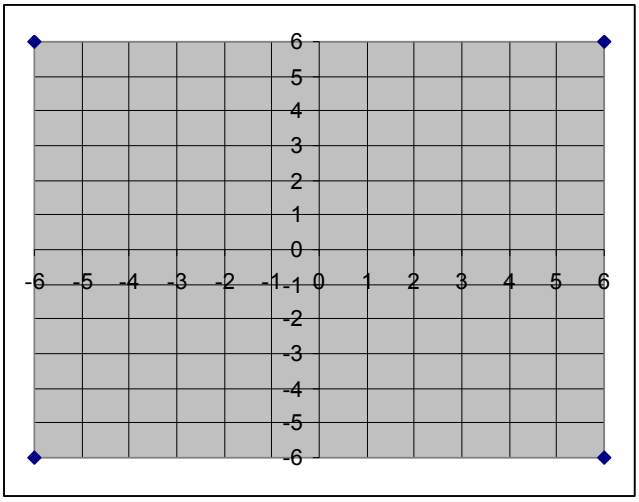
On peut le représenter comme un ensemble d'individus pour lesquels on mesure sur une échelle de –6 à +6 leur intérêt pour le sport et leur intérêt pour l'art.

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 l'algorithme séquentiel BSAS du cours 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 ou une distance pour comparer ces entreprises.

B. Cette deuxième partie vérifie vos capacités de codage en langage de scripts (si vous vous sentez plus à l'aise en Python étudié en L1, vous pouvez répondre aux questions dans ce langage)

Perl /Python (6 points)

http://www.shellunix.com/perl.html

Question 1 : Écrivez le programme PERL suivant – *« firstsearch.pl »* - à l'aide de *gedit*, *gvim, emacs* ou autre en essayant de comprendre ligne par ligne ce qu'il fait (en ajoutant des commentaires dans votre propre code à la suite du caractère spécial de commentaires #) puis exécutez-le : $*perl firstsearch.pl*. Ensuite à l'aide de la commande *man*, étudiez la documentation pour PERL. Éventuellement, surfez sur le WEB.

firstsearch.pl

```
#!/usr/bin/perl -w
# Look for nucleotide string in sequence data

my $target = "ACCCTG";
my $search_string= 'CCAAATTCTTCGGGACCCTGGGGGGGTTAAATTACCCTGACCCTGATG'.
    'CATGGTATGTACAGTAGACTAGGACAACCCTGGGGTAGA';

my @matches;

#Try to find a match in letters 1-6 of $search_string, then look at letters 2-7,
#and so on. Records the starting offset of each match.
foreach my $i (0..length $search_string) {
    if ($target eq substr ( $search_string, $i, length $target)) {
      push @matches, $i;
    }
}
```

```
#Make @matches into a comma-separated list of printing
print "My matches occured at the following offsets : @matches.\n";
print "done\n";
```

Question 2 : Avec cette trame, répondez à l'aide d'un code *perl* ou *python* à ces deux questions en imaginant un test sur une donnée d'intérêt.

1. Recherche de signaux protéine STOP

	.U.	.C.	.A.	.G.
U..	UUU Phe	UCU Ser	UAU Tyr	UGU Cys
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys
	UUA Leu	UCA Ser	UAA Stop	UGA Stop
	UUG Leu	UCG Ser	UAG Stop	UGG Trp
C..	CUU Leu	CCU Pro	CAU His	CGU Arg
	CUC Leu	CCC Pro	CAC His	CGC Arg
	CUA Leu	CCA Pro	CAA Gln	CGA Arg
	CUG Leu	CCG Pro	CAG Gln	CGG Arg
A..	AUU Ile	ACU Thr	AAU Asn	AGU Ser
	AUC Ile	ACC Thr	AAC Asn	AGC Ser
	AUA Ile	ACA Thr	AAA Lys	AGA Arg
	AUG Met	ACG Thr	AAG Lys	AGG Arg
G..	GUU Val	GCU Ala	GAU Asp	GGU Gly
	GUC Val	GCC Ala	GAC Asp	GGC Gly
	GUA Val	GCA Ala	GAA Glu	GGA Gly
	GUG Val	GCG Ala	GAG Glu	GGG Gly

Figure 6.2 *: Code génétique standard. Le code est indiqué dans sa version ARN (avec des U et non des T). Il y a trois codons stop UAA («ocre»), UAG («ambre») et UGA («opale»). La traduction démarre en général sur un codon AUG qui permet l'incorporation de la méthionine N-terminale dans les protéines. Chez les bactéries, le codon GUG est aussi parfois utilisé comme codon de démarrage (~30% des gènes chez E. coli), et plus rarement UUG, mais dans ce cas là, ils codent aussi une méthionine et non pas une valine ou une leucine.*

2. Calcul de taux de A,G,C,T sur plasmodium et streptomyces.

C. Enfin cette partie vérifie vos capacités d'analyse et de recul sur le cours enseigné.

Biologie Computationnelle : définition(s), enjeux, problématiques et solutions ? (8 points)

Sur la base de l'article joint au verso et de l'ensemble des cours et des TP de cette session en informatique, répondez à la question posée ci-dessus en une page ou deux.