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. Calculer la matrice de contingence associée A(x,y). En déduire la distance de Hamming associée. A quoi la valeur trouvée correspond-elle pratiquement ?
3. 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.
4. En quoi tout cela aide-t-il à la découverte de séquences semblables ?
5. Que se passe-t-il si les séquences sont de longueurs différentes ?

Google, indexation, analyse de documents et « text mining »

Essayez : 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.

Soient deux documents V1=(4,2,) et V2=(1,2). A quoi pourrait correspondre cette description-représentation de documents selon vous en terme de mots représentatifs d'un document ? Calculer la distance euclidienne puis du chi-2 entre ces deux vecteurs. Calculer la distance entropique entre ces deux vecteurs.

Valeurs floues.

On travaille dans le cadre de la théorie du flou (théorie des possibilités et des croyances).
A présent, les composantes des vecteurs prennent leurs valeurs dans l'intervalle [0,1]. Une valeur de 1 ou 0 indique une forte croyance dans l'observation correspondante. A l'inverse, une valeur de 0.5 indique une totale ignorance ou une forte imprécision. On travaille donc dans un hypercube de R^n.
Ainsi, donc, x[0.8, 0.4] indique que l'individu x aime le sport avec un degré de possibilité de 0.8 et aime l'art avec un degré de possibilité de 0.4. On considère également les individus y[1,1], z[0.25,0.25], w[0.75,0.75], t[0,0] et u[0.5,0.5].

6. Calculer les similitudes floues inter individus suivantes : sim (t,t) ? sim(u,u) ? sim (y,y) ? sim(w,y) ? sim(z,u) ?
7. Concluez visuellement avec un raisonnement graphique.

Faites un bilan à partir de tous ces exemples sur les notions de distance et de mesure de dissimilarité (différences) et sur les cadres de travail (flou/probabiliste : différence).

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.

9. Proposez une mesure de similarité pour comparer ces entreprises.

Cas 1 : algorithmes séquentiels

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

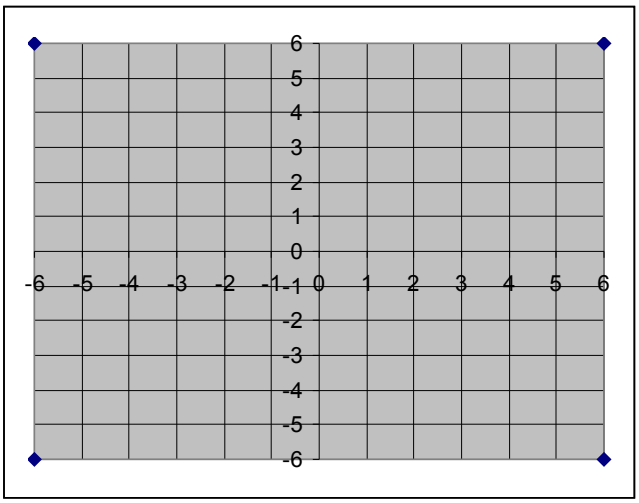
On peut le représenter comme un ensemble d'individus pour lesquels on mesure sur une échelle de –6 à +6 leur intérêt pour le sport et leur intérêt pour l'art.

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.

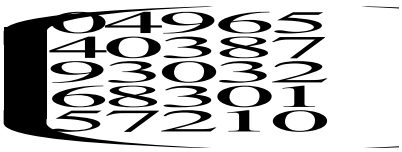
1. Faites tourner l'algorithme BSAS et MBSAS quand les vecteurs sont présentés dans l'ordre lexicographique indiqué. On prendra comme seuil $\Theta = \sqrt{2}$
2. Changer l'ordre de présentation à x_1, x_{10}, x_2, x_3, x_4, x_{11}, x_{12}, x_5, x_6, x_7, x_{13}, x_8, x_{14}, x_9 et refaire tourner les 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.

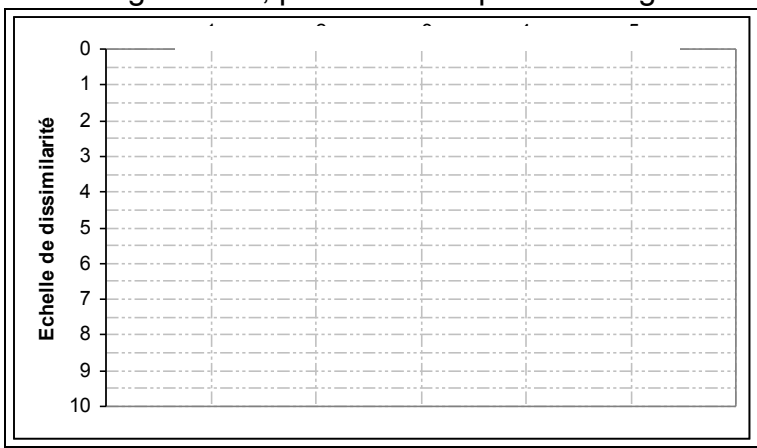


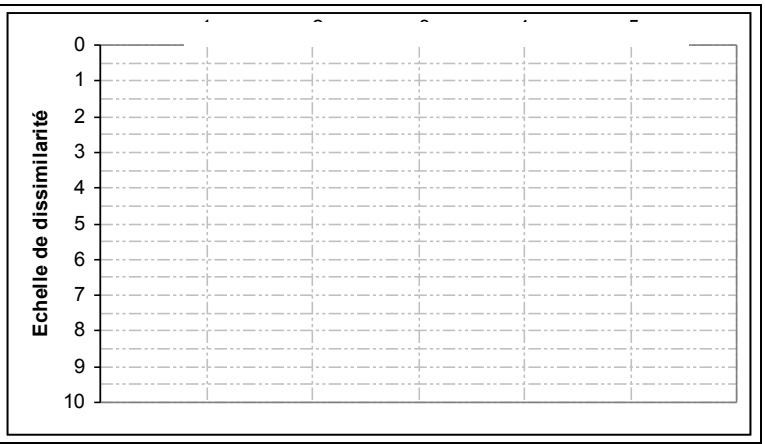
Cas 2 : algorithme hiérarchique

Considérez la matrice de dissimilarité suivante P :



1. Sur combien de points travaille-t-on ? Dans quel espace ?
2. Déterminez les dendrogrammes résultants de l'application du « single link algorithm », puis du « complete link algorithm » sur P et commentez.





3. On modifie un peu la matrice de dissimilarité initiale : P(3,4) = 4 et P(1,2) = 10. On appelle P1 la nouvelle matrice de dissimilarité. Appliquer l'algorithme hiérarchique avec les valeurs suivantes pour la mise à jour de la matrice de proximité :
   b=c = 0 et $a_i = n_i /(n_i+n_j)$ et $a_j = n_j / (n_i+n_j)$.

Clustering : créer des concepts ?

Exercice 0 (si l'objectif principal du cours n'est pas de programmer, on pourra passer cet exercice) : programmation C

Récupérez le fichier *classif.tar.gz* sur mon site. 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 l'algorithme de clustering type *C-moyennes floues* (« fuzzy K-Means ») 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 »* puis l'exécutable visualise « cc –o visualise visualise.o »

Exercice 1 : les critiques du pariscope

Soit le fichier de critique critiquefilms.pdf.

a) Proposer une ou plusieurs problématique de type classification automatique.

b) Modéliser vos problématiques dans le cadre formel d'une classification hiérarchique. Vous testerez dans l'exercice suivant avec HCE.

c) Comment transformer/ faire évoluer le problème/ les données pour proposer une problématique de type prédiction.

Exercice 2 : utilisation du logiciel HCE sous Windows

Author : ***Jinwook Seo*** ()
If you have any comment or question, send an email to the author.

Download HCE 3.0 ou sur mon site
Project Webpage : http://www.cs.umd.edu/hcil/multi-cluster/
Human-Computer Interaction Lab (http://www.cs.umd.edu/hcil/)
University of Maryland, College Park

HCE (Hierarchical Clustering Explorer) est un outil de visualisation pour l'exploration interactive d'ensembles de données multidimensionnelles. Il œuvre comme un téléscope permettant d'explorer et de comprendre ces ensembles en maximisant les habiletés perceptuelles humaines qui ont été sous-utilisées dans les logiciels existant. Il doit permettre d'identifier des caractéristiques cachées et inattendues dans l'espace multidimensionnel.
Ce logiciel sur un principe dit GRID : Graphics, Ranking, and Interaction for Discovery pour aider à découvrir :

- Relations
- Clusters
- Points Manquant
- Points aberrants (outliers)
- Autres

1. **Récupération des données :** Ouvrir le fichier *cereal.txt*. Le format de données doit respecter :
   i. 1ère ligne : nom des attributs
   ii. 2ème ligne : type des attributs
   iii. 1ère colonne : identifiant de type *fieldtype* donc unique.

   Remarque : les variables catégoriques sont considérées comme des codes entiers (1 pour « Cold » , 2 pour « Hot ») et ne sont pas pris en compte pour l'instant dans le logiciel pour le clustering.

2. **Prétraitrement des données :filtrage et transformation** : nous ne nous occupons pas pour l'instant de l'opération de filtrage, même si c'est une étape importante d'un processus de Data Mining. Ici, ils sont trop spécifiques à l'analyse des Puces ADN. Par contre, vous choisissez de normaliser les données **des colonnes** dans l'hypercube $[0,1]^n$ pour visualiser plus de façon plus compact les données (Cocher Colum-by-Column and Scale)
   Sur combien de points et dans quel espace travaillons-nous ?

3. **Déterminer les paramètres de clustering : prendre soin de ne regrouper que les lignes** (décocher Cluster Columns **à chaque fois**) puis faire plusieurs combinaisons possibles (en cliquant sur le bouton Clustering à chaque fois). Vous visualiserez à chaque fois les dendrogrammes correspondant.
   A quoi correspond le code des couleurs par défaut ? Faites le lien avec l'espace dans lequel on travaille ?

4. **Visualiser et utiliser un clustering :** soit le clustering des céréales obtenu par :
   - Complete Linkage

- distance euclidienne
- sur les attributs CALORIES, PROTEIN, FAT

Utilisez la barre de similarités *Min* pour obtenir 4 clusters puis la barre *Detail* pour visualiser les 4 familles de céréales avec la moyenne de leur composition en les 3 attributs de regroupement.

- Quelle est la famille de céréales la plus riche en ces composants ?
- Sélectionnez cette famille en cliquant sur l'arbre concerné. Un retour en jaune est visible. Dans la fenêtre de contrôle à droite, dans l'onglet Detail, visualisez les données textuelles correspondantes.
- En n'affichant que cette famille (Menu contextuel sur l'arbre concerné), combien y a-t-il de céréales correspondant ? Essayez de trouver des liens entre ces céréales d'une même famille ?
- Enregistrez les données sélectionnées sous le nom *cereales_riches.exp* pour une future analyse.

5. **Exploration des données** : *histogram, scatterplot, profile* que nous verrons peut-être une autre fois mais qui montre la richesse des outils d'exploration interactive de données multi-dimmensionnelles offert par ce logiciel fait plus spécifiquement pour des généticiens.
6. Tester votre modélisation b) de l'exercice sur les critiques de films avec ce logiciel.

Association : créer des règles ?

Soit la base de données booléennes suivantes :

	x1	x2	x3	x4	x5
e1	0	1	0	0	1
e2	0	0	1	0	1
e3	0	0	1	0	0
e4	1	1	1	1	1
e5	1	1	1	1	1
e6	1	1	1	1	0
e7	1	0	1	1	0
e8	1	0	1	1	0
e9	1	0	0	0	1
e10	1	0	0	0	1

Faite tourner l'algorithme a priori avec MinCouv=0.3, en prenant soin de ne pas prendre en compte les associations impossibles en cours d'algorithme.
Quelles sont les règles trouvées ?