EXAMEN ANALYSE DE DONNEES
M1 Informatique/MIAGE
12 janvier 2012

Durée de l'examen : 3h
Supports de cours autorisés.

Exercice 1 (ACP) :

Considérons les valeurs suivantes correspondant à la taille et au poids d'une population :

Taille : 170 166 167 182 189 172 177 169 179 191 181 175 176 175 161

Poids : 66 65 75 80 101 74 79 66 73 87 86 80 65 59 54

a. Tracer le graphe correspondant à ces valeurs avec une croix pour chaque couple (avec le poids en abscisse et la taille en ordonnée).
b. Calculer la taille moyenne et le poids moyen.
c. Tracer une ligne horizontale et une ligne verticale passant par le couple moyen (*command line*).
d. Calculer la matrice de covariance S.
e. Calculer les valeurs propres et vecteurs propres de cette matrice S.
f. Calculer la composante principale de nos données, qui correspond à la corpulence.

Exercice 2 (classification hiérarchique) :

Soient 8 individus à 2 variables :

3	2	12	1	10	7	2	8
8	1	3	9	5	3	14	9

La matrice des distances euclidiennes entre ces données est la suivante :

0	7.07	10.30	2.24	…	6.40	6.08	5.10
	0	10.20	8.06	8.94	5.39	13.00	10.00
		0	12.53	2.83	5.00	14.87	7.21
			0	9.85	…	5.10	7.00
				0	3.61	12.04	…
					0	12.08	6.08
						0	7.81
							0

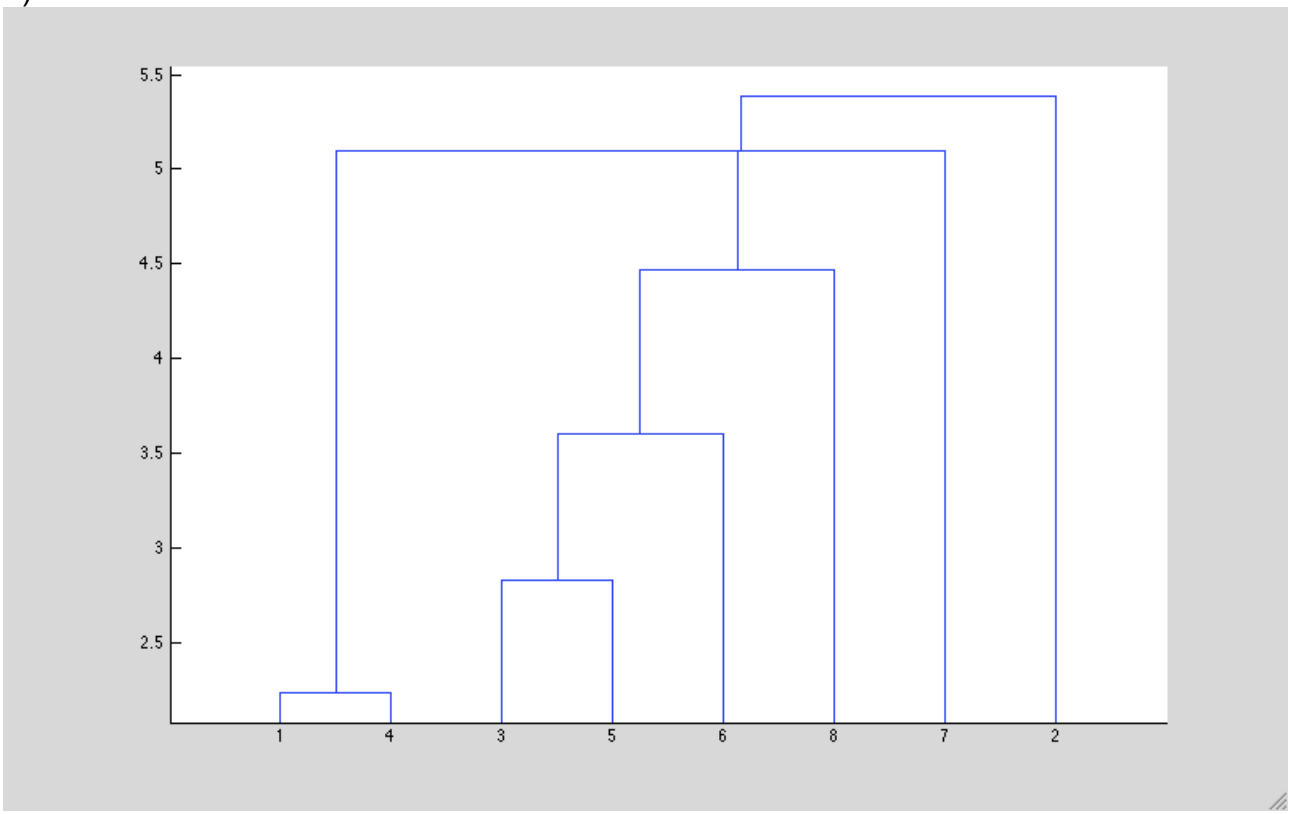
ans =

0	7.0711	10.2956	2.2361	7.6158	6.4031	6.0828	5.0990
7.0711	0	10.1980	8.0623	8.9443	5.3852	13.0000	10.0000
10.2956	10.1980	0	12.5300	2.8284	5.0000	14.8661	7.2111
2.2361	8.0623	12.5300	0	9.8489	8.4853	5.0990	7.0000

7.6158	8.9443	2.8284	9.8489	0	3.6056	12.0416	4.4721
6.4031	5.3852	5.0000	8.4853	3.6056	0	12.0830	6.0828
6.0828	13.0000	14.8661	5.0990	12.0416	12.0830	0	7.8102
5.0990	10.0000	7.2111	7.0000	4.4721	6.0828	7.8102	0

1) Calculer les valeurs manquantes de la matrice des distances.
2) Appliquer l'algorithme de classification hiérarchique ascendant à ces données. Pour cela, utiliser la méthode du lien minimal pour construire le dendrogramme.

3) Si l'on effectue un clustering avec k = 2, quels clusters obtient-on ?
4)



Exercice 3 :

Cet exercice consiste en différentes études de cas. Quatre cas pratiques nécessitant d'appliquer une approche d'analyse de données (AD) vous sont présentés.

1.1 Vous devrez tout d'abord identifier à quelle approche d'AD vue dans le cours chaque cas s'adresse. Expliquez brièvement votre choix.
1.2 Puis, pour chaque cas, vous répondrez aux questions spécifiques, associées à chaque approche d'AD en vous basant sur le cas pratique associé.

Description des cas pratiques :

1er cas:
Un chercheur en santé publique décide d'analyser l'incidence de la pollution radioactive sur le taux de cancer au sein d'une population donnée. Des relevés du taux d'iode radioactif ont été effectués sur 40 ans en prélevant des échantillons d'eau dans une rivière adjacente à la centrale nucléaire. En parallèle, le nombre total de cancers de la thyroïde déclarés, par an, dans l'hôpital avoisinant ont été enregistrés sur la même période.

2ème cas :
Un célèbre scientifique monte une expédition d'exploration de la canopée, la partie la plus élevée de de la forêt amazonienne, celle qui reçoit directement les rayons du soleil. L'objectif de cette expédition consiste à répertorier et comparer la diversité biologique de la faune et de la flore de cet écosystème. L'entomologiste de l'équipe capture (puis relâche) 6532 insectes en 30 jours. Chaque insecte est alors décrit selon sa couleur, sa taille, nombre d'ailes, forme de la tête, forme des antennes, forme de l'abdomen, type de bouche, diurnes ou nocturnes. Ce chercheur souhaite répertorier l'ensemble de ces insectes, les comparer et analyser les relations entre groupes d'individus.

3ème cas :
Un site de vente de livres en ligne souhaite analyser les différents profils « d'acheteurs » de ses clients. Bénéficiant d'une base de 256082 clients, ils ont gardé la trace de tous leurs achats sur une période de 5 ans. La librairie proposée par ce site se divise en 8 catégories : scolaire et académique, jeunesse, roman, essais, bande-dessinée, science-fiction, fantastique et enfin, arts et loisirs. Pour chacun de ses clients, le site connaît le nombre de livre acheté dans chaque catégorie. A partir de ces informations, le site souhaiterait identifier des différents profils types de ses clients.

4ème cas :
Une société souhaite commercialiser un outil de tri et d'empaquetage de fruits (des pommes par exemple) pour les petites exploitations agricoles. Associé à cette machine, un logiciel doit être capable d'identifier la qualité d'un fruit et l'orienter vers le circuit de distribution ou vers les déchets. Pour chaque fruit passant dans la chaîne de tri, une série de mesure est effectuée à l'aide de capteurs mesurant la densité, le gradient de couleur, le poids et le taux de dégagement de méthane (un indicateur du mûrissement). Le but final étant que le logiciel parvienne à reconnaître automatiquement si un fruit est bon ou trop mûr à partir de cette série de mesures. Ainsi, le tri, circuit de distribution/déchet, pourra être effectué de façon automatique.