

Exercice 1 : Calcul de distance (Préparation TP)

Les Iris de différentes familles sont caractérisées par des longueurs et des largeurs de pétales et de sépales différentes.

Une iris est définie par :

- La longueur de son sépale en cm (sepal length)
- La largeur de son sépale en cm (sepal width)
- La longueur de son pétale en cm (petal length)
- La largeur de son pétale en cm (petal width)

Summary Statistics:

Attribute	Min	Max	Mean	SD
sepal length:	4.3	7.9	5.84	0.83
sepal width:	2.0	4.4	3.05	0.43
petal length:	1.0	6.9	3.76	1.76
petal width:	0.1	2.5	1.20	0.76

Q1: Proposer une formule pour le calcul de distance entre 2 iris

Q2 : Calculer la distance entre ces 3 iris

Iris 1 : 5.1,3.5,1.4,0.2

Iris 2 : 7.0,3.2,4.7,1.4

Iris 3 : 6.3,3.3,6.0,2.5

Exercice 2 : Clustering Kmeans

On se place dans un carré de côté 1, contenant les points suivants :

Point	Abscisse	Ordonnée
1	0.1	0.1
2	0.2	0.2
3	0.1	0.3
4	0.2	0.3
5	0.4	0.9
6	0.4	0.7
7	0.5	0.8
8	0.3	0.8
9	0.9	0.3
10	0.8	0.4
11	1.0	0.4
12	0.9	0.5

Q1 : Appliquez l'algorithme des K-Moyennes aux douze premiers points, en recherchant 3 clusters et en prenant comme centres initiaux les points (0.6,0.9), (1.0,0.9),(0.3,0.3). La distance utilisée sera la distance euclidienne.

Q2 : Sur une visualisation graphique, tracer la frontière des clusters

Exercice 3 : Clustering Hiérarchique

Voici une matrice de distance entre 5 points A,B,C,D,E :

	A	B	C	D	E
A	0				
B	7.40	0			
C	7.56	8.62	0		
D	5.01	6.03	12.47	0	
E	12.43	6.55	4.66	9.28	0

On suppose que les objets sont regroupés par un algorithme de clustering hiérarchique où la distance entre deux clusters est définie par la distance minimale entre deux points dans chaque cluster (single link).

Q1 : Dessinez le dendrogramme qu'on obtient si on applique cet algorithme de classification.

Q2 : On utilise ensuite complete link, obtient on les mêmes résultats ?

Exercice 4 : BeeFrox

Action	Situation_famille	Revenu	Catégorie_emission	Action	Heure	Durée*
A1	Couple	60000	Meteo	Passif	20h30	5
A2	Célibataire	20000	Musique	Zap	7h30	15
A3	Famille	35000	Meteo	Eteint	7h45	5
A4	Famille	40000	Emission_pognon	Demande d'info	18h	20
A5	Célibataire	35000	Musique	Zap	19h	6
A6	Couple	50000	Emission_pognon	Demande d'info	18h	15

Ce tableau donne la description des actions, le revenu indique le revenu annuel du foyer en Euros, la durée indique la durée écoulée entre le début de l'émission et l'action (en minute).

Q1. On se positionne dans l'espace tri-dimensionnel défini par les attributs «Catégorie_emission», «Revenu» et «Durée».

- Lesquelles des visites parmi A1, A2 et A3 sont les plus similaires.
- Donner la description, dans ce même espace, du nuplet A_{123} centre de A1, A2 et A3.
- Quel est le problème rencontré ?. Comment y remédier ?

Q2. On se positionne dans l'espace défini par les dimensions « Revenus », « Catégorie_emission » et « Durée ». On souhaite partitionner l'ensemble des visites en trois groupes.

Utiliser la méthode appropriée afin de répondre à cet objectif.

Exercice 5 (Examen 2014)

L'empire cherche à regrouper les planètes et les étoiles selon le nombre d'espèces qu'elles comportent, le nombre de bases rebelles, le diamètre de la planète et la présence de végétation. Les agents de l'empire ont relevé les informations suivantes :

	<i>Nb Espèces</i>	<i>Nb Bases</i>	<i>Diamètre</i>	<i>Végétation</i>
<i>P1</i>	<i>11</i>	<i>1</i>	<i>120536</i>	<i>oui</i>
<i>P2</i>	<i>5</i>	<i>2</i>	<i>6700</i>	<i>non</i>
<i>P3</i>	<i>10</i>	<i>4</i>	<i>12756</i>	<i>oui</i>
<i>P4</i>	<i>7</i>	<i>2</i>	<i>12100</i>	<i>non</i>
<i>P5</i>	<i>2</i>	<i>0</i>	<i>4880</i>	<i>non</i>
<i>P6</i>	<i>0</i>	<i>0</i>	<i>2300</i>	<i>non</i>
<i>P7</i>	<i>8</i>	<i>2</i>	<i>12100</i>	<i>oui</i>

Question 1. Définissez formellement une distance permettant de considérer tous les attributs. Donnez la distance de P1 à P2 avec la distance précédemment définie.

Question 2. Les généraux de l'empire proposent la matrice de distance suivante (ne comparez pas ces distances avec votre réponse à Q1, on ne peut pas faire confiance à l'empire) :

	<i>P1</i>	<i>P2</i>	<i>P3</i>	<i>P4</i>	<i>P5</i>	<i>P6</i>	<i>P7</i>
<i>P1</i>	<i>0</i>	<i>0.462</i>	<i>0.666</i>	<i>0.538</i>	<i>0.478</i>	<i>0.334</i>	<i>0.666</i>
<i>P2</i>		<i>0</i>	<i>0.428</i>	<i>0.466</i>	<i>0.334</i>	<i>0.52</i>	<i>0.786</i>
<i>P3</i>			<i>0</i>	<i>0.5</i>	<i>0.44</i>	<i>0.652</i>	<i>0.846</i>
<i>P4</i>				<i>0</i>	<i>0.482</i>	<i>0.52</i>	<i>0.572</i>
<i>P5</i>					<i>0</i>	<i>0.636</i>	<i>0.76</i>
<i>P6</i>						<i>0</i>	<i>0.487</i>
<i>P7</i>							<i>0</i>

Proposez un regroupement via la méthode single-link, puis via complete-link. Vous détaillerez les calculs et donnerez le dendogramme obtenu.