Université d'Angers **Data mining**
master DS1

TD CNS : Classification non supervisée

Exercice 1 Afin de comprendre l'évolution du singe, des chercheurs souhaitent faire une classification des primates. Pour ce faire, l'ADN d'un individu de chaque espèce est analysé : 1 est un humain, 2 est un chimpanzé, 3 est un bonobo, 4 est un gorille, 5 est un orang-outan et 6 est un gibbon. Les distances euclidiennes entre chacun de ces individus, caractérisant leur ressemblance quant à l'ADN, sont données dans le tableau incomplet suivant :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0	0,84	0.95	1.49	1.85	2.56
ω_2	0.84	0	0.7	1.11	1.87	2.38
ω_3	0.95	0.7	0	1.35	2,05	2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56	2,38	2.15	2.32	2.30	0

1. Compléter le tableau.

2. Faire une classification par l'algorithme CAH muni du saut simple en complétant les tableaux suivants.

$C_{-7} = \{\omega_2, \omega_3\}$ $\Delta_{-7} = 0.7$ $C_{-8} = \{\omega_2, \omega_3, \omega_1\}$ $\Delta_{-8} = 0,84$ $C_{-9} = \{C_{-8}, \omega_4\}$ $\Delta_{-9} = 1.11$

d^2	ω_4	ω_5	ω_6	C_{-7}
ω_1	1.49	1.85	2.56	0,84
ω_4		1.96	2.32	1.11
ω_5			2.30	1.87
ω_6				2.15

d^2	ω_5	ω_6	C_{-8}
ω_4	1.96	2.32	1.11
ω_5		2.30	1.85
ω_6			2.15

d^2	ω_6	C_{-9}
ω_5	2.30	1.85
ω_6		2.15

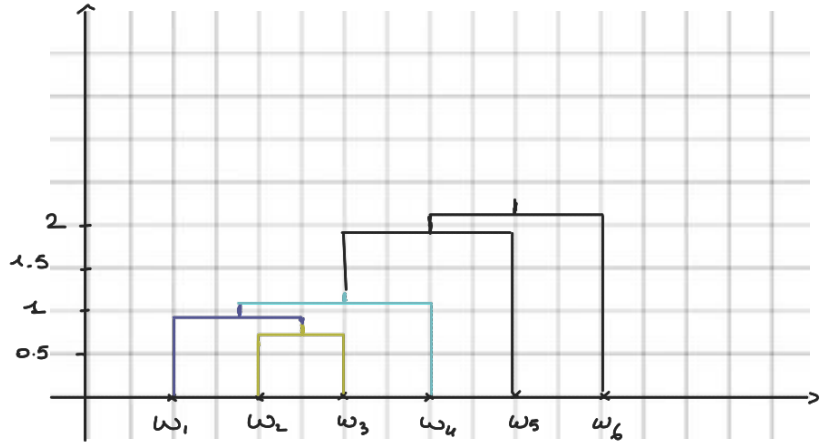
d^2	C_{-10}
ω_6	2.15

$C_{-10} = \{C_{-9}, \omega_5\}$ $\Delta_{-10} = 1.85$

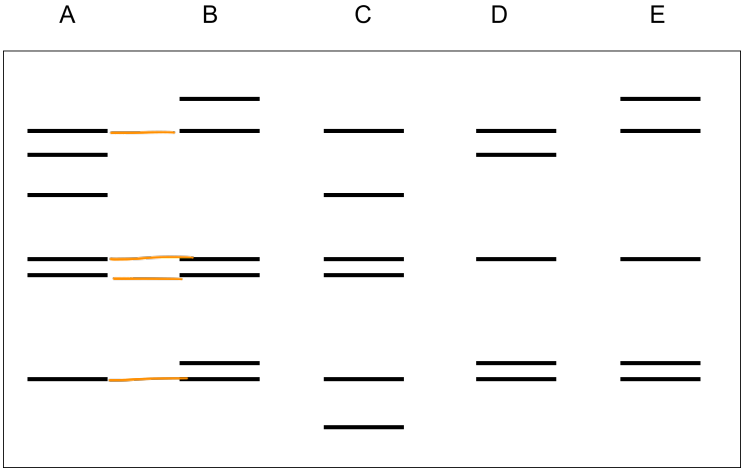
$C_{-11} = \{C_{-10}, \omega_6\}$ $\Delta_{-11} = 2.15$

La plus petite des distances

3. Construire le diagramme des indices et le dendrogramme.



Exercice 2 Un laboratoire veut étudier la ressemblance génétique de cinq lignées de sorgho (A, B, C, D, E). Pour cela, il réalise une analyse par RFLP avec l'enzyme de restriction Eco RI et une sonde d'origine inconnue. Les fragments d ?ADN ainsi amplifiés sont ensuite séparés par electrophorèse. Les résultats de l ?electrophorèse donnent les profils suivants :



NB : Deux lignées de sorgho présentent autant de fragments d ?ADN identiques que de bandes révélées à la même hauteur sur les profils de l ?electrophorèse.

1. Construire le tableau de similarité basé sur l'indice de Dice, S_{xy} défini par

$$S_{xy} = 2\frac{N_{xy}}{N_x + N_y}$$

avec N_{xy} le nombre de bandes communes, N_x et N_y le nombre de bandes de chaque lignée. En déduire l'indice de dissimilarité $D_{xy} = 1 - S_{xy}$.

S_{xy}	A	B	C	D	E
A	1	8/12	10/12	8/11	6/11
B		1	8/12	8/11	10/11
C			1	6/11	6/11
D				1	8/10
E					1

D_{xy}	A	B	C	D	E
A	0	4/12	2/12	5/11	5/11
B		0	4/12	3/11	1/11
C			0	5/11	5/11
D				0	2/10
E					0

pour A - B
$N_{AB} = 4$
$N_A = 6$
$N_B = 6$
$S_{AB} = 2 \times \frac{4}{12} = \frac{8}{12}$

2. Faire une classification par l'algorithme CAH avec la méthode du voisin le plus éloigné et la dissimilarité proposée. Tracer le dendrogramme associé.

D_{xy}	C	D	C-6
A	2/12	5/11	5/11
C		5/11	5/11
D			3/11

D_{xy}	C-6	C-7
D	3/11	5/11
C-6		5/11

D_{xy}	C-8
C-7	5/11

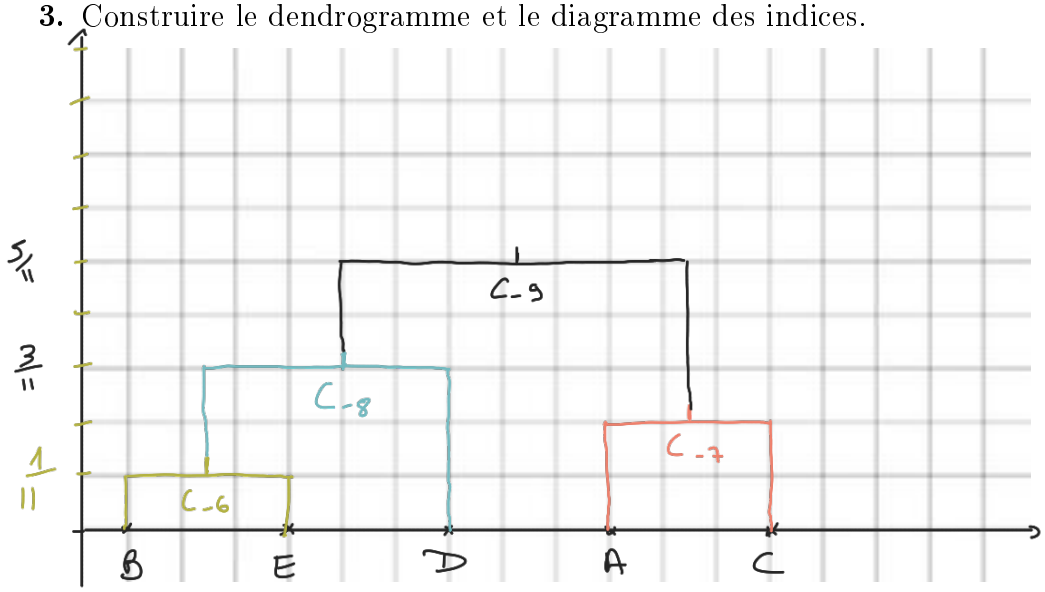
$C_{-6} = \{B, E\}$ $\Delta_{-6} = \frac{1}{11}$

$C_{-7} = \{A, C\}$ $\Delta_{-7} = \frac{2}{12}$

$C_{-8} = \{C_{-6}, D\}$ $\Delta_{-7} = \frac{3}{11}$

$C_{-9} = \{C_{-7}, C_{-8}\}$ $\Delta_{-9} = \frac{5}{11}$

3. Construire le dendrogramme et le diagramme des indices.



4. On appelle définit le coefficient d'aggomération par

$$AC = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{I(\omega_i, A_i)}{I(Q,R)}\right),$$

avec A_i le premier élément avec lequel l'élément ω_i a été regroupé, Q et R les deux éléments rassemblés à l'étape finale de l'algorithme et I l'indice du regroupement.

a. Calculer AC dans l'exercice.

b. Déterminer les valeurs possibles de AC ?

c. Comment interpréter une valeur proche de 0 ? proche de 1 ?

Exercice 3 Valeur test

On considère un ensemble de $n = n_1 + dots + n_q$ individus répartis en q classes. Soit X une variable quantitative mesurée sur chaque individu. Pour établir que cette variable caractérise une classe k donnée, on calcule sa valeur test donnée par :

$$\text{v-test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{n_q}\left(\frac{n-n_q}{n-1}\right)}}.$$

avec $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ la variance de X. On pose comme hypothèse H_0 que les individus de la classe k sont tirés au hasard sans remise parmi l'ensemble des n individus. Soit $\{x_i, i \in E\}$ les n_q variables échantillonnées au hasard parmi les n valeurs possibles. On pose alors la moyenne de l'échantillon E

$$\bar{e} = \frac{1}{n_q}\sum_{i\in E} x_i = \frac{1}{q}\sum_{i=1}^{n}\delta_i x_i$$

avec $\delta_i = 1$ si $i \in S$, 0 sinon. La difficulté vient ensuite de la non indépendance des (δ_i).

1. Montrer $P(\delta_i = 0) = \frac{n-n_q}{n}$ et en déduire $E(\delta_i) = \frac{n_q}{n}$ et $\text{var}(\delta_i) = \frac{n_q}{n}\left(1 - \frac{n_q}{n}\right)$.

$\mathbb{P}(\delta_i = 0) = \dfrac{\binom{n-1}{n_q}}{\binom{n}{n_q}}$ ← sans i ← S $= \dfrac{(n-1)!}{n_q!(n-1-n_q)!} \times \dfrac{n_q!(n-n_q)!}{n!} = \dfrac{1}{n}\cdot\dfrac{(n-n_q)!}{(n-n_q-1)!} = \dfrac{n-n_q}{n}$

$\mathbb{E}(\delta_i) = \frac{n_q}{n}$ car $\delta_i \sim \text{Bernoulli}\left(\frac{n_q}{n}\right)$ $\quad \text{Var} = \frac{n_q}{n}\left(1 - \frac{n_q}{n}\right)$

2. Montrer que $\text{cov}(\delta_i, \delta_j) = \frac{n_q(n_q-1)}{n(n-1)} - \left(\frac{n_q}{n}\right)^2 = \frac{1}{n-1}\frac{n_q}{n}\left(\frac{n_q}{n} - 1\right)$.

$\text{cov}(\delta_i, \delta_j) = \mathbb{E}[\delta_i \delta_j] - \mathbb{E}[\delta_i]\,\mathbb{E}[\delta_j] \quad i \neq j$

$= \dfrac{\binom{n-2}{n_q-2}}{\binom{n}{n_q}}$ ← # i,j $- \left(\dfrac{n_q}{n}\right)^2 = \dfrac{(n-2)!}{(n_q-2)!\,(n-n_q)!} \times \dfrac{(n-n_q)!\,n_q!}{n!} - \left(\dfrac{n_q}{n}\right)^2$

$= \dfrac{n_q(n_q-1)}{n(n-1)} - \left(\dfrac{n_q}{n}\right)^2 =$

3. Montrer que $E(\bar{e}) = \bar{X}$ et

$$\text{var}(\bar{e}) = \text{var}\left(\frac{1}{n_q}\sum_{i=1}^{n} x_i\delta_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} x_i^2 \text{var}(\delta_i) + \sum_{i\neq j}^{n} x_i x_j \text{cov}(\delta_i, \delta_j)\right] = \frac{n-n_q}{n-1}\frac{s^2}{n_q}$$

en remarquant que $\sum_{i=1}^{n} x_i^2 - \frac{1}{n-1}\sum_{i\neq j} x_i x_j = \frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

$$\mathbb{E}(\bar{e}) = \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n}\delta_i x_i\right] = \frac{1}{n_q}\sum_{i=1}^{n} x_i \underbrace{\mathbb{E}[\delta_i]}_{\frac{n_q}{n}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\mathrm{Var}(\bar{e}) = \mathrm{Var}\left(\frac{1}{n_q}\sum_{i=1}^{n}\delta_i x_i\right) = \mathrm{cov}\left(\frac{1}{n_q}\sum_{i=1}^{n}\delta_i x_i, \frac{1}{n_q}\sum_{i=1}^{n}\delta_i x_i\right)$$

$$= \frac{1}{n_q^2}\left[\sum_{i=1}^{n} x_i^2 \mathrm{Var}(\delta_i) + \sum_{i\neq j}^{n} x_i x_j \mathrm{cov}(\delta_i, \delta_j)\right]$$

$$= \frac{1}{n_q^2}\left[\sum_{i=1}^{n} x_i^2 \frac{n_q}{n}\left(1-\frac{n_q}{n}\right) + \sum_{i\neq j}^{n} x_i x_j \frac{1}{n-1}\frac{n_q}{n}\left(\frac{n_q}{n}-1\right)\right]$$

$$= \frac{1}{n n_q}\left(1-\frac{n_q}{n}\right)\left[\sum_{i=1}^{n} x_i^2 - \sum_{i\neq j}^{n} x_i x_j \frac{1}{n-1}\right] = \frac{1}{n n_q}\frac{n-n_q}{n} \times \frac{n}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \frac{n-n_q}{n-1} \times \frac{s^2}{n_q}$$

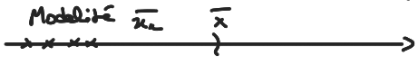
4. En considérant que la moyenne suit une loi normale pour n_q et n assez grands, en déduire une règle de décision pour H_0.

On approche $\bar{x}_{n_q}$ par $\mathcal{N}\left(\mathbb{E}[\bar{e}], \mathrm{Var}(\bar{e})\right)$

$$\mathbb{P}\left(-2 \leq \frac{\bar{x}_{n_q} - \bar{x}}{\sqrt{\frac{n-n_q}{n-1}\frac{s^2}{n_q}}} \leq 2\right) \simeq 0.95$$

5. Ce critère est utilisé dans différentes méthodes. Rappeler les cas où on l'utilise et comment.

ACM : relation entre un axe ou une variable quantitatif supplémentaire et une modalité.



Cas d'une variable qualitative

Quelles modalités d'une variable qualitative caractérise une classe k ?

On obtient des tableaux de contingence de la forme :

Modalité	C_1	C_2	C_3	$\sum$
m	$n_{mk} = 6$	3	0	$n_m = 9$
r	0	4	0	4
s	0	0	4	4
t	0	8	0	8
$\sum$	$n_k = 6$	15	4	$n = 25$

grp k

0 1 0 0 1 | 1 1 0 0 1 1 | 0 1 0 1 0 0

On souhaite tester l'hypothèse H_0 : $\frac{n_{mk}}{n_k} = \frac{n_m}{n}$ ⇒ esp : $\frac{\text{nb de 1 dans grp k}}{\text{nb d'effectifs dans grp k}} = \frac{\text{nb de 1 au tot}}{\text{nb d'ag au tot}}$

1. Interpréter cette hypothèse.

$x_i = \begin{cases} 1 & \text{si } m_i = m \\ 0 & \text{sinon} \end{cases}$

2. Justifier que sous H_0 :

$$\text{v-test} = \frac{\frac{n_{mk}}{n_k} - \frac{n_m}{n}}{\sqrt{\frac{(n-n_k)}{n-1}\frac{s^2}{n_k}}} \sim \mathcal{N}(0,1)$$

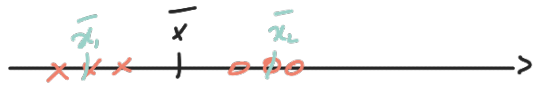
avec $s^2 = \frac{n_m}{n}(1 - \frac{n_m}{n})$.

3. Compléter le tableau suivant et interpréter le :

	cla/mod	mod/cla	global	p.value	v.test
m	0.67 6/9	1 6/6	0.36 9/25	0.00047	3.5
s	0/4	0/6	4/25		

Variable quantitative ou axe factoriel caractérisant une partition

Rappeler la définition de η^2 et construire le test de Fisher correspondant à l'effet de la partition sur la variable quantitative (ANOVA).



$SCT = SC_{intra} + SC_{inter}$

$\eta^2 = \frac{SC_{inter}}{SC_{tot}} = \frac{\sum (\bar{x}_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}$

Variable qualitative caractérisant une partition

Imaginer un test analogue ici.

test de χ^2 entre les groupes et les variables qualitatives.

$F = \frac{SC_{inter}/q-1}{SC_{intra}/n-q}$

$= \frac{\eta^2}{1-\eta^2} \times \frac{n-q}{q-1}$

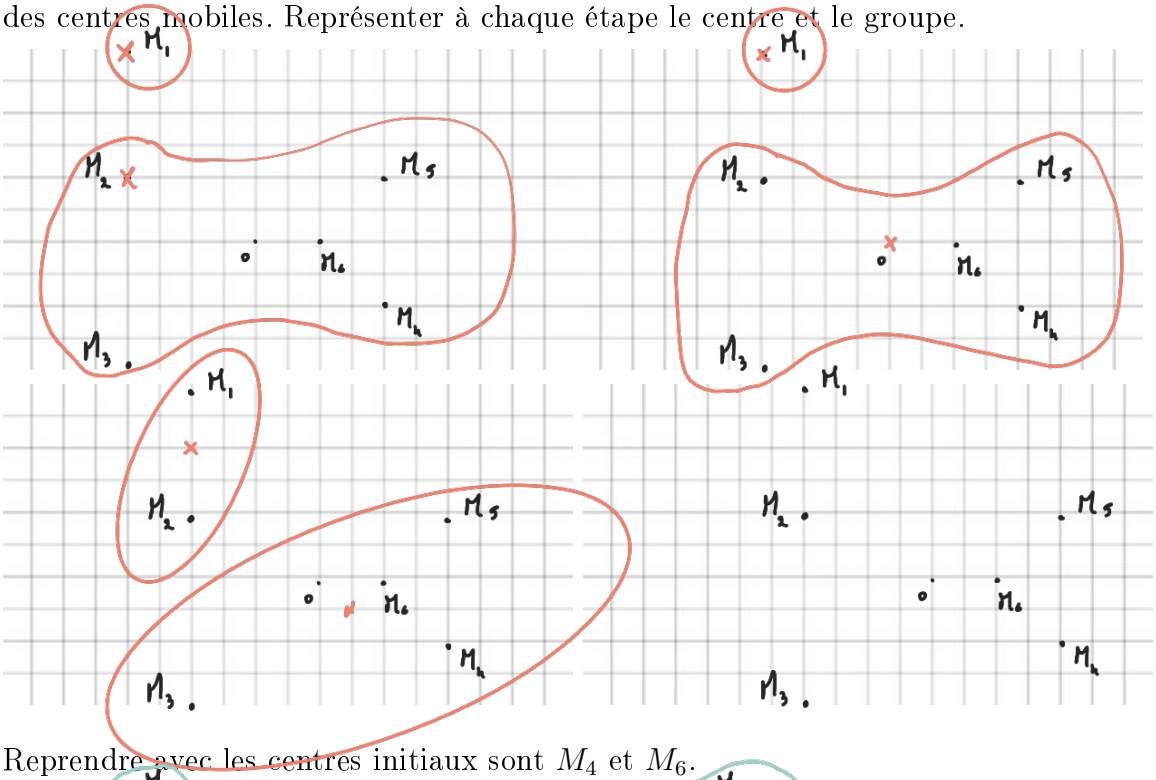
Exercice 4 Soit X le caractère donnant la taille en mètres d'un client qui entre dans un magasin de vêtements masculins au centre ville de Caen. On suppose que :

- si le client est une femme, X suit la loi N (1.65, 0.162),
- si le client est un homme, X suit la loi N (1.75, 0.152),
- la probabilité qu'un client homme rentre est 0.7.

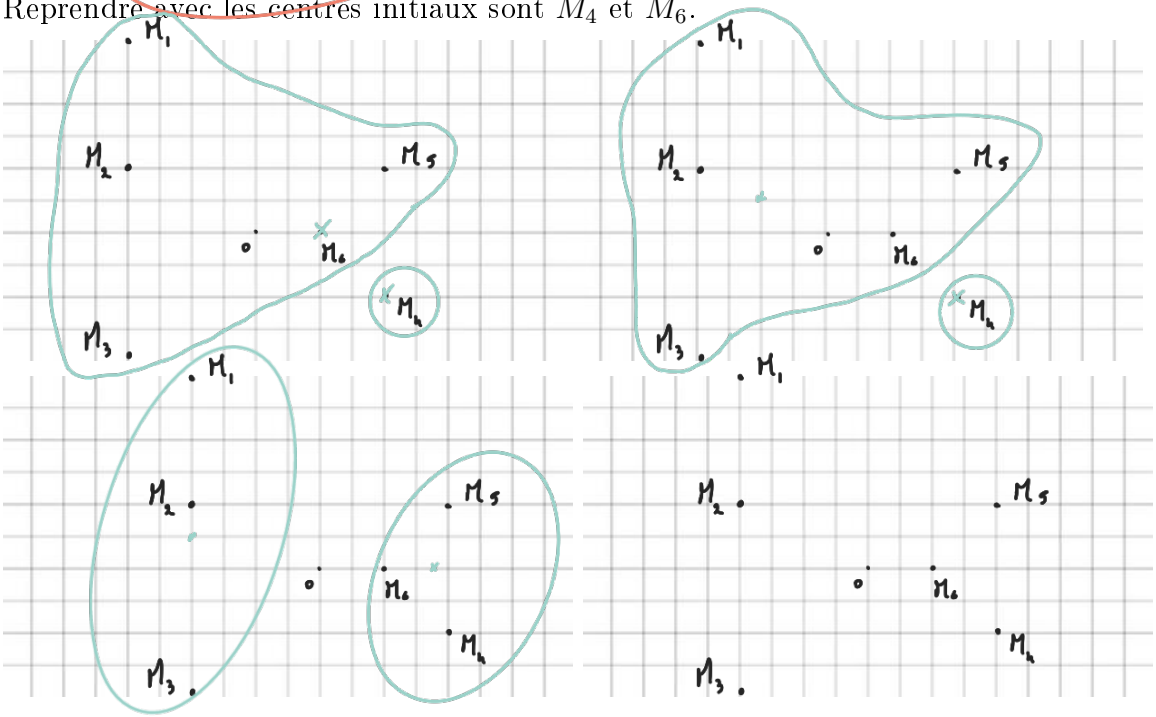
Un client rentre dans le magasin. On sait qu'il mesure 1.60 mètres. Qu'elle est la probabilité que ce soit un homme ?

Exercice 5 Soit 6 points $M_1(-2,3)$, $M_2(-2,1)$, $M_3(-2,-2)$, $M_4(2,-1)$, $M_5(2,1)$ et $M_6(1,0)$.

1. En supposant que les centres initiux sont M_1 et M_2 décrire les étapes de l'algorithme des centres mobiles. Représenter à chaque étape le centre et le groupe.



2. Reprendre avec les centres initiaux sont M_4 et M_6.



3. Quel critère pourrait-on proposer pour choisir la meilleure ?

4. Classifier les points à l'aide d'une CAH utilisant l'indice de Ward $W(i,j) = \dfrac{p_i p_j}{p_i + p_j} d^2(M_i, M_j)$. M_i, M_j représente le point ou le centre de gravité de la classe. On calculera les distances puis l'indice. Construire le dendrogramme.

d^2	M_2	M_3	M_4	M_5	M_6
M_1	4	16	32	20	18
M_2	0	4	32	16	18
M_3		0	16	20	10
M_4			0	4	2
M_5				0	2

W	M_2	M_3	M_4	M_5	M_6
M_1					
M_2	0				
M_3		0			
M_4			0		
M_5				0	

d^2	...	...	...	...
	0			
		0		
			0	

W	...	...	...	...
	0			
		0		
			0	

d^2	...	...	...
	0		
		0	
		0	

W	...	...	...
	0		
		0	
		0	

d^2	...	...
	0	

W	...	...
	0	

d^2	...

W	...

Exercice 6 On étudie le modèle simple de la loi Zero inflated Poisson qui est un mélange entre une loi discrète telle que $P(X = 0) = 1$ (avec une proportion π) et une loi de Poisson classique de paramètre λ (avec une proportion $1 - \pi$).

1. Spécialisation de l'algorithme

   Écrire la vraisemblance des paramètres d'une loi ZIP pour les observations $(x_i)_{1 \leq i \leq n}$. supposées issues d'une telle loi.

   Écrire l'algorithme EM dans le cas spécifique de la loi ZIP.

2. Mise en oeuvre en R

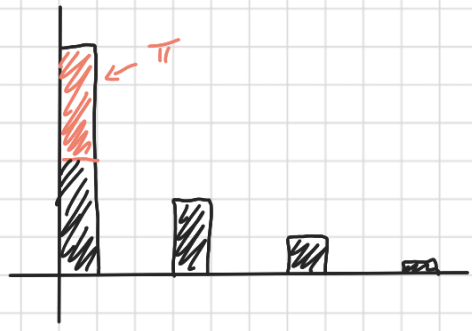
   Écrire une fonction R rzip qui simule des réalisations indépendantes d'une loi ZIP de paramètres π et λ. On s'appuiera sur la fonction rpois qui simule des réalisations d ?une loi Poisson.

   Écrire une fonction R emzip qui estime les paramètres π et λ d'une loi ZIP à partir d'un échantillon en maximisant la vraisemblance par l'algorithme EM.

   En utilisant la fonction optim de R, construire une autre fonction directzip qui estime les paramètres π et λ d'une loi ZIP à partir d'un échantillon en maximisant la vraisemblance de façon directe (sans passer par l'algorithme EM).

Exercice 6 =

1 -



$$\pi \delta_0 + (1-\pi) \cdot \sum_{k=1}^{+\infty} P_\lambda(X=k)\,\delta_k \qquad \text{or } P_\lambda(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\left(l(\underline{X};\phi) = \prod_{i=1}^{n} \left[\pi \delta_0 + (1-\pi) \cdot \sum_{k=1}^{+\infty} \frac{\lambda^k}{k!} e^{-\lambda} \delta_k \right] \right) ?$$

$$ll(\underline{X};\phi) = \sum_{i=1}^{n} \ln$$