TD : Classification

Exercice 1. Soit $\mathbf{X}$ la matrice de données associée à 4 individus $\omega_1, \ldots, \omega_4$ définie par

$$\mathbf{X} = \begin{pmatrix} 2.0 & 3.0 \\ 7.0 & 4.0 \\ 3.5 & 3.0 \\ 0.5 & 5.0 \end{pmatrix}$$

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_4\}$ et donner son inertie totale.
2. Déterminer le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\})$ obtenu par la méthode du plus proche voisin.
3. Au hasard, on forme le groupe d'individus : $\{\omega_3, \omega_4\}$. Déterminer le tableau des écarts associé à $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3, \omega_4\})$ obtenu par la méthode du plus proche voisin.

Exercice 2. Dans une étude en sciences sociale, on a étudié 2 caractères X_1 et X_2 sur 5 individus $\omega_1, \ldots, \omega_5$. Les données recueillies sont :

	X_1	X_2
ω_1	2.4	2.1
ω_2	7.1	4.3
ω_3	3.8	3.0
ω_4	1.2	5.1
ω_5	6.5	4.3

1. Déterminer la matrice des données.
2. Tracer le nuage de points $\mathcal{N}$ formé par $\Gamma = \{\omega_1, \ldots, \omega_5\}$.
3. Déterminer le centre de gravité g de $\mathcal{N}$, le tracer, et calculer les distances entre les individus et celui-ci.
4. Déterminer l'inertie totale de $\mathcal{N}$.
5. (a) Faire une classification avec l'algorithme CAH muni de la méthode du voisin le plus éloigné.
   (b) Tracer le dendrogramme associé. Proposer 2 groupes d'individus semblables.
6. (a) Faire une classification avec l'algorithme CAH muni de la méthode de Ward. À chaque nouvelle partition, indiquer l'inertie intra-classes associée.
   (b) Tracer le dendrogramme associé. Proposer 2 groupes d'individus semblables.
   (c) Déterminer le parangon du groupe $\{\omega_1, \omega_3, \omega_4\}$.
7. Comparer les classifications obtenues par les deux méthodes précédentes.

Exercice 3. Décrire brièvement l'enjeu des commandes R suivantes :

```
x = matrix(c(1, 16, 2, 9, 10, 16, 1, 17, 15, 2, 1, 37, 0, 14, 9, 9, 12, 4, 3, 13), ncol = 5,
nrow = 4)

library(cluster)
ag = agnes(x, method = "average")
ag$merge
```

Exercice 4. 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 : ω_1 est un humain, ω_2 est un chimpanzé, ω_3 est un bonobo, ω_4 est un gorille, ω_5 est un orang-outan et ω_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 :

$\mathbf{E} =$

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0		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		1.35		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		2.15	2.32	2.30	0

1. Recopier et compléter $\mathbf{E}$.
2. Faire une classification par l'algorithme CAH muni de la méthode du plus proche voisin. Tracer le dendrogramme associé.
3. Faire une classification par l'algorithme CAH muni de la méthode du voisin le plus éloigné. Tracer le dendrogramme associé.
4. Faire une classification par l'algorithme CAH muni de la méthode de la distance moyenne. Tracer le dendrogramme associé.
5. Proposer des commandes R pour répondre aux questions précédentes.

Exercice 5. Le tableau des distances associé à la partition initiale $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_4\})$ est

	ω_1	ω_2	ω_3	ω_4
ω_1	0	2	4	7
ω_2	2	0	4	5
ω_3	4	4	0	3
ω_4	7	5	3	0

1. Faire une classification avec l'algorithme CAH muni de la méthode de la distance moyenne.
2. Calculer le coefficient d'agglomération. Est-ce que la classification obtenue est fortement structurée en groupes ?

Exercice 6. Comprendre pourquoi la sortie finale des commandes R suivantes est `TRUE` :

```
x = matrix(c(2, 7.5, 3, 0.5, 6, 2, 4, 3, 5, 4), ncol = 2)

library(cluster)
ag = agnes(x, method = "complete")
pltree(ag, hang = -1)

a = (1 / 5) * ((1 - 1.41 / 7.07) + (1 - 1.5 / 7.07) + (1 - 1.41 / 7.07) + (1 - 1.5 / 7.07)
+ (1 - 3.35 / 7.07))

round(ag$ac, 2) == round(a, 2)
```

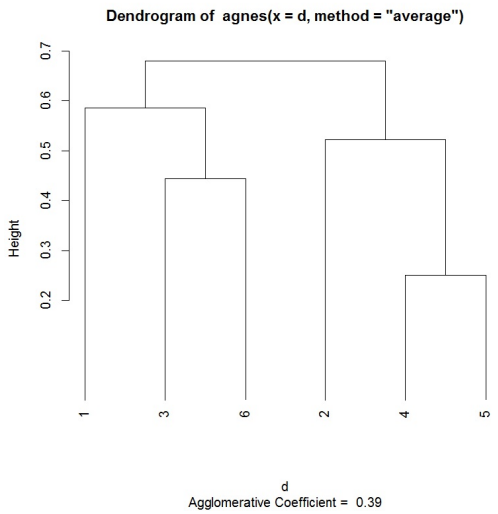
Exercice 7. Thomas, Fatiha, Emilie, Robert, Igor et Mathias sont malades. Ci-dessous quelques données relatives à leur maladie : Sexe $\in$ {Homme, Femme}, Fièvre $\in$ {Oui, Non}, Toux $\in$ {Positif, Négatif} et pour les 4 tests : Test $\in$ {Positif, Négatif} :

Nom	Sexe	Fièvre	Toux	Test 1	Test 2	Test 3	Test 4
Thomas	H	O	P	P	P	N	P
Fatiha	F	O	N	P	N	P	P
Emilie	F	O	P	N	N	N	N
Robert	H	O	N	P	N	P	N
Igor	H	O	N	P	P	P	N
Mathias	H	O	P	P	N	N	N

On désire classer ces individus en 2 groupes suivant la ressemblance de leur diagnostique.

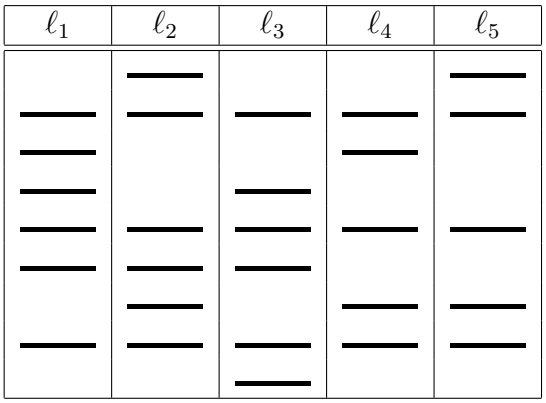
1. Dresser le tableau disjonctif complet.

2. Proposer des commandes R pour faire une classification avec l'algorithme CAH muni de la méthode de la distance moyenne avec la dissimilarité de Jaccard permettant d'obtenir le dendrogramme suivant :



3. Au vu du dendrogamme, quels sont les 2 groupes que vous suggéreriez ?

Exercice 8. La ressemblance génétique de cinq lignées d'une céréale (ℓ_1, ℓ_2, ℓ_3, ℓ_4 et ℓ_5) peut se mesurer par le nombre des bandes à la même hauteur révélées sur des profils obtenu par une certaine technique biologique (électrophorèse). Après analyse, on obtient les profils suivants:



Dans de contexte, on considère l'indice de similarité de Dice défini par :

$$s_{\ell,\ell_*} = \frac{2N_{\ell,\ell_*}}{N_\ell + N_{\ell_*}},$$

où N_ℓ est le nombre de bandes de la lignée ℓ, N_{ℓ_*} est le nombre de bandes de la lignée ℓ_* et N_{ℓ,ℓ_*} est le nombre de bandes à la même hauteur entre les lignées ℓ et ℓ_*. De même, on considère la dissimilarité de Dice défini par :

$$d_{\ell,\ell_*} = 1 - s_{\ell,\ell_*}.$$

1. Recopier et compléter le tableau des indices de similarités :

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1		$\frac{8}{12}$			$\frac{6}{11}$
ℓ_2					
ℓ_3			1		
ℓ_4					$\frac{8}{10}$
ℓ_5					

   En déduire le tableau des dissimilarités, avec des valeurs décimales.

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

3. Calculer le coefficient d'agglomération.

Exercice 9. On a relevé les valeurs de 2 caractères X_1 et X_2 sur 6 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

	X_1	X_2
ω_1	-2	3
ω_2	-2	1
ω_3	-2	-1
ω_4	2	-1
ω_5	2	1
ω_6	1	0

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_6\}$.
2. Faire une classification avec l'algorithme des centres mobiles avec, pour centres initiaux, ω_1 et ω_2.
3. Reprendre la question précédente avec, pour centres initiaux, ω_4 et ω_6.

Exercice 10. On a relevé les valeurs de 2 caractères : X_1 et X_2, sur 7 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

	X_1	X_2
ω_1	1.42	3.58
ω_2	4.23	4.65
ω_3	0.34	1.04
ω_4	1.30	1.95
ω_5	2.56	3.46
ω_6	1.33	4.67
ω_7	3.17	4.59

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_7\}$.
2. Faire une classification par l'algorithme des centres mobiles avec, pour centres initiaux, m_1 de coordonnées $(1, 1)$ et m_2 de coordonnées $(3, 3)$.
3. Proposer des commandes R pour répondre à la question précédente.

Exercice 11. Dans une étude industrielle, on a étudié 2 caractères X_1 et X_2 sur 6 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

Groupe A		
ω_1	ω_2	ω_3
$(0, 0)$	$(1, 1)$	$(2, 2)$

Groupe B		
ω_4	ω_5	ω_6
$(6, 6)$	$(5.5, 7)$	$(6.5, 5)$

Déterminer le groupe d'appartenance d'un individu ω_* vérifiant $X_1 = 4$ et $X_2 = 4$ avec la méthode des k plus proches voisins (kNN) pour $k = 3$. Faire un dessin.

Exercice 12. 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 :

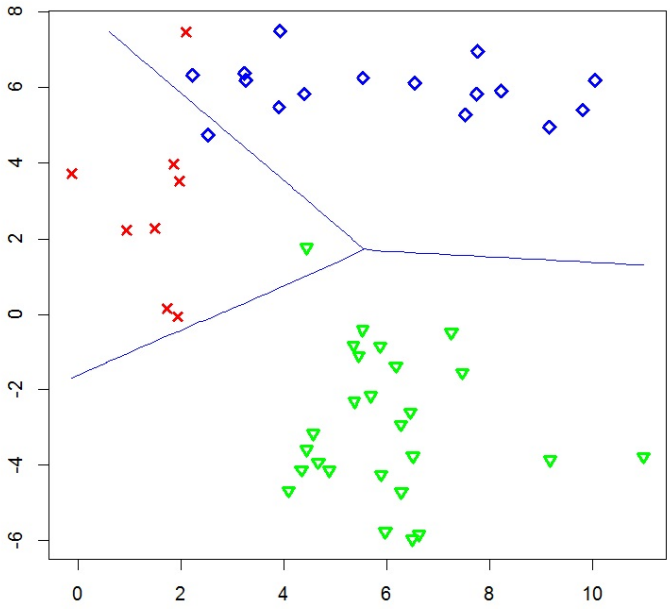
- si le client est une femme, X suit la loi $\mathcal{N}(1.65, 0.16^2)$,
- si le client est un homme, X suit la loi $\mathcal{N}(1.75, 0.15^2)$,
- la probabilité qu'un client homme rentre est 0.7.

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 13. Voici un exemple de commandes R illustrant la classification supervisée par l'analyse discriminante :

```
p1 = 0.2 ; p2 = 0.3 ; p3 = 0.5
n = 50
s1 = c(1, 2) ; s2 = c(3, 1) ; s3 = c(1.5, 2)
s = rbind(s1, s2, s3)

m1 = c(1, 2) ; m2 = c(6, 6) ; m3 = c(6, -2)
m = rbind(m1, m2, m3)

c = sample(c(1, 2, 3), size = n, prob = c(p1, p2, p3), replace = TRUE)

x = cbind(rnorm(n, m[c, 1], s[c, 1]), rnorm(n, m[c, 2], s[c, 2]))

couleur = rep("red", n)
couleur[c == 2] = "blue"
couleur[c == 3] = "green"
x11()

plot(x,col = couleur, pch = 3 + c, lwd = 3, , xlab = " ", ylab = " ")

library(MASS)
T = as.factor(couleur)
x.lda = lda(x, T)
len = 50
xp = seq(min(x[ ,1]), max(x[ ,1]), length = len)
yp = seq(min(x[,2]), max(x[,2]), length = len)
grille = expand.grid(z1 = xp, z2 = yp)

Z = predict(x.lda, grille)
T.lda = predict(x.lda)$class

zp = Z$post[ ,3] - pmax(Z$post[ ,2], Z$post[ ,1])
contour(xp, yp, matrix(zp, len), add = TRUE, levels = 0, drawlabels = FALSE, col = "blue")

zp = Z$post[ ,1] - pmax(Z$post[ ,2], Z$post[ ,3])
contour(xp, yp, matrix(zp, len), add = TRUE, levels = 0, drawlabels = FALSE, col = "blue")
```

On obtient :



Quelle est la proportion de points mal classés ?

Exercice 14. On cherche à modéliser la probabilité de survenue du cancer du poumon chez l'homme en fonction de trois caractères : l'âge, le poids et le tabagisme. Pour ce faire, 10000 hommes dans la population sont considérés. On mesure les valeurs de (Y, X_1, X_2, X_3), où Y est un caractère qui vaut 1 si l'homme présente un tel cancer et 0 sinon, X_1 est l'âge en années, X_2 est le poids de l'individu en kilogrammes et X_3 la quantité de cigarettes consommées ces 10 dernières années.

1. Proposer brièvement une modélisation adaptée au problème.
2. Expliquer par quelle méthode on obtient les estimations des paramètres inconnus du modèle proposé au résultat de la question 1.

Exercice 15. On considère le jeu de données "anesthésie" :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/anesthésie.txt", header = T)
```

Trente patients ont reçu un certain niveau de dosage d'agent anesthésique pendant 15 minutes. Puis une incision leur est faite. Il est ensuite noté si le patient a bougé ou pas lors de l'incision. Ainsi, pour chaque patient, on dispose :

- du dosage de l'agent anesthésique pendant 15 minutes (variable $X1$),
- du fait qu'il ait bougé ou pas (variable Y, avec $Y = 1$ pour bougé).

On souhaite expliquer Y en fonction de $X1$.

1. On exécute les commandes R suivantes :

```
attach(w)
library(stats)
reg = glm(Y ~ X1, family = binomial)
summary(reg)
predict.glm(reg, data.frame(X1 = 1.25), type = "response")
confint.default(reg, level = 0.95)
```

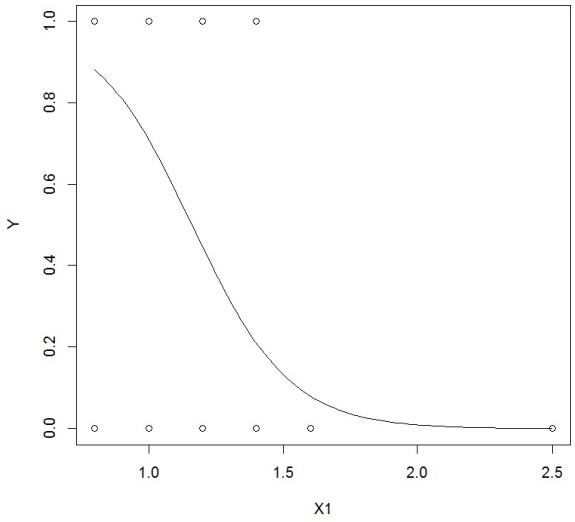
Cela renvoie :

```
> summary(reg)

Call:
glm(formula = Y ~ X1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06900  -0.68666  -0.03413   0.74407   1.76666

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.469      2.418   2.675  0.00748 **
X1            -5.567      2.044  -2.724  0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 27.754  on 28  degrees of freedom
AIC: 31.754

Number of Fisher Scoring iterations: 5

> predict.glm(reg, data.frame(X1 = 1.25), type="response")
        1
0.379946
> confint.default(reg, level=0.95)
                2.5 %    97.5 %
(Intercept)  1.728560 11.208790
X1          -9.572126 -1.561398
```

Quel est le modèle considéré ? Est-ce qu'un patient ayant eut pour dosage $\mathsf{X1} = 1.25$ a plus de chance de bouger que de ne pas bouger ?

2. On exécute les commandes R suivantes :

```
plot(X1, Y)
curve(predict(reg, data.frame(X1 = x), type = "response"), add = T)
```



Que représente ce graphique ?

3. On exécute les commandes R suivantes :

```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob >0.5, "1", "0"))
mc = table(Y, pred.mod)
t = (mc[1, 2] + mc[2, 1]) / sum(mc)
t
```

Cela renvoie $t = 0.2$.

Que représente cette quantité ? Est-ce que le résultat est satisfaisant ?