université de BORDEAUX Collège Sciences & Technologies	Année universitaire : **2021/2022** **Examen Analyse, Classification, Indexation des Données** Code UE : **4TIN703U** Date : **10 décembre** Heure : **9h00** Durée : **1h30** **Documents non autorisés** **4** pages

Nom :
Prénom :
Groupe :

Exercice 1 : Généralités

1. On dispose d'un nuage de points 2D $\{(x_1, y_1), \ldots, (x_m, y_m)\}$. On cherche la droite de régression linéaire correspondant à ce nuage. Il s'agit de la droite $y = \theta_1 x + \theta_0$ qui minimise le critère suivant :

$$\frac{1}{m} \sum (\theta_0 + \theta_1 x_i - y_i)^2$$

   (a) Expliquez à quoi correspond ce critère.

   (b) Proposez deux méthodes permettant de trouver la "meilleure" droite. On vous demande d'expliquer le principe de ces deux méthodes et pourquoi sont elles équivalentes.

2. On dispose d'un corpus de données (X, y). La figure suivante donne un extrait d'un `jupyter notebook` utilisé pour entraîner un model sur ce corpus. Prenez le temps de bien assimiler et comprendre la sortie de chaque cellule.

Entrée [3]:
```
print(X.shape)
print(y.shape)
```
```
(10000, 2)
(10000,)
```

Entrée [4]:
```
import numpy as np
print(np.unique(y))
print(len(y[y==0]))
print(len(y[y==1]))
```
```
[0 1]
9000
1000
```

Entrée [5]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.3)
```

Entrée [6]:
```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
```
Out[6]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')
```

Entrée [7]:
```
from sklearn.metrics import confusion_matrix
y_pred = knn.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
cm
```
Out[7]:
```
array([[2683,   15],
       [   5,  297]])
```

   (a) Expliquez le rôle de la cellule 5. En quoi cette étape est essentielle dans le processus d'entraînement d'un modèle de machine learning ?

   (b) Interprétez le résultat de la cellule 7. Quelle est l'"accuracy" du modèle entraîné ? Dans ce cas particulier, est-ce que l'accuracy est la bonne mesure de la qualité du modèle ?

Exercice 2 : Régression linéaire

On souhaite expliquer la nocivité des cigarettes (teneur en monoxyde de carbone – `CO`) à partir de leur composition : `TAR` (goudron) , `NICOTINE` et `WEIGHT` (poids) , soit $p = 3$ variables explicatives. Nous disposons de $n = 24$ observations.

On fait donc une régression linéaire et on obtient le rapport suivant :

OLS Regression Results

Dep. Variable:	CO(mg)	**R-squared (uncentered):**	0.993
Model:	OLS	**Adj. R-squared (uncentered):**	0.992
Method:	Least Squares	**F-statistic:**	1009.
Date:	Mon, 06 Dec 2021	**Prob (F-statistic):**	7.57e-23
Time:	16:53:27	**Log-Likelihood:**	-35.446
No. Observations:	24	**AIC:**	76.89
Df Residuals:	21	**BIC:**	80.43
Df Model:	3		
Covariance Type:	nonrobust		

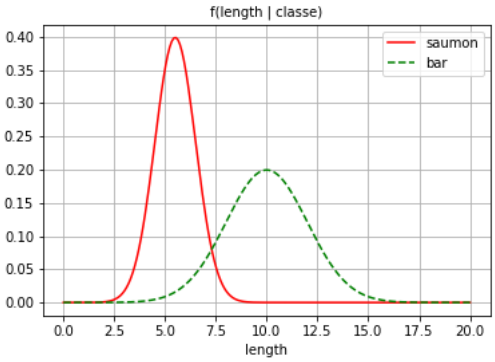
	coef	std err	t	P>|t|	[0.025	0.975]
TAR(mg)	0.8929	0.189	4.727	0.000	0.500	1.286
NICOTINE(mg)	0.4395	3.149	0.140	0.890	-6.110	6.989
WEIGHT(g)	1.5145	0.899	1.685	0.107	-0.355	3.383

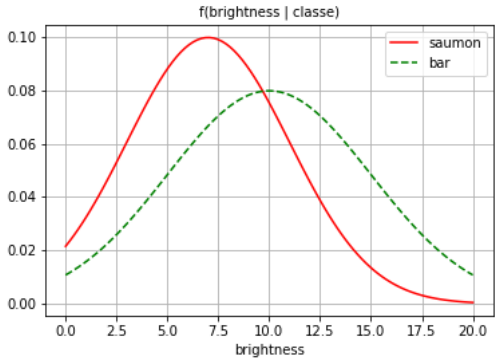
1. Interpréter la première partie du rapport. Le modèle est-il pertinent ?
2. Quelles variables ont un impact sur la variable `CO` ? Justifiez votre réponse.

Exercice 3 : Classifieur de Bayes

Dans cet exercice, on reconsidère le corpus vu en cours et en travaux dirigés : on cherche à distinguer entre les saumons et les bars en se basant sur deux descripteurs : la longueur (`length`) et la brillance (`brightness`).

On a donc estimé les paramètres des lois normales suivies par chacun des descripteurs et on a dessiné les courbes des densités de probabilités :





1. Si on ne considère que le descripteur `brightness`, comment serait classé un poisson dont la valeur du descripteur est 7.5 ? Quelle est la frontière de décision ?

2. En prenant en compte les deux descripteurs, quelle serait la classe d'un poisson dont la valeur des deux descripteurs est 7.5 ?
3. En prenant en compte les deux descripteurs, comment serait classé le même poisson si on sait que :
   (a) $p = 30\%$ des poissons sont des bars.
   (b) $p = 80\%$ des poissons sont des bars.

Exercice 4 : Réduction de dimension

Soit l'extrait suivant d'un `jupyter notebook` :

Entrée [1]:
```
import numpy as np
R = np.array([[6, 8, 6, 14.5, 14, 11, 5.5, 13, 9],
              [6, 8, 7, 14.5, 14, 10, 7, 12.5, 9.5],
              [5, 8, 11, 15.5, 12, 5.5, 14, 8.5, 12.5],
              [5.5, 8, 9.5, 15, 12, 7, 11.5, 9.5, 12],
              [8, 9, 11, 8, 10, 13, 10, 12, 18]])
```

Entrée [2]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
Z = scaler.fit_transform(R)
```

Entrée [3]:
```
p, n = R.shape[0], R.shape[1]
```

Entrée [4]:
```
S = (n-1) * np.cov(Z)
```

Entrée [5]:
```
from numpy.linalg import eig
val, vec = eig(S)
```

Entrée [6]:
```
sorted_index = np.argsort(val)[::-1]
sorted_val = val[sorted_index]
sorted_vec = vec[:,sorted_index]
```

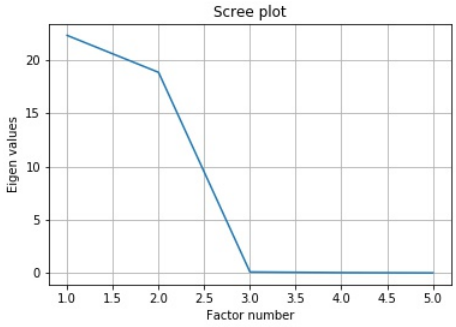
Entrée [7]:
```
sorted_val
```

Out[7]:
```
array([ 2.23427588e+01,  1.88646805e+01,  8.37212500e-02,  2.41371509e-02,
       -1.30373755e-16])
```

Entrée [8]:
```
sorted_vec
```

Out[8]:
```
array([[ 0.39569376,  0.47062549,  0.26564899,  0.59276362,  0.4472136 ],
       [ 0.35061965,  0.29780028, -0.49476593, -0.5861634 ,  0.4472136 ],
       [ 0.0389695 , -0.6890927 , -0.42753763,  0.37529216,  0.4472136 ],
       [ 0.06048053, -0.36511317,  0.70664676, -0.40457983,  0.4472136 ],
       [-0.84576344,  0.28578011, -0.04999219,  0.02268744,  0.4472136 ]])
```

1. Expliquer ce que fait la cellule 2.
2. A quoi correspondent les contenus des variables `val` et `vec` ? Quel est le rôle de la cellule 6 ?
3. On affiche l'éboulie des valeurs propres et on obtient la figure suivante :

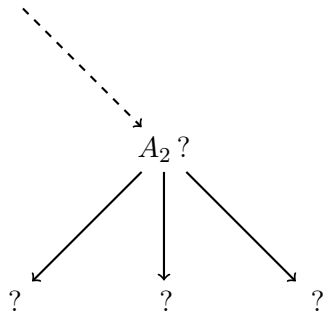


Combien d'axes faut-il garder ? Justifiez votre réponse.

4. A quoi correspondent les valeurs propres ?
5. Expliquez (par des phrases, ou par des formules) les étapes suivantes pour compléter l'ACP.
6. En général, quelles sont les limites de cette approche ? Quelle autre méthode peut-on utiliser pour réduire le nombre de descipteurs ?

Exercice 5 : Arbres de décision

Dans un processus de création d'un arbre de décision, on arrive à la configuration illustrée par la figure suivante :



A_1	A_2	$Classe$
2	3	C_1
1	2	C_1
2	2	C_2
1	1	C_2

Le noeud marqué "A_2 ?" indique que le partitionnement du corpus se fait dans ce noeud en fonction des valeurs prises par le descipteur A_2. La tableau à droite de la figure indique le contenu du corpus une fois sur le noeud "A_2 ?".

1. Quels sont les corpus correspondant aux points d'interrogation issus du noeud "A_2 ?"
2. Expliquez par des calculs détaillés le choix du descripteur A_2 comme descipteur de partitionnement à cette étape de la construction de l'arbre.

FIN.