Examen Introduction aux Sciences des Données - Session 1

Durée 2h - Barême indicatif (note sur 20)
Matériel autorisé : une feuille A4 recto-verso de synthèse personnelle du cours, manuscrite (pas de photocopie) + calculatrice non-programmable (téléphone interdit même pour sa calculatrice)
Tous les programmes donnés et demandés sont en Python3

Ce sujet d'examen est constitué de quatre exercices indépendants, énoncés sur 3 pages.

1 Régression linéaire (6 pts)

Un organisme bancaire a besoin de prévoir, en fonction du montant des prêts accordés aux professionnels, quelles sommes il doit lui même emprunter sur les marchés financiers. Pour cela, il réalise une étude sur les 12 trimestres écoulés.
Le montant global des prêts accordés chaque trimestre est donné en millions d'euros dans le tableau ci-dessous.

Rang du trimestre	1	2	3	4	5	6	7	8	9	10	11	12
Montant y_i	40	42	44	45.	48	50	52	55	58	63	68	70

Par la suite, tous les calculs sont faits à 0.1 près.

1. On cherche à expliquer la variable Montant dans le temps. Pour cela :
   (a) La meilleure droite Δ_1 de régression $y = ax + b$ (où x est le rang du trimestre) qui permet cette explication a pour coefficient directeur $a = 2.7$. En déduire l'équation de la droite Δ_1.
   (b) Déterminer une estimation des prêts accordés au 4ème trimestre 2022 sachant que le point (1,40) est celui du premier trimestre 2019.
   (c) Calculer les résidus pour les trimestres 3, 4 et 5 et pour les trois derniers. Qu'observez-vous ?
2. Cette régression Δ_1 repose sur les 12 trimestres précédents l'étude, mais pour tenir compte de l'évolution récente, on se limite aux quatre dernières observations.
   (a) Utiliser la méthode des moindres carrés pour estimer l'équation de la droite de régression Δ_2 sur ces 4 derniers points seulement.
   (b) Donner la prévision plus réaliste pour le 4ème trimestre 2022. Calculer, en million d'euros, l'écart entre les deux prédictions pour ce trimestre.
   (c) Quelle est la MSE empirique (sur les quatre derniers points) de ce nouvel estimateur ?
3. Dans un repère orhtonormal, représenter le nuage des 12 points, tracer Δ_1 et Δ_2 en deux couleurs différentes.
4. On sait que l'estimateur des moindres carrés est consistant ; qu'est-ce que cela signifie ?

2 Clustering par k-moyenne (7 pts)

[1]

Soit l'échantillon de données S, composé de 8 exemples (points) décrits en deux dimensions : $S = \{A_1(2,10); A_2(2,5); A_3(8,4); A_4(5,8); A_5(7,5); A_6(6,4); A_7(1,2); A_8(4,9)\}$.
L'objectif est d'apprendre un partitionnement de ces données en 3 clusters, à l'aide de l'algorithme des k-moyennes en utilisant la distance euclidienne.
Les distances euclidiennes entre ces points sont données dans la matrice ci-après :

1. Inspiré de AgroParisTech 2003

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
A_1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A_2	-	0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A_3	-	-	0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{4}$	$\sqrt{53}$	$\sqrt{41}$
A_4	-	-	-	0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A_5	-	-	-	-	0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A_6	-	-	-	-	-	0	$\sqrt{29}$	$\sqrt{29}$
A_7	-	-	-	-	-	-	0	$\sqrt{58}$
A_8	-	-	-	-	-	-	-	0

On initialise l'algorithme avec les points A_1, A_4 et A_7.

1. Dérouler une itération de l'algorithme des k-moyennes pour ces données et cette initialisation, et indiquer :
   (a) Les nouveaux clusters (les points parmi les 8 ci-dessus que chaque cluster regroupe)
   (b) Les centres de gravité de ces nouveaux clusters
   (c) Une représentation graphique montrant les points de l'échantillon, les clusters et leurs centres.
2. Combien d'itérations supplémentaires sont-elles nécessaires pour converger ?
3. Indiquer le graphique de la solution finale (sans détailler les calculs éventuels)
4. Calculer l'homogénéité moyenne du regroupement final obtenu.

3 Complétion de données manquantes (4 pts)

Soit le jeu de données suivant, où les 9 données sont décrites en dimension 3, et dans lequel il manque certaines valeurs ; les réponses aux questions doivent être justifiées.

	a_1	a_2	a_3
x_1	0	12.5	4
x_2	0	14.5	6
x_3	1	2.3	-
x_4	1	4.1	9
x_5	0	6.2	9
x_6	0	3.8	6
x_7	1	-	6
x_8	1	-	4
x_9	0	4.9	-

1. Compléter les données de la colonne a_1 en utilisant une méthode stationnaire.
2. Compléter les données de la colonne a_2 par la médiane.
3. Compléter les données de la colonne a_3 en utilisant les 1−plus proches voisins de la colonne a_2.

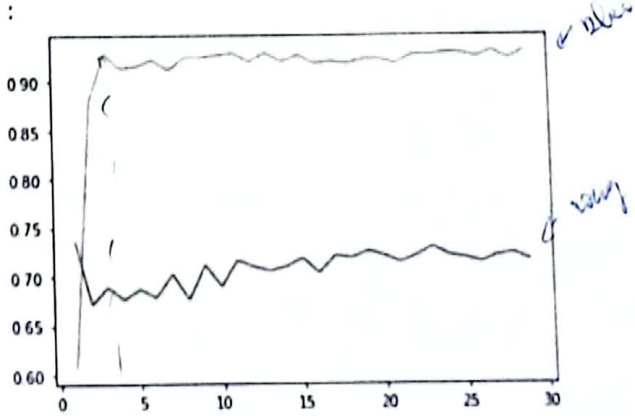
4 Compréhension de code (3 pts)

Soit le code ci-après (on suppose les import déjà réalisés) :

```
digitsData=load_wine() # jeu de données digits
X=digitsData.data # les exemples, un array numpy, chaque élément est aussi un array
y=digitsData.target # les classes

r = range(1, 30)
SV = np.zeros(len(r))
for ik in r:
    kppv = nn.KNeighborsClassifier(ik)
    s = 0
    for iexp in range(10):
        Xtrain, Xtest, ytrain, ytest = train_test_split(X,y,test_size=0.25, random_state=iexp)
        kppv.fit(Xtrain, ytrain)
        s = s + kppv.score(Xtest, ytest)
    SV[ik-1] = s/10

ST = np.zeros(len(r))
for ik in r:
    clf = DecisionTreeClassifier(max_depth = ik)
    s = 0
    for iexp in range(10):
        Xtrain, Xtest, ytrain, ytest = train_test_split(X,y,test_size=0.25, random_state=iexp)
        clf.fit(Xtrain, ytrain)
        s = s + clf.score(Xtest, ytest)
    ST[ik-1] = s/10

plt.plot(r, SV) # bleu (en dessous)
plt.plot(r, ST) # orange (au dessus)
plt.show()
```

1. Expliquer en quelques phrases ce que réalise ce programme
2. Expliquer l'objectif et l'intérêt des boucles internes `for iexp in range(10)`

L'exécution de code fournit le graphique suivant :



3. Commenter ce graphique
4. Quel type de classifieur semble le plus performant, et avec quel hyper-paramétrage (justifier) ?
5. **BONUS** Proposer un code alternatif beaucoup plus court en exploitant les fonctionnalités de sklearn.