Veuillez prendre connaissance des consignes ci-dessous :

- Durée de l'examen: **2h30**
- Une feuille de note recto verso 8.5" X 11" ou A4 et une calculatrice programmable sont permises.
- Le barème est donné à titre indicatif et peut être sujet à modification.
- Vous devez encadrer les résultats et donner les valeurs avec deux chiffres après la virgule en arrondissant au plus proche. Par exemple, $1,947$ doit être noté $\boxed{1{,}95}$.
- Vous pouvez répondre aux questions en français ou en anglais.

Please read carefully the following instructions:

- Exam duration: **2h30**
- A double-sided sheet of size 8.5" X 11" or A4 and a programmable calculator are allowed.
- The marking of the questions is given for references and may be subject to change.
- You must draw a box around the results and give values with two decimal places by rounding to the nearest value. For instance, 1.947 must be given as $\boxed{1.95}$.
- You are allowed to answer in French or in English.

1 Science de données (1 point)

En observant l'augmentation significative de la taille moyenne des joueurs de volleyball au fil des années, on pourrait penser que cela reflète une tendance générale de la population. Cependant, pourquoi ne peut-on pas tirer cette conclusion à partir de ces données ?

L'augmentation de la taille moyenne des joueurs de volleyball au fil des années reflète une tendance spécifique au sport. Les joueurs plus grands ont un avantage compétitif au volleyball, notamment au niveau élite, car la taille peut améliorer des compétences comme le bloc et l'attaque. En conséquence, les équipes et les entraîneurs de volleyball sélectionnent de plus en plus de joueurs grands, ce qui crée une tendance à la hausse de la taille dans ce sport.

Cependant, cette tendance ne peut pas être utilisée pour tirer des conclusions sur la taille moyenne de la population générale pour plusieurs raisons :

Biais de sélection : Les joueurs de volleyball représentent un sous-ensemble spécifique et auto-sélectionné de la population, où la taille est un trait désirable. Les données sont biaisées en faveur des individus plus grands, et ne représentent pas l'ensemble de la population.

Spécialisation dans le sport : La taille est un facteur crucial dans certains sports, mais pas dans la population générale. Les tendances observées dans la taille des joueurs de volleyball sont donc davantage liées à la spécialisation dans ce sport plutôt qu'à une tendance biologique générale.

Échantillon non représentatif : La taille moyenne des joueurs de volleyball ne constitue pas un échantillon aléatoire de la population mondiale. Les changements au sein de ce sous-groupe ne peuvent pas être extrapolés à des changements démographiques plus larges.

En conclusion, bien que la taille moyenne des joueurs de volleyball ait augmenté, il s'agit d'une conséquence des pressions de sélection spécifiques au sport, et non d'un indicateur fiable de la tendance de la taille dans la population générale.

2 Préparation de données (2 points)

Lors de l'analyse de la performance des portefeuilles d'investissement, comment pouvez-vous gérer les données manquantes dues aux jours de fermeture des marchés ? Décrivez deux méthodes spécifiques d'imputation pour assurer la fiabilité des résultats.

Imputation par interpolation :

J'utilise l'interpolation linéaire pour estimer les valeurs manquantes en me basant sur les prix des jours précédents et suivants, ce qui permet de conserver une continuité dans les données.

Utilisation de la valeur précédente :

J'applique une méthode simple en utilisant le prix de clôture du jour précédent pour remplir les jours de fermeture, en supposant que le prix reste stable sur une courte période.

Suppression des jours fermés Dans certains cas, j'exclus simplement les jours de fermeture du calcul des performances, surtout s'ils ne représentent qu'une petite fraction des données.

Analyse des tendances J'examine les séries temporelles des actions et j'emploie des modèles de régression pour prédire les valeurs manquantes en fonction des tendances observées.

3 Réduction de données (1 point)

Vous disposez d'un ensemble de m objets répartis en K groupes, où le i-ème groupe a une taille de m_i. Si l'objectif est d'obtenir un échantillon de taille $n < m$, quelle est la différence entre les deux schémas d'échantillonnage suivants ? (Supposez un échantillonnage avec remplacement).

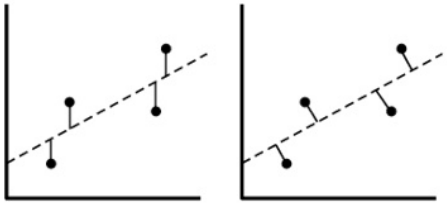
a) Nous sélectionnons aléatoirement $n \times m_i/m$ éléments de chaque groupe.

b) Nous sélectionnons aléatoirement n éléments de l'ensemble de données, sans tenir compte du groupe auquel appartient un objet.

Le premier schéma garantit que chaque groupe est représenté dans l'échantillon de manière proportionnelle à sa taille dans l'ensemble de données original, tandis que pour le second schéma certains groupes pourraient être sur- ou sous-représentés par rapport à leur proportion dans l'ensemble original.

4 Régression (2 points)

(0.5 point) (a) Supposons que l'axe horizontal représente une variable indépendante et l'axe vertical une variable dépendante. Quelle figure représente le mieux l'erreur résiduelle dans l'ajustement de la droite de régression linéaire? Celle de gauche ou celle de droite?



gauche

(0.5 point) (b) On peut calculer les coefficients d'une régression linéaire à l'aide d'une méthode analytique. Parmi les affirmations suivantes, laquelle/lesquelles sont vraies à propos de cette méthode ?

(i) Nous n'avons pas besoin de choisir un taux d'apprentissage.

(ii) Elle devient lente lorsque le nombre d'attributs est très grand.

(iii) Il n'est pas nécessaire de l'itérer jusqu'à la convergence.

(i),(ii) et (iii)

(1 point) (c) Pouvons-nous résoudre les problèmes de classification multi-classe en utilisant la régression logistique? Si oui, comment?

Oui, pour traiter la classification multiclasse en utilisant la régression logistique, la méthode la plus connue est appelée l'approche un-contre-tous (one-vs-all). Dans cette approche, un certain nombre de modèles sont entraînés, égal au nombre de classes.

5 Classification (2 points)

Considérez un ensemble de données étiquetées contenant 100 instances de données, qui est partitionné aléatoirement en deux ensembles A et B, chacun contenant 50 instances. **Nous utilisons A comme ensemble d'entraînement** pour apprendre deux arbres de décision, T_{10} avec 10 nœuds feuilles et T_{100} avec 100 nœuds feuilles. Les valeurs d'*accuracy* des deux arbres de décision sur les ensembles de données A et B sont présentées dans le tableau suivant:

	accuracy	
jeux de données	T_{10}	T_{100}
A	0.86	0.97
B	0.84	0.77

(1 point) (a) En se basant sur les valeurs présentées dans le tableau, quel arbre s'attendriez-vous à avoir une meilleure performance sur des données non étiquetées ?

L'arbre T10 est censé montrer une meilleure performance de généralisation sur des instances non vues que T100. Nous pouvons voir que la précision d'entraînement de T100 sur l'ensemble de données A est très élevée, mais sa précision de test sur l'ensemble de données B, qui n'a pas été utilisé pour l'entraînement, est faible. Cet écart entre les précisions d'entraînement et de test implique que T100 souffre du surapprentissage. Ainsi, même si l'accuracy d'entraînement de T100 est très élevée sur l'ensemble de données A, elle n'est pas représentative de la performance de généralisation sur des instances non vues dans l'ensemble de données B. D'autre part, l'arbre T10 a une accuracy d'entraînement modérément basse sur l'ensemble de données A, mais la précision de test de T10 sur l'ensemble de données B n'est pas très différente. Cela implique que T10 ne souffre pas du surapprentissage et que la performance d'entraînement est effectivement indicative de la performance de généralisation.

(1 point) (b) Maintenant, vous avez testé T_{10} et T_{100} sur l'ensemble de données entier $(A + B)$ et vous avez trouvé que l'accuracy de classification de T_{10} sur l'ensemble de données $(A+B)$ est de 0.85, tandis que l'accuracy de classification de T_{100} sur l'ensemble de données $(A + B)$ est de 0.87. En vous basant sur cette nouvelle information et vos observations du tableau, quel arbre choisiriez-vous finalement pour la classification ?

Nous choisirions toujours T10 plutôt que T100 pour la classification. La haute accuracy de T100 sur l'ensemble de données (A + B) peut être attribuée à la haute précision d'entraînement de T100 sur l'ensemble de données A, qui est un artefact du surapprentissage. Notez que la performance d'un classificateur sur l'ensemble de données combiné (A + B) ne peut pas être considérée comme une estimation de la performance de généralisation, car elle contient des instances utilisées pour l'entraînement (provenant de l'ensemble de données A).

6 *Clustering* (2 points)

Considérons l'ensemble de points de données unidimensionnels suivant : $\{0.1, 0.25, 0.45, 0.55, 0.8, 0.9\}$. Tous les points sont situés dans l'intervalle $[0, 1]$.

(1 point) (a) Supposons que nous appliquons l'algorithme de k-moyennes pour obtenir trois clusters, A, B et C. Si les centroïdes initiaux sont $\mu_A = 0$, $\mu_B = 0.4$ et $\mu_C = 1$, montrez les affectations aux clusters et les emplacements des centroïdes mis à jour après les trois premières itérations en remplissant le tableau suivant.

	Affectations aux clusters (A, B ou C)						Position des centroïdes		
Itération	0.10	0.25	0.45	0.55	0.80	0.90	μ_A	μ_B	μ_C
0	-	-	-	-	-	-	0.00	0.40	1.00
1									
2									
3									

(1 point) (b) Pour l'ensemble de données donné dans la partie (a), est-il possible d'obtenir un cluster vide après la première affectation de clusters ? Si c'est possible, quelles sont les positions des centroïdes initiaux pour produire le cluster vide ? Si ce n'est pas possible, expliquez pourquoi.

Oui. Si deux des centroïdes sont initialisés aléatoirement à gauche des points les plus à gauche ou à droite des points les plus à droite. Par exemple, si les centroïdes initiaux sont (0, 0.05, 0.8), alors le premier cluster est vide.