NOM : Prénom :

Université de Bretagne Sud
Master 1 Informatique

Contrôle Continu – Traitement Numérique des Données – Session 1

6 décembre 2021

Durée de l'épreuve : 1h30, 6 pages
Note : Aucun document n'est autorisé. Calculatrices, assistants personnels, téléphones portables sont interdits. Il sera tenu compte de la pertinence de l'argumentation écrite dans le cadre d'une interprétation limitée et de la qualité de la rédaction. Vous composerez directement sur le sujet, dans les espaces réservés. Le barème est indicatif.

Les 3 premiers exercices sont des questions de cours, ou des exercices appliqués du cours. Le 4ème exercice est un problème de réflexion plus avancé relatif à ce cours.

1. **Régression linéaire (5 points)**

Soit un ensemble de m données d'apprentissage : $\{(x^{(i)}, y^{(i)}), i = 1…m\}$. On souhaite appliquer une méthode de régression linéaire à cet ensemble d'apprentissage pour prédire la sortie estimée y correspondant à une nouvelle entrée x.

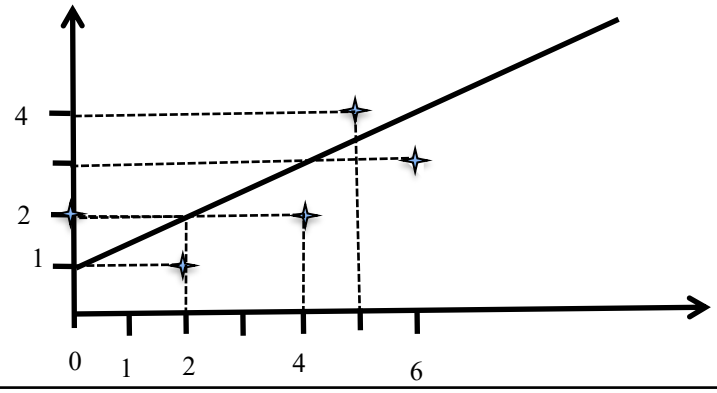
1.1. Comment peut-on définir la fonction de prédiction (encore appelée fonction hypothèse h) ?

1.2. Quels sont les paramètres de cette fonction de prédiction ?

1.3. Comment définissez-vous la fonction coût qui caractérise l'erreur de la droite de régression par rapport aux données d'apprentissage ? Donnez la formule pour un nombre m de points de l'ensemble d'apprentissage.

1.4. Pour l'ensemble d'apprentissage défini par les 5 points $(x^{(i)}, y^{(i)})$ (i=1..5) qui prennent les valeurs {(0,2), (2,1), (4, 2), (5, 4), (6, 3)} (voir figure à la page suivante), calculez la valeur de la fonction coût définie à la question 1.3 en considérant la droite de régression tracée sur la même figure.

NOM : Prénom :



1.5. Donnez le principe de l'algorithme de régression linéaire par descente de gradient (pseudo-algorithme)

2. Régression logistique (5 points)

2.1. Dans la régression logistique, la sortie est-elle continue ou discrète ?

2.2. Quelle forme prend la fonction de prédiction ? (Cochez la bonne réponse).

- ☐ $h_\theta(x) = \theta^T x$
- ☐ $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$
- ☐ $h_\theta(x) = \frac{1}{1+\log(\theta^T x)}$

2.3. Quelle propriété doit respecter la fonction coût pour que l'algorithme de minimisation de ce coût trouve une solution et une seule.

NOM : Prénom :

2.4. Tracez la fonction de prédiction : $h_\theta(x)$ en fonction de $\theta^T x$. Comment utilisez-vous cette fonction pour obtenir la sortie du *classifieur* ? Expliquez votre raisonnement sur la figure.

2.5. Soit la matrice de confusion suivante, qui permet de classer les patients reçus à l'hôpital (10 000 au total) en patients testés positifs à la maladie de Lyme (classe C+) et en patients testés négatifs à cette maladie (classe C-).

Prédiction\Vérité terrain	C+	C-
^C+	142	475
^C-	358	9025

^C+ : estimation de la classe C+
^C- : estimation de la classe C–
C+ : classe C+ réelle (vérité terrain)
C- : classe C- réelle

Expliquez ce que représente chacun des éléments de ce tableau.

2.6. Quels sont les taux de reconnaissance (Accuracy), de Précision (ou exactitude), de Rappel (ou de sensibilité) ?

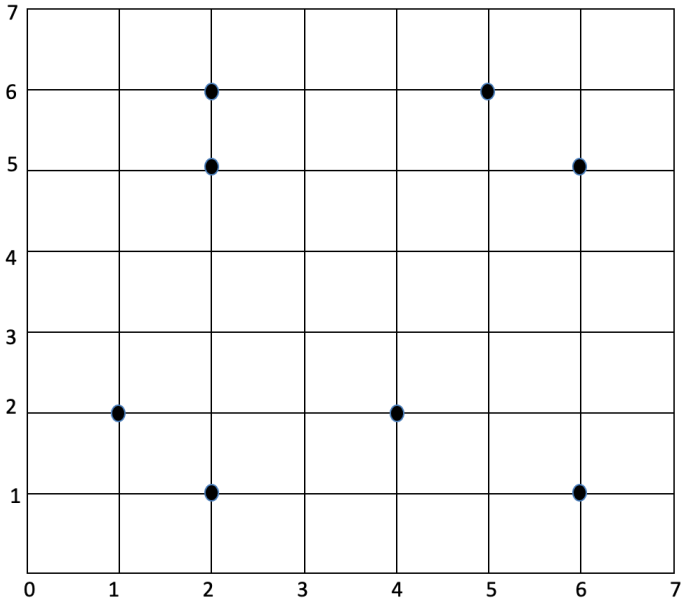
2.7. Donnez une interprétation des résultats obtenus à la question précédente.

NOM : Prénom :

3. Algorithme des k plus proches voisins (k-NN) (4 points)

On cherche à prédire la couleur d'un fruit en fonction de sa largeur (*L*) et de sa hauteur (*H*), grâce à un algorithme k-NN. On dispose des données d'apprentissage suivantes, que l'on place dans un schéma quadrillé (largeur en abscisse, hauteur en ordonnée ; voir tableau et figure suivants).

largeur	hauteur	couleur
2	6	Red (R)
5	6	Yellow (Y)
2	5	yellow
6	5	yellow
1	2	red
4	2	Green (G)
2	1	green
6	1	green



L'objectif est d'étudier l'influence des voisins sur la propriété de couleur d'un fruit. Soit *F* le nouveau fruit de largeur *L* = 1 et de hauteur *H* =4 (exemple de l'ensemble de test).

1. Indiquez pour chaque point sa couleur (inscrivez la lettre correspondante dans le schéma à droite).
2. Quelle est la couleur du nouveau point *F* si on considère 1 voisin, puis 3 voisins ? Expliquez en vous aidant du graphique.

3. Plutôt que le vote majoritaire, on considère la couleur des voisins pondérée par la distance euclidienne. Chaque voisin a ainsi un poids *w* inversement proportionnel au carré de sa distance au point test : $w = 1/d^2$. Si l'on prend 3 voisins, quelle est la couleur de *F* ? Donnez le calcul et expliquez.

NOM : Prénom :

4. Problème de classification (question bonus – 6 points)

Dans une banque, on considère des données de signatures de clients qui sont représentées par des traces de points ($x(t)$,$y(t)$) évoluant au cours du temps. Ces traces sont échantillonnées de telle façon que chacune d'elle est caractérisée par une série temporelle $S = [P_1, P_2, \ldots, P_n]$, avec $P_j = (x_j, y_j)$ le point de la trace à l'instant j. On souhaite faire de la classification sur ces signatures à partir d'un algorithme k-NN. Vous pourrez écrire l'algo de la question 4.3 en considérant que la fonction de la question 4.2 existe.

4.1. Représentez les données de ce problème sous la forme d'une matrice Σ et d'un vecteur z. Vous écrirez ci-dessous la matrice Σ composée de m lignes, chaque ligne étant une série temporelle S^i correspondant à une signature et s'exprimant à partir de la séquence de points $\{P^i_j\}$, j variant de 1 à n. Vous ferez l'hypothèse que les séries temporelles sont de longueur identique n. z est le vecteur des classes, chaque classe étant représentée par le n° de la signature.

4.2. Pour 2 signatures S^i et S^j de 2 clients différents, donnez l'algorithmique (fonction) qui permet de calculer la distance $dist(S^i, S^j)$ en considérant que cette distance est la somme des distances euclidiennes au carré des points constituant chacune des séries temporelles.

NOM : Prénom :

4.3. Écrivez en pseudo-code l'algorithme K-NN qui permet, pour une nouvelle signature donnée *S* de déterminer les *k* plus proches voisins (signatures les plus similaires) et de déterminer la classe de cette signature. À travers ce pseudo-code, vous expliquerez le principe de l'algorithme (*k* est un paramètre de l'algorithme).