République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre universitaire Abdelhafid Boussouf - Mila
Institut des Sciences et de la Technologie – Département des Mathématiques et Informatique

Concours d'accès à la formation de 3ème cycle (Doctorat LMD)
Filière : Informatique - Spécialité : Intelligence artificielle
Épreuve de spécialité : Apprentissage automatique (Sujet 3)
Durée : 2H00

Date : 28-01-2023

Exercice 01 (06 pts) :

Le tableau suivant contient des données sur des individus d'une population décrits selon deux attributs : attribut 1 et attribut 2. La classe d'un individu peut être : C1, C2, ... ou C6.

N°	Attribut 1	Attribut 2	Classe
1	1	2	C1
2	2	6	C1
3	2	5	C2
4	2	1	C3
5	4	2	C5
6	5	6	C4
7	6	5	C3
8	6	1	C6

1) On veut classer un nouvel individu U ayant comme attributs (1, 4) en utilisant la méthode KNN. Quelle sera la classe de U si on choisit k=3 (Utiliser la distance Manhattan). Justifier votre réponse.
2) Refaire la question 1) en utilisant la masure de similarité cosinus.

NB. La similarité cosinus calcule la similarité en mesurant le cosinus de l'angle entre deux vecteurs. Le cosinus entre deux vecteurs A et B est calculé comme suit :

$$\text{Cos}(\vec{a},\vec{b}) = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|.\|\vec{b}\|} = \frac{\sum_i^n a_{i*}b_i}{\sqrt{\sum_i^n a_i^2} * \sqrt{\sum_i^n b_i^2}}$$

Exercice 02 (08 pts) :

Nous souhaitons réaliser un classifieur Bayesien permettant de classifier les emails en « Spam » ou « Ham (not spam) ». Pour ce faire, chaque mot w_i d'un e-mail, quel que soit sa position dans le texte de l'e-mail, est supposé avoir une probabilité $P(W= w_i|Y)$, où W prend des mots dans un dictionnaire prédéterminé (la ponctuation est ignorée). Y prend une valeur binaire (Spam ou Ham).

I. Supposons que nous avons trois emails comme ensemble d'apprentissage.

(Spam) dear sir, if you could answer my questions I would be most grateful.
(Ham) see you at 12
(Ham) well, prepare it for tomorrow.

A partir de cet ensemble d'entrainement, calculer les probabilités Bayésiennes suivantes :

1) $P(W=sir \mid Y = spam)$.
2) $P(W=see \mid Y = ham)$.
3) $P(W=today \mid Y = ham)$.
4) $P(Y = ham)$.

II. Le tableau suivant montre les probabilités estimées d'un ensemble de mots spams entrainés sur un large corpus d'emails.

W	*good*	*to*	*fine*	*luck*	*pay*
P(W|Y=spam)	*1/6*	*1/8*	*1/4*	*1/8*	*1/4*
P(W|Y=ham)	*1/8*	*1/3*	*1/4*	*1/12*	*1/12*

On vous donne un nouvel email à classer, avec seulement deux mots : « *Good luck* »

1. Calculer la décision estimée pour cet email, sachant que : $P(Y = spam) = 1/5$.
2. Quelle est l'intervalle de probabilités de $P(Y = spam)$ pour lequel le classifieur Bayesien classe ce nouvel email comme spam ?

Exercice 03 (06 pts) :

Soit $\mathbf{D} = \{x_1, x_2, \ldots, x_N\}$ un ensemble d'apprentissage de N données unidimensionnelle. On suppose que les données de l'ensemble $\boldsymbol{D}$ sont constituées de K groupes: $\boldsymbol{G_1}, \ldots, \boldsymbol{G_K}$. Chaque groupe $\boldsymbol{G_i}$ suit une loi normale $\mathcal{N}(\boldsymbol{\mu_i}, \sigma_i^2)$. On considère l'algorithme Espérance-Maximisation (EM) appliqué à l'ensemble de données $\boldsymbol{D}$, soit :

$$p(x|\boldsymbol{G_i}, \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

La modélisation du groupement (clustering) par EM est donnée par : $\Phi = \{\pi_i, \mu_i, \sigma_i\}_{i=1}^{K}$. En guise de rappel, la formule de l'espérance de vraisemblance de l'algorithme EM est la suivante:

$$Q(\Phi|\Phi^l) = \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p(x^t|\boldsymbol{G_i}, \Phi^l)$$

où $\pi_i = p(G_i|\Phi)$ est la probabilité *a priori* du groupe G_i et $h_i^t = p(G_i|\mathbf{x}, \Phi)$ est l'appartenance probabiliste de la donnée x au groupe $\boldsymbol{G_i}$.

1) Donner le développement complet permettant de calculer les estimations π_i des probabilités *a priori* des groupes.
2) Donner le développement complet permettant de calculer les estimations $\mathbf{m}_i$ des moyennes μ_i.
3) Donner le développement complet permettant de calculer les estimations s_i^2 des σ_i^2.

Bon courage.

Sujet : 02

Exercice 01 (05 pts)

Ecrire une procédure ***Elmts_communs (X, Y, m, n)*** qui admet en entrée deux tableaux ***X*** et ***Y*** triés dans ordre croissant, de tailles respectives ***m*** et ***n*** puis affiche leurs éléments communs, en utilisant une seu boucle.

Exemple : Soient 2 tableaux *X* et *Y* triés dans un ordre croissant, de tailles respectives 10 et 8:

$X = \{11, \mathbf{15}, 23, \mathbf{25}, 26, \mathbf{28}, 38, 40, 44, 48\}$,

$Y = \{\mathbf{15}, 17, 21, \mathbf{25}, \mathbf{28}, 46, 51, 57\}$

Les éléments affichés après l'appel de la procédure ***Elmts_communs (X, Y, 10, 8)***: 15, 25, 28.

Exercice 02 (07 pts)

1. Quel est le nombre maximal de nœuds parcourus lors de la recherche d'un élément:

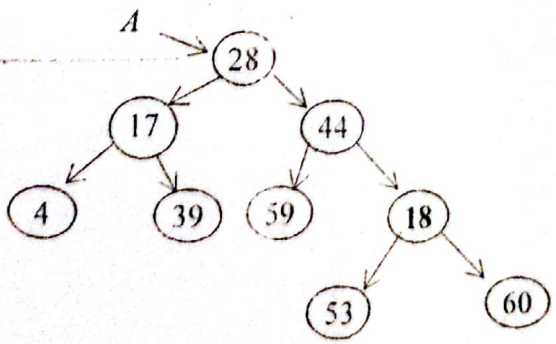
   - Dans un arbre binaire simple.

   - Dans un arbre binaire de recherche.

2. Ecrire une fonction ***nb_nœudsparcourus*** qui retourne le nombre de nœuds parcourus lors de la recherche d'un entier *x* dans un arbre binaire d'entiers *A*.

3. Ecrire une fonction ***nb_nœudsparcourusR*** qui retourne le nombre de nœuds parcourus lors de la recherche d'un entier *x* dans un arbre binaire de recherche *A*.

Exemple : le nombre de nœuds parcourus lors de la recherche de la valeur **18** dans l'arbre A ci-dessous, si on commence toujours par le parcours du sous-arbre gauche de chaque nœud, est **7**.



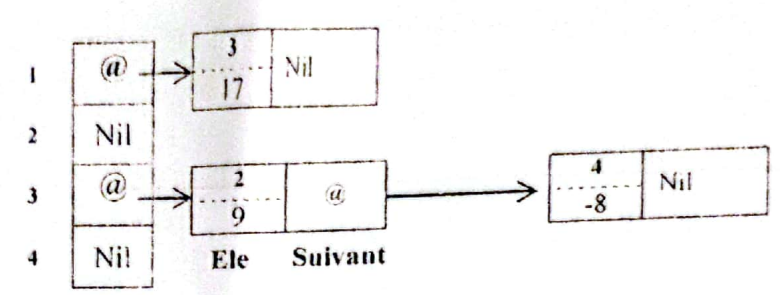
Exercice 03 (08 pts)

Une matrice creuse est une matrice dont la majorité des éléments sont nuls. On peut représenter une matrice creuse, par un tableau de N listes (N est le nombre de lignes de la matrice), chaque élement du tableau représente une ligne de la matrice qui est représentée à son tour par une liste linéaire chainée ordonnée selon l'indice de la colonne des éléments non nuls. Chaque élément de la liste contient l'indice (j) de la colonne et la valeur (v) de l'élément non nul de la matrice.

Exemple : ci-dessous un exemple sur la représentation d'une matrice **A** par un tableau de listes ***Tlistes***.

	1	2	3	4	5
1	0	0	17	0	0
2	0	0	0	0	0
3	0	9	0	-8	0
4	0	0	0	0	0

A (4*5)



Tlistes

1. Ecrire une procédure ***afficher (Tlistes, N, M)*** admettant en entrée un tableau ***Tlistes*** représentant une matrice creuse de N*M éléments qui affiche ligne par ligne tous les éléments (nuls et non nuls) de cette matrice.
2. Ecrire une fonction ***pourcentage (Tlistes, N, M)*** admettant en entrée un tableau ***Tlistes*** représentant une matrice creuse de N*M éléments et qui retourne le pourcentage des éléments nuls dans la matrice.
3. Écrire une fonction ***diagonale(Tlistes, N)*** admettant en entrée un tableau ***Tlistes*** représentant une matrice carrée de N*N éléments et qui permet de vérifier si cette matrice est diagonale ou non. Une matrice carrée est dite diagonale si tous les éléments hors la diagonale sont nuls.