Université Abderrahmane Mira de Bejaia
Faculté des Sciences Exactes
Département de recherche opérationnelle
Année universitaire : 2022/2023

EMD en Fouille de données et apprentissage automatique (M1 SDAD) (Durée : 1h.30mn)

Toute réponse incomplète ou non argumentée ne sera pas notée. Echange d'objets interdit. Documents, Mobile et moyens électroniques sont interdits. (Le sujet est composé de deux feuilles)

Questions (3 Pts)

Q1- Quel est l'intérêt de décomposer l'échantillon de travail en : « training data set : échantillon d'apprentissage » et en « validation data set : échantillon de validation » ?
Q2- Ali veut écrire un programme qui utilise la fréquence des mots « science », « public », « accès », « université », « gouvernement », « financer », « éducation », « budget », « justice » et « loi » pour déterminer si un article traite ou non de politique. Il a commencé par annoter un millier d'articles selon leur sujet. Quel genre de problème d'apprentissage automatique doit-il résoudre ?
Q3- Bilal dispose de 10000 articles de journaux qu'il souhaite classer par leur thématique. Doit-il utiliser un algorithme supervisé ou non supervisé ?
Q4- Ahmed gère un outil qui permet d'organiser les liens HTML qui ont été sauvegardés. Il souhaite suggérer des catégories auxquelles affecter un nouveau lien, en fonction des catégories déjà définies par l'ensemble des utilisateurs du service. Quel type d'algorithme d'apprentissage doit-il utiliser ?
Q5- Amine veut examiner ses spams pour déterminer s'il existe des sous-types de spams. Quel type d'algorithme d'apprentissage doit-il utiliser?
Q6- Que se passe-t-il si votre modèle fonctionne bien sur les données d'apprentissage, mais se généralise mal aux nouvelles instances ?

Exercice 1 (8 Pts)

Soient les informations des symptômes et du diagnostic des patients suivants :

Douleurs	Fatigue	Mal de tête	Fièvre	Grippe ?
Oui	Non	Doux	Oui	Non
Oui	Oui	Néant	Non	Oui
Oui	Non	Fort	Oui	Oui
Non	Oui	Doux	Oui	Oui
Non	Non	Néant	Non	Non
Non	Oui	Fort	Oui	Oui
Non	Oui	Fort	Non	Non
Oui	Oui	Doux	Oui	Oui

En utilisant la classification de Bayes naïve, prédire l'état d'un patient ayant les symptômes suivants :

Douleurs	Fatigue	Mal de tête	Fièvre	Grippe ?
Non	Non	Doux	Oui	?

Exercice 2 (7 Pts)

Soit l'ensemble des entiers suivants E = {3,7,10,12,15,17}. Nous voulons grouper ces entiers en deux groupes G1, G2 en utilisant l'algorithme Kmeans. La distance entre deux nombres x, y est calculée comme suit : d(x,y) = |x-y|.

Déroulez l'algorithme Kmeans avec comme centres initiaux : c1=10 et c2=12 (donner toutes les étapes de calcul).

Exercice 3 (2 Pts)

Donner les grandes lignes d'un notebook jupyter d'un problème résolu avec le machine Learning.