UVF3B403 **Fouille de données**

Examen 23 février 2016 – Durée : 1h30 – Aucun document autorisé

Nom Prénom de l'élève : .....

`[x]` aligné à droite à la fin d'une question indique le barème en point(s).

Pour chacune des questions 1 à 7 à choix multiples (partie I.), plusieurs réponses proposées peuvent être exactes. Vous devez cocher la ou les réponse(s) exacte(s) sans justification. Une bonne réponse complète rapporte **x point**. Une réponse mauvaise, ou partiellement incomplète, enlève $\frac{1}{2}$ **point**. L'absence de réponse ne rapporte aucun point et n'enlève aucun point.

Pour les questions suivantes (parties II. et III.) soyez précis et concis.

I. Questions à Choix Multiples

1. La technique des arbres de décision est une méthode d'apprentissage `[1/2]`
   - ◯ supervisée.
   - ◯ non supervisée.
2. La technique des arbres de décision est une méthode d'apprentissage permettant de prévoir les valeurs prises par `[1/2]`
   - ◯ une variable numérique.
   - ◯ une variable binaire.
   - ◯ une variable catégorielle.
3. Que signifie tfidf ? `[1/2]`
   - ◯ Text frequency and intelligent data framework
   - ◯ Term frequency and inverse document frequency
   - ◯ Translation from indirect to direct features
   - ◯ Text functions and inverse data features
4. À quoi servent les noyaux des SVM ? `[1/2]`
   - ◯ À simplifier les données pour n'en garder que les plus pertinentes
   - ◯ À interpoler les données manquantes
   - ◯ À transformer les données pour que la séparation par hyperplan soit possible
   - ◯ À paralléliser les calculs en attaquant directement les noyaux des processeurs

5. Le taux d'erreur d'un modèle prédictif est toujours une bonne indication des performances du modèle [1/2]

   ◯ oui.

   ◯ non.

6. Vous disposez de p variables qualitatives, chacune ayant 3 modalités, et de k variables continues. Vous devez appliquer les règles d'association sur des données ainsi décrites. Quelle(s) opérations devez-vous réaliser ? [1/2]

   ◯ discrétisation.

   ◯ ne rien faire.

   ◯ transformation disjonctive complète.

7. Une donnée aberrante [1/2]

   ◯ n'est jamais une vraie donnée.

   ◯ est parfois une vraie donnée.

   ◯ peut être une erreur.

II. Questions / Réponses

8. Quelle est la valeur de la mesure `confiance` à l'indépendance ? [1]

   .....

   .....

   .....

9. Donnez trois critères permettant de comparer deux modèles/algorithmes de fouille de données [1]

   .....

   .....

   .....

   .....

   .....

   .....

   .....

   .....

10. Un modèle $\mathcal{M}$ produit les trois règles suivantes pour estimer les valeurs d'une variable numérique Y à l'aide des variables prédictives X_1 et X_2 :

    R1 : si $X_1 \leq 2$ alors $Y = -0.02 \times X_2 + 0.5 \times X_1 - 1.44$

    R2 : If $X_1 > 3$ et $X_2 > 6.5$ alors $Y = -0.2 \times X_2 + 0.05 \times X_1 - 7.5$

    R3 (sinon) : $Y = -0.03 \times X_2 - 0.22$

    Quel(s) familles de modèle(s) (ou hypothèse(s)) sont utilisées, et comment le sont-elles, par le modèle $\mathcal{M}$ (c'est-à-dire, expliquez comment obtenir un tel modèle à partir d'un ensemble de données) ? [2]

    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....

11. Qu'est-ce qu'une échelle ordinale ? Donnez un exemple d'attribut ordinal. [1/2]

    .....
    .....
    .....
    .....
    .....
    .....
    .....
    .....

12. La comparaison de deux classifieurs via une courbe ROC montre qu'aucun ne domine strictement l'autre. Que proposez-vous pour choisir l'un des deux classifieurs ? [1]

13. On considère l'algorithme des plus proches voisins. Donner un exemple de situation où il peut être nécessaire d'utiliser un grand nombre de voisins pour aider à classer un nouvel exemple [1]

14. Le tableau[1] suivant contient 14 individus –un sous-ensemble du tableau d'origine– décrits par 4 variables nominales (age of the patient: young, pre-presbyopic, presbyopic; spectacle prescription: myope, hypermetrope; astigmatic: no, yes; tear production rate: reduced, normal) et une variable de classe (hard contact lenses hck; soft contact lenses scl; no contact lenses ncl). Quelle variable proposez-vous en premier pour construire un arbre de décision ? Pourquoi ? [2]

Id	age	prescription	astigmatic	tear	Classe
1	young	myope	true	reduce	ncl
2	young	myope	true	normal	scl
3	young	myope	false	reduce	ncl
4	young	myope	false	normal	hcl
5	young	hypermetrope	true	reduce	ncl
6	young	hypermetrope	true	normal	scl
7	young	hypermetrope	false	reduce	ncl
8	young	hypermetrope	false	normal	hcl
9	pre-presbyopic	myope	true	reduce	ncl
10	pre-presbyopic	myope	true	normal	scl
11	pre-presbyopic	myope	false	reduce	ncl
12	pre-presbyopic	myope	false	normal	hcl
13	pre-presbyopic	hypermetrope	true	reduce	ncl
14	pre-presbyopic	hypermetrope	true	normal	scl

[1]Cendrowska, J. "PRISM: An algorithm for inducing modular rules", International Journal of Man-Machine Studies, 1987, 27, 349-370

15. Un de vos ami utilise l'algorithme Apriori sur des données transactionnelles d'un supermarché. Il trouve les temps de calcul excessivement longs. La règle <milk, butter, cheese, bread, flour, sugar, salt, chocolate, apples> → vanilla> se trouve être parmi les résultats. Pouvez-vous l'aider à comprendre pourquoi les calculs sont longs ? Justifiez votre réponse. [1]

◯ non je ne peux pas l'aider

.....
.....
.....
.....
.....

◯ oui je peux l'aider

.....
.....
.....
.....
.....

16. Cet ami, encore lui, ne comprend pas l'intérêt des approches ensemblistes. Pouvez-vous le convaincre à l'aide de quelques arguments en faveur des méthodes ensemblistes ? [2]

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

III. Solution(s) pour le RAK Mining

17. Le gérant du RAK vous demande de l'aider à comprendre les habitudes de consommation. Bref, il veut faire du RAK-Mining. Quelle solution pouvez-vous lui proposer ? Décrivez brièvement les données à collecter, les éventuels traitements à appliquer aux données et les méthodes de fouille envisageables. [5]

Feuille supplémentaire