Université Sultan Moulay Slimane
Ecole National des Sciences Appliquées
Khouribga

Module : Data Mining & Texte Mining

1ère Année Master Big Data & Aide à la Décision

Semestre 2 / Année 2018/2019 / Feuille de TD N° 2

Exercice 1

1. Le data mining est le processus de recherche de modèles _____, nouveaux, utiles et exploitables dans un grand volume de données. Lequel des termes suivants remplit le mieux le vide ?

   A. volumineux

   B. hétérogènes

   C. valides

   D. Bruyants

2. La première étape du processus de data mining est généralement :

   A. La visualisation

   B. Le prétraitement

   C. La modélisation

   D. Le déploiement

3. Le nom d'une personne peut être considéré comme un attribut de type :

   A. Nominal

   B. Ordinal

   C. Intervalle

   D. Ratio

4. La hauteur d'une personne peut être considérée comme un attribut de type :

   A. Nominal

   B. Ordinal

   C. Intervalle

   D. Ratio

5. Lesquelles des opérations suivantes peuvent être effectuées sur des attributs nominaux ?

   A. Distinction

   B. Comparaison

C. Addition

D. Multiplication

6. Lesquelles des opérations suivantes peuvent être effectuées sur des attributs ordinaux ?

   A. Distinction

   B. Comparaison

   C. Les deux

   D. Aucune de ces réponses

7. La relation d'amitié des utilisateurs d'un site de réseau social peut être considérée comme un exemple de :

   A. Enregistrement de données

   B. Données ordonnées

   C. Données graphiques

   D. Aucune de ces réponses

8. Les colonnes d'une matrice de données stockant des enregistrements représentent généralement des :

   A. métadonnées

   B. objets

   C. attributs

   D. agrégats

9. Une valeur extrêmale est :

   A. une description d'enregistrements de données

   B. un point de données très différent des autres points de données

   C. un enregistrement avec attributs manquants

   D. un enregistrement en double

10. La réduction de la dimensionnalité est effectuée sur une matrice de données, la matrice de données transformée :

    A. a un nombre de lignes réduit

    B. a un nombre de colonnes réduit

    C. a un nombre de lignes et de colonnes réduit

    D. a le même nombre de lignes et de colonnes

11. L'échantillonnage est effectué sur une matrice de données, la matrice de données transformée :

    A. a un nombre de lignes réduit

B. a un nombre de colonnes réduit

C. a un nombre de lignes et de colonnes réduit

D. a le même nombre de lignes et de colonnes

12. L'analyse en composantes principales (ACP) est une technique pour effectuer :

A. L'échantillonnage

B. La discrétisation

C. La réduction de dimensionnalité

D. L'agrégation

Exercice 2

Classifiez les attributs suivants en binaires, discrets ou continus. Classifiez-les également comme qualitatifs (nominaux ou ordinaux) ou quantitatifs (intervalle ou ratio). Certains cas peuvent avoir plus d'une interprétation, alors indiquez brièvement votre raisonnement (si vous pensez qu'il peut y avoir une certaine ambiguïté). Exemple: Âge en années. Réponse: discret, quantitatif, ratio.

(a) Heure en termes d'AM ou PM.

(b) Luminosité mesurée par un posemètre.

(c) Luminosité mesurée par les jugements des personnes.

(d) Angles mesurés en degrés entre 0° et 360°.

(e) Médailles de bronze, d'argent et d'or attribuées aux Jeux olympiques.

(f) Hauteur au-dessus du niveau de la mer.

(g) Nombre de patients dans un hôpital.

(h) Numéros ISBN pour les livres.

(i) Capacité à laisser passer la lumière en termes des valeurs suivantes : opaque, transparent et translucide.

(j) Grade militaire.

(k) Distance du centre du campus.

(l) Densité d'une substance en grammes par centimètre cube.

(m) Numéro de chèque.

Exercice 3

Pour les attributs binaires, la distance L_1 correspond à la distance de Hamming, c'est-à-dire le nombre de bits différents entre deux vecteurs binaires. La similarité de Jaccard est une mesure de la similarité entre les vecteurs binaires (rapport des nombres de 1 qui coïncident par le nombre de 0 qui coïncident).

1. Calculez la distance de Hamming et la similarité de Jaccard entre les deux vecteurs binaires suivants : x = 0101010001 et y = 0100011000

2. Quelle approche, la similarité de Jaccard ou la distance de Hamming, est plus similaire à la mesure **SMC** (*Simple Matching Coefficient*) égale au quotient de la distance de Hamming par le nombre total de bits ? Quelle approche est plus similaire à la mesure du cosinus ? Expliquez.
3. Si vous souhaitez comparer la constitution génétique de deux organismes de même espèce, par exemple deux êtres humains, utiliseriez-vous la distance de Hamming, le coefficient de Jaccard ou une autre mesure de similarité ou de distance ? Expliquez (Notez que deux êtres humains partagent plus de 99,9% des mêmes gènes).

Exercice 4

1. A et B étant deux ensembles, A − B désigne leur différence. Montrer que la mesure définie par : $d(A, B) = |A – B| + |B – A|$ est une distance. Remarquez que : $d(A, B) = |A| + |B| - 2|A \cap B|$.
2. Etant donné une mesure de la similarité à valeurs dans l'intervalle [0, 1]. Décrire deux manières pour transformer cette mesure de similarité en une mesure de dissimilarité à valeurs dans l'intervalle $[0, +\infty]$.