EISTI	ING2-GI : DATA MINING - EXAMEN DU 17/12/2015
Rédigé par : Astrid Jourdan,Hervé de Milleville et Jean-Paul Forest	Ref : *DAT-MIN-EXA*
A l'intention de : Elèves d'ING2-GI	Créé le : 13/12/2015

Il s'agit de quelques indications de correction

1. Identification de tâches

Avertissement : les questions qui suivent n'ont pas nécessairement une seule réponse possible. De même, nous avons mis à chaque, une tâche à identifier. Il se peut que plusieurs tâches soit aussi une bonne réponse. Toutes vos réponses doivent être justifiées.

1) Dans les tâches de classification et de clustering, on parle de classes. Expliquer la différence des rôles des classes entre les deux tâches.
   Classification : classes données = cible / clustering : construction de classes
2) On demande de faire une première étude sur les cinéphiles. Les films sont entre autres classés par genre. Parmi les questions posées à nos cinéphiles, on leur a demandé de lister leurs genres préférés. On souhaite établir d'éventuels liens entre les genres.
   a. Quelle tâche identifiez-vous ?
   b. Sous quelle forme doit-on présenter les données pour les soumettre au logiciel ?

Instances = cinéphile / Attributs = genre 1, genre 2,…/ Régles d'association pour déterminer des régles du type SI films d'horreur ALORS films de science fiction

3) On reprend cette première étude en supposant que les personnes interviewées ont au moins 18 ans. Dans cette enquête, en plus des genres préférés, on a d'autres attributs comme l'âge, le sexe, le type d'études poursuivies, le type du lieu d'habitation (campagne, grande ville, ville moyenne, banlieue), le nombre de films vus dans une salle de cinéma par mois et d'autres variables. On s'interroge sur la possibilité d'expliquer les genres préférés en fonction de cette série d'attributs.
   a. Quelle tâche identifiez-vous ?
   b. Sous quelle forme doit-on présenter les données pour les soumettre au logiciel ?

   Instances=cinéphiles / Attributs = l'âge, le sexe, le type d'études poursuivies, le type du lieu d'habitation,… /Attribut cible = le genre préféré / prédiction du genre préféré en fonction des autres attributs par exemple avec une méthode arbre de décision si attributs explicatifs qualitatifs

4) On fait maintenant une deuxième étude. Cette étude n'est plus ciblée simplement sur les cinéphiles mais sur les personnes qui vont plus ou moins régulièrement au cinéma. On fait donc la même enquête à la sortie des séances de cinéma. On reprend la liste des attributs introduits dans la question précédente et on veut vérifier si cette population peut être partitionnée en fonction de ces attributs.
   a. Quelle tâche identifiez-vous ?
   b. Sous quelle forme doit-on présenter les données pour les soumettre au logiciel ?

Instances=personnes interviewées / Attributs = l'âge, le sexe, le type d'études poursuivies, le type du lieu d'habitation, leurs genres préférés,… / construire des groupes d'individus présentant des similarités (clustering)

5) On reprend cette deuxième étude. Si un individu interrogé n'est pas un cinéphile, il peut malgré tout avoir des genres de films préférés. On a donc demandé aussi à chaque interviewé sa liste de genres préférés. On suppose

que les résultats de la question précédente donne une partition probante. On souhaite établir un éventuel lien entre les différents sous ensembles de la partition d'une part et les genres de films préférés d'autre part.

a. Que proposez-vous pour répondre à ce souhait ?

Faire une AFC entre les deux variables qualitatives genre préféré d'une part et classe construite en 4) d'autre part

6) On décide de différencier les deux populations par leur taux de fréquentation des salles obscures en fixant un seuil mensuel. Un interviewé sera considéré comme un cinéphile si son taux de fréquentation est supérieur à ce seuil sinon il sera considéré comme un intéressé par le cinéma mais occasionnel. On désire expliquer les différences entre les cinéphiles et les personnes occasionnelles à travers leurs attributs.

a. Quelle tâche identifiez-vous ?

b. Sous quelle forme doit-on présenter les données pour les soumettre au logiciel ?

Instances=personnes interviewées / Attributs = l'âge, le sexe, le type d'études poursuivies, le type du lieu d'habitation, leurs genres préférés,… / Attribut cible binaire = fréquentation (acidu/occasionnel)/ Prédiction de la fréquentation en fonction des attributs à l'aide d'uine méthode supervisé comme un arbre de décision (si attributs explicatifs qualitatifs) ou ppv (si attributs

2. Les méthodes

1) Dans la méthode des k plus proches voisins, que se passe-t-il si on prend autant de voisins que de points dans la base d'apprentissage ? Une seule classe, celle de plus grand effectif
2) Dans la méthode de classification hiérarchique ascendante, nous avons énoncé quatre formules de dissimilarité (ou distance) pour fusionner les classes à chaque itération :

$$d_{\min}(C_1,C_2) = \min_{x\in C_1, y\in C_2} d(x,y)$$

$$d_{\max}(C_1,C_2) = \max_{x\in C_1, y\in C_2} d(x,y)$$

$$d_{moyenne}(C_1,C_2) = moyenne_{x\in C_1, y\in C_2} d(x,y)$$

$$d_{Ward}(C_1,C_2) = \frac{n_1 * n_2}{n_1+n_2} d(g_1,g_2)$$

dmin et dmax ne reposent que sur 2 points donc sensibles aux valeurs extremes des classes. dmoyenne et dward reposent sur l'ensemble des points mais dmoyenne ne tient pas compte des effectifs pouvant etre differents d'une classe à l'autre.

Expliciter en quoi ces formules mesurent des choses différentes.

3) Quelle est la différence entre la distance euclidienne d'une part et n'importe laquelle des quatre distances précédentes d'autre part ? distance euclidienne = distance entre points/ distances question 2 = distance entre classes
4) Dans la méthode des k-means, comment mesure-t-on la qualité de la segmentation obtenue ? inertie
5) Dans cette question, on ne parle que des méthodes supervisées.
   a. Pourquoi ne faut-il pas tester un modèle sur les données qui ont permis de l'obtenir ? sur-apprentissage et pauvre performance en prédiction
   b. Comment fait-on pour tester la qualité d'un modèle ? taux d'erreur sur une base test (n'ayant pas contribuée à l'élaboration du modèle)
6) Qu'appelle-t-on normaliser les données ? Pour quelles méthodes faut-il effectuer ce prétraitement ? Pourquoi ? Normaliser=centrer-réduite les variables pour les ramener au même ordre de grandeur dans le calcul des distances