Cycle Ingénieurs : Deuxième Année

Corrigé de l'Examen de fouille de données

17 février 2014

Durée 1h30 - Documents de cours et de TDs autorisés

Utiliser l'espace blanc prévu pour répondre

Nom :

Notations

Exercice 1.1 : (4Pts)
Exercice 1.2 : (3Pts)
Exercice 1.3 : (4Pts)
Exercice 1.4 : (3Pts)
Exercice 1.5 : (3Pts)
Exercice 1.6 : (3Pts)

Notes globales :

1 Apprentissage supervisé & non supervisé

Nous nous intéressons à un extrait adapté de la base *contact-lenses* contenant 6 exemples ; le but étant de classer les patients par rapport à leurs ports de lentilles de contact ou non. Trois attributs sont considérés selon le fichier . arff suivant :

@relation contact-lenses
@attribute spectacle-prescrip {myope, hypermetrope}
@attribute astigmatism {no, yes}
@attribute tear-prod-rate {reduced, normal}
@attribute contact-lenses {yes, no}
@data
myope,no,reduced,no
myope,no,normal,no
myope,yes,reduced,no
myope,yes,normal,yes
hypermetrope,no,reduced,no
hypermetrope,no,normal,yes

Exercice 1.1 Construire l'arbre de décision en utilisant la fonction de gain basée sur l'entropie. Représenter ensuite le modèle obtenu sous forme d'un ensemble de règles.

Corrigé 1.1 Nous calculons la fonction E(T) : la somme des entropies du partitionnement obtenu par le test T (si on place T à la racine de l'arbre) selon la formule donnée par le cours. Les tests possibles sont : spectacle-prescrip, astigmatism, tear-prod-rate.

$$E(spectacle - prescrip) = \frac{4}{6}I(1,3) + \frac{2}{6}I(1,1)$$

$$E(astimatism) = \frac{4}{6}I(1,3) + \frac{2}{6}I(1,1)$$

$$E(tear - prod - rate) = \frac{3}{6}I(0,3) + \frac{3}{6}I(2,1) = \frac{3}{6}I(2,1)$$

Il est clair que $E(tear - prod - rate)$ a la plus petite valeur donc c'est l'attribut tear-prod-rate qui sera placé dans la racine de l'arbre.

Nous procédons ainsi pour les sous arbres restants et nous obtenons enfin l'arbre suivant :

tear-prod-rate = reduced : no
tear-prod-rate = normal
| spectacle-prescrip = myope
| | astigmatism = no : no
| | astigmatism = yes : yes
| spectacle-prescrip = hypermetrope : yes

Exercice 1.2 Calculer l'erreur de l'arbre sur l'ensemble de test suivant :

@attribute spectacle-prescrip {myope, hypermetrope}
@attribute astigmatism {no, yes}
@attribute tear-prod-rate {reduced, normal}
@attribute contact-lenses {yes, no}
@data
hypermetrope,yes,reduced,no
hypermetrope,yes,normal,no

Corrigé 1.2 L'erreur de l'arbre sur les deux données est égale à 50%, la précision de la classe no est de 100% mais le rappel de la classe no est de 50%.

Exercice 1.3 Considérions de nouveau les 6 premiers exemples. Nous souhaitons étudier les associations entre les différentes valeurs des attributs y compris celles de la classe. Trouver un moyen pour représenter la base sous forme de transactions. Appliquer ensuite l'algorithme Apriori avec minSupp=2.

Corrigé 1.3 Les transactions sont les suivantes :

spectacle-prescrip=myope,astigmatism=no,tear-prod-rate=reduced,contact-lenses=no
spectacle-prescrip=myope,astigmatism=no,tear-prod-rate=normal,contact-lenses=no
spectacle-prescrip=myope,astigmatism=yes,tear-prod-rate=reduced,contact-lenses=no
spectacle-prescrip=myope,astigmatism=yes,tear-prod-rate=normal,contact-lenses=yes
spectacle-prescrip=hypermetrope,astigmatism=no,tear-prod-rate=reduced,contact-lenses=no
spectacle-prescrip=hypermetrope,astigmatism=no,tear-prod-rate=normal,contact-lenses=yes

Sous ensembles fréquents de longueur 1 L(1) :

spectacle-prescrip=myope 4
spectacle-prescrip=hypermetrope 2
astigmatism=no 4
astigmatism=yes 2
tear-prod-rate=reduced 3
tear-prod-rate=normal 3
contact-lenses=yes 2
contact-lenses=no 4

Sous ensembles fréquents de longueur 2 L(2) :

spectacle-prescrip=myope astigmatism=no 2
spectacle-prescrip=myope astigmatism=yes 2
spectacle-prescrip=myope tear-prod-rate=reduced 2
spectacle-prescrip=myope tear-prod-rate=normal 2
spectacle-prescrip=myope contact-lenses=no 3
spectacle-prescrip=hypermetrope astigmatism=no 2
astigmatism=no tear-prod-rate=reduced 2
astigmatism=no tear-prod-rate=normal 2
astigmatism=no contact-lenses=no 3
tear-prod-rate=reduced contact-lenses=no 3
tear-prod-rate=normal contact-lenses=yes 2

Sous ensembles fréquents de longueur 2 L(3) :

spectacle-prescrip=myope astigmatism=no contact-lenses=no 2
spectacle-prescrip=myope tear-prod-rate=reduced contact-lenses=no 2
astigmatism=no tear-prod-rate=reduced contact-lenses=no 2

Exercice 1.4 Extraire les règles d'association ayant une confiance minimale qui est égale à 1.

Corrigé 1.4 1. tear-prod-rate=reduced 3 ==> contact-lenses=no 3 conf :(1)
2. astigmatism=yes 2 ==> spectacle-prescrip=myope 2 conf :(1)
3. spectacle-prescrip=hypermetrope 2 ==> astigmatism=no 2 conf :(1)
4. contact-lenses=yes 2 ==> tear-prod-rate=normal 2 conf :(1)
5. spectacle-prescrip=myope astigmatism=no 2 ==> contact-lenses=no 2 conf :(1)
6. spectacle-prescrip=myope tear-prod-rate=reduced 2 ==> contact-lenses=no 2 conf :(1)
7. astigmatism=no tear-prod-rate=reduced 2 ==> contact-lenses=no 2 conf :(1)

Exercice 1.5 Nous nous intéressons à l'ensemble de règles ayant comme conclusion la valeur de la classe. Comparer avec les règles trouvées par l'algorithme ID3. Expliquer pourquoi certaines règles trouvés par ID3 ne sont pas présentes parmi les règles d'association.

Corrigé 1.5

La règle numéro 1 correspond à la première branche de l'arbre, les deux règles dont les numéros sont 5 et 7 ne se trouvent pas dans l'arbre. En fait la prise ne compte d'un support minimum dans l'algorithme apriori ne nous permet pas de retrouver toutes les règles trouvées par ID3 (comme les règles dont le support est 1 ici). D'un autre côté, le fait que Apriori cherche toutes les possibilités (même en présence de chauvechement), certaines règles comme 5 et 7 ne sont dans l'arbre.

Exercice 1.6 Nous utilisons Weka pour appliquer l'algorithme k-means avec $k = 2$ en mode "classes to cluster evaluation" sur l'ensemble des exemples. Nous obtenons la matrice de correspondance suivantes :

0 1 <– assigned to cluster
2 0 | yes
1 3 | no
Expliquer cette matrice. Quelles sont les caractéristiques des clusters trouvés ?

Corrigé 1.6 Le cluster numéro 1 est pur car il correspond à des exemples appartenant à la même classe, tandis que le cluster numéro 0 contient des exemples des deux classes. 3 exemples parmi 4 de la classe 1 sont similaires (leurs distances sont petites les unes par rapports aux autres) et peuvent être représentés par le barycentre du cluster 1.