TP DE DATA MINING
3 : REGLES D'ASSOCIATION
ET ANALYSE COMPLETE
AVEC SPSS CLEMENTINE

EPF – 4/ 5ème année - Option Ingénierie d'Affaires et de Projets - Finance

Bertrand LIAUDET

TP N°3 DE DATA MINING

Le 3ème TP a pour objectif de mettre en œuvre la modélisation de règles d'association sur un jeu de données correspondant à des tickets de caisse et une analyse complète sur une jeu de données correspondant à la description d'un produit.

Jeu de données de tickets de caisse

1 Compréhension des données

Faire l'analyse des données du fichier « tickets ».

On fera une analyse par variable (signification caractéristiques), puis par corrélation simple. A chaque fois, on tirera les conclusions. Certaines corrélations pourront justifier des analyses supplémentaires.

On finira par une synthèse : variables conservées et principales interprétations.

On fera aussi une analyse des co-occurrences.

2 Recherche d'associations

Pour la recherche d'association, on utilisera d'abord le modèle GRI proposé dans la version d'évaluation de SPSS-Clementine.

On fera une interprétation des résultats en utilisant les indicateurs de valeur d'une règle.

On comparera les résultats avec l'analyse des co-occurrences.

On produira une synthèse des résultats.

Ensuite, on utilisera les deux autres algorithmes proposés par SPSS-Clementine : l'algorithme A Priori et l'algorithme Carma. Quelles différences y a –t-il entre les résultats des différents algorithmes ?

Jeu de données du fichier des Céréales

1 Compréhension des données

Faire l'analyse des données du fichier : « Cereales ».

On fera une analyse par variable (signification, caractéristiques), puis par corrélation simple. A chaque fois, on tirera les conclusions. Certaines corrélations pourront justifier des analyses supplémentaires.

On finira par une synthèse : variables conservées et principales interprétations.

2 Modélisation non supervisée

Pour la modélisation, on va essayer de produire une classification.

On utilisera l'analyse en composantes principales et ou analyse factorielle, le K-mean et les réseaux de Kohonen. Pour l'analyse en composantes principales et ou factorielle, on utilisera au moins 3 méthodes.

On justifiera le nombre de classes finalement choisi en faisant plusieurs tests et en analysant à chaque fois les résultats obtenus.

Pour interpréter les classes obtenues, on utilisera la méthode de l'analyse exploratoire des classes et la méthode de la modélisation des classes par règles de décision mais aussi par réseau de neurones. Pour les réseaux de neurones, on prendra les trois méthodes vues en cours.

On essaiera de trouver une interprétation intuitivement compréhensible.

On conclura par le **choix** d'**une** classification (selon **un** modèle).

3 Modélisation supervisée

Quelle variable cible peut-on envisager ? Faites une modélisation en conséquence.

4 Description des variables du fichier

1. Name: Name of cereal
2. mfr: Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
3. type: cold or hot

4. calories: calories per serving
5. protein: grams of protein
6. fat: grams of fat
7. sodium: milligrams of sodium
8. fiber: grams of dietary fiber
9. carbo: grams of complex carbohydrates
10. sugars: grams of sugars
11. potass: milligrams of potassium
12. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
13. shelf: display shelf (1, 2, or 3, counting from the floor)
14. weight: weight in ounces of one serving
15. cups: number of cups in one serving
16. rating: a rating of the cereals

Rendu

La livraison se fera par mail à **liaudet.bertrand@wanadoo.fr.**

Date limite : dimanche 16 novembre minuit.

Il faut livrer un fichier word et ou pdf ainsi que les flux SPSS-Clementine numérotés de 1 à N. Le rapport fera référence à ces fichiers.

L'ensemble des documents se trouvent dans un dossier compressé dont le nom est : année-DM-noms des membre du binôme.