Examen - UP3

partie Clustering, Classification, Règles d'Association

le 12 décembre 2023

Les documents papier sont autorisés, de même qu'une calculatrice classique (comme celle utilisée au collège). Tout autre dispositif électronique est interdit.
L'ordre de résolution des sujets n'est pas imposé, bien au contraire.

1. **Transactions d'achat**
   La base suivante traduit une série de transactions de type "panier".

Id	Transaction
T100	C, A, F, E
T200	C, A, F, E, I, N
T300	C, I, E, L
T400	L, I, A, N, E
T500	E, L, A, I, T
T600	L, A, T, E

   Soient les limites du support ($min\_support$) à 49% et de la confiance ($min\_confiance$) à 80%.
   — Déroulez un algorithme vu en cours pour calculer tous les itemsets fréquents par rapport à $min\_support$ *(Attention : la limite est donnée en pourcentage !)*
   — A partir du résultat calculé au point précédent calculer toutes les règles d'association de type $X \rightarrow Z$ qui sont au-delà de $min\_confiance$ (X et Z sont des items).
2. **Item trop fréquent**
   Dans une base de transactions l'item F apparaît dans toutes les transactions. Quel est son impact dans le calcul des itemsets fréquents et dans les règles d'association ?
3. **Drôles de fonctions**
   Soit $de : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ la fonction de distance euclidienne.
   (a) Un ensemble de données dans $\mathbb{R}^2$ est donné en coordonnées polaires (r, ϕ). On considére $dp : (\mathbb{R}^+ \times [0, 2\pi[)^2 \rightarrow \mathbb{R}_+$ comme
   $$dp(x, y) = de((r_x, \phi_x), (r_y, \phi_y))$$
   Est-ce que cette fonction dp peut être utilisée dans un algorithme de clustering ?
   (b) Un ensemble de données dans $\mathbb{R}^2$ est donné en coordonnées cartésiennes (x, y). Soient $r1$ et $r2$ deux points fixes du plan. Pour chaque point, on défini $f : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ comme :
   $$f(x) = min(de(x, r1), de(x, r2))$$
   et pour une paire de points $df : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ comme :
   $$df(x, y) = |f(x) - f(y)|$$
   Même question : est-ce que cette fonction df peut être considérée comme une fonction distance ou dissimilarité ?

4. `mammal.dentition`

   Dans le TP sur le dataset `mammal.dentition` on devait mettre en place une classification avec k-NN. Une consigne indiquait de ne pas considérer les classes avec une seule observation. Pourquoi ?

5. **code R et résultats**

   On exécute le code suivant dans la console R :

```
n <- nrow(donnees) ## donnees est un dataframe avec 3 colonnes

I <- sample(1:n,(2*n)/3) # indices
J <- setdiff(1:n,I)  # autres indices

cl <- donnees[I,3] # la classe à apprendre

dlrn <- donnees[I,1:2] #
dtest <- donnees[J,1:2] #

library (class)

mknn3 <- knn(dlrn, dtest,cl, k=3)
t <- table(mknn3, donnees[J,3])
er <- (t[1,2] + t[2,1])/sum(t)
```

   Que fait ce code ?

   Pour une première exécution du code on obtient $er = 0.1174$ et pour une seconde exécution $er = 0.08823$. Est-ce normal ?