Correction - examen écrit du 26 novembre 2015

Règles d'association
Majeure Science des données

1er décembre 2015

1. Dans le script il manquait la bibliothèque `arulesViz`.
2. Pour des nouvelles données incomplètes ou non étiquetées, si on dispose des règles d'associationde type $\alpha \to X$, avec X l'attribut manquant, on peut en déduire la valeur de X ou sa présence/absence avec une probabilité égale à la confiance de la règle.

   On fait donc de la prédiction.
3. Il faut prouver que pour d items distincts le nombre de règles d'association est

   $$3^d - 2^{d+1} + 1$$

   sans sens prendre en compte les règles avec la partie droite ou la partie gauche nulle.

   La solution n'est pas unique et votre solution peut être encore différente de celles-ci.

   Solution 1 (suggérée par un élève à la sortie de l'examen) : par récurrence.

   Pour $d = 2$ on a deux RA possibles. D'autre part $3^2 - 2^3 + 1 = 9 - 8 + 1 = 2$.

   Supposons que l'affirmation est vraie pour d et on démontre pour $d+1$.

Soit $\mathcal{R}_d$ l'ensemble des RA sur d attributs sans parties gauche ou droite nulles ; soient $\mathcal{GV}_d$ et $\mathcal{DV}_d$ les ensembles de RA sur d attributs avec la partie gauche, respectivement, droite nulle.

Remarquons que :

— selon l'hypothèse de récurrence $card(\mathcal{R}_d) = 3^d - 2^{d+1} + 1$

— les ensembles $card(\mathcal{GV}_d) = card(\mathcal{DV}_d) = 2^d$

— les ensembles avec les RA contenant l'ensemble vide à gauche ou à droite sont disjoints : $\mathcal{GV}_d \cap \mathcal{DV}_d = \Phi$

Soit α l'attribut numéro $d+1$, l'ensemble $\mathcal{R}_{d+1}$ contient :

— $\mathcal{R}_d$

— des RA extraites de $\mathcal{R}_d$ auxquelles on a rajouté α sur la partie droite

— des RA extraites de $\mathcal{R}_d$ auxquelles on a rajouté α sur la partie gauche

— des RA de forme : $X \to \alpha$ et $\alpha \to X$ avec X ensemble non nul d'attributs à exception de α

Les sous ensembles cités forment une partition de $\mathcal{R}_{d+1}$.

Alors $card(\mathcal{R}_{d+1}) = 3 \times card(\mathcal{R}_d) + 2(2^d - 1) = 3 \times (3^d - 2^{d+1} + 1) + 2^{d+1} - 2 = 3^{d+1} - 3 \times 2^{d+1} + 3 + 2^{d+1} - 2 = 3^{d+1} - 2^{d+2} + 1$.

Solution 2 (celle que j'avais pensée) - compter les RA ... en base 3.

Soient $\{i_1, i_2, \cdots, i_d\}$ les attributs qui apparaissent dans notre base de transactions.

On peut mettre en évidence une bijection entre les RA et les nombres en base 3 écris sur d chiffres : $\overline{c_1 c_2 \dots c_d}_{(3)}$. La correspondance est faite de la manière suivante :

$$c_k = \begin{cases} 0 & , \ i_k \text{ n'apparait pas dans la RA} \\ 1 & , \ i_k \text{ apparait dans la partie gauche} \\ 2 & , \ i_k \text{ apparait dans la partie droite} \end{cases}$$

Les nombres de ces nombres est de 3^d. Pour compter les RA qui nous intéressent on doit exclure les nombres écrits uniquement avec de 0 et de 1 et, respectivement, uniquement avec de 0 et de 2, aussi le nombre $00\dots0$ écrit avec des 0 uniquement.

Donc on a $3^d - 2 \times (2^d - 1) - 1 = 3^d - 2^{d+1} + 1$.

Solution 3 (suggérée par des élèves forts en algèbre et combinatoire) - basée sur un comptage.

La partie gauche contient entre 1 et d items. Pour un k donné, $0 < k < d$, on peut choisir la partie gauche selon C_n^k possibilités [1], parmi les $n - k$ itemsets restants on peut choisir la partie droite parmi les $2^{d-k} - 1$ possibilités.
Le nombre total des choix est donc

$$\begin{aligned}
\sum_{k=1}^{d-1} C_d^k \times (2^{d-k} - 1) &= \sum_{k=1}^{d-1} C_d^k \times 2^{d-k} - \sum_{k=1}^{d-1} C_d^k \\
&= (\sum_{k=0}^{d} C_d^k \times 2^{d-k} - 1 - 2^d) - \sum_{k=0}^{d} C_d^k + 2 \\
&= (1+2)^d - 1 - 2^d - 2^d + 2 \\
&= 3^d - 2^{d+1} + 1
\end{aligned}$$

4. Les RA pour un itemset dédié

   Avec un seul parcours de la base de transactions on pré-calcule le support de l'itemset $\{i_1, i_2, \ldots, i_k\}$ et on détecte et on garde aussi les items $\{j_1, j_2, \ldots j_r\}$ qui sont différents des items d'entrée et qui apparaissent dans les mêmes transactions que au moins un item $i_l, , 1 \leq l \leq k$. Si le support de l'itemset $\{i_1, i_2, \ldots, i_k\}$ est plus petit que le *min_support*, on ne fait rien.
   Sinon, on travaillera sur un l'ensemble $I' = \{i_1, i_2, \ldots, i_k\} \cap \{j_1, j_2, \ldots j_r\}$ qui est un ensemble réduit de I ($I' \subset I$) et sur des transaction $\mathcal{D}' \subset \mathcal{D}$, $\mathcal{D}' = \{d \in \mathcal{D} | d$ contient au moins un item parmi $i_1, i_2, \ldots, i_k\}$.
   Ensuite on applique soit :
   — l'algorithme A priori, mais lors de la constitution des candidats C_2 on génère uniquement les paires dont au moins un item soit parmi $\{i_1, i_2, \ldots, i_k\}$

1. la notation C_n^k est équivalente à $\binom{n}{k}$ et indique le coefficient binomial

— ECLAT : le parcours en profondeur se fait en explorant uniquement les noeuds $\{i_l\}$, $1 \leq l \leq k$ dans cet ordre.

5. Extraction des RA depuis une base de type "panier d'achat".
   Les données sont :

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	L, A, I, T
T600	C, A, T, E

   Selon les données $min\_support = 49\%$ signifie $min\_count\_support = 3$.
   En déroulant l'algorithme A priori on obtient successivement les candidats et itemsets fréquents suivants :

C_1	Support
C	4
A	5
F	2
E	5
I	4
N	2
L	3
T	2

$\Rightarrow$

L_1	Support
C	4
A	5
E	5
I	4
L	3

$\Rightarrow$

C_2	Support
C,A	3
C,E	4
C,I	2
C,L	1
A,E	4
A,I	3
A,L	2
E,I	3
E,L	2
I,L	3

$\Rightarrow$

L_2	Support
C,A	3
C,E	4
A,E	4
A,I	3
E,I	3
I,L	3

$\Rightarrow$

C_3	Support ou élagage
C, A, E	3
C, A, I	éliminé car C,I pas fréquent
C, E, I	éliminé, idem
A, E, I	2, éliminé
A, I, L	éliminé, car A, L pas fréquent

$\Rightarrow$

L_3	Support
C, A, E	3

   Les itemsets fréquents sont ; C, A, E, I, L, CA, CE, AE, AI, EI, IL, CAE.

Seulement l'itemset C,A,E peut engendrer des RA de type $X, Y \to Z$.
Les RA générées sont :

Règle d'association	Confiance	LIFT	Décision
C,A $\to$ E	1	6/5	
C,E $\to$ A	3/4	9/10	éliminé
O,N $\to$ Y	3/4	9/8	éliminé

On garde donc uniquement la règle d'association $C, A \to E$ en raison de la $min\_confiance = 80\%$.

Grille de correction :

- *2pt d'office*
- `arulesViz` *pas noté*
- *2pt pour la notion de prédiction depuis des RA*
- *4pt démonstration correcte de* $3^d - 2^{d+1} + 1$
- *4pt pour les RA sur une restriction d'items*
- *8pt pour le panier d'achat*
  - $min\_count\_support = 3$ *0.5pt*
  - C_1, L_1 *1,5pt*
  - C_2, L_2 *1,5pt*
  - C_3, L_3 *1,5pt*
  - *affichage explicit des itemsets fréquents 1pt*
  - *les 3 RA avec confiance, LIFT, décision 1.5pt*
  - *justification du choix* $C, A \to E$ *0.5pt*