TP 4
ACQUSITION DE CONNAISSANCES 2

Damien *Crémilleux* - Lauriane *Holy*

26 février 2014

1 Génération de règles d'association

1. On constate que Weka génère les différents k-itemsets fréquents. Weka construit ensuite les règles et ne garde que celle dont la confiance est égale à 1, c'est-à-dire que la règle est vérifiée par chaque instance. Cela est également vérifié par le support (chaque partie de règle à le même support).

2. Confidence

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6

Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf
    :(1)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4
    conf:(1)
 4. outlook=sunny play=no 3 ==> humidity=high 3
    conf:(1)
 5. outlook=sunny humidity=high 3 ==> play=no 3
    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3
    conf:(1)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3
    conf:(1)
 8. temperature=cool play=yes 3 ==> humidity=normal
    3    conf:(1)
```

```
 9. outlook=sunny temperature=hot 2 ==> humidity=
    high 2    conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2
      conf:(1)
```

Lift

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 1

Best rules found:

 1. temperature=cool 4 ==> humidity=normal 4    conf
    :(1) < lift:(2)> lev:(0.14) [2] conv:(2)
 2. humidity=normal 7 ==> temperature=cool 4    conf
    :(0.57) < lift:(2)> lev:(0.14) [2] conv:(1.25)
 3. humidity=high 7 ==> play=no 4    conf:(0.57) <
    lift:(1.6)> lev:(0.11) [1] conv:(1.13)
 4. play=no 5 ==> humidity=high 4    conf:(0.8) <
    lift:(1.6)> lev:(0.11) [1] conv:(1.25)
 5. outlook=overcast 4 ==> play=yes 4    conf:(1) <
    lift:(1.56)> lev:(0.1) [1] conv:(1.43)
 6. play=yes 9 ==> outlook=overcast 4    conf:(0.44)
     < lift:(1.56)> lev:(0.1) [1] conv:(1.07)
 7. humidity=normal windy=FALSE 4 ==> play=yes 4
    conf:(1) < lift:(1.56)> lev:(0.1) [1] conv
    :(1.43)
 8. play=yes 9 ==> humidity=normal windy=FALSE 4
    conf:(0.44) < lift:(1.56)> lev:(0.1) [1] conv
    :(1.07)
 9. humidity=normal 7 ==> play=yes 6    conf:(0.86)
    < lift:(1.33)> lev:(0.11) [1] conv:(1.25)
10. play=yes 9 ==> humidity=normal 6    conf:(0.67)
   < lift:(1.33)> lev:(0.11) [1] conv:(1.13)
```

On constate que moins d'itemsets sont générés (22 contre 104). Les règles générées ne sont donc pas les mêmes qu'avec la confiance (seules deux règles sont communes sur les 10). Seules les 10 règles avec un lift élévé (donc corrélées) sont retenues.

Leverage

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
```

```
Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 1

Best rules found:

 1. temperature=cool 4 ==> humidity=normal 4    conf
    :(1) lift:(2) < lev:(0.14) [2]> conv:(2)
 2. humidity=normal 7 ==> temperature=cool 4    conf
    :(0.57) lift:(2) < lev:(0.14) [2]> conv:(1.25)
 3. humidity=normal 7 ==> play=yes 6    conf:(0.86)
    lift:(1.33) < lev:(0.11) [1]> conv:(1.25)
 4. play=yes 9 ==> humidity=normal 6    conf:(0.67)
    lift:(1.33) < lev:(0.11) [1]> conv:(1.13)
 5. humidity=high 7 ==> play=no 4    conf:(0.57)
    lift:(1.6) < lev:(0.11) [1]> conv:(1.13)
 6. play=no 5 ==> humidity=high 4    conf:(0.8) lift
    :(1.6) < lev:(0.11) [1]> conv:(1.25)
 7. outlook=overcast 4 ==> play=yes 4    conf:(1)
    lift:(1.56) < lev:(0.1) [1]> conv:(1.43)
 8. play=yes 9 ==> outlook=overcast 4    conf:(0.44)
     lift:(1.56) < lev:(0.1) [1]> conv:(1.07)
 9. humidity=normal windy=FALSE 4 ==> play=yes 4
    conf:(1) lift:(1.56) < lev:(0.1) [1]> conv
    :(1.43)
10. play=yes 9 ==> humidity=normal windy=FALSE 4
   conf:(0.44) lift:(1.56) < lev:(0.1) [1]> conv
   :(1.07)
```

On constate que les règles générées avec le leverage sont très proches de celle avec le lift, ce qui est normal, car il ne s'agit que d'une variante de cette mesure.

Conviction

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 26

Size of set of large itemsets L(3): 4

Best rules found:

 1. temperature=cool 4 ==> humidity=normal 4    conf
    :(1) lift:(2) lev:(0.14) [2] < conv:(2)>
 2. outlook=sunny humidity=high 3 ==> play=no 3
    conf:(1) lift:(2.8) lev:(0.14) [1] < conv:(1.93)
    >
```

```
 3. outlook=sunny play=no 3 ==> humidity=high 3
    conf:(1) lift:(2) lev:(0.11) [1] < conv:(1.5)>
 4. temperature=cool play=yes 3 ==> humidity=normal
    3    conf:(1) lift:(2) lev:(0.11) [1] < conv
    :(1.5)>
 5. outlook=overcast 4 ==> play=yes 4    conf:(1)
    lift:(1.56) lev:(0.1) [1] < conv:(1.43)>
 6. humidity=normal windy=FALSE 4 ==> play=yes 4
    conf:(1) lift:(1.56) lev:(0.1) [1] < conv:(1.43)
    >
 7. play=no 5 ==> outlook=sunny humidity=high 3
    conf:(0.6) lift:(2.8) lev:(0.14) [1] < conv
    :(1.31)>
 8. humidity=high play=no 4 ==> outlook=sunny 3
    conf:(0.75) lift:(2.1) lev:(0.11) [1] < conv
    :(1.29)>
 9. outlook=rainy play=yes 3 ==> windy=FALSE 3
    conf:(1) lift:(1.75) lev:(0.09) [1] < conv
    :(1.29)>
10. humidity=normal 7 ==> play=yes 6    conf:(0.86)
   lift:(1.33) lev:(0.11) [1] < conv:(1.25)>
```

Les règles sont là aussi différentes. Les différentes méthodes de mesures ne servent donc pas à rechercher les mêmes caractéristiques. Ainsi, une bonne confiance ne signifie pas forcément un bon lift.

3. On choisit la règle :

```
1. temperature=cool 4 ==> humidity=normal 4    conf
   :(1) < lift:(2)> lev:(0.14) [2] conv:(2)
```

— Vérification de la confiance :

$$confiance_{temperature=cool->humidity=normal} = \frac{N_{temperature=cool,Humidity=normal}}{N_{temperature=cool}} = \frac{4}{4} = 1$$

— Vérification du lift :

$$\begin{aligned} lift_{temperature=cool->humidity=normal} &= \frac{confiance_{temperature=cool->humidity=normal}}{confiance_{univers->humidity=normal}} \\ &= N_{univers} \times \frac{N_{temperature=cool,Humidity=normal}}{N_{temperature=cool} \times N_{humidity=normal}} \\ &= 4 \times \frac{4}{4 \times 7} \\ &= 2 \end{aligned}$$

— Vérification du leverage :

$$
\begin{aligned}
leverage_{temperature=cool->humidity=normal} &= sup_{temperature=cool \cup Humidity=normal} \\
&- sup_{temperature=cool} \times sup_{Humidity=normal} \\
&= \frac{10}{1} - \frac{6}{14} \times \frac{7}{4} \\
&= 0.5
\end{aligned}
$$

Le calcul ne semble pas correspondre aux chiffres donnés par Weka :(

4. La conviction correspond à la probabilité de ne pas avoir B, sur la probabilité que A n'engendre pas B. Plus cette mesure est proche de 1, plus cela signifie que la présence de A implique l'absence de B.

2 Weka pour l'étude de la population américaine

Question 1.8 Nous appliquons l'algorithme APriori, en sélectionnant les règles dont la confiance est supérieure à 0.5. Les régles qui nous intéressent sont celle dont la partie droite est gain>=50K. Nous obtenons :

```
education-num='(12.5-inf)' race=_White 54 ==> gain=_>50
    K 28    conf:(0.52)
 education-num='(12.5-inf)' race=_White native-country=
     _United-States 49 ==> gain=_>50K 25    conf:(0.51)
  education-num='(12.5-inf)' native-country=_United-
      States 56 ==> gain=_>50K 28    conf:(0.5)
```

Ainsi nous constatons qu'être blanc, une éducation poussée, et être originaire des États-Unis sont des facteurs conduisant à un gain élevé.

Question 1.9 Nous n'obtenons que les règles possèdant le gain en partie droite. Le résultat obtenu confirme le résulat précédent.

3 Etude de cas : Articles de presse

Question 2.1 Le nombre d'article modifie la complexité pour l'étape de construction des itemsets les plus fréquents. En effet, le nombre d'itemsets générés est en 2 puissance N.

Question 2.2 Pour chaque mot du fichier mot.lst, nous créeons un attribut binaire qui représente sa présence au sein de l'article. La liste des attributs est générée grâce au script perl script.pl join à ce compte rendu (et un retour à la ligne ajouté manuellement ensuite). Cependant la génération des règles à l'aide de Weka échoue, avec le message d'erreur suivant :

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9
    -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
```

```
Relation:     resultat-weka.filters.unsupervised.
    attribute.NominalToBinary-Rfirst-last
Instances:    468
Attributes:   200
              [list of attributes omitted]
```

Nous n'avons malheureusement pas réussi à résoudre ce problème.