Homework 3

Pattern Mining and Social Network Analysis

BOUYSSOU Gatien , de POURTALES Caroline, LAMBA Ankit

07 novembre, 2020

Contents

Parameters in association rules

There are parameters controlling the number of rules to be generated.

For A => B :

Support

Support is an indication of how frequently the itemset appears in the dataset.

$$Support = \frac{\text{Number of transaction with both A and B}}{\text{Total Number of transaction}} = P(A \cap B)$$

Confidence

Confidence is an indication of how often the rule has been found to be true.

$$Confidence = \frac{\text{Number of transaction with both A and B}}{\text{Total Number of transaction with A}} = \frac{P(A \cap B)}{P(A)}$$

Lift

Lift is the factor by which, the co-occurence of A and B exceeds the expected probability of A and B co-occuring, had they been independent. So, higher the lift, higher the chance of A and B occurring together.

$$Lift = \frac{P(A \cap B)}{P(A) * P(B)}$$

Apriori algorithm

Definition

Apriori searches for frequent itemset browsing the lattice of itemsets in breadth.
The database is scanned at each level of lattice. Additionally, Apriori uses a pruning technique based on the properties of the itemsets, which are: If an itemset is frequent, all its sub-sets are frequent and not need to be considered.

Example on Groceries data

```
##     items
## [1] {citrus fruit,
##      semi-finished bread,
##      margarine,
##      ready soups}
## [2] {tropical fruit,
##      yogurt,
##      coffee}
## [3] {whole milk}
## [4] {pip fruit,
##      yogurt,
##      cream cheese ,
##      meat spreads}
## [5] {other vegetables,
##      whole milk,
##      condensed milk,
##      long life bakery product}
## [6] {whole milk,
##      butter,
##      yogurt,
##      rice,
##      abrasive cleaner}
```

On R

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.2    0.1    1 none FALSE            TRUE       5    0.03      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 295
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
## sorting and recoding items ... [44 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
```

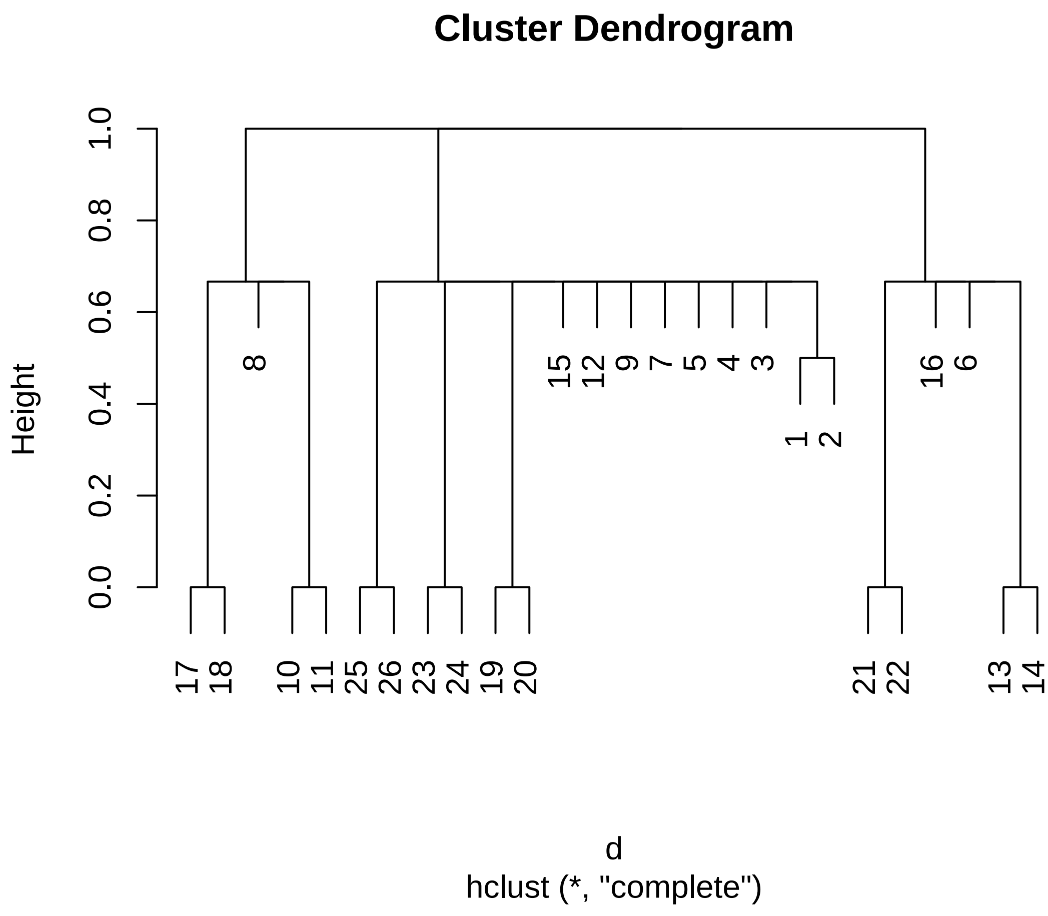
```
## writing ... [26 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
## set of 26 rules
```

```
##     lhs                        rhs                support    confidence coverage
## [1] {whipped/sour cream} => {whole milk}       0.03223183 0.4496454  0.07168277
## [2] {root vegetables}    => {whole milk}       0.04890696 0.4486940  0.10899847
## [3] {root vegetables}    => {other vegetables} 0.04738180 0.4347015  0.10899847
## [4] {tropical fruit}     => {whole milk}       0.04229792 0.4031008  0.10493137
## [5] {yogurt}             => {whole milk}       0.05602440 0.4016035  0.13950178
##     lift     count
## [1] 1.759754 317
## [2] 1.756031 481
## [3] 2.246605 466
## [4] 1.577595 416
## [5] 1.571735 551
```

```
##  [1] 1 1 1 1 1 2 1 3 1 3 3 1 2 2 1 2 4 4 1 1 5 5 1 1 1 1
```

```
##  [1] 1 1 1 1 1 2 1 3 1 3 3 1 2 2 1 2 3 3 1 1 2 2 1 1 1 1
```



Using Frequent itemset to find rules

Concept

TO DO

Example on personal data

We cal also use the ruleInduction method to find closed frequent itemset.

ruleInduction has as attribute a method function.

Closed Frequent itemsets :

An itemset X is a closed frequent itemset in set S if X is both closed and frequent in S.

Eclat algorithm :

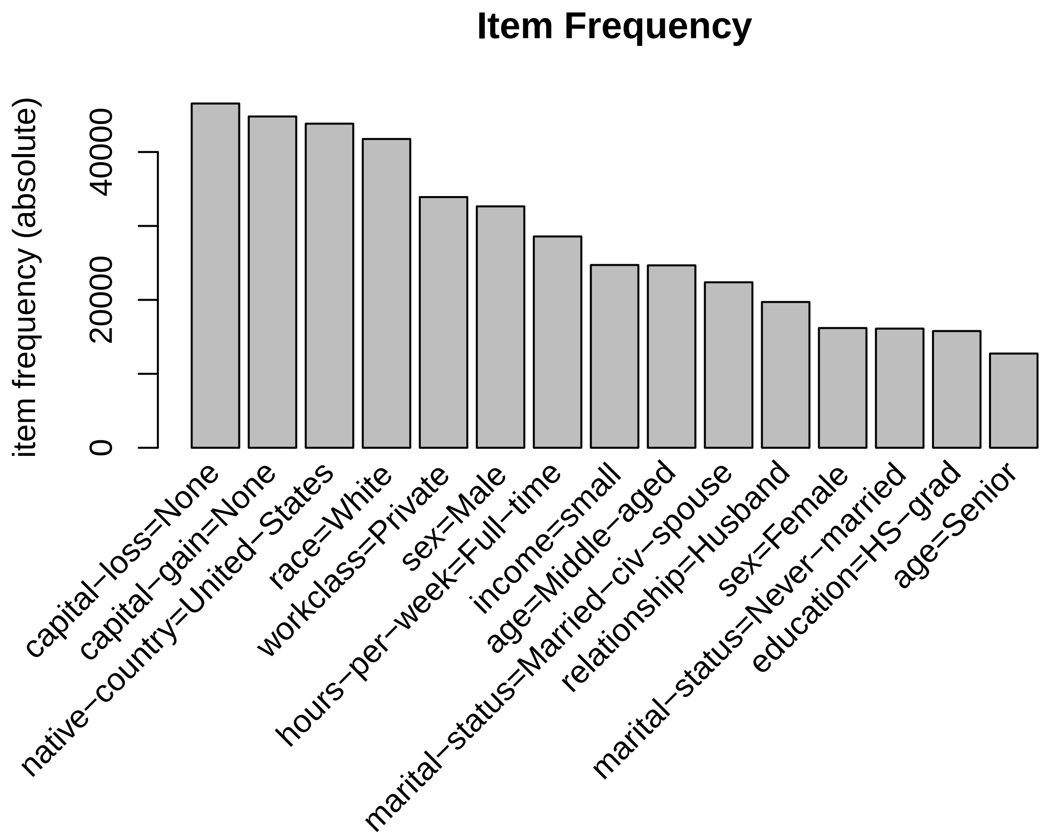
Mine frequent itemsets
This algorithm uses simple intersection operations for equivalence class clustering along with bottom-up lattice traversal.

On R

```
##     items                              transactionID
## [1] {age=Middle-aged,
##      workclass=State-gov,
##      education=Bachelors,
##      marital-status=Never-married,
##      occupation=Adm-clerical,
##      relationship=Not-in-family,
##      race=White,
##      sex=Male,
##      capital-gain=Low,
##      capital-loss=None,
##      hours-per-week=Full-time,
##      native-country=United-States,
##      income=small}                                 1
## [2] {age=Senior,
##      workclass=Self-emp-not-inc,
##      education=Bachelors,
##      marital-status=Married-civ-spouse,
##      occupation=Exec-managerial,
##      relationship=Husband,
##      race=White,
##      sex=Male,
##      capital-gain=None,
##      capital-loss=None,
##      hours-per-week=Part-time,
##      native-country=United-States,
##      income=small}                                 2
## [3] {age=Middle-aged,
##      workclass=Private,
##      education=HS-grad,
##      marital-status=Divorced,
##      occupation=Handlers-cleaners,
##      relationship=Not-in-family,
```

```
##      race=White,
##      sex=Male,
##      capital-gain=None,
##      capital-loss=None,
##      hours-per-week=Full-time,
##      native-country=United-States,
##      income=small}                                    3
## [4] {age=Senior,
##      workclass=Private,
##      education=11th,
##      marital-status=Married-civ-spouse,
##      occupation=Handlers-cleaners,
##      relationship=Husband,
##      race=Black,
##      sex=Male,
##      capital-gain=None,
##      capital-loss=None,
##      hours-per-week=Full-time,
##      native-country=United-States,
##      income=small}                                    4
## [5] {age=Middle-aged,
##      workclass=Private,
##      education=Bachelors,
##      marital-status=Married-civ-spouse,
##      occupation=Prof-specialty,
##      relationship=Wife,
##      race=Black,
##      sex=Female,
##      capital-gain=None,
##      capital-loss=None,
##      hours-per-week=Full-time,
##      native-country=Cuba,
##      income=small}                                    5
```

```
## Eclat
##
## parameter specification:
##  tidLists support minlen maxlen            target  ext
##     FALSE    0.01      1    100 frequent itemsets TRUE
##
## algorithmic control:
##  sparse sort verbose
##       7   -2    TRUE
##
## Absolute minimum support count: 488
##
## create itemset ...
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.06s].
## sorting and recoding items ... [67 item(s)] done [0.01s].
## creating bit matrix ... [67 row(s), 48842 column(s)] done [0.01s].
## writing  ... [80228 set(s)] done [0.30s].
## Creating S4 object  ... done [0.02s].
```



If in control method = "apriori" is used, a very simple rule induction method is used. All rules are mined from the transactions data set using Apriori with the minimal support found in itemsets. And in a second step all rules which do not stem from one of the itemsets are removed. This procedure will be in many cases very slow (e.g., for itemsets with many elements or very low support).

```
##     lhs                                       rhs                     support confidence     lift
## [1] {marital-status=Married-civ-spouse,
##      sex=Female,
##      capital-gain=None,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01095369  0.9870849 20.68263
## [2] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      capital-gain=None,
##      income=large}                         => {relationship=Wife} 0.01076942  0.9868668 20.67806
## [3] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01238688  0.9837398 20.61254
## [4] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      capital-loss=None,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01113796  0.9837251 20.61223
## [5] {marital-status=Married-civ-spouse,
```

```
##      sex=Female,
##      capital-gain=None,
##      income=large}                         => {relationship=Wife} 0.01220261  0.9834983 20.60748
```

If in control method = "ptree" is used, the transactions are counted into a prefix tree and then the rules are selectively generated using the counts in the tree. This is usually faster than the above approach.

```
##     lhs                                    rhs                      support confidence     lift itemset
## [1] {marital-status=Married-civ-spouse,
##      sex=Female,
##      capital-gain=None,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01095369  0.9870849 20.68263    559
## [2] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      capital-gain=None,
##      income=large}                         => {relationship=Wife} 0.01076942  0.9868668 20.67806    558
## [3] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01238688  0.9837398 20.61254    558
## [4] {marital-status=Married-civ-spouse,
##      race=White,
##      sex=Female,
##      capital-loss=None,
##      native-country=United-States,
##      income=large}                         => {relationship=Wife} 0.01113796  0.9837251 20.61223    558
## [5] {marital-status=Married-civ-spouse,
##      sex=Female,
##      capital-gain=None,
##      income=large}                         => {relationship=Wife} 0.01220261  0.9834983 20.60748    559
```

NOW THE BIG QUESTION ???

How to win money ?

```
## Eclat
##
## parameter specification:
##  tidLists support minlen maxlen            target  ext
##     FALSE    0.01      1    200 frequent itemsets TRUE
##
## algorithmic control:
##  sparse sort verbose
##       7   -2    TRUE
##
## Absolute minimum support count: 488
##
## create itemset ...
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.06s].
## sorting and recoding items ... [67 item(s)] done [0.01s].
## creating bit matrix ... [67 row(s), 48842 column(s)] done [0.01s].
## writing  ... [80228 set(s)] done [0.27s].
## Creating S4 object  ... done [0.03s].
```

```
## set of 14 rules
##      lhs                               rhs                     support confidence     lift
## [1]  {capital-loss=None,
##       hours-per-week=Over-time,
##       income=large}                  => {capital-gain=High} 0.01148602  0.1817887 5.253802
## [2]  {race=White,
##       capital-loss=None,
##       hours-per-week=Over-time,
##       income=large}                  => {capital-gain=High} 0.01052373  0.1779778 5.143665
## [3]  {capital-loss=None,
##       hours-per-week=Over-time,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.01046231  0.1779248 5.142132
## [4]  {hours-per-week=Over-time,
##       income=large}                  => {capital-gain=High} 0.01148602  0.1625145 4.696765
## [5]  {capital-loss=None,
##       income=large}                  => {capital-gain=High} 0.02319725  0.1602999 4.632763
## [6]  {capital-loss=None,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.02119078  0.1600680 4.626061
## [7]  {hours-per-week=Over-time,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.01046231  0.1594881 4.609302
## [8]  {race=White,
##       hours-per-week=Over-time,
##       income=large}                  => {capital-gain=High} 0.01052373  0.1591824 4.600466
## [9]  {race=White,
##       capital-loss=None,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.01951190  0.1578860 4.562999
## [10] {race=White,
##       capital-loss=None,
##       income=large}                  => {capital-gain=High} 0.02069940  0.1576977 4.557557
## [11] {sex=Male,
##       capital-loss=None,
##       income=large}                  => {capital-gain=High} 0.01887720  0.1539232 4.448472
## [12] {sex=Male,
##       capital-loss=None,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.01719831  0.1529776 4.421143
## [13] {race=White,
##       sex=Male,
##       capital-loss=None,
##       income=large}                  => {capital-gain=High} 0.01705499  0.1520906 4.395507
## [14] {race=White,
##       sex=Male,
##       capital-loss=None,
##       native-country=United-States,
##       income=large}                  => {capital-gain=High} 0.01605176  0.1518203 4.387696
```

Frequent pattern-based cluster analysis

Clustering with Apriori algorithm as dissimilarity measure

Concept

TO DO

Example on tennis data

On R

```
##     items        transactionID
## [1] {Result=0,
##      ACE.1=Low,
##      UFE.1=Low,
##      ACE.2=Low,
##      UFE.2=Low}               1
## [2] {Result=0,
##      ACE.1=None,
##      UFE.1=Low,
##      ACE.2=High,
##      UFE.2=Low}               2
## [3] {Result=1,
##      ACE.1=Low,
##      UFE.1=Low,
##      ACE.2=Low,
##      UFE.2=High}              3
## [4] {Result=1,
##      ACE.1=High,
##      UFE.1=High,
##      ACE.2=None,
##      UFE.2=High}              4
## [5] {Result=0,
##      ACE.1=Low,
##      UFE.1=High,
##      ACE.2=High,
##      UFE.2=High}              5
```

The associations rules for Player-1 winning :

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.3    0.1    1 none FALSE            TRUE       5    0.15      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 17
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[12 item(s), 118 transaction(s)] done [0.00s].
```

```
## sorting and recoding items ... [11 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [13 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
## set of 13 rules
```

```
##     lhs                        rhs          support   confidence coverage
## [1] {ACE.1=High,UFE.1=Low} => {Result=1} 0.1525424 0.8181818  0.1864407
## [2] {ACE.1=High}           => {Result=1} 0.2881356 0.6938776  0.4152542
## [3] {ACE.1=High,UFE.2=Low} => {Result=1} 0.1694915 0.6451613  0.2627119
## [4] {ACE.2=Low}            => {Result=1} 0.2457627 0.6170213  0.3983051
## [5] {UFE.1=Low}            => {Result=1} 0.3220339 0.6129032  0.5254237
##     lift     count
## [1] 1.532468 18
## [2] 1.299644 34
## [3] 1.208397 20
## [4] 1.155691 29
## [5] 1.147977 38
```

The associations rules for Player-1 loosing :

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.3    0.1    1 none FALSE            TRUE       5    0.15      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 17
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[12 item(s), 118 transaction(s)] done [0.00s].
## sorting and recoding items ... [11 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [10 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
## set of 10 rules
```

```
##     lhs                          rhs          support   confidence coverage
## [1] {ACE.2=High}             => {Result=0} 0.2372881 0.6086957  0.3898305
## [2] {UFE.1=High}             => {Result=0} 0.2627119 0.5535714  0.4745763
## [3] {ACE.1=Low}              => {Result=0} 0.2372881 0.5283019  0.4491525
## [4] {UFE.2=Low}              => {Result=0} 0.2796610 0.5238095  0.5338983
## [5] {UFE.1=High,UFE.2=High} => {Result=0} 0.1525424 0.4864865  0.3135593
##     lift     count
## [1] 1.305929 28
## [2] 1.187662 31
## [3] 1.133448 28
```

```
## [4] 1.123810 33
## [5] 1.043735 18
```

All the rules with Result as association :

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.4    0.1    1 none FALSE            TRUE       5     0.2      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 23
##
## set item appearances ...[2 item(s)] done [0.00s].
## set transactions ...[12 item(s), 118 transaction(s)] done [0.00s].
## sorting and recoding items ... [11 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [14 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
## set of 14 rules
```

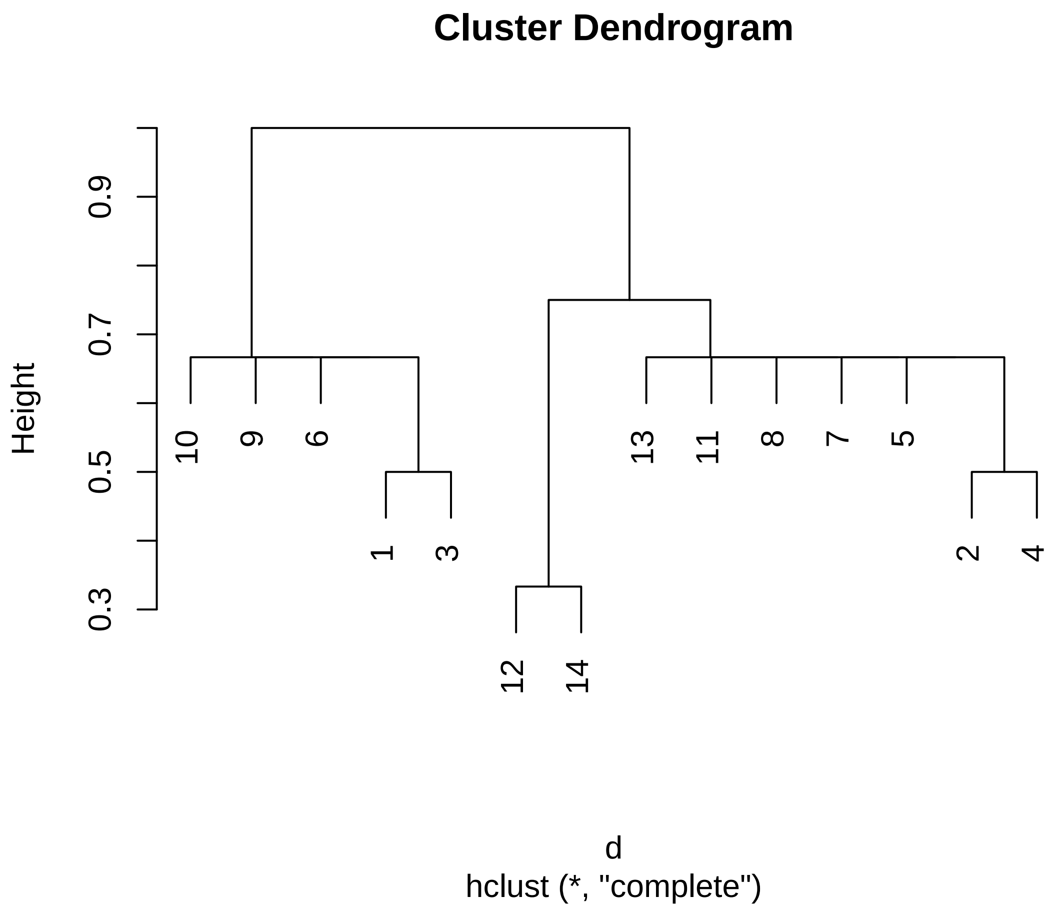
```
##     lhs             rhs        support   confidence coverage  lift     count
## [1] {ACE.2=High} => {Result=0} 0.2372881 0.6086957  0.3898305 1.305929 28
## [2] {ACE.1=High} => {Result=1} 0.2881356 0.6938776  0.4152542 1.299644 34
## [3] {UFE.1=High} => {Result=0} 0.2627119 0.5535714  0.4745763 1.187662 31
## [4] {ACE.2=Low}  => {Result=1} 0.2457627 0.6170213  0.3983051 1.155691 29
## [5] {UFE.1=Low}  => {Result=1} 0.3220339 0.6129032  0.5254237 1.147977 38
```

Cluster the results :

```
##      lhs                     rhs        support   confidence coverage
## [1]  {}                   => {Result=0} 0.4661017 0.4661017  1.0000000
## [2]  {}                   => {Result=1} 0.5338983 0.5338983  1.0000000
## [3]  {ACE.2=High}         => {Result=0} 0.2372881 0.6086957  0.3898305
## [4]  {ACE.2=Low}          => {Result=1} 0.2457627 0.6170213  0.3983051
## [5]  {ACE.1=High}         => {Result=1} 0.2881356 0.6938776  0.4152542
## [6]  {ACE.1=Low}          => {Result=0} 0.2372881 0.5283019  0.4491525
## [7]  {ACE.1=Low}          => {Result=1} 0.2118644 0.4716981  0.4491525
## [8]  {UFE.2=High}         => {Result=1} 0.2796610 0.6000000  0.4661017
## [9]  {UFE.1=High}         => {Result=0} 0.2627119 0.5535714  0.4745763
## [10] {UFE.2=Low}          => {Result=0} 0.2796610 0.5238095  0.5338983
## [11] {UFE.1=High}         => {Result=1} 0.2118644 0.4464286  0.4745763
## [12] {UFE.1=Low}          => {Result=1} 0.3220339 0.6129032  0.5254237
## [13] {UFE.2=Low}          => {Result=1} 0.2542373 0.4761905  0.5338983
## [14] {UFE.1=Low,UFE.2=Low} => {Result=1} 0.2033898 0.5454545  0.3728814
##      lift      count
## [1]  1.0000000 55
## [2]  1.0000000 63
## [3]  1.3059289 28
```

```
## [4]  1.1556906 29
## [5]  1.2996437 34
## [6]  1.1334477 28
## [7]  0.8834981 25
## [8]  1.1238095 33
## [9]  1.1876623 31
## [10] 1.1238095 33
## [11] 0.8361678 25
## [12] 1.1479775 38
## [13] 0.8919123 30
## [14] 1.0216450 24
```

```
##  [1] 1 2 1 2 2 1 2 2 1 1 2 2 2 2
```



This clustering regroups Player-1 winner together very well.

The CLIQUE algorithm

The ENCLUS algorithm

ENtropy-based CLUStering

Frequent pattern-based classification

Classification based on Association

CBA Algorithm

Implementation the CBA algorithm with the M1 or M2 pruning strategy introduced by Liu, et al. (1998).

Candidate classification association rules (CARs) are mined with the standard APRIORI algorithm. Rules are ranked by confidence, support and size. Then either the M1 or M2 algorithm are used to perform database coverage pruning and to determin the number of rules to use and the default class.

TO DO DEFINITION

Example on tennis data

Recall from Homework 1 With Random Forest, the accuracy rate was 0.6931818.
With Logistic regression it was 0.7667.

From classification to associations rules

```
##     lhs                     rhs        support confidence coverage lift count size coveredTransactions
## [1] {ACE.1=[3.5, Inf],
##      ACE.2=[-Inf,1.5),
##      UFE.2=[8.5, Inf]} => {Result=1}   0.125      0.917    0.136 1.61    11    4                  12
## [2] {ACE.1=[3.5, Inf],
##      ACE.2=[-Inf,1.5)} => {Result=1}   0.159      0.875    0.182 1.54    14    3                   4
## [3] {ACE.1=[3.5, Inf],
##      UFE.2=[8.5, Inf]} => {Result=1}   0.216      0.760    0.284 1.34    19    3                  13
## [4] {UFE.1=[7.5, Inf],
##      ACE.2=[-Inf,1.5),
##      UFE.2=[8.5, Inf]} => {Result=1}   0.227      0.690    0.330 1.21    20    4                  18
## [5] {}                  => {Result=0}   0.432      0.432       NA 1.00    88    1                  41
```

```
##                      true
## classifier.prediction  0  1
##                     0 14  5
##                     1  3  8
```

The accuracy rate is :

```
## [1] 0.7333333
```

The sensitivity is the percentage of true output giving Player1-winner among the population of true Player1-winner :

```
## [1] 0.6153846
```

The specificity is the percentage of true output giving Player2-winner (= Player1-looser) among the population of true Player2-winner:

```
## [1] 0.8235294
```

The precision is the percentage of true output giving Player1-winner among all the outputs giving Player1-winner (even if not winner) :

```
## [1] 0.7272727
```

So the F_Mesure is :

```
## [1] 0.6666667
```

From associations rules to classification

```
##     items        transactionID
## [1] {Result=0,
##      ACE.1=Low,
##      UFE.1=Low,
##      ACE.2=Low,
##      UFE.2=Low}             1
## [2] {Result=0,
##      ACE.1=None,
##      UFE.1=Low,
##      ACE.2=High,
##      UFE.2=Low}             2
## [3] {Result=1,
##      ACE.1=Low,
##      UFE.1=Low,
##      ACE.2=Low,
##      UFE.2=High}            3
## [4] {Result=1,
##      ACE.1=High,
##      UFE.1=High,
##      ACE.2=None,
##      UFE.2=High}            4
```

```
##
## Mining CARs...
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.6    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen target  ext
##       5  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 8
##
## set item appearances ...[12 item(s)] done [0.00s].
## set transactions ...[12 item(s), 88 transaction(s)] done [0.00s].
## sorting and recoding items ... [12 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [23 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
##
## Pruning CARs...
## CARs left: 8
```

```
##
## classifier.prediction  0  1
##                     0 11  3
##                     1  6 10
```

The accuracy rate is :

```
## [1] 0.7
```

The sensitivity is the percentage of true output giving Player1-winner among the population of true Player1-winner :

```
## [1] 0.7692308
```

The specificity is the percentage of true output giving Player2-winner (= Player1-looser) among the population of true Player2-winner:

```
## [1] 0.6470588
```

The precision is the percentage of true output giving Player1-winner among all the outputs giving Player1-winner (even if not winner) :

```
## [1] 0.625
```

So the F_Mesure is :

```
## [1] 0.6896552
```

Classification based on Multiple Association Rules

Classification based on Predictive Association Rules

Evaluation

Compare the algorithms