

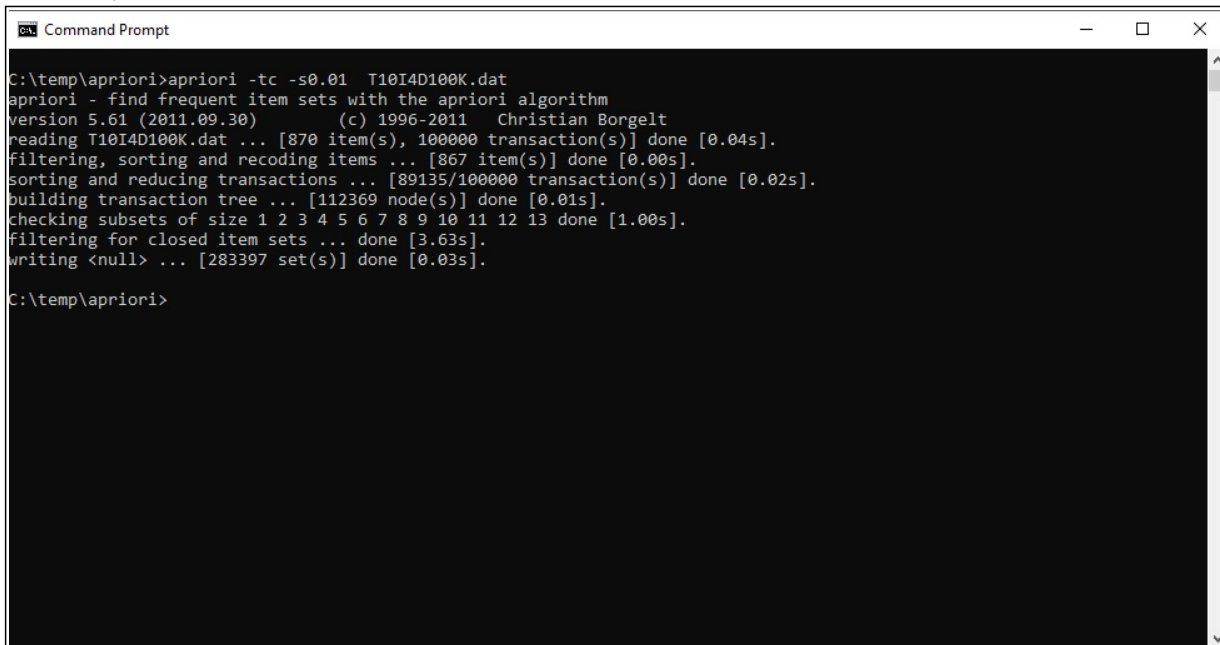
Lab 5: Association Rules

Submit Assignment

Due 25 Oct by 23:59 **Points** 0.4 **Submitting** a file upload **File types** pdf
Available 13 Oct at 13:30 - 25 Oct at 23:59 12 days

Software and Data

- Download the [apriori implementation](#) for Windows
- Download the [apriori implementation](#) for Mac
- It is advisable to read the [documentation](#) first.
- For the newest apriori version please visit: <http://www.borgelt.net/apriori.html> [.\(http://www.borgelt.net/apriori.html\)](http://www.borgelt.net/apriori.html)
- **Note** that you need to run the apriori implementation in command line:



```
C:\temp\apriori>apriori -tc -s0.01 T10I4D100K.dat
apriori - find frequent item sets with the apriori algorithm
version 5.61 (2011.09.30)      (c) 1996-2011  Christian Borgelt
reading T10I4D100K.dat ... [870 item(s), 100000 transaction(s)] done [0.04s].
filtering, sorting and recoding items ... [867 item(s)] done [0.00s].
sorting and reducing transactions ... [89135/100000 transaction(s)] done [0.02s].
building transaction tree ... [112369 node(s)] done [0.01s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 11 12 13 done [1.00s].
filtering for closed item sets ... done [3.63s].
writing <null> ... [283397 set(s)] done [0.03s].




C:\temp\apriori>
```

. To check quickly

all the apriori options just type apriori in command line.

```
C:\temp\apriori>apriori
usage: apriori [options] infile [outfile [appfile]]
find frequent item sets with the apriori algorithm
version 5.61 (2011.09.30)      (c) 1996-2011  Christian Borgelt
-t#      target type                (default: s)
        (s: frequent, c: closed, m: maximal item sets,
        g: generators, r: association rules)
-m#      minimum number of items per set/rule  (default: 1)
-n#      maximum number of items per set/rule  (default: no limit)
-s#      minimum support of a set/rule         (default: 10%)
-S#      maximum support of a set/rule         (default: 100%)
        (positive: percentage, negative: absolute number)
-o       use original rule support definition (body & head)
-c#      minimum confidence of a rule          (default: 80%)
-e#      additional evaluation measure         (default: none)
-a#      aggregation mode for evaluation measure (default: none)
-d#      threshold for add. evaluation measure (default: 10%)
-i#      least improvement of evaluation measure (default: no limit)
        (not applicable with evaluation averaging, i.e. option -aa)
-z       zero evaluation below expected support (default: evaluate all)
-p#      (min. size for) pruning with evaluation (default: no pruning)
        (< 0: weak forward, > 0 strong forward, = 0: backward pruning)
-q#      sort items w.r.t. their frequency     (default: 2)
        (1: ascending, -1: descending, 0: do not sort,
        2: ascending, -2: descending w.r.t. transaction size sum)
-u#      filter unused items from transactions (default: 0.01)
        (0: do not filter items w.r.t. usage in sets,
        <0: fraction of removed items for filtering,
        >0: take execution times ratio into account)
```

- Download the following data sets:

- [mushroom](#) 
- [T10I4D100K](#) 
- [Teams dataset](#) 

Problem 1:

The purpose of this problem is to make you familiar the Apriori algorithm. You will use it on real and synthetic data sets to find frequent itemsets.

1. Run the Apriori algorithm to generate all frequent itemsets from the 'T10I4D100K' data set at a support thresholds of 0.01%, 0.02% and 0.03%, and report the number of frequent itemsets so produced. Use the `-ts` option with Apriori to generate frequent itemsets. Compare the performance of the algorithm in terms of the time taken to produce the results at these thresholds, and comment on the possible reason(s) for this difference in performance. You may estimate the amount of time spent by adding up the time displayed by the program when it is executed. Include all the times displayed by the program.
2. Run Apriori (using the `-ts` option) on the 'mushroom' data set to generate frequent itemsets of sizes 2 through 15 at support thresholds of 5%, 10% and 20%. Generate three plots, one for each threshold, showing the number of frequent sets obtained of size 2 through size 15. (To get the distribution of the number of frequent itemsets use option `-Z`). Comment on the general trends illustrated by the plots, and comment on the reason(s) for these trends. Also comment on how the plots vary between the three thresholds.
3. Use the Apriori algorithm to generate closed (using the `-tc` option) and maximal (using the `-tm` option) frequent itemsets from the 'T10I4D100K' and 'mushroom' data sets. Use a support threshold of 5% for the 'mushroom' data set and 0.01% for the 'T10I4D100K' data set. Compare the total number of closed and maximal frequent itemsets obtained for each dataset individually. How do these numbers compare with the number of frequent itemsets obtained from these data sets using the same threshold? What relationship among closed, maximal and frequent itemsets is revealed by this comparison?

Problem 2:

This question uses the data set 'Teams' that is based on a team A's performance against two teams B and C. The data set is in the file Teams.dat. The data set contains the following items:

- 1: Team A plays against Team B,

2: Team A plays against Team C,

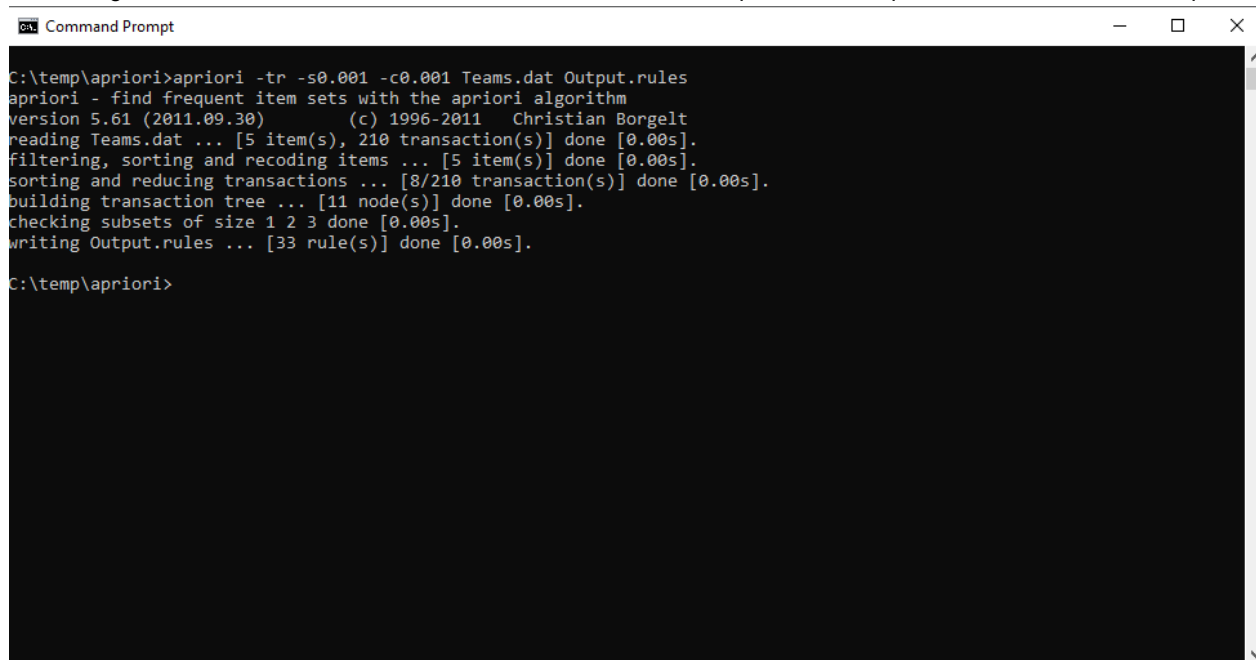
3: Game is played at Team A's home,

4: Game is played away from Team A's home,

5: Game is won by Team A

1. Find the probability of Team A winning against Team B ($\text{Pr}[\text{Won_By_A}|\text{Team_B}]$). Also find the probability of Team A winning against Team C ($\text{Pr}[\text{Won_By_A}|\text{Team_C}]$). Compare these probabilities.
2. Compare the probabilities of Team A winning against Team B and C at a home venue ($\text{Pr}[\text{Won_by_A}|\text{Team_B,Home}]$ and $\text{Pr}[\text{Won_by_A}|\text{Team_C,Home}]$ respectively). State the rules that you based your comparison upon.
3. Compare the probabilities of Team A winning against Team B and C when the games are played away from A's home ($\text{Pr}[\text{Won_by_A}|\text{Team_B,Away}]$ and $\text{Pr}[\text{Won_by_A}|\text{Team_C,Away}]$ respectively). State the rules that you based your comparison upon.
4. Are the results in (b) and (c) consistent with those in (a)? Explain why or why not.

Note: To get the association rules derived from the data, we need to provide an output file name when we run apriori.



```
C:\temp\apriori>apriori -tr -s0.001 -c0.001 Teams.dat Output.rules
apriori - find frequent item sets with the apriori algorithm
version 5.61 (2011.09.30)      (c) 1996-2011  Christian Borgelt
reading Teams.dat ... [5 item(s), 210 transaction(s)] done [0.00s].
filtering, sorting and recoding items ... [5 item(s)] done [0.00s].
sorting and reducing transactions ... [8/210 transaction(s)] done [0.00s].
building transaction tree ... [11 node(s)] done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing Output.rules ... [33 rule(s)] done [0.00s].

C:\temp\apriori>
```

Report: Write a report addressing the above questions.