Association Rules Questions

Objectives

Explore the decision rule generation with a relatively simple, and clean, dataset.

Experiment with the different support and confidence thresholds of the association rule algorithm to identify a set of rules.

Experiment with the Apriori algorithm.

Q1. Consider the data set

Customer ID	Transaction ID	Items Bought
1	0001	*{a, d, e}*
1	0024	*{a, b, c, e}*
2	0012	*{a, b, d, e}*
2	0031	*{a, c, d, e}*
3	0015	*{b, c, e}*
3	0022	*{b, d, e}*
4	0029	*{c, d}*
4	0040	*{a, b, c}*
5	0033	*{a, d, e}*
5	0038	*{a, b, e}*

a) Compute the support for itemsets *{a}*, *{b, c}*, and *{a,b,e}* by treating each transaction ID as a market basket. You have 10 transactions

b) Use the results in part (a) to compute the confidence for the association rules *{b, c}* → *{a}* and *{a}* → *{b, c}*.

c) Repeat part (a) by treating each customer ID as a market basket. You have 5 customers. Eachitem should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

d) Use the results in part (c) to compute the confidence for the association rules *{b, c}* → *{a}* and *{a}* → *{b, c}*.

Q2. Consider the following set of frequent 3-itemsets:

*{*1, 2, 3*}*, *{*1, 2, 4*}*, *{*1, 2, 5*}*, *{*1, 3, 4*}*, *{*1, 3, 5*}*, *{1, 3, 6}*, *{1, 5, 6}*, *{*2, 3, 4*}*, *{*2, 3, 5*}*, *{2,3,6}*, *{*3, 4, 5*}*, *{3, 5, 6}*.
Assume that there are only six items in the data set.

a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_{k-1}$ merging strategy.

b) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

Q3. Consider the following dataset:

Weather Condition	Driver's Condition	Traffic Violation	Seat Belt	Crash Severity
Good	Alcohol-impaired	Exceed speed limit	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Disobey stop sign	Yes	Minor
Good	Sober	Exceed speed limit	Yes	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Disobey stop sign	Yes	Minor
Bad	Alcohol-impaired	None	Yes	Major
Good	Sober	Disobey traffic signal	Yes	Major
Good	Alcohol-impaired	None	No	Major
Bad	Sober	Disobey traffic signal	No	Major

a) Show a binarized version of the data set.

b) What is the maximum width of each transaction now that it is binarized?

c) Assuming that support threshold is 30%, how many candidate and frequent itemsets will begenerated?

d) Create a data set that contains only the following asymmetric binary attributes: (Weather = Bad,Driver's condition = Alcohol-impaired, Traffic violation = Yes, Seat Belt = No, Crash Severity = Major). For Traffic violation, only None has a value of 0. The rest of the attribute values are assigned to 1.

e) Assuming that support threshold is 30%, how many candidate and frequent itemsets will begenerated?

Q4. Consider the following dataset:

TID	Temperature	Pressure	Alarm 1	Alarm 2	Alarm 3
1	95	1105	0	0	1
2	85	1040	1	1	0
3	103	1090	1	1	1
4	97	1084	1	0	0
5	80	1038	0	1	1
6	100	1080	1	1	0
7	83	1025	1	0	1
8	86	1030	1	0	0
9	101	1100	1	1	1

a) Partition the range of the temperature into 3 bins of equal sized range. Show those ranges.

b) If the support threshold is 30%, which ranges from (a) have the support?

c) If you partition the range such that each bin has the same number of transations, show thoseranges.

d) if the support threshold is 30%, which ranges in (c) have the support?