REI502M - Introduction to Data Mining

Solutions to homework 5

Elías Snorrason October 15, 2019

Problem 1

The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

Transaction ID	Items Bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

A. Draw a contingency table for each of the following rules using the transactions shown in the table above. Rules: $\{b\} \rightarrow \{c\}$**,** $\{a\} \rightarrow \{d\}$ **,** $\{b\} \rightarrow \{d\}$ **,** $\{e\} \rightarrow \{c\}$**,** $\{c\} \rightarrow \{a\}$**.**

	c	$\bar{c}$	
b	**3**	**4**	7
$\bar{b}$	**2**	**1**	3
	5	5	10

	d	$\bar{d}$	
a	**4**	**1**	5
$\bar{a}$	**5**	**0**	5
	9	1	10

	d	$\bar{d}$	
b	**6**	**1**	7
$\bar{b}$	**3**	**0**	3
	9	1	10

	c	$\bar{c}$	
e	**2**	**4**	6
$\bar{e}$	**3**	**1**	4
	5	5	10

	a	$\bar{a}$	
c	**2**	**3**	5
$\bar{c}$	**3**	**2**	5
	5	5	10

B. Use the contingency tables in part A to compute and rank the rules in decreasing order according to the following measures.

- Support
- Confidence
- Lift
- Leverage
- Conviction
- Chi-square

Table 1: Rule measures with rankings per measure

Rule	Support	Confidence	Lift	Leverage	Conviction	χ^2
$b \to c$	0.3 (3)	0.429 (3)	0.875 (3)	-0.050 (2)	0.857 (1)	0.4762 (3)
$a \to d$	0.4 (2)	0.800 (2)	0.889 (2)	-0.050 (2)	0.500 (5)	1.111 (2)
$b \to d$	0.6 (1)	0.857 (1)	0.952 (1)	-0.030 (1)	0.700 (4)	0.4762 (3)
$e \to c$	0.2 (4)	0.333 (5)	0.667 (5)	-0.100 (3)	0.750 (3)	1.667 (1)
$c \to a$	0.2 (4)	0.400 (4)	0.800 (4)	-0.050 (2)	0.833 (2)	0.4000 (4)

C. Given the ranking you had obtained above, compute the correlation between the ranking of confidence and the other five measures. Which measure is most highly correlated with confidence? Which measure is least correlated with confidence?

The correlation between two measure rankings X, Y is their covariance devided by their respective standard deviations:

$$\rho_{XY} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

This is done quickly with the function `cor()` in the `Statistics` package in Julia:

```
using Statistics

conf_ranks = [3; 2; 1; 5; 4]

supp_ranks = [3; 2; 1; 4; 4]
lift_ranks = [3; 2; 1; 5; 4]
leve_ranks = [2; 2; 1; 3; 2]
conv_ranks = [1; 5; 4; 3; 2]
chi2_ranks = [3; 2; 3; 1; 4]


cor(supp_ranks, conf_ranks) # 0.970...
cor(lift_ranks, conf_ranks) # 1.00
cor(leve_ranks, conf_ranks) # 0.894...
cor(conv_ranks, conf_ranks) # -0.500
cor(chi2_ranks, conf_ranks) # -0.277...
```

From this, we see that **lift is most highly correlated with confidence** (since it has the same formula for the rule $\{A\} \to \{B\}$, just scaled by $\frac{1}{P(B)}$). Thus, we get the same rankings with confidence and lift.
We see that **conviction is least correlated with confidence**, as it is the inverted lift of the rule $\{A\} \to \{\bar{B}\}$.

Problem 2 (5.12 in textbook)

Given the lattice structure shown in Figure 5.33 and the transactions given in Table 5.22, label each node with the following letter(s):

- M if the node is a maximal frequent itemset,
- C if it is a closed frequent itemset,
- N if it is frequent but neither maximal nor closed, and
- I if it is infrequent.

Assume that the support threshold is equal to 30%.

Table 5.22. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

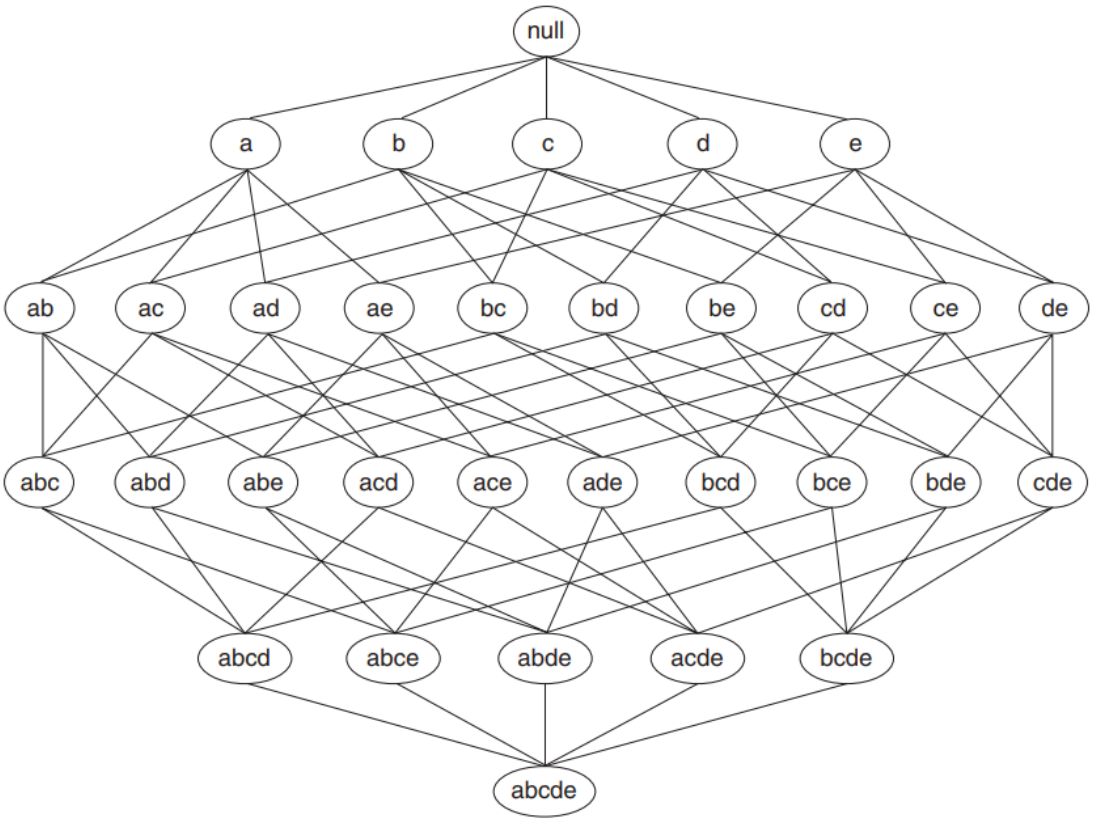

Figure 5.33. An itemset lattice

In the table below are the all the frequent itemsets, with first items in lexicographic order and then sorted by increasing support:

Itemset	Support
$\{a\|b\}$	**3**
$\{a\|d\|e\}$	**4**
$\{a\|d\}$	4
$\{a\|e\}$	4
$\{a\}$	**5**
$\{b\|c\}$	**3**
$\{b\|d\|e\}$	**4**
$\{b\|d\}$	**6**
$\{b\|e\}$	4
$\{b\}$	**7**
$\{c\|d\}$	**4**
$\{c\}$	**5**
$\{d\|e\}$	**6**
$\{d\}$	**9**
$\{e\}$	6
$\{\}$	**10**

Closed itemsets are in bold. From the table we can immediately see that only two 2-itemsets have support equal to the minimum support ($\{a|b\}$ and $\{b|c\}$). These itemsets happen to be maximally frequent (as well as the frequent 3-itemsets and the itemset $\{c|d\}$ with support 4). From the list, we can see that only 4 itemsets are labeled with F.

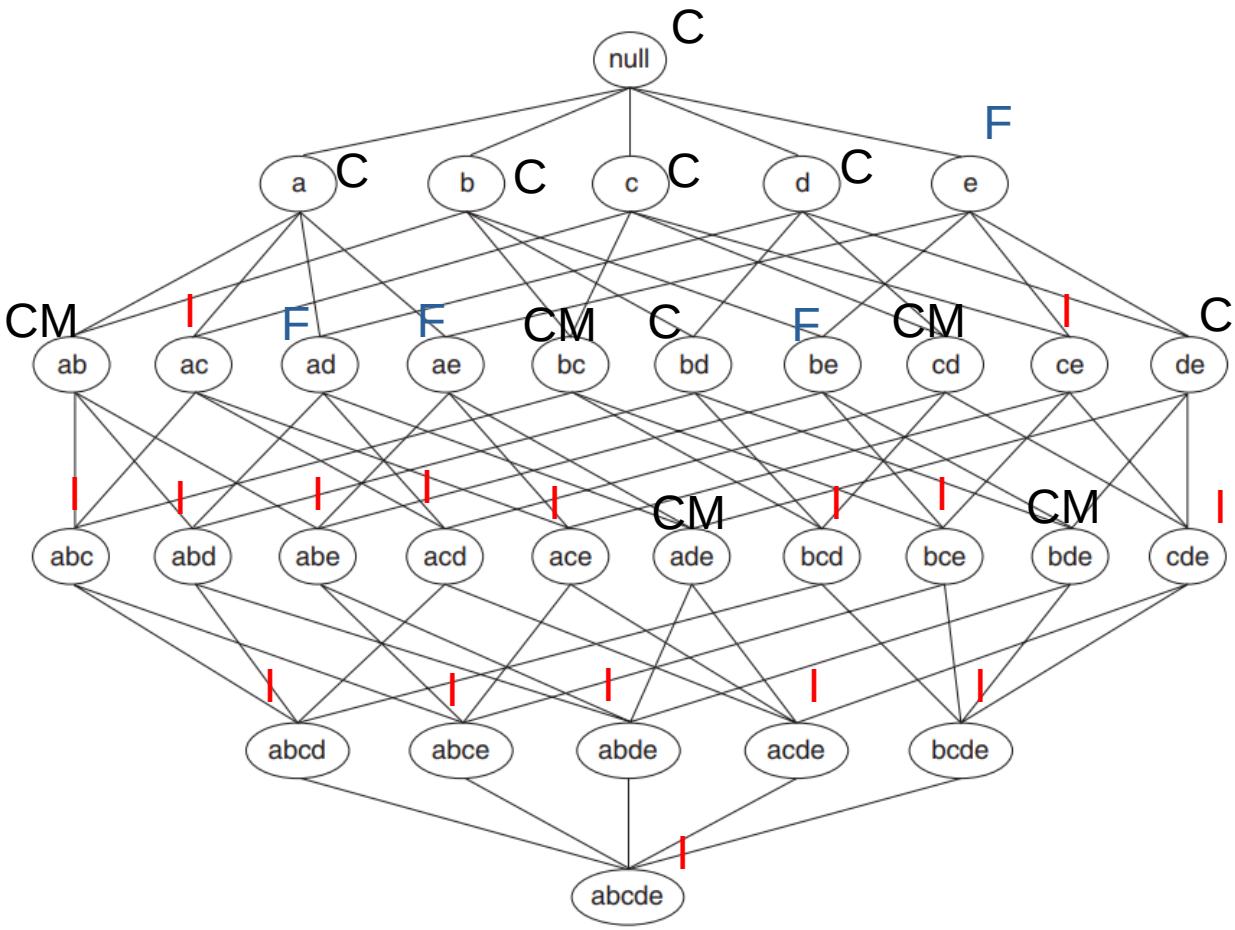

Figure 5.33. An itemset lattice

Problem 3 (5.15 in textbook)

Answer the following questions using the data sets shown in Figure 5.34. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the *Apriori* algorithm to extract frequent itemsets with $minsup = 10\%$ (i.e., itemsets must be contained in at least 1000 transactions).

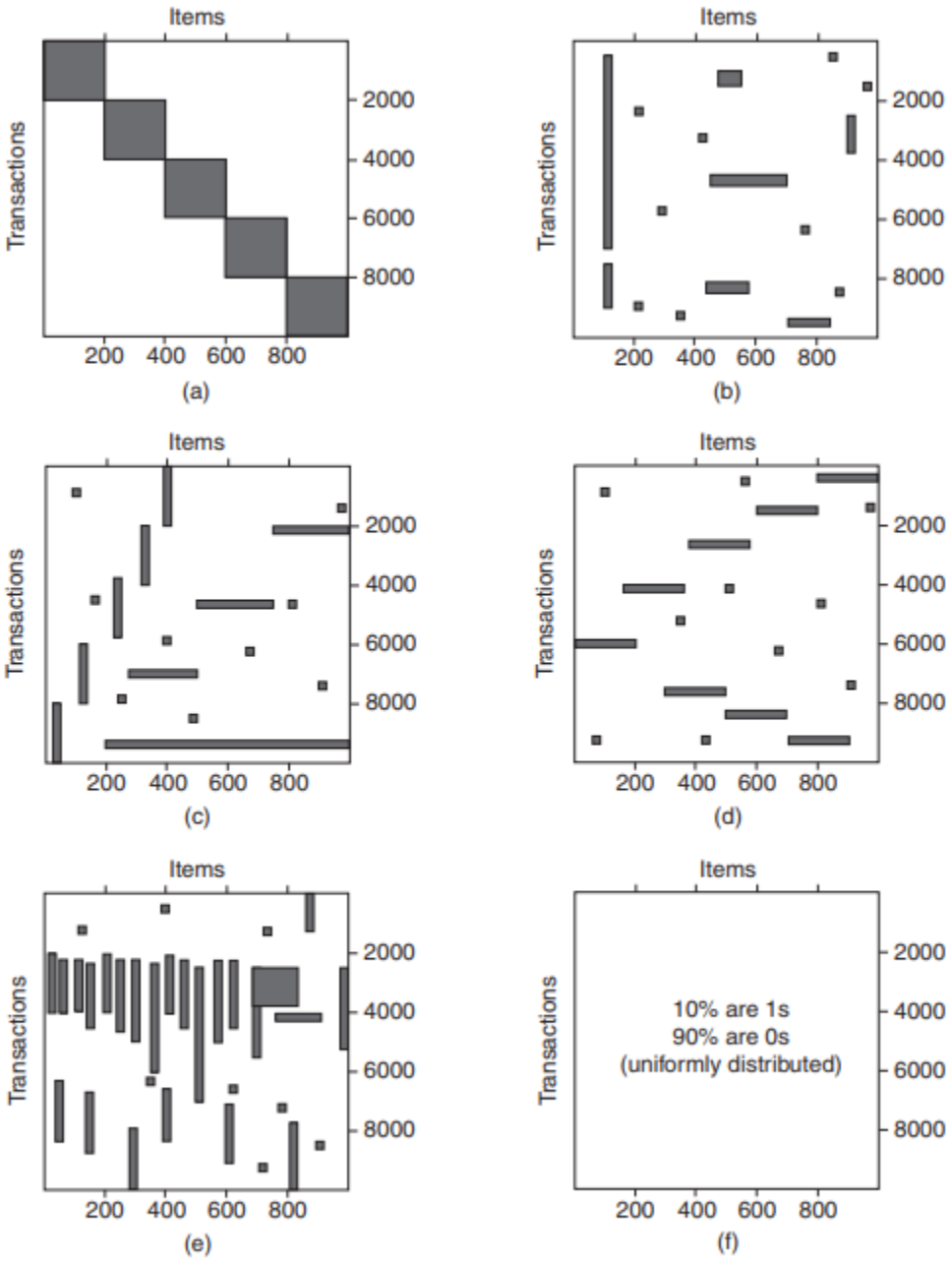

Figure 5.34. Figures for Exercise 15.

A. Which data set(s) will produce the most number of frequent itemsets?

Frequent itemsets are best represented by frequently occurring items (dense columns). Overlapping columns indicate a higher number of possible frequent itemsets. Data set e) will produce most number of frequent itemsets.

B. Which data set(s) will produce the fewest number of frequent itemsets?

Data set d) with no frequent itemsets, as it's columns are sparse and they don't fulfill the minimum support.

C. Which data set(s) will produce the longest frequent itemset?

Data set e) has multiple frequent 1-itemsets overlapping, indicating that it will produce the longest frequent itemset.

D. Which data set(s) will produce frequent itemsets with highest maximum support?

In dataset b), item number around 100 occurs roughly 8000 times, which is the highest support of the data sets.

E. Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%)?

Data set e) has frequent item columns with varying lengths.