University of Southern Denmark
IMADA
Ricardo Campello
(Shortened Version of Arthur Zimek's Originals for DM870)

DM583: Data Mining

Exercise 7: Frequent Itemsets and Association Rules

Exercise 7-1 Combinatoric Explosion

(a) A database contains transactions over the following items: "apples", "bananas", "cherries", "dates", "eggplants", "figs", and "guavas". How many different combinations of these items can exist (i.e., how many different transactions could possibly occur in the database)? How does this number grow by increasing the number of items? **Note**: We do not distinguish whether a transaction contains a fruit once or several times, e.g., if someone bought one apple or several apples would just result in the transaction to contain "apples".

(b) How many transactions with exactly two items (i.e., 2-itemsets) can we have when the database contains 3 items? When it contains 5 items? How many *k*-itemsets do we have when the database contains *n* items?

Exercise 7-2 Itemsets and Association Rules

Given a set of transactions T according to the following table:

Set of transactions T

Transaction ID	items in basket
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies }
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What are the support and the confidence of {Milk} ⇒ {Diapers}?

(b) What are the support and the confidence of {Diapers} ⇒ {Milk}?

(c) What is the maximum number of association rules that can be extracted from this dataset?

(d) What is the maximum size of frequent itemsets that can be extracted (assuming support $\sigma > 0$)?

(e) Find an itemset (of size 2 or larger) that has the largest support.

(f) Find a pair of items, a and b, such that the rules $\{a\} \Rightarrow \{b\}$ and $\{b\} \Rightarrow \{a\}$ have the same confidence.

Exercise 7-3 Apriori Candidate Generation

Given the complete collection of frequent 3-itemsets in a transaction database:

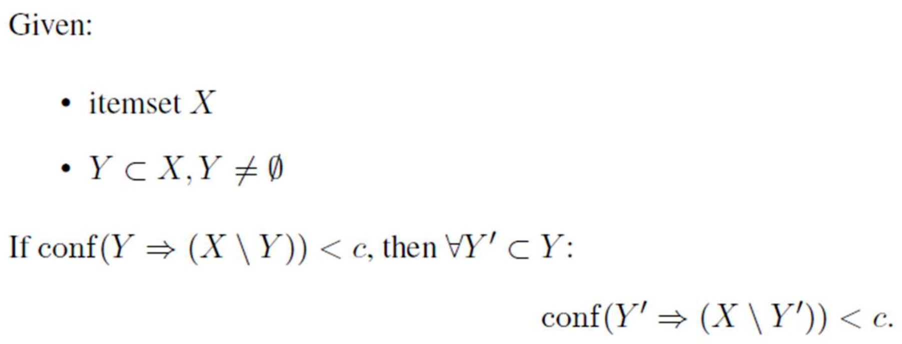
$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}$$

(a) List all candidate 4-itemsets following the Apriori joining and pruning procedure.

(b) Discuss the case $\{1, 2, 4, 5\}$ – we see two reasons why it cannot be frequent, but if we go through systematically, we should only find the first one and stop searching. Which should be the first one to find?

Exercise 7-4 The Monotonicity of Confidence

Theorem 2.1 in the Lecture states:



Given:

- itemset X
- $Y \subset X, Y \neq \emptyset$

If $\mathrm{conf}(Y \Rightarrow (X \setminus Y)) < c$, then $\forall Y' \subset Y$:

$$\mathrm{conf}(Y' \Rightarrow (X \setminus Y')) < c.$$

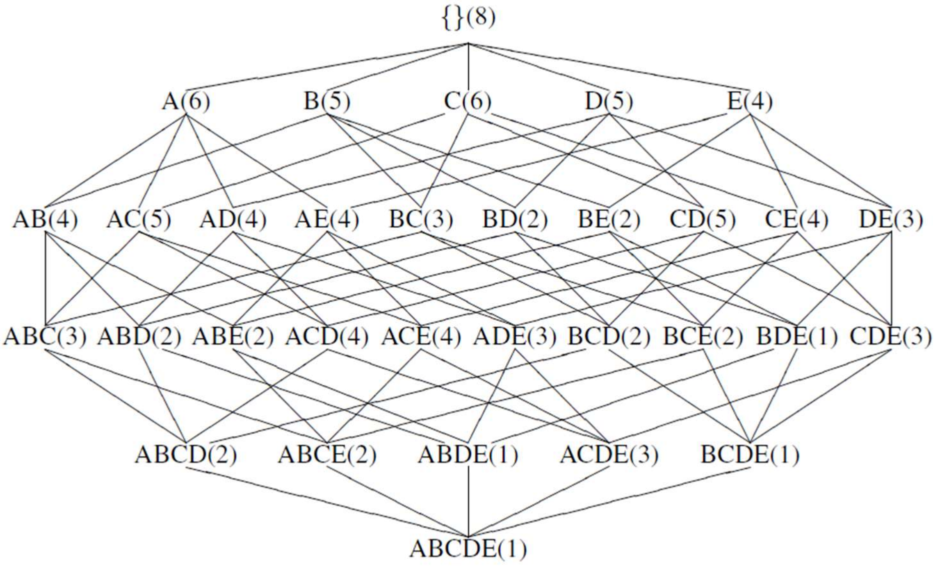
(a) Prove the theorem.

(b) Sketch an algorithm (pseudo code) that generates all association rules with support σ or above and a minimum confidence of c, provided the set F of all frequent itemsets (w.r.t. σ) with their support, efficiently using the pruning power of the given theorem.

Exercise 7-5 Support Based on Closed Frequent Itemsets

(a) The database from the lecture grew by one transaction. We computed the corresponding support of all itemsets in the lattice:

TID	A	B	C	D	E
1	0	1	0	0	0
2	1	0	1	1	1
3	1	1	1	0	1
4	0	0	1	1	0
5	1	1	1	1	1
6	1	0	1	1	1
7	1	1	0	0	0
8	1	1	1	1	0


{}(8)
A(6) B(5) C(6) D(5) E(4)
AB(4) AC(5) AD(4) AE(4) BC(3) BD(2) BE(2) CD(5) CE(4) DE(3)
ABC(3) ABD(2) ABE(2) ACD(4) ACE(4) ADE(3) BCD(2) BCE(2) BDE(1) CDE(3)
ABCD(2) ABCE(2) ABDE(1) ACDE(3) BCDE(1)
ABCDE(1)

Identify the closed frequent itemsets for the support thresholds $\sigma = 4$ and $\sigma = 2$, respectively. What do you observe?

(b) Sketch an algorithm (pseudo code) to find the support for all frequent itemsets, using only the set of closed frequent itemsets as information.

Show in the lattice, how the algorithm works.

Exercise 7-6 Apriori

Consider the following transaction database D over the items $I = \{A, B, C, D, E, F\}$.

TransID	Items
1	A B E
2	B D
3	C D F
4	A B D
5	A C E
6	B C E F
7	A C E
8	A B C E
9	A B C D F
10	B C D E

Given the support threshold $\sigma = 2$, apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Please explain in the solution all the steps that you followed.

In particular, include for each level the candidate set (C_k) (i) after the join step before pruning and (ii) after pruning. Annotate for those objects pruned in (ii) the explicit reason for pruning them.

Also give explicitly the solution of frequent k-itemsets (S_k) for each k.