02450 Introduction to Machine Learning and Data Mining

Week 12: Association mining

Bjørn Sand Jensen

29 April 2025

DTU Compute, Technical University of Denmark

Today

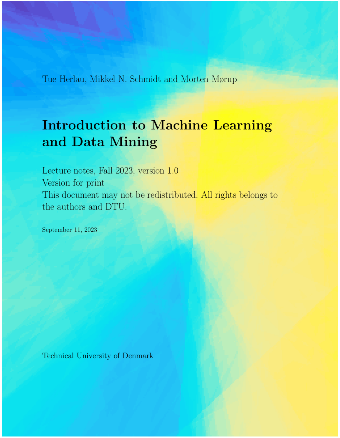
Feedback Groups of the day:

Christian Torp, João Pires de Azevedo Teixeira dos Prazeres, Qiao Chih Lei, Frederik Springer Krehan, Andreas Abildgaard Ryberg, Noah Samuelsen Schütze, Valdemar Sejersen, Ning Sun, Agathe Michot, Cecilia Haslund Oreskov, Dimitrios Papantzikos Moutafis, Konstantinos Tzimoulias, Zhetao Qu, Ignacio Ripoll González, Sebastian Timm, Sumaya Mahamed Mahamed, Rebekka Theilgaard Rosenkilde, Clara Ella Lassen, Aleksander Holm Lauridsen, Frede Søndergaard Møllegaard, Oscar Christen Thorning Engelstoft, Anastasia Maftei, Bogdan-Petru Pascut, Tomasz Gabriel Stepien, Mirka Katuscáková, Konstantin Takors, Nicolai Jess Møllnitz, Andrea Raithel, Zofia Agata Lenarczyk, Eric Torres García, Sarah Abigail Tauro, Marcus Prehn-Chang, Lars Henrik Dæhli Skjolding, Matias Haage Them Nielsen, Amanda Schriver Mårtensson, Anne Reimer, Madeline Shah, Stamatia Papageorgiou, Vasiliki Tsanaktsidou, Edvin Smajlovic, Zahra Soleimani, Leon Maximilian Spohn, Oscar Stilling, Victor Stubgaard, Peter Tang, Christian Holst Thomsen, Magnus Thorsteinsson

Reading/homework material:

Chapter 21

P21.1, P18.2, P18.3



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

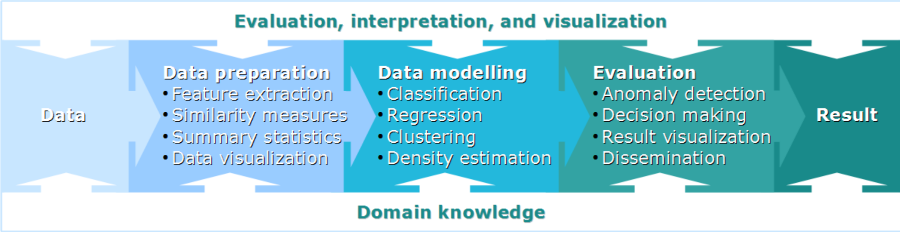
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. **Association mining**
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Calculate support and confidence of association rules
- Describe the Apriori algorithm for association mining and how it is used for efficient estimation of association rules

Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- **Goal:** Produce dependency rules
  - Predict the occurence of an item based on occurences of other items

Association Mining

mining association rules

About 1.850.000 results (**0,05** sec)

[PDF] Fast algorithms for **mining association rules**
R Agrawal, R Srikant - Proc. 20th int. conf. very large data bases, VLDB, 1994 - it.uu.se
We consider the problem of discovering **association rules** between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally di erent from the known algorithms. Experiments with synthetic as well as real …
Cited by 26110 Related articles All 115 versions

Mining association rules between sets of items in large databases
R Agrawal, T Imieliński, A Swami - Proceedings of the 1993 ACM …, 1993 - dl.acm.org
We are given a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. We present an efficient algorithm that generates all significant **association rules** between items in the database. The algorithm incorporates …
Cited by 23230 Related articles All 39 versions

An effective hash-based algorithm for **mining association rules**
JS Park, MS Chen, PS Yu - Acm sigmod record, 1995 - dl.acm.org
In this paper, we examine the issue of **mining association rules** among items in a large database of sales transactions. The **mining** of **association rules** can be mapped into the problem of discovering large itemsets where a large itemset is a group of items which …
Cited by 2465 Related articles All 18 versions

Source: Google Scholar (November, 2020)

Association rule discovery: Example

Market basket analysis

Training set

1. {Bread, Soda, Milk}
2. {Beer, Bread}
3. {Beer, Soda, Diaper, Milk}
4. {Beer, Bread, Diaper, Milk}
5. {Soda, Diaper, Milk}

Rules discovered

{Milk} ►{Soda}
{Diaper, Milk} ►{Beer}

Market basket data

- Representation as:

Transaction table

ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Data matrix

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Association analysis, rules and support

- **Itemset**
  - For example: {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset $\boldsymbol{X}$
  - Percentage of transactions that contain $\boldsymbol{X}$
- **Association rule**
  - Expression of the form: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$
    where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disjoint sets
- **Support** for an association rule $\boldsymbol{X} \rightarrow \boldsymbol{Y}$
  - Percentage of transactions that contain $\boldsymbol{X} \cup \boldsymbol{Y}$

    $sup(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\sigma(\boldsymbol{X} \cup \boldsymbol{Y})}{N} = P(\boldsymbol{X}, \boldsymbol{Y})$

Quiz 1: Support (Spring 2018)

	x_1^L	x_1^H	x_2^L	x_2^H	x_3^L	x_3^H	x_4^L	x_4^H	x_5^L	x_5^H	x_6^L	x_6^H
O1	1	0	1	0	1	0	1	0	1	0	1	0
O2	0	1	0	1	0	1	0	1	0	1	0	1
O3	1	0	0	1	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1	0	1	1	0
O5	0	1	1	0	1	0	1	0	1	0	1	0
O6	0	1	0	1	0	1	0	1	0	1	0	1
O7	0	1	1	0	1	0	0	1	0	1	0	1
O8	1	0	1	0	1	0	1	0	0	1	0	1
O9	0	1	0	1	1	0	1	0	0	1	0	1
O10	1	0	0	1	0	1	0	1	0	1	1	0

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute x_1–x_6.

We consider a dataset of airline safety binarized according to the median value. Values below median is referred to with the superscript L and above the median value using the superscript H. In Table 1 is given the first 10 observations O1–O10. Consider the association rule:

$$\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}.$$

What is the support of the rule?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

Association analysis, confidence

- **Itemset**
  - For example: {Bread, Soda, Milk}, {Milk, Diaper}, {}
- **Support** for an itemset $\boldsymbol{X}$
  - Percentage of transactions that contain $\boldsymbol{X}$
- **Association rule**
  - Expression of the form: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$
    where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disjoint sets
- **Support** for an association rule $\boldsymbol{X} \rightarrow \boldsymbol{Y}$
  - Percentage of transactions that contain $\boldsymbol{X} \cup \boldsymbol{Y}$

    $sup(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\sigma(\boldsymbol{X} \cup \boldsymbol{Y})}{N} = P(\boldsymbol{X}, \boldsymbol{Y})$
- **Confidence** for an association rule $\boldsymbol{X} \rightarrow \boldsymbol{Y}$
  - Percentage of transactions containing $\boldsymbol{X}$ that also contain $\boldsymbol{Y}$

    $conf(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{P(\boldsymbol{X}, \boldsymbol{Y})}{P(\boldsymbol{X})} = P(\boldsymbol{Y} \mid \boldsymbol{X})$

Quiz 2: Confidence (Spring 2018)

	x_1^L	x_1^H	x_2^L	x_2^H	x_3^L	x_3^H	x_4^L	x_4^H	x_5^L	x_5^H	x_6^L	x_6^H
O1	1	0	1	0	1	0	1	0	1	0	1	0
O2	0	1	0	1	0	1	0	1	0	1	0	1
O3	1	0	0	1	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1	0	1	1	0
O5	0	1	1	0	1	0	1	0	1	0	1	0
O6	0	1	0	1	0	1	0	1	0	1	0	1
O7	0	1	1	0	1	0	0	1	0	1	0	1
O8	1	0	1	0	1	0	1	0	0	1	0	1
O9	0	1	0	1	1	0	1	0	0	1	0	1
O10	1	0	0	1	0	1	0	1	0	1	1	0

Table 1: The ten first observations of the airline safety dataset binarized considering the attribute x_1–x_6.

We again consider the airline safety data and the rule

$$\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}.$$

What is the confidence of the rule?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

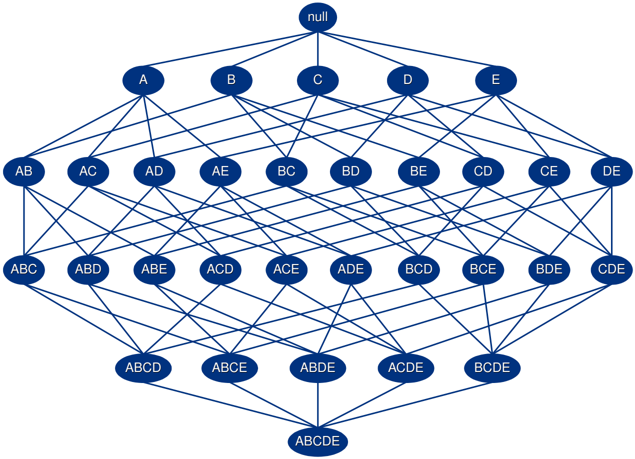
E. Don't know.

Association rule mining

- Find all association rules that have
  - **Support** $\geq$ minsupport
  - **Confidence** $\geq$ minconf
- Approach:
  - **Frequent itemset generation**
    - Generate a list of all **itemsets** with
  - **Association rule generation**
    - Generate all **association rules** with **Confidence** $\geq$ minconf

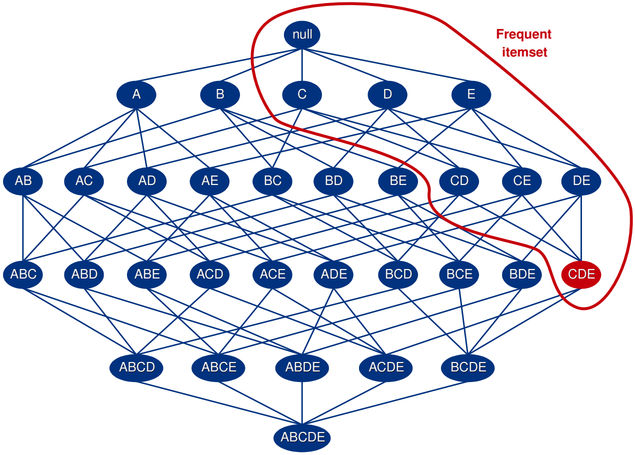
Frequent itemset generation

How many different itemsets can be created for a problem with a total of D items?



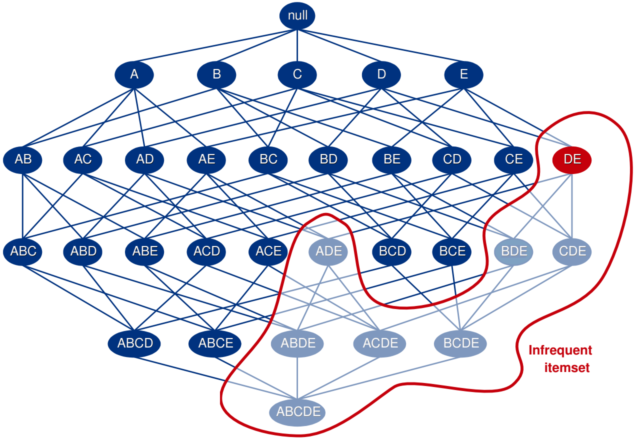
Frequent itemset generation

If an itemset is frequent, then all of its subsets must also be frequent



Frequent itemset generation

If an itemset is infrequent, then all of its supersets must also be infrequent



The Apriori Algorithm

Find all 1-itemsets

Generate k-itemsets by merging single items to the k-1-itemsets

Remove all the generated itemsets for which subsets are not part of the k-1-itemsets

Keep remaining k-itemsets with enough support.

Output all frequent itemsets

Algorithm 8: Apriori algorithm

Given N transactions and let $\epsilon > 0$ be the minimum support count
$L_1 = \{\{j\} | \text{supp}(\{j\}) \geq \epsilon\}$
for $k = 2, \ldots, M$ and $L_k \neq \emptyset$ **do**
    $C'_k = \{s \cup \{j\} | s \in L_{k-1}, j \notin s\}$
    Set $C_k = C'_k$
    for each $c \in C'_k$ **do**
        for each $s \subset c$ such that $|s| = k - 1$ **do**
            if s is not frequent, i.e. $s \notin L_{k-1}$ **then**
                $C_k = C_k \setminus \{c\}$ (Remove c from C_k)
            end if
        end for
    end for
    $L_k = \{c | c \in C_k, \text{supp}(c) \geq \epsilon\}$ (compute support)
end for
$L_1 \cup L_2 \cup \cdots \cup L_k$ are then all frequent itemsets

Quiz 3: A-priori (Fall 2018)

We will consider a binary dataset consisting of the $M = 6$ features $f_1,\ f_2,\ f_3,\ f_4,\ f_5,\ f_6$. We wish to apply the Apriori algorithm to find all itemsets with support greater than $\varepsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

$$L_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Recall the key step in the Apriori algorithm is to construct L_3 by first considering a large number of candidate itemsets C_3', and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose L_2 is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

A. $\{f_2, f_3, f_4\}$

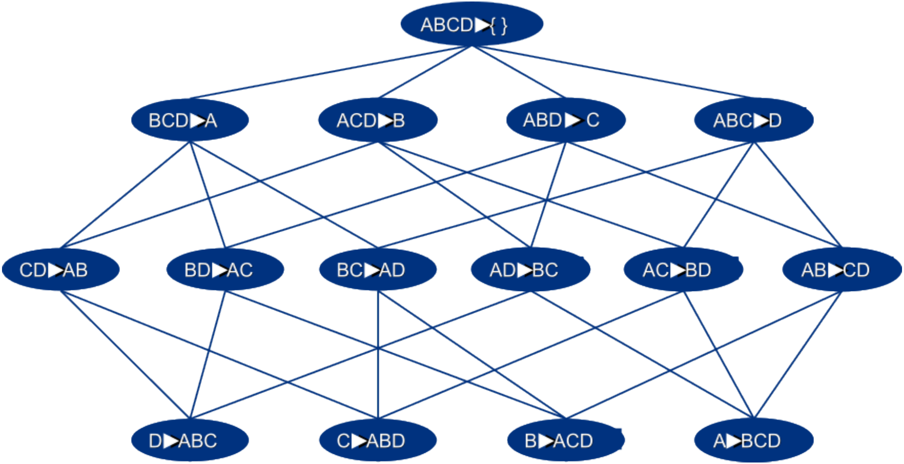
B. $\{f_1, f_2, f_6\}$

C. $\{f_2, f_3, f_6\}$

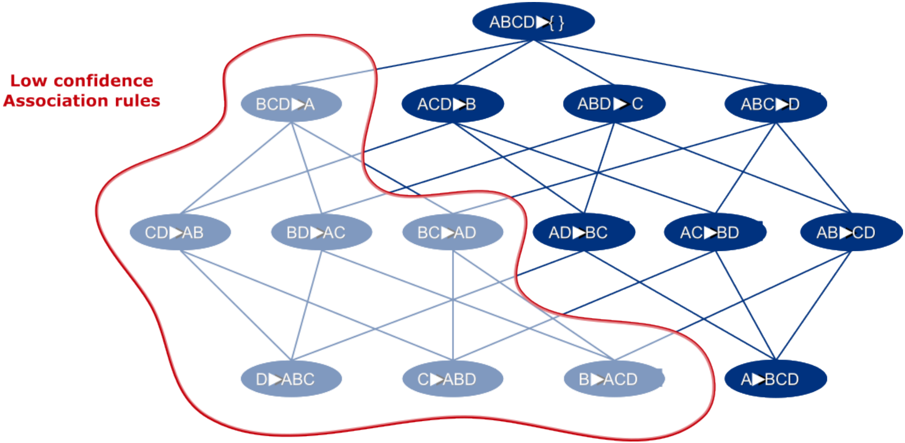
D. $\{f_1, f_3, f_4\}$

E. Don't know.

Association rule generation



Association rule generation



Results for market basket example

Itemset	Support
Milk	80%
Bread	60%
Soda	60%
Beer	60%
Diaper	60%
Diaper Milk	60%
Soda Milk	60%
Bread Beer	40%
Bread Milk	40%
Soda Diaper	40%
Beer Diaper	40%
Beer Milk	40%
Soda Diaper Milk	40%
Beer Diaper Milk	40%

Association rule	Support	Confidence
{} ►Milk	80%	80%
Soda ►Milk	60%	100%
Diaper ►Milk	60%	100%
Soda, Diaper ►Milk	40%	100%
Beer, Diaper ►Milk	40%	100%
Beer, Milk ►Diaper	40%	100%

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

- How can we do association mining for continuous data?

$\mathbf{X}=$

Attribute 1	Attribute 2	Attribute 3
0.3689	0.9827	0.6999
0.4607	0.7302	0.6385
0.9816	0.3439	0.0336
0.1564	0.5841	0.0688
0.8555	0.1078	0.3196
0.6448	0.9063	0.5309
0.3763	0.8797	0.6544
0.1909	0.8178	0.4076
0.4283	0.2607	0.8200
0.4820	0.5944	0.7184
0.1206	0.0225	0.9686
0.5895	0.4253	0.5313
0.2262	0.3127	0.3251
0.3846	0.1615	0.1056
0.5830	0.1788	0.6110
0.2518	0.4229	0.7788
0.2904	0.0942	0.4235
0.6171	0.5985	0.0908
0.2653	0.4709	0.2665
0.8244	0.6959	0.1537

Binarize data according to percentiles

AttributeNames=

X=

Attribute 1	Attribute 2	Attribute 3
0.3689	0.9827	0.6999
0.4607	0.7302	0.6385
0.9816	0.3439	0.0336
0.1564	0.5841	0.0688
0.8555	0.1078	0.3196
0.6448	0.9063	0.5309
0.3763	0.8797	0.6544
0.1909	0.8178	0.4076
0.4283	0.2607	0.8200
0.4820	0.5944	0.7184
0.1206	0.0225	0.9686
0.5895	0.4253	0.5313
0.2262	0.3127	0.3251
0.3846	0.1615	0.1056
0.5830	0.1788	0.6110
0.2518	0.4229	0.7788
0.2904	0.0942	0.4235
0.6171	0.5985	0.0908
0.2653	0.4709	0.2665
0.8244	0.6959	0.1537

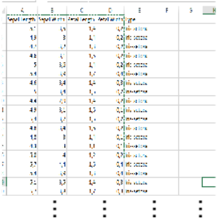
AttributeNamesBin=

Xbinary=

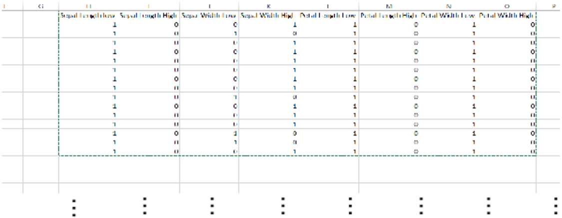
Attribute 1 0-50 %	Attribute 1 50-100 %	Attribute 2 0-33.3 %	Attribute 2 33.3-66.7%	Attribute 2 66.7-100 %	Attribute 3 0-50%	Attribute 3 50-100%
1	0	0	0	1	0	1
0	1	0	0	1	0	1
0	1	0	1	0	1	0
1	0	0	1	0	1	0
0	1	1	0	0	1	0
0	1	0	0	1	0	1
1	0	0	0	1	0	1
1	0	0	0	1	1	0
0	1	1	0	0	0	1
0	1	0	1	0	0	1
1	0	1	0	0	0	1
0	1	0	1	0	0	1
1	0	1	0	0	1	0
1	0	1	0	0	1	0
0	1	1	0	0	0	1
1	0	0	1	0	0	1
1	0	1	0	0	1	0
0	1	0	1	0	1	0
1	0	0	1	0	1	0
0	1	0	0	1	1	0

Recap of association rule discovery on Iris data

X=



Xbinary=



Example of association rules where y is also appended to Xbinary, i.e. [Xbinary Y], i.e. Y has three columns indicating which of the three flowers the observation belongs to.

{Petal length Low, Petal Width low, Iris Setosa} -> {Sepal Length Low}
(Conf 100, sup 33)

{Sepal Length Low, Petal Width low, Iris Setosa} -> {Petal length Low}
(Conf 100, sup 33)

{Sepal length Low, Petal length Low, Iris Setosa} -> {Petal Width Low}
(Conf 100, sup 33)

{Sepal length Low, Sepal Width High, Petal length Low, Petal Width Low} -> {Iris Setosa}
(Conf 100, sup 28)

Quiz 4: A-priori (Bonus)

Consider the dataset below consisting of 10 transactions Find all itemsets with support greater than 35How many are there?

A: 7, B: 9, C: 11, D:13, E: Don't know

	Juice	Milk	Beer	Cheese	Chocolate	Yoghurt	Sugar	Flour	Egg	Wine
Customer 1	0	0	1	0	0	0	1	0	1	0
Customer 2	1	1	0	0	0	1	0	1	1	1
Customer 3	0	1	0	1	1	0	0	0	0	1
Customer 4	1	1	0	0	0	1	0	0	0	1
Customer 5	1	0	1	0	0	0	0	0	1	0
Customer 6	1	0	0	0	0	0	0	0	1	0
Customer 7	1	1	0	0	1	1	0	0	0	1
Customer 8	0	1	0	1	0	0	1	1	0	1
Customer 9	1	1	0	0	1	1	0	0	0	0
Customer 10	0	0	1	0	0	1	0	0	1	1

Practicals

- Post-test on DTU Learn: Quizzes > Post test
  - Already open, closes Sunday at midnight
  - Similar to pre-test
  - We will present the results next week
- Mock exam from Fall 2024 is availble from the 29th April 2025 after the lecture (also no aids)
  - **Optional**: pick up physical answer sheet and hand in next week (for us to test the setup)
- Next week (and beyond)
  - Recap of the course
  - Assistance via Piazza (limited TA availability, so help each other out)

Resources

Fast Algorithms for Mining Association Rules Proc. 20th Int. Conf. Very Large Data Bases, 1994, Rakesh Agrawal and Ramakrishnan Srikan.