Tutorial Business Analytics

Tutorial 11: Association Rules

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial Business Analytics

Outline

Today's topics:

- Association Rule Mining
  - MBA
  - Support
  - Confidence
  - Lift

- Apriori algorithm

Tutorial Business Analytics

Association Rule Mining – "Discover correlation among different attributes"

Discover strong rules which describe the correlation among different attributes.

$$\{A, E\} \Rightarrow B$$

Unsupervised Learning

Market Basket Analysis (MBA)

- boost the sales (up-/cross-selling)
- store layout
- loyalty card / cashback

Other Application

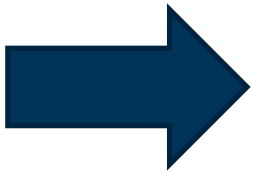
- Intrusion detection
- Bioinformatics

Tutorial Business Analytics

Association Rule Mining

Market basket transactions

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



Binary representation

ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Tutorial Business Analytics

Association Rule Mining – Support

The support of an *item set* is its relative frequency.

ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Example:

- Itemset {Milk, Bread} == {Bread, Milk}
- **supp**({Milk, Bread}) = 2/5 = 0.4
- i.e. the combination Milk & Bread appears in 40% of all transactions.

The support of a rule is the support of all item sets it contains: Let R: {A, B} ⇒ {C, D}; that means **supp**(R) = **supp**({A, B, C, D}).

Example:

supp({Milk ⇒ Bread}) = **supp**({Bread ⇒ Milk})
= **supp**({Milk, Bread})
= 0.4

Tutorial Business Analytics

Association Rule Mining – Confidence

The confidence of a rule (R: X ⇒ Y) is its *likeliness* to apply to the data set.

ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Definition:

- **conf**(R) = **supp**(X ∪ Y) / **supp**(X)

Example:

- **conf**({Milk, Bread} ⇒ {Butter}) = **supp**({Milk, Bread, Butter}) / **supp**({Milk, Bread})
  = 0.2 / 0.4 = 0.5
- i.e. the probability that the rule applies to the given data set is 50%

Association rules with minimum support and confidence are sometimes called "**strong**" rules.

Tutorial Business Analytics

Association Rule Mining – Lift

The lift of a rule (R: X ⇒ Y) indicates by how much the confidence of a rule surpasses the expected value. (Relation of observed confidence and expected confidence.)

ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Definition of lift is

lift(R) = **conf**(R) / **expConf**(R)

= (**supp**(X ∪ Y) / **supp**(X)) / **supp**(Y)

= **supp**(X ∪ Y) / (**supp**(X) * **supp**(Y))

Example:

- **lift**({Milk, Bread} ⇒ {Butter}) = 0.2 / (0.4 * 0.4)
  = 1.25
- i.e. the observed confidence is 25% higher than expected

lift > 1: Item sets X and Y appear more frequent than expected. X has a positive effect on Y. The greater the lift, the higher the probability that the rule is not coincidence.

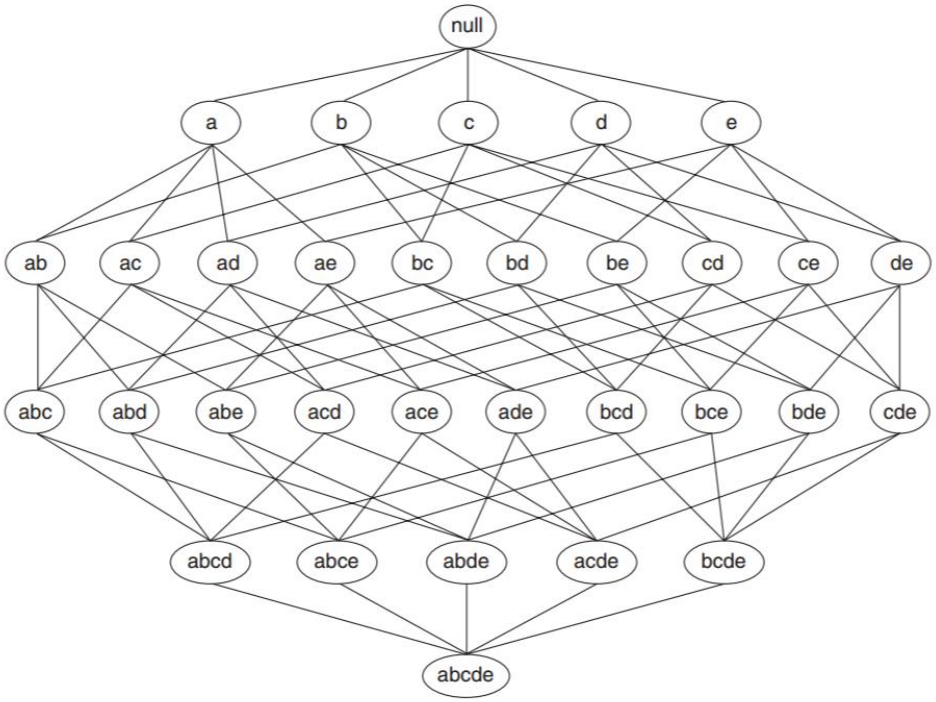
lift ≈ 1: X, Y are independent.

lift < 1: Item sets X and Y appear less frequent than expected. X has a negative effect on Y.

Tutorial Business Analytics

Association Rule Mining - Will brute-force approach work?

Compute support and confidence for all possible rules on items {A, B, C, D, E}



How many rules on d items?

$$3^d - 2^{d+1} + 1$$

#items	2	3	4	5	6	7	8
#rules	2	12	50	180	602	1932	6050

- In the left example of 5 items
  → 180 possible rules
- In our market example of 4 items
  → 50 possible rules

Source: https://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Tutorial Business Analytics

Apriori algorithm

- Iteratively find all frequent i-item sets (**item sets with cardinality i** and **minSup**)
  - First evaluate every **1-item set** (i = 1) on the **minSup**
  - for i > 1
    - evaluate all **i-item sets** which originate from single (i-1)-item sets items on the **minSup**
    - i++
    - Stop as soon as there are no more i-item sets which meet **minSup**

- Generate rules from the found item sets
  - Find rules that meet **minConf** and contain only 1 item on the right (side of the ⇒)
  - Evaluate possible rules that contain several items on the right and are made up of simpler rules
    - e.g. if you found X ⇒ Y and X ⇒ Z, evaluate X ⇒ Y, Z

Tutorial Business Analytics

Association Rule – Exercise 11.1

Regarding a data set about taste in music with 1000 entries below-mentioned association rule has a support of 0.4 and a confidence of 0.8. Answer each of the following questions with an interval as small as possible. ([−∞,+∞] or a single value are valid options.)

$$\{beatles, stones\} \Rightarrow \{dylan, cohen\}$$

- How many people like beatles and stones?
- How many people like stones and dylan?
- What is the support of the rule {beatles, dylan, stones} ⇒ {cohen}?
- What is the lift of the above-mentioned rule (BS ⇒ DC)? Interpret your result.

Tutorial Business Analytics

Apriori algorithm – Exercise 11.2

Have a look the following items {Wine, Noodles, Tomato sauce, Diapers} and transactions and find all item sets that meet min. support = 0.4. Construct all possible rules that meet the min. confidence = 0.8.

Customer	Wine	Noodles	Tomato sauce	Diapers
1	1	1	1	0
2	1	0	0	1
3	0	1	1	1
4	1	1	1	1
5	0	1	1	0
6	1	1	0	1
7	0	0	0	1
8	1	1	1	1
9	0	0	1	1
10	1	1	1	0

Tutorial Business Analytics

Singular Value Decomposition – Exercise 11.3

Gregory registered to an online music platform, where he can stream his favorite songs. The platform uses Singular Value Decomposition (SVD) to predict ratings for songs, based on previous ratings of a user. The following 3 tables are the result of that SVD.

U	Dim1	Dim2
User 1	0.35	0.00
Gregory (G)	0.53	0.84
User 3	0.84	-0.23
User 4	0.20	0.25
…	…	…

Table 1

S	Dim1	Dim2
Dim 1	5.22	0
Dim 2	0	3.12

Table 2

V	Shape of You (S)	One Dance (D)	Closer (C)	Despacito (D)	Faded (F)
Dim 1	0.76	0.54	0.63	0.13	-0.13
Dim 2	0.84	0.45	0.83	-0.32	0.34

Table 3

Tasks

- Predict the ratin of Gregory (G) for the song Faded (F) considering that Gregory‘s average rating is $\bar{r}_G = 2$.
- How do you interpret the values of User 1 (U1) in table 1 and the values of Despacito in table 3?

Tutorial Business Analytics

Summary and outlook

- Association Rule Mining
  - MBA
  - Support
  - Confidence
  - Lift

- Apriori algorithm
- Singular Value Decomposition