Tutorial Business Analytics

Tutorial 11

Exercise 11.1

Regarding a data set about taste in music with 1000 entries the association rule

R: beatles, stones → dylan, cohen

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.
(remark: [−∞,+∞] or a single value are valid options, too)

a) How many people like beatles and stones?

b) How many people like stones and Dylan?

c) What is the support of the rule "beatles, dylan, stones → cohen" ?

d) What is the lift of the above-mentioned rule (BS->DC)? Interpret your result.

Solution

Abbreviations:

B	beatles
S	stones
D	dylan
C	cohen
R	rule

R: B,S → D,C n = 1000

a) How many people like beatles and stones?
conf(B,S → D,C) = supp(B,S,D,C) / supp(B,S)
=> supp(B,S) = supp(B,S,D,C) / conf(B,S → D,C) = 0.4 / 0.8 = 0.5
→ 500 people (n * supp(B,S) = 500)

b) How many people like stones and Dylan?
n * supp(S,D) = unknown => estimation:
supp(S,D) ≥ supp(B,S,D,C) = 0.4 => supp(S,D) ∈ [0.4, 1.0]
→ [400, 1000] people

Occurrence in general is O.K thus single elements must have higher probability or equal

c) What is the support of the rule "beatles, dylan, stones → cohen" ?
The support of a rule depends on the constraints involved. Whether the constraints are on the left or on the right side of the arrow does not affect the support.
The support of (B,S → D,C) is equal to the support of (B,D,S → C)
→ supp(B,D,S → C) = 0.4

d) What is the lift of the above-mentioned rule (R)? Interpret your result.
- lift(R) = supp(B,S,D,C) / (supp(B,S) * supp(D,C)) = 0.4 / (0.5 * supp(D,C))
- supp(D,C) is unknown but can be estimated
  - : lower bound: supp(D,C) ≥ supp(B,S,D,C) = 0,4
  - upper bound: supp(B,S) = 0.5 and supp(B,S,D,C) = 0.4
  - → at least 0.1*n = 100 people like the beatles and stones, but not all four interprets. This can only be the case when at least 100 people don't like dylan and cohen.
- plugging in supp(D,C) results in
  - . lift(R) ∈ [0.4 / (0.5 * 0.9), 0.4 / (0.5 * 0.4)] ∈ [0.89, 2.0]
- cases:
  - lift ∈ [0.89, 1) → people who like B,S tend to like D,C less
  - lift = 1 → liking B,S does not affect liking D,C
  - lift ∈ (1,2] → people who like B,S tend to like D,C more

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

Table 1: 10 customer transactions. 1 = bought, 0 = not bought

Solution

Abbreviations: Wine (W), Noodles (N), Tomato sauce (T), Diapers (D)

At first we need to calculate the support of all item sets. In order to do so, we divide the number of transactions for each item by the overall number of transactions.

1-item set	supp
W	0.6
N	0.7
T	0.7
D	0.7

2-item set	supp
W,N	0.5
W,T	0.4
W,D	0.4
N,T	0.6
N,D	0.4
T,D	0.4

3-item set	supp
W,N,T	0.4
W,N,D	0.3
W,T,D	0.2
N,T,D	0.3

Apriori algorithm:

- Start with the smallest item sets and add one item at a time as long as they meet min. support.
- Red item sets don't meet the min supp. and don't need to be further considered.
- Because only one 3-item set meets the min. supp. there can be no 4-item sets that meet the min. support.

The second step is to generate rules based on the frequent item sets.

1 left, 1 right	Conf
W→ N	5/6
N→W	5/7
W→T	4/6
T→W	4/7
W→D	4/6
D→W	4/7
N→T	6/7
T→N	6/7
N→D	4/7
D→N	4/7
T→D	4/7
D→T	4/7

2 left, 1 right	Conf
W,N→T	4/5
W,T→N	4/4
N,T→W	4/6

0 left, 1 right	Conf
_ → W	6/10
_ → N	7/10
_ → T	7/10
_ → D	7/10

Information from T1 implicates that these rules don't need to be considered

Rules with two items on the right side of the arrow are not evaluated, because W→N and W→T do not meet the min. confidence.

Apriori algorithm:

- Start with building rules from the smallest item sets possible (2-item sets in this case) and add one element at a time as long as they meet min. conf.
- I.e. only the black colored rules mentioned in the first paragraph above.
- Mixing these rules results in the 3 rules stated in the second paragraph above.
- Rules which contain 4-item sets cannot fulfill min conf. because they don't fulfill min. supp. in the first place.

Exercise 11.3 Singular Value Decomposition (SVD)

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 (U1)	0.35	0.00
Gregory (G)	0.53	0.84
User 3 (U3)	0.84	-0.23
User 4 (U4)	0.20	0.25
…	…	…

table 1

S	Dim1	Dim2
Dim1	5.22	0
Dim2	0	3.12

table 2

V^T	Shape of You (S)	One Dance (O)	Closer (C)	Despacito (D)	Faded (F)	…
Dim1	0.76	0.54	0.63	0.13	-0.13	…
Dim2	0.84	0.45	0.83	-0.32	0.34	…

table 3

a) Predict the rating of Gregory (G) for the song Faded (F) considering that Gregory's average rating is $\overline{r_G} = 2$.
b) How do you interpret the values of Users 1 (U1) in table 1 and the values of Despacito (D) in table 3?

Solution

a) $r_{G,F} = 2 + (0.53 \quad 0.84) * \begin{pmatrix} 5.22 & 0 \\ 0 & 3.12 \end{pmatrix} * \begin{pmatrix} -0.13 \\ 0.34 \end{pmatrix} =$

   a. $2 + (2.766 \quad 2.621) * \begin{pmatrix} -0.13 \\ 0.34 \end{pmatrix} = 2 + 0.531 = 2.531$

b) User 1 prefers music having higher values for the first latent factor.
   User 1 does not care about the second latent factor.
   The song Despacito is slightly positively represented in the first latent factor.
   Users having only a preference for the first latent factor will prefer Despacito over Faded.
   A User having a preference for the second latent factor will less like Despacito, since it is negatively represented.