Assignment 4: Association Rules

Low, Daniel Mark (S3120155)
Petre, Bogdan (S3480941)
Xu, Teng Andrea (S3548120)

Group 7

September 26, 2017

1 Introduction: Confidence and Support (30p)

1. Question: What's the importance of lift in association rule mining? Why is confidence and/or support not sufficient?
1. Answer: The function lift measure how many times more often the items X and Y occur together than expected if they are are statistically independent of each other. Is a measure of how the two items are really related rather then coincidentally happening together.
The following is the lift formula:

$$Lift(X \rightarrow Y) = \frac{Support(X \cap Y)}{Support(X) * Support(Y)}$$

If the result is equal to 1 then X and Y are statistically independent of each other, otherwise if the result is greater than 1 than there is a strong association between the two items

Support and Confidence are not sufficient because they just one item and that's it (the first)or consider just the antecedent item of the expression X → Y and the concurrence of X and Y(the latter). Furthermore, confidence cannot tell if a rule contains true implication of the relationship or if the rule is pure coincidence.

2. Question: suppose you want to apply association rule mining on a collection of surveys containing personal details such as: name, age, gender, hobbies, favorite color, income, country etc. How would you convert this raw data into a form suitable for association analysis?
2.Answer: First of all the table name gives us nothing, for example if we have lift (name){Marco} → (hobby){Basketball} equal to 2 it does not mean anything, it's just coincidence that in the dataset many transaction have Marco that play basketball but statistically it does mean nothing. Then I will change the continuous data like "age" and "income" in an interval or categorical, for example for age "youth" if the age is below 30 and the wealthy if income(annual) is like 100.000 euros. Then we will have for example lift like (country,job){USA,engineer} → {wealthy} equal to 2.5, that does make sense.

3. Question: Suppose we already know that the lift of the rule wine => cheese is 2. We also know the that the support of wine is 0.1. What can we say about the support for cheese?
3.Answer: Support(cheese) it has to be 5 times Support(wine ∩ cheese) because from the previous lift formula we obtain:

$$Support(cheese) = \frac{Support(wine \cap cheese)}{Support(wine) * Lift(wine \rightarrow cheese)}$$

That also makes sense because if we have 10% of all transaction contain wine then the value of Support(wine ∩ cheese) has to be lower or at most Support(wine) (case limit when in all occurrence of wine we have also cheese).

References