NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Answers to Association Rules

BC2407 ANALYTICS II SESSION 3

BASED ON CHEW C.H. TEXTBOOK AAD VOLUME 2.

3 Key Concepts in Association Rules X → Y

- Define the following in terms of Probability:
  - Supp (X) ≡ P(contains X)
  - Supp (X and Y) [aka Rule Support] ≡ P(contains X and Y)
  - Conf (X → Y) ≡ P(contains Y|contains X)
  - Lift (X → Y) ≡ $\frac{P(contains\ Y|contains\ X)}{P(contains\ Y)}$

Using the keyword "contains" avoids the awkward definition of P(Y | X) = P(X intersect Y)/P(X). In probability, X and Y are events but in Association Rules, X and Y are itemsets that typically has no items in common and thus no intersection. X and Y in itemsets actually mean items in X together with items in Y, and hence is a union of two itemsets and not the intersection.

A Simple Numerical Example (Answers)

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Assoc Rule: {milk, bread} → Butter

- Supp ({milk}) = 2/5
- Supp ({milk, bread}) = 2/5
- Supp ({milk, bread, butter}) = 1/5 = 0.2
- Confidence of the rule = ½ = 0.5
- Lift of the rule = ½ / 2/5 = 1.25

Answers to Questions for Class Activity 1

1. Explain in your own words, the meaning of the concept "Confidence", and why is this measure useful?

Association Rule, X -> Y is actually a probabilistic concept as only a certain percentage of transactions will contain items in Y, if those transactions already contain items in X. Not every transaction has both X and Y. To precisely capture this probabilistic concept, we use the condition probability P(Y | X). This measures the probability that Y occurs, given that X already occurred, and can be interpreted as the confidence of Y occurring, in the presence of X. It is a measure of the "strength" or "predictability" of the rule.

2. Explain in your own words, the meaning of the concept "Support", and why is this measure useful?

Support is the proportion of occurrence of the items and is a measure of prevalence or popularity of the items. It is a measure of the "applicability" of the rule. It tells you how often you will be able to apply the association rule, whereas confidence tells you the strength of the rule in situations where it is applicable. Thus, they complements each other.

Answers to Questions for Class Activity 1

3. Explain in your own words, the meaning of the concept "Lift", and why is this measure useful? Is Lift still necessary if we have a rule that has high confidence and high support?

Association Rule: X -> Y

Lift measures how useful the rule is, in the context of the existing situation. Even if a rule has high confidence and high support, it may not be useful if P(Y) is already high. Example: Conf(X→Y) = 99% but P(Y) = 99.99%

i.e. The additional information about X has no/minimal impact on Y if P(Y) is already very high i.e. If Y is already popular, why do you need X to boost Y?

If Lift = 1, then X has no impact on Y. [X and Y are independent.]
If Lift > 1, then X can help to boost Y.
If Lift < 1, then X will reduce Y.

Answers to Questions for Class Activity 1

4. Many application of association rules require both Confidence and Support. Explain in your own words how both measures can be used to define good association rules.

Support and Confidence (and Lift) reveals different information and thus, complements each other in understanding the situation. Support measures the applicability of the rule, whereas Confidence measures the strength of the rule in situations where it is applicable.

The real business question is how high is high enough for support and confidence. This depends on the business. If it is a high volume, low profit margin business (e.g. supermarket) then it needs rules with higher support levels, compared to low volume, high profit margin business (e.g. Diamond stores).

Answers to Questions for Class Activity 1

5. The Apriori algorithm is the standard method for generating association rules. (a) Explain in your own words, in a few sentences, how it works. (b) Is this sufficient to compute Confidence? Explain.

(a) It computes the support of all 1-itemsets that meets the minimum support criteria, then use it to compute the support of all 2-itemsets, and so on. By exploiting the fact that adding more items will either reduce the support or stay constant, but never increase support, it smartly reduces the amount of items to consider. The benefit is speed and scale, especially if there are many products to consider. It can assemble different product combinations (that are likely to result in good rules) and calculate their supports in order generate a lot of "good" rules, for a lot of products, very fast.
(b) It is sufficient as confidence can be expressed as the support of m items divided by the support of n items, and Apriori already has all these supports computed.

Answers to Questions for Class Activity 1

6. Is Confidence or Lift a symmetric concept? Explain. Implications?

Confidence is a conditional probability and hence non-symmetrical.

Lift is symmetrical (and hence cannot distinguish directional recommendations per se):

$$Lift(X \rightarrow Y) \equiv \frac{P(Y|X)}{P(Y)} \equiv \frac{P(X\ and\ Y)}{P(X) \times P(Y)}$$

$$Lift(Y \rightarrow X) \equiv \frac{P(X|Y)}{P(X)} \equiv \frac{P(X\ and\ Y)}{P(X) \times P(Y)}$$

7. Provide another potential application of association rules beyond groceries and insurance claims.

Many answers. e.g. Medical diagnosis based on symptoms.

Example: Recommend Burger?

Transaction ID	Items Purchased
1	Burger, Fries
2	Burger
3	Burger
4	Burger

- Conf(Burger → Fries) = ¼
- Conf(Fries → Burger) = P(Burger | Fries) = 1/1 = 100%.
- Hence, recommend burger to all who purchased Fries (w/o burger)?
- Ans: No. Lift(Fries → Burger) = P(Burger | Fries) / P(Burger) = 1.
- i.e. Fries did not boost the sales of Burger.

Reflection

- The Supp, Conf and Lift are easy to calculate.
- What do you think is the main problem/difficulty with doing Association Analysis?

The Support, Confidence and Lift are easy to compute **if the rule is given**. The problem is that we are not given all the "good" rules, and need to generate potentially good rules for a list of (many) items, fast. Imagine NTUC store with 10,000 unique items in the store.

Apriori algorithm is a smarter way than brute force to quickly generate "potentially good" rules, fast, even for many items. i.e. speed and scale. It avoids wasting time considering "useless" rules.