# Association Rules

Prof. Sharon McNicholas

STATS 780/CSE 780





## Introduction

- Consider a (large) data set comprising binary variables.
- One example is a transaction data base (or a market basket).
- A market basket analysis can be carried out.
- Or, more generally, an association rule analysis.
- Note that Assignment 1 is now available, and concerns association rules.

# What is an Association Rule?

- Used to discover relationships in transaction databases.
- Transaction databases contain, exclusively, binary variables.
- Although formally introduced by Agrawal *et al*. (1993), many of the ideas behind association rules can be seen in the literature at least as far back as Yule (1903).
- No underlying statistical model is assumed and no hypotheses are formally proposed.





# Definition of an Association Rule

- Given a non-empty set, *I*, an association rule is a statement of the form $A \Rightarrow B$, where $A, B \subset$ *I* such that $A \neq \emptyset, B \neq \emptyset$, and $A \cap B = \emptyset$.
- The set $A$ is called the antecedent of the rule, the set $B$ is called the consequent of the rule, and *I* is called the itemset. Association rules are generated over a large set of transactions, denoted $\tau_1, \ldots, \tau_n$.
- An association rule is deemed interesting if the items involved occur together often and there is evidence to suggest that one of the sets might in some sense lead to the presence of the other set.
- Association rules are commonly characterized by mathematical notions called support, confidence, and lift.

# Functions of Association Rules

- Functions by which associations are traditionally characterised:

$$\text{Support}: s(A \Rightarrow B) = P(A, B).$$

$$\text{Confidence}: c(A \Rightarrow B) = P(B \mid A) = \frac{P(A, B)}{P(A)}.$$

$$\text{Lift}: L(A \Rightarrow B) = \frac{c(A \Rightarrow B)}{P(B)} = \frac{P(B \mid A)}{P(B)} = \frac{P(A, B)}{P(A)P(B)}.$$

- A variety of other functions have also been introduced.





# Mining Rules

- Association rules are commonly generated using the apriori algorithm or a variant thereof (Agrawal & Skirant 1994, Borgelt & Kruse 2002, Bolgelt 2003).
- An implementation is available in the `arules` package in R.
- The algorithm requires a minimum support threshold, a minimum confidence threshold, and maximum rule length.
- Often, this will result in many association rules being generated.
- Pruning is then used to remove rules that are not “interesting”.

# Interestingness

- Confidence is popular as a measure of interestingness.
- Confidence, combined with support, is also popular.
- Lift is another approach that has some intuitive appeal.
- There are also other options, e.g., Gray and Orlowska's interestingness





# Gray and Orlowska's Interestingness

- An example of another function is Gray and Orlowska's interestingness (Gray and Orlowska, 1998).
- Gray and Orlowska's is defined as follows:

$$\mathsf{Int}(A \Rightarrow B; K, M) = \left[\left(\frac{P(A,B)}{P(A)P(B)}\right)^K - 1\right][P(A)P(B)]^M.$$

- Presents a compromise between the distance of lift (to the power of $K$) from one and the respective magnitudes of $P(A)$ and $P(B)$ (to the power of $M$); $K$ and $M$ can be viewed as weights.
- It is symmetric in the sense that $\mathsf{Int}(A \Rightarrow B; K, M) = \mathsf{Int}(B \Rightarrow A; K, M)$ — because lift is symmetric in the sense that $L(A \Rightarrow B) = L(B \Rightarrow A)$.

# Gray and Orlowska's Int. Contd.

- McNicholas (2007) gives an argument for setting $K = M$, based on:

$$\begin{aligned}\mathsf{Int}(A \Rightarrow B; K, M) &= \left[\left(\frac{P(A,B)}{P(A)P(B)}\right)^K - 1\right](P(A).P(B))^M \\ &= \left[\frac{\left(\frac{P(A,B)}{P(A)}\right)^K - P(B)^K}{P(B)^K}\right](P(A).P(B))^M \\ &= \left[P(B \mid A)^K - P(B)^K\right] P(A)^M P(B)^{M-K} \\ &= \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^M P(B)^{M-K}.\end{aligned}$$

- Now, for $K = M$:

$$\mathsf{Int}(A \Rightarrow B; K) = \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^K.$$





# Lift

- Recall that

$$L(A \Rightarrow B) = \frac{P(A,B)}{P(A)P(B)}.$$

- The lift seems appealing as a measure of interestingness.

- An interesting rule has lift "far" from 1, but the lift is not symmetric about 1.

- One solution is to consider $\log L(A \Rightarrow B)$.

- A better solution would be to find the upper and lower bounds of lift.

# Bounds in $P(A)$ and $P(B)$

- The range of values that the lift of an association rule $A \Rightarrow B$ can take is restricted by the respective values of $P(A)$ and $P(B)$;

$$\frac{\max\{P(A)+P(B)-1, 1/n\}}{P(A)P(B)} \leq L(A \Rightarrow B) \leq \frac{1}{\max\{P(A), P(B)\}}, \quad (1)$$

where $n$ is the number of transactions, $\tau_i$.

- These bounds that are almost identical to those derived by Fréchet (1951) and could be used to standardize the lift.

- What if minimum thresholds for support and confidence were used in the mining process?





# Considering Minimum Support

- Suppose the minimum support threshold is $s$.

- We have

$$\frac{4s}{(1+s)^2} \leq L(A \Rightarrow B) \leq \frac{1}{s}.$$

- Setting $s = 1/n$ gives the bound if no support threshold is used;

$$\frac{4n}{(n+1)^2} \leq L(A \Rightarrow B) \leq n.$$

- These quantities, alone, are useless for standardizing lift with a view to ranking association rules.

# Adding Minimum Confidence

- Suppose the minimum confidence threshold is $c$.

- Considering this and the other bounds, we have

$$\max\left\{\frac{P(A)+P(B)-1}{P(A)P(B)}, \frac{4s}{(1+s)^2}, \frac{s}{P(A)P(B)}, \frac{c}{P(B)}\right\} \leq L(A \Rightarrow B) \leq \frac{1}{\max\{P(A), P(B)\}}. \tag{2}$$

- McNicholas et al. (2008) use $(2)$ to standardize the lift.





# Standardized Lift

- Denote:

$$\upsilon = \frac{1}{\max\{P(A), P(B)\}},$$

$$\lambda = \max\left\{\frac{P(A)+P(B)-1}{P(A)P(B)}, \frac{4s}{(1+s)^2}, \frac{s}{P(A)P(B)}, \frac{c}{P(B)}\right\}.$$

- The standardized lift is given by

$$\mathcal{L}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - \lambda}{\upsilon - \lambda}.$$

# German Social Life Feeling Data

- Consider data taken from a study of German "social life feelings" that appeared in Schuessler (1982) and Krebs & Schuessler (1987).
- These data have been analyzed many many times, including by Bartholomew & Schuessler (1991), Bartholomew (1991), Bartholomew *et al*. (1997), de Menezes & Bartholomew (1996) and Bartholomew & Knott (1999).
- The data used herein represent the answers given by a sample of 1,490 Germans to five questions.
- McNicholas et al. (2008) use these data in a paper on association rules.



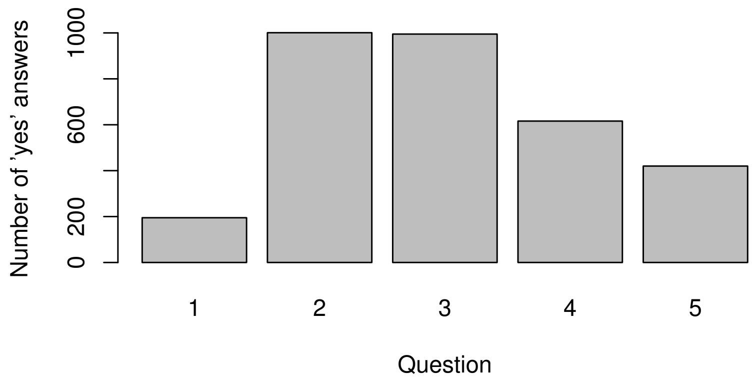


# The Questions

1. Anyone can raise his standard of living if he is willing to work at it.
2. Our country has too many poor people who can do little to raise their standard of living.
3. Individuals are poor because of the lack of effort on their part.
4. Poor people could improve their lot if they tried.
5. Most people have a good deal of freedom in deciding how to live.

# The Answers

- Overall, there were 3,227 “yes” answers and 4,223 “no” answers.




- Only questions 2 and 3 had more than 50% “yes” answers.





# Negations?

- These data raise an interesting point: the fact that “yes” is coded “1” and “no” is coded “0” can be viewed as arbitrary.

- Further, had the questions been worded differently, the “1”s and “0”s could have been be flipped in some or all of the questions.

- The term “negation” can be used to denote the absence of an item from a transaction.

# Negations & German Social Data

- The "1"s were coded y1, y2, y3, y4 and y5, respectively, while the "0"s, or negations, were coded n1, n2, n3, n4 and n5, respectively.
- Association rules were generated using the `arules` package in R with minimum support set at 20% and minimum confidence at 80%.
- This approach led to the generation of 38 association rules.
- We will look at the R code later....





# Negations & German Social Data

- The "best" rule, ranked by standardized lift was $\{$n3,n4$\} \Rightarrow \{$n1$\}$.
- 95.1% of those who did not agree that people were poor because of lack of effort or that poor people could improve their lot if they tried, also did not agree that people could raise their standard of living if they were willing to work at it.
- The ranking of the 38 rules by $\mathcal{L}$ is different than the ranking by either confidence or lift.

# German Social Data: Top 10 Rules

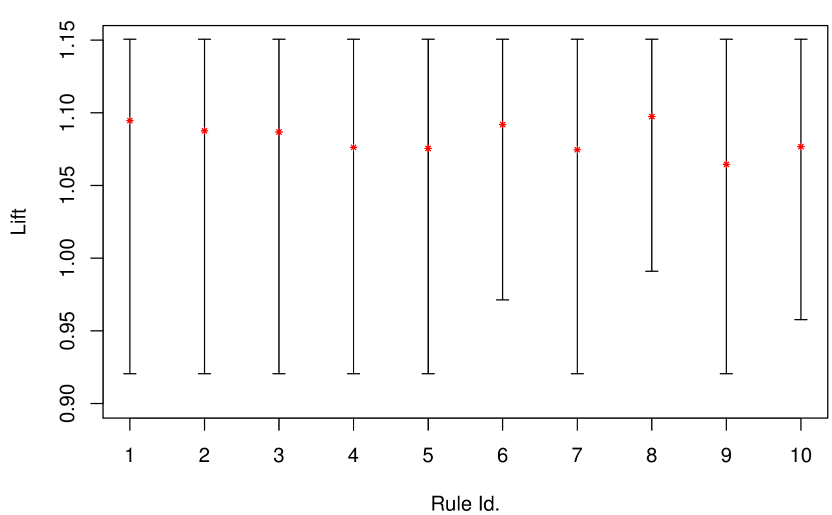
| Id. | Rule | Supp. | Conf. | Lift | $\mathcal{L}$ | LB | UB |
|---|---|---|---|---|---|---|---|
| 1 | {n3,n4} $\Rightarrow$ {n1} | 0.262 | 0.951 | 1.095 | 0.757 | 0.920 | 1.151 |
| 2 | {n3,n5} $\Rightarrow$ {n1} | 0.255 | 0.945 | 1.088 | 0.726 | 0.920 | 1.151 |
| 3 | {n4,n5} $\Rightarrow$ {n1} | 0.446 | 0.945 | 1.087 | 0.723 | 0.920 | 1.151 |
| 4 | {n3} $\Rightarrow$ {n1} | 0.311 | 0.935 | 1.076 | 0.677 | 0.920 | 1.151 |
| 5 | {n4} $\Rightarrow$ {n1} | 0.548 | 0.935 | 1.076 | 0.674 | 0.920 | 1.151 |
| 6 | {n2,n4} $\Rightarrow$ {n1} | 0.225 | 0.949 | 1.092 | 0.673 | 0.971 | 1.151 |
| 7 | {n4,n5,y2} $\Rightarrow$ {n1} | 0.256 | 0.934 | 1.075 | 0.670 | 0.920 | 1.151 |
| 8 | {n3,n4,n5} $\Rightarrow$ {n1} | 0.221 | 0.954 | 1.097 | 0.667 | 0.991 | 1.151 |
| 9 | {n4,y2} $\Rightarrow$ {n1} | 0.323 | 0.925 | 1.064 | 0.626 | 0.920 | 1.151 |
| 10 | {n4,n5,y3} $\Rightarrow$ {n1} | 0.225 | 0.936 | 1.077 | 0.617 | 0.958 | 1.151 |





# Visualization of Std. Lift

- Standardizing the lift shows that rules with the higher lift are not necessarily better rules.



# Redundant Rules

- Looking at the top 10 rules for the German social data, raises an interesting question.
- Are some of the rules redundant.
- That is, are there rules with "extra" elements?
- Let's take a look.





# German Data: Redundant Rules (in Top 10)

| Id. | Rule | Supp. | Conf. | Lift | $\mathcal{L}$ | LB | UB |
|---|---|---|---|---|---|---|---|
| 1 | {n3,n4} $\Rightarrow$ {n1} | 0.262 | 0.951 | 1.095 | 0.757 | 0.920 | 1.151 |
| 2 | {n3,n5} $\Rightarrow$ {n1} | 0.255 | 0.945 | 1.088 | 0.726 | 0.920 | 1.151 |
| 3 | {n4,n5} $\Rightarrow$ {n1} | 0.446 | 0.945 | 1.087 | 0.723 | 0.920 | 1.151 |
| 4 | {n3} $\Rightarrow$ {n1} | 0.311 | 0.935 | 1.076 | 0.677 | 0.920 | 1.151 |
| 5 | {n4} $\Rightarrow$ {n1} | 0.548 | 0.935 | 1.076 | 0.674 | 0.920 | 1.151 |
| 6 | {n2,n4} $\Rightarrow$ {n1} | 0.225 | 0.949 | 1.092 | 0.673 | 0.971 | 1.151 |
| 7 | {n4,n5,y2} $\Rightarrow$ {n1} | 0.256 | 0.934 | 1.075 | 0.670 | 0.920 | 1.151 |
| 8 | {n3,n4,n5} $\Rightarrow$ {n1} | 0.221 | 0.954 | 1.097 | 0.667 | 0.991 | 1.151 |
| 9 | {n4,y2} $\Rightarrow$ {n1} | 0.323 | 0.925 | 1.064 | 0.626 | 0.920 | 1.151 |
| 10 | {n4,n5,y3} $\Rightarrow$ {n1} | 0.225 | 0.936 | 1.077 | 0.617 | 0.958 | 1.151 |

# Why Include Negations?

- Negations were used to facilitate a full analysis of the German social data.

- In general, it will often be the case that the absence of items from the antecedent and the consequent parts of an association rule may of interest.

- The absence of items from the antecedent part can be related to the presence or absence of items from the consequent part and *vice versa*.

- McNicholas et al. (2008) discuss the history of negations and negative association rules.





# How Many More Rules?

- Including negations will lead to the generation of more rules, but how many more?

- Without negations, Hipp et al. (2002) calculate the number of rules that could be generated as:
$$3^n - 2\left(2^n\right) + 1.$$

- When negations are included, McNicholas et al. (2008) calculate the number as:
$$5^n - 2\left(3^n\right) + 1.$$

# How Many More Rules?

- So, the number of 'extra' rules that can be mined when negations are included is given by

$$5^n - 2\,(3^n) + 1 - [3^n - 2\,(2^n) + 1] = 5^n - 3^{n+1} + 2^{n+1}. \qquad (3)$$

- Now, the proportion of rules that contain negations is given by

$$\frac{5^n - 3^{n+1} + 2^{n+1}}{5^n - 2\,(3^n) + 1}. \qquad (4)$$

- Therefore, when an itemset of just $20$ items is considered, it follows that $99.996\%$ of potential association rules involve negations.





# How Many More Rules?

- Of course, the number of rules given by $(3)$ may be unrealistic because a minimum of $2^n$ transactions would be required in order that all potential rules may exist.

- Also, it is unlikely in most practical applications that a transaction involving all, or even most, of the items will occur.

- However, $(4)$ does provide a useful estimate for the proportion of potential rules that will contain negations.

# Comments

- Association rules have been introduced.
- We saw one example, on the German social life feeling data.
- In this example, negations were introduced to facilitate the mining process.
- A general argument was given for the inclusion of negations in the association rule mining process, including a quantification of the amount of rules that can be mined when negations are included.
- Next, we will mine some rules (in R), starting with another look at the German social life feeling data.
- Full bibliographical details for the key references cited herein are given by McNicholas et al. (2008)[a].

[a] McNicholas, P.D., Murphy, T.B. and O'Regan, M. (2008), 'Standardising the lift of an association rule', *Computational Statistics and Data Analysis* **52**(10), 4712–4721.