Association Rules

UBCO MDS — DATA 571

Association Rules Intro

- Association rule mining is primarily used to discover associated items in large databases.
- We are looking for simple rules (and probabilities associated with those rules) that tell us "if this, then that"
- Combinatorially speaking, this is a challenging problem. But conceptually it is pretty straightforward. Let's learn through the most common ARL example — purchasing groceries.

Grocery Example

- We have 30 days of point-of-sale transaction data from a grocery store. Think of your receipt as a single observation in the data.
- We can formulate that receipt (observation) into a binary vector of all categories of items that the grocery store sells.
- For example, if you bought milk and eggs but not cookies or bread then

Observation	Milk	Cookies	Bread	Eggs	...
Receipt 1	·	·	·	·	...
Receipt 2 (yours)	1	0	0	1	...
Receipt 3	·	·	·	·	...
...	·	·	·	·	...

Grocery Example

- For the actual data, we have 9835 transactions (receipts, observations, etc) and have aggregated all of the purchases into 169 categories (also called items).
- So, our hope is to figure out which sets of items (itemsets) are co-purchased
- This will help us generate rules like: "if pasta and chicken are purchased, then we also expect cheese to be purchased"
- From a marketing standpoint, the utility of such information should be obvious. Cheese companies may want to include coupons in the pasta aisle in the hopes that the person will select their cheese over a competitor, for example.

Grocery Example

- Naively, we want to consider all possible itemsets for an association with all other possible itemsets.
- This reasoning gets out of hand quickly!
- Each itemset can either have, or not have, each individual item. So how many unique itemsets are there? Board.
- How many possible association rules exist? Board.

Definitions

- Notation wise, lets let each transaction be

$$t_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$$

  where each x_{ij} equals 0 or 1 indicating whether item j was purchased. Suppose d items and n transactions.

- We seek rules that tell us that itemset $Y \to Z$ with some measurement(s) of that discovered relationship.

- We'll say that itemset $Y = \{x_k = 1, x_l = 1, \ldots\}$ is equivalent to the set $\{k, l, \ldots\}$, meaning that the itemset consists of items k, l, and...

Definitions

- Support is the fraction of transactions that a particular itemset, say $M = \{x_k = 1, x_l = 1, \ldots\}$, appears

$$\text{Supp(M)} = \frac{\sum_{i=1}^{n} I(M \in t_i)}{n}$$

- So, if the itemset of interest only includes one item $M = \{x_j = 1\}$, then the support is $\frac{\sum_{i=1}^{n} x_{ij}}{n}$, or simply the proportion time that item j was purchased.
- So support indicates the proportion of time all items in the itemset are purchased together.

Definitions

- We're seeking measures between itemsets, so
- For itemsets Y and Z, confidence for the relationship $Y \to Z$ is the proportion of transactions that contain both itemsets Y and Z divided by those that only contain Y.

$$\text{Conf(Y, Z)} = \frac{\text{Supp}(Y \cup Z)}{\text{Supp}(Y)} = \frac{\sum_{i=1}^{n} I(Y \in t_i; Z \in t_i)}{\sum_{i=1}^{n} I(Y \in t_i)}$$

- AKA, how likely itemset Z is purchased given that itemset Y was purchased (which implicitly controls for the popularity of Y).
- See board for discussion on confidence...

Definitions

- Lift adjusts confidence by altering the denominator to be the expected $\text{Supp}(Y, Z)$ assuming that Y and Z can be considered independent itemsets

$$\text{Lift(Y, Z)} = \frac{\text{Supp}(Y \cup Z)}{\text{Supp}(Y) \times \text{Supp}(Z)} = \frac{n \sum_{i=1}^{n} I(Y \in t_i; Z \in t_i)}{\sum_{i=1}^{n} I(Y \in t_i) \times \sum_{i=1}^{n} I(Z \in t_i)}$$

- AKA, a measure of how likely itemset Z is purchased given that itemset Y was purchased, controlling for the popularity of Z (and Y).

- $\text{Lift}(Y, Z) = 1$ suggests independence. $\text{Lift}(Y, Z) > 1$ suggests positive relationship. $\text{Lift}(Y, Z) < 1$ suggests negative relationship.

Number of Rules

- Now, back to the problem at hand.
- We can't calculate these measures for all possible $Y \rightarrow Z$ itemset comparisons
- But, especially in cases like the grocery example where most transactions will only include a small subset of items, we can simplify greatly...
- First, we consider only investigating rules that meet minimum thresholds of support and confidence.

Apriori Algorithm

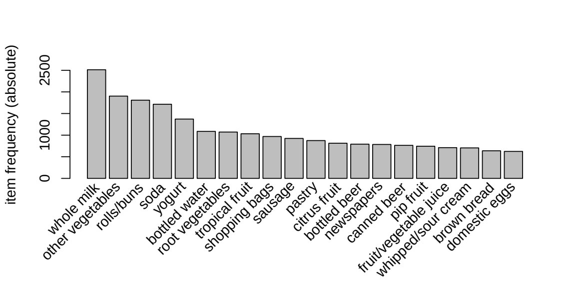
- Next, we recognize that the support of any itemset cannot exceed the support of its subsets.
- So for example, if $Y = \{j\}$ and $Z = \{j, k, \ldots\}$ then Supp(Y) $\geq$Supp(Z)
- And so, when a relatively simple itemset Y is found to have support below the threshold, we don't have to consider any larger itemsets that contain Y.
- See board

Apriori Algorithm

- Furthermore, some softwares only consider single items for the right hand side of the relationship (Z).
- Let's take a look at the grocery data...

AR on Groceries

- Quick exploration of the data



AR on Groceries

```
> summary(arun)
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.329   5.000   6.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001017   Min.   :0.8000   Min.   : 3.131   Min.   :10.00
 1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.: 3.312   1st Qu.:10.00
 Median :0.001220   Median :0.8462   Median : 3.588   Median :12.00
 Mean   :0.001247   Mean   :0.8663   Mean   : 3.951   Mean   :12.27
 3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.: 4.341   3rd Qu.:13.00
 Max.   :0.003152   Max.   :1.0000   Max.   :11.235   Max.   :31.00

mining info:
      data ntransactions support confidence
 Groceries          9835   0.001        0.8
```

AR on Groceries

```
    lhs                          rhs                     support confidence     lift count
[1] {citrus fruit,
     tropical fruit,
     root vegetables,
     whole milk}              => {other vegetables} 0.003152008  0.8857143 4.577509    31
[2] {other vegetables,
     curd,
     domestic eggs}           => {whole milk}       0.002846975  0.8235294 3.223005    28
[3] {hamburger meat,
     curd}                    => {whole milk}       0.002541942  0.8064516 3.156169    25
[4] {herbs,
     rolls/buns}              => {whole milk}       0.002440264  0.8000000 3.130919    24
[5] {tropical fruit,
     herbs}                   => {whole milk}       0.002338587  0.8214286 3.214783    23
```

AR on Groceries

```
> inspect(sort(arun, by="confidence")[1:5])
    lhs                        rhs               support confidence     lift count
[1] {rice,
     sugar}                 => {whole milk} 0.001220132          1 3.913649    12
[2] {canned fish,
     hygiene articles}      => {whole milk} 0.001118454          1 3.913649    11
[3] {root vegetables,
     butter,
     rice}                  => {whole milk} 0.001016777          1 3.913649    10
[4] {root vegetables,
     whipped/sour cream,
     flour}                 => {whole milk} 0.001728521          1 3.913649    17
[5] {butter,
     soft cheese,
     domestic eggs}         => {whole milk} 0.001016777          1 3.913649    10
```

AR on Groceries

```
> inspect(sort(arun, by="lift")[1:5])
    lhs                           rhs                  support confidence      lift c
[1] {liquor,
     red/blush wine}           => {bottled beer}    0.001931876  0.9047619 11.235269
[2] {citrus fruit,
     other vegetables,
     soda,
     fruit/vegetable juice} => {root vegetables} 0.001016777  0.9090909  8.340400
[3] {tropical fruit,
     other vegetables,
     whole milk,
     yogurt,
     oil}                      => {root vegetables} 0.001016777  0.9090909  8.340400
[4] {citrus fruit,
     grapes,
     fruit/vegetable juice} => {tropical fruit}  0.001118454  0.8461538  8.063879
[5] {other vegetables,
     whole milk,
     yogurt,
     rice}                     => {root vegetables} 0.001321810  0.8666667  7.951182
```


UBC
THE UNIVERSITY OF BRITISH COLUMBIA