Market Basket Analysis IN R

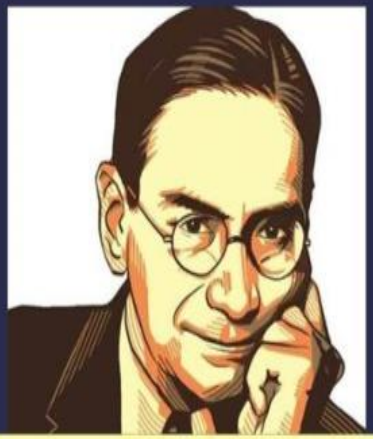
By

Sourav Maji

LEARN, EXPLORE AND

INNOVATE

RESEARCH PROJECTS ON STAT, DS & ML



Training Programs

Research Internships

STAT&ML LAB

EXPLORE MORE AT

HTTPS://WWW.CTANUJIT.ORG/STATML-LAB.HTML

INTRODUCTION

- Frequent pattern mining searches for recurring relationships in a given data. A typical example of frequent pattern mining is Associate rule Mining/ Market-Basket Analysis.
- Market-Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. In other words, it analyzes customer buying habits by finding associations between different items that customers place in their shopping baskets.

Theoretical framework & Concepts

- $I = \{ I_1, I_2, \ldots, I_m \}$ be an itemset.
- A non-empty itemset $T \subseteq I$ is a *transaction*.
- A set of transactions $D = \{T_1, T_2, \ldots, T_n\}$ is a *transaction data base*.
- Market basket transactions:

  T1: {bread, cheese, milk}

  T2: {apple, eggs, salt, yogurt}

  … …

  Tn: {biscuit, eggs, milk}

Theoretical framework & Concepts

- Support: The rule holds with support *sup* in T (the transaction data set) if sup% of transactions contain $X \cup Y$.
  - $sup = \Pr(X \cup Y)$.
- Confidence: The rule holds in T with confidence *conf* if *conf*% of tranactions that contain X also contain Y.
  - $conf = \Pr(Y \mid X)$
- An association rule is a pattern that states when X occurs, Y occurs with certain probability.
- Support count: The support count of an itemset X, denoted by $X.count$, in a data set T is the number of transactions in T that contain X. Assume T has n transactions.
- Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Theoretical framework & Concepts

- *Association rule* is an implication of the form

  $X \Rightarrow Y$, where $X, Y \subset I$, $X \cap Y = \Phi$

- The rule $X \Rightarrow Y$ in a transaction D holds with *support* $s = P(X \cup Y)$, i.e., s% of transactions in D contains both X and Y, and has *confidence* $c = P(Y \mid X)$ that c% of transactions that contain X contain Y.
- An itemset I satisfying a minimum support threshold is called a *frequent itemset*.
- Rules that satisfy both the minimum support threshold and a minimum confidence threshold are called *strong rules*.

Algorithms

- Association rule Mining can be viewed as a two-step process:
- Find all frequent itemsets.
- Generate strong association rules from the frequent itemsets.

Algorithm Apriori(T)

$C_1 \leftarrow$ init-pass(T);
$F_1 \leftarrow \{f \mid f \in C_1, f.\text{count}/n \geq minsup\}$; // n: no. of transactions in T
for ($k = 2$; $F_{k-1} \neq \emptyset$; k++) **do**
    $C_k \leftarrow$ candidate-gen(F_{k-1});
    for each transaction $t \in T$ **do**
        for each candidate $c \in C_k$ **do**
            if c is contained in t **then**
                $c.count$++;
        end
    end
    $F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$
end
return $F \leftarrow \bigcup_k F_k$;

Algorithms: Example –Frequent itemsets

Dataset T

minsup=0.5

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

itemset:count

1. scan T ➔ C_1: {1}:2, {2}:3, {3}:3, {4}:1, {5}:3

   ➔ F_1: {1}:2, {2}:3, {3}:3, {5}:3

   ➔ C_2: {1,2}, {1,3}, {1,5}, {2,3}, {2,5}, {3,5}

2. scan T ➔ C_2: {1,2}:1, {1,3}:2, {1,5}:1, {2,3}:2, {2,5}:3, {3,5}:2

   ➔ F_2: **{1,3}**:2, **{2,3}**:2, **{2,5}**:3, **{3,5}**:2

   ➔ C_3: {2, 3,5}

3. scan T ➔ C_3: **{2, 3, 5}**:2 ➔ F_3: **{2, 3, 5}**

Concept of Lift

- The *lift* between the occurrence of X and Y can be measured by computing
- lift (X,Y) = P(X $\cup$ Y)/(P(X)*P(Y))
- If lift (X,Y) = 1, then occurrence of X is independent of occurrence of Y and vice-a-versa.
- < 1, then the occurrence of X is negatively correlated with the occurrence of Y, i.e., occurrence of one likely leads to absence of the other one
- > 1, then the occurrence of X is positively correlated with the occurrence of Y, i.e., occurrence of one implies the occurrence of the other

APPLICATIONS

Examples of areas in which association rules have been used include:

- Supermarket Purchases: Common combinations of products can be used to inform product placement on supermarket shelves (Hence the name, MBA)
- Insurance Claims: Unusual combinations of insurance claims can be a sign of fraud.
- Medical Patient histories: Certain combinations of conditions can indicate increased complications.

MARKET BASKET ANALYSIS

Importing the dataset

```
>library(arules)
>mbd=read.csv("C:/Users/Sourav/Desktop/Book4.csv)
>dt=split(mbd$Items,mbd$Id)
>dt1=as(dt,"transactions")
```

Creating rules

```
> rules <- apriori(dt1, parameter = list(supp = 0.1, conf = 0.8,minlen=2))
> rview=inspect(rules)
```

rules	support	confidence	lift
{OS} => {Visual Studio}	0.3535354	0.6481481	1.188272
{Visual Studio} => {OS}	0.3535354	0.6481481	1.188272
{OS} => {AntiVirus}	0.4646465	0.8518519	1.249383
{AntiVirus} => {OS}	0.4646465	0.6814815	1.249383
{OS} => {Project Mgmt}	0.4696970	0.8611111	1.217857
{Project Mgmt} => {OS}	0.4696970	0.6642857	1.217857
{OS} => {Office Suite}	0.4747475	0.8703704	1.180365
{Office Suite} => {OS}	0.4747475	0.6438356	1.180365

MARKET BASKET ANALYSIS

Exporting rules

> dataframe=as.data.frame(as.matrix(rview))

>write.csv(dataframe,file="C:/Users/Sourav/Desktop/dataframe.csv")

Interest measures

>intm=interestMeasure(rules,c("chisquared","fishersExactTest"),dt1)

> rulesm=cbind(dataframe,intm)

>rulesm

sl no	lhs	V2	rhs	support	confidence	lift	chiSquared	imbalance
1	{fast}	=>	{really}	0.1111111	1	6	11.25	0.333333333
2	{features}	=>	{phone}	0.1111111	1	3	4.5	0.666666667
3	{latest}	=>	{android}	0.1111111	1	9	18	0
4	{android}	=>	{latest}	0.1111111	1	9	18	0
5	{latest}	=>	{really}	0.1111111	1	6	11.25	0.333333333
6	{android}	=>	{really}	0.1111111	1	6	11.25	0.333333333
7	{also}	=>	{phone}	0.1111111	1	3	4.5	0.666666667
8	{worth}	=>	{good}	0.1111111	1	6	11.25	0.333333333
9	{android,latest}	=>	{really}	0.1111111	1	6	11.25	0.333333333
10	{latest,really}	=>	{android}	0.1111111	1	9	18	0
11	{android,really}	=>	{latest}	0.1111111	1	9	18	0

The value of ChiSquare with 1 degree of freedom is 3.84 (α=0.05). The "chiSquare" values more than 3.84 is significant.

"imbalance"=" Imbalance ratio", which is lower the better.

Thank You