Data Mining
Lecture 10: Market basket Analysis

Jo Grundy

ECS Southampton

17^{th} March 2022

Market Basket - Introduction

Why analyse market baskets?
Get insight:

- do products sell quickly or slowly
- which products are sold together?
- which might need a promotion?

Use that to take action:

- store layout
- promotions
- recommendations

Market Basket - Introduction

Learn Patterns:

- If I buy baking powder, but no flour, what am I baking?
- If I buy a mobile phone and no case, could the shop make more money?

Market Basket - Introduction

What do we mean?
Association Rules:
if X then Y
$X \Rightarrow Y$
Looking for rules to predict if something X is bought, what else is likely to be bought?

Market Basket - Introduction

Beer and Nappies
Back in 1992 A data consultant was using SQL queries to find things were often bought along side nappies (Diapers in the US), as nappies are high margin, they wanted to sell more of them. They were looking to find things to put on the shelves near each other. She found a correlation between beer sales, and nappy sales, and emailed her colleagues about it.
There was no good statistical basis for this link, but the story has become well known, one of the first to 'go viral'

Market Basket - Introduction

Market Basket analysis:

Given a database of transactions
Find groups of items that are frequently bought together

Each transaction is a set of items, a basket, called here an *itemset*
This allows companies to understand why people make certain purchases

Market Basket - Applications

Insight can be gained about the products they sell

- Which sell quickly or slowly?
- Which are bought together?
- Identify possible missed opportunities

This helps companies to decide on:

- How to layout a shop?
- Which products to promote?

E. g. if one specific product (e.g. "Earl Grey Redbush Tea") is only rarely bought, but when it is bought that same customer spends lots of money on other products, is it worth keeping it just for that person?

Market Basket - Applications

Other applications include:

- communication (set of phone calls)
- banks (each account is a transaction)
- Medical Treatment (a patient is a transaction with a set of diseases!)

The maths and algorithms are very similar for all.

Market Basket - Definitions

Definitions:

- $I = i_1, i_2, \ldots, i_n$ is a set of all items
- Transaction t_i is a set of items such that $t_i \subseteq I$ (basket)
- Transaction database D contains all transactions $t_1, \ldots, t_d$
- An **Association Rule** is where $X \implies Y$, i.e. X implies Y
- An **itemset** is a set of items. If it has k items, it is a $k - itemset$

Market Basket - Definitions

- **Support** s of an itemset X is the percentage of transactions in D that contain X
- **Support** of **association rule** $X \implies Y$ is the support of the itemset $X \cup Y$
- **Confidence** of the rule $X \implies Y$ is the ratio between the transactions that contain both X and Y and the number of transactions that have X in D

Market Basket - Problem

Problem: Find association rules
Given:

- a set I of items
- database D of transactions
- minimum support s
- minimum confidence c

Find: Association rules $X \implies Y$ with a minimum support s and minimum confidence c

Market Basket - Problem

Solution

- Find all itemsets that have minimum support
- Generate rules using frequent itemsets

Market Basket - Association Rule Mining

For example:

Transaction	Items
1	coffee, pen
2	coffee, pastry
3	coffee, paper, pen
4	pastry, crisps

If minimum support is 0.5 then only 2-itemset coffee, pen has minimum support

Market Basket - Association Rule Mining

Step 1 : Generate frequent itemsets

frequent itemset	itemset support
coffee	0.75
pen	0.5
pastry	0.5
coffee, pen	0.5

Step 2: Generate Rules
Confidence: ratio of transactions that have both X and Y and the number of transactions that have X in D

rule	support	confidence
coffee => pen	0.5	$2/3 = 0.6$
pen => coffee	0.5	$2/2 = 1$

Market Basket - Association Rule Mining

Using this transaction database D
Find most frequent *itemsets*

Transaction	Itemsets
t_1	A, B, C
t_2	A, C
t_3	A, C, D
t_4	A, E
t_5	D, E

$$support = \frac{freq(item)}{n}$$

Where n = number of transactions

itemsets	frequency	support
$\{A\}$	4	0.8
$\{B\}$	1	0.2
$\{C\}$	3	0.6
$\{D\}$	2	0.4
$\{E\}$	2	0.4
$\{A, B\}$	1	0.2
$\{A, C\}$	3	0.6
$\{A, D\}$	1	0.2
$\{A, E\}$	1	0.2
$\{B, C\}$	1	0.2
$\{D, E\}$	1	0.2
$\{A, B, C\}$	1	0.2
$\{A, C, D\}$	1	0.2

Market Basket - Association Rule Mining

itemsets	frequency	support
$\{A\}$	4	0.8
$\{B\}$	1	0.2
$\{C\}$	3	0.6
$\{D\}$	2	0.4
$\{E\}$	2	0.4
$\{A, B\}$	1	0.2
$\{A, C\}$	3	0.6
$\{A, D\}$	1	0.2
$\{A, E\}$	1	0.2
$\{B, C\}$	1	0.2
$\{D, E\}$	1	0.2
$\{A, B, C\}$	1	0.2
$\{A, C, D\}$	1	0.2

With minimum support 0.4:

itemsets	frequency	support
$\{A\}$	4	0.8
$\{C\}$	3	0.6
$\{D\}$	2	0.4
$\{E\}$	2	0.4
$\{A, C\}$	3	0.6

So the only rules we can examine are $A \implies C$ or $C \implies A$

assn rules	support	confidence
$A \implies C$	0.6	0.75
$C \implies A$	0.6	1.00

Market Basket - A Priori Algorithm

The Apriori Algorithm

We know:

- Any subset of a *frequent itemset* is also frequent
- Any superset of an infrequent itemset is also infrequent

Let:

- L_k = set of frequent $k - itemsets$ (have minimum support)
- C_k = set of candidate $k - itemsets$ (potentially frequent)

Market Basket - A Priori Algorithm

Algorithm 1: A Priori Algorithm

```
Data: D transaction database, minSupport
L_1 = {frequent items};
k = 1;
while L_k not empty do
    C_{k+1} = all possible candidates from L_k;
    for each transaction t in D do
        if candidate in Ck + 1 is in t then
            increment count for candidate;
        end
    end
    L_{k+1} = candidates in C_{k+1} with minSupport;
    k = k + 1;
end
```


Market Basket - A Priori Algorithm

Algorithm 2: A Priori Algorithm - Generating Candidates

```
Data: L_{i-1}
C_i = {};
for each itemset J in L_{i-1} do
    for each itemset K in L_{i-1} such that K ≠ J do
        if i - 2 elements in J and K are equal then
            if all subsets of {K ∪ J} are in L_{i-1} then
                C_i = C_i ∪ {K ∪ J};
            end
        end
    end
end
return C_i;
```


Market Basket - A Priori Algorithm

Simple example:

minSupport = 0.5

Database D:

Transaction	Basket
t_1	A, C, D
t_2	B, C, E
t_3	A, B, C, E
t_4	B, E

$k = 1$,
Go through D:

itemset	support
{A}	0.5
{B}	0.75
{C}	0.75
{D}	0.25
{E}	0.75

So L_1 = {A, B, C, E}
$\therefore C_2 =$

itemset	support
{A, B}	0.25
{A, C}	0.5
{A, E}	0.25
{B, C}	0.5
{B, E}	0.75
{C, E}	0.5

So $L2$ ={ {A, C}, {B, C}, {B, E}, {C, E} }

Market Basket - A Priori Algorithm

$k = 3$
$L2 =$ { {A, C}, {B, C}, {B, E}, {C, E} }
Generating Candidates:
{A, C}, {B, C} are both in L_2, giving {A, B, C}
Not all subsets of {A, B, C} are in L_2
{A, C}, {C, E} are both in L_2 giving {A, C, E}
Not all subsets of {A, C, E} are in L_2
{B, C}, {B, E} are both in L_2 giving {B, C, E}
All subsets of {B, C, E} are in L_2 so:

Go through D:

itemset	support
{B, C, E}	0.5

Market Basket - Generating Rules

Transaction	Basket
t_1	A, C, D
t_2	B, C, E
t_3	A, B, C, E
t_4	B, E

Consider 3-itemset {B, C, E}
Use all permutations of rules from these three items

rule	support	confidence
$\{B, C\} \implies E$	0.5	$2/2 = 1$
$\{B, E\} \implies C$	0.5	$2/3 = 0.66$
$\{C, E\} \implies B$	0.5	$2/2 = 1$
$E \implies \{B, C\}$	0.5	$2/3 = 0.66$
$C \implies \{B, E\}$	0.5	$2/3 = 0.66$
$B \implies \{C, E\}$	0.5	$2/3 = 0.66$

Market Basket - A Priori Algorithm

Advantages of A Priori Algorithm:

- Uses large itemset property
- Can be Parallelised
- Easy to implement

Disadvantages

- Assumes D transaction database is in memory
- Requires many database scans

Market Basket - Improvements

Confidence of a rule is the ratio between transactions with $X \cup Y$ to the number of transactions with X

$$conf(X \implies Y) = \frac{\frac{nTrans(X \cup Y)}{|D|}}{\frac{nTrans(X)}{|D|}} = \frac{p(X \wedge Y)}{p(X)} = p(Y|X)$$

If Y is independent of X: $p(Y) = p(Y|X)$
This means if you have a high probability of p(Y) we have a rule with high confidence that associates independent itemsets
e.g. if p("bread") = 0.8, and "bread" is independent from "sausages", then the rule "bread" $\implies$ "sausages" will have confidence 0.8

Market Basket - Improvements

Alternative measures:
lift measure indicates departure from independence of X and Y
the **lift** of $X \implies Y$ is:

$$lift(X \implies Y) = \frac{conf(X \implies Y)}{p(Y)} = \frac{\frac{p(X \wedge Y)}{p(X)}}{p(Y)} = \frac{p(X \wedge Y)}{p(X)p(Y)}$$

Unfortunately, lift is *symmetric*, the same for $X \implies Y$ as $Y \implies X$

Market Basket - Improvements

Conviction indicates that X and Y are not independent, and takes in to account the direction of implication
The conviction of $X \implies Y$ is: [1]

$$conv(X \implies Y) = \frac{p(X)p(\neg Y)}{p(X \wedge \neg Y)}$$

[1] Brin *et al* SIGMOD 1997

Market Basket - Linked Concepts

"Baskets" = **documents**
"items" = **words** in those documents
If we can find words that appear together more often than others, these are **linked concepts**

	word1	word2	word3	word4
doc1	1	0	1	1
doc2	0	0	1	1
doc3	0	1	1	0

$\therefore word4 \implies word3$
As when *word*4 occurs, there is a large probability that *word*3 will also occur

Market Basket - Linked Concepts

Detecting Plagarism
"Baskets" = **sentences**
"items" = **documents** containing those sentences
Items that appear together could mean that a student has copied work from another document, plagarism!

	doc1	doc2	doc3	doc4
sent1	1	0	1	1
sent2	0	0	1	1
sent3	0	1	1	0

Here..
$\therefore doc4 \implies doc3$
If there is a sentence occurring in document 4, there is a high probability of it occurring in document 3, so if *doc*3 is your coursework, you may be in trouble!

Market Basket - Linked Concepts

Web pages
"Baskets" = **web pages**
"items" = **linked pages**
Pairs of pages with many common references may be about the same topic
"Baskets" = **web pages,** p_1
"items" = **pages that link to** p_1
Pages with many of the same links may be mirrors or about the same topic

Market Basket - Summary

Terms were defined:

- **Association rules**: if X then Y, $X \implies Y$
- **Items** I, set of all possible items i
- **Transaction**: set of items t_i such that $t_i \subset I$
- **Database** D containing all transactions $\{t_i\}_1^d$
- **Itemset**: subset of I, with k items is a $k-itemset$

Measures were defined:

- **Support** of itemset X is % transactions in D that contain X
- Support of Association rule $X \implies Y$ is $\frac{|t \in D; X \cup Y \subset t|}{|t \in D; X \subset t|}$
- **Confidence** is $\frac{Sup(X \cup Y)}{Sup(X)}$
- **Lift** is $\frac{Sup(X \cup Y)}{Sup(X)Sup(Y)}$
- **Conviction** is $\frac{p(X)p(\neg Y)}{p(X \wedge \neg Y)}$

A Priori Algorithm described