Class comparisons association rule mining: Market basket analysis- basic concepts.

Submitted by:
Jay kakdiya: 160410116046
Kaushiki kansara: 160410116048
Devangini kathad : 160410116049
Kaustubh wade : 160410116050

1

Concept description

- Data mining can be classified into two categories: **descriptive** data mining and **predictive** data mining.
- Descriptive data mining **describes the data set** in a **concise and summative manner** and **presents interesting general properties** of the data.
- Predictive data mining **analyzes the data** in order to **construct one or a set of models**, and attempts to **predict the behavior of new data sets**.
- Database is usually storing the large amounts of data in great detail. However users often **like to view** sets of **summarized data in concise, descriptive terms**.

2

Market basket analysis

- Market Basket Analysis is a **modelling technique**.
- It is based on, if you buy a certain group of items, you are more (or less) likely to buy another group of items.
- For example, if you are in a store and you buy a car then you are more likely to buy insurance at the same time than somebody who didn't buy insurance.
- The **set of items** a customer buys is referred to as an **itemset**.
- Market basket analysis seeks to **find relationships between purchases** (Items).
    - **E.g.** IF {Car, Accessories} THEN {Insurance}

{Car, Accessories} → {Insurance}

3

Market basket analysis (Cont..)

{Car, Accessories} → {Insurance}

- The **probability** that a customer will buy car **without** an accessories is referred as the **support** for **rule**.
- The **conditional probability** that a customer will purchase Insurance is referred to as the **confidence**.

4

Association rule mining

- Given a set of transactions, we need rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.
- **Market-Basket transactions**

TID	Items
1	Bread, Milk
2	Bread, Chocolate, Pepsi, Eggs
3	Milk, Chocolate, Pepsi, Coke
4	Bread, Milk, Chocolate, Pepsi
5	Bread, Milk, Chocolate, Coke

Example of Association Rules

{Chocolate} → {Pepsi},
{Milk, Bread} → {Eggs, Coke},
{Pepsi, Bread} → {Milk}

5

Association rule mining (Cont..)

TID	Items
1	Bread, Milk
2	Bread, Chocolate, Pepsi, Eggs
3	Milk, Chocolate, Pepsi, Coke
4	Bread, Milk, Chocolate, Pepsi
5	Bread, Milk, Chocolate, Coke

- **Itemset**
  - A collection of **one or more items**
    - **E.g.** : {Milk, Bread, Chocolate}
  - k-itemset
    An itemset that contains **k** items
- **Support count (σ)**
  - **Frequency** of occurrence of an **itemset**
    - **E.g.** σ({Milk, Bread, Chocolate}) = 2
- **Support**
  - **Fraction of transactions** that **contain an itemset**
    - **E.g.** s({Milk, Bread, Chocolate}) = 2/5
- **Frequent Itemset**
  - An itemset whose **support is greater than or equal to a minimum support** threshold

6

Association rule mining (Cont..)

TID	Items
1	Bread, Milk
2	Bread, Chocolate, Pepsi, Eggs
3	Milk, Chocolate, Pepsi, Coke
4	Bread, Milk, Chocolate, Pepsi
5	Bread, Milk, Chocolate, Coke

Example:

Find support & confidence for {Milk, Chocolate} ⇒ Pepsi

$$s = \frac{\sigma(Milk,Chocolate,Pepsi)}{|T|} = \frac{2}{5} = \mathbf{0.4}$$

$$c = \frac{\sigma(Milk,Chocolate,Pepsi)}{\sigma(Milk,Chocolate)} = \frac{2}{3} = \mathbf{0.67}$$

7

Association rule mining - example

TID	Items
1	Bread, Milk
2	Bread, Chocolate, Pepsi, Eggs
3	Milk, Chocolate, Pepsi, Coke
4	Bread, Milk, Chocolate, Pepsi
5	Bread, Milk, Chocolate, Coke

Calculate **Support** & **Confidence:**
{Milk, Chocolate} → {Pepsi}
{Milk, Pepsi} → {Chocolate}
{Chocolate, Pepsi} → {Milk}
{Pepsi} → {Milk, Chocolate}
{Chocolate} → {Milk, Pepsi}
{Milk} → {Chocolate, Pepsi}

Answer

Support (s) : **0.4**

{Milk, Chocolate} → {Pepsi} c = **0.67**
{Milk, Pepsi} → {Chocolate} c = **1.0**
{Chocolate, Pepsi} → {Milk} c = **0.67**
{Pepsi} → {Milk, Chocolate} c = **0.67**
{Chocolate} → {Milk, Pepsi} c = **0.5**
{Milk} → {Chocolate, Pepsi} c = **0.5**

8

Association rule mining (Cont..)

- A common strategy adopted by many association rule mining algorithms is to decompose the problem into two major subtasks:
  1. **Frequent Itemset Generation**
     - The objective is to find all the item-sets that satisfy the minimum support threshold.
     - These itemsets are called **frequent itemsets**.
  2. **Rule Generation**
     - The objective is to extract all the high-confidence rules from the frequent itemsets found in the previous step.
     - These rules are called **strong rules**.

9

Apriori algorithm

- **Purpose**: The Apriori Algorithm is an influential algorithm for mining **frequent itemsets** for Boolean **association rules**.
- **Key Concepts**:
  - **Frequent Itemsets**:
    The sets of item which has **minimum support** (denoted by L_i for ith-Itemset).
  - **Apriori Property**:
    Any **subset of frequent itemset must** be **frequent**.
  - **Join Operation**:
    To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} itself.

10

Apriori algorithm steps (Cont..)

- **Step 1:**
  - Start with itemsets containing just a **single item (Individual items).**
- **Step 2:**
  - Determine the support for itemsets.
  - **Keep** the itemsets that **meet your minimum support threshold** and **remove** itemsets that **do not support minimum support**.
- **Step 3:**
  - Using the itemsets you have kept from Step 1, **generate all the possible itemset combinations**.
- **Step 4:**
  - **Repeat** steps 1 & 2 until there are **no more new itemsets**.

11

Apriori algorithm - Pseudo code (Cont..)

C_k: Candidate itemset of size k

L_k: Frequent itemset of size k

L_1= {frequent items};

for (k = 1; L_k != Ø; k++) **do begin**

C_{k+1} = candidates generated from L_k;

for each transaction **t** in database do

Increment the count of all candidates in C_{k+1}

That are contained in **t**

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$;

12

THANK YOU

13