Assocation Rules

Prof. Eric A. Suess

April 29, 2020

Introduction

Do you ever make ‘’impulse purchases”?

Why is the product always ‘’nearby’‘and’‘easily available”?

On some of the online retailer websites there is the ‘’wishlist’‘or’‘shopping cart”?

Introduction

Recommendation systems have been based on subjective experience of marketers.

With online retailers machine learning has been used to learn the patterns of **purchasing behavior**.

With barcode scanners, computerized inventory systems, and online shopping there is a lot of **transactional data** available for **data mining**.

Introduction

Do you know what an SKU is?

Answer: **Stock Keeping Unit**

Introduction

In this chapter we will learn about methods for identifying associations among items in transactional data.

This is known as **market basket analysis**.

Understanding association rules

The result of **market basket analysis** is a set of **association rules**.

For example,

```
{peanut butter,jelly}  ->  {bread}
```

Association rules are learned from subsets of **itemsets**.

Understanding association rules

Association rules were developed in the context of **Big Data** and **database science** and **data mining** for **knowledge discovery** (KDD).

Looking for the needle in the haystack.

Association rules are **unsupervised**, so there is **no need** for the algorithm to be **trained**.

And there is no objective measure of performance for such rule learners.

Apriori algorithm

The complexity of transactional data is what makes association rule mining a challenging task.

Transactional datasets are typically **extremely large**, both in terms of the **number of transactions** and the **number of features** or items for sale.

The potential itemsets grows with the number of items for sale.

The good thing is that many itemsets are rare.

By **igoring rare itemsets**, it is possible to limit the search for rules.

Apriori algorithm

The most widely used algorithm is the **Apriori algorithm**.

It employs a simple *a priori* belief as a guideline for reducing the association rule space, *all subsets of a frequent itemset must also be frequent*. This is the **Apriori property**.

See the paper

Fast algorithms for mining association rules, Agrawal and Srikant (1994).

Or

A comparison of association rule discovery and bayesian network causal inference algorithms, Bowes, et. al.

Measuring rule interest

Whether or not an association rule is deemed **interesting** is determined by two statistical measures:

- support

$P(X)$

- confidence

$P(Y|X)$

Measuring rule interest

By providing **minimum thresholds** for each of these metrics and applying the Apriori principle, it is easy to limit the number of rules reported.

Measuring rule interest - support

The **support** of an itemset measures how frequently it occurs in the data.

$$support(X) = \frac{count(X)}{N}$$

where N is the number of transactions in the database and $count(X)$ is the number of transactions that contain the itemset X.

Measuring rule interest - confidence

A rule's **confidence** is a measurement of its predictive power or accuracy.

It is defined as the support of the itemset containing both X and Y divided by the support of the itemset containing only X.

$$confidence(X \rightarrow Y) = \frac{support(X,Y)}{support(X)}$$

The confidence tells us the proportion of transactions where the presence of item or itemset X results in the presence of item or itemset Y.

Measuring rule interest - confidence

Note $X \rightarrow Y$ is not the same as $Y \rightarrow X$.

Rules that have **high support** and **high confidence** are referred to as **strong rules**.

Building a set of rules with the Apriori principle

The **Apriori principle** states that all subsets of a frequent itemset must also be frequent.

The **Apriori algorithm** uses the **Apriori principle** to exclude potential association rules prior to actually evaluating them.

The **process of creating rules** occurs in two phases:

- find all itemsets that meet a minimum **support threshold**
- create rules from these itemsets that meet a minimum **confidence threshold**

Example

The author gives an example of the use of Market Basket Analysis using **transaction data** to identify frequently purchased groceries with association rules.

Recommendation system

Example

The example uses:

- **unstructured data**
- **nosql**
- **sparse matrix**

We will use the R packages

- arules
- arulesViz

Example

Lift is a metric used to measure how much more likely one item is to be purchased relative to its typical purchase rate, given that you know another item has been purchased.

$$lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{support(Y)}$$

Here

$$lift(X \rightarrow Y) = lift(Y \rightarrow X)$$