Unit - III

Page No.
Date: / /

Association Rule and Regression

Association Rule and Regression

Advanced analytical theory & methods : Association Rules

Overview.

Given a large collection of transactions, in which each transaction consists of one or more items. association rules go through the items being purchased to see what items are frequently brought together and to discover a list of rules that describe the purchasing behaviours. The goal with the association rule is to discover a list of rules that describe the purchasing behaviours. The goal with association interesting relationship among the item (the relationship occurs too frequently) to be random and is meaning full from a business prespective which may or may not be obious) The relationship that are interesting depend both on buiness context and the nature of the algorithm being used for the discovery. Each of the uncovered rules is in the form $X \rightarrow Y$, meaning that them item X is observed, item Y is also observed in this case the left hand side (LHS) of the rule is X and the right hand side (RHS) of the Rule Y.

Apriori Algorithm

Apriori algorithm takes a bottom up iterative approach to uncovering the frequent itemsets by first determining all the possible item and then identifying which among them are frequent.

At each iteration the algorithm checks whether the support criterion can be met if it can, the algorithm grows the itemset, repeating the process until it runs out of support or until the itemset reach a predefined length.

The Apriori algorithm is given next. let variable C be the set of candidate k-itemsets and variable L_k be the set of k itemset that satisfy the minimum support. Given a transaction database D a minimum support threshold and an optional parameter N indicating the maximum length an itemset could reach Apriori iteratively compute frequent itemsets (L_{k-1} based on

Algorithm

1. Apriori (D, S, N)
2. $k \leftarrow 1$
3. $L_k \leftarrow$ { i = itemsets that satisfy minimum support of }
4. while $L_k \neq \emptyset$
5. if $k \neq N \vee (\exists N \wedge k < N)$
6. $C_{k+1} \leftarrow$ Candidate itemset generated from
7. for each transaction t in database D do
8. increment the count of C_{k+1} that satisfy minimum support of
9. $k \leftarrow |L_k|$
10. Return $\bigcup L_k$

The first step of the Apriori algorithm is to identify the frequent itemsets by starting with each item in the transaction that meet the predetermined minimum support threshold δ. These itemset are 1- itemset denoted as L_1 as each 1-itemset contains only one item. Next the algorithm grows the itemsets by joining L_1 onto itself to form new grown-2 itemset denoted as L_2 and determines the support of each 2 itemset in L_2. Those itemset that do not meet the minimum support threshold δ are pruned away.

Next a collection of candidate rule is formed based on the frequent itemsets uncovered in the iterative process described earlier. For example a frequent itemset {milk, eggs} may suggest candidate rule

Evaluation of Candidate Rules.

Frequent itemset from the previous section can form candidate rule such as X implies Y ($X \Rightarrow Y$). This section discusses how measures such as confidence lift and leverage can help evaluate the appropriateness of these candidate Rule.

Confidence [C] is defined as the measure of certainty or trustworthiness associated with each discovered rule. Mathematically confidence is the percent of transactions that contain both X & Y out of all the transactions that

Confidence & (see Equations)

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

for example, if {bread, egg, milk} has a support of 0.15 and {bread & eggs} also has a support of 0.15 the confidence of rule {bread, egg} → {milk} is 1 which means 100% of the time a customer buys bread and egg & milk is bought as well the rule is therefore correct for 100% of the transcations containing bread & eggs.

A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold. This predefined threshold is called the minimum confidence. A higher confidence indicates that the rule (X → Y) is more interesting or more trustworthy, based on the sample dataset.

Lift measures how many times more often X and Y occurs together than expected if they are statically independent of each other. Lift is measure of how X & Y are really related rather than coincidently happening together.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{support } X \times \text{support } Y}$$

Date: / /

Case study - transaction in grocery store

An example illustrate the application of the Apriori algorithm to a relatively simple case that generalizes to those used in practice using R and the arules and arulesViz packages. This example show how to use the Apriori algorithm to generate frequent itemsets and rules and to evaluate and visualize rules

The following commands install these two packages and import them into the current R work space

```
install.packages('arules')
install.packages('arulesViz')

library('arules')
library('arulesViz')
```

(i) The Groceries Dataset

The example uses the groceries dataset from the R arules package. The groceries dataset is 9835 transactions and the items are aggregated into 169 categories

```
data(groceries)
Groceries
```

transaction in sparse format with
9835 transactions (rows) and
169 (columns) items

The summary shows that most frequent item in the dataset include item such as whole milk other vegetables, rolls/buns, soda and yogurt. these items are purchased more often than the others

The class of the dataset is transactions as defined by the arules package. The transaction class contains three slots.

Transaction.info : A data frame with vectors of the same length as the number of transcations class contains three slots.

Iteminfo: A data frame to store item labels

data:- A binary incidence matrix that indicates which item labels appear in every transaction

Frequent itemset Generation.

The results are accessed based on the specific business context of the excersise using specific dataset. if the dataset changes or a different minimum support threshold is chosen the Apriori algorithm must run each iteration again to retrieve the updated frequent itemset

Rule Generation & Visualization

The Apriori() function can also be used to generate rule. Assume that the minimum support threshold is now set to a lower value

0.001 and the minimum confidence threshold is set to 0.6 A lower minimum support threshold allows more rule to show up. The following code creates 2918 rules from all transaction in the groceries dataset that satisfy both the minimum support and the minimum confidence

Validation and Testing

After gathering the o/p rules it may become necessary to use one or more methods to validate the results in the business context for the sample dataset The first approach can be established through statistical measures such as Confidence, lift and leverage Rule that involve mutually independent items or com few transactions are considered uninteresting because they may capture spurious relationship

Confidence measures the chance that x & appear together in relation to chance x appears. Confidence can be used to identify the interestingness of the rule

lift & the leverage both compare the support of x & y against their individual support while mining data with association rules some rules generated could be purely coincidental for example if 95% of customers buy x & 90% of customers buy y then x & y would occur together at least 85% of the time even if there is no

relationship between the two measures like lift and leverage ensure that interesting rules are identified rather than coincidental ones.

Another set of criteria can be established through subjective arguments. Even with a high confidence rule may be considered subjectively uninteresting unless it reveals any unexpected profitable actions. For example rules like {paper} → {pencil} may not be subjectively interesting or meaningful despite high support and confidence values. In contrast a rule like {diaper} → {beer} that satisfies both minimum support and minimum confidence can be considered subjectively interesting because this rule is unexpected and may suggest a cross-sell opportunity for the retailer. The incorporation of subjective knowledge into the evaluation of rules can be difficult to do and it requires collaboration with domain experts. The domain experts may serve as the business users or the business intelligence analysts as part of the data science team. The team can communicate the results and decide if it is appropriate to operationalize them.

Diagnostics.

Although the Apriori Algorithm is easy to understand and implement, some of the rules generated are uninteresting or practically useless. Additionally, some of the rules may be generated due to coincidental relationships between the variables measured like confidence, lift and leverage should be used along with human insight to address this problem.

Another problem with association rules is that in phase-3 & 4 of data Analytics lifecycle the team must specify the minimum support prior to the model execution, which may lead to too many or too few rules. In related research a variant of the algorithm can use a predefined targeted range for the number of rules so that the algorithm can adjust the minimum support accordingly.

The Apriori algorithm, which is one of the earliest and the most fundamental algorithms for generating association rule. The Apriori algorithm reduces the computational workload by only examining itemsets that meet the specified minimum threshold. However depending upon the size of the datasets the apriori algorithm can be computationally expensive. For each level of support the algorithm requires a scan of the entire database to obtain the results accordingly.

as the database grows. It takes more time to compute in each run. Here are some approaches to improve Apriori's efficiency.

* Partitioning:-

Any itemset that is potentially frequent in a transaction database must be frequent in at least one of the partitions of the transaction database.

* Sampling:-

This extracts a subset of the data with a lower support threshold and uses the subset to perform association rule mining.

* Transaction reduction:-

A transaction that does not contain frequent k-itemsets is useless in subsequent scans and therefore can be ignored.

* Hash-based itemset counting:-

if the corresponding hashing bucket count of a k-itemset is below a certain threshold the k-itemset cannot be frequent.

* Dynamic itemset counting:-

Only add new candidate itemsets when all of their subsets are estimated to be frequent.

Regression - linear.

linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable. A key assumption is that the relationship between an input variable and the outcome variable is linear. Although this assumption may appear restrictive. it is often possible to properly transform the input or outcome variable is linear. Although this assumption may appear restrictive it is often possible to properly transforming the input or outcome variable to achive a linear relationship between the modified input and outcome variable.

Use cases :-

linear regression is often used in business, government and other scenarios. Some common practical application of linear regression in the real world include the following.

Real estate :-

A simple linear regression analysis can be used to model residential home prices as a function of the home's living area such a model helps set of evaluate the list price of a home on the market. The model could be further improved by including other input variables such as no of bathrooms, number of bedrooms, lot size school district rankings, crime statistics.

Page No.
Date: / /

Demand forecasting:-

Businesses and governments can be use linear regression model to predicts demand for goods and services for example, restaurant chains can appropriately prepare for the predicted type and quantity of food that customer will consume based upon the weather the day of the week, whether an item is offered as a special, the time of day and the reservation volume. Similar models can be built to predict retail sales, emergency room visits and ambulance dispatches.

Medical:-

A liner regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor size. Input variable might include duration of single radiation treatement frequency of radiation treatment, and patient attributes such as age or weight

Logistic - Regression:-

In linear regression modelling, the outcome variable is a continuous variable, As seen in the earlier Income example, liner regression can be used to model the relationship between age and education to income. Suppose a persons actual income was not of interest, but rather weather some one was wealthy or poor. In such case, when the outcome variable is categorical in nature

logistic regression can be used to predict likelihood of an outcome based on the input variables. Although logistic regression can be applied to an outcome variable that represents multiple values the following discussion focuses the case in which the outcome variable represents two values such as true / false, pass / fail or yes / no

Use Cases:

The logistic regression model is a applies to a variety of situations in both the public and the private sector. Some common ways that the logistic regression model is used include the following.

* medical :-

Develope a model to determine the likelihood of a patients successful response to a specific medical treatment or procedure. Input variables could include age weight blood pressure and cholesterol levels.

* Finance :-

Using a loan applicants credit history and the details on the loan, determine the probability that an applicant will default on the loan Based on the prediction the loan can be approved or denied or the terms can be modified

Page No.
Date: / /

Marketing :-

Determine the wireless customer probability of switching carriers (known as churning) based on age, number of family members on the plan months remaining on the existing contract and social network contacts with such insights target the high probability customer with appropriate offer to prevent churn

Engineering :-

Based on operating condition and various diagnostic measurement determine the probability of mechanical part experiencing a malfunction or failure. With this probability estimate schedule the appropriate preventive maintenance activity.

Reasons to choose & Cautions.

linear regression is suitable when the input variable are continuous or discrete include categorical data types but the outcome variable is continuous. If the outcome variable is categorical, logistic regression is better choice

Both models assume a linear additive function of the input variable. If such an assumption does not hold true both regression techniques perform poorly. Furthermore in linear regression the assumption of normally distributed error terms with a constant

Variance is important for many of the statistical inference that can be considered if the various assumption do not appear to hold, the appropriate transformations need to be applied to the data

Although a collection of input variables may be good predictor for the outcome variable the analyst should not infer that the input variables directly cause an outcome. for example it may be identified that those individuals who have regular dentist visits may have a reduced risk of heart attacks. However simply sending someone to the dentist almost certainly has no effect on the persons chance of having a heart attack It is possible that regular dentist visit may indicate a person's overall health and dietary choices which may have direct impact on persons health. This example illustrates the commonly known expression, " Correlation does not imply causation"

Use caution when applying an already fitted model to data that falls outside the dataset used to train dataset for example if income if income was an input variable and the values of income ranged from $35000 to $90000, applying the model to incomes well outside those income could results in inaccurate estimates and predictions.

In a linear regression model, the state of residence could provide a simple additive term to income model but no other impact on the coefficients of the other input variables such as Age and Education However if state does influence the other variables impact to the income model, an alternative approach would be to build 50 seperate linear regression models. one models for each state. Such an approach is an example of the options and decisions that the data scientists must be willing to consider

If several of the input variable are highly correlated to each other, the condition is known as multicollinearity

$$BMI = \frac{weight}{height}$$

where weight is kilograms and height is meters

Additional Regression Models:-

In a case of multicollinearity it may sense to place some restrictions on the magnitudes of the estimated coefficients. Ridge regression. which applies a penalty based on the size of coefficient is one

Techniques that can be applied In fitting a liner regression model the objective is to find the values of the coefficients that minimize the sum of the residuals squared

In ridge regression, a penalty term proportional to the sum of the squares of the coefficient is added to the sum of the residuals squared - Lasso regression is a related modelling technique in where the penalty is proportional to the sum of the absolute values of the coefficients