Name: A. Shiva Surya Saran
Roll No.: BE-B 20

Data Analytics

Assignment 3

Unit 3: Association Rules and Regression

1. Explain Association Rules in details.

Given a large collection of transactions, in which each transaction consists of one or more items, Association rules goes through the items being purchased to see what items are frequently bought together and to discover a list of rules that describe the purchasing behaviour.

The goal with Association rule is to discover list of rules that describe the purchasing behaviour. The goal with Association interest is relationship amongst the items (the relationship that occurs frequently) to be random and his meaning a full form a business perspective which may or may not be obvious.

The relationship that are interesting depend both on business context and the nature of the algorithm being used for the discovery. Each of connected rule is in the form X →Y, meaning that the item X is observed, item Y is also observed in this core, the left hand side of the rule is X and the right hand side of the rule is Y.

2. Explain Apriori algorithm.

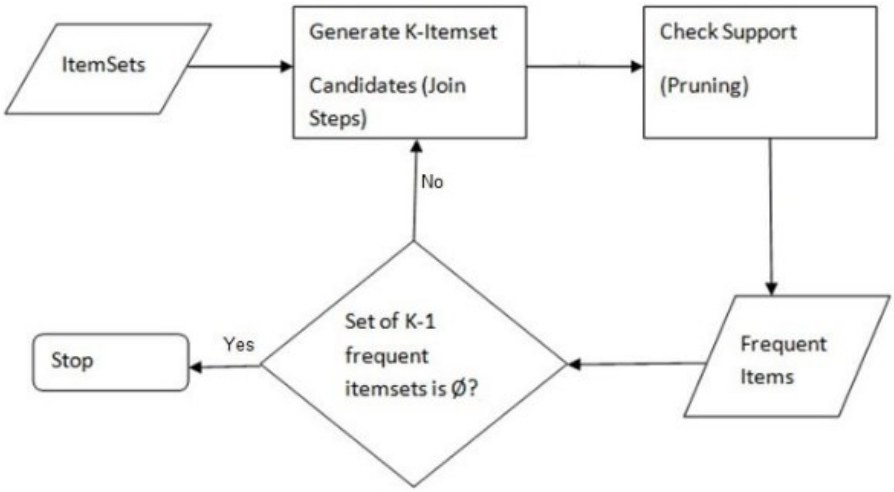
Apriori algorithm Takes a bottom of iteration approach to uncover the frequent item sets by first determining all the possible items and then identifying which among them is the frequent one.

At each iteration, the algorithm checks whether the support criterion on can be met or not. If it can be the algorithm grows the item set repeating the process until it runs out of support or until the itemset reaches up to predefined length.

The Apriori algorithm is given as follows:

1. In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.
2. Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned.

3. Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.
4. The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.
5. The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent, then the superset will be frequent otherwise it is pruned.
6. Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.



3. Explain the evaluation of candidate rules

Frequent item sets from the previous section can form candidate rules such as X implies Y (X => Y). This section discusses how measures such as confidence lift and leverage can help evaluate the appropriateness of these candidate rule.

Confidence Q is defined as the measure of certainty or trustworthiness associated with each discovered rule. Mathematically confidence is the percentage of transaction that contains both X&Y out of all transactions that contain X (See below equation).

Confidence (X → Y) = Support (X & Y) / Support (X)

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

A relationship me be thought 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 on more trustworthy based on the sample data set.

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

Lift (X→Y) = Support (X&Y) / Support X * Support Y

4. Explain Regression details with linear, logistics, reasons to choose and cautions, additional regression models.

Linear regression

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 to be linear. Although this assumption may appear restrictive it is often possible to properly transform the input of the outcome variable to achieve a linear relationship between the modified input and outcome variable.

Use cases:

Linear regression is often used in business, government and other scenarios some comment 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 such as a function of the homes living area. search a model helps set off evaluate the list price of a home on the market.
- *Demand forecasting:*
  businesses and governments can use linear regression model to predict 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 of the day, whether an item is offered as a special, the time of the day and the reservation volume.
- *Medical:*
  A linear regression model can be used analyse the effect of proposed radiation treatment reducing tumour sizes.

Logistic Regression

In linear regression modelling the outcome variable is a continuous variable as seen in the earlier income example, Linear regression can be used to model the relationship between age and education to income. Suppose a person's average income was not of interest but rather than that, to see whether someone was wealthy or poor. In such case, when the outcome variables categorically in native logistic regression can be used to predict likelihood of an outcome based on the input variables. All the logistic regression can be applied do an income variable that represents multiple values the following discussion examines the case in which the outcome variable requests no value such as true/ false, pass/ fail or yes/ no.

Use cases:

The logistic regression model is is applied to a variety of situations in both the public and private sector. Some common ways that the logistic regression model is used include the medical field, in the financial field, in the marketing field and as well as in engineering field.

Reasons to choose and cautions

Linear regression is suitable when the input variables are contagious or discrete including categorical data types, but the outcome variable is continuous and categorical, logistic regression is better choice.

Both models assume linear additive function of the input variable. If such an assumption does not hold prove both regression techniques perform poorly. For the more in linear regression the assumption of normally distribution errors terms with a constant variance is important for many of the statistical inference that can be considered if the various assumptions do not appear to hold appropriate misinformation need to be applied to the data.

All over collection of input variables may be good predictor for the outcome variable the analyst should not infer that the input variable directly causes an outcome. For example, it may be identical 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 person change of having a heart attack. it is possible that regular dentist visits me 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 caution".

Additional Regression Models

In a case of multicollinearity, it may send to play some restrictions on the magnitude of the estimated coefficient. Ridge regression which applies a penalty based on the size of coefficient is one of the techniques that can be applied. In fitting a linear regression model, the objective is to find the values of the coefficient that minimize to form the residual squares.