Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th

Association Rule

Association Rule (Dietrich et al., 2015)

- Association Rule is used to discover a list of rules that describe purchasing behaviour.
- This is for discovering relationships among the items.
- The relationships depend on the business context and the nature of the algorithm being used for the discovery.

Association Rule (Dietrich et al., 2015)

- Association rules are sometimes referred to as the market basket analysis.
- Each transaction can be viewed as the shopping basket of the customer that contains one or more items.

Association Rule (Dietrich et al., 2015)

- Itemset refers to a collection of items that contain some relationship.
- Each itemset contains *k* items (*k*-itemset).
- Frequent itemset refers to items that often appear together (at least minimum support)
- Example,
    - Minimum support = 0.5 means any itemset can be considered a frequent itemset if at least 50% of the transactions contain this itemset.

Apriori Algorithm (Dietrich et al., 2015)

- Apriori algorithm takes a bottom-up iterative approach by first determining all the possible items (1-itemset such as {bread}, {milk}, ...} and identifying which of them are frequent.
- The itemsets that have at least a minimum support threshold will be kept.
- In the next iteration, the frequent 1-itemsets are paired into 2-itemsets such as {bread, milk}

Candidate Rules (Dietrich et al., 2015)

- Confidence measures of certainty associated with each discovered rule.
  - Confidence(X → Y) = Support(X $\wedge$ Y) / Support(X)

    For example, if {bread, eggs, milk} has a support 0.15 and {bread, eggs} also has a support 0.15. Therefore, the confidence of {bread, eggs} → {milk} is 1.

Candidate Rules (Dietrich et al., 2015)

- Lift measures how many times more often X and Y occur together than expected if they are statistically independent of each other.
  - Lift $(X \rightarrow Y)$ = Support$(X \wedge Y)$ / Support(X) × Support(Y)

    For example, if there are 1,000 transactions, {milk, eggs} appears in 300, {milk} appears in 500, and {eggs} appears in 400.
    Therefore, Lift(milk → eggs) = 0.3/(0.5 × 0.4) = 1.5

Candidate Rules (Dietrich et al., 2015)

- Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.
  - $\text{Leverage}(X \rightarrow Y) = \text{Support}(X \wedge Y) - \text{Support}(X) \times \text{Support}(Y)$

    For example, if there are 1,000 transactions, {milk, eggs} appears in 300, {milk} appears in 500, and {eggs} appears in 400.
    Therefore, Leverage(milk → eggs) = 0.3-(0.5 × 0.4) = 0.1

Association Rule Implementation

- Use the groceries_data.csv file as a dataset
- Assume we want to predict the item that would be purchased if the basket contains:
  - [vegetable juice, frozen fruits, packaged fruit]
  - [mayonnaise, butter, buns]

Association Rule Implementation

Import Libraries:

- SparkContext and SparkSession
- FPGrowth from pyspark.ml.fpm

Association Rule Implementation

- Create SparkSession
- Read data
- Group data based on member number
- Show data
- Add a column basket where it contains a list of items
- Show data

Association Rule Implementation

- Create FPGrowth
  - fp = FPGrowth(minSupport = <value>, minConfidence = <value>, itemsCol = ‘basket’, predictionCol = <column name>)
- Fit the created FPGrowth into a model
- Show frequencies using <model>.freqItemsets.show(<value>)

Association Rule Implementation

- Filter the rules
  - <model>.associationRules.filter(<model>.associationRules.confidence > <value>)
  - Show data
- Create a dataframe <dataframe> = [‘basket’]

Association Rule Implementation

- Add new data to be used for the predictions
- Create parallelise RDD
- Use the created Dataframe from RDD
- Show data
- Transform the model using the Dataframe of new data

Assignment (1 point)

- Please implement the association rule analysis and show the results to get 1 point.
- The results include: 5 Dataframes

References

- Dietrich, D., Heller, B., & Yang, B. (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*. Wiley.