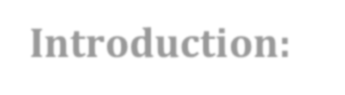
Name: \_\_\_\_\_\_\_\_\_\_\_Archit Jain\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Lab #4: Association Rule Mining**

Submit **a scan copy of answered this Lab #4** to the Lab #4 Drop Box by Friday, 4/3 @11:59PM.

**Introduction:**



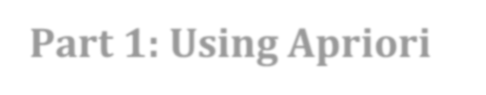
In this lab we will explore association rule mining algorithms in Weka and evaluate them. The tasks include understanding the dataset and implementation of the Apriori algorithm in Weka. It is assumed that you have a basic knowledge of the Apriori algorithm.

To describe briefly, this algorithm is a divide and conquer approach. The implementation works on the concept of maximum support level, which is also referred as confidence level. The algorithm works iteratively and identifies a rule based on the confidence level. For rule generation in subsequent iterations, the dataset that is not covered in the previous rules will be used. For detailed description and pseudo code, refer to study material.

**Dataset used: *supermarket*** dataset(Check data folder under Weka installation)

It is a real-world data set collected over a short duration of time at a supermarket. The attributes are aggregated to the department level. The value “t” indicates that customers shopping cart contained at least one product from that department.

**Part 1: Using Apriori**



Open Weka and navigate to Explorer window.

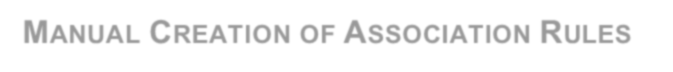
**Q1.1** In the Preprocess tab and select the *supermarket* dataset. Navigate to Associate tab and select and choose ‘Apriori’ in associations folder. Run Apriori using the default settings of the options. The last 10 lines of the output below ‘Best rules found’ are the 10 best rules generated. The confidence of rule 1 is 0.92. How was this confidence value computed? Write down the proportion as division.

\_\_\_723\_\_\_\_\_ / \_\_\_788\_\_\_\_\_

Q1.2 How many instances is the support of rule 7? 701

|  |  |
| --- | --- |
| Q1.3 | Based on rule 10: |
|  | LHS (frozen foods=t fruit=t total=high) ==> RHS (bread and cake=t), are LHS and RHS independent? Justify your answer using a proper metric value. |
|  | Metric Used: lift |
|  | Metric Value: lift:(1.26) |
|  | Justification:  LHS and RHS are not independent as lift is greater than 1. Hence its very likely RHS is purchased when the LHS is purchased. |
| Q1.4 | Using the available Apriori algorithm options, design a method to find the two rules with the highest support count. Your answer should include which options you modified, the values you modified the options to, the resultant rules, and their support count.  Options modified are:  metric type = Lift  min metric = 1.0  numRules = 2  Best rules found:  1. bread and cake=t 3330 ==> milk-cream=t 2337 conf:(0.7) < lift:(1.1)> lev:(0.05) [221] conv:(1.22)  2. milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8) < lift:(1.1)> lev:(0.05) [221] conv:(1.37)  Support count are 2337 for both rules. |
| Q1.5 | What does ‘best rules’ mean? What criterion is used to decide what the best rules are? (Hint: Study the descriptions of the parameters for Apriori by pressing the ‘More’ button in the window that allows you to change the options for ‘Apriori’.) |
|  | Best Rules are the association rules in Apriori which are created and ranked based on the metric selected. The top *n* rules are generated which start from upper minimum-support and incrementally decreased to lower minimum-support threshold value.  Lift should be used as criterion to select the best rule. As confidence metric can give a false positive value as it only choses the most popular association rule. |

**Part 2:** **MANUAL CREATION OF ASSOCIATION RULES**



*Market basket Analysis:*

*The table below shows details of 5 transactions at a supermarket.*

|  |  |
| --- | --- |
| **TID** | **Items** |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Bread, Milk, Diaper, Beer, Coke |
| 4 | Bread, Diaper, Beer |
| 5 | Bread, Diaper, Coke |

Let’s make a tabular using binary representation of the above where A = Bread; B= Beer; C=Coke, D=Diaper; E=Egg; F=Milk:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TID | A | B | C | D | E | F |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 0 | 1 | 1 | 0 |
| 3 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 1 | 1 | 0 | 0 |

Q2.1 Form the item sets. Let’s start by forming the item set containing one item. Write the number of occurrences and the support of each item set. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | # Occurrences | Support  (%) |
| A | 5 | 100 |
| B | 3 | 60 |
| C | 2 | 40 |
| D | 4 | 80 |
| E | 1 | 20 |
| F | 2 | 40 |
|  |  |  |

Q2.2 Now let’s form the item sets containing 2 items. We only take the item sets from Q2.1 whose minimum support is 60%. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | # Occurrences | Support  (%) |
| AB | 3 | 60 |
| AD | 4 | 80 |
| BD | 3 | 60 |
|  |  |  |
|  |  |  |
|  |  |  |

Q2.3 List item sets containing 3 items from Q2.2. (Note: All rows might not be necessary)

|  |  |  |
| --- | --- | --- |
| Item set | # Occurrences | Support  (%) |
| ABD | 3 | 60 |
|  |  |  |
|  |  |  |

Q2.4 Now form the rules and calculate their confidence (c) using minsup = 60%. (Note: All rows might not be necessary)

|  |  |
| --- | --- |
| Rules | Confidence  (c) |
| A->B | 3/5 = 60% |
| B ->A | 3/3 = 100% |
| A->D | 4/5 = 80% |
| D->A | 4/4 = 100% |
| B->D | 3/3 = 100% |
| D->B | 3/4 = 75% |
| AB->D | 3/3 = 100% |
| D->AB | 3/4 = 75% |
| A->BD | 3/5 = 60% |
| BD->A | 3/3 =100% |
| B->AD | 3/3 =100% |
| AD->B | 3/4 = 75% |
|  |  |

Q2.5 Prune the rules in Q2.4 using confidence measure of 80%. (Note: All rows might not be necessary)

|  |  |
| --- | --- |
| Rules | Confidence(c) |
| B ->A | 3/3 = 100% |
| A->D | 4/5 = 80% |
| D->A | 4/4 = 100% |
| B->D | 3/3 = 100% |
| AB->D | 3/3 = 100% |
| BD->A | 3/3 =100% |
| B->AD | 3/3 =100% |
|  |  |

LAB 4 Gradesheet Name: \_\_\_\_Archit Jain\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |  |
| --- | --- | --- | --- |
| Questions | Max. Points | Points Earned | Comments |
| Part 1 |  |  |  |
| Q1.1 | 10 |  |  |
| Q1.2 | 10 |  |  |
| Q1.3 | 10 |  |  |
| Q1.4 | 10 |  |  |
| Q1.5 | 10 |  |  |
| Part 2 |  |  |  |
| Q2.1 | 10 |  |  |
| Q2.2 | 10 |  |  |
| Q2.3 | 10 |  |  |
| Q2.4 | 10 |  |  |
| Q2.5 | 10 |  |  |
| TOTAL | 100 |  |  |