Jay Patel
CS422 - Data Mining
09/28/18

ASSIGNMENT 3 Association Rule

Part 1: Weka

1.1 Supermarket Dataset

Before addressing the Apriori Association Algorithm, Lets look at the data set. There are 217 attributes in total, and 4627 instances. There are many attributes that has no data in them, so they are empty. The data type for all attributes in Nominal. The Dataset can be concluded to be a supermarket's point of sale, and each instance represent a transaction, department, and products involved. We will be using the association rule to analyze the pattern of *what items are purchased together*.

We would be using the Apriori Association Algorithm to find the association rules for this dataset. In a nutshell, the algorithm creates 'Attribute - Item' sets that maximizes trough-out the the data. Press the start button on the attribute tab to run the algorithm. The top 4 association rules on default parameters are :

```
Best rules found:

 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723
<conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696
<conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705
<conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746
<conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
```

We can notice that bread and cake are the most bought together with certain products such as biscuits, frozen fruits, baking needs, and vegetables. Some of the key things to notice in the Attribute Output (Not everything mentioned here), the algorithm starts at 100% minSupport and stops at 15% after running 17 instances. It can be noticed in the association rule that all the transaction total for top rules is high, and we can see that the default association cut off is 0.92. observing the rules, we can conclude that we can convince people that buy frozen food, and biscuits to also buy break and cake obtaining high transaction value. Here, we can emphasis on parameters such as **minMetric** and **Delta** are crucial in changing the form of association rules. and we can device certain settings to achieve accurate results.

lets recalibrate the parameters, and decrease the confidence level. I changed the min. metric to 0.5(low) and re-run the algorithm. 11 cycles of instances were performed, and min support received was 0.45. The rules changes as well, since now, each attribute is one on one associated with bread and cake. Also, the cut off (conf.) value is at 0.78.

Changing parameters to 0.75 (mid), we get 13 cycles were performed, and min. support received was 35%. First thing to notice in the Association Rules, is that the related items to the association changes. We get to see that the top two association has milk-cream and fruits as antecedent and also, covered through out half of the total instances (4627). Lets change the settings to 0.95(high). Keeping a high confidence value puts a lot of work on the machine, and it took lot of time in order to build the model. The number of cycles for which, the algorithm ran increased to 19. and the min. support dropped from 15% to 5%. The top Association Rules are similar to the ones received in default parameters. However, the rules appear to be more detailed, and with varied product list, which can give a high total, as well as an attribute that a customer would likely buy along with it.

1.2 Pruning Attributes

After removing the top occurring attributes from the data set, we removed 7 of them, and the Association Output completely changed. At the default setting of the algorithm, we observe that the model runs for 19 cycles which is higher than the regular model's. The min. support drops to 5%. Another thing to notice in the association rules is that, now we see that, a certain type of food when bought together would give a high transaction total. There is no other attribute provided on the right hand side of '==>' .

Performing same experiments as previous part. I changed the min. support metric to 0.5(low), 0.75(mid) and 0.95(high). In the low settings, The pruned data set run for 13 cycles, and provides min support of 3%. The top rules that were found has conf. cut off of 0.68 and contained, milk-cream as a product that was likely to be bought with certain food items. At medium settings, the cycle ran for 18 instances, and the min support dropped to 1%. The high totals also dropped compared to previous experiments. The association items also differs from previous experiment, as well as the low setting rules. At high setting, the pruned data set returned zero rules. This could be due to high confidence level, and no attribute that was spread out to meet the algorithm's requirement.

1.3 Vote Dataset

The voting data set was also used in the previous assignment, where we used the decision tree classifier in order to analyze the data set. In this part, we would compare the results we achieved from the previous assignment, with the association rules we obtained during or test/experiment. In the decision Stump classifier, the tree was build off of 'physician-fee-freeze' attribute of the data set, and we could also notice that there were larger number of instances for democrats. Our top three Association rules conveys that if attributes - adoption of the budget, physician fee freeze and aid to Nicaraguan follow the nominal pattern it is likely to favor Democrat party.

Part 2: Short Answers

2.1 Table 1 : GINI Index

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

The Gini Index is calculated by following:

$$GINI = 1 - \sum_j [p(j/t)]^2$$

a) Gini index for all examples.

All the examples are distributed into two classes. C0 and C1. with total of 20 examples, C1 = 10 and C0 = 10. $p(c_1) = \frac{10}{20}$ $p(c_2) = \frac{10}{20}$. Therefore, $G_{index} = 1 - \frac{1}{4} - \frac{1}{4} = 0.5$. This answer is quite obvious by just observing the table, since the split between classes is at equal 10 and 10.

b) Gini index for Customer Id

The Gini index for the Customer id is 0 since each ID is distinct and hence there will be no

c) Gini index for Gender

The Gini index for male as well as female (since num Male = num Female) is :

$$G_{male,female} = 1 - \frac{6}{10} - \frac{4}{10} = 0.48$$

$$G_{index} = \frac{10}{20} \times 0.48 + \frac{10}{20} \times 0.48 = 0.48$$

d) Gini index for Car Type

There are 4 Family cars, 8 sports car, and 8 luxury cars. We will use multiway split to calculate car type index.

$$G_{family} = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$$

$$G_{sport} = 1 - (\frac{8}{8})^2 - (\frac{0}{8})^2 = 0.0$$

$$G_{luxury} = 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.2188$$

$$G_{index} = \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0.218 = 0.1625$$

e) Gini index for Shirt Size
There are 5 small shirt , 7 medium, 4 Large, 4 Extra Large. We will use multiway split to calculate :

$$G_{small} = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.48$$

$$G_{medium} = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.4898$$

$$G_{large} = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$$

$$G_{extralarge} = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$$

$$G_{index} = \frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.4898 + \frac{4}{20} \times 0.5 + \frac{}{20} \times 0.5 = 0.4914$$

f) Conclusion: As calculated the Gini Index for the car types is the least, and hence that would become the first split of our data set. It would provide a better idea, of which class it belongs to. A Particular attribute is useful or not, can be determined by the *overfitting* factor. Here, we saw that customer ID can not be used as one of the attributes, because the its arbitrary and it will change every time a new customer is added. If we also look at the relation between the type of cars and the gender, we can see that many of the male prefer sport cars, while female prefers family cars, and we can also see the relation on the sizes of shirt associated with the gender.

2.2 Table 2 : Entropy

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

The Entropy is calculated by

$$Entropy(t) = -\sum p(j|t) log_2 p(j|t)$$

a)Entropy with respect to class "+".
there are 4 '+' class and 5 '-' class. So the entropy will be calculated as:

$$Entropy('+') = -\frac{4}{9} log_2 \frac{4}{9} - \frac{5}{9} log_2 \frac{5}{9} = 0.9911$$

The entropy calculation are similar based on gini computation as well. Since the value of our entropy is closer to 1, we can conclude that the data set is evenly distributed. also, there is least information that is available, and its not homogeneous.

b) Entropy for a1 is

$$\frac{4}{9}[\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4}] + \frac{5}{9}[\frac{1}{5} log_2 \frac{1}{5} - \frac{4}{5} log_2 \frac{4}{5}] = 0.7616$$

Hence, the information gain for a1 is $0.9911 - 0.7616 = 0.2294$.

a1	+	-	entropy

T	3	1	0.8115
F	1	4	0.7211

The Entropy for a2 is

$$\frac{5}{9}[\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}] + \frac{4}{9}[\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}] = 0.7616$$

Hence the information gain for a2 is $0.9911 - 0.9839 = 0.0072$.

a2	+	-	entropy
T	2	3	0.0970
F	2	2	1

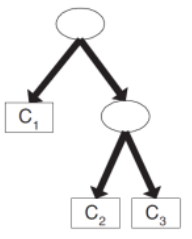
Using the following information we can create a table for a1 such as:

P(+) for T	3/4
P(+) for F	1/5
P(-) for T	1/4
P(-) for F	4/5

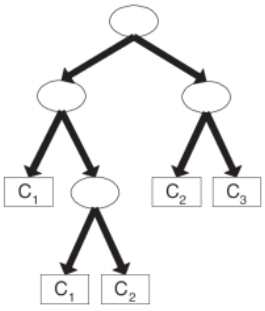
We can use the probability for T for class '+', probability for F for class '-' total number of both '+' and '-' instances in the equation to get the entropy.

c) The error rate for a1 is 2/9 while the error rate for a2 is 4/9 therefore a1 is a better option for split. Due the information gain, it is clear from part (B) that the information gain from a1 is much higher than from a2, and hence the error rate of a1 is smaller and it is a better choice.

2.3 Decision Trees



(a) Decision tree with 7 errors



(b) Decision tree with 4 errors

a)MDL Principle: The given tree has 16 attributes, and the cost for each internal node in the decision tree is

$$log_2(m) = log_2(16) = 4$$

also, there are 3 distinct classes, therefore, for each leaf node. the cost is :

$$log_2(k) = log_2(3) = > log_2(4) = 2$$

Total description length of the given tree can be calculated using :

Cost(tree, data) = Cost(tree) + Cost(data | tree)

Cost(tree) is the cost of encoding the nodes in the tree. In order to simplify the solution, we can assume that the total cost can be obtained by adding up the cost encoding each internal node, and each leaf node.

Considering the errors as well at the rate of logn.
Tree A is $2 \times 4 + 3 \times 2 + 7 \times log(n) = 14 + 7log(n)$,
and for Tree B $4 \times 4 + 5 \times 2 + 4 \times log(n) = 26 + 4log(n)$. Hence, we can conclude that A is better than B if and only if $n > 16$.

b) Post - Pruning process refers to the process of making (growing) the entire decision three and removing nodes for which we do not have a lot of information. Lets consider a Node S , we can prune the children of S if all the children are leaves, and the accuracy on the validation set does not decrease if we assign class label to all the items of S. There are several methods for evaluating trees to prune.
one of the most efficient method would be the ten fold cross validation. The data is split into 10 or N equal data set and only part of it, is a test data, which will provide an average classification .