# Homework 3



## Association Rule Mining

1. Trace the results of using the Apriori algorithm on the grocery shop with support threshold 33.34% and confidence threshold 60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

| Transaction ID | Items |
|---|---|
| T1 | HotDogs, Buns, Ketchup |
| T2 | HotDogs, Buns |
| T3 | HotDogs, Coke, Chips |
| T4 | Chips, Coke |
| T5 | Chips, Ketchup |
| T6 | HotDogs, Coke, Chips |

2. Trace the results of using the Apriori algorithm on the computer shop with support threshold 70% and confidence threshold 80%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

| Transaction ID | Items |
|---|---|
| T1 | Tri-pod, Lens, bag |
| T2 | Camera, Lens, bag |
| T3 | Camera, Tri-pod, Lens, Memorycard |
| T4 | Camera, Tri-pod, Lens, bag |
| T5 | Lens, Memorycard, bag |

3. Describe the important of support and confidence thresholds in finding association rules ? And what should be their most appropriate values ?

## Association Rule Mining with WEKA

Apriori works with categorical values only. Therefore, if a dataset contains numeric attributes, they need to be converted into nominal before applying the Apriori algorithm. Hence, data preprocessing must be performed. Repeat homework 2 (Data Preprocessing), if you don't know how to deal with numeric to nominal conversion.

- weather.nominal.arff
  1. Load weather.nominal.arff into a text editor and analyze the attribute types and values.
  2. Is this dataset appropriate for association rule mining ? if not, modify it. You may use WEKA's "Preprocessing" capability.
  3. Apply Apriori algorithm to the dataset.
     a. Goto Association tab
     b. Choose Apriori as Associator
     c. Accept all default values. You may click on More button to see the synopsis for the different parameters.

d. Click on Start button to run

4. Study the output in the right panel. It should look something similar to the following :

```
Apriori
=======

Minimum support: 0.15
Minimum metric : 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6

Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
 4. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
 5. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 8. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
 9. outlook=sunny temperature=hot 2 ==> humidity=high 2    conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    conf:(1)
```

5. Can you explain what the output says ?
6. Try vary value of parameters; for example, minimum support, minimum confidence and number of rules.
7. What do you find ?

## WEKA's Apriori (*ref: web.mac.com)*

The default values for *Number of rules*, the decrease for *Minimum support* (delta factor) and *minimum Confidence* values are 10, 0.05 and 0.9. Rule *Support* is the proportion of examples covered by the LHS and RHS while *Confidence* is the proportion of examples covered by the LHS that are also covered by the RHS. So if a rule's RHS and LHS covers 50% of the cases then the rule has 0.5 support, if the LHS of a rule covers 200 cases and of these the RHS covers 50 cases then the confidence is 0.25. With default settings Apriori tries to generate 10 rules by starting with a minimum support of 100%, iteratively decreasing support by the delta factor until minimum non-zero support is reached or the required number of rules with at least minimum confidence has been generated. If we examine Weka's output, a *Minimum support* of 0.15 indicates the minimum support reached in order to generate the 10 rules with the specified minimum metric, here confidence of 0.9. The item set sizes generated are displayed; e.g. there are 6 four-item sets having the required minimum support. By default rules are sorted by confidence and any ties are broken based on support. The number preceding ==> indicates the

number of cases covered by the LHS and the value following the rule is the number of cases covered by the RHS. The value in parenthesis is the rule's confidence.

- bank.arff
  1. Load bank.arff into a text editor and analyze the attribute types and values.
  2. Is this dataset appropriate for association rule mining ? if not, modify it. You may use WEKA's "Preprocessing" capability.
  3. Apply Apriori algorithm to the dataset.
  4. Study the output in the right panel.
  5. Check out output from various different sets of parameters.
  6. Is it something you expected ?

- market-basket.arff
  1. Perform similar steps against market-basket.arff.

---

You don't have to turn in anything. However, be prepared to discuss results and findings in class individually. I will randomly call you guys to give explanation.