# Association Rule Mining

## Ali Nauman (18L-1863) & Ammar Hussain (18L-1834)

## Assignment – 4

### Association Rule Mining

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.[1] This rule-based approach also generates new rules as it analyzes more data.

### Data set

We have used the Bank data set as provided for the previous assignments. We have used the smaller version, containing 10% examples (4521), randomly selected from the full data set.

### Pre-Processing

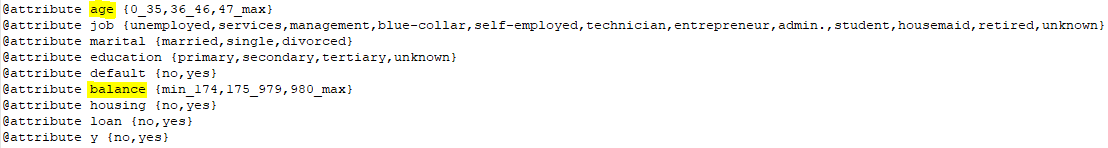
**Missing Value** – There were no missing attributes in the dataset. This was clearly stated in the read-me document that came along the dataset as well.

**Data Reduction** – There were a number of attributes that were meaningless and had no impact on the final version. Below are the attributes that were removed,

* *contact: contact communication type (categorical: "unknown","telephone","cellular")*
* *day: last contact day of the month (numeric)*
* *month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")* – This was included in the first instance. However, the rules generated were of less lift and were more month oriented.
* *duration: last contact duration, in seconds (numeric)*
* *campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)*
* *pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)*
* *previous: number of contacts performed before this campaign and for this client (numeric)*
* *poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")* – Unknown values accounted for ‘3705’. Thus, evaluating based on this attribute would result in

**Discretization** – Now we are left with two attributes that are numeric. Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes. There are 2 such attributes in this data set: *“age”*, *“balance”*. We will rely on WEKA to perform discretization on the *"age"* and *"balance"* attributes. In this example, we divide each of these into 3 bins (intervals). Furthermore, we have selected the option of *“Equal Frequency”* to be true.

Please find below the final processed attribute list. We have saved a *‘.csv’* version of this to be used in the Map-Reduce part of the assignment. Also, now we shall apply ***Association Rule Mining*** on this data.



### Association Rule Mining

WEKA allows the resulting rules to be sorted according to different metrics such as confidence, leverage, conviction and lift. For this data set, we have selected High Lift and High Confidence as the two criteria to generate the best rules.

We have entered 1.2 as the minimum value for lift (or improvement) is computed as the confidence of the rule divided by the support of the right-hand-side (RHS). In a simplified form, given a rule L => R, lift is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R.

Confidence is an indication of how often the rule has been found to be true. Initially, we generated the rules with minimum confidence as 0.8. However, for almost all the best rules generated by this the value of the lift was ‘1’. Given that the value of the lift = 1, transcends that X and Y are independent. Thus, we continued to decrease the value of the confidence, in pursuit to increase the value of the Lift more than 1. Lift value more than ‘1’, translates that X and Y are positively correlated. Also, we needed to make sure that we did not decrease the value of confidence way too much. As confidence, is a measure of how often items in Y appear in transactions that contain X.

For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction. Note that leverage and lift measure similar things, except that leverage measures the difference between the probability of co-occurrence of L and R as the independent probabilities of each of L and R.

In other words, leverage measures the proportion of additional cases covered by both L and R above those expected if L and R were independent of each other. Thus, for leverage, values above 0 are desirable, whereas for lift, we want to see values greater than 1. Finally, conviction is similar to lift, but it measures the effect of the right-hand-side not being true. It also inverts the ratio.

### Best Rules and Explanation

We have included all the best rules generated via WEKA. Given, the top 5 most interesting rules are most likely not the top 5 in the result set of the Apriori algorithm. They are rules that, in addition to having high support, lift, and confidence, also gives some non-trivial, useful information based on the underlying business objectives. Thus, below are a few of the best rules generated and their explanations.

*Best Rules*

All the rules generated must comply with the fact that the client has subscribed a term deposit or not.

**marital=married 2797 ==> y=no 2520 <conf:(0.9)> lift:(1.02) lev:(0.01) [45] conv:(1.16)**

This is a very important rule with a lift of 1.02, signifying that the married people mostly do not subscribe for a term deposit. There are multiple rules also vindicating this phenomena.

**marital=married default=no 2761 ==> y=no 2487 <conf:(0.9)> lift:(1.02) lev:(0.01) [44] conv:(1.16)**

This is also a very similar rule yet it also strongly says that the married people do not have credit in default and thus also do not subscribe to a term deposit.

**marital=married default=no loan=no 2320 ==> y=no 2069 <conf:(0.89)> lift:(1.01) lev:(0) [16] conv:(1.06)**

Overall we can conclude from this assumptions of rules which have very high confidence and high lift that married people tend to take very low risks and do not tend to divulge into loans and keep it balanced.

**education=secondary default=no 2260 ==> y=no 2021 <conf:(0.89)> lift:(1.01) lev:(0) [21] conv:(1.09)**

**education=secondary default=no loan=no 1858 ==> y=no 1643 <conf:(0.88)> lift:(1) lev:(-0) [0] conv:(0.99)**

People with good educational background rarely take loan as these rules imply that they are self-sufficient. The confidence of these rules are high. Along with the lift, which signify that they are positively correlated.

**job=management 969 ==> education=tertiary 787 conf:(0.81) < lift:(2.72)> lev:(0.11) [497] conv:(3.71)**

**job=management default=no 955 ==> education=tertiary loan=no 680 conf:(0.71) < lift:(2.74)> lev:(0.1) [431] conv:(2.56)**

People with management jobs are with the highest education and thus are more than self-sufficient and do not require any loan from the bank. These rules have very high confidence and very high lift, signifying positive correlation between them.

**age=47\_max default=no 1456 ==> marital=married loan=no y=no 817 conf:(0.56) < lift:(1.21)> lev:(0.03) [143] conv:(1.22)**

**age=47\_max 1481 ==> marital=married default=no loan=no y=no 817 conf:(0.55) < lift:(1.21)> lev:(0.03) [139] conv:(1.21)**

These rules signify that people falling in the age group of 47-max are mostly married people with low risk. They rarely take a loan under extreme circumstances and also they have not subscribed for a term deposit.

**housing=yes 2559 ==> y=no 2339 <conf:(0.91)> lift:(1.03) lev:(0.02) [74] conv:(1.33)**

People who have taken the housing loan, do not tend to subscribe for a term deposit. This is genuine as the loan itself is a big thing in itself.

**default=no housing=yes 2514 ==> y=no 2297 <conf:(0.91)> lift:(1.03) lev:(0.02) [72] conv:(1.33)**

This is a very positive rule. Firstly, the count of people with housing loan are very good on their installments and are thus not default in payments. This count is very high as compared to the total entries. Secondly, these people do not subscribe to term deposit, and thus are intelligent enough to not divulge into more issues.

### References:

1. Big Data Analytics, Association Rule Mining Slides by Dr. Zareen Alamgir
2. <https://en.wikipedia.org/wiki/Association_rule_learning> - Wikipedia explanation of the Association Rule Mining
3. <http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/preprocess.html> WEKA tutorial for Data Preprocessing
4. <http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/associate.html> WEKA tutorial for Associate Rule Mining