Laboratory 5 – Association rules

1. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.
2. Following the original definition by Agrawal et al. the problem of association rule mining is defined as follows. Let $I=\{i_1,i_2,\ldots,i_n\}$ be a set of *n* binary attributes called *items*. Let $D=\{t_1,t_2,\ldots,t_m\}$ be a set of transactions called the *database*. Each transaction in *D* has a unique transaction ID and contains a subset of the items in *I*. A *rule* is defined as an implication of the form X→Y, where $X, Y \in I$and $X \cap Y = \emptyset$. The sets of items (for short *itemsets*) *X* and *Y* are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.
3. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.
    a. The *support supp(X)*of an itemset *X* is defined as the proportion of transactions in the data set which contain the itemset.
    b. The *confidence* of a rule is defined as: $conf(X \to Y) = \frac{supp(X \cup Y)}{supp(X)}$, where supp(X∪Y) means "*support for occurrences of transactions where X and Y both appear*".
4. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:
    a. First, minimum support is applied to find all *frequent itemsets* in a database.
    b. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

    Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations). The set of possible itemsets is the power set over *I* and has size 2^n-1 (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items *n* in *I*, efficient search is possible using the *Apriori* property, which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its supersets must also be infrequent.

    Many algorithms for generating association rules were presented over time. The best-known algorithm to mine association rules is Apriori. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*), and groups of candidates are tested against the data. It generates candidate item sets of length *k* from item sets of length *k-1*. Then it prunes the candidates which have an infrequent sub pattern. After that, it scans the transaction database to determine frequent item sets among the candidates. The algorithm terminates when no further successful extensions are found.

    Next, the frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database.
5. One of the most common domain of application of association rules is market basket analysis.