Universidad de La Sabana Felix Mohr
Facultad de Ingenería 2020.2

Business Intelligence

- Sheet 8 -

Exercise 1 (4 Points)

1. Show the following:
   a) if X, Y are itemsets and $X \subseteq Y$, then $\sup(X) \geq \sup(Y)$
   b) if X, Y are itemsets and $X \subseteq Y$, then $\mathbf{t}(X) \supseteq \mathbf{t}(Y)$
   c) if S, T are tidsets and $S \subseteq T$, then $\mathbf{i}(S) \supseteq \mathbf{i}(T)$
2. For a combination of two tidsets T_1, T_2, derive a criterion that is true iff Eclat is preferable over dEclat in terms of runtime when creating the child node.
3. Given two k-itemsets $X_a = \{x_1, .., x_{k-1}, x_a\}$ and $X_b = \{x_1, .., x_{k-1}, b\}$ that share the common prefix $(k-1)$-itemset $X = \{x_1, x_2, .., x_{k-1}\}$ as a prefix, prove that $sup(X_{ab}) = sup(X_a) - |\mathbf{d}(X_{ab})|$ where $X_{ab} = X_a \cup X_b$ and $\mathbf{d}(X_{ab})$ is the diffset of X_{ab}.
4. Let Z be an itemset and $\emptyset \neq X, Y \subset Z$ such that $Y = Z \setminus X$. Suppose that the rule $X \longrightarrow Y$ is not strong. Show that then every rule $X' \longrightarrow Y'$ with $X' \subset X$ and $Y' = Z \setminus X'$ is also not strong.

Exercise 2 (4 Points) In this exercise, we will use the transaction view on itemsets. You can use the given implementations to read in a transaction database and set utilities such as computing the difference of sets/lists. Note that we treat itemsets as python lists (not numpy arrays). In the following `db` refers to a *list* of *lists* (a list of transactions, which are encoded as lists).

1. Write functions `apriori(db, minsup)`, `eclat(db, minsup)`, and `declat(db, minsup)` that run the respective algorithms and return the set of frequent itemsets (with threshold `minsup`). Run the three algorithms on the shop dataset and report the runtimes.
2. Write functions `createAssociationRules(fsets, minsup)` and `getStrongRules(db, minsup, minconf)`. The first should implement the AssociationRules algorithm and compute all association rules with a given support from a set of frequent sets. The second should compute all frequent sets and then derives the strong rules from it with respect to the given minimum confidence.

   Your boss is interested in a list of strong rules of the shop dataset. He considers rules strong if their support is at least 500 and their confidence is at least .75. Report a list of such strong rules sorted by support. He is also interested in a sub-list of these strong rules in which at least two items appear in the conclusion of the rule. Report also this sub-list, ordered descendingly by the number of items in the conclusion. State explicitly, e.g. in a markdown cell, what the first of these rules means, and explain what the respective support and confidence value of that specific rule mean.