Data Mining Homework 2

Discovery of Frequent Itemsets and Association Rules

Group 22 - Blanca Bastardés Climent, Alice Ciallella

Short description

The implementation of this assignment has been done in Scala using Spark. All the code is provided in the notebook **HW2.ipynb**, in which the following functions can be found:

- **apriori_alg:** it implements the Apriori algorithm that finds frequent itemsets given a support threshold; for this task a support of 1000 was chosen, which corresponds to 1% of the baskets provided. The function first calculates the frequent singletons and, through recursive calls, search for all the frequent sets with size > 1. The output is a list of the frequent itemsets.
- **association_rule:** it investigates all the possible association rules given an itemset, and prints the ones whose confidence is above a certain threshold **c**. For this application a value of 0.50 has been chosen.

Instructions to run the code

We provide the following folders and files:

- **src/HW2.ipynb:** A Spark notebook containing code for finding frequent itemsets and association rules from a sales transaction database.
- **data/T10I4D100K.dat:** provided file for the task.

Results

The output of the **apriori_alg** is the following:

(217, 346), (829, 789), (829, 368), (825, 704), (825, 39), (682, 368), (704, 39), (722, 390), (390, 227), (825, 704, 39)

The output of the **assiciation_rule** is the following:

{(704)} -> {(825)}: 0.6142697881828316
{(704)} -> {(39)}: 0.617056856187291
{(227)} -> {(390)}: 0.577007700770077
{(704)} -> {(825, 39)}: 0.5769230769230769
{(825, 704)} -> {(39)}: 0.9392014519056261
{(825, 39)} -> {(704)}: 0.8719460825610783
{(704, 39)} -> {(825)}: 0.9349593495934959