Data Mining: Frequent Itemsets and Association rules

Braulio Grana Gutiérrez Adrián Ramírez del Río

November 20, 2016

1 Description

For this assignment we had to implement the Apriori algorithm to find frequent itemsets with at least a given support in a dataset of sales transactions. As a bonus task we developed the generation of association rules with a given confidence. We implemented our solution to this exercise using Python3.

1.1 Frequent itemsets

We created a class named *Frequent_items* that takes a file path to the dataset as parameter in its constructor. Then we compute the first iteration of candidate items and filter them using the support threshold.

Then, we keep generating iteration of candidates increasing their length by one in every iteration and filter them by their support and the support of all their subsets.

Finally, we stop iterating when no more candidates are generated.

1.2 Association rules

In this bonus task we added a method to the original class to get the rules of association that have a level of confidence above the given threshold according to the following formula:

$$confidence(X \rightarrow Y) = \frac{support(X \bigcup Y)}{support(X)}$$

For this method we receive the result of the previous part of the execise and eliminate all itemsets of size 1 that are not subsets of other itemsets of size greater than 1 in the answer of the first part. Then we calculate all possible rules by doing all the permutations of the list and, finally we check which of those rules have a confidence above the given threshold.

2 Instructions

In order to execute our solution certain libraries need to be installed beforehand. To install them, run the following commands (it is possible that you will need admin access):

```
apt-get install python3 python3-pip
pip3 install numpy
```

To run the code, simply go to the project folder and run:

```
python3 src/main.py --in PATHFILE --s SUPPORT --c CONFIDENCE
```