# Report 2: Itemsets

Group 150: Lorenzo Deflorian, Riccardo Fragale

November 15, 2025

## 1 Introduction

This homework was about discovering frequent itemsets and generating association rules. This two problems are important especially in the field of sales transaction databases since companies wants to discover and understand the logic that stands behing customers' behaviours. The procedure is in a way quite straightforward and will be tested on a dataset given by the teacher that includes generated transactions (baskets) of hashed items.

## 2 Our methods

We needed to develop a way to solve the following two questions;

- How to find frequent itemsets with support at least $s$
- How to generate association rules with confidence at least $c$ from the itemsets found before

In order to solve the first problem we implemented the A-Priori algorithm in the class **apriori**. After testing its functionalities we used that methods and set of new ones in the class called **association_rules** to address the second question. Finally we implemented a set of tests on the dataset given to show that our procedure it is working and also measuring how much time do our algorithms take to find solutions.

## 3 Results

The two main tests are called **test_apriori** and **test_rule_generator**. The first one gave us the following results:

- 381 frequent itemsets
- itemsets by size: [375, 6]

It took 0.460 seconds for the procedure.

Regarding the second test, where actually the first part includes the Apriori procedure as explained before, we found 12 association rules, that are shown below.

```
    AssociationRule({227} -> 390, s=0.0105, c=0.5770, i=0.5502)
INFO     AssociationRule({390} -> 227, s=0.0105, c=0.3907, i=0.3725)
INFO     AssociationRule({346} -> 217, s=0.0134, c=0.3850, i=0.3313)
INFO     AssociationRule({390} -> 722, s=0.0104, c=0.3881, i=0.3296)
INFO     AssociationRule({217} -> 346, s=0.0134, c=0.2486, i=0.2139)
INFO     AssociationRule({682} -> 368, s=0.0119, c=0.2887, i=0.2104)
INFO     AssociationRule({789} -> 829, s=0.0119, c=0.2771, i=0.2090)
INFO     AssociationRule({722} -> 390, s=0.0104, c=0.1783, i=0.1514)
INFO     AssociationRule({829} -> 789, s=0.0119, c=0.1753, i=0.1322)
INFO     AssociationRule({368} -> 682, s=0.0119, c=0.1524, i=0.1111)
INFO     AssociationRule({829} -> 368, s=0.0119, c=0.1753, i=0.0971)
INFO     AssociationRule({368} -> 829, s=0.0119, c=0.1525, i=0.0844
```

Statistically speaking we have an average confidence of **0.2824** and an average interest of **0.2328**. Looking at the time requested it is **0.483** which implies that the longer procedure is the apriori algorithm with respected to the rule generation phase.

# 4 Conclusion

This assignment helped us thoroughly understand the steps required to identify assocaition rules. We also realized how important these techniques are, what might be their economical impact and why they are used especially looking at the possible scalability.

Moreover, we could try in the future to test our modules on bigger datasets, maybe even on data streams to verify in reality whether our expectations on performance and scalability are respected.