Data Mining Assignment - Pattern Mining

Sam Pinxteren, Prof. Goethals

Deadline: March 16, 23:59

This assignment document consists of 3 sections; the project description, the tasks and questions, and the notices. Please read all carefully before beginning your work.

1 Project Description

In this assignment, we will use a dataset derived from kaggle[1]. Reading the source can give you deeper insight into the data but it is not allowed to download the full kaggle dataset to solve these assignments. All the data needed here can be found in the file *dataset.csv*.

The dataset consists of events on an online cosmetics store where each line describes a single interaction with an item. A *user_id* is provided such that multiple interactions by the same person (or account) can be linked. The sessions described by *session_id* can be ignored for this assignment.

The main work in this assignment involves association rules. We will try to find association rules in the interactions of each user. In terms of the typical *market basket* used in association rules, each market basket is all events from a single user. The meaning of a rule $X \Rightarrow Y$ during this assignment is therefore: "People who interact with items in set X also interact with items Y with a certain confidence".

You are allowed to use any programming language of your choice, though we recommend Python. It is convenient to load the dataset in Python using Pandas:

```
import pandas as pd

# Load the dataset as a pandas DataFrame
dataset = pd.read_csv("dataset.csv")

# Convert the dataset to baskets (a list of sets)
baskets = dataset.groupby("user_id").product_id.apply(set).tolist()
```

[1] https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop

2 Tasks and Questions

Complete the following tasks and report your process together with the answers to each question in a pdf. It is advised to read all the questions before starting the assignment.

1. Implement an association rule mining algorithm, or use an existing online implementation. Show that you understand the method by describing its function (without using code) in your report. Make sure you are able to get the confidence and support of any found association rules.

2. Run the association rule mining algorithm on the given dataset. At this point, use only the *user_id* and *product_id* columns. What are the top 10 association rules in terms of support your method finds? Also include the confidence of these rules. What can you say about the number of items in these rules?

3. If you were asked to give the 10 most interesting rules, which 10 would you give and why?

4. A lot of information from the dataset was omitted in the current association rules, such as the event types, which describe whether an item was viewed, purchased, added or removed from the cart and the prices of items. Find a way to incorporate the additional information provided into your association rules. Describe what you have added in your report.

5. After adding additional information, which rules would you deem most interesting now, and why?

3 Notices

Your final submission should be a zip-archive called submission.zip which contains:

1. Your report, a PDF file called *report.pdf.* Although it is not a requirement, we recommend you use LaTeX.
2. The code you have written to solve the questions. The report should stand on its own as your submission. The code should not be necessary for us to evaluate your work. You are allowed to send in Jupyter Notebooks as your code, but please avoid having long outputs which fill the entire notebook.

Important Notes:

- Any file with the wrong file name will not be considered, including capital letters and file extensions. Please double check that this is correct.
- Upload your submission to the assignments section on blackboard for this course and assignment before the announced deadline, any delay will result in a score of 0 for this assignment.
- If you have any questions or remarks, feel free to send an email to: sam.pinxteren@uantwerpen.be. Alternatively, I can be found in office M.G.323.