Data Mining Assignment - Recommender Systems

Sam Pinxteren, Prof. Goethals

Deadline: June 18, 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 again look at the cosmetics store dataset from the first assignment on association rules. The goal is to create a recommender which the store could use. That is, you have some previous interactions from a user, when they come to the store again, you recommend them certain items. This items could, for example, be displayed on the home page after a user has logged in. Your goal is to recommend the user a certain item which they end up purchasing. For this work, we will assume there are 10 spots on the homepage in which to recommend an item, so you can recommend 10 items each time.

There are many ways to do this, some of which you have seen in the theory lectures. In this assignment, however, we will use the association rules which you have found before. Because these rules infer which items could be present based on a set of given items, we can use them to find a missing item. In this case, we will try to predict

In many machine learning tasks, the time travel problem can give false confidence in certain results. You will notice that all the data you are provided is timestamped. For this assignment, you are allowed to ignore this problem, this means you do not have to take into account the order of events. You are allowed to use the dates and times provided, but you do not have to consider that a certain prediction might use future data for that event.

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:

```python
import pandas as pd

# Load the datasets as pandas DataFrames
dataset = pd.read_csv("dataset.csv")
test = pd.read_csv("recommend.csv")
```

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. We will start with a simple recommender. In this case, the homepage is completely static, meaning the recommendations are the same for each user. You also are not allowed to use the *recommend.csv* data to find these 10 items. How do you find the best items to recommend? Which items do you recommend? create a file called *baseline.csv* which contains a line with a *product_id* for each item you think should be on the homepage.

2. For this baseline recommender, how often do you expect someone to buy the item you've recommended, can you give a reasonable upper or lower limit?

3. We will now introduce the association rules which you have mined on the dataset in the first assignment. How can you use these rules to create individual recommendations to a user? In the file *recommend.csv* you will find the event data for 2333 users. Each of these users will come to the platform and purchase 1 item. Your task is to recommend them this item so they can find it on the homepage without searching for it. Hand in your recommendations for each user in a file called *recommendations.csv.* In this file, each line corresponds to one user. Format this file in CSV form, where the first value on each line is the *user_id*, followed by the 10 items you would recommend. The order of the recommended items does not matter in this case. You do not have to use all 10 recommendations, but there is little reason not to.

4. How well do you expect your recommender to perform? Do you have a reasonable estimation, or an upper or lower limit? How do you think this method compares to the baseline in the first question?

3 Minimal Requirement

These points indicate the minimal work needed to pass this assignment. In other words, you should pass the assignment if you've completed these tasks correctly. If you finish only these minimal requirements and make mistakes you might not pass. Gaining higher grades requires creativity on your part, this generally consists of putting above minimal effort into solving the given problems.

- Answer *every* question posed, describe your process and reasoning clearly.
- Create the baseline recommender, with your solution in *baseline.csv*
- Create a recommender which in some way uses association rules, with your solution in *recommendations.csv.* This recommender needs to have a higher hit rate at 10 than the baseline.

4 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 fill the entire notebook.

3. The file *baseline.csv*. This file describes the files chosen in your baseline recommender. This file should consist of 10 lines of text, where each line describes a single *product_id*. The order of these items does not matter in this case.

4. The file *recommendations.csv*, which consists of lines describing your recommendations for each *user_id* in the *recommend.csv* file. Each line should be formatted as such:

   ```
   <user_id>,<product_id 1>,<product_id 2>,...,<product_id 10>
   ```

   For example:

   ```
   580009457,5851894,4229,4807,5859270,5526,4809,585854,566996,6718,4621
   ```

   Since there are 2333 users for who need a recommendation, there should be 1810 lines in this file. Be sure to do a quick sanity check to make sure this is the case.

Important Notes:

- Any file with the wrong file name will not be considered, including capital letters and file extensions. The same goes for files which are formatted incorrectly. 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.