Data Analytics 874
Post-Block Assignment 2: Rule Association
Department of Industrial Engineering

Deadline: 5 November 2021, 23:59

Instructions

For the purposes of this assignment, note the following instructions:

- Assignments are to be completed by students individually.

- Submit your own work.

- You will submit a pdf document addressing all of the aspects listed in the specifications below. Name your pdf document `????????PBA2_874.pdf`, where the question marks are replaced with your student number.

- Make sure that your pdf document has a title page containing the assignment number as the title, and your initials, surname and student number.

- Submit your pdf document no later than the deadline.

Rule Association (100 marks)

You are given a groceries dataset, uploaded to SUNlearn as `groceries.csv`. This is a comma separated file. Your task is to find correlations among items purchased, and to report on these correlations. The assignment requires the application of a rule association algorithm and an investigation into visualization of the most frequent item sets and the extracted association rules.

The dataset contains 9835 transactions, one per row in the dataset. The first entry per row indicates the number of items purchased, with the rest of the entries indicating the items that were purchased.

Complete the assignment in the following steps:

1. Complete a data understanding and visualization of the provided data, before you apply any rule association analysis. Here, you are required to think out of the box in terms of how a better understanding of the data can be obtained, and what aspects can be visualized. Ideas are to provide descriptive statistics on the number of items purchased, including a histogram to indicate the distribution of the number of items purchased. Visualize item frequency to illustrate the number of times that each item has been purchased. Think of approaches to visualize the frequency in which item pairs have been purchased. Google to find different ways in which a better understanding of the data set can be obtained. In your report, you have to provide a detailed discussion of the approaches used, and your observations from these visualizations. (30)

2. Decide on the rule association algorithm that you will use and provide a short description of the algorithm. (5)

3. For the next step you have to pre-process the provided dataset into a representation that can be provided as input to the chosen rule association algorithms. Describe in your report the pre-processing steps implemented. (5)

4. Decide on the values to use for the control parameters of the chosen rule association algorithm. Provide these in your report, and motivate your choices. Then report on all of the associations that were found in a table. Indicate for each of the rules the rule support and the rule confidence. (30)

5. For the last part of the assignment, you have to explore approaches to visualize the extracted associations. In your report, describe the approaches that you have found and used, present these visualizations and discuss them. (30)