Topics in Data Analytics S23
Assignment 2

June 8, 2023

1 Submission

You need to submit your analysis as an executable Python Jupyter Notebook file, to onq. This Jupyter Notebook file should be named "1234-Assn2.ipynb", where 1234 stands for the last 4 digits of your Queen's student ID.

You should use Markdown cells in Jupyter Notebook to describe the motivation of questions and findings. Make sure the TA can find your answer to each question easily.

An "I uploaded the wrong file" excuse will result in a mark of zero.

2 Background

We will reuse the dataset from assignment 1, with a different purpose. This assignment aims to practice unsupervised analysis, i.e., clustering and frequent pattern mining.

3 Clustering Analysis

Q1 (30 points) Our goal is to cluster users based on their spending behaviors. Your task is to perform a clustering analysis leveraging the K-means method and report your findings. You should specify how you select the right "K" for the k-means method, and how you create features. Explain the resultant clusters (the meaning of each cluster) and judge the quality of resultant clusters.

Q2 (5 points) Based on your results from Q1, report the statistics of loyalty scores for each cluster of users. Report your findings.

Q3 (30 points) Discuss whether you need to reduce the dimensions to improve clustering performance and build another clustering model (if you believe dimension reduction would help, you can apply it or choose another clustering method) to improve your analysis from Q1. You should explain why the resultant clusters are better than the ones you got in Q1.

Q4 (5 points) Based on your results from Q3 report the statistics of loyalty scores for each cluster of users. Report your findings.

4 Frequent Pattern Mining

You can consider the mlxtend library[1] or the spmf library[2]

Q5 (20 points) Use frequent pattern mining to find frequently co-occurring transaction types. Report your findings and justify how you pick the min support threshold.

Q6 (10 Points) Redo Q5 by splitting users based on their loyalty scores. The goal is to explore if different frequent patterns exist in users with high/low loyalty scores.

[1] https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/
[2] https://www.philippe-fournier-viger.com/spmf/