Assignment 3: Frequent Itemsets, Clustering, Advertising

Formative, Weight (10%), Learning objectives $(1, 2, 3)$,
Abstraction (4), Design (4), Communication (4), Data (5), Programming (5)

Due date: $11:59$ **pm,** 28 **May,** 2021

1 Overview

Read the following carefully as it differs from the last assignment.

For students who are enrolled in the course COMP SCI 3306 (i.e. undergraduate students), the assignment must be done in groups consisting of **TWO** *students. Please use* A3-groups *on MyUni to organise yourselves into groups. If you have problems/questions regarding grouping or require assistance, please contact the teaching assistant Mahdi (mahdi.kazemimoghaddam@adelaide.edu.au).*

For other students who are enrolled in the course COMP SCI 7306 (i.e. postgraduate students), this assignment must be done **individually. Do not join a group in this case.**

References to sections, examples, etc. refer to the book of "Leskovec, Rajaraman and Ullman: Mining Massive Datasets **(Second Edition)**".

2 Assignment

Exercise 1 Frequent Itemsets (15+15+10+10 points)

For this exercise, you have to read Section 6.4 up to 6.4.3.

1. Implement the simple, randomized algorithm given in 6.4.1

2. Implement the algorithm of Savasere, Omiecinski, and Navathe (SON algorithm) in 6.4.3

3. Compare the two algorithms on the datasets T10I4D100K, T40I10D100K, chess, connect, mushroom, pumsb, pumsb star provided at

   http://fimi.ua.ac.be/data/

   and report the outcomes.

4. Experiment with different sample sizes in the simple randomized algorithm such as 1, 2, 5, 10% and compare your results (including the result produced by the SON algorithm).

   Your approach should be as efficient as possible in terms of runtime and memory requirements.
   Report on challenges that you might have observed in the implementation and by running experiments.

Exercise 2 Clustering (10+20 points)

1. Perform a hierarchical clustering on the one-dimensional set of points

   $1, 4, 9, 16, 25, 36, 49, 64, 81.$

   assuming the clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged. (Exercise 7.2.1)

2. Implement the K-means algorithm and carry out experiments on the Iris dataset (note that you are not allowed to use the libraries such as scikit-learn to implement the algorithm itself, but you are free to compare your results with such). The dataset can be accessed from scikit-learn library. You may follow the instructions at the following link:

   https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

   a) Plot the K-means clustering results by plotting the first 2 dimensions of the input data as well as the converged centroids.

   b) Provide some discussions about how you picked the value of K in the K-means algorithm.

   Note: You should only use the 4 input **features** in the Iris dataset to cluster them, and not the **labels**.

Exercise 3 Advertising (Exercise 8.4.1) (10+10 points)
Consider Example 8.7. Suppose that there are three advertisers A, B, and C. There are three queries x, y, and z. Each advertiser has a budget of 2. Advertiser A only bids on x, B bids on x and y, and C bids on x, y, and z. Note that on the query sequence $xxyyzz$, the optimal offline algorithm would yield a revenue of 6, since all queries can be assigned.

1. Show that the greedy algorithm will assign at least 4 of the 6 queries $xxyyzz$.

2. Find another sequence of queries such that the greedy algorithm can assign as few as half the queries that the optimal offline algorithm would assign to that sequence.

3 Procedure for handing in the assignment

Work must be handed in using Canvas (MyUni). The submission should include:

- a PDF file of your solutions for theoretical assignments. The solutions should contain a detailed description of how to obtain the result.

  For Exercise 2.2, you should properly provide comments in your code to show your understanding.

- all source files, all the project files.

- a README.txt file containing instructions to run the code, the names, student numbers, and email addresses of the group members (or individuals in the case of postgraduates).