 ­­­­­­



ECS766P: DATA MINING

Dr. Dimitrios Kollias, Elona Shatri

***Assignment 2 (worth 20% of total mark)***

­­

**Due Date: 15/12/2023 (16:00)**

**A general note:**

Most of the below tasks will require code in order to be addressed. Please **do make sure to show your workings** - i.e., how did you derive the result by showing the code that was used to generate the result that addresses the question and by writing down your thinking. The code ***should not*** be a screenshot-image; it should be copied and pasted (e.g. from your notebook) to the pdf/doc/docx that you will submit (see deliverables below). The code should be in a correct & readable format, e.g. with correct indentation for python.

# Tasks

1. Questions 1(a) to 1(f) are pen-and-paper exercises (brief answers and justifications are expected). In all responses, please show your workings (equations, justifications). Please try to use an editor instead of taking pictures or scanning actual paper. Pen and paper here refer to not using any programming.   
   1. [pen&paper] - What is the advantage of using the Apriori algorithm in comparison with computing the support of every subset of an itemset in order to find the frequent itemsets in a transaction dataset?

[5 marks]

* 1. [pen&paper] - Let denote the set of frequent -itemsets. For why must every frequent -itemset be a superset of an itemset in ?

[6 marks]

* 1. [pen&paper] - Let . Compute the set of candidates that is obtained by joining every pair of joinable itemsets from

[5 marks]

* 1. [pen&paper] - Let denote the support of the association rule:  
       
       
     Let denote the support of the association rule:  
       
        
       
     What is the relationship between and ?

[5 marks]

* 1. [pen&paper] - What is the support of the rule:  
     in the transaction dataset below shown in **Figure 1**?

A group of black text

Description automatically generated

Figure 1. Dataset 1

[5 marks]

* 1. [pen&paper] - In the transaction dataset shown in **Figure 1**, what is the maximum length of a frequent itemset for a support threshold of 0.2?

[5 marks]

1. For your answers to the assignment, please include your workings (e.g. equations, code) when this is relevant to the question. Questions 2a & 2b are pen and paper exercises. Question 2c can be addressed either on paper or using code. Question 2d is a coding exercise.
   1. [pen&paper] - For a system designed to prevent identity theft in online transactions, we are focusing on identifying unusual transaction patterns. Propose 2 possible contextual attributes and 2 possible behavioural attributes that could be integrated into this system's algorithm. Provide a rationale for classifying each attribute as either contextual or behavioural.

[6 marks]

* 1. [pen&paper] - Assume that you are provided with the [University of Wisconsin breast cancer dataset](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data) from the Week 3 lab, and that you are asked to detect outliers from this dataset. Additional information on the dataset attributes can be found [online](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names). Explain one possible outlier detection method that you could apply for detecting outliers for this particular dataset, explain what is defined as an outlier for your suggested approach given this particular dataset, and justify why would you choose this particular method for outlier detection.

[7 marks]

* 1. [Coding or pen&paper] - The monthly rainfall in the London borough of Tower Hamlets in 2019 had the following amount of precipitation (measured in mm, values from January-December 2018):   
       
     {22.93, 20.69, 25.75, 23.84, 25.34, 3.25, 23.55, 28.28, 23.72, 22.42, 26.83, 23.82}.  
       
     Assuming that the data is based on a normal distribution, identify outlier values in the above dataset using the maximum likelihood method.

[7 marks]

* 1. [Coding] - Using the stock prices (stocks.csv included in the supplementary material) dataset used in sections 1 and 2 of Week 9 lab, estimate the outliers in the dataset using the one-class SVM classifier approach. As input to the classifier, use the percentage of changes in the daily closing price of each stock, as was done in section 1 of the notebook. Use the same SVM settings as in the lab notebook. Plot a 3D scatterplot of the dataset, where each object is color-coded according to whether it is an outlier or an inlier. Also compute a histogram and the frequencies of the estimated outlier and inlier labels. In terms of the plotted results, how does the one-class SVM approach for outlier detection differ from the parametric and proximity-based methods used in the lab notebook? What percentage of the dataset objects are classified as outliers?

[8 marks]

1. Questions 3(a)(I) is a pen and paper exercise and 3(a)(II) is a coding exercise. Questions 3(b) is a pen-and-paper exercise.
   1. You are provided with the following URL: <http://eecs.qmul.ac.uk/~emmanouilb/income_table.html>.

This webpage includes a table on individuals' income and shopping habits.

* + 1. [pen&paper] - Inspect the HTML code of the above URL and provide a short report on the various tags present in the code. What is the function of each unique tag present in the HTML code?

[6 marks]

* + 1. [Coding] - Using Beautiful Soup, scrape the table and convert it into a pandas dataframe. Perform data cleaning when necessary to remove extra characters (no need to handle missing values). In the report include the code that was used to scrape and convert the table and provide evidence that the table has been successfully scraped and converted (e.g. by displaying the contents of the dataframe).

[6 marks]

b. [pen&paper] Consider the graph in the figure below as displaying the links for a group of 5 webpages. Which of the 5 nodes would you consider hubs and which would you consider authorities?

A diagram of a diagram

Description automatically generated

[6 marks]

1. Question 4a. is a pen-and-paper exercises; questions 4b is a coding exercise. For all your answers please show your workings (equations or code when applicable).   
   1. [pen&paper] - Consider the following sentences related to data mining theory, and assume that each of the below sentences corresponds to a different document:

*\* Data refers to characteristics that are collected through observation.*

*\* A dataset can be viewed as a collection of objects.*

*\* Data objects are described by a number of attributes.*

*\* An attribute is a characteristic or feature of an object.*

1. Construct and display the document-term matrix for the above documents. Remove all stop words (here consider as stop words: articles, prepositions, conjunctions, pronouns, and common verbs) and punctuation marks; convert any plural nouns/adjectives to their singular form; and convert verbs to the present tense and first-person singular form, before you construct the matrix.

[7 marks]

1. Using the above constructed document-term matrix, calculate the inverse document frequency for all words you have identified from the previous question (I).

[6 marks]

* 1. [Coding] - Using the daily births dataset from Week 11 lab notebook, smooth the timeseries using trailing moving average smoothing and a window size that corresponds to one week; then replace any NaN values with zeros. Perform timeseries forecasting using the smoothed dataset in order to predict daily births for the first 5 days of 1960, using the models below. Show your forecasting results.  
       
      AR model with   
     ARMA model with and

[10 marks]

**Submission Requirements - Deliverables**

You are asked to submit a **report** that should answer the above questions, show the obtained results and include code (not as screenshot-image). The report should be in **PDF, .doc or .docx format** (so ***not*** notebook etc), named as:  
  
Assignment2-StudentName\_StudentSurname-StudentNumber.pdf or .docx/.doc

**Marking Criteria**

* Correct and sufficient explanations, plots and written code that answer the questions and show understanding.
* Clearly and succinctly written report

## **Important notes about the assignment:**

# - This is an individual assignment. **Plagiarism** is an irreversible non-negotiable failure in the course (if in doubt of what constitutes plagiarism, please ask).

- The submission cut-off date will be 7 days after the deadline. Late submissions will   
 receive late penalties in line with the late penalty policy, see EECS handbook and   
 QMUL assessment handbook.

- Cases of **Extenuating Circumstances (ECs)** have to go through the proper procedure   
 of the School in due time. Only cases approved by the School in due time can be   
 considered.

- No other means of submission other than submitting your assignment through the   
 appropriate QM+ link are acceptable at any time. Submissions sent via email   
 will **not** be considered.

**Submission Checklist**

* Has your file been saved in **PDF or .doc/.docx format**?
* Have you clicked **[Submit]** after uploading?
* Have you checked that the file you uploaded is the correct version?

The first time you submit, you will be required to accept the Turnitin End User Licence Agreement.

**After uploading, it is your responsibility to check that your file is in the correct format and that it is readable.**