Queen Mary
University of London

Winter Examination Period 2021 — January — Semester A

ECS766P DATA MINING ***Duration: 3 hours***

This is a 3-hour open-book exam, which must be started within a 24-hour period.

You MUST submit your answers within 3 hours of the time that you started the exam.

Follow all instructions on the download page.

You can refer to textbooks, notes and online materials to facilitate your working, but normal referencing and plagiarism rules apply, and you must cite any sources used.

You must upload a **SINGLE PDF** file containing your solutions. These can be typed or hand-written, or a combination of the two. Multiple submissions are not permitted, so be sure that you check your submission before uploading it.

Calculators are permitted in this examination. You should not use code (e.g. python) to respond to questions.

Answer ALL questions.

You MUST adhere to the word limits, where specified in the questions. Failure to do so will lead to those answers not being marked.

YOU MUST COMPLETE THE EXAM ON YOUR OWN, WITHOUT CONSULTING OTHERS.

Examiners:
Dr. Emmanouil Benetos and Prof. Ioannis Patras

Question 1

(a) Consider the following data mining problems to be addressed by an online sales company:

1. Predicting the amount of money a customer would spend for the next month based on their purchase history.
2. Assuming that the company has created customer categories based on existing customers' shopping habits, predicting the customer category for a new customer.
3. Creating targeted offers to customers for items frequently purchased together.
4. Flagging customers that exhibit unusual buying patterns.

Which data mining task should the company use for each of the above cases? Justify your response for each case. **[8 marks]**

(b) A bookstore needs to perform the following two tasks:

1. Calculating the count and sum of its daily transactions; transactions are recorded in the store's database.
2. Grouping the books in the bookstore according to topics, using keywords which are stored for each book in the store's database.

For each of the above two tasks, explain whether it can be considered as a data mining task or not; justify your response. **[4 marks]**

(c) Consider a student feedback survey for a module, where students can provide one of the following ratings with respect to module satisfaction: very dissatisfied, dissatisfied, neutral, satisfied, very satisfied.

1. Suggest and formulate a dissimilarity measure to compare student ratings.
2. Provide an example using two student feedback ratings (e.g. comparing dissatisfied with very satisfied) to show how your dissimilarity measure works in practice.

[4 marks]

(d) Consider the following two datasets:
$x = [1, 1, 1, 2, 3, 3, 25, 50]$ and $y = [2, 4, 6, 6, 6, 8, 10]$.

1. Calculate the mean, median, and mode for each of the above two datasets.
2. For each dataset, compare its previously calculated mean, median, and mode values. For each of the above two datasets, is the data distribution symmetric or skewed, and if the latter to which direction? Justify your response.

[4 marks]

(e) Consider the following two binary vectors: $x = (0, 1, 1, 0, 1, 1)$ and $y = (0, 0, 0, 1, 1, 1)$.

1. Compute the SMC (simple matching coefficient) and Jaccard coefficients for the above vectors.

2. Assume that the 6 dimensions in both vectors represent 6 different items found in a store. Each vector represents a customer transaction, indicating whether a customer has purchased a particular item (1) or not (0). We are interested in developing a customer similarity metric, which considers two customers similar when they both purchase the same item. The similarity metric does not need to consider the case when both customers do not purchase any particular item. Which of the above two coefficients should we use and why?

[5 marks]

Question 2

(a) Consider the below table of weather measurements in London boroughs for certain dates:

Date	Borough	Temperature (°C)	Precipitation (mm)	Wind (mph)
1/11/2018	Newham	10.1	5.6	4
1/11/2018	Tower Hamlets	15.0		7
1/11/2019	Newham	9.7	20.3	-1
1/11/2019	Tower Hamlets	?	45.8	5
1/11/2020	Newham	8.5	31.4	11
1/11/2020	Tower Hamlets	11.1	38.1	6

1. Identify cells in the above table which require data cleaning.
2. Describe two methods that can be applied to handle the data identified in the previous question.

[5 marks]

(b) Describe the difference between dimensionality and numerosity reduction. Assuming we have a dataset with *M* objects and *N* attributes, how does numerosity reduction affect *M* and *N*? How does dimensionality reduction affect *M* and *N*?

[4 marks]

(c) Consider a dataset *A* that contains a single numeric attribute. The minimum value in the dataset is denoted as $\min_A$, the maximum value is denoted as $\max_A$, the dataset mean is denoted as μ_A, and the dataset's standard deviation is denoted as σ_A. If we apply z-score normalisation, which is the new value range? What are the values of the mean and standard deviation on the normalised dataset?

[4 marks]

(d) An online store needs to develop infrastructure for two tasks: (1) to register daily store transactions; (2) to study customers' shopping trends over the past 10 years. Which data infrastructure and system should the company use for each of the above two tasks and why?

[4 marks]

(e) Consider a data warehouse for an airline which includes dimensions on passenger, time of the flight, and flight code. The data warehouse also has the measure fee, which refers to the amount for a flight for a passenger, and count, which counts how many times has the passenger flown with the airline.

1. Draw a star diagram for the data warehouse. Populate the dimension tables with attributes relevant to the application.
2. Starting with the base cuboid [day, passenger,flight code], which OLAP (online analytical processing) operations should be performed in order to list the total amount collected for a specific flight code (e.g. BA 208) over a specific year (e.g. 2019)?

3. Assume that the time dimension has 3 levels: day $<$ month $<$ year. How many cuboids will the cube contain, including the base and apex cuboids?

[8 marks]

Question 3

(a) Give a numerical example of each of the following data summarisation scenarios:

1. The mean exhibits vulnerability to outliers while the median exhibits robustness to outliers.
2. The 80-th percentile of a feature is the value 8.
3. The (Pearson) correlation coefficient between a pair of features is equal to one.

[6 marks]

(b) Consider the following dataset represented by a table.

ID	Feature 1	Feature 2	Feature 3	Feature 4
0	15	13	4	26
1	9	17	13	36
2	13	10	7	23
3	12	11	18	25
4	13	5	11	12
5	16	1	9	3
6	11	16	18	32
7	2	3	2	9

1. Create a scatterplot to visualise features 2 and 4. Do not use code to generate the plot. What can you say about their (Pearson) correlation coefficient based on this visualisation? You don't need to compute the correlation coefficient.
2. Draw an equal-width histogram with four bins for feature 3. Do not use code to generate the plot. Assume the feature ranges between 1 and 20.

[4 marks]

(c) Answer each of the following questions related to learning algorithms for classification:

1. Use an example to explain why a K-nearest neighbours classifier may require a tie-breaking policy.
2. Use an example to explain why a feature whose range is much larger than the range of the other features may disrupt the behavior of a K-nearest neighbours classifier.

[4 marks]

(d) Answer each of the following questions related to the evaluation of classifiers:

1. Suppose a classifier with perfect recall classifies a patient as disease-positive. Can the patient be confident that they have the disease?
2. Suppose a classifier with perfect precision classifies a patient as disease-positive. Can the patient be confident that they have the disease?
3. Explain the sentence: improving model selection after using the test set defeats its purpose.

[5 marks]

(e) Assume that the K-means algorithm is employed in a clustering task using a Euclidean distance between observations:

1. Suppose the cluster centers for a given iteration of the algorithm are $\boldsymbol{\mu}_1 = (1, 2)$ and $\boldsymbol{\mu}_2 = (-1, 1)$. To which of the clusters would the observation $\mathbf{x} = (1, -1)$ be assigned? Show your calculations.
2. Suppose a cluster center is associated to the observations $\mathbf{x}_1 = (3, 4)$, $\mathbf{x}_2 = (-1, 2)$, $\mathbf{x}_3 = (2, 3)$ after the assignment step. What would be the new position of this cluster center after the movement step?
3. Explain why choosing a number of clusters K that solely minimizes the sum of squared errors is not a good strategy for finding an informative clustering.

[6 marks]

Question 4

(a) Consider the following transaction dataset represented by a table:

ID	Transaction
1	$\{1,2,3,4\}$
2	$\{1,3\}$
3	$\{1,2,5\}$
4	$\{2,5\}$
5	$\{1,2,3\}$
6	$\{1,4\}$
7	$\{1,2,3,5\}$
8	$\{4,5\}$

1. What is the *support count* of the itemset $\{1,3\}$? What is the *support* of the itemset $\{2,4\}$?
2. For a support threshold of $\tau_S = 0.4$, is the itemset $\{1,2\}$ considered frequent?
3. What is the support of the association rule $\{1,2\} \implies \{5\}$?
4. Compute the Kulczynski measure of the itemsets $\{1,3\}$ and $\{2\}$.
5. Assume that the itemset $\{1,3,4\}$ is frequent for some support threshold τ_S. List all the association rules that can be derived from this itemset.

[10 marks]

(b) Answer the following questions regarding outlier detection:

1. Give an example of a scenario where it is important to identify outliers for specific contexts. Make the distinction between contextual features and behavioural features clear.
2. What are the two main difficulties in obtaining a representative dataset for supervised outlier detection methods?

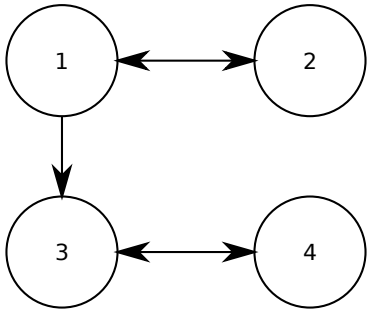
[4 marks]

(c) Answer the following questions regarding outlier detection:

1. Explain why the density around an observation **x** is inversely related to the distance between **x** and its *k*-th nearest neighbour (for an appropriate choice of *k*).
2. Explain how a predictive model (*e.g.*, a classification or regression model) can be used to detect contextual outliers even in the absence of observations labeled as outliers.

[4 marks]

(d) Present the system of linear equations that could be solved in order to obtain the PageRank $\pi(i)$ for each webpage *i* represented by a node in the graph below. Let α denote the probability of teleportation.



[4 marks]

(e) Consider the set of frequent 2-itemsets $\mathcal{L}_2 = \{\{1,2\},\{1,5\},\{2,3\},\{2,4\}\}$. Compute the set of candidates $\mathcal{C}_3$ that would be generated by the Apriori algorithm by joining every possible pair of joinable itemsets from $\mathcal{L}_2$. **[3 marks]**

End of questions