The University of Melbourne

Semester Two 2019 Exam

School: Computing and Information Systems
Subject Number: COMP20008
Subject Title: Elements of Data Processing

Exam Duration: 2 hours

Reading Time: 15 minutes

This paper has 7 pages

Authorised Materials:
No calculators may be used.
Instructions to Invigilators:
Supply students with standard script books.

Instructions to Students:
Answer all 12 questions. The maximum number of marks for this exam is 50. Start each question (but not each part of a question) on a new page.

Formulae and Notation

Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Pearson's correlation coefficient: $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Entropy: $H(X) = -\sum_{i=1}^{\#categories} p_i \log_2 p_i$
where p_i is the proportion of points in the ith category.

Conditional entropy: $H(Y|X) = \sum_{x \in \mathcal{X}} p(x)H(Y|X = x)$
where $\mathcal{X}$ is the set of all possible categories for X

Mutual information:
$MI(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

Normalised mutual information:
$NMI(X, Y) = \frac{MI(X,Y)}{min(H(X),H(Y))}$

Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$ where
TP is number of true positives
TN is number of true negatives
FP is number of false positives
FN is number of false negatives

Set Union: $X \cup Y$ is the union of sets X and Y (elements in X or in Y or in both X and Y)

Set intersection: $X \cap Y$ is the intersection of sets X and Y (elements in both X and Y)

1. Consider the following XML file:

```
<?xml version="1.0"?>
<subject code="COMP20008">
    <URL> https://handbook.unimelb.edu.au/subjects/comp20008 </url>
    <name> Elements of Data Processing </name>
</subject>
<semester>1</semester>
<year/>
```

   (a) (1 mark) Identify the errors present in the XML file. Modify the XML so that it is well formed.

   (b) (1 mark) Is this an example of unstructured, semi-structured or structured data? Explain why.

   (c) (1.5 marks) Write a JSON file that represents the same information.

   (d) (0.5 mark) Explain one advantage of JSON over XML.

2. Consider the following temperature data for one week in Melbourne:

```
12,   15,   18,   13,   51,   10,   16
```

   (a) (1 marks) Calculate the median and interquartile range (IQR) of the data

   (b) (0.5 mark) Draw a Tukey boxplot of the data.

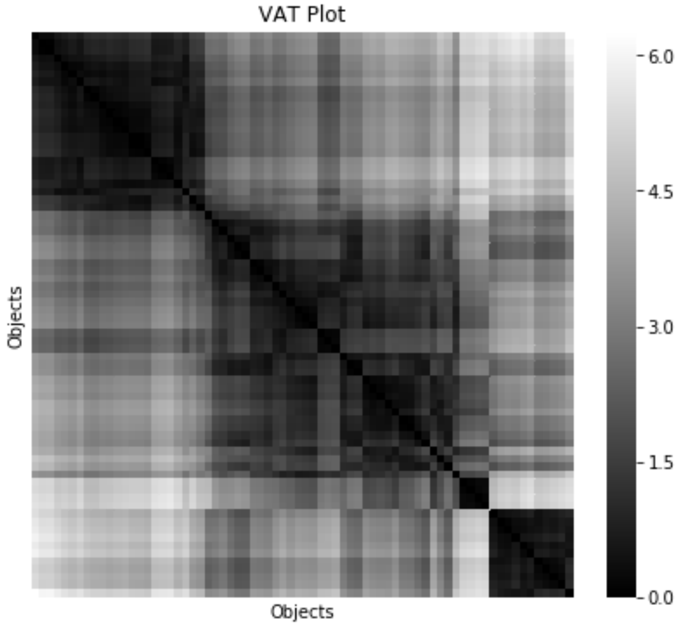
   (c) (1 mark) Will the 51 value be classified as an outlier on the Tukey plot? Why or why not.

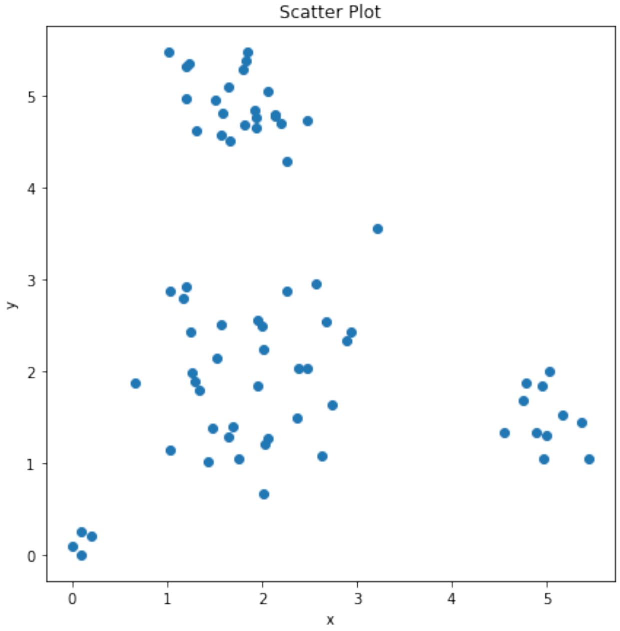
   (d) (1 mark) Suppose we consider the 51 value to be infeasible and discard it. One method for calculating a replacement value is imputation by mean. Considering only the values of the datapoints and not the real-world context of the data, do you think this method would work well for this data? Why or why not?

   (e) (1 mark) Consider the real-world context of the data. Do you think imputation by mean is likely to work well? Why or why not?

   (f) (0.5 mark) Suggest one alternative imputation approach that might work better than imputation by mean for this dataset

3. Consider the following two plots:





   (a) (1.5 mark) Are these likely to have come from the same dataset? Give two reasons why or why not.

   (b) (1.5 marks) Consider a dataset with 10000 rows and 500 features. Give three reasons why we might want to apply PCA while analysing the dataset.

4. Consider the following data points:

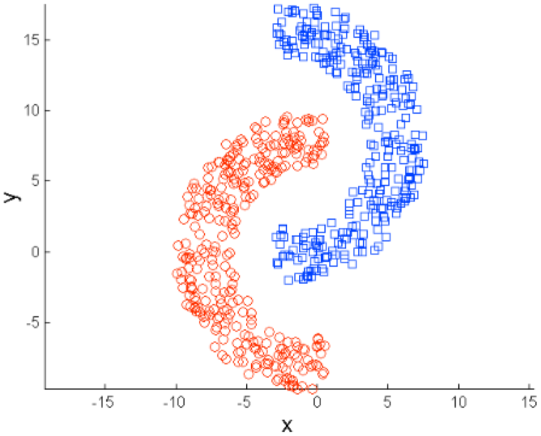
   `(0,0),   (0,3),   (4,0),   (4,3)`

   (a) (1 mark) Calculate the proximity matrix using Euclidean distance between points.

   (b) (2 marks) Apply hierarchical clustering to this dataset using the single linkage method. Draw the resulting Dendrogram.

   (c) (1 mark) Describe one situation where hierarchical clustering using the complete linkage method might perform better than hierarchical clustering using the single linkage method.

   (d) (1 mark) Consider the dataset represented by the scatter plot below, where one class of data is indicated by circular points and another class of data is indicated by square points:



Do you think that hierarchical clustering or K-Means clustering would better separate the two classes? Why?

(e) (1 mark) Describe how clustering can be used to detect outliers in a dataset.

5. Pearson Correlation and Mutual Information are two widely used correlation measures. Assume you have collected the following information for 10,000 Australian adults: Age in years (ranging from 20 to 100) and salary in dollars (ranging from 0 to 100,000).

   (a) (1 mark) When you calculate the Pearson correlation between age and salary for this data set, the value you obtain is 0.8. Name two conclusions you can draw from this result.

   (b) (2 marks) Pearson Correlation has a number of limitations. Name two of them and illustrate with the example above.

   (c) (2 marks) If you want to apply mutual information to this example, the first step is to create appropriate bins for the data. Briefly explain why this is necessary and show one suitable strategy of creating such bins for the example above.

   (d) (1 mark) Another data wrangler applied Mutual Information to the data set and obtained a value of 0.9. Does this together with the Pearson co-efficient of 0.8 mean we can conclude that age determines the salary of a new employee?

6. Cross-validation is a common experimental design tool.

   (a) (1 mark) Briefly explain how 5-fold cross-validation works using a small illustration.

   (b) (2 marks) Classification data sets can often be imbalanced, i.e. there may be far more training examples of one class versus another. Can cross-validation help in this situation? Why or why not?

7. Decision trees consist of a number of elements. Two of them are internal nodes and leaf nodes.

   (a) (1 mark) Explain the difference between these two node types.

   (b) (2 marks) Using decision trees, what happens to a data instance during testing if has an attribute value that was never seen in training? Give two possible scenarios.

8. Consider an online 'Movie' recommender system. The system collects users' ratings for movies and uses collaborative filtering techniques to generate recommendations to the users. Ratings are between 1 and 5 (1, 2, 3, 4, or 5).

   (a) (2 marks) Explain the steps an item-based collaborative filtering approach takes to predict the likely rating of a movie M_1 by a particular user U_1 using the user rating matrix.

   (b) (2 marks) Give one advantage and one disadvantage of item-based collaborative filtering method.s

   We want to offer external people aggregate statistics with some global differential privacy guarantee. It has been decided that we will allow people to query the following two aggregate statistics for a specific movie M_1 from the user rating matrix: (1) the total sum of the ratings for M_1. (2) the number of ratings for M_1. In this context,

   (c) (2 marks) Explain the goal of a global differential privacy guarantee and the general strategy applied to achieve it.

   (d) (1 mark) What are the global sensitivity values for the two aggregate statistics?

9. (4 marks) Consider two parties A and B. Each has a dataset of names, including personal and company names of about 1 million records.

   Explain a 3-party protocol for a privacy preserving data linkage between A and B which can prevent both dictionary and frequency attacks. Explanations should include both data processing and the roles of each of the three parties in performing the processing.

10. Blockchain technology enables data sharing in a peer to peer system.

    (a) (3 marks) Explain how to produce a message digest in a digital signature and give 2 properties of a message digest.

    (b) (2 marks) Describe the data structure of the header of a block in a blockchain; and explain how the mechanism of chaining the blocks in a blockchain is a part of the design for supporting data immutability

11. (2 marks) Given the anonymised table below, what is the best achievable k-anonymity and l-diversity for the data? Justify your answer.

Gender (non-sensitive)	Subject (non-sensitive)	Grade (sensitive)
male	COMP20008	H1
male	COMP20008	H1
female	COMP20008	H1
male	COMP20008	H2
male	COMP20008	H2
female	COMP20008	H3
female	COMP20008	P
male	COMP20003	H1
male	COMP20003	H2
male	COMP20003	P

12. (2 marks) Give two considerations for responsible big data research.

End of Exam