The University of Melbourne

Semester One 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 8 pages

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

Instructions to Students:
Answer all 14 questions. The maximum number of marks for this exam is 50.

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: $H(X) = -\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 p_i$
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 and $|\mathcal{X}|$ is the number of categories.

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. This question concerns representations of data in XML/JSON.

   (a) (0.5 mark) Consider the XML data below. Is it well-formed? Explain your answer.

```
<?xml version="1.0"?>
<subject xmlns:uom="http://UoM/offer/"
         xmlns:open_u="http://OpenUniversity/subjects/" >
COMP20008</subject>
<semester>1</semester>
<year/>
<name> Elements of Data Processing</name>
```

   (b) (0.5 mark) Explain the purpose of '**uom**' and '**open_u**' in the XML data above.

   (c) (1 mark) Provide an effective representation of the following data as a list of 'person' records in JSON format.

Name	**Age**	**Interests**
Tom	45	Gardening, Bird watching
Jack	34	Movies, Video Games
Susan	29	Cycling, Chess

2. Consider the task of processing readings from multiple temperature sensors.

   (a) (1 mark) Explain the difference between readings which are noise and readings which are outliers.

   (b) (2 marks) Using a simple example, explain two ways in which using a histogram for outlier detection in temperature sensor data might provide misleading results.

   (c) Consider the following dataset of readings from a set of 10 temperature sensors collected at a particular time (one reading per sensor).

```
175, 179, 5, 185, 190, 191, 193, 193, 201, 33
```

      i. (1 mark) Calculate the interquartile range (IQR) of this data.

      ii. (1 mark) Explain why readings 5 and 33 would be categorised as outliers in a Tukey boxplot representation.

      iii. (2 marks) There is in fact one sensor reading that has a missing value and was not shown above. Suggest an imputation method for this reading, in light of the data above, and justify your choice.

3. Consider a recommender system for an online bookstore *Books R Us* which records ratings of books from users, and makes recommendations to users about books they might be interested to purchase. The system uses a popularity based strategy for its recommendations, randomly recommending a book from the top 50 books. The top 50 books are determined according to total number of purchases in the store during the last month.

   (a) (2 marks) Explain one advantage and one disadvantage of this popularity based strategy, compared to using an item based collaborative filtering strategy for making recommendations. Explain any assumptions made.

   (b) (2 marks) A data science consultant suggests to *Books R Us* that an alternative strategy would be to train a decision tree classification model to predict whether a user will like a particular book. The features of the classification model would include information about the customer such as (age, gender, location) and also information about the book (cost, topic, year of publication). The class label to be predicted would be $\{$*like, not_like*$\}$.

   Explain one advantage and one disadvantage of this scheme. Explain any assumptions made.

4. Consider a dataset of four data instances $\{x_1, x_2, x_3, x_4\}$ which has the following pairwise dissimilarity (distance) matrix.

	x_1	x_2	x_3	x_4
x_1	0	1	4	5
x_2	1	0	2	6
x_3	4	2	0	3
x_4	5	6	3	0

   (a) (3 marks) Apply single link agglomerative hierarchical clustering to group the data instances represented by this matrix. Draw the resulting dendrogram. Show all working.

   (b) (2 marks) Explain one general advantage and one general disadvantage of single link agglomerative hierarchical clustering compared to k-means clustering.

5. Consider the task of detecting a strong negative (linear) correlation between two features, using visualisation.

   (a) (2 marks) Explain, with the aid of a simple diagram, how a parallel coordinates plot could help a user detect a strong negative (linear) correlation between two features.

   (b) (1 mark) Explain, with the aid of a simple diagram, how a scatter plot could help a user detect a strong negative (linear) correlation between two features.

   (c) (1 mark) Which of the two plots do you believe would be more effective for this task? Explain and justify your answer.

6. (2 marks) Consider a dataset D_1 with 4000 instances and 10 features. Consider two possible strategies to cluster this dataset, using $k = 3$.

   - (S1) Apply PCA to D_1 and retain the top two principal components. Let the resulting 2-dimensional dataset be D_2. Apply k-means on D_2.
   - (S2) Apply k-means directly on D_1.

   In what circumstances would you expect the first strategy (S1) to be more effective than the second strategy (S2)? In what circumstances would you expect the second strategy (S2) to be more effective than the first strategy (S1)? Explain any assumptions made.

7. (4 marks) Given a dataset with four classes, A, B, C and D, suppose the root node of a decision tree has 125 instances of class A, 355 instances of class B, 200 instances of class C and 500 instances of class D. Consider a candidate split of this root node into four children, using a numeric feature F with values from 0 to 100.

   - The first child contains instances with $0 \leq F < 20$ and has class proportions (50 class A and 75 class B),
   - The second child contains instances with $20 \leq F < 40$ and has class proportions (50 class A and 225 class B),
   - The third child contains instances with $40 \leq F < 60$ and has class proportions (25 class A, 30 class B, 50 class C, 100 class D)
   - The fourth child contains instances with $60 \leq F \leq 100$ and has class proportions (25 class B, 150 class C, 400 class D)

   Using these numbers, write an expression for computing the utility of this candidate split using the information gain criterion. The expression may be complex and you do not need to simplify it to a single number.

8. A data scientist collects a large number of data pairs (age, height) of people from birth to 80 years old. She computes a Pearson correlation coefficient between age and height.

   (a) (1 mark) Would you expect the correlation to be positive or negative? Why?

   (b) (1 mark) Would you expect the correlation to be similar in value if it was computed using a different set of people? Explain your answer and any assumptions made.

9. (3 marks) Suppose I have a dataset of 5000 customers (instances). Each customer is described by 10 features. There is a binary class label of {*highvalue, lowvalue*}.

   Suppose we wish to empirically decide whether using k-NN will be a more accurate model than decision tree, for predicting the value of a new customer. Explain the steps one should follow in order to make this decision. Explain any assumptions made.

10. Business X and Business Y have decided to conduct a joint marketing campaign. For this marketing campaign, they first need to determine how many customers they have in common (how many people are in the customer list of both businesses). They involved a trusted 3^{rd} party, Company Z, and implemented the following 3-party privacy preserving protocol, making use of the SHA-256 one way hashing function.

   #In the following, the '+' symbol indicates string concatenation (joining two strings)

   #Business X does the following
   SetX=empty
   For each customer at Business X
   SetX=SetX ∪ SHA-256("First Name"+"Last Name")
   Send SetX to Company Z

   #Business Y does the following
   SetY=empty
   For each customer at Business Y
   SetY=SetY ∪ SHA-256("First Name"+"Last Name")
   Send SetY to Company Z

   #Company Z does the following
   result=count(SetX ∩ SetY)
   Share result with Business X and Business Y

(a) (1 mark) Suppose Company Z is malicious and intends to compromise the privacy of the data. Explain two possible privacy attacks that Company Z can perform.

(b) (1 mark) Explain what changes to the protocol would be required for resisting these two possible attacks.

(c) (2 marks) According to the protocol, Company Z can only compute exact matches between pairs of records from Business X and Business Y. Propose changes to the protocol so that approximate similarity can be computed in a privacy preserving manner by Company Z.

11. In a data de-duplication scenario, blocking can be used to reduce the cost of finding matching entities.

(a) (1 mark) Using the records below, propose a simple blocking strategy that results in at most two records being in the same block, when applied to the four records below.

Lee Cheng, 7 George St, Carlton
Lee Cheng, 7 George Rd, Brunswick
Jan Cheng, 7 Haddock Rd, Carlton
Jan Lee, 14 George Rd, Carlton

(b) (1 mark) Based on your proposed strategy, show the blocks having two records. (Do not show blocks with only one record.)

(c) (2 marks) Provide two reasons as to why your proposed blocking strategy might perform poorly when applied to a large real dataset.

12. Consider the following dataset which has been anonymised. The quasi-identifiers are {*Sex,Age,Postcode*} and the sensitive attribute is *Diagnosis*.

ID	Sex	Age	Postcode	Diagnosis
1	male	20-25	318*	skin rash
2	male	20-25	318*	flu
3	male	40-45	318*	headache
4	female	30-35	318*	skin rash
5	male	20-25	318*	flu
6	female	30-35	318*	flu
7	female	30-35	318*	headache
8	male	40-45	318*	headache
9	male	40-45	318*	headache
10	male	20-25	318*	skin rash
11	female	30-35	318*	headache
12	male	40-45	318*	headache

(a) (2 marks) In the context of k-anonymity: Is this data 1-anonymous? Is it 2-anonymous? Is it 3-anonymous? Is it 4-anonymous? Is it 5-anonymous? Is it 6-anonymous? Is it 7-anonymous? Is it 8-anonymous? Explain your answers.

(b) (2 marks) In the context of l-diversity: Is this data 1-diverse? Is it 2-diverse? Is it 3-diverse? Is it 4-diverse? Is it 5-diverse? Is it 5-diverse? Is it 6-diverse? Is it 7-diverse? Is it 8-diverse? Explain your answers.

13. (2 marks) Suppose Bob signs a document with his digital signature. Fred receives the document and changes its contents, but leaves the digital signature unchanged. How could a third party (Alice), know that the document has been modified from its original version, by someone other than Bob?

14. (3 marks) Consider the following quote from William Mougayar about blockchain *"Online identity and reputation will be decentralized. We will own the data that belongs to us"*.

Using the example of a blockchain for storing education related data, argue 3 distinct reasons why this quote could be true.

End of Exam