The University of Melbourne

Semester Two 2018 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 9 pages

Authorised Materials:
No calculators may be used.
Instructions to Invigilators:
Supply students with standard script books.
This exam paper can be taken away by the students after the exam.
This paper may be held by the Baillieu Library.

Instructions to Students:
Answer all 17 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: $d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

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

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. (1 mark) Describe what kind of strings will be matched by the following regular expression: **@([A-Za-z0-9_-]+)**

   In your answer, break it up into parts and explain what each one does.

2. Consider the following XML fragment:

```
<?xml version="1.0" encoding="UTF-8"?>
<Burger type="beef">
    <Bun>
        <Pickles/>
        <Cheese/>
        <Patty/>
    </Bun>
</Burger>
<Cola>
    <Sugar/>
    <Water/>
</cola>
<Fries kind="French"/>
```

   (a) (1 mark) Identify two errors in the above XML fragment.

   (b) (1 mark) Add a default namespace with URI ”http://burger.com” to the element Burger in the corrected XML fragment.

   (c) (1 mark) Provide a JSON representation of the corrected XML fragment

3. Given the following age values (in increasing order):

   6, 15, 15, 17, 18, 20, 21, 24, 25, 25, 27, 30, 33, 35, 35, 36, 40, 45, 52, 78, 96.

   (a) (2 marks) What is the first quartile (Q1) and the third quartile (Q3) of the data?

   (b) (2 marks) Show a 3-bin equal width discretization

4. Briefly answer each of the following questions:

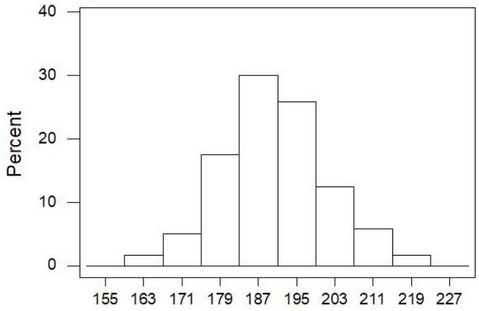
   (a) (1 mark) Explain one difference between a global and a contextual outlier.

   (b) (1 mark) Explain two different strategies used to handle missing data.

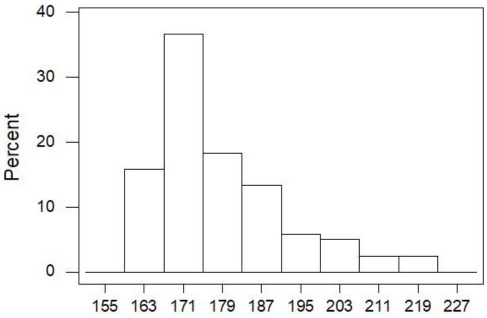
   (c) (1 mark) Explain the relationship between recommender systems and missing data. Support your answer with an example.

5. (2 marks) Match each of the following histograms to one boxplot letter.

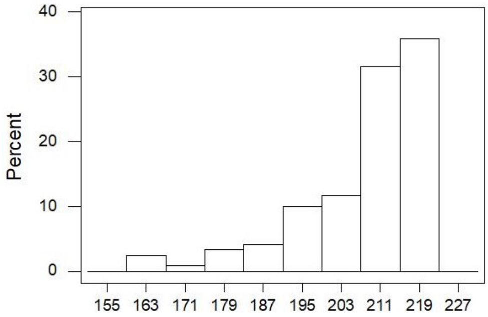
Histogram	H_1	H_2	H_3	H_4
Boxplot letter	...	...	...	...



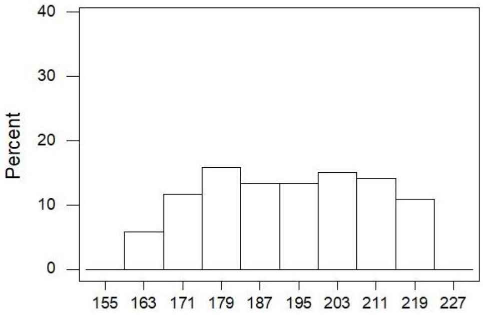
H_1



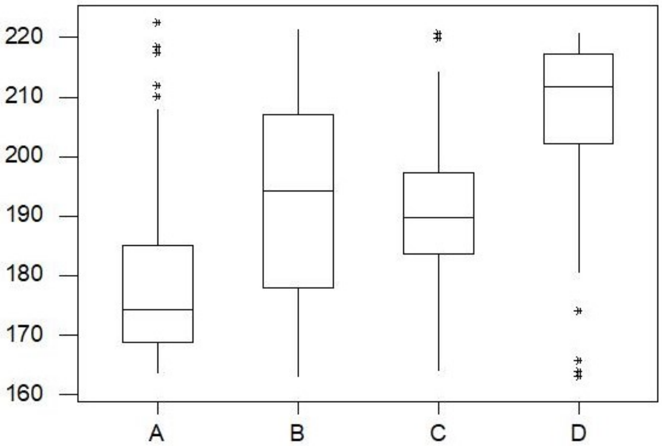
H_2



H_3



H_4



Boxplot

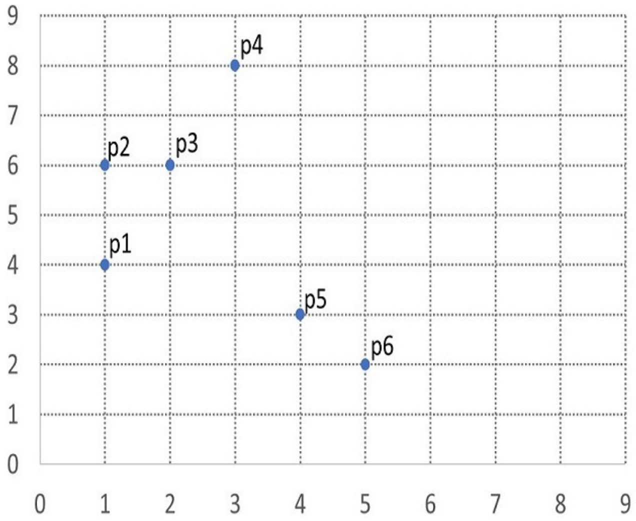
6. Richard is a new data wrangler. He uses a dataset about skin disease. For each patient, he recorded four features and one label saying whether disease severity is high or low. Three features: thickness, redness and scaling have a value of 0, 1, 2, 3, or 4. While the area feature is a percentage value which ranges from 0 to 100. Richard wants to provide doctors with an assessment for the importance of each feature in regard to its relationship with the severity label. He chooses to use NMI for feature ranking, where he computes the NMI between each feature and the severity label. The last row in the following table shows the calculated NMI scores:

Patient ID	Feature 1: area (%)	Feature 2: redness	Feature 3: thickness	Feature 4: scaling	Label: severity
205489					low
754330					high
...					...
NMI	0.14	0.57	0.82	0.74	

(a) (2 marks) Richard expects the lesion area to have the highest NMI score, based on background knowledge in dermatology. Surprisingly, lesion area has the lowest NMI score across other features (NMI = 0.14). Suggest two reasons that explain the mismatch between what Richard expects (high NMI score) and what he obtains (low NMI score) for the lesion area feature. State any assumptions made.

(b) (3 marks) Richard decides to find the strength of the relationship between feature1 and feature3 (area and thickness). He does not want to use NMI any more. But Richard is not sure whether Euclidean distance or Pearson correlation would work better in this case. To help Richard make the right decision, provide at least two limitations for each technique. Which technique you think would work better? Why?

7. (2 marks) Consider the following 2-dimensional data set. Apply ***one iteration*** of the K-means clustering algorithm (k=2) using squared Euclidean distance. Start with $p4$ and $p5$ as initial seeds, and compute the resulting clusters and their centroids.



8. For each of the following techniques, briefly explain i) two benefits, ii) two limitations:

   (a) (2 marks) Parallel coordinates

   (b) (2 marks) Principal component analysis

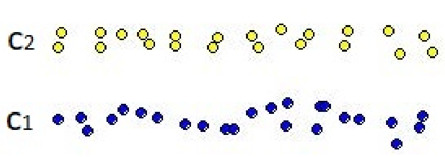
9. Given the following dissimilarity matrix for 6 objects:

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

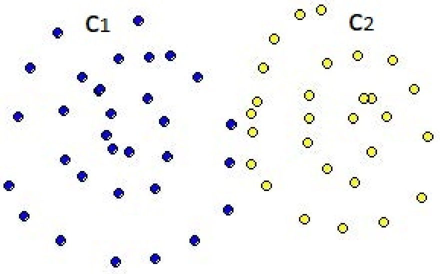
(a) (2 marks) Draw the dendrogram of hierarchical agglomerative clustering with single linkage after ***two iterations*** (i.e. two merges). Show all working.

(b) (1 mark) Explain briefly how a dendrogram can be used to split the objects into k-clusters.

10. (2 Marks) The Figures below show the result of clustering two datasets using either K-means or hierarchical agglomerative clustering methods (k=2). The blue circle objects are in cluster c1 and the yellow objects are in cluster c2. For each dataset, state the most likely clustering method to produce the shown output and explain why the method will work better on this dataset and the other will not.



(a) Dataset1



(b) Dataset2

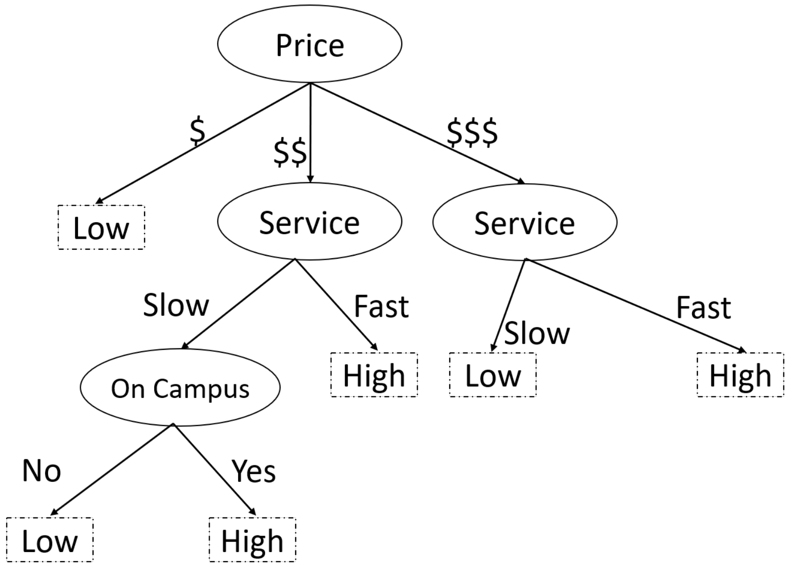
11. Alice and Bob are working on their project for COMP20008. They want to build a model to help students pick restaurants with the highest quality of food. They use three features: the price, the location of the restaurant and the speed of service. The training data is given below, where the column ”Food Quality” is the class label.

#	Price	Service	On Campus?	Food Quality
1	$	Slow	No	Low
2	$	Fast	Yes	Low
3	$$	Slow	No	Low
4	$$	Fast	Yes	High
5	$$	Fast	No	High
6	$$	Slow	Yes	High
7	$$$	Slow	No	Low
8	$$$	Fast	Yes	High

(a) (2 marks) What is the information gain of the attribute Price? (You may leave any logarithm terms unsimplified)

(b) (2 marks) Given a node with a set of training instances S, a decision tree learning algorithm recursively selects an attribute to split instances into subsets. Explain how to determine the best split and when to stop splitting.

After training stage, Alice and Bob obtain the following decision tree model:



(c) (2 marks) Given the following 5 test instances, use the trained decision tree to predict the label for each instance (i.e. fill in the ? cells).

Test Instances				Food Quality	
#	Price	Service	On Campus?	Actual Label	Predicted Label
1	$$	Fast	No	***High***	?
2	$$$	Slow	Yes	***High***	?
3	$	Slow	No	***Low***	?
4	$$	Slow	No	***Low***	?
5	$$	Fast	Yes	***Low***	?

(d) (2 marks) Use the Actual Label column in the above table to compute each of the following performance measures: TP, TF, TN, FN, and accuracy.

12. (1 mark) Describe one scenario where a 2-class k-NN classifier predicts all test instances as positive even though there are negative instances in the training and testing data.

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. Consider the 3 party protocol for privacy preserving linkage with approximate matching using bloom filters of length l and using k hash functions. As the ratio l/k increases:

    (a) (1 mark) Would you expect the matching accuracy of the system to become better or worse? Why?

    (b) (1 mark) Would you expect the robustness of the system to frequency attack (by the trusted 3^{rd} party) to become better or worse as l/k increases? Why?

15. (3 marks) Suppose we are using SHA-256 one way hashing function with the 3 party exact matching scheme for privacy preserving data linkage. Describe two drawbacks of this protocol and explain how to overcome these drawbacks.

16. Consider the quasi-identifier {marital status, gender} for the following dataset:

gender	**postcode**	**marital status**	***diagnosis***
male	1072	unmarried	HIV
female	1268	divorced	HIV
male	1276	unmarried	Hepatitis
female	1073	married	Hepatitis
female	1262	married	Chest pain
female	1077	divorced	Anemia

    (a) (1 mark) what is the maximal k for which it satisfies k-anonymity?

    (b) (1 mark) what is the maximal l for which it satisfies l-diversity?

    (c) (1 mark) Explain two disadvantages of using k-anonymity for protecting privacy.

17. (2 marks) Assume we have a dataset with three features: Gender, Resident and Education. Gender is a categorical feature with two possible values (Male, Female) and Resident is a categorical feature with two possible values (Yes, No) and Education is a categorical feature with four possible values (High-school, Diploma, Undergraduate, Postgraduate). If we are interested in 'count' queries, what is the value of global sensitivity? Explain your answer.

End of Exam