The University of Melbourne

Semester One 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 8 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 15 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. (2 marks) Given the following JSON fragment, convert it to an XML representation. The first line of your XML should be `<?xml version="1.0" encoding="UTF-8"?>`

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

2. Given two instances represented by the tuples $(3, 2, 2, 3, 1, ?, 7)$ and $(?, 4, 6, 5, 9, ?, ?)$

   a) (1 mark) Write an expression for the Euclidean distance between these two instances using mean imputation (Method 1 in lectures).

   b) (1 mark) Write an expression for the Euclidean distance between these two instances using scaling (Method 2 in lectures).

   In both 2a) and 2b), you may leave any square root terms unsimplified.

3. (3 marks) Consider the following claim "Data cleaning cannot be done by computers alone, as they lack the domain knowledge to make informed decisions on what changes are important. On the other hand, manual approaches are time consuming and tedious."

   Using a real world example, explain why it is difficult to fully automate data cleaning and explain why domain knowledge is needed.

4. (3 marks) University X is planning to build a recommender system for its students. Based on subjects they have enrolled in, the system will recommend new subjects they might consider studying in the future.

   A table showing a fragment of the data input to the system is below. The columns correspond to the codes of all the subjects in the University handbook. Rows correspond to students and whether they have enrolled in a subject ("Yes" if they have previously enrolled and "-" otherwise). The dataset covers the period 2010-2017, with 100,000 students and 3000 subjects. It is proposed that subject recommendations should be made using user based collaborative filtering.

PersonName	Subject1	Subject2	Subject3	Subject4	Subject5	...
Alice	Yes	-	-	-	-	...
Bob	Yes	Yes	-	Yes	-	...
Margaret	Yes	Yes	-	-	-	...
...	...	...	...	...	...	...

   Explain three challenges for making this recommendation approach effective.

5. a) (2 marks) What is classification? What is the difference between classification and k-means clustering?

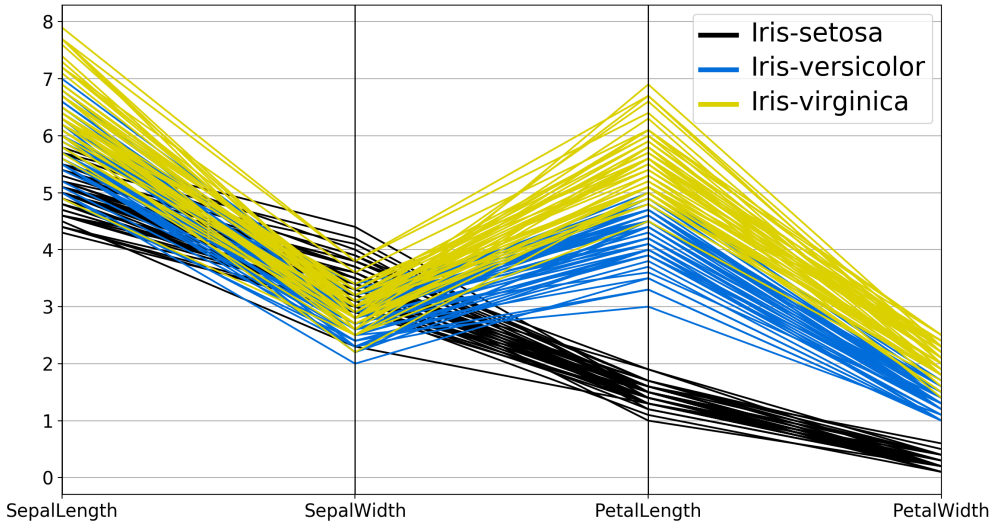
   b) (3 marks) Sketch how the k-means algorithm could be used to perform hierarchical clustering (i.e. construct a dendrogram).

6. (2 marks) Consider a dataset D, with 10 features and 500 instances. Each feature has been normalised so its values are in the range 0 to 1. Suppose we construct a dissimilarity matrix for D and then visualise it as heat map.

   Explain how this visualisation could be used to identify outliers.

7. (3 marks) Suppose we have the Iris dataset (discussed in lectures) which has 150 instances, each of which corresponds to a flower, where the features are *Sepal-Length*, *Sepal-Width*, *Petal-Length* and *Petal-Width*, and where there are three classes of flower: *Iris-setosa*, *Iris-versicolor* and *Iris-virginica*. A parallel co-ordinates visualisation plot for this dataset is shown below.

   Explain three inferences which can be made more easily from this visualisation, compared to browsing the raw CSV data.



8. Given a dataset with three classes, A, B and C, suppose the root node of a decision tree has 50 instances of class A, 50 instances of class B and 100 instances of class C. We are evaluating a candidate split of this root node into three children using a categorical feature F that has three possible values. The first child has distribution (25 class A, 25 class B, 25 class C), the second has (20 class A, 5 class B and 0 class C), the third has (5 class A, 20 class B, 75 class C).

   a) (3 marks) Based on 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.

   b) (2 marks) Explain the relationship between the split utility and mutual information.

9. (2 marks) Alice collects a dataset D about students enrolled in subject X during 2015. It includes 4 features that record performance in subject X coursework assignments: *Mark in Project 1*, *Mark in Project 2*, *Mark in Project 3* and *Mark in Project 4*. There is is a binary class label recording subject X exam performance with categories: *Honours grade in exam* or *Non-Honours grade in exam*. Alice collects another dataset D', which has the same features and class as D, but is for students enrolled in subject X during 2016. Alice also collects a third dataset D'', which has the same features and class as D, but is for students enrolled in subject X during 2017.

   If a decision tree is trained using all data in D and then evaluated using all instances in D' as testing data, the accuracy is 90%. If a decision tree is trained using all data in D and then evaluated using all instances in D'' as testing data, the accuracy is 70%.

   Suggest an explanation for this large difference in accuracies and state any assumptions made.

10. (3 marks) Consider the following steps of the k-NN algorithm to classify a single test instance

```
1. Compute distance of the test instance to each of the instances
   in the training set and store these distances
2. Sort the calculated distances
3. Store the K nearest points
4. Calculate the proportions of each class
5. Assign the class with the highest proportion
```

    Step 1 may be very slow if the training set is large.

    Suggest three possible strategies that might be used to speed up this step and describe a disadvantage of each.

11. Consider the following property of blockchains: "Once a record is entered into the ledger, it is virtually impossible to alter it".

    a) (2 marks) Explain two mechanisms of blockchains that help ensure this property.

    b) (2 marks) Explain, using a real world example, why this property is an advantage.

    c) (2 marks) Explain, using a real world example, why this property could also be a disadvantage.

12. Business X and Business Y have decided to conduct a joint marketing campaign. For this marketing campaign, they need to determine how many customers they have in common (how many people are in the customer list of both businesses). They implement the following 2 party privacy preserving protocol, making use of the SHA-256 one way hashing function.

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

#Business X does the following
SetX=empty
For each customer at Business X
    SetX=SetX $\cup$ SHA-256("First Name"+"Last Name")
Send SetX to Business Y

#Business Y does the following
SetY=empty
For each customer at Business Y
    SetY=SetY $\cup$ SHA-256("First Name"+"Last Name")
result=count(SetX $\cap$ SetY)
Share result with Business X

a) (2 marks) Explain a privacy drawback of this protocol.

b) (2 marks) Explain how the protocol could be modified to eliminate this privacy drawback.

13. Suppose you are conducting data linkage between two databases, one with m records and the other with n records (assume $m < n$). You are using a blocking strategy.

a) (2 marks) Assuming that each record is assigned to exactly one block and that the blocking strategy uniformly distributes records across blocks, show that a total of mn/b record comparisons will be needed on average.

b) (1 mark) Explain why this formula (mn/b record comparisons) would not be applicable if a record could be assigned to multiple blocks.

c) (2 marks) Explain an advantage of assigning a record to multiple blocks, instead of to a single block. Explain a disadvantage.

14. (2 marks) The owner of a pharmacy "Drugs are Us" wishes to publically release the following dataset (a fragment is shown below). This dataset has 3 features and records all purchases of medicine and other products at the pharmacy during 2017.

Date of purchase	Time of Purchase	Cost of Purchase
...	...	...
11/1/2017	08:23	$10.50
...	...	...
2/3/2017	17:20	$18.37
...	...	...
11/11/2017	08:53	$23.50
...	...	...

Describe a privacy attack that could be made on this dataset. In your answer, include description of any knowledge or capability that the attacker would need to have, what private information the attacker could infer, as well as any assumptions made.

15. (3 marks) Given a user query to be applied on a dataset, differential privacy involves adding noise to the true result for the query, and then returning the noisy result to the user.

Explain two factors which influence how much noise should be added to the query result. For each factor, you should explain how it is related to the level of noise that gets added.

End of Exam