The University of Melbourne

Semester One 2017 Sample Exam

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

Exam Duration: 2 hours

Reading Time: 15 minutes

This paper has 6 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 5 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

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

1. a) (3 marks) What is the difference between XML and HTML ? When would it would be more appropriate to use XML instead of HTML ? When would it be more appropriate to use HTML instead of XML ?

   b) (2 marks) List two advantages of using JSON to represent data, compared to using a CSV file representation.

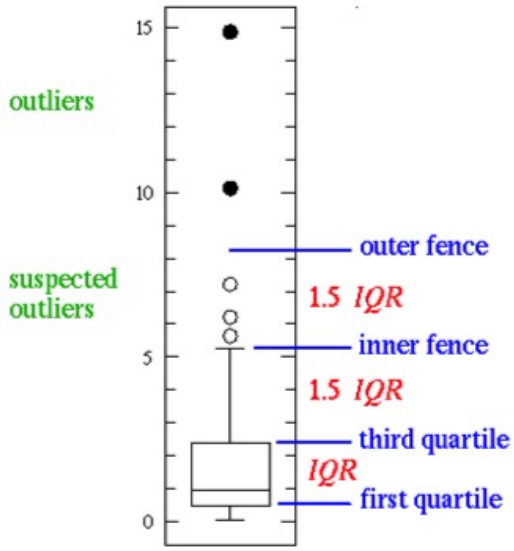
   c) (2 marks) Explain two advantages of using blockchain technology for storing data, compared to storing information in a single, centralised database.

   d) (3 marks) Imagine the following dataset, recording the ages of students enrolled in a hypothetical computer science subject at the University of Melbourne:

StudentID	Gender	Age
1	Male	22
2	Female	29
3	Female	29.5
4	Female	32.5
5	Female	28.5
6	Male	20
7	Male	20.5
8	Male	20.3
9	Male	?

   Write the formula that would be used to impute the value of the "?" using the category (gender) mean. Explain why this would be more appropriate than using the mean for all students. What could be a disadvantage?

2. a) (3 marks) Consider the following box plot. Explain the meaning of each of the following items and the formula used to calculate it: *outliers, suspected outliers, outer fence, inner fence, third quartile, first quartile.*



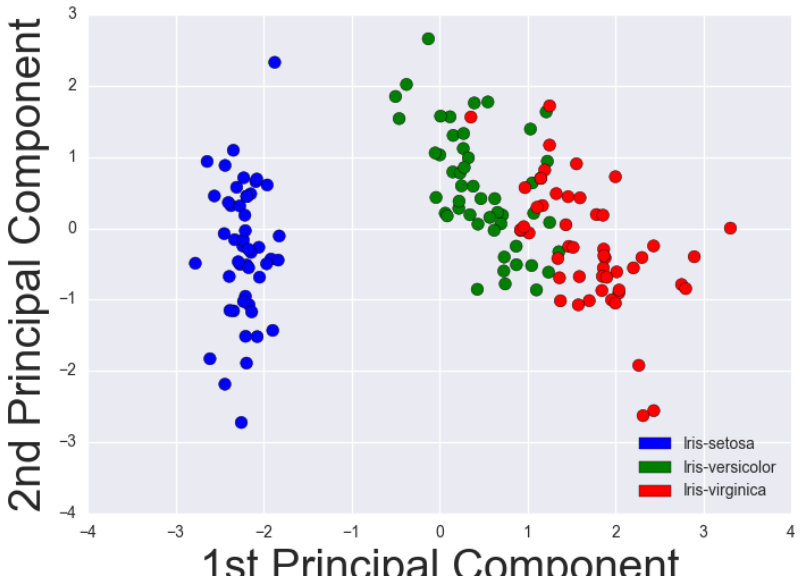
b) Imagine a dataset about the 3 million people living in Melbourne. Its features are *Age (years), Height (metres), Weight (kilograms)* .

i) (3 marks) Alice wants to know whether the correlation between Age and Height, is stronger than the correlation between Age and Weight. Which measure would be more appropriate to use for this task, Pearson correlation coefficient or normalised mutual information? Explain your reasoning and justify any assumptions made.

ii) (2 marks) Bob tells Alice that before computing any correlations, she should first remove outliers from the dataset (throw away all records with Age greater than 90, or height greater than 1.9 metres, or weight greater than 110 kg.). Is this a good idea? Explain why or why not and any assumptions made.

iii) (2 marks) Next, the dataset is processed, such that each feature is separately normalised so its values are now in the range $[0, 1]$. Explain two possible benefits of this processing step.

3. 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 three are classes of flower - setosa, versicolor and virginica. Suppose we apply PCA to this dataset and obtain the following plot.



a) (3 marks) Explain why it was useful to apply PCA. What are the benefits of the PCA plot?

b) (3 marks) Suppose instead we had taken the *Iris* dataset, computed a dissimilarity matrix and then applied the VAT algorithm. Compared to PCA, what are the differences in information that might be revealed or not revealed?

c) (3 marks) Suppose Alice takes some other dataset D with 4 features, class labels and 100 instances. She computes the correlation of each feature with the class label using mutual information and discards the two features with lowest correlation. She now has a processed version D' of the dataset (2 features, class labels and 100 instances). She splits D' into two - 80% training (80 instances) and 20% testing (20 instances). She learns a decision tree model on the training set and evaluates the model accuracy on the testing set. She reports the accuracy as being 90%.

Why might this estimate of 90% accuracy be over-optimistic? Give reasons.

d) (1 mark) Consider the k nearest neighbor classification method. Describe how the parameter k might be chosen in practice.

4. For each of the following concepts, briefly explain i) its meaning, ii) why the concept is important.

   a) (2 marks) sharding (in the context of NoSQL databases)

   b) (2 marks) blocking (in the context of record linkage)

   c) (2 marks) digital signature (in the context of blockchain data)

   d) (2 marks) l-diversity (in the context of privacy)

   e) (2 marks) k-anonymity (in the context of privacy)

5. a) (2 marks) Given the strings *wrangling* and *wrapping*, compute their (approximate) similarity using 2-grams. Show all working.

   b) (2 marks) Suppose a 14 bit bloom filter is used as a privacy preserving representation for strings. The bloom filter representation of *wrangling* is 11000010001111 and the bloom filter representation for *wrapping* is 11100001101111. What is the approximate similarity of these strings using the bloom filter representation?

   c) (3 marks) Explain how bloom filters can help provide a private representation for strings, in the context of privacy preserving data linkage.

   d) (3 marks) Consider the 3 party protocol for privacy preserving linkage with exact matching, discussed in lectures. What is a salt? Explain why a salt is used with the hash functions. Who chooses and knows the salt?

End of Exam