# Sample/practice exam June 2016, questions and answers

Elements of Data Processing (University of Melbourne)

# The University of Melbourne

# Semester One 2016 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: $H(\mathbf{X}) = -\sum_{i=1}^{\#bins} p_i \log_2 p_i$
where $p_i$ is the proportion of points in the $i$th cell.

Joint entropy: $H(\mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{\#binsX} \sum_{j=1}^{\#binsY} p_{ij} \log_2 p_{ij}$
where $p_{ij}$ is the proportion of points in the $\{i, j\}$th cell.

Mutual information:
$MI(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{\#binsX} \sum_{j=1}^{\#binsY} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j}$

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

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.

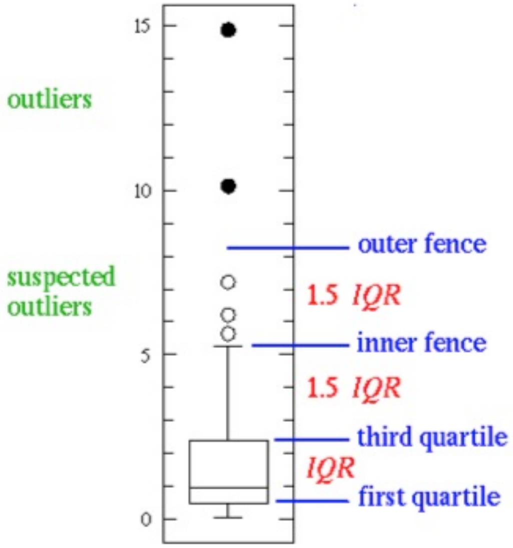
c) (2 marks) Two students, Bob and Alice, are having a conversation about data wrangling. Alice argues that "UNIX is fantastic for data wrangling, because it provides pipes, as well as standard input and output redirection facilities". Provide reasons to support Alice's claim, using concrete examples.

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.*



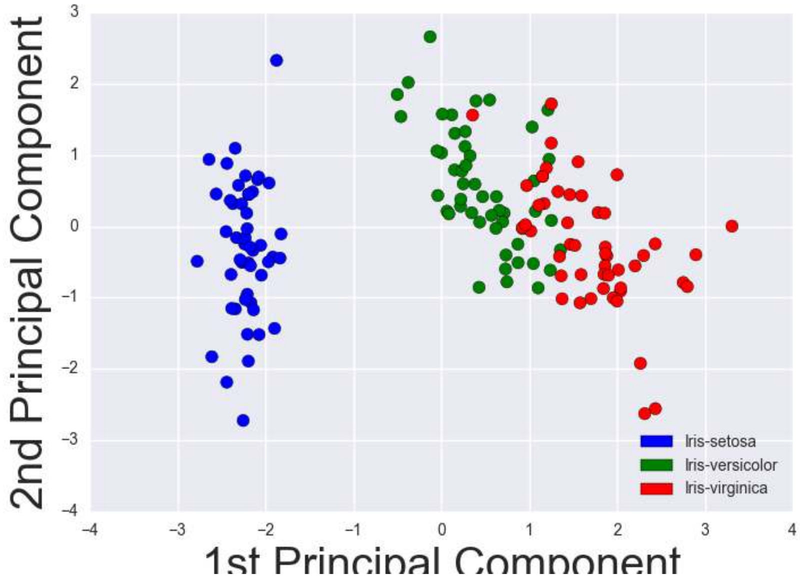
b) Imagine a dataset about the 3 million people living in Melbourne. Its features are *Age (years), Height (metres), Weight (kilograms)* .

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) Map-Reduce (in the context of distributed systems)

   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

Some sketch answers and pointers for the sample exam.   These are not the only ways to answer each question,
but provide some idea of points to consider.
--------------------------------------------------------------

1a) XML allows one to define vocabulary of elements/attributes, whereas for HTML this is fixed
and uses a vocabulary suitable for presenting in a browser.  XML better for generic applications
(e.g. ChemML) not dependent on presentation.  HTML well suited to presentation.   [Note that HTML
is a subset of XML, so one could argue that when using HTML one is also using XML]

b) JSON relatively easier for data interchange by Web servers, more natural to represent hierarchial
data than CSV.

c) Could consider a pipeline

```
E.g. grep xxx bigfile | wc >outputfile
```

How might this be done without a pipe or output redirection in Windows?   Would it be easier or harder and
why?   Consider how it generalises to multiple programs

```
prog1<inputfile|prog2|prog3|grep xxx|wc>outputfile
```

d) Compute mean of column age, restricted to rows which are Male.  Such a mean incorporates the background
knowledge about the missing value's gender, providing a more specific estimate.   However sample size is lower
and estimation may be less reliable.

2a) Follow slide 27 of lecture 6

bi) Need to explain whether relationship between variables is linear or non linear and why.   For a non linear
relationship such as Age and Height, the (normalised) mutual information is more appropriate, since
Pearson correlation won't be able to detect it.   Could draw your estimate of the relationship (curve) between
a person's Age and Height to support the reasoning

bii) Could argue either way here.    The examples given are not very extreme, so throwing the information
away could be viewed as harsh and might result in misleading analysis.    It would of course depend
on the population being analysed (the range of values for people in Japan would be very different from those
in the USA)

biii) See conversation in discussion forum

3a) Allows immediate visualisation of the dataset.   Helps show the cluster structure, helps show the overlap between classes, helps identify potential anomalies, extreme individuals from each class.

b) VAT might reveal more clearly how many clusters there are and their respective sizes. More difficult to relate VAT info to class structure.   VAT provides less idea about *why* an instance is different/similar to other instances.

c) The 90% estimate would be biased, since the testing data (class label info) was looked at when doing feature selection. This provided information to the feature selection process that should not have been seen.  (like seeing the final exam before it is held).   Consequently the model that was trained using the results from the feature selection was developed on information that should not have been seen.   The reported accuracy will thus likely be over optimisitic.

d) Could be domain knowledge, or evaluating accuracy using different choices of k and choosing the one that works best.

4) a) slide 15 of lecture 12

b) Technique used to improve efficiency in record linkage of large datasets.  Blocks are formed based on some property of each record (e.g. first letter of surname), then only blocks with matching properties are compared.

c) A programming  model for processing large datasets in a distributed fashion. Processing is broken down into two steps Map() and Reduce(). Programmers just need to write code for each of these functions and do not have to concern themselves with the details of the infrastructure on which they are run.   Makes development of (some) distributed applications much easier.

d) A model for data anonymisation, following on from k anonymity (an individual should be indistinguisable from at least (k-1) other individuals on the non sensitive attributes).  Furthermore, there should be at least l different values for the sensitive attribute.    This reduces the risk of privacy attacks on data which only satisfies k anonymity.

e) A table satisfies k- anonymity if every record in the table is indistinguishable from at least k − 1 other records with respect to every set of quasi-identifier attributes; such a table is called a k- anonymous table.

5a) Break each string into its two grams, e.g.
wrangling-> wr, ra, an, ng .....

wrapping-> wr, ra, ap ....
Use Dice coefficient for similarity
Lecture 17 slide 11

b) Lecture 17 slide 21

c) Explain how string information is represented in bloom filter (generation of 2 grams, hashing of 2 grams with multiple hash functions).  Explain how a single bloom filter might map to multiple possible input strings, can't easily reverse engineer.

d) Extra string that is appended to the information being encoded, so that hashed value is not susceptible to a dictionary attack (need to explain dictionary attack).   The two parties doing the linkage would agree on a salt, the 3rd party would not know it.