UNIVERSITY OF OTAGO EXAMINATIONS 2019

Information Science

INFO204

Introduction to Data Science
Semester Two

(TIME ALLOWED: 2 HOURS)

This examination paper comprises 6 pages.

Candidates should answer questions as follows:

This exam has two sections (A and B). Follow the instructions given in each section and answer *all* questions.
The exam is marked out of 100. It is worth 50% of the course.

The following material is provided:

Nil

Use of calculators:

Any model of calculator may be used provided it is battery powered, silent, truly portable and free of communication capabilities.

(Subject to inspection by the examiners)

Candidates are permitted copies of:

Nil

Other Instructions:

Nil

TURN OVER

Section A

ANSWER ALL QUESTIONS (TOTAL 50 MARKS).
Allocate between 5 and 7 minutes to each question.

1. Explain the interdisciplinary nature of *data science* and identify its main purposes. (4 marks)

2. Give *four* approaches that we may use to avoid overfitting in machine learning scenarios. (4 marks)

3. What are the main components of the statistical learning framework? Draw a diagram and explain how they are related to each other. (5 marks)

4. Explain the bias-variance trade-off when selecting models. (5 marks)

5. Outline the basic idea of principal component analysis and explain why it is useful for data visualisation. (5 marks)

6. What are the benefits of conducting feature selection? Outline *three* major approaches for feature selection. (6 marks)

7. Clustering data relies on the use of a distance metric to determine if data points are close or distant from each other. State *two* of the criteria for a distance metric and explain what each criterion means. (4 marks)

8. In text mining, a "Bag-of-Words" (BoW) is one possible output from pre-processing raw text. What is a BoW and how is it subsequently used for feature construction for a supervised machine learning task? (5 marks)

9. Answer the following questions:

   (a) What are the *two* principal methods by which recommender systems work? (2 marks)

   (b) What type of question is each method best suited to answering? Provide examples of how they are used by companies that apply such approaches. (4 marks)

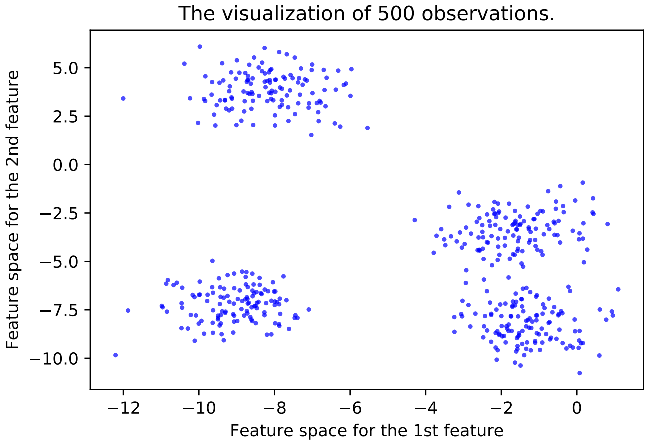
10. The ethical considerations pertaining to data science are somewhat different to those required for more traditional data processing scenarios. Using appropriate examples, discuss *two* problems relating to ethics that are prominent in data science, and identify strategies that may be used to manage these problems. (6 marks)

[SECTION A TOTAL 50 MARKS]

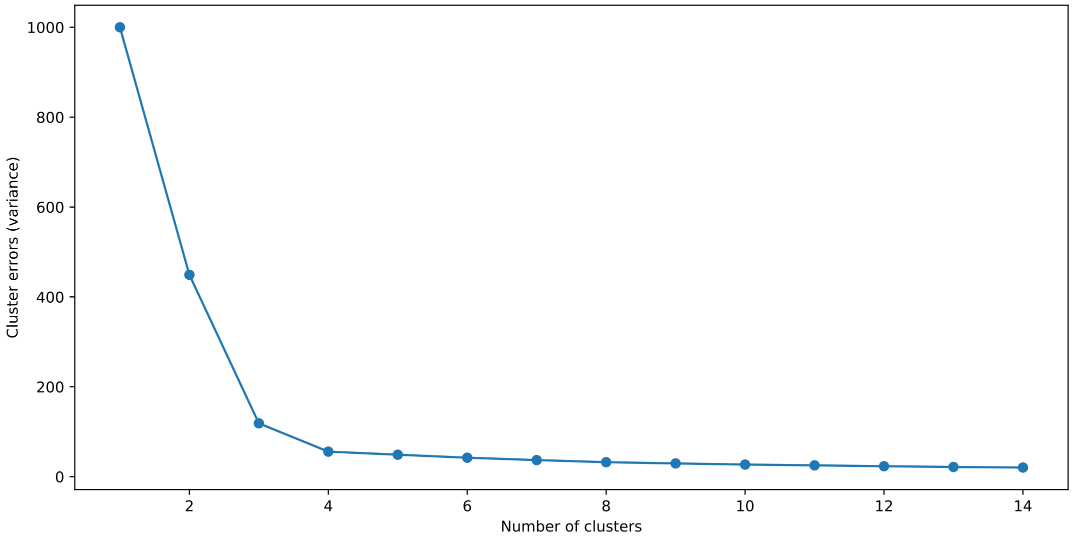
Section B

ANSWER ALL QUESTIONS (TOTAL 50 MARKS).

Each question should be answered within one or two pages of your answer book.



(a)



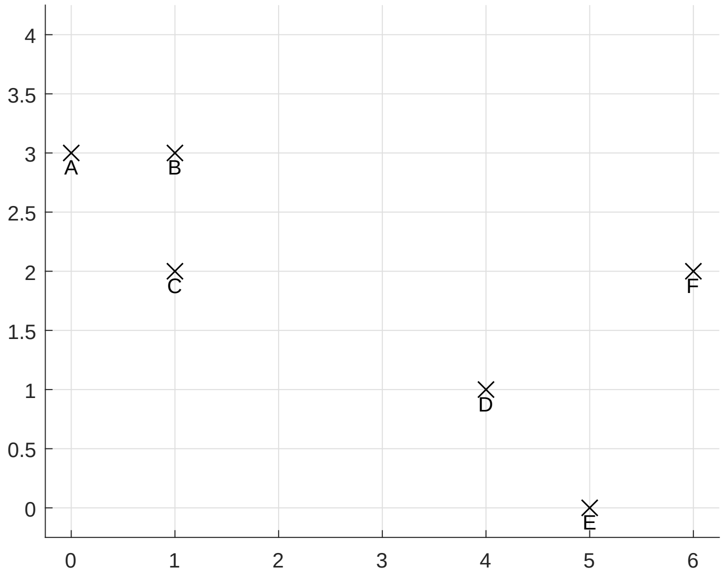
(b)

11. You have been provided with a data set that is shown above in Figure (a) and your task is to find clusters in it using k-means clustering. To determine an appropriate value for k you have performed an Elbow Analysis on the data set and the results of this analysis are shown in Figure (b).

    (a) What is the objective of Elbow Analysis and how are the cluster errors in Figure (b) related to the number of clusters? (4 marks)

    (b) What would be a good recommendation for an appropriate value for k given the result of Elbow Analysis and your interpretation of the number of clusters in Figure (a)? Explain your reasoning. (2 marks)

12. Manually perform k-means clustering on the following points for $k = 2$: $(0,3)$, $(1,3)$, $(1,2)$, $(4,1)$, $(5,0)$, $(6,2)$. These points, labelled ‘A’ to ‘F’, are shown below:



Initially, data points $D = (4,1)$ and $E = (5,0)$ are chosen as cluster centres. Use the following procedure to carry out the clustering algorithm and show the outcome.

(a) For iteration 1, calculate the coordinates of the cluster centres, and report cluster membership of all data points. Reproduce the plot, and roughly indicate the locations of the cluster centres and the scopes of the clusters. (4 marks)

(b) Do the same as above for iteration 2. (4 marks)

(c) How many iterations does it take for the algorithm to converge? Report the final clustering membership and the locations of cluster centres. (4 marks)

13. In a medical classification problem, there are three classes labelled as “A”, “B”, “C”. The prior probabilities are known: $P(A) = 0.1, P(B) = 0.2, P(C) = 0.7$. For an input x to take a “High” value based on belonging to one of these three classes, the conditional probabilities are: $P(H|A) = 0.8$, $P(H|B) = 0.3$, and $P(H|C) = 0.1$, respectively. Answer the following questions and *show your working*.

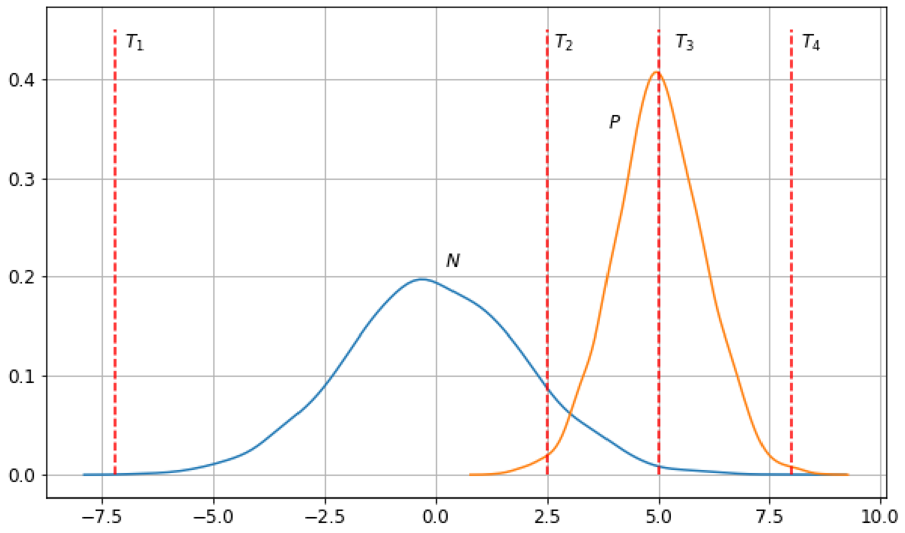
(a) What is the joint probability, for x to be High and also belong to class A, i.e. $P(H, A)$? (2 marks)

(b) How much is $P(H)$, the overall probability for x to be High, across all three classes? (2 marks)

(c) Given that x is High, what is the probability for it to belong to class C, i.e. $P(C|H)$? (3 marks)

(d) Clinically, class A means “malignant”, while classes B and C mean “benign”. Given x is High, what will be your diagnosis? (3 marks)

14. In a classification problem, there are two classes, “negative” (N) and “positive” (P). Their probability density functions are shown below.



A classifier using a threshold value is adopted. Given threshold T, all cases with $x < T$ are classified as “N”, otherwise as “P”. Four possible thresholds, T_1, ..., T_4, are indicated by the vertical dash lines.

(a) Estimate the false positive rates (FPR) and the true positive rates (TPR) for thresholds $T_1, \cdots, T_4$. (4 marks)

(b) Using the TPR and FPR values, plot out a likely “receiver operating characteristic” (ROC) curve for the classifier. (4 marks)

(c) Within the ROC plot, indicate what AUC stands for. (2 marks)

15. Reflect on the data science problem you dealt with in Assignment 1, which consisted of a few data analysis and modelling processes performed on a cartographic dataset. Describe the overall approach, and the specific algorithms used for data exploration and problem-solving. How can your computational model(s) be configured, validated and tested?

(12 marks)

[SECTION B TOTAL 50 MARKS]

END