UNIVERSITY OF OTAGO EXAMINATIONS 2019

Information Science

INFO411

Machine Learning and Data Mining
Semester Two

(TIME ALLOWED: 3 HOURS)

This examination paper comprises 5 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 40% 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 9 minutes to each question.

1. (a) Explain the neural network model of self-organizing maps (SOM) with the aid of a diagram, and describe its learning rule. (3 marks)

   (b) During the learning process, what are the key parameters of a SOM model and how should they evolve over time? (2 marks)

2. Why would the L_1 norm be useful in generating sparse models? Give an example. (4 marks)

3. Outline *two* methods of feature selection for classification problems. (6 marks)

4. Outline an incremental learning algorithm that finds the optimal k value for the k-means algorithm. (6 marks)

5. Suppose we need to deal with live streaming data, $X = \{\mathbf{x}_t\}$, with $\mathbf{x}_t$ being the input vector arriving at time t.

   (a) Design a learning rule that derives the online mean vector (denoted by $\mathbf{m}$) of the input data X. (3 marks)

   (b) Design a learning algorithm that carries out online principal component analysis (PCA) for data visualisation. (3 marks)

6. Describe how the Bayesian Information Criterion (BIC) is used in X-means clustering to estimate the true number of clusters in the underlying distribution of a dataset. (5 marks)

7. What are the advantages and potential disadvantages of using the area-under-the-curve (AUC) as a performance index? (4 marks)

8. How is a K-Fold Cross Validation Paired t-Test used to compare the performance of two classifiers? (4 marks)

9. When cascading is used to combine multiple classifiers, what impact does it have on the overall complexity of the resulting model? (4 marks)

10. Compare and contrast the design of a Convolutional Neural Network (CNN) with that of a traditional feed-forward neural network. Discuss the similarities and differences of the models, making particular reference to the potential benefits that CNNs have over feed-forward neural networks. (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.

11. (a) What factors may affect the performance of the *k*-means algorithm? (4 marks)

    (b) Suppose we are handling a genetic dataset that contains 10 million rows and 1000 columns of data. Suggest a few possible technical solutions for using *k*-means to carry out cluster analysis. (4 marks)

12. Design an online algorithm for stream data clustering. Propose possible mechanisms for configuring the key parameters of the algorithm, and outline a scheme to validate your algorithm. (10 marks)

13. A probabilistic classifier reports the posterior probability $P(+|x)$ for a two-class ("+"/"−") dataset. The outcome for eight instances is given in the following table:

Instance x	1	2	3	4	5	6	7	8
Class ground truth	+	+	+	−	−	−	+	−
$P(+\|x)$	0.75	0.55	1.00	0.35	0.95	0.65	0.85	0.45

    A threshold is then employed to classify the instances into either of the two classes.

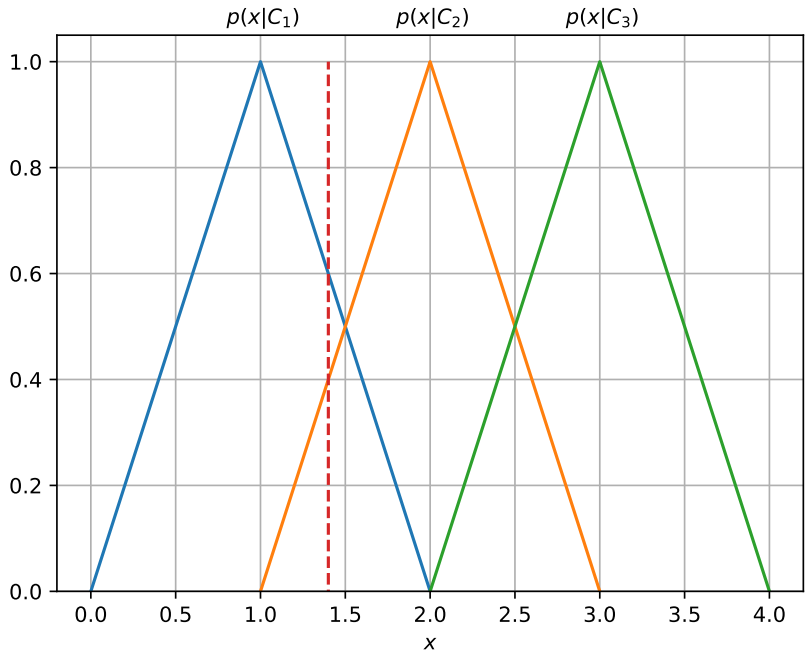
    (a) Use a new table to calculate classification performance metrics necessary for ROC generation. (5 marks)

    (b) Draw the ROC curve. (3 marks)

14. Give an example of problem-solving you have performed using machine learning techniques. Describe the overall procedure, and the necessary techniques (including algorithmic and methodological details) for relevant data analysis, modelling and validation. (12 marks)

15. A chemical testing process reports a measurement x with three modes (denoted by C_1, C_2, C_3). Their probability density functions, $p(x|C_1)$, $p(x|C_2)$, and $p(x|C_3)$, are triangular, centred at 1.0, 2.0, and 3.0 respectively. For each mode, the density values outside the triangle are all zero. It is known the prior probabilities are all equal:

$$P(C_1) = P(C_2) = P(C_3) = 1/3.$$



There is a testing case, where the x value ($x = 1.4$) is indicated by the dashed vertical line shown above.

(a) Give the formula for calculating $P(C_i|x), i = 1, 2, 3$. (3 marks)

(b) Give the discriminant function $g_i(x)$ that classifies x into one of the three modes. (2 marks)

(c) Based on the discriminant function, what is your classification decision? Show your working. (3 marks)

(d) After collecting more samples, it is found that the priors need to be modified to

$$P(C_1) = P(C_3) = 1/6, P(C_2) = 2/3.$$

Do you need to change your classification decision for x? Show your working. (4 marks)

[SECTION B TOTAL 50 MARKS]

END