Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Machine Learning

Resit Examinations 2019

Time allowed: 1:30

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1

(a) Logistic regression, linear discriminant analysis, and k-nearest neighbours are three machine learning methods.

(i) What type of machine learning problem are these methods used to solve?

(ii) With reference to the other two methods, list one advantage and one disadvantage of each method.

[9 marks]

(b) In a problem that is being solved using LDA, there are two classes A and B of data points in a two-dimensional space. A and B are normally distributed with identical covariance, and are centred at $A = (-3, 0)$ and $B = (3, 0)$ respectively. The two classes each contain the same number of data points.

(i) By considering the symmetry of the problem, approximately where is the class boundary and what shape is it?

(ii) What would happen to the class boundary if class A contained twice as many points as class B, whilst maintaining the same covariance?

(iii) If the two classes have different covariance, how should the approach be adjusted, and what would the effect on the decision boundary be?

[6 marks]

(c) A dataset is to be used for a machine learning task in which the goal is to predict the risk of a person developing diabetes. The variables in the dataset are shown in the following table together with information about their possible values. "Diabetes Risk" is the dependent variable that is to be predicted.

What type of machine learning problem is this?

Explain what steps should be considered in order to prepare this dataset for machine learning?

Variable	Values
Age	Minimum: 32 Mean: 55 Maximum: 76
Height (metres)	Minimum: 1.45m Mean: 1.75m Maximum: 2.01m
Weight (kg)	Minimum: 43kg Mean: 82kg Maximum: 125kg
Exercise (discrete)	<1 hour/week 1-3 hours/week 3-6 hours/week >6 hours/week
Diabetes Risk (Discrete)	Low risk Medium risk High risk

[5 marks]

Question 2

(a) Sketch an example of an unlabelled dataset containing two clusters that cannot be separated by k-means clustering with $k = 2$ using Euclidean distance, and explain why this is the case.

Suggest a clustering method that would work on this dataset, explaining what options you have considered and why you have chosen one in preference to the others.

[8 marks]

(b) High dimensional space can cause problems in classification and clustering problems. Explain why this is the case and suggest how these problems can be overcome.

[6 marks]

(c) Explain the Adaboost algorithm.

[6 marks]

Question 3

(a) Explain the meaning of *bias* and *variance* in the context of a regression problem, illustrating your answer with appropriate diagrams.

[6 marks]

(b) The regularised least-squares loss function can be written as

$$\mathcal{L}(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}(\mathbf{x}, \mathbf{w})\|^2 + \lambda R(\mathbf{w})$$

Explain the purpose of the term $R(\mathbf{w})$.

In a regression problem, it is decided that it would be desirable for $\sum_i w_i^2 = 1$. Write down a regularisation term $R(\mathbf{w})$ that would encourage this, and explain your choice.

[8 marks]

(c) Describe a procedure for selecting the optimal value of the regularisation parameter λ, sketching any relevant graphs and explaining your reasoning.

[6 marks]

End of Paper

This page intentionally left blank.

Machine Learning

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.