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 2020

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) In the notation used in the lectures, the quantities needed to solve a univariate unregularised least squares regression problem are:

- The vector of *independent* variables $\mathbf{x}$ with components $\{x_i\}_{i=1}^{N}$.
- The vector of *dependent* variables $\mathbf{y}$ with components $\{y_i\}_{i=1}^{N}$.
- The vector of model parameters $\mathbf{w}$ with components $\{w_i\}_{i=1}^{M}$.
- The basis states $\{\phi_i(x)\}_{i=1}^{M}$.

Explain how to construct the *normal equations* for unregularised regression from these quantities. You do *not* need to derive the normal equations from first principles. **[5 marks]**

(b) Explain the meaning of *bias* and *variance* in the context of a regression problem, illustrating your answer with appropriate diagrams. **[7 marks]**

(c) Explain the principle of regularisation and write down the general expression for the regularised least-squares loss function. Give two examples of regularisation functions and explain their effect. **[8 marks]**

Question 2

(a) A *decision stump* is a decision tree containing only one split on the most informative variable. Using the principle of maximising information gain, determine which variable should be used to form a decision stump for the data shown in the table below. **[5 marks]**

x_0	x_1	x_2	y
0	1	1	0
0	0	1	1
1	1	0	0
1	0	1	1

(b) Explain the random forest algorithm for classification. **[7 marks]**

(c) A *labelled* dataset contains 500 samples, each of which is from a 5-dimensional space. It is known that there are three (3) classes of data (A, B, C) in this dataset and each sample is drawn from one of those classes. The number of training points in each of the classes is A: 50; B: 250: C:200 . The classes are known to not be fully separable by three hyperplanes.

Explain how you would choose an algorithm to classify this dataset, what difficulties may be encountered, and how you would overcome them. **[8 marks]**

Question 3

(a) The Johnson-Lindenstrauss lemma can be stated as:

$$1 - \varepsilon \leq \frac{\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2}{\|\mathbf{x}_1 - \mathbf{x}_2\|^2} \leq 1 + \varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^M$; $f : \mathbb{R}^M \mapsto \mathbb{R}^K$; $0 < \varepsilon < 1$ and $K < M$.

Explain the implications of this lemma and their relevance to machine learning. **[6 marks]**

(b) The table below contains a set of data with two variables. Each column contains one datapoint. *Sketch* the dendrogram for agglomerative hierarchical clustering using single-linkage on this dataset.

x_0	2.0	2.0	3.0	2.0	1.0	5.0	6.0
x_1	1.0	2.0	3.0	5.0	6.0	6.0	6.0

[7 marks]

(c) A common modification to the *k*-nearest neighbours algorithms is the *weighted k*-nearest neighbours algorithm.

(i) Describe how the *weighted k*-nearest neighbours algorithm works.

(ii) Sketch one example of a situation in which this method will give rise to an incorrect decision. Explain your reasoning.

[7 marks]

This page intentionally left blank.

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.