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

Main Summer 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) Explain what is meant by the terms "supervised" and "unsupervised" learning.

List two types of supervised learning problems and explain the similarities and differences between them. You do not need to provide details of specific algorithms.

[6 marks]

(b) The expected value of the least squares loss-function is

$$\mathbb{E}[\mathcal{L}] = \sigma^2 + \text{var}[f] + (h - \mathbb{E}[f])^2$$

where h is the true function underlying the data, f is an estimate of h and σ is the standard deviation in the data.

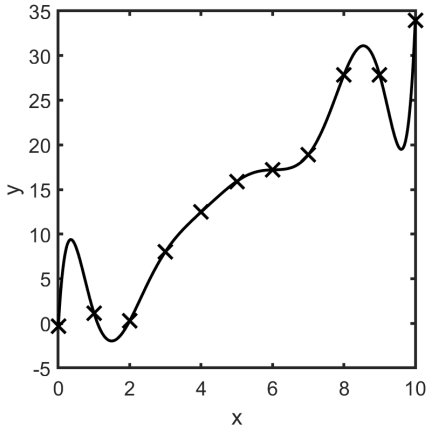
Explain the terms in this expression and their implications for learning.

[7 marks]

(c) The graph below shows the fit of a curve $f(x) = \sum_{i=0}^{9} w_i x^i$ to a set of points $\{(x_i, y_i)\}_{i=1}^{11}$ generated by a noisy underlying process. The independent variable is x and the dependent variable is y. This fit was obtained by solving the normal equations $\mathbf{\Phi}^{\mathsf{T}}\mathbf{y} = \mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}\mathbf{w}$ for the model weights $\mathbf{w} = (w_0, \ldots, w_9)^{\mathsf{T}}$, where $\mathbf{\Phi}$ is the basis matrix and $\mathbf{y} = (y_1, \ldots, y_{11})^{\mathsf{T}}$.

Comment on this fit with reference to the equation given in Question 1(b).

Suggest three ways by which the result could be improved, explaining your reasoning.



[7 marks]

Question 2

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

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

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

Explain the implications of this lemma and their relevance to machine learning.

[6 marks]

(b) An *unlabelled* dataset contains 500 samples, each of which is from a 1000-dimensional space. It is known that there are three (3) classes of data in this dataset and each sample is drawn from one of those classes. The classes are known to be fully separable by three hyperplanes.

Explain what method you would use to separate the dataset into its three classes, what difficulties may be encountered, and how you would overcome them.

[7 marks]

(c) 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_1	1.0	2.0	3.0	4.0	4.0	1.0	2.0
x_2	2.0	1.0	1.0	5.0	6.0	5.0	6.0

[7 marks]

Question 3

(a) A statistical process has three outcomes which occur with probabilities p_0, p_1, and p_2 respectively. Given that $\text{logit}(p_0) = \ln \frac{1}{5}$ and $\text{logit}(p_1) = 0$; compute p_0, p_1, and p_2.

[4 marks]

(b) You are working on a classification problem and have implementations of LDA (Linear Discriminant Analysis), Logistic Regression and k-nearest neighbours available. Describe how you would select which is the most suitable approach for your problem.

[7 marks]

(c) Compare and contrast the AdaBoost and Random Forest algorithms, explaining the key similarities and differences between them.

[9 marks]

End of Paper

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.