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 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) Describe two ways in which the tendency of decision trees to overfit their training data can be overcome **[5 marks]**

(b) Provide the mathematical definition of *information entropy* and sort the following binary strings from lowest to highest entropy. **[5 marks]**

(i) 1101101110

(ii) 1101010010

(iii) 1001000101

(c) The following table describes a binary classification dataset $\mathcal{D} = \{x_i, y_i, z_i, t_i\}_{i=0}^{7}$ with independent variables x, y, and z; and dependent variable t.

i	x_i	y_i	y_i	t_i
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0
3	0	1	1	1
4	1	0	0	0
5	1	0	1	1
6	1	1	0	1
7	1	1	1	1

The data generating process returns $t = 1$ when two or more of the independent variables are 1.

Using the principle of maximum information gain to determine the order of the variable splits, construct the full decision tree for this dataset, using data points $i = \{0, 1, 2, 4, 6, 7\}$ as the training set. Test your tree on data points $i = \{3, 5\}$ and comment on your result. **[10 marks]**

You may find the following table of logarithms helpful.

x	1/4	1/3	1/2	2/3	3/4	1
$\ln(x)$	−1.386	−1.099	−0.693	−0.405	−0.288	0

Question 2

(a) L_2 regularisation is sometimes used to control the solution to regression problems. It is often referred to as a *shrinkage* method.

(i) What is meant by the term "shrinkage method"?

(ii) Explain why L_2 regularisation is classified as a shrinkage method, with reference to its probabilistic interpretation.

(iii) Give another example of a shrinkage method and explain how it influences the solutions of regression problem.

[5 marks]

(b) The expectation value of the least squares loss function can be written as

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

Explain the meaning of the symbols σ, f, and h; and of the terms i, ii, and iii. **[6 marks]**

(c) Classification problems can be solved using a regression-type approach with the *logistic regression* algorithm. Explain the principles of binary logistic regression, how it can be extended to multi-class problems, and what advantages and disadvantages it has over other methods. **[9 marks]**

Question 3

(a) Sketch an example of a 2-dimensional dataset containing three classes of data point (unlabelled) that could not be separated by k-means clustering using Euclidean distance, indicating on your sketch how k-means would incorrectly partition the data. **[5 marks]**

(b) A researcher in the School of Chemistry has asked for your help. They have been collecting samples of water from different places around the world and have been trying to measure what is in the samples to understand the effects of environmental pollution. They have analysed all of the samples using a technique called mass spectrometry, which measures the number of molecules of a particular mass, for a range of different masses. For each sample, this produces a histogram that shows how many molecules of each mass were in the sample. This histogram can be represented as a vector, where each component corresponds to a mass value, and the value of the component is the number of molecules of that mass. The instrument used to obtain this data can measure the number of molecules at each of 1.2 million different mass values.

The researcher has samples from 10,000 different locations and wishes to separate them into groups of similar water composition. Suggest how you would do this, highlighting any potential problems you might encounter, and how you would solve them. **[7 marks]**

(c) The researcher uses another technique to identify 12 distinct types from 500 randomly chosen samples. Suggest how you might use this information to improve your results from part (b), again highlighting any potential problems you might encounter, and how you would solve them. **[8 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.