IMPERIAL COLLEGE LONDON

MEng and MSc EXAMINATIONS 2022

Part IV and Advanced Mechanical Engineering

for Internal Students of the Imperial College of Science, Technology and Medicine
This paper is also taken for the relevant examination for the Associateship or Diploma

MACHINE LEARNING (SAMPLE PAPER)

Duration: 2 hours

This paper contains SEVEN questions. Attempt every question.

The numbers shown by each question are for your guidance; they indicate approximately how the examiners intend to distribute the marks for this paper.
A Data and Formulæ Book is provided.

This is an OPEN BOOK Examination.

This time-limited remote assessment has been designed to be open book. You may use resources which have been identified by the examiner to complete the assessment and are included in the instructions for the examination. You must not use any additional resources when completing this assessment.

The use of the work of another student, past or present, constitutes plagiarism. Giving your work to another student to use constitutes an offence. Collusion is a form of plagiarism and will be treated in a similar manner. This is an individual assessment and thus should be completed solely by you. The College will investigate all instances where an examination or assessment offence is reported or suspected, using plagiarism software, vivas and other tools, and apply appropriate penalties to students. In all examinations we will analyse exam performance against previous performance and against data from previous years and use an evidence-based approach to maintain a fair and robust examination. As with all exams, the best strategy is to read the question carefully and answer as fully as possible, taking account of the time and number of marks available.

Turn over

Instructions

A Jupyter notebook has been provided which contains some initial code, including providing access to the necessary datasets. You should use this as a starting point for your work. Upload it to Colab via 'File -> Upload Notebook'. Through this exam you should only make use of the libraries numpy, matplotlib and pandas, i.e. your code must not use other libraries such as keras and scikit-learn.

You should write out any necessary working in the Jupyter notebook (either as code or text) and should add comments as necessary as part of this working. Note that this exam is not to assess programming but rather the understanding of the material on the course, so comments will be assessed in line with working given in a standard exam, rather than for programming correctness. Also note that the quality of the code (including how optimised it is) will not be assessed, but the focus will be on performing the calculations necessary to answer the questions correctly.

At the end of the exam, submit your Jupyter notebook (.ipynb file), which you can download from Colab via 'File -> Download -> Download .ipynb'. Do not submit any other files.

1. My company has been making carbon fibre components and I have found that 1 in 10 are breaking. I am aware that there are variations in our production process, and I wish to use information from this to predict failure. In particular I wish to use the temperature at which the carbon fibre has been cured. In dataset [s1.csv] I provide example data which contains as x the curing temperature in °C, and as y whether the component did not break (0) or did break (1). Note that this data has been selected to give an equal number of points within each class but is otherwise an unbiased representation.

   Assuming that both of the probability distributions for the breaking and non-breaking cases can be well captured by a normal distribution, produce a plot of probability that the component will break as x varies between 150 and 280, and give the specific value where the curing temperature was 240 °C. [21%]

2. Define a grid of 200 points in 2 dimensions, from -1 to 1 in each direction. A function has been provided to help with this. Produce three images for this grid, showing the distance of each point from the origin:
   (a) under the L_1 norm [2%]
   (b) under the L_∞ norm [3%]
   (c) under the $L_{1.5}$ norm. [4%]
   In each case ensure that you include a colour scale with your plot.

3. (a) I have some data measured in time in [s2.csv]. I want to detect a step change in the signal. Use a first order differentiation scheme to identify where the most likely position of this change is. [6%]
   (b) I now want to apply a threshold to the output of the differentiation to automatically identify any steps.
   (i) What happens if the threshold is too low? [2%]
   (ii) What happens if the threshold is too high? [2%]

4. (a) Why, in general, should neural network activation functions incorporate nonlinearity? [4%]

(b) Plot the ReLU activation function for inputs from -2 to +2. [4%]

(c) I have the neural network illustrated in Figure Q4. All activation functions are sigmoid functions $S(x)=\frac{1}{(1+e^{-x})}$. Produce a plot of z_1 as the input, x_1, varies from -2 to 2. [9%]

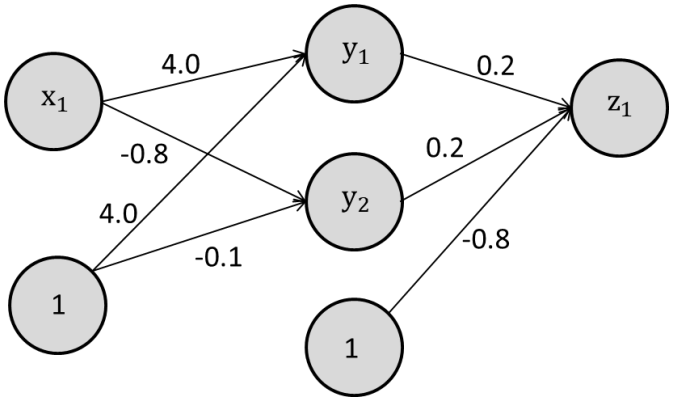

Figure Q4

5. I have a set of work tools, and data recorded about each one. This includes the following parameters: mass (kg), length (cm), type (spanner, hammer etc.) and age (years). I have had these valued, and I wish to estimate for a general tool whether the value will be more or less than £30.

(a) Explain why a random forest/decision tree approach may be most suitable for this task. [5%]

(b) I now only consider a subset, consisting of just my collection of screwdrivers. Would a random forest still necessarily be the most appropriate? Explain your answer. [4%]

6. (a) I have two principal components of a system, which are $\mathbf{p}_1$ = (2/3, 1/3, 2/3) and $\mathbf{p}_2$ = (1/3, 2/3, -2/3). Calculate all other principal components. [5%]
(b) Express the point (3, 2, 1) as a linear combination of these principal components, showing your working. Please note that for this calculation you must exploit the properties of the principal components to avoid any unnecessary working. [6%]

7. (a) When defining Parzen windows to estimate the probability density function, there is a trade-off between making the window small and big.

(i) What is beneficial about making the window small? [3%]
(ii) What is beneficial about making the window big? [3%]

(b) I have a dataset consisting of points at -0.6 and 1.1. Using a rectangular 'top hat' Parzen window of width 2, plot a graph from -2.5 to 2.5 of what the resulting probability density function (PDF) estimate will look like. [8%]

(c) Consider a Parzen window defined as:

$$\phi(\mathbf{u})=\begin{cases} 1 & |u_j|<\frac{1}{2} \quad j=1,\ldots,d \\ 0 & \text{otherwise} \end{cases}$$

for d dimensions, with $\mathbf{u}=(\mathbf{x}-\mathbf{x}_i)/h$ for each data point $\mathbf{x}_i$ with h=0.5. A dataset is given in [s3.csv] for two parameters, x_1 and x_2.

Estimate the probability density centred at point (1, 0.5) based on the defined Parzen window and the points given. [9%]