## 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

Wednesday, 19th January 2022: 14.00 to 16.00

*This paper contains FOUR 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. Your boss has given you a dataset. Each point in the dataset should consist of two continuous input parameters x1 and x2, and a class to which the point is assigned, y. Your boss says that she wants you to use the data to make classification predictions for similar data in the future.

   Unfortunately, part of the data you needed has been wiped. You have just been left with the two continuous values x1 and x2, as given in [d1.csv]. In a bid to impress your boss, you have a look at the data and think you might be able to identify the classes anyway.

   You estimate that the means lie around (-1, -1) and (2, 1) for classes 1 and 2 respectively. Write a k-means routine with an $L_2$ norm which performs 5 iterations to better estimate these means and hence separate the data into their two classes. What are the two final mean locations? [25%]

2. The full dataset described in Q1 (x1, x2 and y) has now been recovered and this is given in [d2.csv]. The class value y is defined as 1 or 2, and the values are sampled without bias with respect to any of the input or output parameters. The data associated with each class has zero covariance between the two parameters.

   (a) Using a maximum likelihood approach, estimate the means and covariance matrices for each of the two classes. Note that you may assume without proof that the zero covariance between the two parameters enables the mean and standard deviation in each parameter to be estimated independently. [15%]

   (b) Define a grid with 200 points from -2.5 to 2.5 in each parameter. For each grid point calculate the probability, based on the data provided, that this point will belong to class 2 over class 1, then plot the complete grid as an image. Functions are provided to help generate the sample grid and also calculate a 2D probability density function. You may also wish to add a scatter plot of the training data to this. [13%]

   (c) Without performing any calculations, based on the information provided above, what are the directions of the principal components for the training data associated with each class, and why? [6%]

   (d) Your boss then says she has an important situation she wants to make a prediction for, where x1 = -2 and x2 = 2. Do you think your prediction will be reliable, and why? [6%]

3. Your boss now comes in and she says that the problem has completely changed, and you'll need to start again. This time you have one continuous input parameter, but the output is now also continuous, rather than being a discrete classifier. The data is in [d3.csv]. You want to fit the curve y = Ax + B cos (πx) + C. Use a suitable linear regression approach minimising the $L_2$ error to determine the constants A, B and C. Produce a plot of the curve you have produced across the range x = -3 to 3 on top of a scatter plot of the data you were provided, with the points and the curve in different colours. [15%]

4. Now, your colleague is trying to solve a two parameter two-class classification problem with a hard-boundary support vector machine. The first class has support vectors at (0, 0) and (0.4, 1). The second class has a single vector at (1, 0).

   (a) What is the equation of the line separating the two datasets, in the form y = mx + c, where m and c are constants (to be written as decimals) to be determined? Note that this problem can be solved by hand but you must write it into your notebook as either commented Python code or in a separate text field. [15%]

   (b) Ignoring the presence of any other datapoints which could become support vectors, how can we move the support vector (0, 0) in the positive x direction such that the support vector machine will be described by just two support vectors? [5%]