IMPERIAL COLLEGE LONDON

MEng EXAMINATIONS 2020

Part IV

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

Friday, 1st May: 14.00 to 15.30

This paper contains TEN questions. Attempt all questions.
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.

While this time-limited remote assessment has not been designed to be open book, in the present circumstances it is being run as an open-book examination. We have worked hard to create exams that assesses synthesis of knowledge rather than factual recall. Thus, access to the internet, notes or other sources of factual information in the time provided will not be helpful and may well limit your time to successfully synthesise the answers required.
Where individual questions rely more on factual recall and may therefore be less discriminatory in an open book context, we may compare the performance on these questions to similar style questions in previous years and we may scale or ignore the marks associated with such questions or parts of the questions. 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

1. (a) Given a metric D(., .), write out mathematically in terms of vectors **a**, **b** and **c** and D the triangle inequality which must hold. [2%]

   (b) I wish to perform a nearest neighbour classification. In my training set I have three points, (2, 5), (5, 3) and (3, 1), which are classified 1, 2 and 3 respectively. Calculate the L_∞ distance metric between point **x** = (3.0, 3.1) and each point in the training set, and hence classify point **x**. [4%]

   (c) Now using the L_1 metric, what are the distances from the training points to, and resulting classification of point (2.1, 2.7)? [4%]

2. I am making headlights for a car. I can buy bulbs from two companies. I must pay the purchasing cost, which from company A is £2/bulb, and from B is £3/bulb. Both are equally easy to install initially (i.e. the installation can be considered effectively free), but the bulb from A is much more difficult to replace, costing me £30 per replacement of a bulb compared to replacing the one from B which just costs me £15.

   (a) Taking the actions α_1, α_2 as purchasing from A and B respectively, and the two states of nature being the bulb not failing (ω_1) or failing once (ω_2), respectively, define the four loss terms $\lambda(\alpha_1|\omega_1)$, $\lambda(\alpha_1|\omega_2)$, $\lambda(\alpha_2|\omega_1)$ and $\lambda(\alpha_2|\omega_2)$ based on the information above. [4%]

   (b) Assuming a probability of failure from company A as 7% and company B as 5%, calculate the risk associated with each action. Based on this, which company would you advise purchasing from? [8%]

3. I have prior probabilities of 0.3 for ω_1, 0.4 for ω_2 and 0.3 for ω_3. Based on a specific observation x, I have likelihoods of 0.2, 0.2 and 0.4 for the three cases respectively. Calculate the posterior probabilities for each state of nature and indicate which one, given no other information, you would select and why. [6%]

4. (a) Given a discriminant hyperplane $\mathbf{w}^t\mathbf{x} + w_0 = 0$, show that the vector **w** is normal to the hyperplane. [4%]

   (b) For a discriminant function $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ with weightings $w_0 = 10$ and $\mathbf{w} = (2,-7,1)^T$, assuming $g(\mathbf{x}) < 0$ corresponds to class 1 and $g(\mathbf{x}) > 0$ to class 2, how would a point at **x** = (-3, 2, 9) be classified? [2%]

   (c) What is the distance of point $\mathbf{x} = (3, 2, 1)^T$ from the hyperplane? [5%]

5. (a) Given a dataset of points (1, -2), (2, -4) and (3, -6):

(i) What is the first principal component? [2%]
(ii) What is the second principal component? [3%]

(b) A dataset has principal components in directions $p_1=(0.667,0.667,0.333)^T$, $p_2=(0.596,-0.298,0.745)^T$ and $p_3=(0.333,0.667,-0.667)^T$. Express the point $(-2,3,1)^T$ as a linear combination of these three principal components. [5%]

6. (a) Why, in general, should neural network activation functions incorporate nonlinearity? [2%]

(b) Sketch the ReLU activation function. Make it clear on your graph what output value you obtain for an input of 1. [2%]

(c) I have the neural network illustrated in Figure Q6. Defining the cost function as the L_2 error, what is the gradient of this cost function relative to the weighting underlined, when, as shown, the true output (i.e. value to be fitted to) is 0.8 for an input of 0.2? Take all activation functions as the sigmoid function, $S(x)=1/(1+e^{-x})$. [8%]

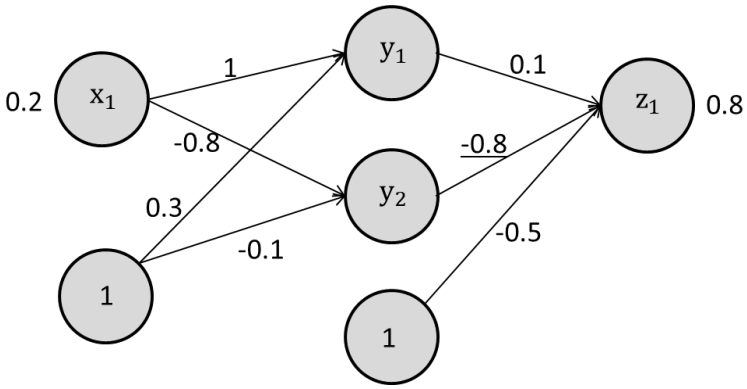

Figure Q6

7. I wish to use the k-means clustering algorithm, with 2 means, operating on the following data:

(-1.5, 0.9), (0.3, 0.7), (0.5, -0.9), (-0.1, -1.2), (-0.2, 0.6), (1.3, -1.1), (0.2, 1.4), (-0.8, 0.7), (1.2, -1.6), (0.6, -0.3)

Take a starting point where mean 1 is at position (-1, 1) and mean 2 is at position (1, -1). Calculate where the revised mean centres will be after one iteration through the k-means algorithm. [8%]

8. Consider a cost function you wish to minimise in 2D space expressed by the variables x_1 and x_2 as

$$C = \exp\left(x_1^2 + 0.9x_1x_2 + 0.7x_2^2\right)$$

where $\exp(x) = e^x$ corresponds to the exponential function. Start from the point $(x_1,x_2) = (0.8, 0.8)$ and calculate the first two steps of the alternating descent function to minimise C, initially moving along x_1. [6%]

9. A hard-margin SVM has basis functions x_1, x_2, and x_1x_2. The support vectors are identified in (x_1, x_2) space as (0, 0.8) and (-1, -1) and are respectively classified as 0 and 1. How is point (-1, 0) classified? Note that you must show suitable working to receive marks. [9%]

10. (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? [2%]
(ii) What is beneficial about making the window big? [2%]

(b) I have a dataset consisting of points at -0.6 and 1.1. Using a rectangular 'top hat' Parzen window of width 2, sketch a graph of what the resulting probability density function (PDF) estimate will look like. [6%]

(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. There are 200 points in a data set, and all the points in this dataset lying within range $0 < x_1 < 3$ and $0 < x_2 < 3$ are given as:

(0.1, 0.8), (1.1, 2.8), (0.2, 1.5), (0.8, 2.1), (1.2, 2.2), (2.7, 2.1), (1.9, 0.1), (1.1, 2.0), (2.3, 1.8), (1.4, 0.9), (0.2, 2.9), (0.7, 1.3), (0.3, 2.8), (1.6, 1.9), (1.1, 2.1), (0.5, 2.0), (1.0, 0.9), (2.3, 0.2), (2.2, 1.5), (0.8, 0.2), (2.8, 1.3), (0.8, 2.2), (1.9, 0.9), (1.4, 2.0), (0.3, 0.1), (2.9, 2.8), (0.8, 1.8), (0.6, 2.4)

Estimate the probability density centred at point (1, 2) based on the defined Parzen window and the points given. [6%]