IMPERIAL COLLEGE LONDON

BEng, MEng and MSc EXAMINATIONS 2020

Part III, 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

90 minute paper.

This paper contains THIRTEEN 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.

This is a CLOSED BOOK Examination.

Turn over

1. A colleague comes to you and tells you that they have identified all the principal components of their dataset as (0.5, 0.5, 0, 0.4) (0.2, 0.1, -0.2, 0.2) and (-0.3, 0, -0.5, -0.1). Give all the reasons why these cannot be correct, including mathematical calculations where necessary to justify these. [6%]

2. Given a set of data sampled the normal distribution $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$, show that the maximum likelihood estimate of mean and standard deviation from this data matches the mean and standard deviation of the sampled data. [14%]

3. Variable a has standard deviation 2 and mean 0, and variable b has standard deviation 1 and mean 1. The covariance between them is -0.7.
(a) Write out the covariance matrix. [2%]
(b) What is the first principal component of the data? [8%]

4. A parzen window is defined as

$$\phi(x) = \begin{cases} \frac{3}{4}(1-x^2) \; if -1 < x < 1 \\ 0 \; otherwise \end{cases}$$

(a) Show that this is a valid parzen window. [4%]
(b) I have a dataset consisting of a single point at $x = -1$. Write out an expression for the resulting probability distribution estimated by the parzen window defined. [2%]
(c) I sample data for a two class problem. I estimate data within the first class as having a probability distribution of $1 - \cos \pi x$ in the range $0 < x < 1$ and zero outside, and data within the second class as being uniform in the same range and zero outside. The probability of any data being in class 1 is 0.6. If I have a sampled point at a position x in the range $0 < x < 1$, what is the probability that it comes from the first class? [6%]

5. I have a single parameter classification problem. Training values for the input parameter x and the output classification y are given as

x	y
5.3	1
2.1	0
3.9	0
1.8	1
1.2	1
0.6	0
0.1	1
3.2	1
4.3	0

(a) Using the k nearest neighbours algorithm with 3 points, classify a point at x=2, using the L_1 metric. Note that you should show full working to receive marks. [2%]

(b) In general, how will increasing the number of nearest neighbours affect the classification behaviour? [2%]

(c) For a single parameter problem like this, what effect would using a L_2 metric instead have on the result? [1%]

6. A linear support vector classifier has points (2, 1), (0, -1) as the support vectors classified as 0 and 1 respectively.

(a) Expressing the resulting decision function in the form $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ with $g(\mathbf{x}) > 0$ being class 1 and $g(\mathbf{x}) < 0$ corresponding to class 0, what are $\mathbf{w}$ and w_0? [6%]

(b) Now assume it is a soft-margin classifier. Taking a training point at (0.5, 1) classified as 1, what is the value of ϵ_i for this point? If it was now classified as 0, what would the value of ϵ_i be? [6%]

7. 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. [3%]

(b) I now only consider a subset, consisting of just my collection of screwdrivers. Would a random forest still necessarily be the most appropriate? [2%]

8. I have intermittent failures in a new design of engine and have recorded the age, temperature and RPM at failure, as well as mean temperature and RPM through the life. In each of the following cases indicate whether unsupervised or supervised learning is most suitable:

(a) I wish to identify whether there are any patterns which could indicate the cause of the failure. [2%]

(b) I wish to estimate the age at failure based on the other parameters. [2%]

(c) Taking a new engine and expected usage measures (temperature and RPM) I wish to predict whether or not it will fail within the warrantee period. [2%]

9. I define a decision tree based on some training data. At one of the nodes, I have 10 points in my dataset which reach that point. There is 1 from category 1, 6 from category 2 and 3 from category 3. What is the Gini impurity at that node? [3%]

10. (a) I have some data measured in time with a sampling rate of 1 kHz listed as 0.106, 0.066, 0.228, 0.609, 0.548, 0.667, 0.657, 0.772. 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. [3%]
(b) I 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%]

11. I have the following points to which I wish to fit a second-order polynomial.

x_1	x_2	y
5.3	8.3	6.2
2.1	2.8	5.6
3.9	5.5	7.2
1.8	2.3	3.4

(a) We wish to do this fitting by adding additional parameters then fitting a linear function to all of these parameters. Write out algebraically what these additional parameters are. [3%]

(b) Now assume we wish to just fit a linear function to the data, i.e. we just have the parameters x_1 and x_2. Take the linear function as $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2$. Define a cost function based on the L_2 norm and calculate what the gradient is, taking a starting point as $\beta_1 = 2$, $\beta_2 = -1$ and $\beta_3 = 0$. For a step length of 0.001, what are the values for β_1, β_2 and β_3 after the first step? [7%]

12. (a) I have a support vector classifier. The bias is too high. Should I reduce or increase the softness of the margin? [2%]

(b) I now have a neural network. I find variance to be too high. If the only thing I can control is the number of layers, should I add or reduce these to address this? [2%]

13. I have the neural network in Figure Q13. Activation functions at each node are unscaled linear functions, i.e. $f(x) = x$. Show that this network is equivalent to a single linear discriminant function of form $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$, and find what $\mathbf{w}$ and w_0 are. [6%]

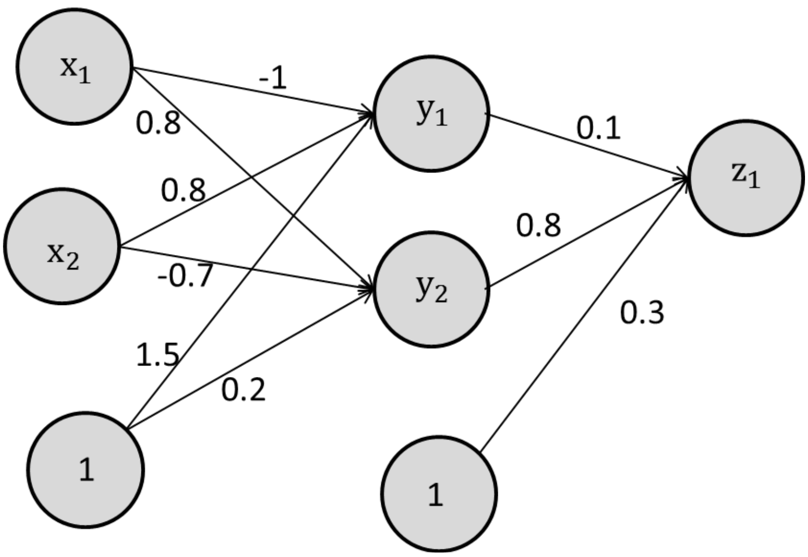

Figure Q13

Turn over