King's College London

This paper is part of an examination of the College counting towards the award of a degree. Examinations are governed by the College Regulations under the authority of the Academic Board.

Degree Programmes	MSc, MSci
Module Code	7CCSMPNN
Module Title	Pattern Recognition
Examination Period	May 2018 (Period 2)
Time Allowed	Three hours
Rubric	ANSWER FOUR OF SIX QUESTIONS.

All questions carry equal marks. If more than four questions are answered, the answers to the first four questions in exam paper order will count.

ANSWER EACH QUESTION ON A NEW PAGE OF YOUR ANSWER BOOK AND WRITE ITS NUMBER IN THE SPACE PROVIDED.

Calculators	Calculators may be used. The following models are permitted: Casio fx83 / Casio fx85.
Notes	Books, notes or other written material may not be brought into this examination

PLEASE DO NOT REMOVE THIS PAPER FROM THE EXAMINATION ROOM

TURN OVER WHEN INSTRUCTED

1.

a. Explain how Bayes decision rule for minimum error is used to determine the class of an exemplar.

[4 marks]

Table 1, below, shows samples that have been recorded from three univariate class-conditional probability distributions:

class	samples of x
ω_1	3.54 4.83 0.74 3.86 3.32 1.69 2.57 3.34 6.58
ω_2	15.85 4.75 17.17 5.63 1.68 5.57 0.98
ω_3	3.75 4.00 4.53 5.34 1.59

Table 1

b. Apply k_n nearest neighbours to the data shown in Table 1 in order to estimate the density of $p(x|\omega_1)$ at location $x = 4$, using $k = 3$.

[4 marks]

c. The class-conditional probability distribution for class 2 in Table 1, $p(x|\omega_2)$, is believed to be a Normal distribution. Calculate the parameters of this Normal distribution using Maximum-Likelihood Estimation; use the "unbiased" estimate of the covariance.

[4 marks]

QUESTION 1 CONTINUES ON NEXT PAGE

SEE NEXT PAGE

d. For the data in Table 1, use Parzen-window density estimation to calculate the density of $p(x|\omega_3)$ at the location $x = 4$. Use a Gaussian window with width $h_n = 2$. For a Gaussian Parzen window the estimate of the probability density, p_n, at a value, x, is given by the formula:

$$p_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

where

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-0.5u^2\right)$$

[4 marks]

e. Using the number of samples given in Table 1 estimate the prior probabilities of each class, *i.e.*, calculate P(ω_1), P(ω_2) and P(ω_3).

[3 marks]

f. Use the answers to sub-questions 1.b, 1.c, 1.d and 1.e to determine the class of a new sample with value $x = 4$.

[6 marks]

2.

a. Give a brief definition of each of the following terms in the context of classification:

i. Supervised Learning

ii. Generalisation

[4 marks]

b. A dichotomizer is defined using the following linear discriminant function: $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+w_0$ where $\mathbf{w}^T = (-1, -2)$ and $w_0 = 3$. Using this linear discriminant function determine the class predicted for the feature vector $\mathbf{x}^T = (2.5, 1.5)$.

[4 marks]

c. Sketch the decision boundary defined by the linear discriminant function specified in sub-question 2.b.

[4 marks]

d. Table 2, shows a linearly separable dataset.

$\mathbf{x}^T$	Class
(2,-1)	0
(-1,0)	1
(0,0)	1
(1,1)	0
(0,-1)	1

Table 2

i. Re-write the dataset shown in Table 2 using augmented vector notation.

[2 marks]

ii. A linear threshold neuron has a transfer function which is a linear weighted sum of its inputs and an activation function that applies a threshold (θ) and the Heaviside function. For a linear threshold neuron with parameter values $\theta = -1$, $w_1 = 3$ and $w_2 = 0.5$ calculate the output of the neuron in response to each of the samples in the dataset given in Table 2.

[4 marks]

iii. One method for learning the parameters of a linear threshold neuron, is the *sequential* delta learning algorithm. Give the learning rule for this algorithm, and apply the learning algorithm to find the parameters of a linear threshold neuron that will correctly classify the data in Table 2. Assume initial values of $\theta = -1$, $w_1 = 3$ and $w_2 = 0.5$, and a learning rate of 1.

[7 marks]

3.

a. Give brief definitions of the following terms:

i. Feature selection

ii. Feature extraction

[4 marks]

b. Principal component analysis (PCA) is a commonly used method for feature extraction in pattern recognition problems. Briefly describe what principal components are, and explain why performing PCA on a dataset before learning may increase classifier accuracy.

[4 marks]

c. Write pseudo-code for the Karhunen-Love Transform method for PCA.

[5 marks]

d. When performing PCA using neural networks, Oja's update rule is often used in place of standard Hebbian learning. Give Oja's update rule, and briefly explain why this rule is preferred.

[5 marks]

e. Linear Discriminant Analysis (LDA) is a supervised projection method for finding linear combinations of features in a dataset. Briefly explain Fisher's LDA method for estimating the weight parameters of a decision boundary to separate two classes in a dataset, and give the objective function.

[3 marks]

QUESTION 3 CONTINUES ON NEXT PAGE

f. Figure. 1 shows a dataset containing two classes of two-dimensional data, with one class labelled with crosses, and the other class labeled with circles. An axis on which the data could be projected is shown with a black arrow.

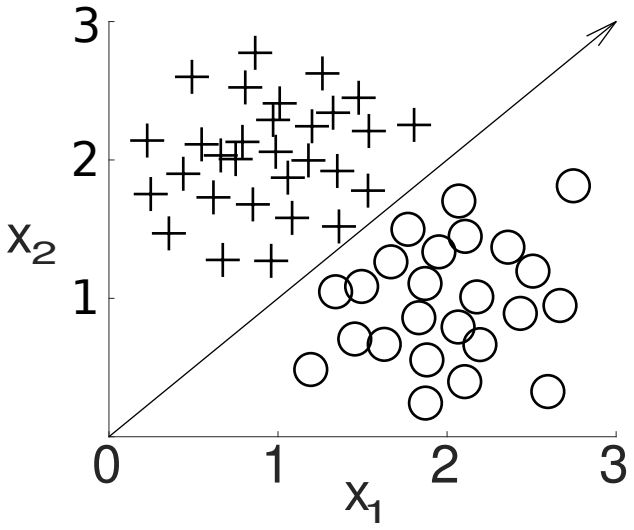

Figure 1

Is this projection a suitable axis for maximally discriminating between the two classes of data? Justify your answer.

[4 marks]

4.

a. Draw a diagram to show each of the following activation functions.

- Symmetric hard limit (signum)
- Linear function
- Symmetric sigmoid function
- Logarithmic sigmoid function
- Radial basis function

[5 marks]

b. Referring to the activation functions listed in Question 4.a, which activation function is not suitable for training a feedforward neural network using the backpropagation algorithm? Explain your answer.

[5 marks]

c. When the backpropagation algorithm is employed to train a neural network, the dataset is usually partitioned into three sub-datasets. Name these three sub-datasets.

[3 marks]

d. Figure 2 shows a diagram of a feedforward neural network with two inputs and two outputs. All activation functions indicated by "/" are a linear function. In the feedforward operation, Table 3 shows the input-output relationship between inputs (x_1 and x_2) and outputs (z_1 and z_2). For example, referring to Table 3, when $x_1 = 2$ and $x_2 = -0.5$, the neural network's outputs are $z_1 = 98$ and $z_2 = 7.5$. Determine the connection weights w_{11}, w_{12}, w_{21} and w_{22} which will produce the results in Table 3.

[12 marks]

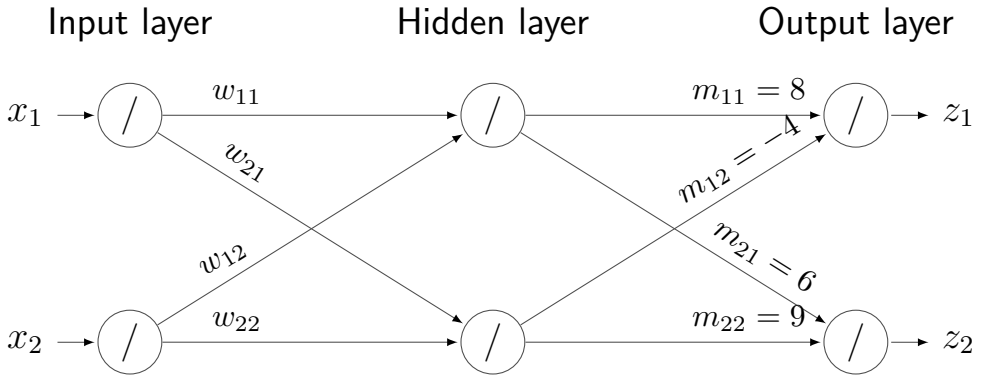

Figure 2: A diagram of 3-layer neural network.

x_1	x_2	z_1	z_2
2	-0.5	98	7.5
-4	-6	-168	-246

Table 3: Inputs x_1 and x_2, and outputs z_1 and z_2.

5.

a. Briefly explain what a "weak classifier" is.

[3 marks]

b. Explain briefly the idea of "ensemble methods" for classification.

[3 marks]

c. Name two methods used for designing ensemble classifiers.

[2 marks]

d. The Adaboost algorithm is given in Table 4.

Algorithm: Adaboost

Given $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, $y_i = \{-1, 1\}, i = 1, \ldots, n$
begin initialise $k_{\max}$, $W_1(i) = 1/n, i = 1, \ldots, n$
  $k \leftarrow 0$
  do $k \leftarrow k + 1$
    $\hat{h}_k = \arg\min_{h_j \in \mathcal{H}} E_j$ where $E_j = \sum_{i=1, y_i \neq h_j(\mathbf{x}_i)}^{n} W_k(i)$ (Weighted error rate)†
    $\epsilon_k =$ overall weighted error rate of classifier $\hat{h}_k$ (i.e., the minimum E_j)
    if $\epsilon_k > 0.5$
      $k_{\max} = k - 1$
      Stop
    end
    $\alpha_k = \frac{1}{2} \ln \left(\frac{1 - \epsilon_k}{\epsilon_k} \right)$
    $W_{k+1}(i) = \frac{W_k(i) e^{-\alpha_k y_i \hat{h}_k(\mathbf{x}_i)}}{Z_k}$, $i = 1, \ldots, n$ (Update $W_k(i)$)‡
  until $k = k_{\max}$
end

† Finds the classifier $h_k \in \mathcal{H}$ that minimises the error ϵ_k with respect to the distribution $W_k(i)$.
‡ Z_k is a normalization factor chosen so that $W_{k+1}(i)$ is a normalised distribution.

Table 4: Adaboost Algorithm.

i. Consider a classification problem with 5 samples. The dataset is denoted as $\mathcal{D} = \{(\mathbf{x}_1, -1), (\mathbf{x}_2, -1), (\mathbf{x}_3, +1), (\mathbf{x}_4, +1), (\mathbf{x}_5, +1)\}$. Three weak classifiers, $h_1(\mathbf{x})$, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$, are designed where the classification results are shown in Table 5. Referring to this table, taking $h_1(\mathbf{x})$ and $\mathbf{x}_1$ as an example, the classification given by the classifier $h_1(\mathbf{x})$ with the input sample $\mathbf{x}_1$ is $+1$. Determine α_1 by runing the first iteration of Adaboost algorithm shown in Table 4 with $k_{\max} = 3$.

[14 marks]

Classifier	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$h_1(\mathbf{x})$	$+1$	-1	$+1$	$+1$	$+1$
$h_2(\mathbf{x})$	-1	$+1$	-1	$+1$	$+1$
$h_3(\mathbf{x})$	-1	-1	$+1$	-1	-1

Table 5: Classification Results.

ii. When the Adaboost algorithm continues, it is assumed that $\hat{h}_2(\mathbf{x}) = h_3(\mathbf{x})$, $\hat{h}_3(\mathbf{x}) = h_2(\mathbf{x})$, $\alpha_2 = 0.8$ and $\alpha_3 = 0.9$ are obtained. With the results obtained in Question 5.d.i, give the formula for this ensemble classifier.

[3 marks]

6.

a. In the context of linear support vector machines (SVMs), considering the linearly separable case, the design of a linear SVM classifier is formulated as the following optimisation problem:

$$\min_{\mathbf{w}} \quad J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \quad i = 1, 2, \ldots, N$$

where $\mathbf{w}$ and w_0 are the parameters of the linear SVM classifier; $y_i \in \{-1, +1\}$ denotes the class label of the sample $\mathbf{x}_i$; N is the number of samples and $\|\cdot\|$ denotes the Euclidean norm operator.

i. Briefly explain why "$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$" has to be minimised.

[4 marks]

ii. Briefly explain why the constraint "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1$" has to be satisfied.

[4 marks]

iii. If the samples are nonlinearly separable, what can be done to the samples so that the linear SVM classifier becomes a nonlinear one?

[2 marks]

b. A training dataset is given as follows:

Class 1: $x_1 = -3$; $x_2 = -2$; $x_3 = 6$; $x_4 = 8$.
Class 2: $x_5 = -1$; $x_6 = 0$; $x_7 = 1$; $x_8 = 2$.

i. Is the dataset linearly separable? Justify your answer.

[4 marks]

ii. A feature mapping function $\mathbf{z}_i = \Phi(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}$, $i = 1, 2, 3, 4, 5, 6, 7, 8$, is considered. It is found that the samples $\mathbf{z}_1$ to $\mathbf{z}_8$ in the feature space after mapping are linearly separable. Using $\mathbf{z}_2$, $\mathbf{z}_5$ and $\mathbf{z}_8$ as support vectors, design an SVM classifier to correctly classify all the given samples x_1 to x_8.

[11 marks]