—— SOLUTIONS ——

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 2017

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.
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

Answer each question on a new page of your answer book and write its number in the space provided.

PLEASE DO NOT REMOVE THIS PAPER FROM THE EXAMINATION ROOM

DO NOT TURN OVER UNTIL INSTRUCTED TO DO SO

1. a. Give a brief definition of each of the following terms:

i. Minimum Error Rate classifier

ii. Minimum Loss classifier

[4 marks]

Answer

i) Minimum Error Rate classifier: a classifier that minimises the number of misclassified exemplars
ii) Minimum Loss classifier: a classifier that minimises the costs associated with misclassifying exemplars

Marking scheme

2 marks for each correct definition.
Tests knowledge of basic terminology used in pattern recognition.

b. Write down the equation for the Likelihood Ratio Test for minimum error, and explain how this equation is used to determine the class of an exemplar.

[4 marks]

Answer

Likelihood Ratio Test:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

Exemplar, x, is predicted to be in class ω_1 if the above expression is true.

Marking scheme

4 marks.
Tests knowledge of Bayesian decision theory.

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

Table 1

class	samples of x									
ω_1	0.82	1.42	1.53	0.21	0.72	1.73	0.49	1.03	0.69	2.12
ω_2	2.54	1.36	3.77	2.86	0.92	6.01				

c. 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 $x = 1.5$, using $k = 3$.

[4 marks]

Answer

The closest three samples to $x = 1.5$ are at: 1.53, 1.42, and 1.73. We need to make a "volume" of radius $|1.5 - 1.73| = 0.23$ (diameter 0.46) around $x = 1.5$ to encompass these three samples.

Therefore density is:

$$p(x = 1.5|\omega_1) = \frac{k/n}{V} = \frac{3/10}{0.46} = 0.652$$

Marking scheme

2 marks for correct method, 2 marks for correct application.

Tests ability to perform density estimation.

d. 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]

Answer

Maximum-Likelihood estimate of mean is the arithmetic mean of the sample values. Therefore,

$$\hat{\mu_2} = \frac{(2.54 + 1.36 + 3.77 + 2.86 + 0.92 + 6.01)}{6} = 2.91$$

Maximum-Likelihood "unbiased" estimate of the covariance for univariate data, is:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

Therefore,

$$\begin{aligned}\hat{\sigma_2}^2 = \frac{1}{5}[&(2.54 - 2.91)^2 + (1.36 - 2.91)^2 + (3.77 - 2.91)^2 \\ &+ (2.86 - 2.91)^2 + (0.92 - 2.91)^2 + (6.01 - 2.91)^2] = 3.37\end{aligned}$$

Marking scheme

2 marks for correct methods, 2 marks for correct application.
Tests ability to perform density estimation.

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

[2 marks]

Answer

$$P(\omega_1) = \frac{10}{16} = 0.625$$
$$P(\omega_2) = \frac{6}{16} = 0.375$$

Marking scheme

1 mark for each correct answer.

Tests ability to apply Bayesian decision theory.

f. Using the Likelihood Ratio Test and the results from question 1.c, question 1.d, and question 1.e determine the class in which a new sample with value $x = 1.5$ should be placed.

[3 marks]

Answer

From part 1.c, $p(x = 1.5|\omega_1) = 0.652$
From part 1.d, $p(x|\omega_2)$ is given by $N(2.91, \sqrt{3.37})$, so $p(x = 1.5|\omega_2) = \frac{1}{\sqrt{2\pi \times 3.37}} \exp\left(-\frac{(1.5-2.91)^2}{2 \times 3.37}\right) = 0.162$

Likelihoods ratio states, choose ω_1 if: $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$
Hence, when $x = 1.5$, choose ω_1 if: $\frac{0.652}{0.162} = 4.02 > \frac{P(\omega_2)}{P(\omega_1)}$
From part 1.e, $\frac{P(\omega_2)}{P(\omega_1)} = \frac{0.375}{0.625} = 0.6$
Therefore we should choose ω_1 if: $4.02 > 0.6$
This is true, so we should choose class 1.

Marking scheme

1 mark for correctly calculating $p(x = 1.5|\omega_2)$, 2 marks for correct application of Likelihood Ratio Test.
Tests ability to apply Bayesian decision theory.

g. Using the Likelihood Ratio Test *with losses* and the results from question 1.c and question 1.d determine in what class a new sample with value $x = 1.5$ should be placed, when $P(\omega_1) = 0.4P(\omega_2)$ and the loss, λ_{ij}, associated with choosing class i when the actual class is j, is as follows: $\lambda_{11} = 0$, $\lambda_{12} = 2$, $\lambda_{22} = 0$, $\lambda_{21} = 1$.

[4 marks]

Answer

Likelihoods ratio with losses states, choose ω_1 if: $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \left(\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\right)\frac{P(\omega_2)}{P(\omega_1)}$
Therefore, we should choose ω_1 if: $4.02 > \left(\frac{2}{1}\right) \times 2.5$
This is false, so we should choose class 2.

Marking scheme

2 marks for knowing the equation, 2 marks for correct application.
Tests ability to apply Bayesian decision theory.

2. a. In augmented feature space, a dichotomizer is defined using the following linear discriminant function $g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$ where $\mathbf{a}^T = (-2, 1, 2, 1, 2)$ and $\mathbf{y} = \left(1, \mathbf{x}^T\right)^T$. Determine the class of the following feature vectors, $\mathbf{x}_1 = (1, 1, 1, 1)^T$ and $\mathbf{x}_2 = (0, -1, 0, 1)^T$.

[4 marks]

<u>Answer</u>

To classify a point $\mathbf{x}$ we calculate $g(\mathbf{x})$, $\mathbf{x}$ is in class 1 if $g(\mathbf{x}) > 0$.

$$(-2\ 1\ 2\ 1\ 2)\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 4$$

Therefore class 1.

$$(-2\ 1\ 2\ 1\ 2)\begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix} = -2$$

Therefore class 2.

<u>Marking scheme</u>

4 marks for correct method, 1 mark each for correct application.

Tests ability to apply linear discriminant functions.

b. One method of learning the parameters of a linear discriminant function is by using the perceptron learning algorithm. Write pseudo-code for the *sequential* perceptron learning algorithm.

[4 marks]

Answer

Augment the feature vectors
set $\omega_k = 1$ for each sample in class 1, and $\omega_k = -1$ for each sample in class 2.
Initialise $\mathbf{a}$
For each sample, $\mathbf{y}_k$, in the dataset:
    Calculate $g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$
    If $\mathbf{y}_k$ is misclassified (*i.e.*, if $sign(g(\mathbf{x})) \neq \omega_k$
    Update solution such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\omega_k\mathbf{y}_k$
Repeat until $\mathbf{a}$ stops changing.

OR

Augment and apply sample normalisation to the feature vectors.
Initialise $\mathbf{a}$
For each sample, $\mathbf{y}_k$, in the dataset:
    Calculate $g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$
    If $\mathbf{y}_k$ is misclassified (*i.e.*, if $g(\mathbf{x}) < 0$)
    Update solution such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\mathbf{y}_k$
Repeat until $\mathbf{a}$ stops changing.

Marking scheme

4 marks
Tests ability to describe a learning algorithm for defining linear discriminant functions.

Table 2

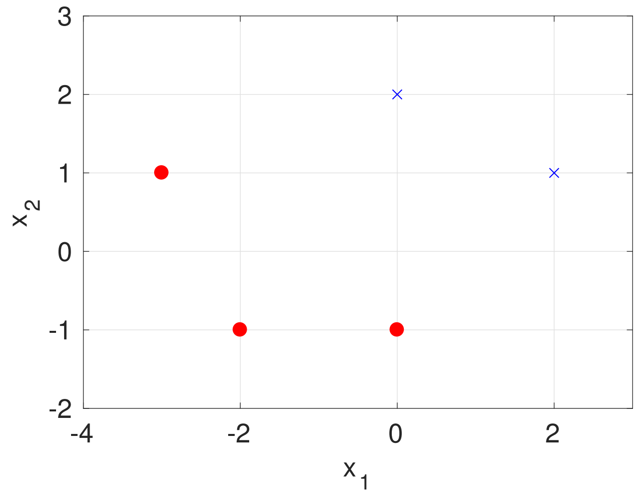
$\mathbf{x}^T$	$class$
$(5, 1)$	1
$(5, -1)$	1
$(7, 0)$	1
$(3, 0)$	-1
$(2, 1)$	-1
$(1, -1)$	-1

c. Determine if the dataset shown in Table 2 is linearly separable.

[2 marks]

Answer

By plotting the data it can be seen that the dataset is linearly separable.



Marking scheme

2 marks.

Tests knowledge of basic terminology used in pattern recognition.

d. Apply the Sequential Perceptron Learning Algorithm to determine the parameters of a linear discriminant function that will correctly classify the data shown in Table 2. Assume an initial value of $\mathbf{a} = (w_0, \mathbf{w}^T)^T = (-25, 5, 2)^T$, and use a learning rate of 1.

[6 marks]

Answer

For the Sequential Perceptron Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\omega\mathbf{y}_k$, where $\mathbf{y}_k$ is a misclassified exemplar, and ω is the class label associated with $\mathbf{y}_k$. Here, $\eta = 1$.

iteration	$\mathbf{a}_{old}^T$	$\mathbf{y}^T$	$g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$	ω	$\mathbf{a}_{new}^T = \mathbf{a}_{old}^T + \omega\mathbf{y}^T$ if misclassified
0	$[-25, 5, 2]$	$[1, 5, 1]$	2	1	$[-25, 5, 2]$
1	$[-25, 5, 2]$	$[1, 5, -1]$	-2	1	$[-24, 10, 1]$
2	$[-24, 10, 1]$	$[1, 7, 0]$	46	1	$[-24, 10, 1]$
3	$[-24, 10, 1]$	$[1, 3, 0]$	6	-1	$[-25, 7, 1]$
4	$[-25, 7, 1]$	$[1, 2, 1]$	-10	-1	$[-25, 7, 1]$
5	$[-25, 7, 1]$	$[1, 1, -1]$	-19	-1	$[-25, 7, 1]$
6	$[-25, 7, 1]$	$[1, 5, 1]$	11	1	$[-25, 7, 1]$
7	$[-25, 7, 1]$	$[1, 5, -1]$	9	1	$[-25, 7, 1]$
8	$[-25, 7, 1]$	$[1, 7, 0]$	24	1	$[-25, 7, 1]$

Learning has converged, so required parameters are $\mathbf{a} = (-25, 7, 1)^T$.

Marking scheme

6 marks.
Tests ability to apply a learning algorithm for defining linear discriminant functions.

e. A support vector machine (SVM) is to be used to classify the data shown in Table 2.

i. Identify the support vectors by inspection.

[3 marks]

Answer

By inspection, the two classes can be separated by constructing a hyperplane in the region of $1 \leq x_1 \leq 3$. The support vectors are: $\mathbf{x}_1 = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$, $\mathbf{x}_4 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$.

Marking scheme

3 marks, 1 mark for each.

Tests understanding of Support Vector Machines.

ii. Design a SVM classifier to classify all given points.

[6 marks]

Answer

Define the label for Class 1 as $+1$ and Class 2 as -1 so we have $y_1 = y_2 = y_3 = 1$ and $y_4 = y_5 = y_6 = -1$.

The hyperplane is $\mathbf{w}^T\mathbf{x} + w_0 = 0$ where $\mathbf{w} = \lambda_1 y_1 \mathbf{x}_1 + \lambda_2 y_2 \mathbf{x}_2 + \lambda_4 y_4 \mathbf{x}_4 = \lambda_1 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 5 \\ -1 \end{bmatrix} - \lambda_4 \begin{bmatrix} 3 \\ 0 \end{bmatrix}$. 1 mark

$y_i(\mathbf{w}^T\mathbf{x} + w_0) = 1$ when $\mathbf{x}$ is a support vector. Hence:

For $\mathbf{x} = \mathbf{x}_1$, $\left(\lambda_1 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 5 \\ -1 \end{bmatrix} - \lambda_4 \begin{bmatrix} 3 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 5 \\ 1 \end{bmatrix} + w_0 = 1$

$\Rightarrow 26\lambda_1 + 24\lambda_2 - 15\lambda_4 + w_0 = 1.$

For $\mathbf{x} = \mathbf{x}_2$, $\left(\lambda_1 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 5 \\ -1 \end{bmatrix} - \lambda_4 \begin{bmatrix} 3 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 5 \\ -1 \end{bmatrix} + w_0 = 1$

$\Rightarrow 24\lambda_1 + 26\lambda_2 - 15\lambda_4 + w_0 = 1.$

For $\mathbf{x} = \mathbf{x}_4$, $\left(\lambda_1 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 5 \\ -1 \end{bmatrix} - \lambda_4 \begin{bmatrix} 3 \\ 0 \end{bmatrix}\right)^T \begin{bmatrix} 3 \\ 0 \end{bmatrix} + w_0 = -1$

$\Rightarrow 15\lambda_1 + 15\lambda_2 - 9\lambda_4 + w_0 = -1.$ 2 marks

Combining with the conditions $\sum_{i=1}^{6} \lambda_i y_i = 0 \Rightarrow \lambda_1 + \lambda_2 - \lambda_4 = 0$, we have the following system of linear equations:

$$\begin{bmatrix} 26 & 24 & -15 & 1 \\ 24 & 26 & -15 & 1 \\ 15 & 15 & -9 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_4 \\ w_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \end{bmatrix}$$

. 1 mark

This gives $\lambda_1 = \lambda_2 = 0.25$, $\lambda_4 = 0.5$ and $w_0 = -4$.

As a result, $\mathbf{w} = 0.25 \begin{bmatrix} 5 \\ 1 \end{bmatrix} + 0.25 \begin{bmatrix} 5 \\ -1 \end{bmatrix} - 0.5 \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

The hyperplane is $\mathbf{w}^T\mathbf{x} + w_0 = x_1 - 4 = 0$. 2 marks

Marking scheme

Tests understanding of Support Vector Machines.

3. a. Give brief definitions of the following terms:

i. an artificial neural network

ii. action potential

iii. firing rate

[6 marks]

Answer

i) a parallel architecture composed of many simple processing elements interconnected to achieve certain collective computational capabilities
ii) the signal outputted by a biological neuron
iii) the number of action potentials emitted during a defined time-period.

Marking scheme

2 marks for each correct definition.
Tests knowledge of basic terminology used in neural networks.

Table 3, below, shows a linearly separable dataset.

Table 3

$\mathbf{x}^T$	$class$
$(0,2)$	1
$(2,1)$	1
$(-3,1)$	0
$(-2,-1)$	0
$(0,-1)$	0

b. Is the dataset shown in Table 3:

i. Univariate or multivariate?

ii. Continuous or discrete?

[2 marks]

Answer

i) multivariate

ii) discrete

Marking scheme

1 mark for each correct answer.

Tests knowledge of basic terminology used in pattern recognition.

c. Re-write the dataset shown in Table 3 using augmented vector notation.

[2 marks]

Answer

Using Augmented notation the dataset is:

$\mathbf{x}^T$	t
$(1, 0, 2)$	1
$(1, 2, 1)$	1
$(1, -3, 1)$	0
$(1, -2, -1)$	0
$(1, 0, -1)$	0

Marking scheme

2 marks.

Tests ability to apply a method used in neural networks.

d. 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 $\theta = -2$, $w_1 = 0.5$ and $w_2 = 1$ calculate the output of the neuron in response to each of the samples in the dataset given in Table 3.

[4 marks]

Answer

Using Augmented notation, output is $y = H(\mathbf{wx})$ where $\mathbf{w} = [-\theta, w_1, w_2]$, and $\mathbf{x} = [1, x_1, x_2]^T$. Here, $\mathbf{w} = [2, 0.5, 1]$.

Hence, response to each sample in dataset is:

$\mathbf{x}^T$	$\mathbf{wx}$	$y = H(\mathbf{wx})$
$(1, 0, 2)$	4	1
$(1, 2, 1)$	4	1
$(1, -3, 1)$	1.5	1
$(1, -2, -1)$	0	1(or 0.5)
$(1, 0, -1)$	1	1

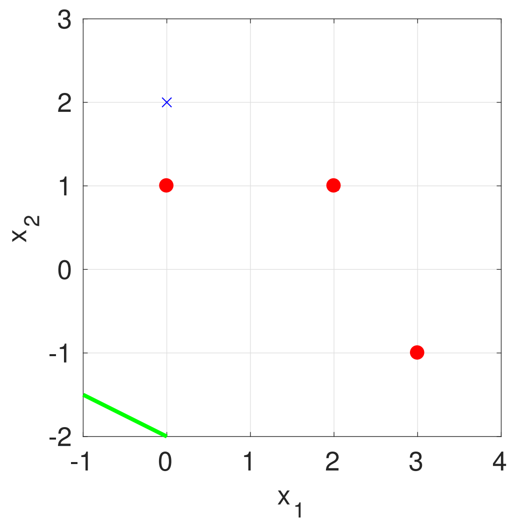
Marking scheme

2 marks for correct method, 2 marks for correct application.
Tests ability to apply a method used in neural networks.

e. Plot the data given in Table 3 and on the same axes plot the decision boundary defined by the linear threshold neuron with parameter values $\theta = -2$, $w_1 = 0.5$ and $w_2 = 1$.

[4 marks]

<u>Answer</u>



<u>Marking scheme</u>

4 marks.

Tests understanding of neural networks.

f. Apply the sequential delta learning algorithm to find the parameters of a linear threshold neuron that will correctly classify the data in Table 3. Assume initial values of $\theta = -2$, $w_1 = 0.5$ and $w_2 = 1$, and a learning rate of 1.

[7 marks]

Answer

Initial weight values are $\mathbf{w} = [2, 0.5, 1]$

For the Sequential Delta Learning Algorithm, weights are updated such that: $\mathbf{w} \leftarrow \mathbf{w} + \eta(t - y)\mathbf{x}^T$. Here, $\eta = 1$.

it	$\mathbf{w}$	$\mathbf{x}^T$	$\mathbf{wx}$	y	t	$\mathbf{w} \leftarrow \mathbf{w} + (t-y)\mathbf{x}^T$
0	$[2, 0.5, 1]$	$[1, 0, 2]$	4	1	1	$[2, 0.5, 1]$
1	$[2, 0.5, 1]$	$[1, 2, 1]$	4	1	1	$[2, 0.5, 1]$
2	$[2, 0.5, 1]$	$[1, -3, 1]$	1.5	1	0	$[2, 0.5, 1] - [1, -3, 1] = [1, 3.5, 0]$
3	$[1, 3.5, 0]$	$[1, -2, -1]$	-6	0	0	$[1, 3.5, 0]$
4	$[1, 3.5, 0]$	$[1, 0, -1]$	1	1	0	$[1, 3.5, 0] - [1, 0, -1] = [0, 3.5, 1]$
5	$[0, 3.5, 1]$	$[1, 0, 2]$	2	1	1	$[0, 3.5, 1]$
6	$[0, 3.5, 1]$	$[1, 2, 1]$	8	1	1	$[0, 3.5, 1]$
7	$[0, 3.5, 1]$	$[1, -3, 1]$	-9.5	0	0	$[0, 3.5, 1]$
8	$[0, 3.5, 1]$	$[1, -2, -1]$	-8	0	0	$[0, 3.5, 1]$
9	$[0, 3.5, 1]$	$[1, 0, -1]$	-1	0	0	$[0, 3.5, 1]$

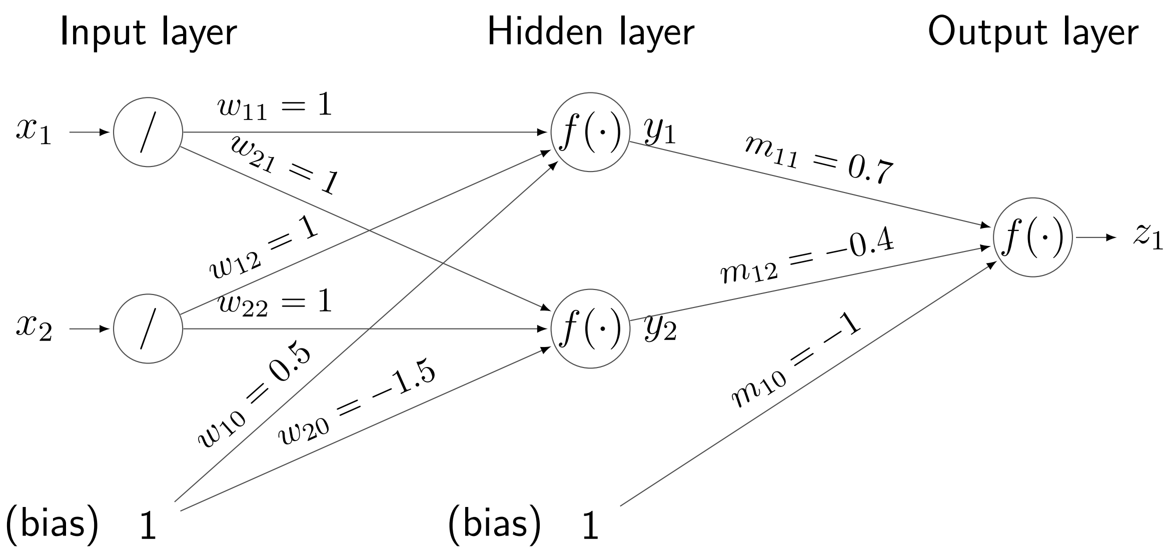
Learning has converged, so required parameters are $\mathbf{w} = (0, 3.5, 1)$.

Marking scheme

2 marks for knowing update rule, 5 marks for correct method.

Tests ability to apply a learning algorithm for defining a neural network.

4. a. A diagram of 2-2-1 fully connected multilayer neural network is shown below, where "/" denotes a linear function, $f(\cdot) = \text{sgn}(\cdot)$ denotes the sign function, $\text{sgn}(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ -1 & \text{if } a < 0 \end{cases}$.



Show that this neural network is able to solve the XOR problem as defined in Table 4.

Table 4

Class	$\mathbf{x}^T = [x_1, x_2]^T$
1	$[-1, -1]$
2	$[-1, +1]$
2	$[+1, -1]$
1	$[+1, +1]$

[10 marks]

<u>Answer</u>

$$\begin{aligned} y_1 &= \text{sgn}(w_{11}x_1 + w_{12}x_2 + w_{10}) \\ &= \text{sgn}(x_1 + x_2 + 0.5) \end{aligned}$$

2 marks

$$\begin{aligned} y_2 &= \text{sgn}(w_{21}x_1 + w_{22}x_2 + w_{20}) \\ &= \text{sgn}(x_1 + x_2 - 1.5) \end{aligned}$$

2 marks

$$\begin{aligned} z_1 &= \text{sgn}(m_{11}y_1 + m_{12}y_2 + m_{10}) \\ &= \text{sgn}(0.7y_1 - 0.4y_2 - 1) \end{aligned}$$

2 marks

Class	$\mathbf{x}^T$	y_1	y_2	z_1
1	$[-1, -1]$	sgn$(-1 - 1 + 0.5) = -1$	sgn$(-1 - 1 - 1.5) = -1$	sgn$(0.7 \times -1 - 0.4 \times -1 - 1) = -1$
2	$[-1, +1]$	sgn$(-1 + 1 + 0.5) = +1$	sgn$(-1 + 1 - 1.5) = -1$	sgn$(0.7 \times +1 - 0.4 \times -1 - 1) = +1$
2	$[+1, -1]$	sgn$(+1 - 1 + 0.5) = +1$	sgn$(+1 - 1 - 1.5) = -1$	sgn$(0.7 \times +1 - 0.4 \times -1 - 1) = +1$
1	$[+1, +1]$	sgn$(+1 + 1 + 0.5) = +1$	sgn$(+1 + 1 - 1.5) = +1$	sgn$(0.7 \times +1 - 0.4 \times +1 - 1) = -1$

2 marks

It can be seen from the table that the output z_1 is -1 for class 1 and the output z_1 is $+1$ for class 2. The output $-1/+1$ clearly shows that the input samples of the XOR problem are classified correctly.

2 marks

Marking scheme

Tests understanding of multilayer neural networks.

—— SOLUTIONS ——

b. A radial basis function (RBF) neural network is to be used to correctly classify the XOR dataset shown in Table 4.

i. There are two hidden units that both employ a Gaussian activation function with a standard deviation of $\sigma_j = \frac{\rho_{\max}}{\sqrt{2n_H}}$. The centres of these hidden units are given by the first and the last input patterns (*i.e.*, the two input patterns in class 1). Compute the outputs at the hidden units for all four input patterns and list them in a table.

[7 marks]

<u>Answer</u>

$\mathbf{c}_1 = [-1 \quad -1]$
$\mathbf{c}_2 = [1 \quad 1]$.
$\sigma_1 = \sigma_2 = \frac{\rho_{\max}}{\sqrt{2n_H}} = \frac{\sqrt{(-1-1)^2+(-1-1)^2}}{\sqrt{2\times 2}} = \sqrt{2}$.
$y_1(\mathbf{x}_p) = exp\left\{-\frac{\|\mathbf{x}_p-\mathbf{c}_1\|^2}{2\sigma_1^2}\right\}$
$y_2(\mathbf{x}_p) = exp\left\{-\frac{\|\mathbf{x}_p-\mathbf{c}_2\|^2}{2\sigma_2^2}\right\}$.

p	x_1	x_2	$y_1(\mathbf{x})$	$y_2(\mathbf{x})$
1	-1	-1	1.0000	0.1353
2	-1	1	0.3679	0.3679
3	1	0	0.3679	0.3679
4	1	1	0.1353	1.0000

<u>Marking scheme</u>

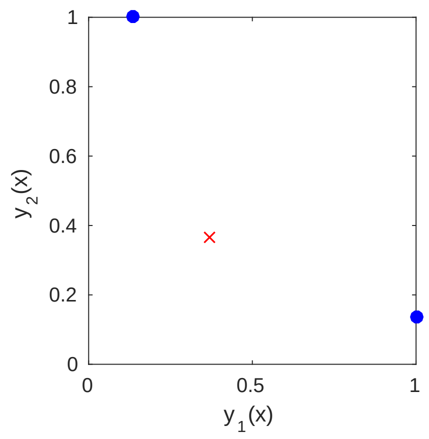
2 marks for correctly calculating σ, 3 marks for correct method of calculating y_1 and y_2, 2 marks for correct application.
Tests ability to apply multilayer neural networks.

ii. Show that the dataset when represented by the hidden units is linearly separable.

[3 marks]

Answer

By plotting $y_2(\mathbf{x})$ against $y_1(\mathbf{x})$ it can be seen that the feature vectors corresponding to the two classes can be separated by a straight line, and so are linearly separable.



Marking scheme

3 marks.

Tests knowledge of basic terminology used in pattern recognition.

iii. The RBF network has one output unit which employs a linear activation function, with a threshold. Write down the equation that would be used to determine the weights of this output neuron using the least squares method. Hence, using the answer you obtained for question 4.b.i write an expression in matrix form that could be used to calculate the weights of the output neuron (you do not need to calculate the pseudo-inverse).

[5 marks]

Answer

The output of the RBF neural network is: $z_1(\mathbf{x}) = \sum_{j=0}^{2} w_{1j} y_j(\mathbf{x}) = \mathbf{w}\mathbf{y}$. Where $\mathbf{w} = [-\theta, w_1, w_2]$ where θ is the threshold. We can calculate the output for all input patterns, in matrix notation, as:

$$\mathbf{Z} = \mathbf{w}\mathbf{Y}$$

If the weights are correct,

$$\mathbf{Z} = \mathbf{w}\mathbf{Y} = \mathbf{T}$$

Hence,

$$\mathbf{w} = \mathbf{T}\mathbf{Y}^\dagger$$

Where $\mathbf{Y}^\dagger$ is the pseudo-inverse of $\mathbf{Y}$.
Therefore,

$$(w_{10} \quad w_{11} \quad w_{12}) = (0 \quad 1 \quad 1 \quad 0) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.3679 & 0.3679 & 0.1353 \\ 0.1353 & 0.3679 & 0.3679 & 1 \end{pmatrix}^\dagger$$

OR

$$(w_{10} \quad w_{11} \quad w_{12}) = (-1 \quad 1 \quad 1 \; -1) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.3679 & 0.3679 & 0.1353 \\ 0.1353 & 0.3679 & 0.3679 & 1 \end{pmatrix}^\dagger$$

Marking scheme

3 marks for correct equation, 2 marks for correct application.

Tests ability to apply multilayer neural networks.

5. a. Give a brief definition of each of the following terms.

i. Feature Selection

ii. Feature Extraction

iii. Dimensionality Reduction

[6 marks]

Answer

i) Feature Selection: choosing a subset of measured features to use for classification.
ii) Feature Extraction: projecting the original feature vectors into a new feature space.
iii) Dimensionality Reduction: selecting/extracting feature vectors of lower dimensionality (length) than the original feature vectors.

Marking scheme

2 marks for each correct definition.
Tests knowledge of basic terminology used in pattern recognition.

b. Write pseudo-code for applying Oja's learning rule to find the first principal component of a multi-dimensional data set.

[4 marks]

Answer

Initialise $\mathbf{w}$ and subtract the mean from all data vectors.
For each zero-mean sample, $\mathbf{x}_k$, in the dataset:
    Update $\mathbf{w}$ such that: $\mathbf{w} \leftarrow \mathbf{w} + \eta y(\mathbf{x}^T - y\mathbf{w})$
Repeat until $\mathbf{w}$ converges or maximum number of iterations is reached.

Marking scheme

4 marks.
Tests ability to describe an algorithm for performing feature extraction.

c. Consider the following 2-dimensional dataset:

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 5 \\ 4 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 8 \\ 7 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 11 \\ 7 \end{pmatrix}.$$

i. Calculate the equivalent zero-mean dataset.

[4 marks]

Answer

Subtract the mean from all data vectors.

The mean of the data is $\mu = \begin{pmatrix} 5.6 \\ 5 \end{pmatrix}$, hence, the zero-mean data is:

$\mathbf{x}'_1 = \begin{pmatrix} -4.6 \\ -3 \end{pmatrix}, \mathbf{x}'_2 = \begin{pmatrix} -2.6 \\ 0 \end{pmatrix}, \mathbf{x}'_3 = \begin{pmatrix} -0.6 \\ -1 \end{pmatrix}, \mathbf{x}'_4 = \begin{pmatrix} 2.4 \\ 2 \end{pmatrix},$
$\mathbf{x}'_5 = \begin{pmatrix} 5.4 \\ 2 \end{pmatrix}.$

Marking scheme

4 marks.

Tests ability to apply an algorithm for performing feature extraction.

ii. Apply two epochs of the batch version of Oja's learning rule to the zero-mean data calculated in answer to question 5.c.i. Use a learning rate of 0.01 and an initial weight vector of [-1,0].

[7 marks]

Answer

For the Batch Oja's learning rule, weights are updated such that: $\mathbf{w} \leftarrow \mathbf{w} + \eta \sum_p y_p(\mathbf{x}_p^T - y_p\mathbf{w})$. Here, $\eta = 0.01$.

Epoch 1, initial $\mathbf{w} = [-1, 0]$

$\mathbf{x}^T$	$y = \mathbf{wx}$	$\mathbf{x}^T - y\mathbf{w}$	$\eta y(\mathbf{x}^T - y\mathbf{w})$	$\mathbf{w}$
$(-4.6, -3)$	4.6	$(0, -3)$	$(0, -0.138)$	
$(-2.6, 0)$	2.6	$(0, 0)$	$(0, 0)$	
$(-0.6, -1)$	0.6	$(0, -1)$	$(0, -0.006)$	
$(2.4, 2)$	-2.4	$(0, 2)$	$(0, -0.048)$	
$(5.4, 2)$	-5.4	$(0, 2)$	$(0, -0.108)$	
		total weight change=	$(0, -0.3)$	$(-1, -0.3)$

Epoch 2, initial $\mathbf{w} = [-1, -0.3]$

$\mathbf{x}^T$	$y = \mathbf{wx}$	$\mathbf{x}^T - y\mathbf{w}$	$\eta y(\mathbf{x}^T - y\mathbf{w})$	$\mathbf{w}$
$(-4.6, -3)$	5.5	$(0.9, -1.35)$	$(0.0495, -0.0743)$	
$(-2.6, 0)$	2.6	$(0, 0.78)$	$(0, 0.0203)$	
$(-0.6, -1)$	0.9	$(0.3, -0.73)$	$(0.0027, -0.0066)$	
$(2.4, 2)$	-3	$(-0.6, 1.1)$	$(0.018, -0.033)$	
$(5.4, 2)$	-6	$(-0.6, 0.2)$	$(0.036, -0.012)$	
		total weight change=	$(0.1062, -0.1055)$	$(-0.8938, -0.4055)$

Marking scheme

7 marks.

Tests ability to apply an algorithm for performing feature extraction.

iii. Using the answer obtained to question 5.c.ii, project the zero-mean data onto the first principal component.

[4 marks]

Answer

Projection of the data onto the subspace spanned by the 1st principal component is given by: $y_i = \mathbf{w}(\mathbf{x}_i - \mu)$, where μ is the mean of the data.

Hence, $y_i = \begin{pmatrix} -0.8938 & -0.4055 \end{pmatrix} \mathbf{x}'_i$

Therefore, the new dataset is:

5.3281, 2.3239, 0.9418, -2.9562, -5.6376.

Marking scheme

4 marks.

Tests ability to apply an algorithm for performing feature extraction.

6. a. Apart from clustering, name one other application for unsupervised learning in pattern recognition.

[1 marks]

Answer

Either embedding or factorisation.

Marking scheme

1 marks.
Tests knowledge of basic terminology used in pattern recognition.

Table 6, below, shows unlabelled 2-dimensional feature vectors which are to be clustered into two classes.

Table 6

i	$\mathbf{x}_i^T$
1	$(1, 0)$
2	$(0, 2)$
3	$(1, 3)$
4	$(3, 0)$
5	$(3, 1)$

b. Apply the k-means algorithm to the dataset shown in Table 6. Use the Euclidean distance criterion and start with initial cluster centres of $\mu_1^T = (3, 2)$ and $\mu_2^T = (4, 0)$. Determine the final cluster centres and the cluster to which each training sample is allocated.

[8 marks]

<u>Answer</u>

First iteration:

Compute $\|\mathbf{x}_i - \mu_j\|$, $i = 1, 2, \ldots, 5$; $j = 1, 2$ and classify samples.

i	$\mathbf{x}_i^T$	$\|\mathbf{x}_i - \mu_1\|$	$\|\mathbf{x}_i - \mu_2\|$	Assigned Cluster
1	$(1, 0)$	2.8284	3	1
2	$(0, 2)$	3	4.4721	1
3	$(1, 3)$	2.2361	4.2426	1
4	$(3, 0)$	2	1	2
5	$(3, 1)$	1	1.4142	1

Recompute μ_j.

$$\mu_1^T = \frac{(1, 0) + (0, 2) + (1, 3) + (3, 1)}{4} = (1.25, 1.5)$$

,

$$\mu_2^T = \frac{(3, 0)}{1} = (3, 0)$$

Second iteration:

Compute $\|\mathbf{x}_i - \mu_j\|$, $i = 1, 2, \ldots, 6$; $j = 1, 2$ and classify samples.

i	$\mathbf{x}_i^T$	$\|\mathbf{x}_i - \mu_1\|$	$\|\mathbf{x}_i - \mu_2\|$	Assigned Cluster
1	$(1, 0)$	1.5207	2	1
2	$(0, 2)$	1.3463	3.6056	1
3	$(1, 3)$	1.5207	3.6056	1
4	$(3, 0)$	2.3049	0	2
5	$(3, 1)$	1.8200	1	2

Recompute μ_j.

$$\mu_1^T = \frac{(1,0) + (0,2) + (1,3)}{3} = (0.6667, 1.6667)$$

,

$$\mu_2^T = \frac{(3,0) + (3,1)}{2} = (3, 0.5)$$

Third iteration: Compute $\|\mathbf{x}_i^T - \mu_j\|$, $i = 1, 2, \ldots, 6$; $j = 1, 2$ and classify samples.

i	$\mathbf{x}_i^T$	$\|\mathbf{x}_i - \mu_1\|$	$\|\mathbf{x}_i - \mu_2\|$	Assigned Cluster
1	$(1, 0)$	1.6997	2.0616	1
2	$(0, 2)$	0.7454	3.3541	1
3	$(1, 3)$	1.3744	3.2016	1
4	$(3, 0)$	2.8674	0.5	2
5	$(3, 1)$	2.4267	0.5	2

Allocation of samples to clusters is unchanged, therefore algorithm has converged and final cluster centres are: $\mu_1^T = (0.6667, 1.6667)$, $\mu_2^T = (3, 0.5)$.

According to the above table, training samples 1 to 3 are in cluster 1, and training samples 4 and 5 are in cluster 2.

Marking scheme

8 marks.

Tests knowledge of and ability to apply an algorithm used for clustering.

c. Apply hierarchical agglomerative clustering to the dataset shown in Table 6, in order to find two clusters. Use the Euclidean distance as the similarity metric and determine the distance between clusters using single-link clustering (where the distance between clusters is the minimum distance between constituent samples). Determine the cluster to which each training sample is allocated.

[8 marks]

Answer

Iteration 1:

Calculate the distance between each pair of clusters.

$\mathbf{x}^T$	cluster	1	2	3	4	5
$(1,0)$	1	-	-	-	-	-
$(0,2)$	2	2.2361	-	-	-	-
$(1,3)$	3	3.0000	1.4142	-	-	-
$(3,0)$	4	2.0000	3.6056	3.6056	-	-
$(3,1)$	5	2.2361	3.1623	2.8284	1.0000	-

clusters 4 and 5 have minimum distance between them, so merge these into a new cluster called "4+5".

Iteration 2:

Calculate the distance between each pair of clusters.

$\mathbf{x}^T$	cluster	1	2	3	4+ 5
$(1,0)$	1	-	-	-	-
$(0,2)$	2	2.2361	-	-	-
$(1,3)$	3	3.0000	1.4142	-	-
$(3,0),(3,1)$	4+5	2.0000	3.1623	2.8284	-

clusters 2 and 3 have minimum distance between them, so merge these into a new cluster called "2+3".

Iteration 3:

Calculate the distance between each pair of clusters.

$\mathbf{x}^T$	cluster	1	2+3	4+ 5
$(1,0)$	1	-	-	-
$(0,2),(1,3)$	2+3	2.2361	-	-
$(3,0),(3,1)$	4+5	2.0000	2.8284	-

clusters 1 and 4+5 have minimum distance between them, so merge these into a new cluster called "1+4+5".

There are two clusters, so terminate.

One cluster contains samples 1, 4, and 5 and the other cluster contains samples 2 and 3.

Marking scheme

8 marks.
Tests knowledge of and ability to apply an algorithm used for clustering.

d. Apply the competitive learning algorithm (without normalisation) to the dataset shown in Table 6. Perform five iterations of the algorithm with the samples chosen in the order of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$. Start with initial cluster centres of $\mathbf{w}_1^T = (1,1)$ and $\mathbf{w}_2^T = (2,2)$. Use a learning rate of $\eta = 0.5$. Determine the final cluster centres and the cluster to which each training sample is allocated.

[8 marks]

Answer

It.	$\mathbf{x}^T$	$\lVert\mathbf{x}-\mathbf{w}_1\rVert$	$\lVert\mathbf{x}-\mathbf{w}_2\rVert$	$j \leftarrow \arg\min_{j'} \lVert\mathbf{x}-\mathbf{w}_{j'}\rVert$	$\mathbf{w}_j \leftarrow \mathbf{w}_j + \eta(\mathbf{x}-\mathbf{w}_j)$
1	$(1,0)$	1	2.2361	1	$\mathbf{w}_1^T = (1,1) + 0.5 \times [(1,0)-(1,1)] = (1,0.5)$
2	$(0,2)$	1.8028	2	1	$\mathbf{w}_1^T = (1,0.5) + 0.5 \times [(0,2)-(1,0.5)] = (0.5,1.25)$
3	$(1,3)$	1.8200	1.4142	2	$\mathbf{w}_2^T = (2,2) + 0.5 \times [(1,3)-(2,2)] = (1.5,2.5)$
4	$(3,0)$	2.7951	2.9155	1	$\mathbf{w}_1^T = (0.5,1.25) + 0.5 \times [(3,0)-(0.5,1.25)] = (1.75,0.625)$
5	$(3,1)$	1.3050	2.1213	1	$\mathbf{w}_1^T = (1.75,0.625) + 0.5 \times [(3,1)-(1.75,0.625)] = (2.375,0.8125)$

After 5 iterations, we have $\mathbf{w}_1^T = (2.375, 0.8125)$, $\mathbf{w}_2^T = (1.5, 2.5)$.

The cluster allocated to each training sample is as follows:

$\mathbf{x}^T$	$\|\mathbf{x}-\mathbf{w}_1\|$	$\|\mathbf{x}-\mathbf{w}_2\|$	$j \leftarrow \arg\min_{j'} \|\mathbf{x}-\mathbf{w}_{j'}\|$
$(1,0)$	1.5971	2.5495	1
$(0,2)$	2.6553	1.5811	2
$(1,3)$	2.5838	0.7071	2
$(3,0)$	1.0251	2.9155	1
$(3,1)$	0.6525	2.1213	1

Marking scheme

8 marks.

Tests knowledge of and ability to apply an algorithm used for clustering.