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 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

—— SOLUTIONS ——

1.

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

[4 marks]

Answer

Calculate $P(\omega_j|x)$ for each class ω_j. Classify exemplar x in class ω_i for which $P(\omega_i|x) > P(\omega_j|x) \forall j \neq i$.

Marking scheme

2 marks for knowing the need to calculate the posterior. 2 marks for knowing how to choose the class.

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]

Answer

The closest three samples to $x = 4$ are at: 3.86, 3.54, and 3.34. We need to make a "volume" of radius $|4 - 3.34| = 0.66$ (diameter 1.32) around $x = 4$ to encompass these three samples.

Therefore density is:

$$p(x = 4|\omega_1) = \frac{k/n}{V} = \frac{3/9}{1.32} = 0.2525$$

Marking scheme

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

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]

Answer

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

$$\hat{\mu_2} = \frac{(15.85 + 4.75 + 17.17 + 5.63 + 1.68 + 5.57 + 0.98)}{7} = 7.3757$$

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,

$$\hat{\sigma_2}^2 = \frac{1}{6}[(15.85 - 7.3757)^2 + (4.75 - 7.3757)^2 + (17.17 - 7.3757)^2 + (5.63 - 7.3757)^2 + (1.68 - 7.3757)^2 + (5.57 - 7.3757)^2 + (0.98 - 7.3757)^2] = 42.3817$$

Marking scheme

2 marks for correct methods, 2 marks for correct application.

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]

Answer

Estimate of probability density is:

$$p_n(x|\omega_3) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

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

For this question, $n = 5$, $h_n = 2$, so:

$$\begin{aligned} p(x = 4|\omega_3) &= \frac{1}{5}\sum_{i=1}^{5}\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-0.5\left(\frac{4 - x_i}{2}\right)^2\right) \\ &= \frac{1}{10\sqrt{2\pi}}\left[\exp\left(-0.5\left(\frac{4 - 3.75}{2}\right)^2\right)\right. \\ &\quad + \exp\left(-0.5\left(\frac{4 - 4}{2}\right)^2\right) \\ &\quad \left. + \ldots\cdots + \exp\left(-0.5\left(\frac{4 - 1.59}{2}\right)^2\right)\right] \end{aligned}$$

$$= \frac{1}{10\sqrt{2\pi}}\left[0.9922 + 1.0000 + 0.9655 + 0.7990 + 0.4838\right]$$

$$= 0.1692$$

Marking scheme

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

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]

Answer

$$P(\omega_1) = \frac{9}{21} = 0.4286$$
$$P(\omega_2) = \frac{7}{21} = 0.3333$$
$$P(\omega_3) = \frac{5}{21} = 0.2381$$

Marking scheme

1 mark for each correct answer.

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]

Answer

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

Ignoring $p(x)$:

$$P(\omega_1|x = 4) \propto p(x = 4|\omega_1)P(\omega_1) = 0.2525 \times 0.4286 = 0.1082$$

$$P(\omega_2|x = 4) \propto p(x = 4|\omega_2)P(\omega_2)$$

From sub-question 1.c, $p(x|\omega_2)$ is given by $N(7.3757, \sqrt{42.3817})$, so $p(x = 4|\omega_2) = \frac{1}{\sqrt{2\pi \times 42.3817}} \exp\left(-\frac{(4-7.3757)^2}{2\times 42.3817}\right) = 0.0536$

Therefore,

$$P(\omega_2|x=4) \propto 0.0536 \times 0.3333 = 0.0179$$

$$P(\omega_3|x=4) \propto p(x=4|\omega_3)P(\omega_1) = 0.1692 \times 0.2381 = 0.0403$$

Therefore most likely class for sample $x=4$ is ω_1.

Marking scheme

6 marks. 3 for correctly calculating the $P(\omega_j|x)$ values, 2 for correctly calculating $p(x=4|\omega_2)$, 1 for choosing correct class label.

2.

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

i. Supervised Learning

ii. Generalisation

[4 marks]

Answer

i) Supervised Learning: methods of finding patterns in data when the dataset includes class information for each exemplar.
ii) Generalization: how well a model/classifier performs on new data (i.e. data not used for training).

Marking scheme

2 marks for each correct definition.

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]

Answer

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = (-1, -2) \times \begin{pmatrix} 2.5 \\ 1.5 \end{pmatrix} + 3 = -2.5$$

Classification rule is that $\mathbf{x}$ is assigned to class 1 if $g(\mathbf{x}) > 0$, and $\mathbf{x}$ is assigned to class 2 otherwise. So exemplar is predicted to be in class 2.

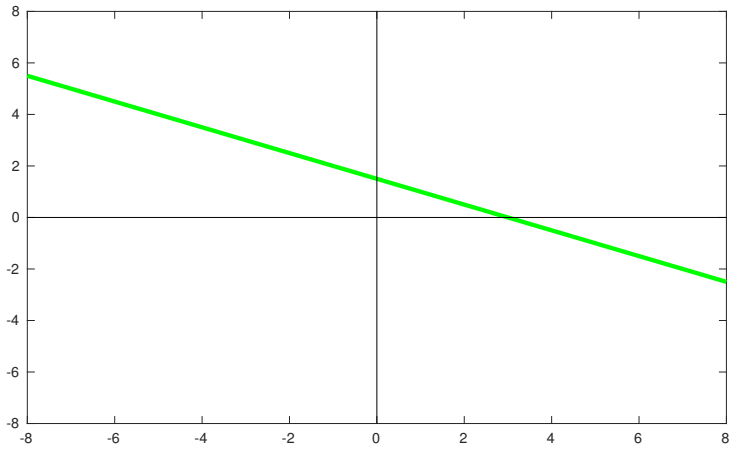
Marking scheme

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

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

[4 marks]

Answer



Marking scheme

2 marks for slope of approximately -0.5 ($= \frac{-w_1}{w_2} = \frac{-(-1)}{-2}$), 2 marks for an intercept of approximately 1.5 ($= \frac{-w_0}{w_2} = \frac{-3}{-2}$).

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]

Answer

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

Marking scheme

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]

Answer

Using augmented output notation, the output is $y = H(\mathbf{wx})$, where $\mathbf{w} = [-\theta, w_1, w_2]$, and $\mathbf{x} = [1, x_1, x_2]^T$. Here, $\mathbf{w} = [1, 3, 0.5]$.
For each sample in the dataset, the response is:

$\mathbf{x}^T$	$\mathbf{wx}$	$y = H(\mathbf{wx})$
(1,2,-1)	6.5	1
(1,-1,0)	-2	0
(1,0,0)	1	1
(1,1,1)	4.5	1
(1,0,-1)	0.5	1

Marking scheme

2 marks for correct method, 2 marks for correct application

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]

Answer

The initial weight vector is $\mathbf{w} = [1, 3, 0.5]$. In the sequential delta learning algorithm, the learning rule is $\mathbf{w} \leftarrow \mathbf{w} + \eta(t - y)\mathbf{x^T}$. Here $\eta = 1$.

i	$\mathbf{w}$	$\mathbf{x}$	$\mathbf{wx}$	y	t	$\mathbf{w} \leftarrow \mathbf{w} + \eta(t - y)\mathbf{x^T}$
0	[1,3,0.5]	[1,2,-1]	6.5	1	0	[0,1,1.5]
1	[0,1,1.5]	[1,-1,0]	-1	0	1	[1,0,1.5]
2	[1,0,1.5]	[1,0,0]	1	1	1	[1,0,1.5]
3	[1,0,1.5]	[1,1,1]	2.5	1	0	[0,-1,0.5]
4	[0,-1,0.5]	[1,0,-1]	-0.5	0	1	[1,-1,-0.5]
5	[1,-1,-0.5]	[1,2,-1]	-0.5	0	0	[1,-1,-0.5]
6	[1-1,-0.5]	[1,-1,0]	2	1	1	[1,-1,-0.5]
7	[1,-1,-0.5]	[1,0,0]	1	1	1	[1,-1,-0.5]
8	[1,-1,-0.5]	[1,1,1]	-0.5	0	0	[1,-1,-0.5]
9	[1,-1,-0.5]	[1,0,-1]	1.5	1	1	[1,-1,-0.5]

Learning has converged at weights $\mathbf{w} = [1, -1, -0.5]$

Marking scheme

2 marks for correct learning rule, and 5 marks for correct method.

3.

a. Give brief definitions of the following terms:

i. Feature selection

ii. Feature extraction

[4 marks]

Answer

1. Feature selection - choosing a subset of features from a given dataset to use for the learning problem.
2. Feature extraction - projecting the original features from a given dataset, into a new feature space.

Marking scheme

2 marks for each correct definition

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]

Answer

The principal components of a dataset are a set of linearly uncorrelated orthogonal bases, where the first principal component is the basis with the greatest variance, and subsequent bases have decreasing variance. Performing PCA allows for one to find these principal components and discard principal components with low variance. By then projecting the dataset onto the remaining principal components, learning can be performed in a lower-dimensional space, which only contains components that contain the majority of the variance of the dataset.

Marking scheme

2 marks for correct explanation of principal components, 2 marks for correct explanation of their use in learning

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

[5 marks]

Answer

Given a dataset $\mathbf{x}$

1. Calculate the mean of $\mathbf{x}$, and subtract from the dataset.
2. Calculate the covariance matrix of the mean-zeroed dataset.
3. Find the eigenvectors and eigenvalues of the covariance matrix
4. Order the eigenvalues from largest to smallest, and discard eigenvectors corresponding to small eigenvalues
5. Form a matrix $\hat{V}$ of the remaining principal components.
6. Project the mean-zeroed dataset onto the PCA subspace by the transpose of $\hat{V}$.

Marking scheme

5 marks for correct method. Partial marks for partially correct answers.

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]

Answer

Oja's update rule is given as: $\mathbf{w} \leftarrow \mathbf{w} + \eta y(\mathbf{x}^T - y\mathbf{w})$.

In standard Hebbian learning, the learnt weights approach infinity as the number of iterations increase (for a positive learning rate). Oja's rule controls learning by causing the weight vector to approach unit length.

Marking scheme

2 marks for correct update rule, 3 marks for correct explaination

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]

Answer

Fisher's method for defining a decision boundary finds a weight vector that maximises the between class scatter, while minimising the within class scatter. This is given by the objective function $J(\mathbf{w}) = \frac{sb}{sw}$, where sb is the between class scatter, and sw is the within class scatter.

Marking scheme

2 marks for correct definition, 1 mark for correct objective function

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.

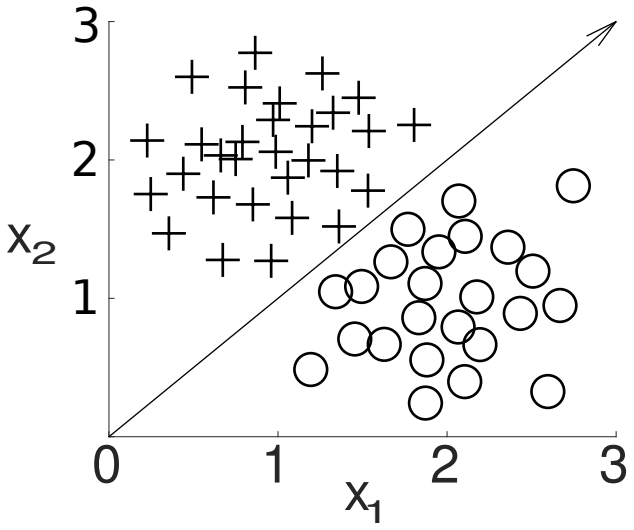


Figure 1

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

[4 marks]

Answer

This projection is not a suitable axis for maximally discriminating between classes. Projections made along this axis will result in large amounts of overlap between the two classes. As the decision boundary is orthogonal to the projection axis, this means that the decision boundary will run through the centre of both clusters of data, instead of seperating them.

Marking scheme

1 mark for correct answer, 3 marks for justification

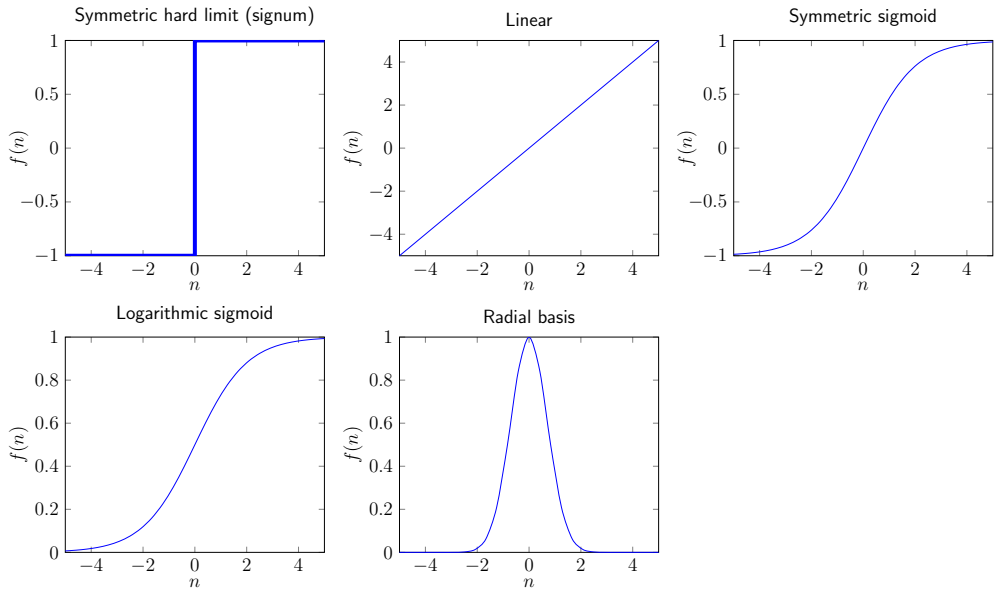
4.

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]

Answer



Marking scheme

1 mark for each figure

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]

Answer

Symmetric hard limit (signum) function is not suitable. 2 marks.

In backpropagation algorithm, the learning rule depends on the rate of change between the output error and the node inputs, which is related to the derivative of the activation functions. Symmetric hard limit (signum) function is not differentiable and its derivative is not defined at some points. 3 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]

Answer

Training set	1 mark
Validation set	1 mark
Test set	1 mark

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]

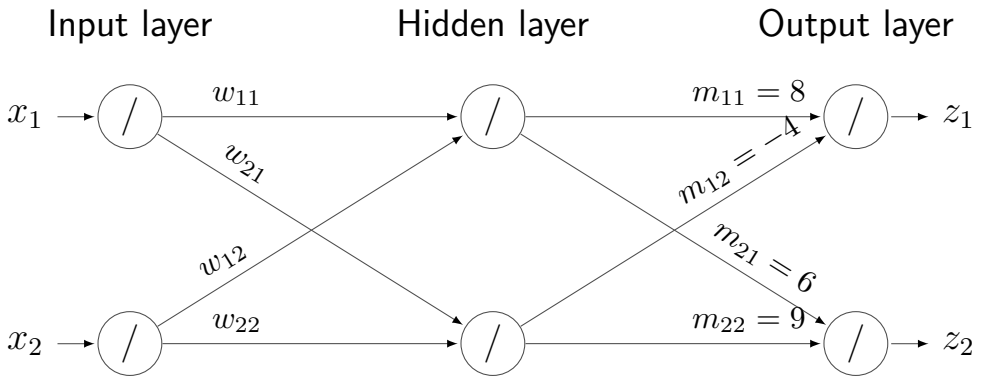


Figure 2: A diagram of 3-layer neural network.

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.

Answer

The output of neural network can be expressed as

$$\mathbf{z} = \mathbf{W}_{kj}\mathbf{W}_{ji}\mathbf{x}$$

where $\mathbf{x} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \end{bmatrix}$, $\mathbf{z} = \begin{bmatrix} z_1(1) & z_1(2) & \cdots & z_1(N) \\ z_2(1) & z_2(2) & \cdots & z_2(N) \end{bmatrix}$,

$\mathbf{W}_{ji} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$ and $\mathbf{W}_{kj} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \begin{bmatrix} 8 & -4 \\ 6 & 9 \end{bmatrix}$.

From $\mathbf{z} = \mathbf{W}_{kj}\mathbf{W}_{ji}\mathbf{x}$, we have $\mathbf{W}_{kj}^{-1}\mathbf{z} = \mathbf{W}_{ji}\mathbf{x} \Rightarrow \mathbf{W}_{ji} = \mathbf{W}_{kj}^{-1}\mathbf{z}\mathbf{x}^T(\mathbf{x}\mathbf{x}^T)^{-1}$.

4 marks

From the table, we have $\mathbf{x} = \begin{bmatrix} 2 & -4 \\ -0.5 & -6 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} 98 & -168 \\ 7.5 & -246 \end{bmatrix}$.

4 marks

$$\mathbf{W}_{ji} = \mathbf{W}_{kj}^{-1}\mathbf{z}\mathbf{x}^T(\mathbf{x}\mathbf{x}^T)^{-1} = \begin{bmatrix} 8 & -4 \\ 6 & 9 \end{bmatrix}^{-1} \times \begin{bmatrix} 98 & -168 \\ 7.5 & -246 \end{bmatrix} \times \begin{bmatrix} 2 & -4 \\ -0.5 & -6 \end{bmatrix}^T \times \left(\begin{bmatrix} 2 & -4 \\ -0.5 & -6 \end{bmatrix} \times \begin{bmatrix} 2 & -4 \\ -0.5 & -6 \end{bmatrix}^T \right)^{-1} = \begin{bmatrix} 5 & 1 \\ -2 & 3 \end{bmatrix}.$$

So, $w_{11} = 5$, 1 mark

$w_{12} = 1$, 1 mark

$w_{21} = -2$, 1 mark

and $w_{22} = 3$. 1 mark

5.

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

[3 marks]

Answer

A weak classifier is a classifier with recognition performance that is slightly better than choosing randomly/blindly.

Marking scheme

3 marks.

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

[3 marks]

Answer

The idea of ensemble methods is to generate a collection of weak classifiers from a given dataset and combine the weak classifiers to be a final stronger one, e.g., by averaging or voting.

Marking scheme

3 marks.

c. Name two methods used for designing ensemble classifiers.

[2 marks]

Answer

Bagging	1 mark
Boosting	1 mark

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)†
    ϵ_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.

Answer

In the first iteration, i.e., $k = 1$, $W_1(i) = 1/5$ for $i =$ 1, 2, 3, 4, 5. 2 marks

The overall weighted error rate for each classifier is listed below: Classifier $h_1(\mathbf{x})$: $E_1 = 1/5 = 0.2$ (due to the misclassification of $\mathbf{x}_1$) 2 marks

Classifier $h_2(\mathbf{x})$: $E_2 = 1/5+1/5 = 0.4$ (due to the misclassification of $\mathbf{x}_2$ and $\mathbf{x}_3$) 2 marks

Classifier $h_3(\mathbf{x})$: $E_3 = 1/5+1/5 = 0.4$ (due to the misclassification of $\mathbf{x}_4$ and $\mathbf{x}_5$) 2 marks

Classifier $h_1(\mathbf{x})$ gives the minimum overall weighted error rate. So, it is chosen to be the classifier $\hat{h}_1(\mathbf{x})$, i.e., $\hat{h}_1(\mathbf{x}) = h_1(\mathbf{x})$. 2 marks

$\epsilon_1 = E_1 = 0.2$ 2 marks

$\alpha_1 = \frac{1}{2}\ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) = \frac{1}{2}\ln\left(\frac{1-0.2}{0.2}\right) = 0.6931$ 2 marks

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]

Answer

$H(\mathbf{x}) = \text{sgn}\big(\alpha_1\hat{h}_1(\mathbf{x}) + \alpha_2\hat{h}_2(\mathbf{x}) + \alpha_3\hat{h}_3(\mathbf{x})\big) = \text{sgn}\big(0.6931h_1(\mathbf{x}) + 0.8h_3(\mathbf{x}) + 0.9h_2(\mathbf{x})\big)$

Marking scheme

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]

Answer

$\frac{2}{\|\mathbf{w}\|}$ is the margin of the hyperplane, which is going to be maximised. Minimising $\frac{\|\mathbf{w}\|}{2}$ is equivalent to maximising $\frac{2}{\|\mathbf{w}\|}$. Making it squared is to make the analysis easier when doing calculus.

Marking scheme

4 marks. Partial marks available for partially correct answers.

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

[4 marks]

Answer

"$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1$" can be separated into two cases: "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$" and "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 1$".

Considering a sample $\mathbf{x}_i$ with $y_i = +1$, $\mathbf{w}^T\mathbf{x}_i + w_0 = 1$ is achieved if it is a support vector and correctly classified which leads to $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$. Considering a sample $\mathbf{x}_i$ with $y_i = -1$, $\mathbf{w}^T\mathbf{x}_i + w_0 = -1$ is achieved if it is a support vector and correctly

classified which leads to $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$. As a result, it can be concluded that "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$" happens when the sample $\mathbf{x}_i$ is a support vector, which is lying at the hyperplane "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$" and correctly classified.

Considering a sample $\mathbf{x}_i$ with $y_i = +1$, $\mathbf{w}^T\mathbf{x}_i + w_0 > 1$ is achieved if it is correctly classified which leads to $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 1$. Considering a sample $\mathbf{x}_i$ with $y_i = -1$, $\mathbf{w}^T\mathbf{x}_i + w_0 < -1$ is achieved if it is correctly classified which leads to $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 1$. As a result, "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 1$" happens when the sample $\mathbf{x}_i$ lying on the correct side of the hyperplane, i.e., the sample is classified correctly.

Summarising the above, "$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1$" is a constraint to be satisfied when all samples are correctly classified.

Marking scheme

4 marks. Partial marks available for partially correct answers.

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]

Answer

A feature mapping function $\mathbf{z}_i = \Phi(\mathbf{x}_i)$ can be applied to the samples. SVM classifier is to be designed in the $\mathbf{z}$ domain after mappings.

Marking scheme

2 marks

b. A training dataset is given as follows:

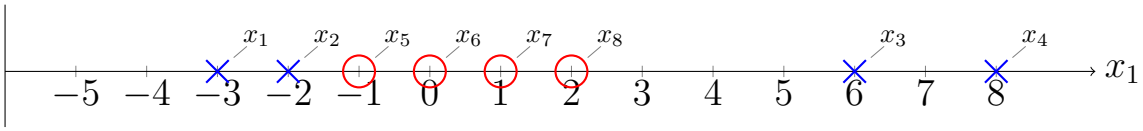
$$\text{Class 1: } x_1 = -3;\ x_2 = -2;\ x_3 = 6;\ x_4 = 8.$$
$$\text{Class 2: } x_5 = -1;\ x_6 = 0;\ x_7 = 1;\ x_8 = 2.$$

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

[4 marks]

Answer

The dataset is not linearly separable as it cannot be separated completely into two classes with a linear classifier.



Marking scheme

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]

Answer

Define the label for Class 1 as $+1$ and Class 2 as -1 so we have $y_1 = y_2 = y_3 = y_4 = +1$ for x_1, x_2, x_3 and x_4; and $y_5 = y_6 = y_7 = y_8 = -1$ for x_5, x_6, x_7 and x_8.

After feature mapping, the dataset in $\mathbf{z}$ feature space is linearly separable. The support vectors are given as $\mathbf{z}_2 = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$, $\mathbf{z}_5 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $\mathbf{z}_8 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$.

The hyperplane is $\mathbf{w}^T\mathbf{z} + w_0 = 0$ where $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} x \\ x^2 \end{bmatrix}$.

As $\mathbf{z}_2$, $\mathbf{z}_5$ and $\mathbf{x}_8$ are support vectors, we have $\mathbf{w} = \lambda_2 y_2 \mathbf{z}_2 + \lambda_5 y_5 \mathbf{z}_5 + \lambda_8 y_8 \mathbf{z}_8 = \lambda_1 \begin{bmatrix} -2 \\ 4 \end{bmatrix} - \lambda_5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \lambda_8 \begin{bmatrix} 2 \\ 4 \end{bmatrix}$. 2 marks

Recall that $y_i(\mathbf{w}^T\mathbf{z} + w_0) = 1$ when $\mathbf{z}$ is a support vector.

For $\mathbf{z} = \mathbf{x}_2$, $\left(\lambda_2 \begin{bmatrix} -2 \\ 4 \end{bmatrix} - \lambda_5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \lambda_8 \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} -2 \\ 4 \end{bmatrix} + w_0 = 1$

$\Rightarrow 20\lambda_2 - 6\lambda_5 - 12\lambda_8 + w_0 = 1.$ 2 marks

For $\mathbf{z} = \mathbf{x}_5$, $\left(\lambda_2 \begin{bmatrix} -2 \\ 4 \end{bmatrix} - \lambda_5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \lambda_8 \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} -1 \\ 1 \end{bmatrix} + w_0 = -1$

$\Rightarrow 6\lambda_2 - 2\lambda_5 - 2\lambda_8 + w_0 = -1.$ 2 marks

For $\mathbf{z} = \mathbf{x}_8$, $\left(\lambda_2 \begin{bmatrix} -2 \\ 4 \end{bmatrix} - \lambda_5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \lambda_8 \begin{bmatrix} 2 \\ 4 \end{bmatrix}\right)^T \begin{bmatrix} 2 \\ 4 \end{bmatrix} + w_0 = -1$

$\Rightarrow 12\lambda_2 - 2\lambda_5 - 20\lambda_8 + w_0 = -1.$ 2 marks

As $\sum_{i=1}^{8} \lambda_i y_i = 0$, we have $\lambda_2 - \lambda_5 - \lambda_8 = 0 \Rightarrow \lambda_2 = \lambda_5 + \lambda_8$. So, we have

$14\lambda_5 + 8\lambda_8 + w_0 = 1,$

$4\lambda_5 + 4\lambda_8 + w_0 = -1,$

$10\lambda_5 - 8\lambda_8 + w_0 = -1$

It follows that $\begin{bmatrix} 14 & 8 & 1 \\ 4 & 4 & 1 \\ 10 & -8 & 1 \end{bmatrix} \begin{bmatrix} \lambda_5 \\ \lambda_8 \\ w_0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$.

It gives $\lambda_2 = 0.2500$, $\lambda_5 = 0.1667$, $\lambda_8 = 0.0833$ and $w_0 = -2$. As a result, $\mathbf{w} = 0.2500 \begin{bmatrix} -2 \\ 4 \end{bmatrix} - 0.16667 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - 0.0833 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} -0.4999 \\ 0.5001 \end{bmatrix}$

2 marks

The hyperplane is

$$\mathbf{w}^T\mathbf{z} + w_0 = -0.4449x + 0.5001x^2 - 2 = 0.$$

1 mark