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 2019 (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. Write down Bayes' Rule, giving an expression for the posterior, $P(\omega_j|x)$, in terms of the likelihood, prior and evidence. Where ω_j represents category j, and x is a sample.

[3 marks]

Answer

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

Marking scheme

3 marks.

b. Briefly describe what is meant by a "Minimum Error Rate classifier".

[2 marks]

Answer

A classifier that minimises the number of misclassified exemplars.

Marking scheme

2 marks.

c. Explain how Bayes' Rule can be used to determine the class of an exemplar so as to minimise error rate.

[2 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.

d. Show, mathematically, that the k-nearest-neighbour classifier provides an estimate of $P(\omega_j|x)$

[5 marks]

Answer

From Bayes theorem:

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

Assume we have a dataset with n samples, where n_j samples come from class ω_j. To classify an unknown sample x, place a volume V around x that captures k samples, this gives estimates of:
the likelihood: $p(x|\omega_j) = \frac{k_j}{n_j V}$
the evidence: $p(x) = \frac{k}{nV}$
the prior: $P(\omega_j) = \frac{n_j}{n}$.
Substituting these values into Bayes rule:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} = \frac{\frac{k_j}{n_j V}\frac{n_j}{n}}{\frac{k}{nV}} = \frac{k_j}{k}$$

Hence, we can estimate the posterior probability that x is in class ω_j by counting the fraction of samples within a region centred on x that are labelled ω_j.

Marking scheme

5 marks

Table 1, below, shows samples from two classes:

class	$\mathbf{x}^T$
ω_1	(5,1)
ω_1	(5,-1)
ω_2	(3,0)
ω_2	(2,1)
ω_2	(4,2)

Table 1

e. Apply the k-nearest-neighbour classifier to the data given in Table 1 to determine the class of a new sample with value $\mathbf{x}^T = (4, 0)$. Use Euclidean distance to measure the distance between samples, and use:

i. k=1

ii. k=3

[4 marks]

Answer

class	$\mathbf{x}^T$	Euclidean Distance to (4,0)
ω_1	(5,1)	$\sqrt{2}$
ω_1	(5,-1)	$\sqrt{2}$
ω_2	(3,0)	1
ω_2	(2,1)	$\sqrt{5}$
ω_2	(4,2)	2

i) For k=1. The nearest-neighbour is (3,0). This is in class 2, so the new sample is given the class label ω_2.

ii) For k=3. The three nearest-neighbours are (3,0), (5,1), and (5,-1). These are in classes 2, 1, and 1. As the majority are in class 1 the new sample is given the class label ω_1.

Marking scheme

2 marks for calculating the Euclidean Distances, 2 marks for correctly applying the k-NN to allocate the new sample to the correct class.

f. Apply k_n nearest neighbours to the data shown in Table 1 in order to estimate the likelihoods, or class-conditional probability densities, at location (4,0) for each class, *i.e.*, calculate $p(x|\omega_1)$ and $p(x|\omega_2)$. Use Euclidean distance to measure the distance between samples, and use k=1.

[4 marks]

Answer

For $p(x|\omega_1)$, the closest sample to (4,0) is at (5,1) (or (5-1)). We need to make a circle of radius $\sqrt{2}$ ("volume" $= \pi r^2 = 2\pi$) around (4,0) to encompass this sample.

Therefore density is:

$$p(x|\omega_1) = \frac{k/n}{V} = \frac{1/2}{2\pi} = 0.0796$$

For $p(x|\omega_2)$, the closest sample to (4,0) is at (3,0). We need to make a circle of radius 1 ("volume" $= \pi$) around (4,0) to encompass this sample.

Therefore density is:

$$p(x|\omega_2) = \frac{k/n}{V} = \frac{1/3}{\pi} = 0.1061$$

Marking scheme

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

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

[2 marks]

Answer

$$P(\omega_1) = \frac{2}{5} = 0.4$$
$$P(\omega_2) = \frac{3}{5} = 0.6$$

Marking scheme

1 mark for each correct answer.

h. Using the answers to sub-questions 1.f and 1.g apply Bayes' Rule to determine the posterior for each class for the sample with feature vector (4,0), *i.e.*, calculate $P(\omega_1|x)$ and $P(\omega_2|x)$. If you have failed to answer sub-questions 1.f and 1.g use the following values: $p(x|\omega_1) = 0.2, p(x|\omega_2) = 0.1, P(\omega_1) = 0.3$, and $P(\omega_2) = 0.7$.

[3 marks]

Answer

Using correct answers to sub-questions 1.f and 1.g:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$
$$P(\omega_1|x) = \frac{0.0796 \times 0.4}{p(x)} = \frac{0.0318}{p(x)}$$
$$P(\omega_2|x) = \frac{0.1061 \times 0.6}{p(x)} = \frac{0.0637}{p(x)}$$
$$p(x) = \sum_j p(x|\omega_j)P(\omega_j) = 0.0318 + 0.0637 = 0.0955$$
$$P(\omega_1|x) = \frac{0.0318}{0.0955} = 0.3333$$

$$P(\omega_2|x) = \frac{0.0637}{0.0955} = 0.6667$$

Using given (incorrect) values:

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

$$P(\omega_1|x) = \frac{0.2 \times 0.3}{p(x)} = \frac{0.06}{p(x)}$$

$$P(\omega_2|x) = \frac{0.1 \times 0.7}{p(x)} = \frac{0.07}{p(x)}$$

$$p(x) = \sum_j p(x|\omega_j)P(\omega_j) = 0.06 + 0.07 = 0.13$$

$$P(\omega_1|x) = \frac{0.06}{0.13} = 0.4615$$

$$P(\omega_2|x) = \frac{0.07}{0.13} = 0.5385$$

Marking scheme

3 marks.

2.

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

i. Feature Space

ii. Decision Boundary

[4 marks]

Answer

i) Feature Space: the (multidimensional) space defined by the feature vectors in the dataset.
ii) Decision boundary: the boundary that separates different regions of feature space which are classified into different classes.

Marking scheme

2 marks for each correct definition.

b. A classifier consists of three linear discriminant functions: $g_1(\mathbf{x})$, $g_2(\mathbf{x})$, and $g_3(\mathbf{x})$. The parameters of these three discriminant functions, in augmented notation, are: $\mathbf{a}_1^T = (1, 0.5, 0.5)$, $\mathbf{a}_2^T = (-1, 2, 2)$, $\mathbf{a}_3^T = (2, -1, -1)$. Determine the class predicted by this classifier for a feature vector $\mathbf{x}^T = (0, 1)$.

[3 marks]

Answer

Writing the feature vector in augmented notation: $\mathbf{y}^T = (1, 0, 1)$
$g_j(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$. Therefore,

$$g_1(\mathbf{x}) = (1, 0.5, 0.5) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = 1.5$$

$$g_2(\mathbf{x}) = (-1, 2, 2) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = 1$$

$$g_3(\mathbf{x}) = (2, -1, -1) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = 1$$

The sample belongs to the class of the linear discriminant function that has the maximum response, so $\mathbf{x}^T = (0, 1)$ is predicted to be in class 1.

Marking scheme

3 marks.

Table 2, below, shows a dataset consisting of five exemplars that come from three classes:

Class	Feature Vector, $\mathbf{x}^T$
1	(0,1)
1	(1,0)
2	(0.5,0.5)
2	(1,1)
3	(0,0)

Table 2

c. Do the linear discriminant functions defined in sub-question 2.b correctly classify the data shown in Table 2? Justify your answer.

[4 marks]

Answer

No. For justification, either:

(A) calculate the predicted class for each sample (until a mis-classified exemplar is found):

Class	Feature Vector, $\mathbf{x}^T$	$g_1(\mathbf{x})$	$g_2(\mathbf{x})$	$g_2(\mathbf{x})$	predicted class
1	(0,1)	1.5	1	1	1
1	(1,0)	1.5	1	1	1
2	(0.5,0.5)	1.5	1	1	1
2	(1,1)	2	3	0	2
3	(0,0)	1	-1	2	3

And note that the third sample is mis-classified so the linear discriminant functions do not correctly classify the data.

Or:

(B) plot data in feature space, and note that as linear discriminant functions define decision boundaries that are straight lines, and the data is not linearly separable (sample 3 can not be separated from both samples 1 and 2 using a straight line), the linear discriminant functions can not correctly classify the data.

Marking scheme

1 mark for correct answer, 3 marks for justification.

d. It is suggested that a quadratic discriminant function of the form $g(\mathbf{x}) = w_0 + \sum_i w_i x_i + \sum_{i,j} w_{ij} x_i x_j$ could successfully classify the data. To learn such a quadratic discriminant function we can learn a linear discriminant function in an expanded feature space. Re-write Table 2 showing the new feature vectors in this expanded feature space.

[2 marks]

Answer

The feature space needs to be expanded by adding an element equal to $x_1 \times x_2$ to each feature vector:

Class	Feature Vector, x^T
1	(0,1,0)
1	(1,0,0)
2	(0.5,0.5,0.25)
2	(1,1,1)
3	(0,0,0)

Could alternatively (or additionally) add elements equal to x_1^2 or x_2^2 but these do not make the data linearly separable.

Marking scheme

2 marks.

e. Apply the sequential multiclass perceptron learning algorithm to find the parameters for three linear discriminant functions that will correctly classify the data in the expanded feature space defined in answer to sub-question 2.d. Assume that the initial parameters of these three discriminant functions, in augmented notation, are:
$\mathbf{a}_1^T = (1, 0.5, 0.5, -0.75)$, $\mathbf{a}_2^T = (-1, 2, 2, 1)$, $\mathbf{a}_3^T = (2, -1, -1, 1)$.
Use a learning rate of 1. If more than one discriminant function produces the maximum output, choose the one with the lowest index (*i.e.*, the one that represents the smallest class label).

[7 marks]

Answer

Sequential Multiclass Perceptron Learning algorithm:

- Initialise $\mathbf{a}_j$ for each class.
- For each exemplar $(\mathbf{y}_k, \omega_k)$ in turn
  - Find predicted class $j = \arg\max_{j'}(\mathbf{a}_{j'}^T \mathbf{y}_k)$
  - If predicted class is not true class (*i.e.*, $j \neq \omega_k$), update weights:

    $\mathbf{a}_{\omega_k} = \mathbf{a}_{\omega_k} + \eta \mathbf{y}_k$

    $\mathbf{a}_j = \mathbf{a}_j - \eta \mathbf{y}_k$
- repeat until weights stop changing

current parameters			$\mathbf{y}^T$	$g_i(\mathbf{x}) = \mathbf{a}_i^T \mathbf{y}$			ω
$\mathbf{a}_1^T$	$\mathbf{a}_2^T$	$\mathbf{a}_3^T$		g_1	g_2	g_3	
$(1, 0.5, 0.5, -0.75)$	$(-1, 2, 2, 1)$	$(2, -1, -1, 1)$	$(0, 1, 0)$	1.5	1	1	1
$(1, 0.5, 0.5, -0.75)$	$(-1, 2, 2, 1)$	$(2, -1, -1, 1)$	$(1, 0, 0)$	1.5	1	1	1
$(1, 0.5, 0.5, -0.75)$	$(-1, 2, 2, 1)$	$(2, -1, -1, 1)$	$(0.5, 0.5, 0.25)$	1.3125	1.25	1.25	2
$(0, 0, 0, -1,)$	$(0, 2.5, 2.5, 1.25)$	$(2, -1, -1, 1)$	$(1, 1, 1)$	-1	6.25	1	2
$(0, 0, 0, -1,)$	$(0, 2.5, 2.5, 1.25)$	$(2, -1, -1, 1)$	$(0, 0, 0)$	0	0	2	3
$(0, 0, 0, -1,)$	$(0, 2.5, 2.5, 1.25)$	$(2, -1, -1, 1)$	$(0, 1, 0)$	0	2.5	1	1
$(1, 0, 1, -1,)$	$(-1, 2.5, 1.5, 1.25)$	$(2, -1, -1, 1)$	$(1, 0, 0)$	1	1.5	1	1
$(2, 1, 1, -1,)$	$(-2, 1.5, 1.5, 1.25)$	$(2, -1, -1, 1)$	$(0.5, 0.5, 0.25)$	2.75	-0.1875	1.25	2
$(1, 0.5, 0.5, -1.25)$	$(-1, 2, 2, 1.5)$	$(2, -1, -1, 1)$	$(1, 1, 1)$	0.75	4.5	1	2
$(1, 0.5, 0.5, -1.25)$	$(-1, 2, 2, 1.5)$	$(2, -1, -1, 1)$	$(0, 0, 0)$	1	-1	2	3
$(1, 0.5, 0.5, -1.25)$	$(-1, 2, 2, 1.5)$	$(2, -1, -1, 1)$	$(0, 1, 0)$	1.5	1	1	1
$(1, 0.5, 0.5, -1.25)$	$(-1, 2, 2, 1.5)$	$(2, -1, -1, 1)$	$(1, 0, 0)$	1.5	1	1	1
$(1, 0.5, 0.5, -1.25)$	$(-1, 2, 2, 1.5)$	$(2, -1, -1, 1)$	$(0.5, 0.5, 0.25)$	1.1875	1.375	1.25	2

Learning has converged, so required parameters are $\mathbf{a}_1^T = (1, 0.5, 0.5, -1.25)$, $\mathbf{a}_2^T = (-1, 2, 2, 1.5)$, $\mathbf{a}_3^T = (2, -1, -1, 1)$.

Marking scheme

4 marks for correct method, 3 marks for correct application.

f. Write pseudo-code for the sequential Widrow-Hoff learning algorithm when applied to learning a discriminant function for a two-class problem.

[5 marks]

Answer

- Initialise $\mathbf{a}$ to an arbitrary solution.
- For each exemplar $(\mathbf{y}_k, \omega_k)$ in turn
  - Update solution: $\mathbf{a} \leftarrow \mathbf{a} + \eta \left(b_k - \mathbf{a}^T \mathbf{y}_k\right) \mathbf{y}_k$
- repeat until $\sum_k | \left(b_k - \mathbf{a}^T \mathbf{y}_k\right) \mathbf{y}_k| < \theta$

Marking scheme

5 marks.

3.

a. A linear threshold neuron has a threshold $\theta = -2$ and weights $w_1 = -1$ and $w_2 = 3$. The activation function is the Heaviside function. Calculate the output of this neuron when the input is $x = (2, 0.5)^T$.

[2 marks]

Answer

Using Augmented notation, $y = H(wx)$ where $w = [-\theta, w_1, w_2]$, and $x = [1, x_1, x_2]^T$. Hence, $y = H((2, -1, 3)(1, 2, 0.5)^T) = H(1.5) = 1$

Marking scheme

2 marks

b. Write pseudo-code for updating the parameters of a single linear threshold neuron using the *sequential* delta learning algorithm.

[4 marks]

Answer

1. Initialise w
2. For each exemplar, x:
3. Calculate the neuron's output: $y = H(wx)$ where H is the heaviside function.
4. Update weights such that: $w \leftarrow w + \eta(t - y)x^T$, where t is the desired output for the current exemplar.
3. Repeat from step 2 until w stops changing.

Marking scheme

4 marks

c. Briefly explain what is meant by a competitive neural network, and describe two methods for implementing the competition.

[4 marks]

Answer

In a competitive neural network the neurons compete for the right to respond to the input (in other words the activity of some neurons is suppressed by other neurons). This can be achieved using inhibitory lateral weights between the neurons, or by using a selection process that chooses the "winning" neuron(s).

Marking scheme

2 marks for correctly defining a competitive neural network, 1 mark for each method of implementing competition.

d. A negative feedback network has three inputs and two output neurons, that are connected with weights $\mathbf{W} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$. Determine the activation of the output neurons after 4 iterations when the input is $\mathbf{x} = (1, 1, 0)^T$, assuming that the output neurons are updated using parameter $\alpha = 0.25$, and the activations of the output neurons are initialised to be all zero.

[6 marks]

Answer

The activation of a negative feedback network is determined by iteratively evaluating the following equations:

$$\mathbf{e} = \mathbf{x} - \mathbf{W}^T\mathbf{y}$$

$$\mathbf{y} \leftarrow \mathbf{y} + \alpha\mathbf{W}\mathbf{e}$$

it	$\mathbf{e}^T$	$(\mathbf{We})^T$	$\mathbf{y}^T$	$(\mathbf{W}^T\mathbf{y})^T$
1	$(1, 1, 0)$	$(2, 2)$	$(0.5, 0.5)$	$(1, 1, 0.5)$
2	$(0, 0, -0.5)$	$(0, -0.5)$	$(0.5, 0.375)$	$(0.875, 0.875, 0.375)$
3	$(0.125, 0.125, -0.375)$	$(0.25, -0.125)$	$(0.563, 0.348)$	$(0.906, 0.906, 0.344)$
4	$(0.094, 0.094, -0.344)$	$(0.188, -0.156)$	$(0.609, 0.305)$	$(0.914, 0.914, 0.305)$

So output is $\begin{pmatrix} 0.609 \\ 0.305 \end{pmatrix}$

Marking scheme

4 marks for correct equations, 2 marks for correct application.

e. Write down the objective function used in sparse coding, and explain the role played by each of the components of the function in representing data efficiently.

[4 marks]

Answer

The sparse coding objective function is given as:

$$min_y \left(||\mathbf{x} - \mathbf{V^T y}||_2 + \lambda ||\mathbf{y}||_0 \right)$$

In this, the first component of the function is the L2-norm of the error between a data point $\mathbf{x}$, and its reconstruction using the dictionary (or weights) $\mathbf{V^T}$ and sparse coefficients $\mathbf{y}$. This component computes how well the weights represent the original data. The second component of this function shows the sparsity of the coefficients vector, *i.e.*, the length of the vector, with the objective function preferring small length vectors.

Marking scheme

2 marks for correct objective function, 1 mark for explanation of each component.

f. Given a dictionary, $\mathbf{V}^T$, what is the best sparse code for the signal $\mathbf{x}$ out of the following two alternatives:
i) $\mathbf{y}^T = (1, 2, 0, -1)$
ii) $\mathbf{y}^T = (0, 0.5, 1, 0)$
Where $\mathbf{V}^T = \begin{pmatrix} 1 & 1 & 2 & 1 \\ -4 & 3 & 2 & -1 \end{pmatrix}$, and $\mathbf{x} = (2, 3)^T$. Assume that sparsity is measured as the count of elements that are non-zero. Use a trade-off parameter of $\lambda = 1$ in order to calculate the costs.

[5 marks]

Answer

Computing the cost $||\mathbf{x} - \mathbf{V^T y}||_2 + \lambda ||\mathbf{y}||_0$, for each set of coefficients gives the following results:

i)

1. The reconstruction is

$$\mathbf{V^T y} = \begin{pmatrix} 1 & 1 & 2 & 1 \\ -4 & 3 & 2 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

2. The L2-norm of the reconstruction error is $||(2,3)^T - (2,3)^T||_2 = ||(0,0)^T|| = \sqrt{0^2+0^2} = 0$
3. The sparsity is $||\mathbf{y}||_0 = 3$
4. The cost is $0 + 1 \times 3 = 3$;

ii)

1. The reconstruction is

$$\mathbf{V^T y} = \begin{pmatrix} 1 & 1 & 2 & 1 \\ -4 & 3 & 2 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 3.5 \end{pmatrix}$$

2. The L2-norm of the reconstruction error is $||(2,3)^T - (2.5,3.5)^T||_2 = ||(-0.5,-0.5)^T|| = \sqrt{0.5^2+0.5^2} = 0.7071$
3. The sparsity is $||\mathbf{y}||_0 = 2$
4. The cost is $0.7071 + 1 \times 2 = 2.7071$;

The cost is lower for the second set of coefficients, so $\mathbf{y}^T = (0, 0.5, 1, 0)$ is the better sparse code.

Marking scheme

5 marks.

4.

a. Figure 1 shows a decision tree of a binary-class classification problem where the classes are denotes as ω_1 for class 1 and ω_2 for class 2. The sample representation is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

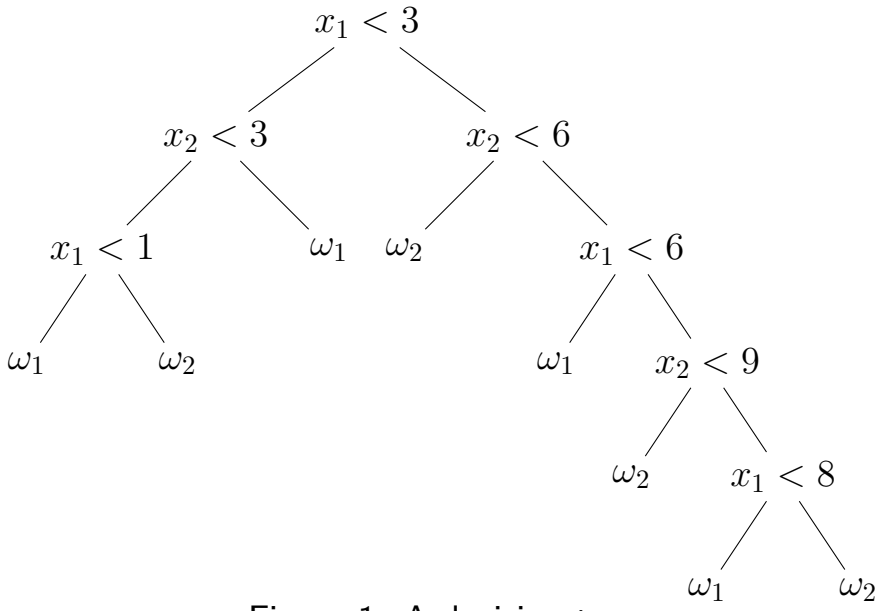
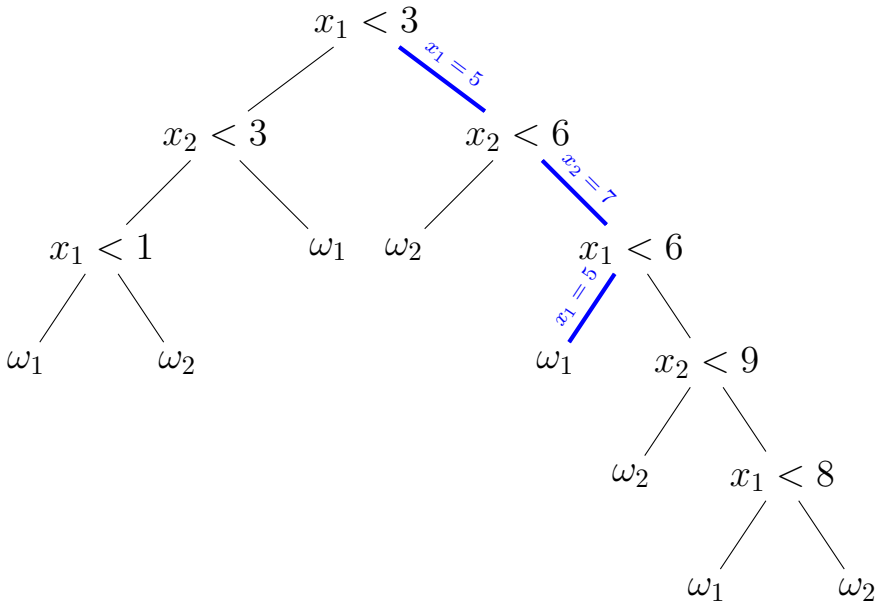


Figure 1: A decision tree.

i. Given a sample $\mathbf{x} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$, predict its class.

[3 marks]

Answer

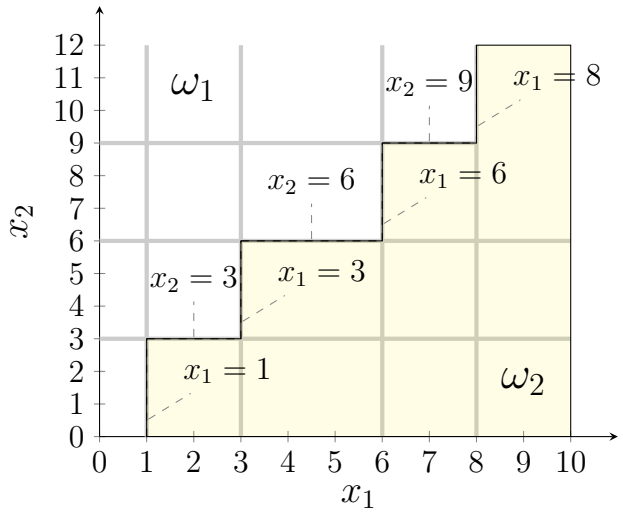


The predicted class is ω_1. 3 marks.

ii. Sketch the decision boundary given by the decision tree where the values of x_1 and x_2 in the axes should be shown.

[7 marks]

Answer



Marking scheme

1 mark for each side.

b. Figure 2 shows a 3-layer partially connected feed-forward neural network. Linear function is used as the activation function in all the input, hidden and output units. The training dataset is given in Table 3 where the sample representation is in the form of $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, output representation is in the form of $\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ and its target representation is in the form of $\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$. *Batch Backpropagation* is employed to perform training using the cost $J = \sum_{p=1}^{N} J_p$ where $J_p = \frac{1}{2}\|\mathbf{t}_p - \mathbf{z}_p\|^2$; N denotes the number of training samples; $\mathbf{z}_p$ is the network output due to the p-th sample and $\mathbf{t}_p$ denotes its target. Assume that only w_{11} is updated. Given the learning rate $\eta = 0.1$, determine the updated value of w_{11} at the next iteration.

[15 marks]

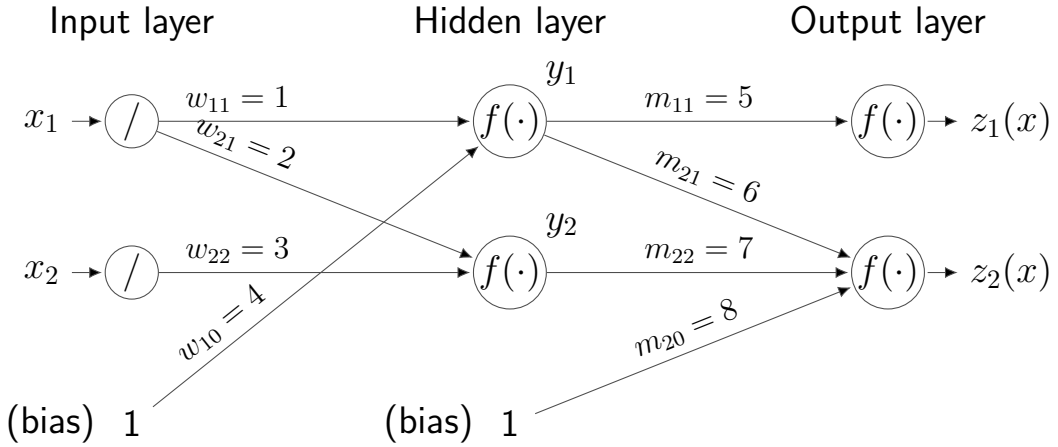


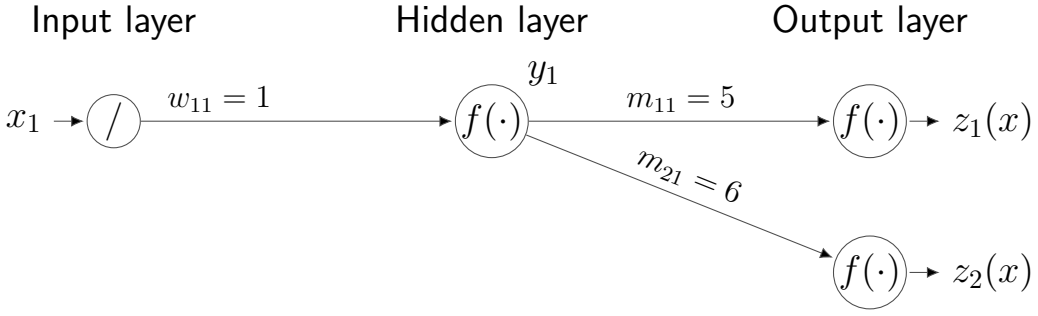
Figure 2: A diagram of neural network.

$\mathbf{x}$	$\mathbf{t}$ (target output)
$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} -6 \\ 8 \end{bmatrix}$
$\begin{bmatrix} -3 \\ 4 \end{bmatrix}$	$\begin{bmatrix} -2 \\ 4 \end{bmatrix}$

Table 3: Input samples $\mathbf{x}$ and its target output $\mathbf{t}$.

Answer

When w_{11} is updated, only the paths in the following figure will be affected.



Denote the net function in the hidden layer as net_{y_1} associated with y_1; output layer as net_{z_1} associated with z_1 and net_{z_2} associated with z_12, where the activation function $f(\cdot)$ is a linear function.

$y_1 = f(net_{y_1}) = net_{y_1} = w_{11}x_1 + w_{10}$

$z_1 = f(net_{z_1}) = net_{z_1} = m_{11}y_1$

$z_2 = f(net_{z_2}) = net_{z_2} = m_{21}y_1 + m_{22}y_2 + m_{20}$

Cost: $J = J_1 + J_2$, $J_1 = \frac{1}{2}((t_1(1) - z_1(1))^2 + (t_2(1) - z_2(1))^2)$, $J_2 = \frac{1}{2}((t_1(2) - z_1(2))^2 + (t_2(2) - z_2(2))^2)$; $\frac{\partial J}{\partial w_{11}} = \frac{\partial J_1}{\partial w_{11}} + \frac{\partial J_2}{\partial w_{11}}$.

$$\begin{aligned}\frac{\partial J_p}{\partial w_{11}} &= \frac{\partial J_p}{\partial y_1}\frac{\partial y_1}{\partial net_{y_1}}\frac{net_{y_1}}{w_{11}} \\ &= \sum_{k=1}^{2}\Big(\frac{\partial J_p}{\partial z_k}\frac{\partial z_k}{\partial y_1}\Big)\frac{\partial y_1}{\partial net_{y_1}}\frac{net_{y_1}}{w_{11}} \\ &= -\sum_{k=1}^{2}\Big((t_k(p)-z_k(p))\frac{\partial z_k}{\partial y_1}\Big)\frac{\partial y_1}{\partial net_{y_1}}\frac{net_{y_1}}{w_{11}} \\ &= -\Big((t_1(p)-z_1(p))m_{11}+(t_2(p)-z_2(p))m_{21}\Big)x_1(p); p=1,2.\end{aligned}$$

5 marks

Network Outputs:

$$\begin{aligned}y_1 &= w_{11}x_1 + w_{10} \\ &= x_1 + 4 \\ y_2 &= w_{21}x_1 + w_{22}x_2 \\ &= 2x_1 + 3x_2 \\ z_1 &= m_{11}y_1 \\ &= 5(x_1+4) \\ &= 5x_1 + 20 \\ z_2 &= m_{21}y_1 + m_{22}y_2 + m_{20} \\ &= 6y_1 + 7y_2 + 8 = 6(x_1+4) + 7(2x_1+3x_2) + 8 \\ &= 20x_1 + 21x_2 + 32\end{aligned}$$

Considering the first sample ($p=1$): $\mathbf{x}_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and its target $\mathbf{t}_1 = \begin{bmatrix} -6 \\ 8 \end{bmatrix}$, we have $\mathbf{z}_1 = \begin{bmatrix} 5x_1 + 20 = 5 \times -1 + 20 = 15 \\ 20x_1 + 21x_2 + 32 = 20 \times -1 + 21 \times 2 + 32 = 54 \end{bmatrix}$.

$$\begin{aligned}\frac{\partial J_1}{\partial w_{11}} &= -\Big((t_1(p) - z_1(p))m_{11} + (t_2(p) - z_2(p))m_{21}\Big)x_1(p) \\ &= -\Big((-6 - z_1(1)) \times 5 + (8 - z_2(1)) \times 6\Big) \times -1 \\ &= -\Big((-6 - 15) \times 5 + (8 - 54) \times 6\Big) \times -1 \\ &= -\Big(- 21 \times 5 + (-46) \times 6\Big) \times -1 \\ &= -381\end{aligned}$$

4 marks

Considering the second sample ($p = 2$): $\mathbf{x}_2 = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$ and its target $\mathbf{t}_2 = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$, we have $\mathbf{z}_2 = \begin{bmatrix} 5x_1 + 20 = 5 \times -3 + 20 = 5 \\ 20x_1 + 21x_2 + 32 = 20 \times -3 + 21 \times 4 + 32 = 56 \end{bmatrix}$

$$\begin{aligned}\frac{\partial J_2}{\partial w_{11}} &= -\Big((t_1(2) - z_1(2))m_{11} + (t_2(2) - z_2(2))m_{21}\Big)x_1(2) \\ &= -\Big((-2 - z_1(2)) \times 5 + (4 - z_2(2)) \times 6\Big) \times -3 \\ &= -\Big((-2 - 5) \times 5 + (4 - 56) \times 6\Big) \times -3 \\ &= -\Big(- 7 \times 5 - 52 \times 6\Big) \times -3 \\ &= -1041\end{aligned}$$

4 marks

Update of w_{11}: learning rate $\eta = 0.1$

$$
\begin{aligned}
m_{11} &\leftarrow m_{11} - \eta \frac{\partial J}{\partial w_{11}} \\
&= 1 - 0.1 \times \Big(\frac{\partial J_1}{\partial w_{11}} + \frac{\partial J_2}{\partial w_{11}} \Big) \\
&= 1 - 0.1 \times (-381 - 1041) \\
&= 143.2000
\end{aligned}
$$

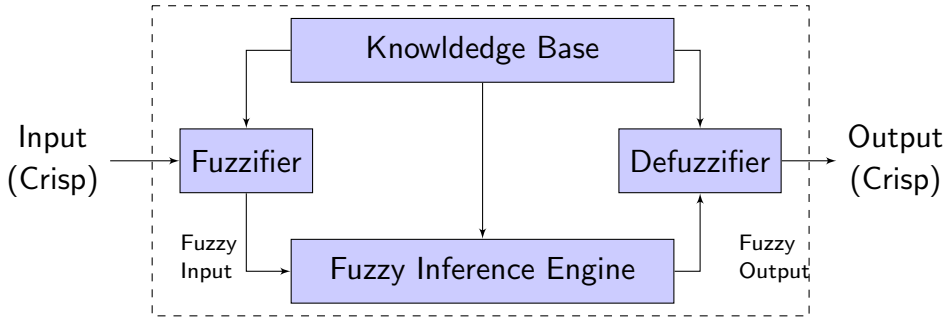
2 marks

5.

a. Draw the diagram of *Fuzzy Inference System* with labels. Briefly describe each component.

[8 marks]

Answer



4 marks

- **Fuzzifiers:** It maps the crisp (real-valued) input into a fuzzy set defined in the universe of discourse $\mathbf{X}$ characterised by membership functions. This process is called *fuzzification*. 1 mark

- **Knowledge Base:** It is a database consisting of linguistic rules in IF-THEN format. 1 mark

- **Fuzzy Inference Engine:** Using the If-Then rules in *Knowledge base*, it performs reasoning by producing a fuzzy output according to the fuzzy input given by the *fuzzifier*. 1 mark

- **Defuzzifiers:** It converts the fuzzy output given by the *fuzzy inference engine* to produce a crisp (real-valued) output. This process is called *defuzzifiaction*. 1 mark

b. A fuzzy inference system using *max-product inference method* has 2 rules as shown below:

Rule 1: IF x is *Zero* THEN y is *Big*
Rule 2: IF x is *Nonzero* THEN y is *Small*

where the fuzzy sets are given as:

$$Zero = \{\frac{0.1}{-5} + \frac{0.3}{-3} + \frac{0.7}{-1} + \frac{1}{0} + \frac{0.7}{1} + \frac{0.3}{3} + \frac{0.1}{5}\},$$

$$Nonzero = \{\frac{0.9}{-5} + \frac{0.7}{-3} + \frac{0.3}{-1} + \frac{0}{0} + \frac{0.3}{1} + \frac{0.7}{3} + \frac{0.9}{5}\},$$

$$Big = \{\frac{1}{0} + \frac{1}{1} + \frac{0.9}{2} + \frac{0.5}{3} + \frac{0.2}{4} + \frac{0}{5}\},$$

$$Small = \{\frac{0}{0} + \frac{0.2}{1} + \frac{0.5}{2} + \frac{0.9}{3} + \frac{1}{4} + \frac{1}{5}\}.$$

i. Identify the linguistic variables and linguistic terms.

[4 marks]

Answer

Linguistic variables: x, y 2 mark

Linguistic terms: *Zero*, *Non-zero*, *Small* and *Big* 2 mark

ii. Considering input $x = 1$, what are the firing strengths for rules 1 and 2? Sketch the inferred output membership function for each rule. Sketch the overall inferred output membership function. The grades of membership must be shown.

[8 marks]

Answer

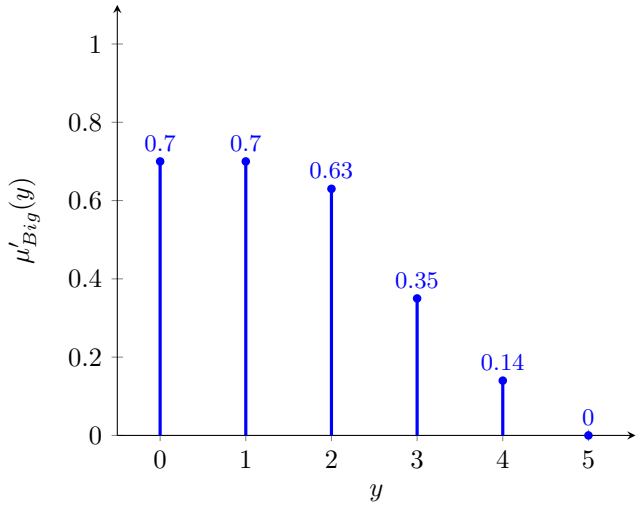
Firing strengths:

Rule 1: 0.7 1 mark

Rule 2: 0.3 1 mark

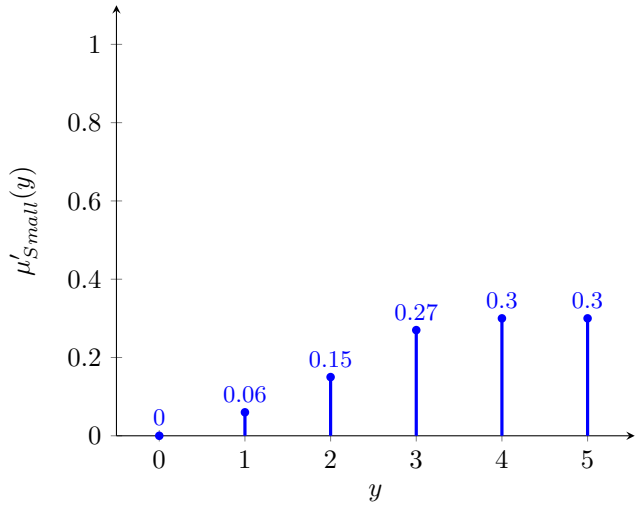
Output Membership functions:

Rule 1:



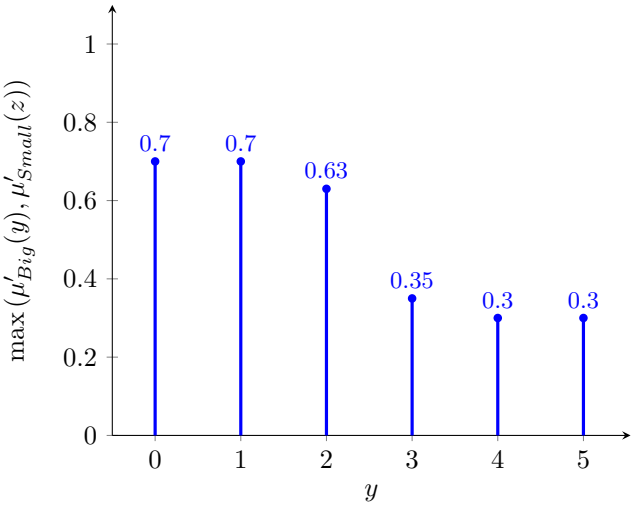
2 marks

Rule 2:



2 marks

Overall inferred output membership functions:



2 marks

iii. Based on the overall inferred output membership function obtained in Q5.b.ii, compute the defuzzified output using the *Centroid* method.

[5 marks]

Answer

Centroid method:

$$z^* = \frac{\sum_{z_i \in Z} \mu_{\tilde{C}}(z_i) z_i}{\sum_{z_i \in Z} \mu_{\tilde{C}}(z_i)}$$

$$\frac{0.7 \times 0 + 0.7 \times 1 + 0.63 \times 2 + 0.35 \times 3 + 0.3 \times 4 + 0.3 \times 5}{0.7 + 0.7 + 0.63 + 0.35 + 0.3 + 0.3}$$

$$= 1.9161$$

5 marks

6.

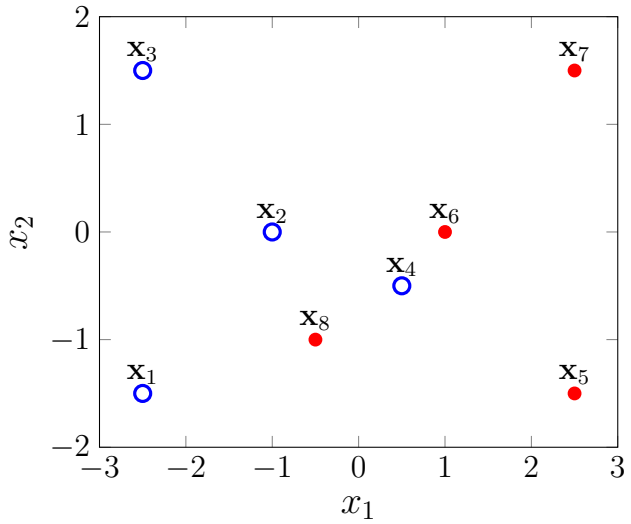
a. In a classification problem, the sample representation is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

The following dataset is considered: Class 1: $\mathbf{x}_1 = \begin{bmatrix} -2.5 \\ -1.5 \end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$, $\mathbf{x}_3 = \begin{bmatrix} -2.5 \\ 1.5 \end{bmatrix}$, $\mathbf{x}_4 = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$; Class 2: $\mathbf{x}_5 = \begin{bmatrix} 2.5 \\ -1.5 \end{bmatrix}$, $\mathbf{x}_6 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{x}_7 = \begin{bmatrix} 2.5 \\ 1.5 \end{bmatrix}$, $\mathbf{x}_8 = \begin{bmatrix} -0.5 \\ -1 \end{bmatrix}$, where class 1 is denoted by class label "-1" (i.e., $y_1 = y_2 = y_3 = y_4 = -1$) and class 2 is denoted by class label "$+1$" (i.e., $y_5 = y_6 = y_7 = y_8 = +1$). A support vector machine (SVM) *hard* classifier with hyperplane $x_1 = 0$ was designed to classify the dataset.

i. Is the datasets linearly separable? Explain your answer.

[5 marks]

Answer



3 marks

From the plot of the dataset above, it cannot find a linear line which can separate perfectly the two classes. The dataset is thus not linearly separable. 2 marks

ii. Determine the margin of the SVM classifier $f(\mathbf{x})$.

[6 marks]

Answer

The hyperplane is $f(\mathbf{x}) = x_1 = 0$ and the hard classifier is sgn$(f(\mathbf{x})) =$ sgn(x_1).

2 marks

$\mathbf{x}_2$ from class 1 and $\mathbf{x}_6$ from class 2 are the support vector as they give $y_i f(\mathbf{x}_i) = 1$. 2 marks

The perpendicular distances from $\mathbf{x}_2$ to $f(\mathbf{x}) = x_1 = 0$ and from $\mathbf{x}_6$ to $f(\mathbf{x}) = x_1 = 0$ are both 1. So the margin is 2. 2 marks

iii. In the context of SVM, the constraint for samples is given as

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i$$

where $\mathbf{w}$ is the weight vector of hyperplane, w_0 is the bias of hyperplane and ξ_i is known as *slack variable*. Identify the samples with non-zero ξ_i and determine their values.

[6 marks]

Answer

Input all samples ($\mathbf{x}_1$ to $\mathbf{x}_8$) one by one to the classifier sgn(x_1) with the hyperplane $f(\mathbf{x}) = x_1 = 0$, it is found that there are two samples which are wrongly predicted, i.e., $\{\mathbf{x}_4,\ y_4 = -1\}$ and $\{\mathbf{x}_8,\ y_8 = +1\}$.

For $\{\mathbf{x}_4,\ y_4 = -1\}$, $f(\mathbf{x}_4) = 0.5 > 0$ predicts the sample class being "$+1$".

For $\{\mathbf{x}_8,\ y_4 = -1\}$, $f(\mathbf{x}_4) = -0.5 < 0$ predicts the sample class being "-1".

Only the wrongly classified samples will have non-zero value of ξ_i, i.e., $\mathbf{x}_4 = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$ and $\mathbf{x}_8 = \begin{bmatrix} -0.5 \\ -1 \end{bmatrix}$. 2 marks

With $\{\mathbf{x}_4,\ y_4 = -1\}$ and $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1 - \xi_i$, we have $-x_1 = 1 - \xi_i$ which gives $\xi_i = 1 + x_1 = 1 + 0.5 = 1.5$. 2 marks

With $\{\mathbf{x}_8,\ y_8 = +1\}$ and $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1 - \xi_i$, we have $x_1 = 1 - \xi_i$ which gives $\xi_i = 1 - x_1 = 1 - (-0.5) = 1.5$. 2 marks

iv. In the context of SVM, the primal problem can be described as

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 + \xi_i))$$

where $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_N \end{bmatrix}$; $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_N \end{bmatrix}$ are Lagrange multipliers; $\mu_i \geq 0$, $\lambda_i \geq 0$ for all $i = 1, 2, \cdots, N$; N denotes the number of training samples. What are the values of λ_i corresponding to the samples $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$. Explain your answer.

[8 marks]

Answer

The weights of hyperplane are contributed by the support vectors, i.e., samples with $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$ and samples with non-zero ξ_i. These samples will have non-zero value of λ_i. 2 marks

$\mathbf{x}_2$ and $\mathbf{x}_6$ give $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$ and thus they have non-zero λ_i. $\mathbf{x}_4$ and $\mathbf{x}_8$ have non-zero ξ_i and thus they have non-zero λ_i.

2 marks

The rest samples $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$ will have $\lambda_i = 0$. 4 marks