# 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**

**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 $\omega_j$ represents category $j$, and $x$ is a sample.

[3 marks]

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

[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]

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

[5 marks]

Table 1, below, shows samples from two classes:

| class | $\mathbf{x}^T$ |
|---|---|
| $\omega_1$ | (5,1) |
| $\omega_1$ | (5,-1) |
| $\omega_2$ | (3,0) |
| $\omega_2$ | (2,1) |
| $\omega_2$ | (4,2) |

Table 1

QUESTION 1 CONTINUES ON NEXT PAGE

**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]

**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]

**g.** Using Table 1 estimate the prior probabilities of each class, *i.e.*, calculate P($\omega_1$) and P($\omega_2$).

[2 marks]

**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]

**2.**

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

i. Feature Space

ii. Decision Boundary

[4 marks]

**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]

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]

QUESTION 2 CONTINUES ON NEXT PAGE

**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]

**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]

**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]

**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]

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

[4 marks]

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

[4 marks]

**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]

**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]

QUESTION 3 CONTINUES ON NEXT PAGE

**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]

**4.**

**a.** Figure 1 shows a decision tree of a binary-class classification problem where the classes are denotes as $\omega_1$ for class 1 and $\omega_2$ for class 2. The sample representation is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

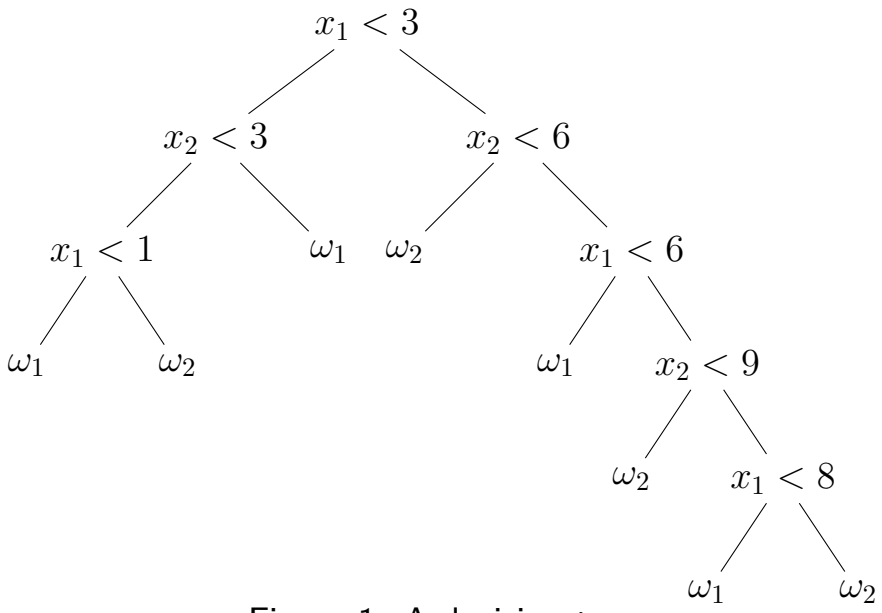


Figure 1: A decision tree.

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

[3 marks]

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]

**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]

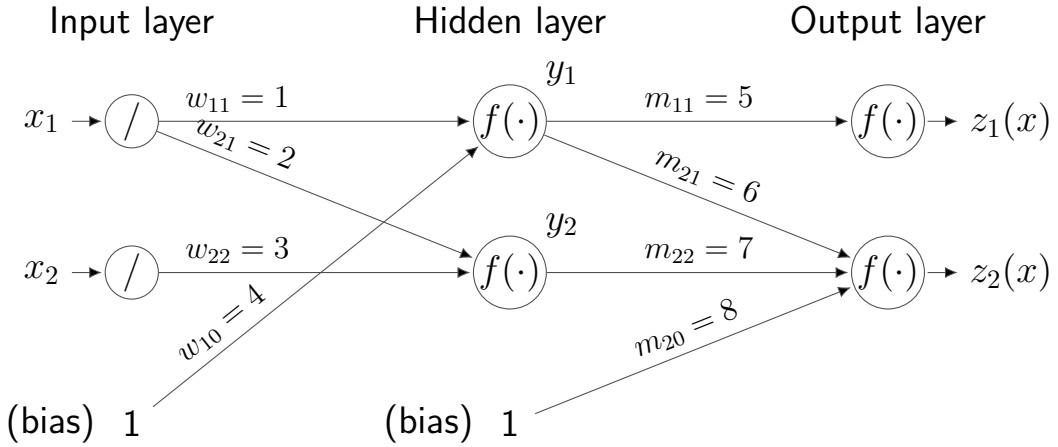


Figure 2: A diagram of neural network.

| $\mathbf{x}$ | $\mathbf{t}$ (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 $\mathbf{t}$.

**5.**

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

[8 marks]

**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]

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]

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

[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]

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

[6 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 $\xi_i$ is known as *slack variable*. Identify the samples with non-zero $\xi_i$ and determine their values.

[6 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 $\lambda_i$ corresponding to the samples $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_5$ and $\mathbf{x}_7$. Explain your answer.

[8 marks]