SECTION A

a) Explain the relevance of optimization in machine learning. Illustrate your answer with two examples. (**4**)

b) What is the Hessian matrix of a cost function? Explain why it is important in optimization algorithms. (**4**)

c) Draw a carefully labelled diagram of a standard 3-layer multilayer Perceptron (MLP) for classifying inputs $x_1, \ldots, x_d$ to a single target value t. (**4**)

d) Write down an expression for the discriminant function for the MLP in part (c). (**2**)

e) In neural network learning, explain the difference between a training set and a test set. Explain why performance of a network should be quoted using the test set rather than the training set. (**2**)

f) Explain the goal of Principal Component Analsysis, identifying the underlying optimization problem. (**4**)

SECTION B

Question B1

Consider the following data sequence:

$$\{x(1), \dots, x(6)\} = \{1.0, \ -0.7, \ 0.9, \ 0.2, \ 0.7 \ -0.3\}$$

a) Describe the key steps in the k-means clustering algorithm indicating how it can be used to group $\{x(n)\}$ into two classes. **(4)**

b) Show that the means $\mu_1 = 0.7$ and $\mu_2 = -0.5$ form a possible k-means solution for the above data. **(4)**

c) Write down the class sequence $\{v(1), \dots, v(6)\}$ generated by the solution for part (b). **(1)**

The data is also known to exhibit dynamic behaviour such that the k-means solution in part (c) can be modelled as the visible sequence of a two state Hidden Markov Model (HMM). Suppose that the transition probabilities are given by $a_{12} = a_{21} = 3/4$ and that the observation probabilities are $b_{11} = b_{22} = 2/3$. Assume that the initial state at $t = 1$ is known to be $\omega(1) = \omega_1$.

d) Calculate the number of possible state sequences $\{\omega(1), \dots, \omega(6)\}$ that may have occurred. **(2)**

e) Explain why an exhaustive search of all state sequences is not necessary in order to find the most probable state sequence. **(3)**

f) Calculate the most probable state sequence given the observed visible symbol sequence in part (c). **(6)**

Question B2

a) Explain what is meant by the term *Maximum Likelihood* and discuss its role in probabilistic data models for both *supervised* and *unsupervised* scenarios. **(5)**

b) Let x be drawn from a uniform density:

$$p(x|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta, \\ 0 & \text{otherwise} \end{cases}$$

(i) Write down an expression for the maximum likelihood estimate for the parameter θ given the data, $D = \{x_1, \ldots, x_d\}$ drawn independently according to $p(x|\theta)$.

(ii) Hence, or otherwise, calculate the maximum likelihood estimate for θ given the data: $D = \{0.1, 0.4, 0.2, 0.8, 0.45\}$. **(5)**

c) Let **s** denote an d-dimensional vector of independent random variables, s_i, each with the same probability density function, $p(s_i)$. Write down an expression for the joint probability density function, $p(\mathbf{s})$ for s. **(1)**

d) Let **x** be defined as: $\mathbf{x} = \mathbf{A}\mathbf{s}$, where the matrix **A** is assumed to be square and invertible, $\mathbf{W} = \mathbf{A}^{-1}$ and **s** is assumed to follow the model in part (c). Hence show that the log likelihood for the data vectors $\mathbf{x}(1), \ldots \mathbf{x}(n)$ can be written as:

$$\mathcal{L}(\mathbf{W}) = n \log \det \mathbf{W} + \sum_{i=1}^{n} \sum_{j=1}^{d} \log f([\mathbf{W}\mathbf{x}(i)]_j)$$

where $[\mathbf{W}\mathbf{x}(i)]_j$ denotes the jth element of $\mathbf{W}\mathbf{x}(i)$ and f is a scalar nonlinear function. **(7)**

e) What form does the function f take when the independent components s_i are assumed to have been drawn from the following Laplace distribution:

$$p(s_i) = \frac{1}{2} \exp(-|s_i|).$$

(2)

Question B3

a) Write down the fixed increment single-sample perceptron learning rule for a 2-category classifier. **(2)**

b) Consider the following data points $(x_1, x_2)_i$, taken from two categories, ω_1 and ω_2:

$$\omega_1 : (0,1)(2,0)(1,2)$$

$$\omega_2 : (2,2)(3,4)(1,6)$$

Calculate the weight updates for when one pass of the data is presented to a fixed-increment single-sample perceptron learning algorithm with initial weights $\{w_0, w_1, w_2\} = \{1.0, 2.0, 0.0\}$, and a learning rate of 0.5. Assume that the samples are presented in the order given. **(5)**

c) Is the single-sample fixed-increment perceptron learning algorithm able to correctly classify these samples? Justify your answer. **(2)**

d) Suppose that for the samples in part (b) the class condition probabilities are modelled as Gaussian densities with means μ_1 and μ_2 and common covariance matrix, $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Using the maximum likelihood principle calculate the Bayes decision boundary for these samples. **(6)**

e) Sketch the decision boundary from part (d) along with the samples. Hence determine whether it correctly classifies the data. **(3)**

f) Discuss to what extent a Bayes decision boundary is optimal and how the choice of prior data model affects the performance of the classifier. **(2)**

END OF PAPER