SECTION A (Compulsory Question)

Question 1

(a) Explain and formalize with equations how the assumption of Gaussian-distributed noise contaminating observations leads to the sum-of-squares error function.

(8 marks)

(b) Briefly explain what is meant by the *kernel trick* and how it may be used to obtain a nonlinear classifier from a linear one. Name and provide expressions for at least two (positive definite) kernel functions.

(10 marks)

(c) Explain the difference between *covariance* and *correlation*. The covariance matrix for two variables is given by

$$\mathbf{S} = \begin{bmatrix} 10.2^2 & -45.6 \\ -45.6 & 8.1^2 \end{bmatrix}$$

Calculate the correlation between them and provide comments on the outcome.

Briefly explain why two decorrelated variables need not be *independent.*

(8 marks)

(d) Explain what is meant by the terms *generalisation error* and *over-fitting*. How can *over-fitting* be combatted?

(6 marks)

(e) With the aid of a carefully labelled diagram, explain what is meant by a *convex* function. Why is it desirable for an error function to be convex?

(8 marks)

(Total 40 marks)

<u>SECTION B</u> (Additional Questions)

<u>Question 2</u>

(a) Write the equation and draw a diagram describing the perceptron model. Specify all variables involved in the model (using formal notation) and describe their role.

(10 marks)

(b) Write the equations describing a three-layer multilayer perceptron (MLP), i.e. an MLP with one hidden layer. Specify all variables and functions involved in the model and describe their role. Assume one hidden layer with L neurons and one output neuron.

(10 marks)

(c) Describe the Least-Mean-Square (LMS) algorithm and write the weight update rule (commenting on all variables involved in the expression).

(10 marks)

(Total 30 marks)

Question 3

A start-up wishes to use machine learning to predict the polluting emissions in car exhaust from the characteristics of the car, such as engine size, fuel, driving speed, etc. They have collected a large data set of measured emissions and the corresponding features for each measurement.

(a) In supervised learning, an error function, $E(\mathbf{w})$, may be defined as:

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \log p(t_n \,|\, \mathbf{x}_n; \mathbf{w}).$$

State the meaning of each of the symbols $\mathbf{w}, N, \mathbf{x}_n$ and t_n in this definition and carefully explain how the error function may be derived from the principle of maximum likelihood.

(11 marks)

(b) The start-up uses a radial basis function neural network with a mean squared error function to learn the relationship between features and emissions. Explain what the use of the mean squared error function implies about the noise or measurement errors in the data.

(4 marks)

(c) Explain how the parameters of the radial basis function network may be determined with the mean squared error function.

(7 marks)

(d) In order to cope with occasional outliers the start-up decides to use a *sum of absolute errors* function. Explain why the sum of absolute errors function may help in this context.

(4 marks)

(e) Suggest how the parameters of the radial basis function network could be learned using the sum of absolute errors.

(4 marks)

(Total 30 marks)

Question 4

(a) Explain the difference between a *hierarchical* clustering algorithm and a clustering algorithm that yields a single *partition*. Explain when you might prefer to use a hierarchical algorithm.

(5 marks)

(b) Explain how *agglomerative clustering* is performed.

(6 marks)

(c) Carefully defining the symbols you use, give an expression for the objective function minimised by the k-means algorithm.

(6 marks)

(d) Giving pseudo-code, explain how the k-means clustering is achieved.

(7 marks)

(e) Two users run the k-means algorithm on the same data set, but find that it gives two different partitions. Explain how this may occur and what may be done in practice to cope with this phenomenon.

(6 marks)

(Total 30 marks)