Question 1. (15 marks)

Answer each of these questions in no more than half a page and give illustrative examples where appropriate. Each question is worth 1.5 marks.

(1) Briefly explain the difference between supervised and unsupervised learning methods.

(2) Briefly explain the terms "overfitting" and "generalization".

(3) Categorize the following methods as "parametric" or "non-parametric" : (i) Maximum likelihood, (ii)Linear discriminant analysis, (iii) k-nearest neighbor method.

(4) Explain the difference between the maximum likelihood and Bayesian estimation methods.

(5) Explain the essence of Bayesian classification.

(6) Explain the Naïve Bayes classification.

(7) Explain the idea of a linearly separable data when considering support vector machines.

(8) Explain three methods that can be used to extend two-class classification methods to multi-class versions.

(9) What is the significance of support vectors in a support vector classification method?

(10) Describe two ways of achieving dimensionality reduction.

Question 2. (6 marks)

(1) Suppose a bank classifies customers as either good or bad credit risks. On the basis of extensive historical data, the bank has observed that 1% of good credit risks and 10% of bad credit risks overdraw their account in any given month. A new customer opens a cheque account at this bank. On the basis of a check with a credit bureau, the bank believe that there is a 70% chance that the customer will turn out to be a good credit risk. Suppose that this customer's account is overdrawn in the first month. How does this alter the bank's opinion of this customer's creditworthiness? (3 marks)

(2) The logistic sigmoid function is defined as:

$$\sigma(a) = \frac{1}{1 + exp(-a)}$$

and the 'tanh' function is denoted as:

$$tanh(a) = 2\sigma(2a) - 1$$

Prove that a general linear combination of logistic sigmoid functions of the form

$$y(x, \boldsymbol{w}) = \omega_0 + \sum_{j=1}^{M} \omega_j \sigma\left(\frac{x - \mu_i}{s}\right)$$

is equivalent to a linear combination of 'tanh' functions of the form

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \, tanh(\frac{x - \mu_i}{s})$$

and **find** expressions to relate the new parameters $\{u_1, \ldots, u_M\}$ to the original parameters $\{\omega_1, \ldots, \omega_M\}$.

(3 marks)

Question 3. (10 marks)

This question explores the topic classifier performance evaluation.

(1) Explain the significance of area under the receiver operating characteristics (ROC) curve as it pertains to classifier performance measure. (3 marks)

(2) Suppose that we have a data set and each data item describes objects that can be classified as either belonging to class ω_1 or ω_2. Two classifiers (X and Y) were designed using 500 training samples from the data set. Some samples were set aside for testing and performance measurement. The result of the test is shown in Table 1 below. **Calculate** the respective AUC for the ROC of the two classifiers and **comment** on their relative performance. Also **calculate** the apparent error rates of the two classifiers and comment on the implication. The area, $\tilde{A}$, under the ROC curve (AUC) is given by

$$\tilde{A} = \frac{1}{N_1 N_2}\left\{S_0 - \frac{1}{2}N_1(N_1 + 1)\right\}$$

where N_1 and N_2 are the number of samples from ω_1 and ω_2 respectively. S_0 is the sum of the ranks of class ω_1 test patterns. (3 marks)

(3) The McNemar's or Gillick test can be used to compare two classifiers with respect to the classification errors they make. **Compute** the statistic z and **state** its implication (with respect to the null hypothesis) for the data shown in Table 1. Also **compare** your result with that of part (2). The z statistic is computed using the following formula:

$$z = \frac{|n_{01} - n_{10}| - 1}{\sqrt{n_{01} + n_{10}}}$$

where

n_{01} = number of samples misclassified by X but not by Y

n_{10}= number of samples misclassified by Y but not by X

Table 1: Table of classification test result

Test sample	Classifier X	Classifier Y	True Class	Ranking of Likelihood ratios
x_1	ω_2	ω_1	ω_2	1
x_2	ω_2	ω_2	ω_2	2
x_3	ω_2	ω_2	ω_2	3
x_4	ω_2	ω_2	ω_2	4
x_5	ω_1	ω_2	ω_2	5
x_6	ω_2	ω_1	ω_1	6
x_7	ω_1	ω_1	ω_1	7
x_8	ω_1	ω_1	ω_1	8
x_9	ω_1	ω_1	ω_1	9
x_{10}	ω_1	ω_2	ω_1	10

Question 4. (5 marks)

Four measurements are made on each of a random sample of 500 animals. The first three variables were different linear dimensions, measured in centimeters, while the fourth variable was the weight of the animal measured in grams. The sample covariance was calculated and its four eigenvalues found to be 14.1, 4.5, 1.2, and 0.2. The eigenvectors corresponding to the first and second eigenvalues were:

$$u_1^t = [0.39 \quad 0.42 \quad 0.44 \quad 0.69]$$

$$u_2^t = [0.40 \quad 0.39 \quad 0.42 \quad -0.72]$$

where t denotes transpose.

(1) What is percentage of the variance in the original data accounted for by the first two principal components? Describe the results? (3 marks)

(2) Comment on the use of these two principal values, instead of the original four measurements, as input in a machine learning algorithm. (2 marks)

Question 5. (9 marks)

(1) Explain the purpose of clustering and when it shall be used in machine learning. (2 marks)

(2) Describe the procedure of k-means clustering. (2 marks)

(3) Describe the way(s) you can use to select the k value of k-means clustering. (1 mark)

(4) Why we usually run the k-means clustering multiple times for a given k value? (1 mark)

(5) Explain back propagation algorithm, providing the necessary equations for the weight updates. (3 marks)