Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

END OF SEMESTER EXAMINATION

MASTER OF SCIENCE IN BIOMATHEMATICS

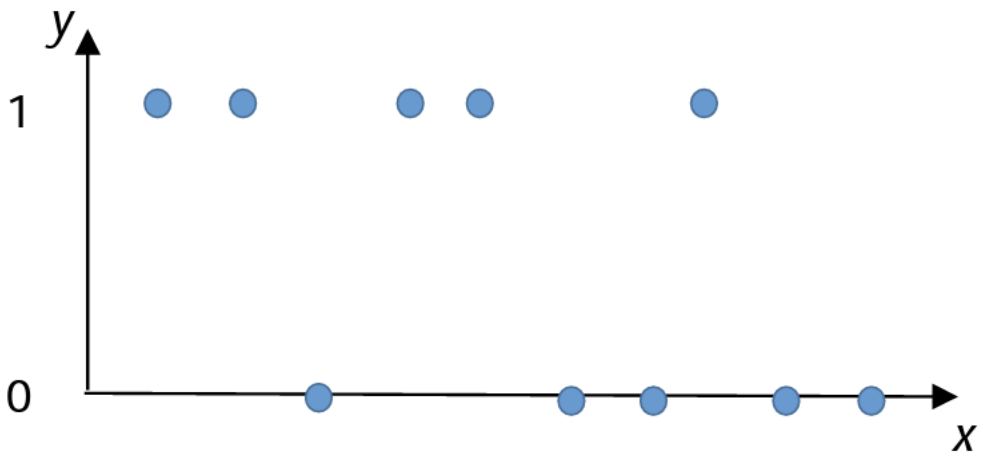
BMA 8104 STATISTICAL MODELLING WITH APPLICATION TO BIOLOGY

Date: 9th December, 2024 **Time: 3 Hours**

Instruction: *Answer Question one and any other two*

Question One (20 Marks)

a. Suppose you are using a Majority Classifier on the following training set containing 10 examples where each example has one real-valued feature, x, and a binary class label, y, with value 0 or 1. Define this Majority Classifier to predict the class label that is in the majority in the training set, regardless of the input value. In case of ties, predict class 1.



i. What is the training set accuracy? (1 mark)
ii. What is the Leave-1-Out Cross-Validation accuracy. (2 marks)
iii. What is the 2-fold Cross-Validation accuracy? (2 marks)

b. Suppose we are given the following dataset, where A, B, C are input binary random variables, and y is a binary output whose value we want to predict.

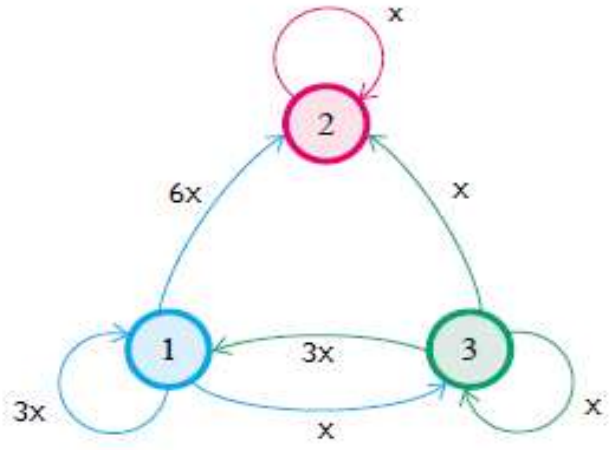
A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

i. How would a naive Bayes classifier predict given this input: A = 0, B = 0, C = 1? Assume that in case of a tie, the classifier always prefers to predict 0 for y. (5 Marks)

ii. Suppose you know for a fact that A, B, and C are independent random variables. In this case, is it possible for any other classifier (e.g., a decision tree or a neural network) to do better than a naive Bayes classifier? (5 Marks)

c. Find the maximum likelihood estimator (MLE) of θ: Xi~Bionmial(n,θ), and we have observed X_1, X_2, X_3, …, X_n. (5 marks)

d. Find the transition matrix from the transition diagram below (3 marks)



e. You are tasked with estimating the integral of the function f(x) = 1/(1+x^2) over the interval [0, 1] using Monte Carlo methods. In addition, use Markov Chain Monte Carlo (MCMC) to sample from a uniform distribution on the interval [0, 1] and estimate the integral. (6 marks)

Question TWO (20 marks)

Data of Clarke et al. (1959) reported excess of gastric ulcers in individuals with blood type O as follows: $n_A = 186$, $n_B = 38$, $n_{AB} = 36$, $n_O = 284$.

a. Write out the likelihood for these data. (7 marks)
b. What are complete data categories? (3 marks)
c. Express the complete data "counts" as a function of allele frequency estimates and the observed data. (5 marks)
d. Apply E-M algorithm to determine the genotype frequencies. (5 marks)

Question THREE (20 marks)

a. In a medical experiment, patients with a chronic condition are asked to say which of two treatments, A, B, they prefer. (You may assume for the purpose of this question that every patient will express a preference one way or the other). Let the population proportion who prefer A be θ. We observe a sample, *n*, patients. Given θ, then responses are independent and the probability that a particular patient prefers A is θ.

Our prior distribution for θ is a beta (*a*, *a*) distribution with a standard deviation of 0.25.

i. Find the value of *a*. [5 marks]
ii. We observe n= 30 patients of whom 21 prefer treatment A. Find the posterior distribution of θ. [5 marks]
iii. Find the posterior mean and standard deviation of θ. [5 marks]
iv. Using R, plot a graph showing both the prior and posterior probability density functions of θ. [5 marks]

Question FOUR (20 marks)

Given a vector data that is drawn from a Poisson distribution with unknown μ.

a. Derive the Poisson likelihood for observation $y_1, y_2, \ldots, y_n$, and hence find the MLE of μ. (7 marks)

b. Write an R function to:

   i. Declare the Poisson log-likelihood function (6 marks)

   ii. Estimate the unknown Poisson parameter using the BFGS (Broyden, Fletcher, Goldfarb, and Shanno) algorithm. (3 marks)

c. Distinguish between Newton Raphson and Quasi Newton Raphson methods. (4 marks)