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: 21st August 2023 **Time: 3 Hours**

Instruction: Answer Question one and any other two

Question One (20 Marks)

a. Clearly define and contrast the following terms as used in practice.

i. Bootstrap and Jackknife methods (2 marks)

ii. Newton Raphson and Quasi Newton Raphson methods (2 marks)

b. Consider the toothache and cavity classification table below.

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

¬ = no cavity.

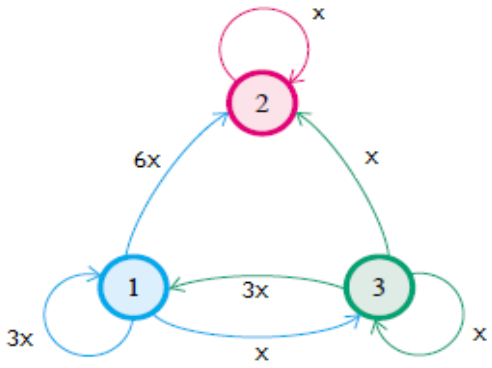
Find

i. P(toothache) (2 marks

ii. P(cavity or toothache) (2 marks)

iii. P(¬cavity | toothache) (3 marks)

c. Find the transition matrix from the transition diagram below (4 marks)



d. Write an R code that will return the minus log likelihood of a normal distribution with mean μ and variance σ. (5 marks)

Question Two (20 marks)

Sociologists have long been interested in *social mobility* – the transition of individuals between social classes defined on the basis of income or occupation. Consider a society with three social classes. Each individual may belong to the lower class (state 1), the middle class (state 2), or the upper class (state 3). Thus, the social class occupied by an individual in generation t may be denoted by $s_t \in \{1,2,3\}$. suppose that *intergenerational mobility* is described by the transition matrix P

$$\begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}$$

a. Determine the transition diagram from this transition matrix (3 marks)

b. Find the transition probabilities after 3 years? (5 marks)

c. Write and run an **R code** to find the long term trend of the transition matrix. (9 marks)

d. State the Markov property and explain why it is important in MCMC methods? (3 marks)

Question Three (20 marks)

a. Consider a vector of binary random variables, $x \in \{0, 1\}$. Assume each variable x is drawn from a Bernoulli(p) distribution, so P(x= 1) = p. Write an expression for P(x|p). (2 marks)

b. Now suppose we have a mixture of K Bernoulli distributions: each vector $x^{(i)}$ is drawn from some vector of Bernoulli random variables with parameters $p^{(k)}$, we will call this Bernoulli($p^{(k)}$). Let $\{p^{(1)}, \ldots, p^{(K)}\} = \mathbf{p}$. Assume a distribution $\pi^{(k)}$ over the selection of which set of Bernoulli parameters $p^{(k)}$ is chosen. Write an expression for $P(x^{(i)} |\mathbf{p}, \pi)$. (3 marks)

c. Finally, suppose we have inputs $X = \{x^{(i)}\}$ i=1...n. Using the above, write an expression for the log likelihood of the data X, log P(X|π, p). (4 marks)

d. Now, we introduce the latent variables for the EM algorithm. Let $z^{(i)} \in \{0, 1\}$ be an indicator vector, such that $z^{(i)}_k = 1$ if $x^{(i)}$ was drawn from a Bernoulli($p^{(k)}$), and 0 otherwise. Let $Z = \{z^{(i)}\}$ i=1...n. What is $P(z^{(i)} |\pi)$? What is $P(x^{(i)}|z(i), p, \pi)$? (5 marks)

e. Using the above two quantities, derive the likelihood of the data and the latent variables, P(Z, X|π, p). (6 marks)

Question Four (20 marks)

a. 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 use a naïve Bayes classifier predict y 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 fact that A, B, C are independent random variables. In this case is it possible for any other classifier (e.g., a decision tree or a neural net) to do better than a naïve Bayes classifier? (The data set is irrelevant for this question) (5 marks)

b. Find the maximum likelihood estimator (MLE) of θ: $X_i \sim \text{Binomial}(m, \theta)$, and we have observed $X_1, X_2, X_3, \ldots, X_n$. Hence, using a gamma prior for θ, determine the posterior distribution. (10 marks)

Question Five (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)