CHUKA UNIVERSITY

UNIVERSITY EXAMINATIONS

FIRST YEAR EXAMINATION FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY IN APPLIED STATISTICS

MATH 943: GENERALIZED LINEAR MODELS AND APPLICATIONS

STREAMS: PhD **TIME: 3 HOURS**

DAY/DATE: TUESDAY 13/8/2019 **2.30 P.M. – 5.30 P.M**

INSTRUCTIONS

- Answer any **THREE** questions.
- Do not write anything on the question paper.

QUESTION ONE (20 Marks)

Suppose X is a binary random variable that takes value 0 with probability p and value 1 with probability 1-p. let $X_1, \ldots, X_n$ be iid samples of X.

(i) Compute a maximum likelihood estimation (MLE) estimate of p. (5 marks)

(ii) Is $\hat{p}$ an unbiased estimate of p? Prove the answer. (5 marks)

(iii) Compute the expected square error of $\hat{p}$ in terms of p. (5 marks)

(iv) Prove that if you know that p lies in the interval $\left[\frac{1}{4}; \frac{3}{4}\right]$ and you are given only $n = 3$ samples of X, then $\hat{p}$ is an inadmissible estimator of p when minimising the expected square error of estimation. (5 marks)

QUESTION TWO (20 Marks)

(a) Suppose that the p.d.f of a random variable X has a 2-component mixture form:

$$p_\alpha(x) = \alpha * p_1(x) + (1 - \alpha) * p_2(x)$$

One component is the density model $p_1(x)$ and the other component is the density model $p_2(x)$. We know both $p_1(x)$ and $p_2(x)$. We do not know α. Given that $\{x_1, x_2, \ldots, x_n\}$ are iid samples from the distribution of X, give EM algorithm for estimating α. Describe the E-Step and M-step clearly in your answer (give clear step by step derivation).

(12 Marks)

(b) Consider independent binary features

$\boldsymbol{X} = (x_1, \ldots, x_d)^t$ $\quad\quad$ $p_i = Pr[x_i = 1 \mid \omega_1)$ $\quad\quad$ $q_i = Pr[x_i = 1 \mid \omega_2)$

and assuming conditional independence

$$P(\boldsymbol{X} \mid \omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}$$

$$P(\boldsymbol{X} \mid \omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

(i) Derive the likelihood ratio. (4 marks)

(ii) Derive the discriminant function. (4 marks)

QUESTION THREE (20 Marks)

The data in Table 1 relate to grain yield (Y), plant height (X_1), and tiller number (X_2) of sorghum.

Table 1: Performance of Sorghum with respect to grain yield (Kg/ha), plant height (cm) and tiller numbers/hill

Grain yield Kg/ha (Y)	Plant height cm (X_1)	Tiller No./hill (X_2)
5755	110.5	14.5
5939	105.4	16.0
6010	118.1	14.6
6545	104.5	18.2
6730	93.6	15.4
6750	84.1	17.6
6899	77.8	17.9
7862	75.6	19.4

(i) Fit a multiple linear regression model of Y on X_i and X_2. (8 Marks)

(ii) Determine variance of β_o, β_1 and β_2. (6 Marks)

(iii) Test if the fitted model is adequate. Take $\alpha = 0.05$. (6 Marks)

QUESTION FOUR (20 Marks)

Linear regression models a real-valued output Y given an input vector X as

$Y/X \sim N(\mu(X), \sigma^2)$

where the mean is a linear function of the input: $\mu(X) = \beta^T X = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Logistic regression models a binary output Y by

$$Y \mid X \sim Bernoulli(\theta(X))$$

where the Bernoulli parameter is related to $\beta^T X$ by the logit transformation

$$logit(\text{ө}(X)) \equiv log\left(\frac{\text{ө}(X)}{1-\text{ө}(x)}\right) = \beta^T X$$

Given data $\{(x_1, y_1), (x_2, y_2), \dots, (x_n x_n)\}$, for each of the two regression models above, show that at the MLEβ

$$\sum_{i=1}^{n} x_i * y_i = \sum_{i=1}^{n} x_i * E[Y \mid X = x_i, \beta = \hat{\beta}]$$
