Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN STATISTICAL SCIENCES

END OF SEMESTER EXAMINATION

STA 8203: PREDICTIVE MODELING AND STATISTICAL LEARNING

DATE: April 28, 2023 Time: 2 Hours

Instructions

1. This examination consists of **FIVE** questions and an appendix to one of the questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks)

a) In statistical learning, distinguish between supervised and unsupervised learning. Give appropriate examples of methods that fall into each of these categories.

(**5** Marks)

b) Exploratory Data Analysis (EDA) is an approach/philosophy employed in data analysis. Explain how this approach is employed, how it differs from classical and Bayesian analysis and hence enumerate the EDA assumptions.

(**6** Marks)

c) Explain the significance of the concept of *Bias-variance trade-off* in a statistical learning algorithm.

(**4** Marks)

d) Suppose that we have a training set consisting of a set of points $x_1, \ldots, x_n$ and real values y_i associated with each point x_i. We assume that there is a function with noise $y = f(x) + \varepsilon$, where the noise, ε, has zero mean and variance σ^2.

For a function $\hat{f}(x)$, that approximates the true function $f(x)$ as well as possible, by means of some learning algorithm, show that $\hat{f}(x)$we can decompose its expected error on an unseen sample as follows:

$$E\left[\left(y - \hat{f}(x)\right)^2\right] = Bias\left[\hat{f}(x)\right]^2 + Var\left[\hat{f}(x)\right] + \sigma^2,$$

where $Bias\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right]$ and $Var\left[\hat{f}(x)\right] = E\left[\hat{f}(x)^2\right] - E\left[\hat{f}(x)\right]^2$.

[**5** Marks]

Question 2 (20 Marks)

a) For each of parts (i) through (iv), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
   i) The sample size n is extremely large, and the number of predictors p is small.
   ii) The number of predictors p is extremely large, and the number of observations n is small.
   iii) The relationship between the predictors and response is highly non-linear.
   iv) The variance of the error terms, that is, $\sigma^2 = \mathrm{Var}()$, is extremely high.

(6 Marks)

b) Mean absolute error (MAE), Mean absolute percentage error (MAPE), and residual mean square error (RMSE) are metrics used to assess the performance of a general linear model.
   i) Provide mathematical expressions for each of these metrics.
   ii) Explain the issue in predictive modeling that each of these metrics evaluates and how they would be used to decide on the best model from a candidate set.
   iii) Explain the difference between MAE and MAPE.

(8 Marks)

c) Consider five regression models, each fit to the same training data.
   - Model 1: `y~1`
   - Model 2: `y~x1`
   - Model 3: `y~x1+x2+x3`
   - Model 4: `y~x1+x2+x3+I(x1^2)+I(x2^2)`
   - Model 5: `y~x1*x2*x3+I(x1^2)+I(x2^2)`

   Assume that the true model is given by

$$Y = 10 + 3x_1 + 5x_2 - 4x_3 + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma^2 = 10).$$

Data on a sample of 100 individuals we split into two parts, with a training data set of size 80 and a testing data set of size 20. The following table provides the within sample and out of sample MAE, MAPE and RMSE for these five models:

	Training data			Testing data		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE
Model 1	166.43	0.66	198.65	199.90	0.80	248.48
Model 2	161.92	0.67	196.09	189.34	0.76	232.64
Model 3	7.18	0.05	9.02	8.03	0.07	10.40
Model 4	6.81	0.05	8.66	7.99	0.06	10.95
Model 5	6.07	0.04	7.51	8.12	0.04	10.69

   i) Comment on the within sample performance of the five models and select the best model.
   ii) Comment on the out of sample performance of the five models and select the best model.

(6 Marks)

Question 3 (20 Marks)

a) Cross-validation is an important tool in predictive modeling. Describe how the following cross-validation techniques work: Leave-One-Out Cross-Validation; and k-Fold Cross-Validation.

(**9** Marks)

b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i) The validation set approach.

(**3** Marks)

ii) LOOCV?

(**3** Marks)

c) Suppose that we use some data mining method to make a prediction for the response **Y** for a particular value of the predictor **X**. Carefully describe how we might estimate the standard deviation of our prediction.

(**5** Marks)

Question 4 (20 Marks)

a) Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

i) Explain what the hat-matrix is.

ii) Explain how standardized residuals are used in regression diagnostics and use a mathematical approach to show that

$$Z = \frac{e_i}{s.e.(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \sim N(0,1)$$

iii) Explain how outliers, high-leverage values, and influential observations on the basis of the hat-matrix.

(**12** Marks)

b) The Dixon and the generalized (extreme Studentized deviate) ESD (Rosner) tests are approaches used in exploratory data analysis. Distinguish between them and explain in (mathematical) detail how each approach works.

(**8** Marks)

Question 5 (20 Marks)

a) Let Y be a Bernoulli random variable with probability density function

$$f(y;\theta) = \theta^y(1-\theta)^{1-y}, y = 0, 1,$$

where θ is the probability of success.

i) Prove that this distribution belongs to the exponential dispersion family.

ii) Based on part (i), what is the canonical link function of this distribution.

(**4** Marks)

b) Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

i) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

ii) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

(**4** Marks)

c) Explain the meaning of the following terminologies as applied to classification problems: sensitivity; specificity; false positive rate; and the false negative rate.

(**2** Marks)

d) A team of researcher at **CDC-Kenya** would like to develop *a predictive model* for HIV serostatus (y) in Kenya using two explanatory variables: age at first sex (x_1); and sex for gifts (x_2). The data used in the study was the most recent Kenya Aids Indicator Survey.

Confusion matrices for the logistic regression models fitted to these data are presented below.

- **Fit 1**: $\text{logit}(y) = \beta_0 + \beta_1 x_1$
- **Fit 2**: $\text{logit}(y) = \beta_0 + \beta_2 x_2$
- **Fit 3**: $\text{logit}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Table 1 Confusion matrices for the three models considered

Fit 1: age1stsex as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	5877	5101
Positive	398	390

Fit 2: sex4gifts_ever as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	376	10602
Positive	37	611

Fit 3: age1stsex and sex4gifts_ever predictors

	Predicted status	
Actual status	Negative	Positive
Negative	6005	497
Positive	408	240

i) From the results present in Fit 1: $\text{logit}(y) = \beta_0 + \beta_1 x_1$

- **Fit 2**: $\text{logit}(y) = \beta_0 + \beta_2 x_2$
- **Fit 3**: $\text{logit}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

ii) Table 1, compare the 3 models. Compare your results.

(**4** Marks)

iii) For the best fitting model, compute the following measures: sensitivity, specificity and the false positive rate.

(**4** Marks)