M.Sc. Examination, 2022
Semester-II
Statistics
Course: MSC-21
(Statistical Inference II)
Time: 3 Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. (a) What do you mean by likelihood ratio ordering? Give an example.
   (b) State and prove Neyman-Pearson lemma 2+8

2. (a) Let $X_1, X_2, \ldots X_n$ be a random sample of size n drawn from $Poisson(\theta)$. Find the most power test of size α for the following testing problem
   $$H: \theta = \theta_0 \quad against\ k: \theta = \theta_1$$
   (b) Find the uniformly most powerful (UMP) test for testing the following hypothesis
   $$H: \theta = \theta_0 \quad against\ k: \theta > \theta_1$$
   on the basis of a random sample of size n drawn from an one parameter exponential distribution.
   5+5

3. (a) Show that UMP test does not always exist.
   (b) Let $X_1 \sim Bin(n_1, \theta_1)$ be independently distributed with $X_2 \sim Bin(n_2, \theta_2)$. Find uniformly most powerful unbiased (UMPU) test for the testing of the following hypothesis
   $$H: \theta_1 = \theta_2\ against\ \ k: \theta_1 > \theta_2$$
   4+6

4. (a) Define Kolmogorov-Smirnov one-sample statistic. Justify this test statistic as "functions of the deviations". When will it be used?
   (b) Show that Kolmogorov-Smirnov one sample test statistic is distribution free.
   5+5

5. (a) Define two-sample linear rank statistics. Find the mean and variance of two-sample general linear rank statistic.
   (b) Define two sample Wilcoxon Rank-Sum test statistic and describe in which context can it be used? Find minimum and maximum of Wilcoxon Rank-Sum test statistic under suitable assumptions. 5+5

6. (a) For a two-sample problem frame a hypothesis to test whether two populations are identical. Describe Wald-Wolfowitz run test for two-sample location problem.
   (b) Define Kolmogorov-Smirnov two-Sample statistic. Describe under which context we may use Kolmogorov-Smirnov two-Sample test and under which context we may use Wald-Wolfowitz run test (state suitable assumptions also, if any).
   5+5

M.Sc. Examination 2022
Semester II
Statistics
Course: MSC-22
(Applied Multivariate Analysis)

Full Marks: 40 Time: 3 hours

Answer any four of the following six questions.
(Notations carry usual meanings)

1. (a) Define Principal Components (PC) and write down the objectives of Principal Component Analysis. Would you perform standardisation of data before finding PCs? If yes, why?
   (b) Write down orthogonal factor model with assumptions and show how the dispersion matrix of the data can be written in terms of factor loadings and specific variances.

   5+5

2. (a) What is Multivariate Analysis of Variance (MANOVA)? When would you use this technique? Give at least two examples.
   (b) Explain one-way MANOVA method along with the MANOVA table.

   3+7

3. (a) Explain discriminant analysis and classification. Write down at least two real life examples where such techniques are used.
   (b) Derive Fisher's linear discriminat function for differenciating two multivariate populations. State necessary assumptions and prove any results you use for this.

   3+7

4. (a) What is cluster analysis? Explain the K-means clustering algorithm. How the number of clusters is determined?
   (b) Let $\mathbf{X} = (x_1, x_2, \ldots, x_p)'$ be partitioned as $X^{(1)} = x_1$ and $X^{(2)} = (x_2, x_3, x_4, \ldots, x_p)'$. Find the first Canonical Correlation between $X^{(1)}$ and $X^{(2)}$. Identify the expresion and name it? You may take the dispersion matrix of $\mathbf{X}$ as

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

   5+5

5. (a) Show that $Var(Y_i) = \lambda_i$ where Y_i is the i^{th} principal component (PC) of $\mathbf{X} = (x_1, x_2, \ldots, x_p)'$ and λ_i is the i^{th} eigenvalue of the dispersion matrix of $\mathbf{X}$.
   (b) What is factor rotation? Why do we need that? Explain the rationale behind factor rotation.

   5+5

6. Derive the expression of the optimum error rate (Minimum Total Probability of Misclassification) when $p_1 = p_2 = \frac{1}{2}$ and $f_1(x)$ and $f_2(x)$ are the multivariate normal densities with common covariance matrix Σ. You may use result of optimum classification technique.

   10

M.Sc. Examination, 2022
Semester-II
Statistics
Course: MSC-23
(Regression Techniques)
Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin.
Notations have their usual meanings

Answer any four questions.

1. (a) Consider the simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, with $E(\epsilon) = 0$, $var(\epsilon) = \sigma^2$, and and ϵ uncorrelated. Show that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$. Also show that $Cov(\bar{y}, \beta_1) = 0$. .

(b) Consider multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Illustrate the significance of hat matrix in multiple regression. Prove that the matrices H and $I - H$ are idempotent. Show that in the multiple linear regression model $Var(\hat{Y}) = \sigma^2 H$.

$5 + 5$

2. Suppose we wish to find the least-square estimator of β in the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ subject to a set of equality constraints on β, say $\mathbf{T}\beta = \mathbf{c}$. Show that the estimator is

$$\tilde{\beta} = \hat{\beta} + (\mathbf{X'X})^{-1}\mathbf{T'}[\mathbf{T}(\mathbf{X'X})^{-1}\mathbf{T}]^{-1}(\mathbf{c} - \mathbf{T}\hat{\beta}),$$

where $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'y}$. Discuss situations in which this constrained estimator might be appropriate.

10

3. (a) What is the studentized residual and when is it used? Show that, the studentized residuals (r_i) can be expressed as $r_i = \dfrac{e_i}{\sqrt{MS_{Res}\left[1 - \left(\dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}$, for $i = 1, 2, \cdots, n$. (notations have their usual meaning).

(b) Write short note on PRESS Residual and PRESS statistic.

$5 + 5$

4. (a) What is Generalized Linear Model? When do we use logistic regression? What is "logit transformation"?.

(b) Write short note on: leverage point and influential point.

$5 + 5$

5. Suppose that there are only two regressor variables, x_1 and x_2 . Let r_{jy} is the simple correlation between x_j and y , $j = 1, 2$. The model, assuming that x_1, x_2 , and y are scaled to unit length, is $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

(a) Find $(\mathbf{X^T X})^{-1}$ in terms of r_{jy} (the simple correlation between x_j and y , $j = 1, 2$.)

(b) Show that, strong multicollinearity between x_1 and x_2 results in large variances and covariances for the least - squares estimators of the regression coefficients.

$5 + 5$

CONTINUED

6. (a) Define "piecewise polynomial fitting" and spline regression. When do we use ridge regression? Does the correlation matrix give any indication of multicollinearity? Illustrate your answer.

(b) Define the ridge estimator $\hat{\beta}_{\mathbf{R}}$. Obtain the mean square error of the ridge estimator. Justify the following statement:

Ridge estimate will not necessarily provide the best " fit " to the data.

$5 + 5$

7. (a) Write short note on: nonlinear regression model.

(b) Illustrate how the linearization can be accomplished by a Taylor series expansion of the nonlinear Regression function, followed by iteration method of parameter estimation.

$5 + 5$

M.Sc. Examination 2022
Semester II
Statistics
Course: MSC-24
(Design of Experiments)

Full Marks: 40 Time: 3 Hrs.

Answer any four questions.

1. (a) For a BIBD with parameters (b, v, r, k, λ), if b is divisible by r, then prove that $b \geq v + r - 1$.
   (b) Let N be the incidence matrix of a symmetric BIBD with parameters (v, r, λ). If v is even, show that $(r - \lambda)$ is a perfect square.
   (c) Give the layout of a $(3^3, 3^2)$ experiment confounding (i) ABC^2, AB^2C, (ii) AB^2C^2, AC.

   3+3+4

2. (a) Construct BIBD s with the following parameters:
      i. $b = v = 11, r = k = 5, \lambda = 2$
      ii. $v = 15, b = 35, r = 7, k = 3, \lambda = 1$.

      You have to clearly state the appropriate results used.
   (b) What are the values of the parameters of a residual BIBD obtained from a symmetric BIBD with parameters (v, r, λ)?

   (4+4)+2

3. (a) In the context of an incomplete block design with b blocks and v treatments, find the inter-block estimate of the treatment contrast $\boldsymbol{p'\tau}$ and also find its variance.
   (b) In the context of the combined inter and intra block analysis of a block design, show that $E(\boldsymbol{Q}) = C\boldsymbol{\tau}$, where $\boldsymbol{Q} = \boldsymbol{T} - \frac{N\boldsymbol{B}}{k}$ stands for the vector of adjusted treatment total and C is the C-matrix of the design.

   7+3

4. What do you mean by the connectedness of a block design? Give the rank definition and the structural definition of connectedness. Show that these definitions are equivalent.

   1+2+7

5. (a) What do you mean by a variance-balanced block design?
   (b) State and prove a necessary and sufficient condition for a connected block design to be variance-balanced.
   (c) In the context of a general block design with b blocks and v treatments, prove that $b + Rank(C) = v + Rank(D)$, symbols having their usual meanings.

   1+5+4

6. (a) What are the advantages and disadvantages of a split-plot design? Mention some of its uses.
   (b) Distinguish between split-plot design and factorial experiments.
   (c) Find the efficiency of a split-plot design with respect to a RBD.

   4+2+4

M.Sc. Examination, 2022

Semester-II

Statistics

Course: MSC-25 (Practical)

Time: 4 hours Full Marks: 40

Questions are of values as indicated in the margin

1. The sample correlations for five stocks- Allied Chemical(AC), Du-Point(DP), Union Carbide (UC), Exxon(Exx) and Texaco (Tex) are given below rounded to two decimal places

	AC	DP	UC	Exx	Tex
AC	1				
DP	0.58	1			
UC	0.51	0.60	1		
Ex	0.39	0.39	0.44	1	
Tex	0.46	0.32	0.45	0.52	1

Cluster the stocks using the single-linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results. **7**

2. Suppose that $n_1 = 11, n_2 = 12$ observations are made from two random variables X_1 and X_2 where X_1 and X_2 are assumed to have a bivariate normal distribution with a common covariance matrix Σ, but possibly different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. The sample mean vectors and pooled covariance matrix are

$$\overline{\boldsymbol{x}}_{\mathbf{1}} = [-4, -5]', \qquad \overline{\boldsymbol{x}}_{\mathbf{2}} = [2, 1]'$$

$$\mathrm{S}_{Pooled} = \begin{bmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{bmatrix}$$

(a) Construct Fisher's (sample) linear discriminant function.

(b) Consider the observations

$$\boldsymbol{x}_{01} = [5, 4]', \qquad \boldsymbol{x}_{02} = [-1, 1]'$$

and state in which the observations belong to using the function derived in (a). You may assume equal costs and equal prior probabilities. **7**

3. Suppose that $X_1, X_2, ..., X_15$ are iid Poisson with unknown mean λ. Derive the most powerful test at level $\alpha = 0.10$ test for $H : \lambda = 0.40$ vs $K : \lambda = 0.30$. **6**

4. Consider the following 16 samples generated from a normal distribution with mean μ and variance 1.
8.936 10.872 10.231 7.914 11.293 9.046 11.218 10.089 10.002 9.002 7.411 9.274 9.147 9.933 10.293 9.065
Find the uniformly most powerful test at level $\alpha = 0.05$ to test the following hypothesis:
$H : \mu = 10$ vs $K : \mu > 10$. **6**

5. Suppose we want to test whether a standard chi-square variable with large degree-of-freedom (dof) can be approximated by a standard normal. Two mutually independent random sample, each of size 8 were generated, one from standard normal (variable Y), another from chi-square (dof=18) (variable C), followed by standardization using mean 18, standard deviation $\sqrt{2 \times 18} = 6$ (variable X). The data is given below:

Y	-1.91	-1.22	-0.96	-0.72	0.14	0.82	1.45	1.86
C	4.90	7.25	8.04	14.10	18.30	21.21	23.10	28.12
X	-2.18	-1.79	-1.66	-0.65	0.05	0.54	0.85	1.69

Perform Kolmogorov-Smirnov two sample test for above problem. **5**

6. Consider the following dataset (time required to pour molten metal into the die). It was suspected that pouring time before lunch were shorter than pouring time after lunch. Perform Wilcoxon Rank-Sum Test for this problem.

Pouring time in sec (before lunch)	12.6	11.2	11.4	9.4	13.2	12.0
Pouring time in sec (after lunch)	16.4	15.4	14.1	14.0	13.4	11.3

4

7. Practical Note book & Viva-Voce **5**

M.Sc. Examination 2022
Semester II
Statistics (Practical)
Course: MSC-26 (Practical on MSC-23 and MSC-24)

Full Marks: 40 Time: 4 Hours

(1) Consider an incomplete block design with v = 9 treatments, b = 7 blocks andblock contents as follows:

Block 1	1 35
Block 2	2 4
Block 3	357
Block 4	46
Block 5	1 57
Block 6	2 6
Block 7	8 9

(i) Find the C-matrix of the design.
(ii) Hence or otherwise, check whether the design is connected or not.
(iii) Check whether the design is variance-balanced or not. 3+2+3=8

(2) The following table provides the yields of 10 treatments in 15 blocks, the layout being a BIBD in blocks of 4 plots. Make a complete intra-block analysis. 10

		VARIETIES									
B L O C K S		1	2	3	4	5	6	7	8	9	10
	1	9.7	8.7	----	5.4	5.0	----	----	----	----	----
	2	----	9.6	8.8	----	----	6.6	----	----	----	3.6
	3	----	9.0	----	7.2	----	3.8	4.3	----	----	----
	4	9.3	----	8.7	----	6.8	----	3.8	----	----	----
	5	10.0	----	----	7.5	----	----	----	4.2	----	2.8
	6	----	9.6	----	----	----	----	5.1	4.6	3.6	----
	7	----	9.8	----	----	7.4	----		4.4		3.8
	8	----	----	----	----	9.4	----	6.3	----	5.1	3.0
	9	9.3	9.3	8.2	----	----	----	----	----	3.3	----
	10	----	----	----	8.7	9.0	6.0	----	----	3.5	----
	11	9.1	----	----	----	----	6.7	6.6	----		2.8
	12	----	----	9.3	8.1	----		----	----	3.9	2.4
	13	9.8	----	----	----	----	7.4	----	5.4	4.0	----
	14	----	----	9.0	8.3	----	----	4.8	3.8	----	----
	15	----	----	9.3	----	8.3	6.3	----	3.7	----	----

(3) Download the data on the concentration of $NbOCl_3$ in a tube along with several controllable variables using the following R-code.

```
library("MPV")
attach(table.b6)
```

(i) Using R, fit a multiple regression model relating concentration of $NbOCl_3(y)$ to concentration of $COCl_2(x_1)$, and mole fraction (x_4). Also carry out a test for the significance of regression.

(ii) Calculate R^2 and adjusted$R^2$for this model.

(iii) Using appropriate test(s), determine the contribution of x_1 and x_4 to the model. Are both regressorsx_1 and $x_4$necessary?

(iv) Justify if multicollinearity a potential concern in this model or not.

2+2+2+2=8

(4) Download the data on the relationship of an abrasion index (y) for a tire tread compound in terms ofhydrated silica level(x_1), silane coupling agent level(x_2) andsulfur level(x_3) using the following R-code.

```
library("MPV")
data(p4.19)
attach(p4.19)
```

(i) Frame the regression model and write down the R codes for performing regression. Perform a thorough regression analysisand interpret the outcomes.

(ii) Write the appropriate R code for ”residual vs. fitted” plots.

(iii) Write down appropriate R code for testing whether or not here a full regression model offers a significantly better fit to a dataset than some reduced version of the model.

2+1+2=5

(5) Download the data on the test-firing results for 25 surface-to-air antiaircraft missiles at targets of varying speeds, using the following R-code.

```
library("MPV")
data(p13.1)
```

(i) Write down the appropriate R code for fitting a logistic regression model to the response variable *y*. Use a simple linear regression model as the structure for the linear predictor.

(ii) Write down the appropriate R code for obtaining model deviance here. Does the model deviance indicate that the logistic regression model from the previous part is adequate? 2+2=4

(6) Practical Note Book and Viva-Voce. 5

M.Sc. Examination, 2022
Semester-IV
Statistics
Course: MSC-41
(Reliability Theory)
Time: 3 Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. (a) Define reliability of a system. How does it differ from quality?
   (b) What is a reliability function ($R(t)$)? How does it related to failure density function ($f(t)$).
   (c) Define mean time to failure (MTTF) and show that MTTF alone can not uniquely characterize a failure distribution. 3+3+4

2. (a) Define hazard rate ($h(t)$). Show that if any one of the four quantities ($f(t)$, $F(t)$, $R(t)$ and $h(t)$) are given, all others can obtained (symbols having usual meanings)
   (b) Prove that a particular hazard rate function will uniquely determine a reliability function.
   (c) Suppose a component has a reliability function given by
   $$R(t) = 1 - \frac{t^2}{a^2}; 0 \leq t \leq a$$
   where a is the parameter of the distribution representing the component's maximum life. Find $h(t)$, MTTF, median time, average failure rate (AFR) and residual MTTF.
   3+2+5

3. (a) What is a bathtub curve. Give an example of a bathtub curve based on linear and constant hazard rates. What is a piecewise linear bathtub curve?
   (b) For the reliability function
   $$R(t) = \frac{a^2}{(a+t)^2}; t \geq 0, a > 0$$
   Find out residual mean time to failure. 6+4

4. (a) For a series system, show that if every component has a Weibull failure rate, the system may not exhibit Weibull type failures.
   (b) Show that, in general, the low level redundancy gives a higher system reliability.
   (c) Suppose n systems are connected (i) through a series-parallel system and (ii) through a parallel-series system. Which one will give better reliability?
   4+3+3

5. (a) Define IFR and DFR. Assume F has a density f, with $F(0-) = 0$. Then Show that F is DFR if and only if $R(t)$ is decreasing. Also show that $\log(1 - F(t))$ is convex for t in $\{t: F(t) < 1, t \geq 0\}$

(b) Show that if F is IFR, then

$$\left(1 - F(t)\right)^{\frac{1}{t}}$$

is decreasing in t. 7+3

6. (a) What are Accelerated Life Testing Models? For a two-parameter Weibul distribution find the expression for CDF and reliability function under accelerated life testing models. (b) Give some examples of stress-related failures. Find the general expression of the reliability function. Find out the reliability function if both stress and strength follow a log-normal distribution with different location and scale parameter.

4+6

M.Sc. Examination, 2022
Semester-IV
Statistics
Course: MSS-3 (Demography)
Time: Three Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any five** questions

1. (a) Assuming an exponential distribution for time to first conception and choosing an appropriate prior for the parameter involved show that the resulting distribution is decreasing failure rate.
(b) Starting from suitable assumptions derive the distribution of time to the nth child birth. 4+4

2. Derive Lotka's integral equation for birth function and hence explain a method of finding out the solution. 5+3

3. (a) Distinguish between curtate expectation e_x and complete expectation of life e_x^0 at age x and find approximate relation between them.
(b) Write down the probability generating function of Life table values $l_1, l_2, ..., l_w$ given the cohort l_0. Discuss its application. 2+6

4. How one can adjust age data? Discuss the uses of (i) Whipple's index, (ii) Meyer's blended index and (iii) Age dependency ratio. 2+2+2+2

5. What is social mobility? Explain the special situations of 'perfect mobility' and 'perfect immobility'. Discuss some possible measures of social mobility when the socio-economic categories can be ordered in some sense. 2+2+4

6. (a) Give the Logistic equation for population growth and interpret the parameters involved.
(b) Describe a method for fitting this equation to observed population data. 5+3

7. Differentiate between the Gross Reproduction Rate (GRR) and the Net Reproduction Rate (NRR), and interpret the situations when for a society (i) GRR=NRR and (ii) NRR=1. 5+3

8. Write short notes on any two of the following: 4+4
(i) Chandrasekharan-Deming balancing equation, (ii) Use of Leslie matrix, (iii) Estimation of net migration based on census and registration data.

M.Sc. (Honours) Examination, 2022
Semester-IV
Statistics
MSS-08-Econometrics
Time: 3 hrs Full Marks:40

Answer any **four** questions of the following.

1. (a) Suppose that linear regression model is $y_i = \alpha + \beta x_i + \epsilon_i$ where $f(\epsilon_i) = \frac{1}{\lambda} e^{-\lambda \epsilon_i}$, $\epsilon_i \geq 0$. Show that the least squares slope is unbiased.

   (b) For a linear regression model $Y = X\beta + \epsilon$ where $\beta_{k\times 1}$ is the vector of coefficient. Propose a test for $H_0 : \beta_2 + \beta_3 = 2$ against $H_1 : \beta_2 + \beta_3 \neq 2$ provided $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I)$.

   (c) Suppose a regression model is $y_i = \mu + \epsilon_i$ where $E(\epsilon_i/x_i) = 0$, $cov(\epsilon_i, \epsilon_j/x_i, x_j) = 0$ for $i \neq j$ but $V(\epsilon_i/x_i) = \sigma^2 x_i$. Find out the variance of OLSE of μ.

   3+4+4

2. (a) For a linear regression model $Y = X\beta + \epsilon$ where $\epsilon \sim N(\mathbf{0}, \sigma^2 \wedge)$, deduce the maximum likelihood estimates of β and σ^2.

   (b) In case of heteroscedastic structure of error how would you find an efficient estimator of β? Describe the technique.

   5+5

3. (a) How do you assume graphically that residuals are negatively autocorrelated?

   (b) Define Durbin Watson test with its range.

   (c) Write down the null hypothesis where the Durbin Watson test is used as a test statistic.

   (d) Is it an exact test?if not why?

   (e) State the rejection criterion on the basis of Durbin Watson test statistic.

   2+2+2+2+2

4. State with reason whether the following statements are TRUE/FALSE.

   (a) Despite perfect multicollinearity , ordinary least square estimates are BLUE.

   (b) The higher the VIF is, the larger is the variances of OLS estimator.

   (c) Multicollinearity is harmless if the objectives of the analysis is prediction only.

   (d) In logit model the slope coefficient of variable interprets the same asthe slope coefficient of probit model.

   (e) Linear regression model without intercept term fails to estimate non zero expectation of error term.

   5×2

5. (a) Consider the least square regression of Y on k variables $X_1, X_2, \cdots, X_k$. Consider a new set of regressors $\mathbf{Z} = \mathbf{X}P$ where P is nonsingular matrix. Prove that the residual vectors in the regression of Y on $\mathbf{X}$ and regression of Y on $\mathbf{Z}$ are identical.

(b) Suppose for a regression model, disturbance follows ARMA(1,1). Describe a test process to check validity of this statement, clearly writing the null hypothesis, alternative hypothesis and test statistic.

5+5

6. (a) Prove that for a linear regression model with autocorrelated disturbances, OLS is unbiased but inconsistent.

(b) Suppose a labour economist postulates an earning function as earning=f(sex,race,age, educational level). Caucasian, black and other are the ethnicity of labour class while mainly middle school pass and high school pass are working as the labour. Propose a linear regression model without any interaction.

(c) In construction of test of heterogeneity which test do you prefer- Breusch-Pagan or Goldfeldt Quandt test when the data is full of outliesrs? Give reason.

5+3+2

M.Sc. Semester IV Examination 2022

Subject: Statistics

Paper: MSS 09

(Introductory Data Science and Statistical Machine Learning)

Time: Three Hours Full Marks: 40

Answer any **four** questions.

1. a) Describe different types of machine learning algorithms with one example of each type.

   b) What do you understand by data wrangling? Give a brief description of steps to be followed for making a dataset ready to be analyzed. 5+5

2. a) Explain the Adaboost algorithm.

   b) What do you mean by text analysis? Mention two applications of it. 7+3

3. a) Explain support vector machine (SVM) algorithm in detail.

   b) What are support vectors in SVM? Name different types of kernels used in SVM. 6+4

4. What is decision tree? Why is it called decision tree? How to decide on which of the attributes are to be chosen first? Define the terms e.g. Ginni index, information gain, entropy in this regard. 10

5. a) Explain association rules with examples.

   b) Define the terms support, confidence, and lift regarding association rule. How would you interpret lift? 4+6

6. a) What is ensemble learning? Why do we use ensemble learning? In this context describe Bagging.
   b) What do you mean by sentiment analysis? Write down the steps for obtaining sentiment of a statement with examples. 6+4

M.Sc Semester IV Examination, 2022

Statistics

MSC-44(Practical)

Time: Four Hours Full Marks: 40

1. One may use Computer Laboratory, if necessary.
2. MSC-41 is compulsory for all. From other three Special papers choose out two of your papers and answer.

MSC-41 Reliability Analysis

1.(a) 10 hypothetical electronic components are placed on life tests. Failure times for the components are 5, 10, 19.5, 30, 42, 53, 67, 82.5, 100.5, 117
Find the empirical estimates of reliability function and plot. Comment on your findings.

(b) The failure time of a certain component has a Weibull distribution with location parameter 100, shape parameter 4 and scale parameter 200.
Find the reliability of the component and the hazard rate for an operating time of 100 hours.

(c) The failure time of a certain component is log-normally distributed with mean 5 and sd 1. Find the reliability of the component and hazard rate for a life of 150-time units.

(d) A system consists of three subsystems A,B and C. The system is primarily used on a certain mission that last 8 hours. The information we have

Subsystem	Requires operating time during mission in hours	Type of failure distribution	Reliability information
A	8	Exponential	50% of subsystem will last at least 14 hours
B	3	Normal	Average life is 6 hours with s.d. of 1.5 hours
C	4	Two parameter Weibull with shape parameter 1	Average life is 40 hours

Assuming independence of the subsystems, calculate the reliability for a mission.

3+3+3+6

MSS-3 Demography

1. Given the following one generation transition probability matrix, calculated two different measures of mobility in a society with an initial distribution = (0.4,0.3,0.1,0.2) among four social classes

0.24	0.26	0.26	0.24
0.10	0.50	0.10	0.30
0.10	0.15	0.70	0.05
0.10	0.25	0.25	0.40

   Interpret the results obtained. 4

2. Fit a logistic curve of following observed data and comment. 6

Year	Population (in million)
1800	0.551
1810	0.725
1820	0.964

1830	1.278
1840	1.708
1850	2.320
1860	3.145
1870	3.856
1880	5.016
1890	6.304
1900	7.600
1910	9.198

MSS-9 Introductory Data Science and Machine Learning

Consider the following data related to choose of a job offer with respect to different criterion.

Provides relocation cost	Year end bonus	Provides accomod-ation	Accept the offer
yes	yes	no	no
yes	yes	yes	no
may	yes	no	yes
no	yes	no	yes
no	no	no	yes
no	no	yes	no
may	no	yes	yes
yes	yes	no	no
yes	no	no	yes
no	no	no	yes
yes	no	yes	yes
may	yes	yes	yes
may	no	no	yes
no	yes	yes	no
yes	no	yes	no
no	yes	no	yes

Let you are asked to apply decision tree for reaching out to a decision of accepting or rejecting the offer. Then answer the following questions:
a) What is the entropy of the sample?
b) Out of the three criterion (relocation cost, bonus and accommodation) which one would you use as the root node fo the decision tree and why? Give details calculations.
c) Solve the above problem using R or Python and show the associated decision tree plot. 1+6+3

MSS 8 Econometrics

A. Write all your answer in a word file with your code saved by your roll number in your laptop
B. In Answer Script first mention the name of the data you are working with.

1. From cardata package extract Arrest data. How many qualitative independent variables are there?
2. Fit a suitable regression model on "released" to sex and citizen. Report a measure by which you can judge the fitting of the regression.
3. Include "age" variable in your model. Do you think that model is improved?
4. Interpret the coefficient attached to age with respect to the previously taken dependent variable.
5. Do you think employed citizens play significant role in being arrested for having marijuana? Justify your answer. 5 X2 =10

Practical Note Book + Viva Voce 5