M. Sc. Semester IV, Mid-semester Exam 2: CSM 402: Introduction to Data Science
Full Marks: 30, Duration: 1 hour 30 minutes

2 marks x 5

1. **Answer any *five* questions from the following:**

a) The proportions of blood phenotypes in a population are as follows:

A	B	AB	0
0.40	0.11	0.04	0.45

Assuming that the phenotypes of two randomly selected individuals are independent of one another, what is the probability that both phenotypes are O?

b) Define unbiased estimator. Give an example of an unbiased estimator.

c) Define p-value. What is the significance of p-value.

d) Define the central limit theorem and explain it's significance.

e) Assuming porosity of certain material follows a normal distribution with standard deviation of 0.75. Compute a 95% CI for the true average porosity of the material if the average porosity for 20 specimens is 4.85.

f) Define the one-sample CI for μ considering t distribution.

g) Define Null and alternative hypothesis and explain their significance.

5 marks x 4

2. **Answer any *four* questions from the following:**

a) Derive the estimated model parameters β_0 and β_1 for linear regression model. Given n=14 , , , , . Compute SSE, SST and coefficient of determination r^2. Comment on the linear relationship between x and y.
[2+1 +2]

b) Define binomial random variable and distribution. When circuit boards used in the manufacture of compact disc players are tested for defective and non defective, the long-run percentage of defectives is 5%. Let X = the number of defective boards in a random sample of size n=25. i) What is the probability that none of the 25 boards is defective? ii) Calculate the expected value and standard deviation of X. [2+1+ 2]

c) There is a maximum speed test going on for small bikes. The speed assumed to follow a normal distribution with mean value 46.8 km/hr and standard deviation of 1.75 km/hr. For a randomly selected bike, i) What is the probability that maximum speed is at most 50 km/hr ? ii) What is the probability that maximum speed differs from the mean value by at most 1.5 standard deviations? iii) what is the 91st percentile of the distribution? n = 20 [2+2+1]

d) i)What is a Normal probability plot and what is it's significance? ii)A sample of 15 female golfers was selected and the clubhead velocity (km/hr) while swinging a driver was determined for each one, resulting in the following data: 69.0, 69.7, 72.7, 80.3, 81.0, 85.0, 86.0, 86.3, 86.7, 87.7, 89.3, 90.7, 91.0, 92.5, 93.0. The corresponding z percentile are: -1.81, -1.28, -0.97, -0.73, -0.52, -0.34, -0.17, 0.0, 0.17, 0.34, 0.52, 0.73, 0.97, 1.28, 1.83. Construct a normal probability plot. Justify if the population distribution is normal. [2+3]

e) If the mean temperature of discharged water of a factory is at most 150°F, there will be no negative effects on the river ecosystem. To check if the factory is complying to the norms, 50 temperatures were recorded. Give a suitable Null and alternative hypothesis and describe Type I and Type II errors in the context. Which type of error would you consider more serious? Explain. [2+3]

2024

COMPUTER SCIENCE

Paper : CSME-401

(Introduction to Data Science)

Full Marks : 70

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

Answer ***question nos.*** **1, 2,** and ***any four*** from the rest.

1. Answer ***any five*** questions from the following : 2×5

(a) Can k-NN be used for a regression problem? Justify.

(b) Can eigenvectors of a matrix form basis vectors? Justify your answer.

(c) Explain what are principal components of a collection of data points in R^3 in principal component analysis.

(d) Between SVM and logistic regression, which algorithm is most likely to work better in the presence of outliers? Why?

(e) If the p-value is 2.78% and the significance level is 5%, do you reject the null hypothesis?

(f) What is the purpose of Linear Discriminant Analysis (LDA)?

(g) Differentiate between Accuracy and Precision.

2. Answer ***any five*** questions from the following : 4×5

(a) Compare and contrast K-means and K-medoid Algorithms.

(b) You have a dataset of a webpage which gives the number of clicks in particular region of the webpage. State one visualization tool that you will use to understand the distribution of clicks with justification.

(c) Define and state the significances of coefficient of determination r^2 for linear regression.

(d) State Central limit theorem and explain its significance.

(e) Define Orthonormal vectors. Find a unit vector in R^2 that is orthogonal to $[-1\ 2]^T$.

(f) Define linear independence of vectors. If a 5 × 5 matrix has rank 3, what can you say about the number of linearly independent vectors?

(g) Justify the use of sigmoid function in logistic regression.

3. (a) Derive the expression for coefficients β_0 and β_1 in simple linear regression.

(b) A linear regression analysis of Birth-Weight (grams) and Gestational-Age (weeks) gave the following output :

Model	Beta Coefficient	95% CI	p-value
Gestational Age	96.56	14.41 to 178.72	0.02
Constant	– 230.34	– 3340.0 to 3180.30	0.39

Compute the birth-weight of a baby born at 40 weeks gestational age. Formulate the Null hypotheses and Alternate Hypotheses for β_0 and β_1 of this model. At 95% level of significance do you reject the null hypotheses? 5+(1+2+2)

4. (a) Give two different methods of choosing the best attribute in Decision Tree construction.

(b) For the following data set draw a decision tree up to two level using GINI Index.

Age	Gender	BP	Cholesterol	Drug Type
23	F	HIGH	HIGH	DrugX
47	M	LOW	HIGH	DrugA
47	M	LOW	HIGH	DrugA
28	F	NORMAL	HIGH	DrugA
61	F	LOW	HIGH	DrugX
22	F	NORMAL	HIGH	DrugX
49	F	NORMAL	HIGH	DrugX
41	M	LOW	HIGH	DrugA
60	M	NORMAL	HIGH	DrugX
43	M	LOW	NORMAL	DrugX
47	F	LOW	HIGH	DrugA
34	F	HIGH	NORMAL	DrugX
43	M	LOW	HIGH	DrugX

4+6

5. (a) What is odds ratio in logistic regression?

(b) Give the logistic regression model for binary classification problem.

(c) Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$. Estimate the probability that a student who studies for 40 hours and has undergrad GPA of 3.5 gets an A class. 3+4+3

6. (a) Explain the steps of Principal Component Analysis (PCA) for an n×p matrix X justifying each step towards dimensionality reduction.

(b) Is it possible to reconstruct the original matrix from the principal components? 8+2

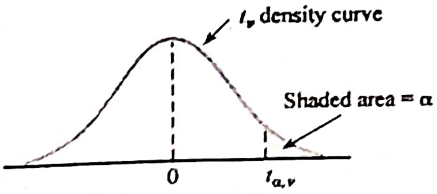
7. Discuss how a Support Vector Machine is formulated highlighting computation of width around seperating hyperplane, formulation of constraint equation and the objective function. 10

8. (a) Define Confusion matrix and the performance measures : Accuracy, Precision, Recall and Fl-score.

(b) Explain Receiver Operating Characteristic (ROC) with a diagram.

(c) Discuss the performance of K-Nearest Neighbour algorithm for small and large of k values. 5+3+2

Table A.5 Critical Values for *t* Distributions



				α			
ν	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725