Roll Number: _____________

National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	Statistical and Mathematical Methods for Data Science	**Course Code:**	DS 501
Program:	MS Data Science	**Semester:**	Fall 2019
Duration:	180 Minutes	**Total Marks:**	80
Paper Date:	December 10, 2019.	**Weight**	40
Section:	N/A	**Page(s):**	6
Exam Type:	Final Exam		

Student : Name:_____________________________ **Roll No.**________________

Instruction/ Notes:

1. Solve in the space provided. Extra sheets will NOT be collected or marked.
2. One A4 handwritten help sheet is allowed.
5. Sharing calculators is NOT allowed
6. In case of any ambiguity make a reasonable assumption.

Good luck!

Problem 1 **(Marks: 1+1+4+4)**

$z_{0.10} = 1.282$, $z_{0.05} = 1.645$, $z_{0.025} = 1.960$, $z_{0.01} = 2.326$, $z_{0.005} = 2.576$

i. $P(z \geq 1.645) =$ ______

ii. $P(-1.96 \leq z \leq 1.645) =$ _____

(For part iii and iv) We have to accept or reject the proposal for using a data science software. We take a sample of 200 test points and find 24 misclassifications in it.

iii. Construct a 90% confidence interval for the proportion of misclassifications made by the system. Working required.

iv. The developer claims that at the most one in 10 labels are misclassified by the system. Do we have sufficient evidence to disprove this claim at the 5% level of significance? Working required.

Problem 2 (Marks 5+5)

x_1	-1	1	3	-2	-1	1	2	3	1	3
x_2	1	2	3	1	3	-1	1	3	-2	-1
Label	-1	-1	-1	-1	-1	1	1	1	1	1

i. Build a MAP classifier by fitting a 2D-Gaussian distribution to the given set of points. How would you classify the test point (1,2)? Specify all parameters of the classifier. Use unbiased estimates of covariance and variance.

ii. Plot the contours of the distributions for +1 and -1 classes and identify the +1 and -1 regions.

Problem 3 **(Marks: (4+2) +4)**

i. Given the function $f(\mathbf{x}) = x_2 - x_1^2 - 1$ and the constraint $x_2 + 2x_1 + 2 = 0$. Find the stationary points of this function using the method of Lagrange multipliers. Also, draw the contours of this function and the feasible set. Show working.

ii. When training a perceptron, apply one iteration of the stochastic gradient descent algorithm when the momentum is set at 0, learning rate is fixed at 1/4 and initial weights are (-1,1,-2) (-1 is the bias). The training point is (3,4) with the target being +2. Show formula and working.

Problem 4 (Marks: 5)

Find two vectors orthogonal to $[1/\sqrt{6} \quad -1/\sqrt{6} \quad 2/\sqrt{6}]^T$ so that the resulting three vectors form an orthonormal set. Show all working.

Problem 5 (Marks: 2 + (1+1+1) +5)

Given the following actual labels and output from a classifier:

Label	+1	-1	-1	-1	+1	+1	-1	-1	+1	+1
Output	0.01	0.2	0.45	0.6	0.65	0.7	0.9	0.95	0.98	0.99

i. Suppose we predict the label as +1 if classifier output > 0.5, then compute:

TP =	FP =	Precision =	Balanced error rate =
FN =	TN =	Recall =	

ii.Plot the ROC curve for this classifier by computing 5 points on the curve. Write down the computed points and the threshold value for each point.

Problem 6 (Marks: 4+3+3)

Consider a hypothetical kitchen with cooks from various cities, wearing three different apron colors, i.e., red, green and blue.

• 70% cooks are from Lahore • 20% cooks are from Karachi • 10% cooks are from Multan	We have the following observations regarding the aprons • For Lahori cooks, 30% wear red and 20% wear green aprons • For Karachi cooks, 5% wear red and 45% wear green aprons • For Multani cooks, 10% wear red and 50% wear green aprons

Use the laws of probability to compute the following. For each part, clearly write down the mathematical formula and show the working clearly. No marks without proper working.

i. If we spot a cook wearing a red apron, what is the probability that he is from Lahore?

ii. If a cooks is wearing a green apron, then use maximum likelihood classifier to find the cook's corresponding city.

iii. If a cook is wearing a blue apron, then use maximum aposteriori classifier to find the cook's city.

Problem 7 **(Marks: 2+4+2+4+2+3+2+2+4)**

i. Given observations of x = {0,0,1,1,1}. What is estimate of P(x=1) with Laplacian smoothing? _____

ii. Suppose in a building, there are 30% people who are both short and have brown eyes. There are 60% people who are short. What fraction of short people have brown eyes? Write the rule of probability and show working.

iii. The L_3 distance between (1,0,1) and (-1,1,-3) = __________________

iv. Find the singular values and matrix of right singular vectors of: $\begin{bmatrix} 25 & 0 \\ 0 & 16 \end{bmatrix}$ (working not required)

v. What is the span of $[1 \quad 1 \quad 2]^T$ and $[1\ 1\ 4]^T$?

vi. Use Chebyshev's inequality to find the bounds on the probability that a point would lie outside the interval [6,10] when the population mean is given by 8 and variance is given by 1. Working required.

vii. Do the vectors $[2 \quad 2 \quad 2]^T$, $[1\ 1\ 4]^T$, $[4\ 4\ 10]^T$ form an independent set? Give reason for your answer. A simple yes or no is not acceptable.

viii. What is the orthogonal projection of $[1 \quad -1 \quad 2]^T$ onto $[3 \quad 1 \quad 4]^T$. Show working.

ix. Use least square regression to fit a straight line to the two points (1,3) and (2,0)? Show working and write the final equation of the line. What are the residuals in this case?