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 2018
Duration:	180 Minutes	Total Marks:	95
Paper Date:	December 31, 2018	Weight	
Section:	N/A	Page(s):	6
Exam Type:	Final Exam		

Student : Name:_____________________________ **Roll No.**________________ **Section:** **MS DS**

Instruction/Notes:
1. One A3 cheat sheet is allowed in the exam
2. Mobile phones, laptops and other electronics that connect to the internet are NOT allowed
3. Use of calculators is allowed. Sharing calculators is NOT allowed
4. Answer in provided space. Extra sheets will NOT be collected and graded.

QUESTION 1 **(Marks: 5+2+5+2)**

Given the following predictions and their corresponding labels:

prediction	0.9	-0.2	0.82	0.01	0.97	0.2	-0.7	0.6	-0.95	0.8
label	+1	+1	+1	+1	-1	-1	-1	-1	-1	-1

Suppose predictedLabel = +1 if prediction ≥ threshold, otherwise predictedLabel = -1.

a. Find 3 points on the ROC curve for thresholds 0.85, 0.5, 0. Write the points and plot them (working not required).

b. Make the confusion matrix for threshold = 0

c. Find for threshold = 0

i. balanced error rate = __________ ii. sensitivity = ______

iii. specificity = ______________ iv. precision = _________ v. recall = _________

d. For what maximum value of threshold do you get maximum recall? ___________

QUESTION 2 **(Marks: 2+2+2+2+2)**

Given the following transition matrix for a Markov chain:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.9 & 0 & 0.1 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

a. Is it an absorbing Markov chain? ☐ yes ☐ no

b. P(S_2 at time t=3 | S_1 at time t=2) = ____________________

c. P(S_3 at time t=4 | S_2 at time t=2) = ______________________

d. P(S_4 at time t=5 | S_2 at time t=2) = ______________________

e. Probability of the Sequence $S_2S_3S_3S_4$ = ______________________

QUESTION 3 **(Marks: 6+2)**

a. Find the steady state vector of the following transition matrix P of a regular Markov chain. Show all working.

$$P = \begin{pmatrix} 1/5 & 4/5 \\ 3/10 & 7/10 \end{pmatrix}$$

b. What is the value of P^{1000}?

QUESTION 4 **(Marks: 3+3+3+3+4)**

a. Do the following set of vectors form a linearly independent set?

i. $[1\ 10\ 20]^T$ $[3\ 10\ 5]^T$ $[4\ 20\ 25]^T$ ☐ yes ☐ no

ii. $[1\ 10]^T$ $[3\ 31]^T$ $[2\ 22]^T$ ☐ yes ☐ no

iii. $[1\ 10\ 20\ 1]^T$ $[3\ 11\ 15\ 2]^T$ $[14\ 20\ 26\ 4]^T$ $[19\ 90\ 66\ 4]^T$ ☐ yes ☐ no

b. What is the span of $[1\ 2]^T$ and $[5\ 8]^T$

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

d. Find the projection of (2,1,0,1) onto **w** = (0,1,2,-1).

e. Find the Mahalonobis distance between the point (1,2,1) and (0,0,0) when given the following covariance matrix. Show working/formula:

$$\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}$$

QUESTION 5 **(Marks: 6+2+2)**

Given the following parameters after training an SVM machine.

x1	1	2	1	3	2	3
x2	1	2	3	4	5	6
target	+1	+1	+1	-1	-1	-1
α	0	1	3	2	2	0

a. Find the weights of the SVM machine and the equation of the separating boundary.

b. Suppose the weights found by the SVM machine are: $w_0 = 1$, $\mathbf{w} = [1\ 2]^T$. Find the classification of the points: (1,4) and (0,6)

c. Draw the decision boundary for weights $w_0 = -1$, $\mathbf{w} = [2\ \ 1]^T$

QUESTION 6 **(Marks: 6)**

Diagonalize the following matrix. Show working and write the final answer.

$$\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$$

QUESTION 7 **(Marks: 6+6+6+6)**

a. Given the following values of the standard Normal variable:

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

Find the 95% confidence interval for the errors made by a machine learning algorithm, which was tested on 10 OCR samples and committed two errors.

b. Suppose a security system is used to authorize people to enter a building. In 70% cases the system allows a person to enter. Suppose the probability of an un-authorized person entering the building is 0.1. The system only allows 60% of authorized persons to enter the building. Suppose the system allows a person to enter then what are the chances that the person is authorized? Show all working/formula.

c. Suppose a lab technician observed a total of 240 people. 80 people had cancer. 25% people with cancer were exposed to high nitrogen levels. Out of 240 people observed, 60 people were exposed to high nitrogen levels. Using probability theory can we conclude from this data that having cancer depends on being exposed to high nitrogen levels or are the two independent? How do you make your conclusion? Show your justification/formulas clearly.

d. Given the following data for 3 attributes x_1,x_2,x_3 and their corresponding labels

x_1	0	1	0	1	0
x_2	0	1	1	1	1
x_3	1	0	1	0	1
label	+1	+1	+1	-1	-1

Give the classification of the point (0,1,0) when a naive Baye's classifier is used. Show all working.

QUESTION 8 **(Marks: 1+6)**

Draw the feasible region and find the stationary points of the given function using the method of Lagrange multipliers. Show working:

$f(\boldsymbol{x}) = 2x_1^2+x_2$

subject to $x_1 = -x_2$