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:	60 Minutes	**Total Marks:**	30
Paper Date:	November 07, 2019	**Weight**	
Section:	N/A	**Page(s):**	2
Exam Type:	Midterm Exam 2		

Student : Name:__________________________ Roll No.______________

Instruction/Notes: 1. Solve in the space provided. Extra sheets are NOT allowed
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: 5)

1. Find the eigen vectors and eigen values of the given matrix. Show all working: $\begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix}$

Solution

(Working not shown)

Eigen values are {4,1}. The corresponding Eigen vectors are: (2,1) and (-1,1)

PROBLEM 2 (Marks: 2+2+2+2+2)

2a. Write (4,5) as a linear combination of vectors in the set: {(2,0),(1,1)}. How many combinations are possible?

-1/2*$\mathbf{v}_1$ + 5*$\mathbf{v}_2$

There is only one possible combination

2b. Add vectors to the set {(1,1,1),(1,2,1)}, so that it forms a basis for R^3.
Can be any vector linearly independent of the vectors in the given set. For example: (2,5,7)

2c. What is the span{(1,1,0,0), (1,2,0,0)}. Is the span a vector space? Give reason.
The span is a plane in R^4. If we label the coordinates in R^4 as (x,y,z,w) then it is the xy plane in R^4.
Yes, the span is a vector space as it is closed under multiplication with a scalar and addition, and it includes the zero vector.

2d. The regression co-efficients for a 2 variable problem are given by (10,5,-1). Here 10 is the intercept. What is prediction for the point (2,5). (show working)
prediction = 10+2*5+5*-1 = 15

2e. What would the regression co-efficients be close to if the shrinkage/regularization constant is set at infinity.
They will be zero (or very close to zero)

PROBLEM 3 (Marks: 2.5+2.5)

We have a Gaussian distribution for positive class given by μ_1 = (3,2) and Σ_1 = I. For the negative class, the Gaussian has μ_2 = (-3,-2) and Σ_2 = I. Draw the contours of the two Gaussian distributions. Suppose we build a classifier using equal priors and using the two Gaussians for computing the likelihood. Plot the decision boundary and show the predicted classes for each region.

PROBLEM 4 (Marks: 5)

How many minimum test samples do you require to guarantee that the accuracy (ratio of total correct to total samples) is estimated with an error of at the most 0.05 with a 90% confidence. Show formula and working.

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

Solution

$n >= 0.25*1.645^2/.05^2$

We need at least 271 samples

PROBLEM 5 (Marks: 5)

The average time for running an SVM on 50 test samples was 20ms with a variance 2.5ms and the average time taken by naive Bayes' was 10ms with a variance of 4ms on 100 test samples. Can we claim that the running time for both algorithms is the same? Show formula and working. Justify your reasoning for proving or disproving this claim. (See problem 4 for z-values)

Solution

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

H_0 is equivalent to claiming that the difference in mean running times of the two algorithms is zero. Perform a two sample test.

Compute $z_{obs} = 33.33$

We can see that z_{obs} is a very large value and hence it would not be in the acceptance region even if the confidence interval is taken to be 99%. Hence we have no sufficient evidence to accept the hypothesis that the average running times are the same and we reject the null hypothesis in favor of the alternative that the running times of the two algorithms are significantly different.