First Mid Term Examination, ME-781, Feb 22, 2019

Name: **Roll No:**

Total Time 2 hours; Total Marks 100

Open notes (self-hand-written) examination.

15 1. Average mass of a car is 1000 kg. Such a car gives an average mileage of 20 km per liter. The mileage of any car can be shown to have a linear relation with the deviation of its weight from the average mass.
If a linear and a nonlinear regression model of the type
$Y = \alpha_0 + \alpha_1 X + \varepsilon_2$, and $Y = \beta_0 + \beta_1 X^2 + \varepsilon_1$ is fitted between the variables X (which represents the deviation of a cars mass from the average mass of a car) and its mileage Y, then derive an expression for β_0, β_1 and compare the coefficients of the linear and nonlinear model.

10 2. The p-value represents the probability that your observed results could be random given the null hypothesis. So a small p-value means that there is a small chance that your results are random. Thus, implying that the results are not random; which means that the null hypothesis is not true. Typically a p-value of 0.05 is considered good enough to discard the null hypothesis.
If in an experiment of tossing of a coin, one gets four consecutive heads. Then is this observation good enough to discard the null hypothesis that the coin being tossed is a fair coin? Provide reasons for your answer.

10 3. Multiple linear regression model has the form
$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$,
with $RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$
This can be written in matrix form as:

$$Y = X\beta$$

Where, *Y*, *X* and β are matrix of the size nx1, nx(p+1) and (p+1)x1, respectively. (Note that n is the number of training data points, and p is the number of predictors)
And

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta)$$

And

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X^T (Y - X\beta)$$

Show that the choice of β which minimizes the RSS leads to residual vector $(Y - X\beta)$ becoming orthogonal to column space of X. Write all the steps and expand the matrix notations for the derivatives.

8

4. Probability of heart disease (Y) is seen to be linearly related to age (X1), gender(X2), body mass index (X3), and ethnicity(X4) (caucasian, african American or south Asian) of a person. Write the typical form for the relation between X and Y variables (please note that some of the X variables have more than one Level). How would this equation be different if Y had a nonlinear relation with X (with no terms having more than one X variable)

18

5. Let set A= $b^{int}((4,4),2)$ and set B=b((0,0),1) in $\boldsymbol{R^2}$ space. Then draw the following and comment as to which of the following sets are Open, Closed, Open precisely and Closed precisely.

a.)	C = A U B	b.)	D = A $\bigoplus$ B	c.)	E = A $\bigominus$ B
d.)	C - D	e.)	D^{cl} – C	f.)	D - E

9

6. On a flat ground several flags have been hosted on poles. Each flag has an (x, y) co-ordinate location and a height z. Hence, this arrangement of flags leads to a data set of an ordered triplets (x, y, z). If there are n such flags, then it leads to a nx3 matrix. Each row of the nx3 data matrix corresponds to feature row vector of the data set or data point X_k k = 1,…., n.
   i) Which Data Scales do the three variables of the ordered triplets (x, y, z) belong to?
   ii) If a norm is used as a dissimilarity measures to quantify dissimilarity between two flags, then comment on the meaning of the following two norms:

   (a) $\|X_i - X_j\| = \sqrt{(X_i - X_j)\mathrm{A}(X_i - X_j)^{\mathrm{T}}}$, where $A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

   (b) $\|X_i - X_j\| = \sqrt{(X_i - X_j)\mathrm{A}(X_i - X_j)^{\mathrm{T}}}$, where $A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1000 \end{bmatrix}$

10

7. Let x be a uniform random variable between 0 and 1. Then, drive an expression for probability density function of x^2.

10

8. In a Cross-Validation process of 3N data points, the data set is divided in to 2N data points for training and the remaining N points for testing (validation). In this process how much more effort would be required as compared to Leave-One-Out Cross-Validation strategy if we want to try all possible combinations in both cases.

10

9. In a regression analysis, how would the performance of a 5^{th} degree polynomial model compare with a linear model when the underlining data has a linear relation with large random error in data for the following two cases: The training data is a very large set, the training data is not very large set.