First Mid Term Examination, ME-781, September 12, 2017

Name: **Roll No:**

Total Time 2 hours; Total Marks 150

Open notes (self hand-written) examination.

10 1. Let a linear regression model of the type

$Y = \beta_0 + \beta_1 X + \varepsilon_1$

approximate the true relation between X and Y. Then

a.) Derive and compare the relation between the coefficient for the following two models:

$$Y = \beta_0 + \beta_1 X$$

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2$$

b.) Which would have more residual sum of squares (RSS) and test Mean Squared Error (MSE) and why.

20 2. Provider brief reasoning to show whether the k[th] nearest neighbor regression would perform better, similar or worst when compared to a linear regression model for the following cases:

a.) The underlying true model being linear and the test data is very sparse with small random error

b.) The underlying true model being linear and extensive test data is available with small random error

c.) The underlying true model being linear and extensive test data is available with large random error

d.) The underlying true model is non-linear and extensive test data is available with small random error

Assume the best possible value of k is chosen in the k[th] nearest neighbor regression and the random error has zero mean and is not a function of predictor.

15 3. The test MSE is a function of variance and bias of the fitting model $\hat{f}(x_0)$ and the variance of the error in data. Please state as to what would happen to variance and bias of the fitting model (with a brief schematic pictorial reason) when:

a.) A polynomial model with increasing degrees is fitted to an underlining linear model with small error in data.

b.) A polynomial model with increasing degrees is fitted to an underlining linear model with large error in data.

c.) A polynomial model with increasing degrees is fitted to an underlining cubic model with small error in data.

15 4. Suppose we have a data set with five predictors, X1 = GPA,X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ (i.e. X4 is the non-linear term which is the product of X1 and X2) , and X5 = Interaction between GPA and Gender (i.e. X5 is the non-linear term which is the product of X1 and X3). The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get

$\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$

a.) Which answer is correct, and why?

i.) For a fixed value of IQ and GPA, males earn more on average than females.

ii.) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

b.) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

c.) State whether true or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

(Please note that the range of GPA is from 0 to 4)

10 5. The probability of an accident of an automobile is a function of the age of the car, the number of kilometers it has travelled, the gender of the driver and the type of vehicle (there are four types of automobiles: Compact-Car, Sedan-Car, Sports-Car and luxury-Car). Then write the form for the expression for the simplest model to predict the probability of accident.

10 6. Multiple linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

With

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.

10 7. Which data scale do the following variables belong to (provide a brief reason):

a.) Name of a person
b.) Cell number of a person
c.) Marks obtained in an exam
d.) Ranking of a student in his/her class
e.) Rate of change of fever of a hospitalized patient (in deg C per minute)

20 8. Let set A={0,1,2} and set B={-1,0,1}. Then draw the following sets:

a.) C = A x B
b.) D=A $\oplus$ b(0,0.5)
c.) $D^{cl} - D^{in}$
d.) Smallest convex set which contains D

10 9. In a random health checkup a person is tested positive for a rare disease. Less than 1 person in million in the world have this disease. However, the medical test which was conducted is very accurate with 99.9% accuracy. What is the probability that the person has this disease.

15 10. A fair dice has numberings from 1 to 6. Random events are elements of the σ-algebra $\mathcal{F}$.

a.) Write the smallest σ-algebra for the above probability space.
b.) Write the smallest σ-algebra which contains an event that the number on the dice is a prime.
c.) Calculate the probability of each of the random event contained in the above two σ-algebras.

15 11. Show whether the following are dissimilarity measures (in a 2D space) or not:

a.) $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
b.) $\sqrt{|(x_1 - x_2).(y_1 - y_2)|}$
c.) $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} + \sqrt{|(x_1 - x_2).(y_1 - y_2)|}$
d.)