Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2019-2020
Comprehensive Examination (Regular)

Course No.	: PCAM* ZC211
Course Title	: REGRESSION
Nature of Exam	: Closed Book
Weightage	: 40%
Duration	: 3 Hours
Date of Exam	: **02/11/2019 (FN)**

No. of Pages = 1
No. of Questions = 7

Q1. Suppose you asked to build a uni-variate regression model with 'N' training points. You are asked to fit a polynomial of degree 2 by minimizing the sum of squares of errors of training data points. [3 + 3 + 4 Marks]

(a) Show that the error function is convex function.

Sol:

Let $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$ be 'N' training data points.

Let $y = w_0 + w_1 x + w_2 x^2$ be the polynomial of degree 2.

$$E(w_0, w_1, w_2) = \frac{1}{2}\sum_{n=1}^{N} \left[(w_0 + w_1 x_n + w_2 x_n^2) - t_n \right]^2$$

$$\nabla E(w_0, w_1, w_2) = \begin{bmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \\ \frac{\partial E}{\partial w_2} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{n=1}^{N} \left[(w_0 + w_1 x_n + w_2 x_n^2) - t_n \right] \\ \sum_{n=1}^{N} \left[(w_0 + w_1 x_n + w_2 x_n^2) - t_n \right] x_n \\ \sum_{n=1}^{N} \left[(w_0 + w_1 x_n + w_2 x_n^2) - t_n \right] x_n^2 \end{bmatrix}$$

$$H = \begin{bmatrix} \frac{\partial^2 E}{\partial w_0^2} & \frac{\partial^2 E}{\partial w_0 \partial w_1} & \frac{\partial^2 E}{\partial w_0 \partial w_2} \\ \frac{\partial^2 E}{\partial w_1 \partial w_0} & \frac{\partial^2 E}{\partial w_1^2} & \frac{\partial^2 E}{\partial w_1 \partial w_2} \\ \frac{\partial^2 E}{\partial w_0 \partial w_2} & \frac{\partial^2 E}{\partial w_1 \partial w_2} & \frac{\partial^2 E}{\partial w_2^2} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{n=1}^{N} 1 & \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 \\ \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 & \sum_{n=1}^{N} x_n^4 \end{bmatrix}$$

~~We show~~

The principal minors of H are non-negative and hence H is positive semi-definite

~~Hence H is~~

Hence the error function is $E(w_0, w_1, w_2)$ is convex function.

(b) Find out the exact (not an approximate) regression model that minimizes error the least.
Hint: Gradient methods might bot help you here!

Sol:

Let $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$ be 'N'

training data points.

Let $y = w_0 + w_1 x + w_2 x^2$ be the polynomial

of degree 2.

$$E(w_0, w_1, w_2) = \frac{1}{2}\sum_{n=1}^{N}\left[\left(w_0 + w_1 x_n + w_2 x_n^2\right) - t_n\right]^2$$

$$\frac{\partial E}{\partial w_0} = 0 \Rightarrow \sum_{n=1}^{N}\left[\left(w_0 + w_1 x_n + w_2 x_n^2\right) - t_n\right] = 0$$

$$\Rightarrow w_0 \sum_{n=1}^{N} 1 + w_1 \sum_{n=1}^{N} x_n + w_2 \sum_{n=1}^{N} x_n^2 = \sum_{n=1}^{N} t_n \longrightarrow (1).$$

$$\frac{\partial E}{\partial w_1} = 0 \Rightarrow \sum_{n=1}^{N}\left[\left(w_0 + w_1 x_n + w_2 x_n^2\right) - t_n\right] x_n = 0$$

$$\Rightarrow w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{N} x_n^2 + w_2 \sum_{n=1}^{N} x_n^3 = \sum_{n=1}^{N} t_n x_n \longrightarrow (2).$$

$$\frac{\partial E}{\partial w_2} = 0 \Rightarrow \sum_{n=1}^{N}\left[\left(w_0 + w_1 x_n + w_2 x_n^2\right) - t_n\right] x_n^2 = 0$$

$$\Rightarrow w_0 \sum_{n=1}^{N} x_n^2 + w_1 \sum_{n=1}^{N} x_n^3 + w_2 \sum_{n=1}^{N} x_n^4 = \sum_{n=1}^{N} t_n x_n^2 \longrightarrow (3)$$

The equations ①, ②, ③ Can be written as follows:

$$N w_0 + \left(\sum_{n=1}^{N} x_n\right) w_1 + \left(\sum_{n=1}^{N} x_n^2\right) w_2 = \sum_{n=1}^{N} t_n$$

$$\left(\sum_{n=1}^{N} x_n\right) w_0 + \left(\sum_{n=1}^{N} x_n^2\right) w_1 + \left(\sum_{n=1}^{N} x_n^3\right) w_2 = \sum_{n=1}^{N} t_n x_n$$

$$\left(\sum_{n=1}^{N} x_n^2\right) w_0 + \left(\sum_{n=1}^{N} x_n^3\right) w_1 + \left(\sum_{n=1}^{N} x_n^4\right) w_2 = \sum_{n=1}^{N} t_n x_n^2$$

The above system of equations Can be written as $A w = b$ where

$$A = \begin{bmatrix} N & \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 \\ \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n^3 & \sum_{n=1}^{N} x_n^4 \end{bmatrix} \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \quad b = \begin{bmatrix} \sum_{n=1}^{N} t_n \\ \sum_{n=1}^{N} t_n x_n \\ \sum_{n=1}^{N} t_n x_n^2 \end{bmatrix}$$

Hence w can be obtained it A and b by making use of $w = A^{-1} b$. A^{-1}, b Can be Computed using the given data set.

(c) If you are asked fit to a polynomial of degree 100, then list out the practical issues that would be faced in the above procedure.

Sol:

If the polynomial of degree 100 is to be fit to the data points $(x_1, t_1), (x_2, t_2), ---, (x_N, t_N)$ then the polynomial is

$$y = w_0 + w_1 x + w_2 x^2 + --- + w_{100} x^{100}$$

The number of parameters in the polynomial is 101 and the parameters are $w_0, w_1, w_2, ---, w_{100}$. The set of normal equations to be solved to find parameters $w_0, w_1, ---, w_{100}$ will be

$Aw = b$ where A is 101×101 size.

$$w = A^{-1} b$$

we have to find the inverse of the matrix $A_{101 \times 101}$ which is a computationally involved task.

Q2. Build the following linear regression models for the data set given below [4 Marks]

(a) $y = w_0$

(b) $y = w_0 + w_1 x$

Data Set:

x	-1	0	2
y	1	-1	1

Sol:

x	-1	0	2
y	1	-1	1

(i) $y = w_0$.

$$E(w_0) = \frac{1}{2}\left[(w_0-1)^2 + (w_0+1)^2 + (w_0-1)^2\right]$$

$$= \frac{1}{2}\left[2(w_0-1)^2 + (w_0+1)^2\right]$$

$$= \frac{1}{2}\left[2(w_0^2+1-2w_0) + (w_0^2+1+2w_0)\right]$$

$$= \frac{1}{2}\left[3w_0^2 - 2w_0 + 3\right]$$

$$\frac{\partial E}{\partial w_0} = 0 \Rightarrow 6w_0 - 2 = 0 \Rightarrow w_0 = \frac{2}{6}$$

$$\Rightarrow w_0 = \frac{1}{3}$$

$y = \frac{1}{3}$ is the linear regression model (with degree 0) fitting the given data.

(ii) $y = w_0 + w_1 x$

$$E(w_0, w_1) = \frac{1}{2}\Big[\big((w_0 + w_1(-1)) - 1\big)^2 + \big((w_0 + w_1(0)) + 1\big)^2 + \big((w_0 + 2w_1) - 1\big)^2\Big].$$

$$= \frac{1}{2}\left[(w_0 - w_1 - 1)^2 + (w_0 + 1)^2 + (w_0 + 2w_1 - 1)^2\right]$$

$$= \frac{1}{2}\left[\left(w_0^2 + w_1^2 + 1 - 2w_0w_1 + 2w_1 - 2w_0\right) + \left(w_0^2 + 1 + 2w_0\right) + \left[w_0^2 + 4w_1^2 + 1 + 4w_0w_1 - 4w_1 - 2w_0\right]\right]$$

$$= \frac{1}{2}\left[3w_0^2 + 5w_1^2 + 2w_0w_1 - 2w_1 - 2w_0 + 3\right]$$

$$\frac{\partial E}{\partial w_0} = 0 \Rightarrow 6w_0 + 2w_1 - 2 = 0$$

$$\Rightarrow 6w_0 + 2w_1 = 2$$

$$\Rightarrow 3w_0 + w_1 = 1 \rightarrow (1).$$

$$\frac{\partial E}{\partial w_1} = 0 \Rightarrow 10w_1 + 2w_0 - 2 = 0$$

$$\Rightarrow 5w_1 + w_0 = 1 \rightarrow (2).$$

$$15w_0 + 5w_1 = 5 \quad \text{(by multiplying (1) by 5)}.$$

$$-14w_0 = -4 \Rightarrow w_0 = \frac{4}{14}$$

$$\Rightarrow w_0 = \frac{2}{7}$$

$$3\left(\frac{2}{7}\right) + w_1 = 1 \Rightarrow w_1 = 1 - \frac{6}{7} \Rightarrow w_1 = \frac{1}{7}$$

Hence the best fit linear regression model of degree 1 for the given data is

$$y = \frac{2}{7} + \frac{1}{7}x$$

i.e., $7y = 2 + x$

Q3. Write down the steps to find R^2 value in single variate linear regression. Find R2 value for for the problem in the above question i.e., Q2. [4 Marks]

Sol:

Steps to find R^2 value in single variate linear regression.

Let $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ be 'N' training examples.

Let $y = \hat{w}_0 + \hat{w}_1 x$ be the best fit linear regression model for the given data.

$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N$ be the 'N' predicted values of $x_1, x_2, \dots, x_N$. [i.e., $\hat{y}_n = \hat{w}_0 + \hat{w}_1 x_n$]

(i) Find the average of $y_1, y_2, \dots, y_N$

$$\bar{y} = \frac{\sum_{n=1}^{N} y_n}{N}$$

(ii) Find the total sum of squares from the training data.

$$SST = \sum_{n=1}^{N} (y_n - \bar{y})^2$$

(iii) Find the sum of squares of errors/residuals of the training data

$$SSE = \sum_{n=1}^{N} (\hat{y}_n - y_n)^2$$

(iv) $R^2 = 1 - \frac{SSE}{SST}$

Let us find out the R^2 values for the linear regression model

$$y = \frac{2}{7} + \frac{1}{7}x$$

$$\bar{y} = \frac{1-1+1}{3} = \frac{1}{3}$$

The predicted values of $x = -1, 0, 1$ are as follows:

x	-1	0	2
y	1	-1	1
$\hat{y}$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{4}{7}$

$$SST = \sum_{n=1}^{3} (y_n - \bar{y})^2$$

$$= \left(1 - \frac{1}{3}\right)^2 + \left(-1 - \frac{1}{3}\right)^2 + \left(1 - \frac{1}{3}\right)^2$$

$$= \left(\frac{2}{3}\right)^2 + \left(-\frac{4}{3}\right)^2 + \left(\frac{2}{3}\right)^2$$

$$= \frac{4 + 16 + 4}{9} = \frac{24}{9} = \frac{8}{3}$$

$$SSE = \sum_{n=1}^{3} (y_n - \hat{y}_n)^2$$

$$= \left(1 - \frac{1}{7}\right)^2 + \left(-1 - \frac{2}{7}\right)^2 + \left(1 - \frac{4}{7}\right)^2$$

$$= \left(\frac{6}{7}\right)^2 + \left(-\frac{9}{7}\right)^2 + \left(\frac{3}{7}\right)^2$$

$$= \frac{36 + 81 + 9}{49} = \frac{\overset{18}{\cancel{126}}}{\underset{7}{\cancel{49}}}$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\frac{18}{7}}{\frac{8}{3}}$$

$$= 1 - \frac{\overset{9}{\cancel{18}} \times 3}{7 \times \underset{4}{\cancel{8}}} = 1 - \frac{27}{28}$$

$$= \frac{1}{28} = 0.036.$$

Q4. Can R^2 value be 1? If so, provide an example for which R^2 is 1. [4 Marks].

Sol: Yes, R^2 can be 1.

Suppose dependent variable, y, is linearly related to the independent variable, x, and in reality they are related by y = 500 + 4x. If the training data set contains "N' data points they are on the true line i.e., y = 500 + 4x. Equivalently speaking there is no noise in the data points and they are taken from the actual linear relationship between x and y. Then R^2 will be 1.

Q5. What are the two techniques to implement regularization for polynomial fitting? What is the difference between these two techniques? Explain the two techniques with all mathematical rigor. [6 Marks]

Sol:

Regularization is to limit the growth of the parameters $w_0, w_1, \ldots$ to combat Overfitting problem in polynomial regression. There are two techniques to implement regularization ~~(1) Ridge~~ and they are Ridge and Lasso regression.

Ridge Regression:

The error function is

$$\min \frac{1}{2}\left[\sum_{n=1}^{N}\left\{\left(w_0 + w_1 x_n + \cdots + w_D x_n^D\right) - t_n\right\}^2\right]$$

$$\text{s.t.} \sum_{d=0}^{D} w_d^2 \leq M$$

where $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$ are 'N' data points

and $y(x, w) = w_0 + w_1 x + \cdots + w_D x^D$ is the polynomial of degree D

The corresponding unconstrained optimization problem is

$$\min \frac{1}{2}\left[\sum_{n=1}^{N}\left\{(w_0+w_1x_n+\cdots+w_Dx_n^D)-t_n\right\}^2\right] + \lambda\sum_{d=0}^{D} w_d^2$$

Lasso Regression:

The error function is

$$\frac{1}{2}\left[\sum_{n=1}^{N}\left\{(w_0+w_1x_n+\cdots+w_Dx_n^D)-t_n\right\}^2\right]$$

with the constraint that

$$\sum_{d=0}^{D}|w_d| \leq M$$

The corresponding unconstrained optimization problem is

$$\min \frac{1}{2}\sum_{n=1}^{N}\left\{(w_0+w_1x_n+\cdots+w_Dx_n^D)-t_n\right\}^2 + \lambda\sum_{d=0}^{D}|w_d|$$

The solution to the ridge regression are real valued i.e., $w_0, w_1, \ldots, w_D$ takes real numbered values where as in lasso regression they are integer valued.

Q6. Do you agree that forward or backward stepwise selection algorithm guarantees the best optimal solution? If so, prove it? Otherwise what the issues are there in figuring out the best feature subset? [6 Marks]

Sol: No. The forward or backward stepwise selection algorithm will not guarantee optimal solution because they are heuristics based on greedy approach. In general it is difficult to find the optimal subset as there are 2^D feasible solutions (or equivalently 2^D subsets). This is a computationally heavy task and hence exact optimal solution van not be found in polynomial time.

Q7. Suppose you are a machine learning consultant and are given census data of 1,00,00,000 people containing 20 features to predict mortality. After doing few basic experiments, you decided to go ahead with regression. Discuss how you finalize the degree of the polynomial that you will be fitting for linear regression (whether you will be fitting linear, quadratic, cubic, curve to model the data). [6 Marks]

Sol: Divide Data - 60% Training Data, 20% Validation Data, 20% Testing Data.

Build regression models (degree of polynomial 1,2,3,4,…) with 60% of training data. Find Validation errors of each model with 20% of validation error. Select regression model with polynomial degree k if its validation is the lowest among all other models.