Data Science & Artificial Intelligence

Machine Learning

Regression

DPP-01 **Discussion Notes**

By- Astha Ma'am

[MCQ]

#Q. The parameters acquired through linear regression:

[✓ A] can take any value in the real space

B are strictly integers

C always lie in the range [0,1]

D can take only non-zero values

→ Real space

Model parameters

Real Values

+ve

0

-ve

[MCQ]

#Q. Which of the statements is/are True ?

A Ridge has sparsity constraint, and it will drive coefficients with low values to 0.

B Lasso has a closed form solution for the optimization problem, but this is not the case for Ridge.

C Ridge regression does not reduce the number of variables since it never leads a coefficient to zero but only minimizes it.

D If there are two or more highly collinear variables, Lasso will select one of them randomly

[MCQ]

#Q. The relation between studying time (in hours) and grade on the final examination (0-100) in a random sample of students in the Introduction to Machine Learning Class was found to be : Grade = 30.5 + 15.2 (h) How will a student's grade be affected if she studies for four hours ?

A It will go down by 30.4 points.

B It will go down by 30.4 points.

[✓ C] It will go up by 60.8 points.

D The grade will remain unchanged.

E It cannot be determined from the information given

h

h = 4

$30.5 + 15.2 \times (4)$

Grade = 91.3

C

[MCQ]

#Q. Which of the following statements about principal components in Principal Component Regression (PCR) is true ?

A Principal components are calculated based on the correlation matrix of the original predictors.

B The first principal component explains the largest proportion of the variation in the dependent variable.

C Principal components are linear combinations of the original predictors that are uncorrelated with each other.

D PCR selects the principal components with the highest p-values for inclusion in the regression model.

E PCR always results in a lower model complexity compared to ordinary least squares regression.

[MCQ]

#Q. Which statement is true about outliers in Linear regression ?

- A Linear regression model is ~~not~~ sensitive to outliers
- [✓ B] Linear regression model is sensitive to outliers
- C Can't say
- D None of these

[MCQ]

#Q. What does the slope coefficient in a linear regression model indicate ?

A The point where the regression line intersects the y-axis

[✓ B] The dependent variable changes for every one-unit change in the independent variable

C The average value of the dependent variable

D The dispersion of the dependent variable

Y

x

Value

one unit

[MCQ]

#Q. Find the mean of squared error for the given predications :

Y	F(X)
1	2
2	3
4	5
8	9
16	15
32	31

1 $(2-1)^2$

2 $(3-2)^2$

3 $(5-4)^2$

4 $(9-8)^2$

8

6 $(15-16)^2 = 1$

$(31-32)^2 = 1$

Hind : Find the squared error for each predication and take the mean of that.

[✓ A 1]

B 2

C 1.5

D 0

$[1 + \dots + 1]$

$= 6/6 = 1$

[MCQ]

#Q. What is the primary assumption of linear regression regarding the relationship between the independent and dependent variables?

A Non-linearity

B Independence of errors

C Homoscedasticity

[✓ **D** Linearity]

Linear Relationship

b/w th $\underline{\underline{X}}$ (Predictors)

(outcomes)

[MCQ]

#Q. Which of the following statements is true regarding Partial Least Squares (PLS) regression ?

A PLS is a dimensionality reduction technique that maximizes the covariance between the predictors and the dependent variable.

B PLS is only applicable when there is no multicollinearity among the independent variables.

[✓ C] PLS can handle situations where the number of predictors is larger than the number of observations.

D PLS estimates the regression coefficients by minimizing the residual sum of squares.

E PLS is based on the assumption of normally distributed ~~residuals~~.

Reduction is

dimensionality

F All of the above.

G None of the above.

[MCQ]

#Q. The confidence interval is an interval which is an estimate of -

[✓ A] The mean value of the dependent variable →

B The standard deviation value of the dependent variable

C The mean value of the independent variable

D The standard deviation value of the independent variable

Range within which true mean of the depend. variable likely fall

[MCQ]

#Q. In the regression model ($y = a + bx$) where $x = 2.50$, $y = 5.50$ and $a = 1.50$ (x and y denote mean of variables x and y and a is a constant), which one of the following values of parameter 'b' of the model is correct ?

A 1.75

[✓ B] 1.60

C 2.00

D 2.50

$\underline{y = a + bx}$

$x = 2.50$

$y = 5.50$

$a = 1.50$

$5.50 = 1.50 \times b \times 2.50$

$\boxed{b = 1.60}$

[NAT]

#Q. There is no value of x that can simultaneously satisfy both the given equations. Therefore, find the 'least squares error' solution to the two equations, i.e ., find the value of x that minimize the sum of squares of the errors in the two equations. ___________.

$2x = 3$

$4x = 1$

$2x = 3$

$4x = 1$

$2x = 3$

$\boxed{2x - 3 = 0 = 0}$

$4x = 1$

$\boxed{4x - 1 = 0}$

$\frac{dR}{dx} = 0$

$\Rightarrow 2\times2(2x-3)+4\times2 \times(4x-1)$

$= \cdot 12$

[MCQ]

#Q. For a bivariate data set on (x, y), if the means, standard deviations and correlation coefficient are x = 1.0, y = 2.0, sx = 3.0, sy = 9.0, r = 0.8.

Then the regression line of y on x is:

A y = 1 + 2.4 (x – 1)

B y = 2 + 0.27 (x – 1)

C y = 2 + 2.4 (x – 1)

D y = 1 + 0.27 (x – 1)

$y - 2 = 0.8 \times 9(x-1)/3$

$y-2 \quad 2.4 (x-1)$

$y = 2 + 2.4 (x-1$

[MCQ]

#Q. What is the purpose of regularization in linear regression ?

- **A** To make the model more complex
- **B** ~~To avoid~~ underfitting
- **C** To encourage overfitting ✗
- **D** To reduce the complexity of the model ✓

underfitting

→ more complex

[NAT]

#Q. A set of observations of independent variable (x) and the corresponding dependent variable (y) is given below :

X	5	2	4	3
Y	16	10	13	12

$\Sigma x = 14$

$\Sigma y = 51$

$y = a + b x$

$51 = 4a + 54b$

Based on the data, the coefficient a of the linear regression model.

y = a + bx is estimated as 6.1

$a = 6.1$

The coefficient b is __________ .(round off to one decimal place)

$b\ 1.9$

x	y	x^2	xy
5	16	25	80
2	10	4	20
4	13	16	52
3	12	4	36

$xy \rightarrow \Sigma xy = 188$

$\Sigma x^2 = 54$

[MCQ]

#Q. The purpose of using a dummy variable in the regression model is ?

[✓ A] Some of the independent variables are categorical data

B The dependent variable is categorical data

C Both independent and dependent variable may have a categorical data value

D The dependent and independent variable must be a numerical data value

group

[MCQ]

#Q. The random error ε in multiple linear regression model $y = X\beta + \varepsilon$ are assumed to be identically and independently distributed following the normal distribution with zero mean and constant variance. Here y is a $n \times 1$ vectors of observations on response variable, X is a $n \times K$ matrix of n observations on each of the K explanatory variables, β is a $K \times 1$ vectors of regression coefficients and ε is a $n \times 1$ vectors of random errors. The residuals $\hat{\varepsilon}=y-\hat{y}$ based on the ordinary least squares estimator of β have , in general.

A Zero mean, constant variance and are independent

B Zero mean, constant variance and are not independent

[✓ C] Zero mean, non constant variance and are not independent

D non Zero mean, non constant variance and are not independent

[MCQ]

#Q. A residual is defined as ?

A The difference between the actual Y values and the mean of Y.

[✓ B] The difference between the actual Y values and the predicted Y values.

C The predicted value of Y for the average X value.

D The square root of the slope.

[MCQ]

#Q. The linear regression model y=a0+a1x1+a2x2+...+apxp is to be fitted to a set of N training data points having p attributes each. Let X be Nx(p+1) vectors of input values (augmented by 1's), Y be Nx1 vector of target values, and theta (0) be (p+1)×1 vector of parameter values (a0, a1, a2,...,ap). If the sum squared error is minimized for obtaining the optimal regression model, which of the following equation holds ?

A XTX=XY

B XO=XTY

C XTXO=Y

[✓ D] XTX0=XTY

XTx0 = XTy

x = Nx(p+1)

[NAT]

#Q. Use the regression equation to predict the glucose level given the age. Consider the following is the data set for understanding the concept of Linear Regression Numerical Example with One Independent Variable.

SUBJECT	AGE X	GLUSCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	?

$$b_0 = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma y}{n \Sigma x^2 (\Sigma x)^2}$$

$$b_1 = \boxed{0.385}$$

[MCQ]

#Q. In linear regression, what is the primary difference between Lasso (L1 regularization) and Ridge (L2 regularization) ?

[✓ **A** Lasso tends to produce sparse coefficient vectors, while Ridge does not.]

B Ridge tends to produce sparse coefficient vectors, while Lasso does not.

C Both Lasso and Ridge produce sparse coefficient vectors.

∝

D Both Lasso and Ridge tend to produce non-sparse coefficient vectors.

∝

Spare coefficient - ∝

[NAT]

#Q. Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}

(a) Find the least square regression line for the given data points.

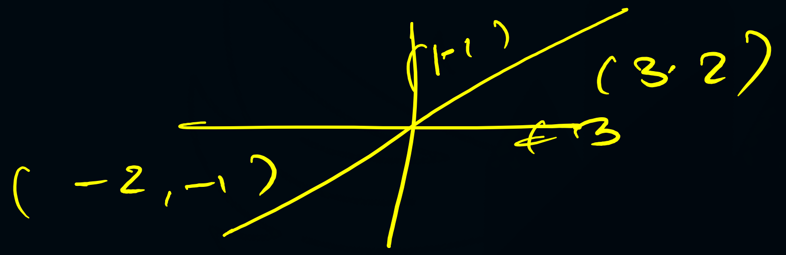
(b) Plot the given points and the regression line in the same rectangular system of axes.

x	y	xy	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9

$x = 2$

$x^2 = 14$

$y = ax + b$



[NAT]

#Q. In the table below, the xi column shows scores on the aptitude test. Similarly, the yi column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each measurement. The last two rows show sums and mean scores.

Find the regression equation

Student	x_i	y_i	$\left(x_i - \bar{x}\right)^2$	$\left(y_i - \bar{y}\right)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

$b_1 = 2(x_1 - \bar{y}$

$/ n(y_1 - y^2$

[MCQ]

#Q. What does $(x^{(5)}, y^{(5)})$ represent or imply ?

0

1

A There are 5 training examples

B The values of x and y are 5

$x^n \quad y^n$

C The fourth training examples

D [✓ The fifth training example.]

[NAT]

#Q. The values of y and their corresponding values of y are shown in the table below.

X	0	1	2	3	4
y	2	3	5	4	6

$= \Sigma x \quad 10$

$\Sigma y = 20$

Σx^2

$\Sigma xy^2 =$

(a) Find the least square regression line y = a x + b.

(b) Estimate the value of y when x = 10.

$$\boxed{y = 0.9 \times 10 + 2.2 = 11.2}$$

[MCQ]

#Q. In the context of linear regression, what is the purpose of the F - test ?

A To determine the significance of individual coefficients.

[✓ B] To test the overall significance of the regression model. Compare

C To assess the presence of multicollinearity among independent variables.

D To evaluate the normality of residuals.

[MCQ]

#Q. When performing linear regression, multicollinearity can be problematic. Which of the following statements about multicollinearity is true ?

A Multicollinearity occurs when there is no correlation between independent variables.

B Multicollinearity makes it easier to interpret the individual coefficients in the regression model.

C Multicollinearity inflates the standard errors of the regression coefficients.

D Multicollinearity always improves the predictive performance of the model.

[MCQ]

#Q. What is heteroscedasticity, and how does it affect the assumptions of linear regression?

A Heteroscedasticity refers to the presence of outliers in the dataset, violating the assumption of linearity

[✓ B] Heteroscedasticity refers to the non-constant variance of residuals, violating the assumption of homoscedasticity

C Heteroscedasticity occurs when there is perfect multicollinearity among independent variables, violating the assumption of independence.

D Heteroscedasticity refers to the presence of correlated errors, violating the assumption of normality.

[MCQ]

#Q. When should one prefer ridge regression over lasso regression ?

A When the goal is to select a subset of important predictors.

B When the coefficients of irrelevant predictors should be exactly zero.

[✓ C] When there is multicollinearity among the independent variables.

D When the dataset has a large number of observations.

x not- be deserted

THANK - YOU