Data Science & Artificial Intelligence

Machine Learning

Regression

DPP-01

Discussion Notes

By - Siddharth Sabharwal Sir

[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 ?

Sol $G = 30.5 + 15.2h$

Studied 4 hours

G will be effected as

$15.2 \times 4 = 60.8$

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

[MCQ]

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

Sol $y = mx + c$ (1D data)

m → Slope

$m = \frac{dy}{dx} \Rightarrow dy = m\,dx \Rightarrow$ if $dx = 1$ i.e change in independent variable $= 1$ then $dy = m$

b)

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

[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	32

Sol y: actual value

f(x): Predicted value

error y-y'	error²
-1	1
-1	1
-1	1
-1	1
1	1
0	0

Mean of error²

$\Rightarrow \frac{1+1+1+1+1+0}{6}$

$\Rightarrow 5/6$

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

A. 1

B. 5

C. 1.5

D. 5/6 ✓

[MCQ]

#Q. In the regression model ($y = a + bx$) where $\bar{x} = 2.50$, $\bar{y} = 5.50$ 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

Sol We know that

$y = a + bx$

So $\bar{y} = a + b\bar{x}$

$b = \frac{\bar{y} - a}{\bar{x}} \Rightarrow \frac{5.5 - 1.5}{2.5} = 1.6$

[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

Based on the data, the coefficient a of the linear regression model, y = a + bx is estimated as 6.1

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

• Sol $a = 6.1$, we know that $\bar{y} = a + b\bar{x}$

$b = \frac{\bar{y} - a}{\bar{x}}$

$b = 1.9$

$\bar{x} = 7/2$

$\bar{y} = 51/4$

[MCQ]

#Q. The linear regression model y=a0+a1×1+a2×2+...+ap×p is to be fitted to a set of N training data points having p attributes each. Let X be N× (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 ?

Sol: done in class

$(X^T X A = X^T Y)$

A. XTX = XY

B. XA = XTY

C. XTXA = Y

D. XTXA = XTY

[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	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	?

Sol So here we have to Create model

- $y = mx + c$

$$X = \begin{bmatrix} 1 & 43 \\ 1 & 21 \\ 1 & 25 \\ 1 & 42 \\ 1 & 57 \\ 1 & 59 \end{bmatrix} \quad X^TX = \begin{bmatrix} 6 & 247 \\ 247 & 11409 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} 486 \\ 20485 \end{bmatrix}$$

$$(X^TX)^{-1} = \begin{bmatrix} 11409 & -247 \\ -247 & 6 \end{bmatrix} \frac{1}{7445}$$

$$\begin{bmatrix} c \\ m \end{bmatrix} = (X^TX)^{-1}[X^TY] = \begin{bmatrix} 11409 & -247 \\ -247 & 6 \end{bmatrix} \begin{bmatrix} 486 \\ 20485 \end{bmatrix} \frac{1}{7445}$$

$$\begin{bmatrix} 484979 \\ 2868 \end{bmatrix} \frac{1}{7445}$$

$$\Rightarrow c = 65.14$$
$$m = .3852$$

So @ $x = 55$, $y = mx + c \Rightarrow \underline{\underline{86.32}}$

[MCQ]

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

Sol

- Lasso → Sparse sol
  ↓
  Because it tend to make more coefficients β's to zero.

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.

[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.

Sol

$$\left.\begin{array}{ccc} x & -2 & 1 & 3 \\ y & -1 & 1 & 2 \end{array}\right\} \text{ model } y = mx + c$$

$$\text{So } X = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \qquad X^TX = \begin{bmatrix} 3 & 2 \\ 2 & 14 \end{bmatrix} \qquad (X^TX)^{-1} = \begin{bmatrix} 14 & -2 \\ -2 & 3 \end{bmatrix} \frac{1}{38}$$

$$X^TY = \begin{bmatrix} 2 \\ 9 \end{bmatrix}, \qquad \begin{bmatrix} c \\ m \end{bmatrix} = (X^TX)^{-1}(X^TY)$$

$$= \begin{bmatrix} 19/38 \\ 23/38 \end{bmatrix}$$

$C = 10/38$

$m = 23/38$

$y = mx + c$

$\Rightarrow y = .605x + .2631$

[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	xi	yi	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
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 ✓		

Sol $y = mx + C$

$m = \frac{Cov(x,y)}{\sigma_x^2}$

$Cov(x,y) = \frac{\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})}{5}$

$\Rightarrow 17\times8 + 7\times18 + 2\times(-7)$
$-8\times-12 + -18\times-7/5$

$$Cov = \frac{470}{5} = 94$$

$$m = \frac{Cov\ of\ x, y}{Var\ of\ x} = \frac{94}{\underbrace{\Sigma(x_i - \bar{x})^2}_{5}} = \frac{94}{730/5} = .6438$$

$$c = \bar{y} - m\bar{x} \Rightarrow 77 - .6438 \times 78$$

$$\Rightarrow 26.78$$

So $(y = .6438x + 26.78)$

[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

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

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

Sol $X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \Rightarrow X^T X = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}$ $(X^T X)^{-1} = \frac{1}{50}\begin{bmatrix} 30 & -10 \\ -10 & 5 \end{bmatrix}$

$X^T Y = \begin{bmatrix} 20 \\ 49 \end{bmatrix}$

$y = ax + b$ is model

$\begin{bmatrix} b \\ a \end{bmatrix} = (X^T X)^{-1}(X^T Y) = \begin{bmatrix} 110 \\ 45 \end{bmatrix}\frac{1}{50}$

$b = 110/50$

$a = 45/50$

$y = .9x + 2.2$

@ $x = 10$

$y = 9 + 2.2 = 11.2$ ✓

[MCQ]

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

Sol due to Multicollinearity model become poor.

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. Suppose you have m = 14 training examples with n = 3 features (excluding the additional all-ones feature for the intercept term, which you should add), the normal equation is $\theta = (X^TX)^{-1}X^Ty$. For the given values of m and n, what are the dimensions of θ, X and y in this equation?

A. X is 14 × 4, y is 14 × 1, θ is 4 × 1 ✓

B. X is 14 × 3, y is 14 × 1. θ is 3 × 3

C. X is 14 × 4, y is 14 × 4, θ is 4 × 4

D. X is 14 × 3, y is 14 × 1, θ is 3 × 1

• Sol $X =$ 14 Rows, 4 Column — 14x4

$Y =$ 14 Rows, 1 Column — 14x1

$\theta = \theta_0 \longrightarrow \theta_3$

$\Rightarrow$ 4 Row, 1 Column — 4x1

[MCQ]

#Q. The closed form solution for linear regression is $\theta = (X^TX)^{-1}X^Ty$. Suppose you have N = 35 training examples and M = 5 features (excluding the bias term). Once the bias term is now included, what are the dimensions of X, y, θ in the closed form equation?

A. X is 35×6, y is 35×1, θ is 6×1

B. X is 35×6, y is 35×6, θ is 6×1

C. X is 35×5, y is 35×1, θ is 5×1

D. X is 35×5, y is 35×5, θ is 5×5

[NAT]

#Q. Suppose the score of the midterm, homework, and final of 4 students are given by

Midterm	Homework	Final
80	85	96
72	68	74
94	97	90
69	70	78

Sol we have to predict

$y \rightarrow$ final score

using mid term & Homework

score $\rightarrow$ 2 dimensions

x^1, x^2

You'd like to use linear regression to predict a student's final exam score from his/her midterm exam score and homework score. The normal equation of linear regression is $\theta = (X^TX)^{-1}X^Ty$. What are X and y in this problem.

Final score $\rightarrow y$, mid term Home work are x

[NAT]

#Q. Assume that you have a dataset composed of N observations: the target t and features X. You want to fit a linear regression model and find the weights w, but you also know that more data is always helpful. Instead of fitting a model with $t \in R^n$ and $X \in R^{nxd}$, you concatenate the data fit a model using $\begin{bmatrix} t \\ t \end{bmatrix} \in R^{2n}$ and $\begin{bmatrix} X \\ X \end{bmatrix} \in R^{2n \times d}$. Running linear regression on this new dataset will give the same weights as on the original dataset.

Sol

data

P_1	x_1	Y_1
P_2	x_2	Y_2

$$\text{loss fxn} = \sum_{i=1}^{2} (y_i - y_i')^2$$

Now new data

data

P_1	x_1	y_1
P_2	x_2	y_2
P_3	x_1	y_1
P_4	x_2	y_2

Now $\alpha = \sum_{i=1}^{4} (y_i - y_i')^2$

$$= 2 \sum_{i=1}^{2} (y_i - y_i')^2$$

Since loss fxn in both case is same (only multiplied by 2 in case 2) the result to minimize L will be same.

Yes the model will be same.

[MCQ]

#Q. First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.

Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors?

A. Mean Training Error: Increase
Mean Testing Error: Decrease

B. Mean Training Error: Increase
Mean Testing Error: Increase

C. Mean Training Error: Decrease
Mean Testing Error: Increase

D. Mean Training Error: Decrease
Mean Testing Error: Decrease

Sol data : 1000 points

Test data Training data

Now if we inc the Training data →
what will happen

* our target is to find pattern of data,
if we have larger data then we get better fit.

* when training data is very small then model
over fit,

Concept

Training data very small

→ over fit model

→ Training error = 0

→ Testing error is high

Training data inc

→ model generalize better

→ Training error ≠ 0

→ Testing error reduce

[NAT]

#Q. Suppose now that Bob records each feature in a different unit; that is, for each coordinate $z_j^{(i)} = cjx_j^{(i)}$ for all i, where $c_j > 0$ is a scalar, and the c_j's are not all equal. Do Alice and Bob still have the same training loss and is $w_{Alice} = W_{Bob}$? If yes, justify your answer. If no, provide an example of a dataset where this is not the case.

Sol 1D data

X	Y
x_1	y_1
x_2	y_2
⋮	
x_N	y_N

model
$y = mx + c$

modify data
1D data

X	Y
$c_1 x_1$	y_1
$c_1 x_2$	y_2
⋮	
$c_1 x_N$	y_N

- Since y value has not changed but x has changed, so for same y we need to update the coeff "m" to get same y for new x

$w_{Alice} \neq w_{Bob}$.

THANK - YOU