Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 04

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** Representation of data
- **Topic** 2D data
- **Topic** final expression of B.
- **Topic**
- **Topic**

Topics to be Covered

- **Topic** R^2 Factor
- **Topic** MSE
- **Topic** Gradient descent.
- **Topic**
- **Topic**

How to do it

don't create your limit

don't think you will be able to do it/not

Basics of Machine Learning

How the data is represented in matrix format

N points, D dimension

$$X_{matrix} = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^D \\ 1 \\ 1 \\ \vdots \end{bmatrix}_{N\times(D+1)} \begin{matrix} \leftarrow 1^{st} \text{ data} \\ \leftarrow 2^{nd} \text{ data} \\ \vdots \\ \\ \end{matrix}$$

Each Row show a datapoint

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}_{D+1 \times 1}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1} ✓$$

REVISION

The final expression of Beta matrix..

So $\beta = (X^T X)^{-1} (X^T Y)$

If we have any data point

$$\left(1 \;\; x_i^1 \;\; x_i^2 \;\; x_i^3 \; \cdots \; x_i^D\right)$$

$\Rightarrow$ So Predicted Value $\Rightarrow$ $\left(\beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots \beta_D x^D\right)$

$$\Rightarrow X\beta \Rightarrow \begin{bmatrix} 1 & x_1^1 & x_1^2 & x_1^3 & \cdots & x_1^D \\ 1 & x_2^1 & x_2^2 & x_2^3 & \cdots & x_2^D \\ \vdots & & & & & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_D \end{bmatrix} \Rightarrow \begin{bmatrix} \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 \cdots \beta_D x_1^D \\ \beta_0 + \beta_1 x_2^1 + \beta_2 x_2^2 \cdots \beta_D x_2^D \\ \vdots \end{bmatrix}$$

$X\beta \Rightarrow$ we get $\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \vdots \\ \hat{y}_N \end{bmatrix}$

So $Y = \begin{bmatrix} y_1 \checkmark \\ y_2 \checkmark \\ \vdots \\ y_N \end{bmatrix} \Leftarrow$ actualy values given in data

$X\beta \Rightarrow \hat{Y} \begin{bmatrix} \hat{y}_1 \checkmark \\ \hat{y}_2 \checkmark \\ \vdots \\ \hat{y}_N \end{bmatrix}$

- Training data

- $\underbrace{\text{loss fxn}}_{\text{Calculated on training data.}} \Rightarrow \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

How to represent the Loss function in the matrix format

Example

$$M = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Now I want $(a^2+b^2+c^2)$

$$M^T M = \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = (a^2+b^2+c^2)$$

So if $X \Rightarrow$ Training data

$Y \Rightarrow$ actual Y values of tr. data

$\beta \Rightarrow$ linear reg. Parameters.

$(X\beta) \Rightarrow \hat{Y}$

Linear Regression

How to represent the Loss function in the matrix format

So $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}$ $\quad \hat{Y} = X\beta = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_N \end{bmatrix}$

So, $(Y - \hat{Y}) = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix}$

Since loss fxn $\Rightarrow \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \Rightarrow (Y - \hat{Y})^T (Y - \hat{Y})$

$\Rightarrow (Y - X\beta)^T (Y - X\beta)$

Linear Regression

How to represent the Loss function in the matrix format

done

$$\mathcal{L} = (Y - X\beta)^T (Y - X\beta)$$

Linear Regression

How to represent the derivative of L by Beta in matrix format

Recall the 2D data

$$\left(\hat{y} = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2\right)$$

$$L = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2)\right)^2$$

$$\frac{\partial L}{\partial \beta_0} = -2\left[\sum_{i=1}^{N} y_i - \beta_0 \sum 1 - \beta_1 \sum x_i^1 - \beta_2 \sum x_i^2\right]$$

$$\frac{\partial L}{\partial \beta_1} = -2\left[\sum_{i=1}^{N} x_i^1 y_i - \beta_0 \sum x_i^1 - \beta_1 \sum (x_i^1)^2 - \beta_2 \sum x_i^1 x_i^2\right]$$

$$\frac{\partial L}{\partial \beta_2} = -2\left[\sum x_i^2 y_i - \beta_0 \sum x_i^2 - \beta_1 \sum x_i^1 x_i^2 - \beta_2 \sum (x_i^2)^2\right]$$

Linear Regression

How to represent the derivative of L by Beta in matrix format

$$\begin{bmatrix} \frac{\partial L}{\partial \beta_0} \\ \frac{\partial L}{\partial \beta_1} \\ \frac{\partial L}{\partial \beta_2} \end{bmatrix} = -2 \begin{bmatrix} \sum y_i - \left(\beta_0 \sum 1 + \beta_1 \sum x_i^1 + \beta_2 \sum x_i^2\right) \\ \sum x_i^1 y_i - \left(\beta_0 \sum x_i^1 + \beta_1 \sum (x_i^1)^2 + \beta_2 \sum x_i^1 x_i^2\right) \\ \sum x_i^2 y_i - \left(\beta_0 \sum x_i^2 + \beta_1 \sum (x_i^1) x_i^2 + \beta_2 \sum (x_i^2)^2\right) \end{bmatrix}$$

$$= -2 \left[\left[X^T Y \right] - \left[(X^T X) \beta \right] \right]$$

Linear Regression

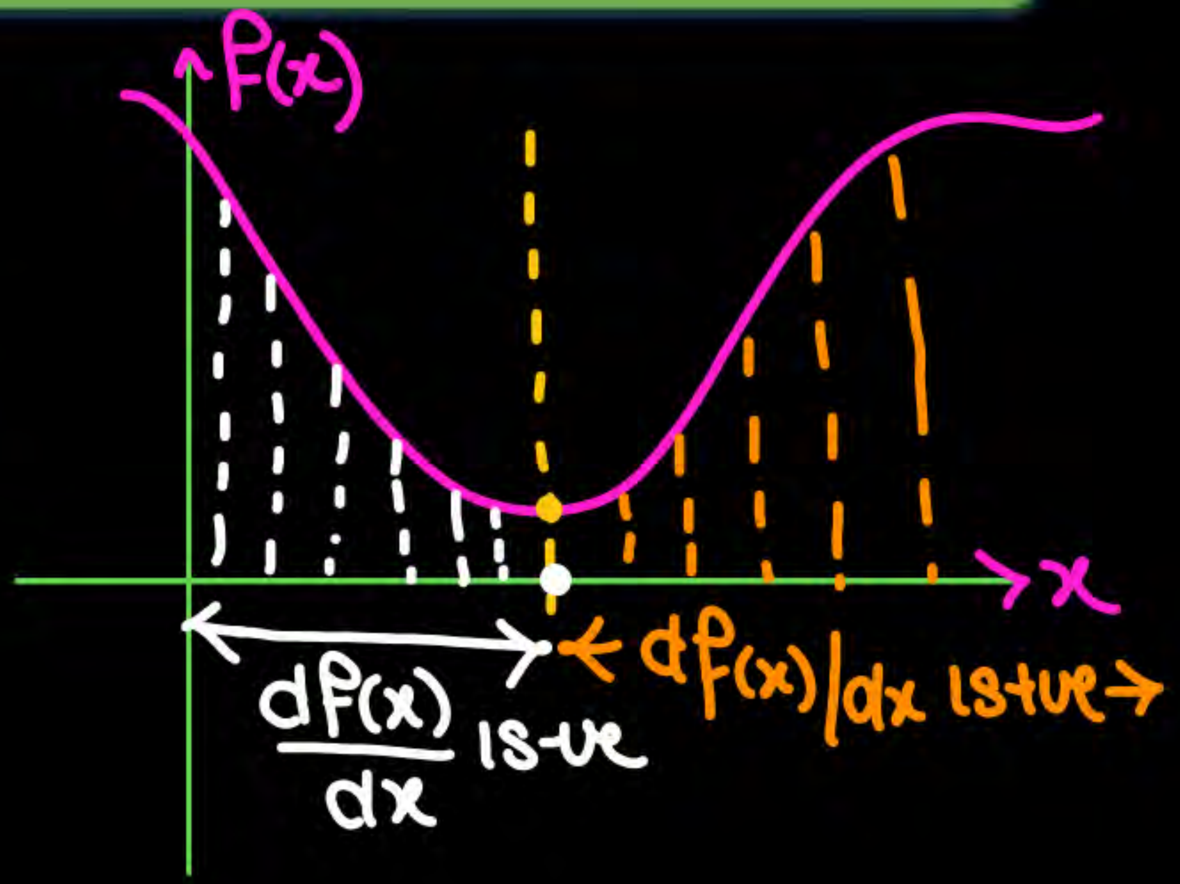
How to represent the derivative of L by Beta in matrix format

- So $L = (Y-\hat{Y})^T (Y-\hat{Y})$
- $$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial L/\partial \beta_0 \\ \partial L/\partial \beta_1 \\ \partial L/\partial \beta_2 \\ \vdots \\ \vdots \end{bmatrix} = -2 \left[(X^T Y) - (X^T X)\beta \right]$$

What is gradient descent method

Concept ⇒

- when any function is decaying then at that location derivative of fxn ⇒ -ve
- when any function is Rising then at that location derivative of fxn ⇒ +ve



What is gradient descent method

* my objective is to reach minima location

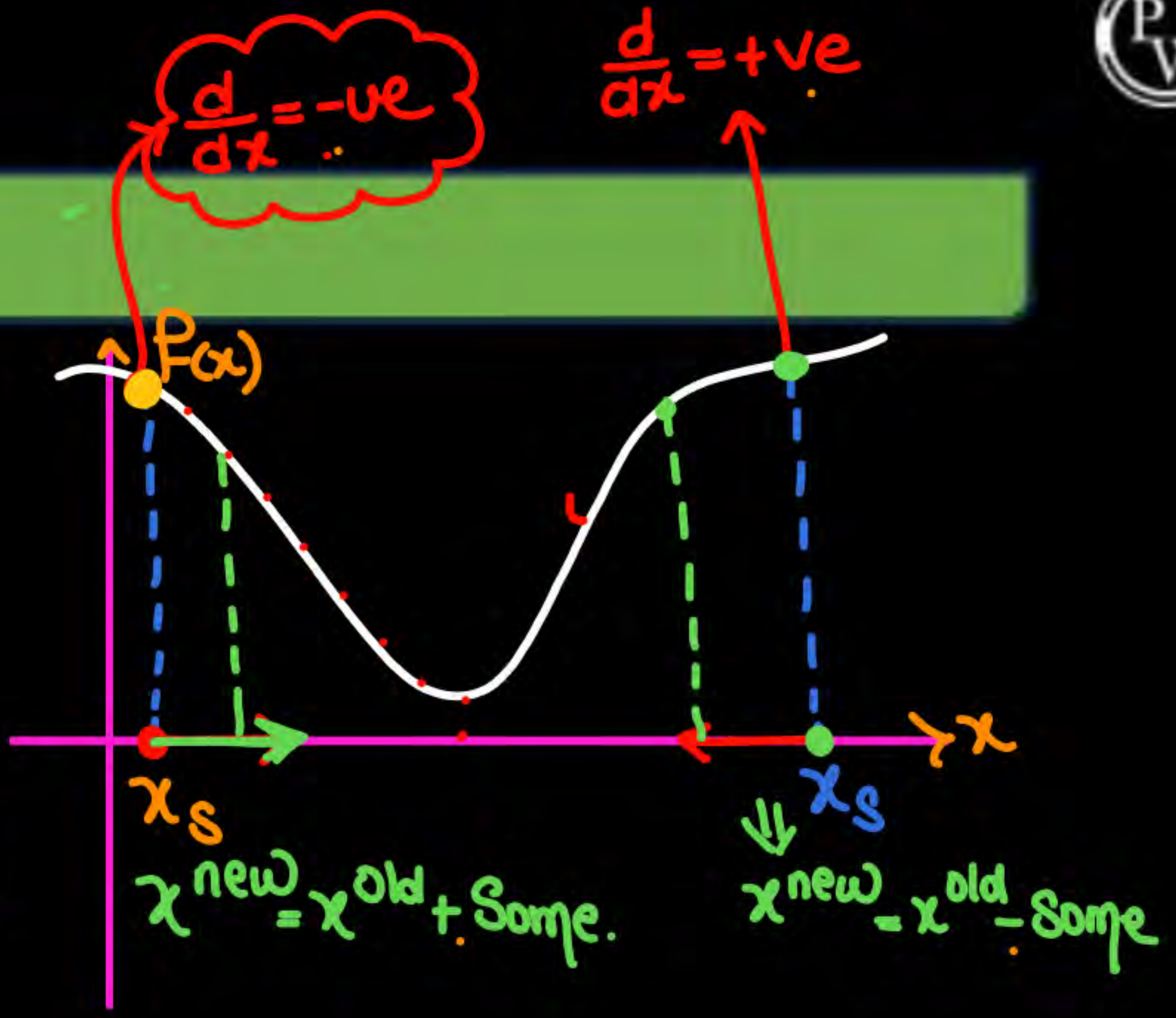
we have a method $df(x)/dx = 0$

So we use gradient descent ⇒

⇒ Start with any Random value of x

⇒ then we move on to next value of x

$$x^{new} = x^{old} - \eta \frac{df(x)}{dx}\bigg|_{x^{old}}$$



What is gradient descent method

So gradient descent method is as follows

So we have to find the min location of any fxn ⇒

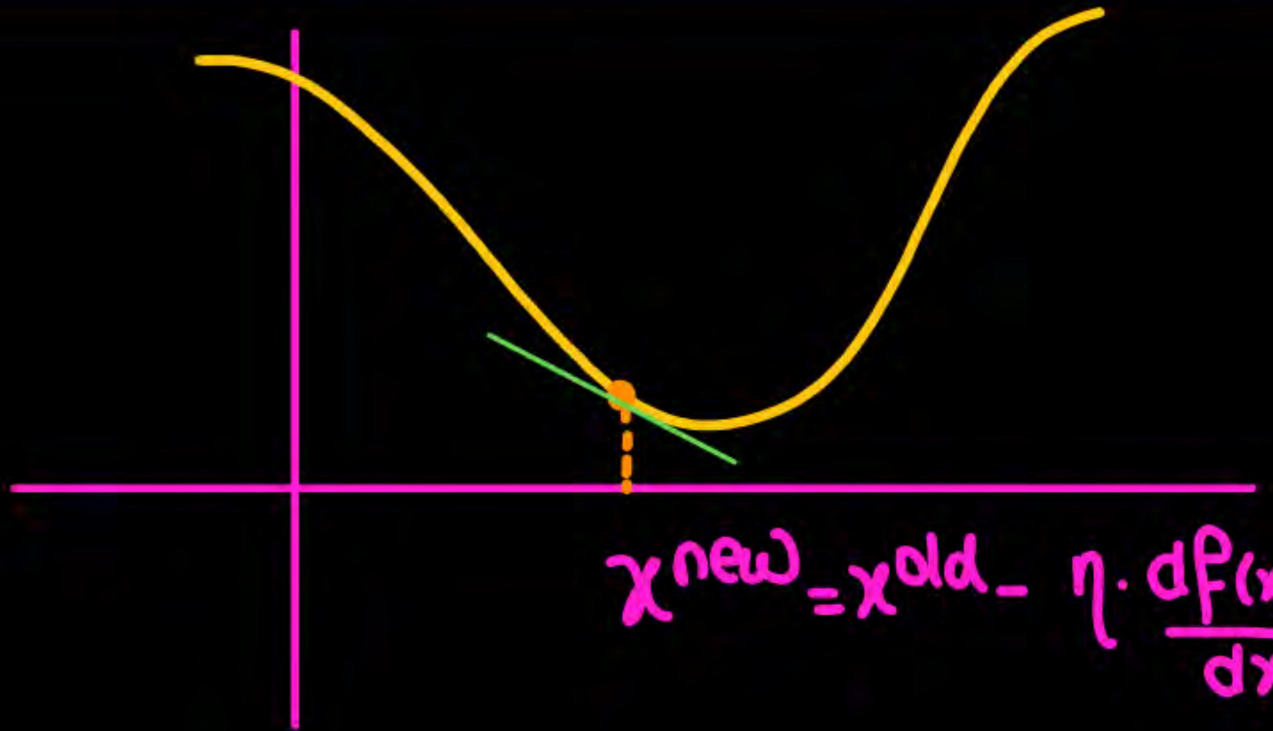
So step 1. Start with any x value

Step 2. find new x ⇒ $x^{new} = \left(x^{old} - \eta \cdot \frac{df(x)}{dx}\Big|_{x^{old}} \right)$

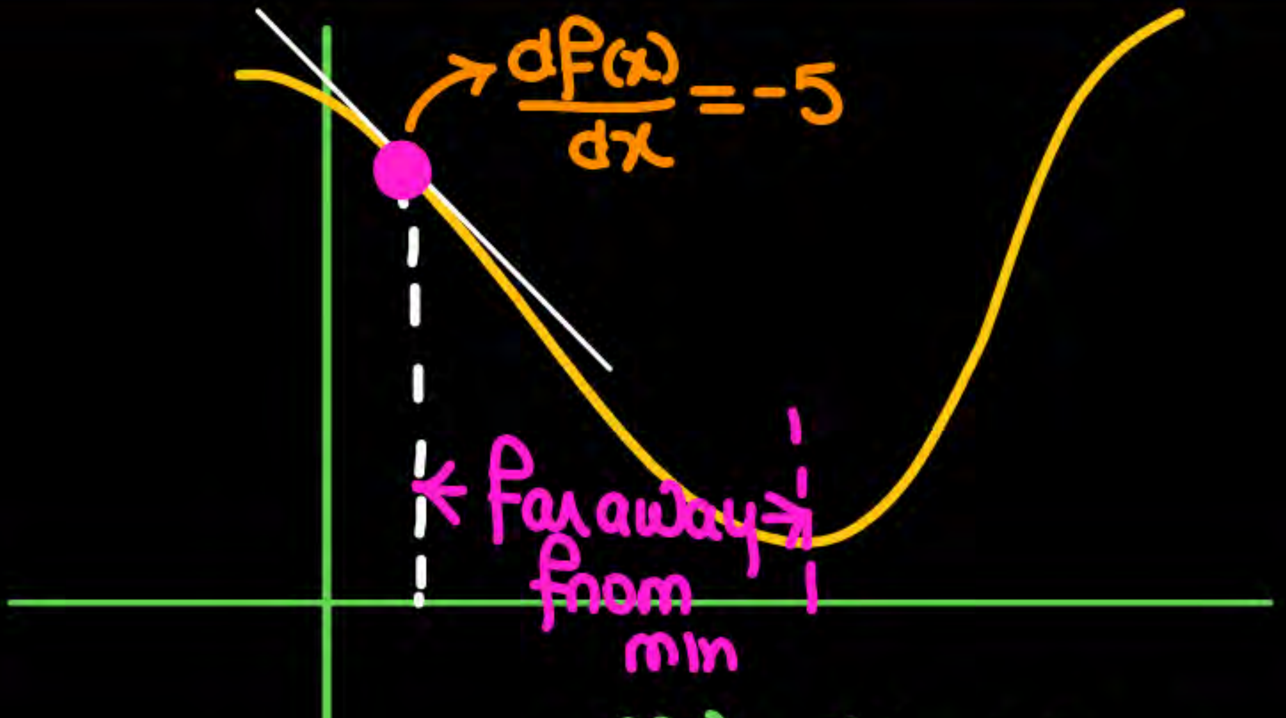
η : learning Rate

⇒ This is a iterative method to reach to Result

What is gradient descent method



$$x^{new} = x^{old} - \eta \cdot \frac{df(x)}{dx}$$



$$x^{new} = x^{old} - \frac{d}{dx} f(x) \times \eta$$

What is gradient descent method

So in linear regression $\Rightarrow$ So we have to minimize

$$\mathcal{L} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\left[y_i - \left(\beta_0 + \beta_1 x_i^1 + \cdots \beta_D x_i^D\right)\right]^2$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

So gradient descent method

So we start with Random β $\Rightarrow$

$$\underbrace{\beta^{new}}_{\text{matrix}} = \underbrace{\beta^{old}}_{\text{matrix}} - \eta \underbrace{\left[\frac{\partial L}{\partial \beta}\right]}_{\text{matrix}}\Bigg|_{\text{at } \beta^{old}}$$

$$\frac{\partial L}{\partial \beta} \Rightarrow -2\left[X^T Y - (X^T X)\beta\right] \rightarrow \beta^{old}$$

#Q. If $g(x, y) = x^2 + y^2 - 4x$, find the gradient vector $\nabla g(1, 2)$

example

Normal maths function

$x = 1$
$y = 2$

So $\nabla g = \begin{bmatrix} \partial g / \partial x \\ \partial g / \partial y \end{bmatrix} = \begin{bmatrix} 2x - 4 \\ 2y \end{bmatrix}$ @ $x = 1, y = 2$

$= \begin{bmatrix} -2 \\ 4 \end{bmatrix}$

#Q. Let f be the function of two variables given by

$$f(x,y) = xy(x+y) = x^2y + xy^2$$

(a) Calculate the gradient vector ∇f and evaluate at the point (1, 2).

H.W.

• Why Gradient descent ⇒

$$\beta = (X^T X)^{-1} (X^T Y)$$

→ we know β solution for min. loss function

→ Problem in above eq is that the Calculation of $(X^T X)^{-1}$ v.v. difficult becoz data is huge.

→ that's why we use Gradient descent.

Significance of learning Rate

↳ Since We donot want that $\frac{df(x)}{dx}$ Completely Control the movement of x.

↳ thus we multiply $\frac{df(x)}{dx}$ by learning Rate so that movement from $x^{old} \rightarrow x^{new}$ is v. slow and we donot cross minima loc.

$\eta \Rightarrow$ v. small

$$\left(x^{new} = x^{old} - \frac{df(x)}{dx}\right)$$

Learning Rate $= 1$

- if learning Rate is v. large $\Rightarrow$ we will never reach min loc.

#Q. Let's consider regression in one dimension, so our inputs $x^{(i)}$ and outputs $y^{(i)}$ are in $\mathbb{R}$.

(a) (4 points) Linny uses regular linear regression. Given the following dataset, (x, y) H.W.

$D = \{((1), 1), ((2), 2), ((3), 4), ((3), 2)\}$

What value of θ and θ_0 optimize the mean squared error of hypotheses of the form $h(x; \theta, \theta_0) = \cancel{\theta x + \theta_0}$? $\theta_1 x + \theta_0$

find θ_1, θ_0

#Q. Suppose we have data about 5 people shown below.

Name	Level	Trials	Phase
Megda	1	10	1
Valerie	5	20	-1
Kumar	2	15	1
Octavia	6	30	1
Dorete	6	5	-1

H.W

(a) Suppose we want to model the level of each person, and use the following constant model:
$f_\theta(x) = \theta_1$. What is $\hat{\theta}_1$, the value that minimizes the average L_2 loss?

#Q. Consider a one-dimensional regression problem with training data $\{x_i, y_i\}$. We seek to fit a linear model with no bias term:

(a) Assume a squared loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and solve for the optimal value of ω^*.

H.W

$\hat{y}_i = \omega x$

"find ω"

#Q. Consider the following 4 training examples:

X	Y
–1	0.0319
0	0.8692
1	1.9566
2	3.0343

H.W

We want to learn a function $f(x) = ax + b$ which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

#Q. The linear regression model $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p$ is to be fitted to a set of N training data points having p attributes each. Let X 1 point be $N \times (p + 1)$ vectors of input values (augmented by 1's), Y be $N \times 1$ vector of target values, and θ be $(p + 1) \times 1$ vector of parameter values $(a_0, a_1, a_2, \ldots\ldots, a_p)$. If the sum squared error is minimized for obtaining the optimal regression model, which of the following equation holds?

d ✓

(a) $X^TX = XY$

(b) $X\theta = X^TY$

(c) $X^TX\theta = Y$

(d) $X^TX\theta = X^TY$

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TSS = total sum of squares

n = number of observations

y_i = value in a sample

$\bar{y}$ = mean value of a sample

Linear Regression

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The most important thing we do after making any model is evaluating the model.
- R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The value of R-square lies between 0 to 1.
- Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.
- However, we get R-square equals 0 when the model does not predict any variability in the model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.
- The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The goodness of fit of regression models can be analyzed on the basis of the R-square method. The more the value of the r-square near 1, the better the model is.

- Note: The value of R-square can also be negative when the model fitted is worse than the average fitted model. .

Considering data of P Dimensions

Adjusted R - Squares

- Adjusted R-Squared is an updated version of R-squared which takes account of the number of independent variables while calculating R-squared.
- n is the total number of observations in the data
- k is the number of independent variables (predictors) in the regression model

$$Adjusted R^2 = 1 - \frac{(1-R^2)\cdot(n-1)}{n-k-1}$$

Considering data of P Dimensions

Lets solve a question

Question 2: Given a simple linear regression model with an R-squared value of 0.64, what percentage of the variation in the dependent variable is explained by the predictor variable?

Linear Regression

Considering data of P Dimensions

Lets solve a question

Question 6: In a simple linear regression model, if the coefficient of determination (R-squared) is 0.81 and the total sum of squares (SST) is 400, what is the sum of squared errors (SSE)?

a)76
b)77
c)54
d)33

What is Mean Square Error

4) Start with the initial guess of $[w_1, w_2] = [5, 5]$. Take the value of learning rate = 0.3. The value of w_1 after ~~4~~ 2 iterations of gradient descent will be ________.

$$J(\omega) = \omega_1^2 + \omega_2^2 - 6\omega_1 + 8\omega_2 - 9$$

1st iteration

Solve. $\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$

Step 1 ⇒ Start with initial value

Step 2 ⇒ $\omega^{new} = \omega^{old} - .3 \left(\frac{\partial J}{\partial \omega}\right)_{\omega^{old}}$

$$\Rightarrow \frac{\partial J}{\partial \omega} = \begin{bmatrix} \partial J / \partial \omega_1 \\ \partial J / \partial \omega_2 \end{bmatrix} = \begin{bmatrix} 2\omega_1 - 6 \\ 2\omega_2 + 8 \end{bmatrix}$$

$$= \begin{bmatrix} 5 \\ 5 \end{bmatrix} - .3 \begin{pmatrix} 4 \\ 18 \end{pmatrix} \Rightarrow \begin{bmatrix} 3.8 \\ -.4 \end{bmatrix}$$

2nd Iteration ⇒

$$W_{new} = W_{old} - .3 \left(\partial J / \partial w \right) \Big|_{@ W_{old}}$$

$$= \underbrace{\begin{bmatrix} 3.8 \\ -.4 \end{bmatrix}} - .3 \begin{bmatrix} 1.6 \\ 7.2 \end{bmatrix}$$

$$= \begin{bmatrix} 3.32 \\ -2.56 \end{bmatrix}$$

What is Mean Square Error

Consider the function $J(w) = w_1^2 + w_2^2 - 6w_1 + 8w_2 - 9$. Answer questions (1-6):

1) The theoretical value of $min(J(w))$ is ________.

So find ω_1, ω_2 to $\min(J(\omega))$

$$\frac{\partial J}{\partial \omega_1} = 2\omega_1 - 6 = 0, \; \omega_1 = 3$$

$$\frac{\partial J}{\partial \omega_2} = 2\omega_2 + 8 = 0, \; \omega_2 = -4$$

$\rightarrow$ Value of ω_1, ω_2

min Value of $J(\omega) \Rightarrow -34$

2 mins Summary

Topic

Topic

Topic

Topic

Topic

THANK - YOU