TASK - 6

REGISTRATION/REFERRAL ID: AIRSS1129

NIKHIL SHRESTHA

Q1. Calculate/ derive the gradients used to update the parameters in cost function

optimization for simple linear regression.

Ans: Simple linear Regression model is in the form $\beta_1 X_i + \beta_0 = y_i$

- β_1 = Slope / coeff
- β_0 = constant / intercenpt
- y_i = dependent variable
- X_i = independent variable

Also, to evaluate the performance of the model we use Loss function (Mean Squared Error) which is basically the sum of squared differences between original target variable values and predicted values. Mathematically represented as:

$$J = \frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - (\beta_1 X_i + \beta_0))^2$$

$\hat{y}_i$ = Prediction

J = Cost function

$(\beta_1 X_i + \beta_0))^2$ = Target variable

Our target is to find best values for β_1 and β_0, such that we minimize this loss function.

Many books and article use $\frac{1}{2n}$ instead of $\frac{1}{n}$, which is basically taken to cancel out 2 which we get after the first derivative.

Now after this base understanding, we can see that we can reduce loss function by finding its minima. To do that we can use optimization algorithm called **Gradient Descent.**

- So, Gradient Descent is an iterative algorithm which finds the minima of a function.
- We use this function to find the minima of the Cost function, which actually finds the best fit line for given training dataset in smaller number of iterations.
- The values for which the MSE will be minimum (Global minima) will be the final values for our model which will ultimately give us the best fit line in predicting the dependent variable.
- Gradient descent starts with a random value of β_1 *or* β_0, which finally reaches the minima through iterations by updating the parameters.

Updating the Parameters in Simple Linear Regression using Gradient Descent:

- $For\ \beta_0$:
  - $\widehat{\beta_0} = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$
- $For\ \beta_1$:
  - $\widehat{\beta_1} = \beta_1 - \alpha \frac{\partial}{\partial \beta_1} J(\beta_1)$

Where α is learning rate and

$\frac{\partial}{\partial \beta_0} J(\beta_0)\ and\ \frac{\partial}{\partial \beta_1} J(\beta_1)$ are Gradients

Gradient's sign (-ve or +ve) will determine the relationship as to when $\beta_0\ or\ \beta_1$ increases how will cost function will be affected.

Positive value of Gradient (+ve) = $\beta_0\ or\ \beta_1$ will be directly proportional to J (Cost Function), as $\beta_0\ or\ \beta_1$ increases J will increase and vice-versa.

Negative value of Gradient (-ve) = $\beta_0\ or\ \beta_1$ will be inversely proportional to J (Cost Function), as $\beta_0\ or\ \beta_1$ decreases J will increase and vice-versa.

Now Learning Rate α determines the rate at which the gradients should be updated. Also, we can observe that when α is small gradient will progress very slowly towards the minima of Cost Function. On the other side, if we take a bigger value for α then function may skip the global minima or even oscillate around a certain value.

As we can understand from the equation that the -ve sign actually serves the purpose of getting the relationship correct between ($\beta_0\ or\ \beta_1$) and Cost Function.

Deriving the Gradients Simple Linear Regression using Gradient Descent:

Now let's solve for $\frac{\partial}{\partial \beta_0} J(\beta_0)$ which is nothing but derivation of our cost function w.r.t to β_0 (all other variables will be considered as constant):

- $\frac{\partial}{\partial \beta_0} J(\beta_0) = \frac{1}{n} \sum_{i=0}^{n} 2(\hat{y}_i - (\beta_1 X_i + \beta_0))(-1)$
- $\frac{\partial}{\partial \beta_0} J(\beta_0) = \frac{-2}{n} \sum_{i=0}^{n} (\hat{y}_i - (\beta_1 X_i + \beta_0))$

Then let's solve for $\frac{\partial}{\partial \beta_1} J(\beta_1)$ which is nothing but derivation of our cost function w.r.t to β_1 (all other variables will be considered as constant):

- $\frac{\partial}{\partial \beta_0} J(\beta_1) = \frac{1}{n} \sum_{i=0}^{n} 2(\hat{y}_i - (\beta_1 x_i + \beta_0))(-x)$

- $\frac{\partial}{\partial \beta_0} J(\beta_1) = \frac{-2x}{n} \sum_{i=0}^{n} (\hat{y}_i - (\beta_1 X_i + \beta_0))$

Finally, we put the Values of $\frac{\partial}{\partial \beta_0} J(\beta_0)$ and $\frac{\partial}{\partial \beta_1} J(\beta_1)$ in $\beta_0 = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$

and $\beta_1 = \beta_1 - \alpha \frac{\partial}{\partial \beta_1} J(\beta_1)$ respectively. We will get the updated parameters of Simple Linear Regression using Gradient Descent.

For β_0:

$\widehat{\beta_0} = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$

$\widehat{\beta_0} = \beta_0 - \alpha \left(\frac{-2}{n} \sum_{i=0}^{n} (\hat{y}_i - (\beta_1 X_i + \beta_0)) \right)$

For β_1:

$\widehat{\beta_1} = \beta_1 - \alpha \frac{\partial}{\partial \beta_1} J(\beta_1)$

$\widehat{\beta_1} = \beta_0 - \alpha \left(\frac{-2x}{n} \sum_{i=0}^{n} (\hat{y}_i - (\beta_1 X_i + \beta_0)) \right)$

Q2. What does the sign of gradient say about the relationship between the parameters and cost function?

Ans: Gradient's sign (-ve or +ve) will determine the relationship as to when β_0 or β_1 increases how will cost function will be affected.

Positive value of Gradient (+ve) = β_0 or β_1 will be directly proportional to J (Cost Function), as β_0 or β_1 increases J will increase and vice-versa.

Negative value of Gradient (-ve) = β_0 or β_1 will be inversely proportional to J (Cost Function), as β_0 or β_1 decreases J will increase and vice-versa.

Using this understanding let's understand the Parameter updating function of Gradient Descent in case of Simple Linear Regression:

$$\widehat{\beta_0} = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$$

Case 1: When $\frac{\partial}{\partial \beta_0} J(\beta_0)$ is positive (+ve) $\beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$ will be a positive value which will be lesser than previous β_0 which actually means that in next iteration this β_0 will be smaller and hence the cost function will also be small (since gradient is +ve)

Case 2: When $\frac{\partial}{\partial \beta_0} J(\beta_0)$ is negative (-ve) $\beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$ will be a positive value which will be greater than previous β_0 which actually means that in next iteration this β_0 will be larger and hence the cost function will be small (since gradient is -ve)

Q3. Why Mean squared error is taken as the cost function for regression problems?

Ans: The Mean Squared Error (MSE) is defined by the equation:

- $MSE = \frac{1}{n} \sum_{i=0}^{n} (\hat{y}_i - y)^2$

Mean Squared Error advantages are:

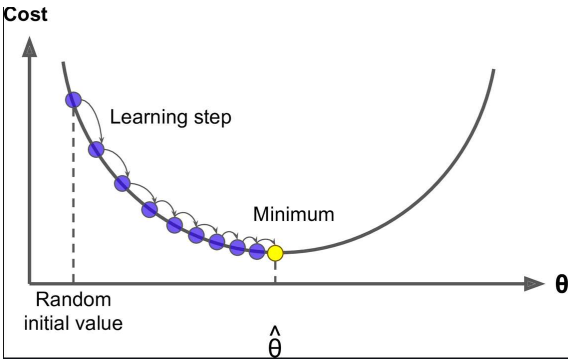
- Ease of differentiation: As we have seen above that Cost Function needs to be differentiable and MSE is very easy for differentiation when compared to the absolute function (Mean Absolute Error)
- Next, we want to avoid the negative values from when we take the difference, that is ensured by squaring the values of individual differences.
- This function gives more weights to larger differences.

Q4. What is the effect of learning rate on optimization, discuss all the cases?

Ans: Learning Rate (α) determines the rate at which the gradients should be updated.

- $\widehat{\beta_0} = \beta_0 - \alpha \frac{\partial}{\partial \beta_0} J(\beta_0)$
- $\widehat{\beta_0} = \beta_1 - \alpha \frac{\partial}{\partial \beta_1} J(\beta_1)$

- It determines how fast or slow we will move towards the optimal weights.
- When α is small gradient will progress very slowly towards the minima of Cost Function, as $\beta_1 \; and \; \beta_0$ will take smaller steps to reach the global minima.
- On the other side, if we take a bigger value for α then function may skip the global minima or even oscillate around a certain value. as $\beta_1 \; and \; \beta_0$ will take larger steps to reach the global minima.
- Usually kept between 0.0 and 1.0, but for better approach decide according to the problem needed to be tackle.



- Image Ref: https://github.com/SoojungHong/MachineLearning/wiki/Gradient-Descent

Here $\theta = \beta$