Underlying Mathematics behind Gradient Descent for Supervised Regression Learning

Raunak Joshi - Dept of Artificial Intelligence and Data Science - VCET

July 2023

The regression based learning models use linear functions for computation of the output layer values. The representation of the same can be given by the formula as follows,

$$\hat{y_i} = W * X_i + b \tag{1}$$

where the $\hat{y_i}$ is the dependent variable and X_i is the independent variable. The W and b are the weights and bias that are arbitrarily declared to form the linear equation. These can be termed as a regression co-efficient and constant for calculations. The cost function for 1 used is either Mean Squared Error or Mean Absolute Error. For now we consider the Mean Squared Error which can be represented by the formula,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{2}$$

where the $\hat{y_i}$ is the predicted value and y_i is the ground truth label for respective record. The process of gradient descent starts by performing a partial derivative on the cost function and will be done with respect to weights and bias. Starting with weights, the equation below is a derivation for the it over the cost function.

$$\frac{\partial}{\partial W} = \frac{1}{N}\sum(y_i - \hat{y_i})^2 \tag{3}$$

which can be also represented as given in equation 1 to give better intuition of the weights parameter.

$$\frac{\partial}{\partial W} = \frac{1}{N}\sum(y_i - (W * X_i + b))^2 \tag{4}$$

The equation 4 is in the form of x^n which requires chain rule for calculating the gradient. The chain rule will equate to $nx^{n-1} * \frac{\partial}{\partial x}x$. The chain rule equates all the gradients sequentially to the point it reaches the parameter for which the partial derivative is to be calculated. The derivation for the same has been given below as follows.

$$\begin{aligned}\frac{\partial}{\partial W} &= \frac{1}{N}\sum(y_i - (W * X_i + b))^2 \\ &= \frac{1}{N}\sum 2 * (y_i - (W * X_i + b)) * \frac{\partial}{\partial W}(y_i - (W * X_i + b)) \\ &= \frac{2}{N}\sum(y_i - (W * X_i + b)) * (-X_i) \\ &= -\frac{2}{N}\sum(X_i) * (y_i - (W * X_i + b))\end{aligned} \tag{5}$$

The equation 5 is the complete derivation of for the partial derivative taken with respect of the weight parameter. The similar can be performed for bias.

$$\begin{aligned}\frac{\partial}{\partial b} &= \frac{1}{N}\sum(y_i - \hat{y_i})^2 \\ &= \frac{1}{N}\sum(y_i - (W * X_i + b))^2 \\ &= \frac{1}{N}\sum 2 * (y_i - (W * X_i + b)) * \frac{\partial}{\partial b}(y_i - (W * X_i + b)) \\ &= \frac{2}{N}\sum(y_i - (W * X_i + b)) * (-1) \\ &= -\frac{2}{N}\sum(y_i - (W * X_i + b))\end{aligned} \tag{6}$$

The calculated derivatives can be now used to generate the second step of gradient descent which is using learning rate to be multiplied with gradient and subtracted from original parameters. The updated weight will be represented by,

$$W = W - (\alpha * -\frac{2}{N}\sum(X_i) * (y_i - (W * X_i + b))) \tag{7}$$

Similarly using the derivative calculated for the bias will be used to update the bias as is given as follows,

$$b = b - (\alpha * -\frac{2}{N}\sum(y_i - (W * X_i + b))) \tag{8}$$

This process is iterative and is supposed to be performed till the point of convergence.