Q.How does the learning rate affect the training of the Neural Network?

Answer:

Learning rate is used to scale the magnitude of parameter updates during gradient descent. The choice of the value for learning rate can impact two things: 1) how fast the algorithm learns and 2) whether the cost function is minimized or not.The variation in cost function with a number of iterations/epochs for different learning rates.

It can be seen that for an optimal value of the learning rate, the cost function value is minimized in a few iterations (smaller time). This is represented by the blue line in the figure. If the learning rate used is lower than the optimal value, the number of iterations/epochs required to minimize the cost function is high (takes longer time). This is represented by the green line in the figure. If the learning rate is high, the cost function could saturate at a value higher than the minimum value. This is represented by the red line in the figure. If the learning rate selected is very high, the cost function could continue to increase with iterations/epochs. An optimal learning rate is not easy to find for a given problem. Though getting the right learning is always a challenge, there are some well-researched methods documented to figure out optimal learning rates. Some of these techniques are discussed in the following sections. In all these techniques the fundamental idea is to vary the learning rate dynamically instead of using a constant learning rate.

Decaying Learning rate

In the decaying learning rate approach, the learning rate decreases with increase in epochs/iterations. The formula used for the decaying learning rate is shown below.

$$\alpha = \frac{\alpha_o}{1+\delta \times Epoch\#}$$ **Equation-5**

In the above equation, o is the initial learning rate, is the decay rate and is the learning rate at a given Epoch number. Figure 3 shows the learning rate decay with the epoch number for different initial learning rates and decay rates.

Scheduled Drop Learning rate

Unlike the decay method, where the learning rate drops monotonously, in the drop learning rate method, the learning rate is dropped by a predetermined proportion at a predetermined frequency. The formula used to calculate the learning rate for a given epoch is shown in the below equation.

$$\alpha_n = \alpha_o \times D^{Quotient\left(\frac{n}{\rho}\right)}$$ **Equation-6**

In the above equation, o is the initial learning rate, '*n*' is the epoch/iteration number, '*D*' is a hyper-parameter which specifies by how much the learning rate has to drop, and ρ is another hyper-parameter which specifies the epoch-based frequency of dropping the learning rate. Figure 4 shows the learning rate variation with epochs for different values of *'D'* and *'ρ'*.