Machine Learning

Assignment 7.4

Submitted By: Ranji Raj

December 12, 2020

Learning Rate

- The role of the hyperparameter:learning rate or *step size* is to moderate the degree to which weights are changed at each step.
- It is almost always set to a very small value and is also sometimes made to decay as the number of weight-tuning iterations increases.
- The learning rate controls how quickly or slowly the model is adapted to the problem.
- Summing over multiple training examples, in standard gradient descent it requires more computations per weight-tuning step. due to which it is often used with a **larger** step size per weight than stochastic gradient descent.
- The weight update rule for gradient descent is given by,

$$\Delta w_i = \eta \sum_{d \in D} (t_d - o_d) x_{i,d}$$

  In this, because the error surface contains only a single global minimum, this algorithm will converge to a weight vector with minimum error, regardless of whether the training examples are linearly separable or not, given a sufficiently **smaller** step size.
- If η is made too large, the gradient descent search runs the risk of *overstepping* the minimum surface rather than converging to it. So a common thumb rule is to gradually reduce the value of η as the number of steps grows.

Effect of fixed learning rate

- If the step size is kept same across along all the layers of the neural network it may fall into the problem of **vanishing gradient**.
- Meaning, the weight may stop changing as weight change backpropogates itself to first layer (since there are lot multiplications of derivatives and these derivatives itself attain decimal values *less* than 1 and there products are even smaller if we observe mathematical analysis of backpropagation of neural networks and as a result learning will not take place and saturate immaturely) so we **MUST** assign variable learning rate to each layer.

Too high vs. Too low learning rate

- A learning rate that is too large can cause the model to converge too quickly to a sub-optimal solution.
- Whereas a learning rate that is too small can cause the process to get stuck in local minimal.

Learning Rate Schedule-An alternative to fixed learning rate

- The way in which the learning rate changes over time (training epochs) is referred to as the **learning rate schedule** or **learning rate decay**.
- The simplest learning rate schedule is to decrease the learning rate *linearly* from a large initial value to a small value. This allows large weight changes in the beginning of the learning process and small changes or fine-tuning towards the end of the learning process.

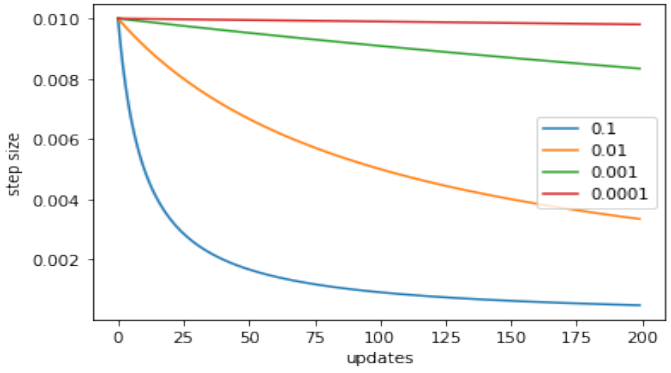

Figure 1: Effect of decay on learning rate