AdaGrad Optimizer

Gradient Descent

Gradient Descent algorithm is a first order derivative optimization algorithm for finding a local minimum of a differentiable function. It is used by machine learning algorithms (especially Neural Networks) to optimize the parameters of a model by minimizing the loss function. However, the issue with gradient descent is that the learning rate remains constant throughout all iterations and for all parameters of the model.

Why Adagrad?

AdaGrad (or Adaptive Gradient) is an improvement over the regular Stochastic Gradient Descent. The algorithm uses a second-order derivative information in the parameter updates and provides adaptative learning rates for each parameter. It adapts the learning rate for each feature depending on the estimated geometry of the problem i.e., it updates the parameters in a way that assigns higher learning rates to features that have been updated frequently and smaller if otherwise.

Update rule for weights:

$$w(t) = w(t-1) - \eta' \frac{\partial L}{\partial w(t-1)}$$

where, η_t is updated as

$$\eta_t' = \frac{\eta}{\sqrt{\alpha_t + \epsilon}}$$

Here, η is a constant, and ε is a small positive number which is used to prevent a divide by zero error. α_t is given as:

$$\alpha_t = \sum_{i=1}^{t} g_i^2$$

$$g_i = \frac{\partial L}{\partial w(old)}$$

Hence, we see that α_t is the sum of all the gradients for that given parameter up until time ‘t’. Since, g_i^2, which is the derivative of loss function is always positive, the value of α_t keeps on increasing with each iteration. Hence, the learning rate reduces adaptively with each iteration.

Advantages and Uses

- There is no need to tune the parameters manually.
- More reliable the stochastic gradient descent.
- Works well with sparse data like the ones used for NLP tasks as it tries to give a fair chance to less frequently occurring words.

Disadvantages

- Since, α_t keeps on increasing with each iteration, after a certain point it becomes too large, and the weight update values become negligible which means that the algorithm