Gradient Descent variations

There are variations for gradient descent like mini batch gradient descent and stochastic gradient descent.

Mini batch gradient descent has a better performance in deep learning when the number of training examples is too large. Stochastic gradient descent is a special case of mini batch where the size of mini batch is 1.

Exponentially weighted average

To plot a moving average of some value say V

Let us assume this trend:

V0 = 0

V1= 0.9V0 + 0.1theta1 and so on..

Chart, scatter chart

Description automatically generated

When B has a value we can assume Vt as an average of V over 1/1-B days

Chart, scatter chart

Description automatically generated

This process is called the exponentially weighted average

Diagram

Description automatically generated

Diagram, letter

Description automatically generated

This is how it is implemented.

Bias correction:

Chart, scatter chart

Description automatically generated

We want the green curve to come when B = 0.98. But we get the purple curve. Why is it so? If the first days temperature is 40. We get v0 = 0. V1 = Bv0 + 1-B\*theta1. Here it will give you 8 degrees which is very less.

There is a way to modify the estimate

Chart, scatter chart

Description automatically generated

So when we divide the vt by 1-Bt, it will find the weighted average and will give you the right answer.`

Gradient descent with momentum:

Instead of subtracting W with dW like normal gradient descent

Diagram

Description automatically generated

It takes small oscillations in the vertical direction but a greater speed in the horizontal direction.

Text

Description automatically generated

Sometimes the 1-B parameter is chucked. The one on the left is better than the one on the right.

RMSProp: In gradient descent there are oscillations in the vertical and horizontal directions. So RMS prop, on iteration t, it will compute dW, db every mini batch.

Text

Description automatically generated

Normally sdw will be relatively small and sdb will be relatively large.

A picture containing dishware

Description automatically generated

The green line will be the trend because of rmsprop. This will allow us to give a larger learning rate and get the learning faster. Sdw should never go to 0 so we add a very small value of the order 10^-8.

Text

Description automatically generated

Adam optimization : It is basically taking momentum and rms prop and putting them together. Here we introduce bias correction also.

Text, letter

Description automatically generated

Learning rate decay:

Why to implement this ?

As mini batch gradient descent decreases the cost, it will never converge correctly to the minima. As the convergence happens, changing the learning rate to a smaller value makes it oscillate in a tighter region to the minima. We can make this as follows

Diagram

Description automatically generated

There is also exponential decay

Graphical user interface, application, Word

Description automatically generated

Local optima in deep learning: most of the points of 0 gradient are not local optima they are also saddle points

Chart

Description automatically generated

Problem of plateaus : they are regions close to 0 for a long long time

A picture containing chart

Description automatically generated