AI Coffe Club (04/12/2019)
Backpropagation

Last modification: December 4, 2019

x is the input vector.
z^l input sum of a neuron
a^l activations column-vector for the layer l
W^l weights or connections from layer $l-1$ to l
w^l_{kj} weight from neuron j from layer $l-1$ to neuron k from layer l
b^l neurons' biases for layer l
σ activation or transfer function
$C(\hat{y}, y)$ the error function where y is the ground truth and $\hat{y}$ is the prediction

Three steps:

1. Forward pass: calculate the activations of each layer.
2. Cost calculation: compute the value of $C(\hat{y}, y)$.
3. Backpropagation: updating the weights and biases of the network.

1 Forward Pass

The forward pass of a layer can be formalized as:

$$a^l = \sigma(z^l) = W^l a^{l-1} + b^l \tag{1}$$

So for the first layer we take the input:

$$a^1 = \sigma(z^1) = W^1 x + b^1 \tag{2}$$

The whole network forward pass can be expressed as:

$$a^L = [\sigma(W^L[\ldots[\sigma(W^2[\sigma(W^1 x_+ b^1)])])] \tag{3}$$

More specifically, the input sum z^l_k of a neuron k in layer l:

$$z^l_k = \sum_j w^l_{kj} a^{l-1}_j + b^l_k \tag{4}$$

Then we use the transfer or activation function:

$$a_k^l = \sigma(z_k^l) \tag{5}$$

So the input sum z_m^{l+1} of a neuron m in a layer $l+1$:

$$z_m^{l+1} = \sum_k w_{mk}^{l+1} a_k^l + b_m^l \tag{6}$$

And so on ...

2 Cost Calculation

Now we have predictions but we need a way to quantitatively determine how good or bad they are: the cost function $C(\hat{y}, y)$, which is a function of the predicted values $\hat{y}$ and the ground truth y. Also referred as loss or just error.

The goal is to find the set of weights and biases that minimize the cost function. The selection of the cost function is critical for the problem, one usual loss function is the cross-entropy:

$$C(\hat{y}, y) = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{7}$$

The cost can be computed just by iterating over all the samples of the training set and pluggint the output of the forward pass of the network ($a^L = \hat{y}$) for each one of them to the cost function.

3 Backpropagation

The goal of backpropagation is to compute the partial derivatives of the cost function w.r.t. to any weight or bias in the network. Using those partial derivatives we can update the weights and biases of the network by multiplying it by some constant α named learning rate.

In other words, this is the gradient descent algorithm since the partial derivatives give us the direction of greatest ascent. By taking a step (whose size is determined by the learning rate) in the opposite direction we can get to the minimum (which may or may not be global):

$$W^l = W^l - \alpha \frac{\partial C}{\partial W^l} \tag{8}$$

$$b^l = b^l - \alpha \frac{\partial C}{\partial b^l} \tag{9}$$

So the derivative of the cost w.r.t. to set of weights in layer l is:

$$\frac{\partial C}{\partial W^l} = \frac{\partial C}{\partial z^l}\frac{\partial z^l}{\partial W^l} = \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l}\frac{\partial z^l}{\partial W^l} \tag{10}$$

If we apply the chain rule, the derivative of the cost function w.r.t. to the set of weights in layer $l-1$ is:

$$\frac{\partial C}{\partial W^{l-1}} = \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l}\frac{\partial z^l}{\partial a^{l-1}}\frac{\partial a^{l-1}}{\partial z^{l-1}}\frac{\partial z^{l-1}}{\partial W^{l-1}} \tag{11}$$

And so on until we reach the last set of weights w^1. As you might have noticed, this process starts from the last layer (opposite to the forward pass which starts from the first). The same development is applied to the biases.

In the end, our general rule can be expressed as:

$$\frac{\partial C}{\partial W^l} = \frac{\partial C}{\partial z^l}\frac{\partial z^l}{\partial w^l} \tag{12}$$

$$\frac{\partial C}{\partial b^l} = \frac{\partial C}{\partial z^l}\frac{\partial z^l}{\partial b^l} \tag{13}$$

Which means that we need four fundamental partial derivatives required for backpropagation:

$$\frac{\partial C}{\partial z^L}, \frac{\partial C}{\partial z^l}, \frac{\partial z^l}{\partial W^l}, \frac{\partial z^l}{\partial b^l} \tag{14}$$

Assume we use a sigmoid activation function for each layer:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{15}$$

3.1 Partial derivative of C w.r.t. z^L

$$\frac{\partial C}{\partial z^L} = \frac{\partial C}{\partial a^L}\frac{\partial a^L}{\partial z^L} = \frac{\partial C}{\partial a^L}\sigma'(z^L)\ , \tag{16}$$

where

$$\frac{\partial C}{\partial a^L} = \frac{\partial(-(y\log(a^L) + (1-y)\log(1-a^L)))}{\partial a^L} = -(\frac{y}{a^L} - \frac{1-y}{1-a^L}) \tag{17}$$

Furthermore, the derivative of the activation function is:

$$\sigma'(z^L) = \sigma(z^L)(1-\sigma(z^L)) = a^L(1-a^L) \tag{18}$$

So putting it all together:

$$\frac{\partial C}{\partial z^L} = -(\frac{y}{a^L} - \frac{1-y}{1-a^L})a^L(1-a^L) = a^L - y \tag{19}$$

3.2 Partial derivative of C w.r.t. z^l

Ideally, we would like to have the partial derivative of C w.r.t. to z^l in terms of the partial derivative of C w.r.t. z^{l+1} so that once we have z^L we can compute $z^{L-1}, z^{L-2}, ...$ and so on. We can write:

$$\frac{\partial C}{\partial z^l} = \frac{\partial C}{\partial z^{l+1}} \frac{\partial z^{l+1}}{\partial a^l} \frac{\partial a^l}{\partial z^l} \tag{20}$$

Since we have calculated first the partial derivative of C w.r.t. to the last input sum z^L we can calculate this derivative for every $l = L-1, L-2, ...$ by substituting the values on the equation above. However, we still need to calculate the partial derivative of z^{l+1} w.r.t. a^l and the partial derivative of a^l w.r.t. z^l. Taking into account that:

$$z^{l+1} = W^{l+1} a^l + b^{l+1} \ , \tag{21}$$

we can substitute it:

$$\frac{\partial z^{l+1}}{\partial a^l} = \frac{\partial}{\partial a^l}(W^{l+1} a^l + b^{l+1}) = W^{l+1} \tag{22}$$

and also

$$\frac{\partial a^l}{\partial z^l} = \sigma'(z^l) \tag{23}$$

If we put everything together:

$$\frac{\partial C}{\partial z^l} = \frac{\partial C}{\partial z^{l+1}} W^{l+1} \sigma'(z^l) \tag{24}$$

3.3 Partial derivative of z^l w.r.t. W^l

Taking into account that:

$$z^l = W^l a^{l-1} + b^l \ , \tag{25}$$

we can substitute it:

$$\frac{\partial z^l}{W^l} = \frac{\partial}{\partial W^l}(W^l a^{l-1} + b^l) = a^{l-1} \tag{26}$$

3.4 Partial derivative of z^l w.r.t. b^l

Again:

$$z^l = W^l a^{l-1} + b^l \ , \tag{27}$$

so:

$$\frac{\partial z^l}{b^l} = \frac{\partial}{\partial b^l}(W^l a^{l-1} + b^l) = 1 \tag{28}$$

3.5 Putting it all together

$$\frac{\partial C}{\partial W^l} = \frac{\partial C}{\partial z^l}\frac{\partial z^l}{\partial W^l} = \frac{\partial C}{\partial z^{l+1}} W^{l+1}\sigma'(z^l)a^{l-1} \tag{29}$$

$$\frac{\partial C}{\partial b^l} = \frac{\partial C}{\partial z^l}\frac{\partial z^l}{\partial b^l} = \frac{\partial C}{\partial z^{l+1}} W^{l+1}\sigma'(z^l) \tag{30}$$