BackPropagation

Algorithm

Definitions

- K = number of output units

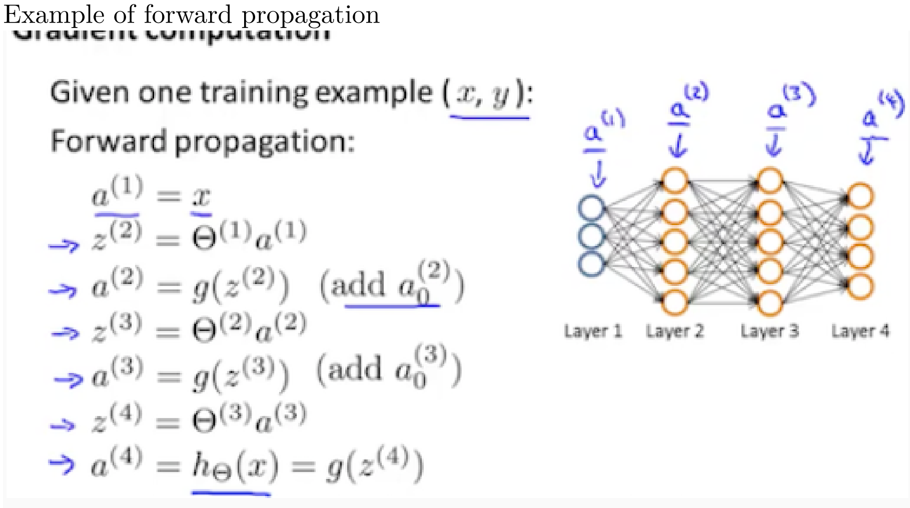
- L = number of layers in network

- s_l = number of units in layer l
- m = number of training examples
- E - another symbol for cost function J

Steps

1. Perform Forward Propagation -> Result is going be in the size of (m * K)

Example of forward propagation



2. Calculate the cost
   - Definitions
     - θ - Hypothesis function parameters
     - $h_\theta()_k$ Hypothesis function
     - $x^{(i)}$ i-th training data
   - $J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[y_k^{(i)}\log((h_\Theta(x^{(i)}))_k) + (1 - y_k^{(i)})\log(1 - (h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{j,i}^{(l)})^2$
3. Calculate derivatives of Cost Function according to every z_s^l $\frac{\partial E}{\partial z_s^l}$

4. Calculate derivatives for every theta

Derivatives

3. Step - Calculating $\frac{\partial E}{\partial z_s^l} = \delta^l$

1 Calculating $\frac{\partial E}{\partial z L-2} = \delta^L$

It is important to note that δ is pretty much $\frac{\partial E}{\partial z}$ for all z.
Calculating δ^L is different from calculating δ^{L-1} δ^{L-2} $\ldots$ δ^2

1. You have to calculate $\frac{\partial E}{\partial a}$

2. Then you have to calculate $\frac{\partial E}{\partial a^L} = \frac{y}{a^L} + \frac{(1-y)}{(1-a^L)}$ and $\frac{\partial a}{\partial z^L} = (a * (1-a)) = \sigma(z)(1-\sigma(z))$ _

$$
\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] \\
&= \frac{d}{dx}\left(1+\mathrm{e}^{-x}\right)^{-1} \\
&= -(1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \left(1-\frac{1}{1+e^{-x}}\right) \\
&= \sigma(x) \cdot (1-\sigma(x))
\end{aligned}
$$

_

3. $\frac{\partial E}{\partial z^L} = \frac{\partial E}{\partial a^L} * \frac{\partial a^L}{\partial z^L} = \delta^{(L)} = a^L * (1-a^L) * (\frac{y}{a^L} + \frac{(1-y)}{(1-a^L)}) = a^{(L)} - y$

2 Calculating $\frac{\partial E}{\partial z^{L-1}} = \delta^{(L-1)}, \frac{\partial E}{\partial z^{L-2}} = \delta^{(L-2)}, \ldots, \frac{\partial E}{\partial z^{L-2}} = \delta^{(2)}$ GENERAL CASE

Reminder: In each level l derivative we can have multiple δ^l, like we have multiple z^l

Calculation is done somewhat recursively. For every smaller level δ^l we need to use δ^{l+1} in our calculation of the derivative because of the chain derivative rule: $\frac{d}{dx}\left[f\left(u\right)\right] = \frac{d}{du}\left[f\left(u\right)\right]\frac{du}{dx}$. If, inside the formula we go further from the output towards the input we need to always take functions in between under consideration.

In other words

$\frac{\partial E}{z^l} = \frac{\partial E}{z^{l+1}}\ \frac{\partial z^{l+1}}{a^l}\frac{\partial a^l}{z^l}$

This translates to

$\frac{\partial E}{\partial Z^l} = \delta^{(l)} = ((\Theta^{(l)})^T\delta^{(l+1)})\ \ .*\ \ a^{(l)}\ .*\ \ (1 - a^{(l)})$

Where each part stands for

+ $((\Theta^{(l)})^T\delta^{(l+1)}) = \frac{\partial z^{l+1}}{\partial a^l}$ + This rule is somewhat more interesting, because it applies a somewhat complicated concept where 1 variable in a function affects the end result through multiple other functions. E.g $E = E(d(x), u(x))$ Such a function, when taking a derivative is solved by just summing up all the derivatives. **This is somewhat intuitive but I wont go to deep into it**
+ Explanation in Estonian

Olgu $w = f\left(u, v\right)$ kahe muutuja funktsioon määramispiirkonnaga Q, kus argumendid u, v on omakorda kahe muutuja funktsioonid

$$u = \varphi_1\left(x, y\right),\ v = \varphi_2\left(x, y\right).$$

Eeldame, et funktsioonidel φ_1 ja φ_2 on ühine määramispiirkond D ning $\left(\varphi_1\left(\mathbf{X}\right), \varphi_2\left(\mathbf{X}\right)\right) \in Q$ iga $\mathbf{X} \in D$ korral. Defineerime hulgas D liitfunktsiooni F seosega

$$F\left(x, y\right) := f\left(\varphi_1\left(\mathbf{X}\right), \varphi_2\left(\mathbf{X}\right)\right) = f\left(\varphi_1\left(x, y\right), \varphi_2\left(x, y\right)\right). \qquad (3.10)$$

Lause 3.4. *Kui funktsioonidel φ_1 ja φ_2 eksisteerivad punktis $\mathbf{A} = (a, b)$ lõplikud osatuletised $\frac{\partial \varphi_1}{\partial x}\left(\mathbf{A}\right)$ ja $\frac{\partial \varphi_2}{\partial x}\left(\mathbf{A}\right)$ ning funktsioon f on punktis $\mathbf{B} := \left(\varphi_1\left(\mathbf{A}\right), \varphi_2\left(\mathbf{A}\right)\right)$ diferentseeruv, siis liitfunktsioonil F on punktis $\mathbf{A}$ osatuletis*

$$\frac{\partial F}{\partial x}\left(\mathbf{A}\right) = \frac{\partial f}{\partial u}\left(\mathbf{B}\right)\frac{\partial \varphi_1}{\partial x}\left(\mathbf{A}\right) + \frac{\partial f}{\partial v}\left(\mathbf{B}\right)\frac{\partial \varphi_2}{\partial x}\left(\mathbf{A}\right).$$

+ $a^{(l)}\ .*\ \ (1 - a^{(l)}) = \frac{\partial a^l}{z^l}$

4. Step - Calculating $\frac{\partial E}{\partial \theta^l}$

$\frac{\partial E}{\partial \theta_{ij}^l} = \frac{\partial E}{z_i^{l+1}}\frac{\partial z_i^{l+1}}{\partial \theta_{ij}^l}$

This in regular derivative form translates to

$\frac{\partial E}{\partial \theta_{ij}^l} = \delta_i^{l+1}a_j^l$

This translates in vectorial form to

How to Check whether Gradient is correct.