

In the lectures notes, one has $\theta_{\text{here}}^{[l]} = \theta_{\text{there}}^{[l-1]}$, where there refers to the machine learning course by Andrew. We do this change to make the connexion with the deep learning specialization lectures.

1 Structure of dense layers

$$a_{\mu}^{[0]} \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \quad (1)$$

$$a_{\mu}^{[l]} \equiv \delta_{\mu 0} + \delta_{\mu i} g^{[l]} \left(\theta_{i\nu}^{[l]} a_{\nu}^{[l-1]} \right) \quad (2)$$

$$z_i^{[l]} \equiv \theta_{i\nu}^{[l]} a_{\nu}^{[l-1]} \quad (3)$$

It should be understood that in $a_i^{[l]}$ one has $i \in \{1, \dots, s_l\}$, and that in $a_{\mu}^{[l]}$ one has $\mu \in \{0, \dots, s_l\}$. The binary cross entropy cross function is

$$cost = \frac{1}{m} \sum_n^m J \Big|_{x=x_n, y=y_n} \quad (4)$$

where

$$J \equiv - \sum_k^{s_L} \left(y_k \log \left(a_k^{[L]} \right) + (1 - y_k) \log \left(1 - a_k^{[L]} \right) \right) \quad (5)$$

2 Backward propagation. Recursion relation

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta_{j_l j_{l-1}}^{[l]}}}_{d\theta_{j_l j_{l-1}}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{j_l}^{[l]}}}_{dz_{j_l}^{[l]}} \underbrace{\frac{\partial z_{j_l}^{[l]}}{\partial \theta_{j_l j_{l-1}}^{[l]}}}_{a_{j_{l-1}}^{[l-1]}} \quad (6)$$

$$\underbrace{\frac{\partial J}{\partial z_{j_l}^{[l]}}}_{dz_{j_l}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{j_{l+1}}^{[l+1]}} \frac{\partial z_{j_{l+1}}^{[l+1]}}{\partial a_{j_l}^{[l]}}}_{da_{j_l}^{[l]}} \frac{\partial a_{j_l}^{[l]}}{\partial z_{j_l}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{j_{l+1}}^{[l+1]}} \theta_{j_{l+1} j_l}^{[l+1]}}_{da_{j_l}^{[l]} = (\theta^{[l+1]T})_{j_l j_{l+1}} dz_{j_{l+1}}^{[l+1]}} g^{[l]'} \left(z_{j_l}^{[l]} \right) \quad (7)$$

Hence, in an abbreviated notation one has:

$$d\theta_{j_l j_{l-1}}^{[l]} = dz_{j_l}^{[l]} a_{j_{l-1}}^{[l-1]} \quad (8)$$

$$da_{j_{l-1}}^{[l-1]} = (\theta^{[l]T})_{j_{l-1} j_l} dz_{j_l}^{[l]} \quad (9)$$

$$dz_{j_l}^{[l]} = da_{j_l}^{[l]} g^{[l]'} \left(z_{j_l}^{[l]} \right) \quad (10)$$

Hence, by providing $\{da^{[l]}, \theta^{[l]}\}$ one gets $\{d\theta_{jl\mu}^{[l]}, da_{jl-1}^{[l-1]}\}$ and this process can be iterated. The initial condition is

$$d\theta_{jl\mu}^{[L]} = \frac{\partial J}{\partial \theta_{jl\mu}^{[L]}} = \frac{(a_{jL}^{[L]} - y_{jL})}{\underbrace{a_{jL}^{[L]}(1 - a_{jL}^{[L-1]})}_{da_{jL}^{[L]}}} g^{[L]'}(z_{jL}^{[L]}) a_{\mu}^{[L-1]} \quad (11)$$

$$dz_{jL}^{[L]} = da_{jL}^{[L]} g^{[L]'}(z_{jL}^{[L]}) \quad (12)$$

$$d\theta_{jl\mu}^{[L]} = dz_{jL}^{[L]} a_{\mu}^{[L-1]} \quad (13)$$

This layer actually simplifies because one chooses $g^{[L]} = \sigma_{\text{sigmoid}}$, which means that $\sigma'_{\text{sigmoid}} = \sigma_{\text{sigmoid}}(1 - \sigma_{\text{sigmoid}})$.

3 Forward propagation convolutional neural networks

No padding case.

$$a_{ij}^{[0]} \equiv x_{ij} \quad (14)$$

$$z_{ij}^{[l]} = \theta_{rs}^{[l]} a_{(iS^{[l]}+r)(jS^{[l]}+s)}^{[l-1]} + b^{[l]} \delta_{ij} \quad (15)$$

$$a_{ij}^{[l]} \equiv g^{[l]}(z_{ij}^{[l]}) \quad (16)$$

It should be understood that indices of the activation of the l layer, $a_{ij}^{[l]}$, run over $i \in \{0, \dots, n_H^{[l]} - 1\}$ and $j \in \{0, \dots, n_W^{[l]} - 1\}$, where $n_X^{[l]} = (n_X^{[l-1]} - f^{[l]})/S^{[l]} + 1$. Finally, the indices of the weights of the l layer, $\theta_{rs}^{[l]}$, runs over $r, s \in \{0, \dots, f^{[l]} - 1\}$.

4 Back prop

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta_{rs}^{[l]}}}_{d\theta_{rs}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{ij}^{[l]}}}_{dz_{ij}^{[l]}} \times \underbrace{\frac{\partial z_{ij}^{[l]}}{\partial \theta_{rs}^{[l]}}}_{a_{(iS^{[l]}+r)(jS^{[l]}+s)}^{[l]}} \quad (17)$$

$$\underbrace{\frac{\partial J}{\partial z_{ij}^{[l]}}}_{dz_{ij}^{[l]}} = \frac{\partial J}{\partial z_{uv}^{[l+1]}} \underbrace{\frac{\partial z_{uv}^{[l+1]}}{\partial a_{(uS^{[l]}+r)(vS^{[l]}+s)}^{[l]}} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}}_{da_{ij}^{[l]}} \frac{\partial a_{ij}^{[l]}}{\partial z_{ij}^{[l]}} \quad (18)$$

$$= \underbrace{\frac{\partial J}{\partial z_{uv}^{[l+1]}} \theta_{rs}^{[l+1]} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}}_{da_{ij}^{[l]} = \theta_{rs}^{[l+1]} dz_{uv}^{[l+1]} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}} g^{[l]'}(z_{ij}) \quad (19)$$

Hence, in an abbreviated notation one has:

$$d\theta_{j\mu}^{[l]} = dz_{ji}^{[l]} a_{ji-1}^{[l-1]} \quad (20)$$

$$da_{ji-1}^{[l-1]} = (\theta_{jl-1}^{[l]})_{jl-1} dz_{ji}^{[l]} \quad (21)$$

$$dz_{ji}^{[l]} = da_{ji}^{[l]} g^{[l]'}(z_{ji}^{[l]}) \quad (22)$$

5 Propagation with softmax

$$a_{\mu}^{[0]} \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \quad (23)$$

$$a_{\mu}^{[l]} \equiv \delta_{\mu 0} + \delta_{\mu i}^{[l]} g^{[l]}(\theta_{i\nu}^{[l]} a_{\nu}^{[l-1]}), \quad l < L \quad (24)$$

$$z_i^{[l]} \equiv \theta_{i\nu}^{[l]} a_{\nu}^{[l-1]}, \quad l \leq L \quad (25)$$

$$\hat{y}_i \equiv \frac{e^{z_i^{[L]}}}{\sum_c e^{z_c^{[L]}}} \quad (26)$$

It should be understood that in $a_i^{[l]}$ one has $i \in \{1, \dots, s_l\}$, and that in $a_{\mu}^{[l]}$ one has $\mu \in \{0, \dots, s_l\}$. The loss function is

$$J \equiv - \sum_i y_i \log(\hat{y}_i) \quad (27)$$

6 Back propagation

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta_{ji}^{[l]}}}_{d\theta_{ji}^{[l]}} = \frac{\partial J}{\partial z_{ji}^{[l]}} \underbrace{\frac{\partial z_{ji}^{[l]}}{\partial \theta_{ji}^{[l]}}}_{\frac{\partial \theta_{ji}^{[l]}}{\partial \theta_{ji}^{[l]}}} \underbrace{\frac{\partial \theta_{ji}^{[l]}}{\partial \theta_{ji}^{[l-1]}}}_{a_{ji-1}^{[l-1]}} \quad (28)$$

$$\underbrace{\frac{\partial J}{\partial z_{j_l}^{[l]}}}_{dz_{j_l}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{j_{l+1}}^{[l+1]}} \frac{\partial z_{j_{l+1}}^{[l+1]}}{\partial a_{j_l}^{[l]}}}_{da_{j_l}^{[l]}} \frac{\partial a_{j_l}^{[l]}}{\partial z_{j_l}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{j_{l+1}}^{[l+1]}} \theta_{j_{l+1} j_l}^{[l+1]}}_{da_{j_l}^{[l]} = (\theta^{[l+1]T})_{j_l j_{l+1}} dz_{j_{l+1}}^{[l+1]}} g^{[l]'} \left(z_{j_l}^{[l]} \right) \quad (29)$$

Hence, in an abbreviated notation one has:

$$d\theta_{j_l j_{l-1}}^{[l]} = dz_{j_l}^{[l]} a_{j_{l-1}}^{[l-1]} \quad (30)$$

$$da_{j_{l-1}}^{[l-1]} = (\theta^{[l]T})_{j_{l-1} j_l} dz_{j_l}^{[l]} \quad (31)$$

$$dz_{j_l}^{[l]} = da_{j_l}^{[l]} g^{[l]'} \left(z_{j_l}^{[l]} \right) \quad (32)$$

Hence, by providing $\{da^{[l]}, \theta^{[l]}\}$ one gets $\{d\theta_{j_l \mu}^{[l]}, da_{j_{l-1}}^{[l-1]}\}$ and this process can be iterated. The initial condition is

$$d\theta_{j_L j_{L-1}}^{[L]} = \frac{\partial J}{\partial \theta_{j_L \mu}^{[L]}} = \left(-\frac{y_i}{\hat{y}_i} \right) \left(\frac{\partial \hat{y}_i}{\partial z_k^{[L]}} \right) \left(\frac{\partial z_k^{[L]}}{\partial \theta_{j_L j_{L-1}}^{[L]}} \right) \quad (33)$$

$$\frac{\partial \hat{y}_i}{\partial z_k^{[L]}} = \hat{y}_i (\delta_{ik} - \hat{y}_k) \quad (34)$$

$$\frac{\partial z_j^{[L]}}{\partial \theta_{j_L j_{L-1}}^{[L]}} = a_{j_{L-1}}^{L-1} \quad (35)$$

7 Old, don't look!

8 Machine Learning coursera Andrew NG

In the lectures notes, one has $\theta_{\text{here}}^{[l]} = \theta_{\text{there}}^{[l-1]}$. We do this change to make the connexion with the deep learning specialization lectures.

$$a_{\mu}^{[0]} \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \quad (36)$$

$$a_{\mu}^{[l]} \equiv \delta_{\mu 0} + \delta_{\mu i}^{[l]} g^{[l]} \left(\theta_{i\nu}^{[l]} a_{\nu}^{[l-1]} \right) \quad (37)$$

$$z_i^{[l]} \equiv \theta_{i\nu}^{[l]} a_{\nu}^{[l]} \quad (38)$$

It should be understood that in $a_i^{[l]}$ one has $i \in \{1, \dots, s_l\}$, and that in $a_\mu^{[l]}$ one has $\mu \in \{0, \dots, s_l\}$.

The cost function has the structure

$$cost = \frac{1}{m} \sum_n^m J \Big|_{x=x_n, y=y_n} \quad (39)$$

where

$$J \equiv - \sum_k^{s_L} \left(y_k \log \left(a_k^{[L]} \right) + (1 - y_k) \log \left(1 - a_k^{[L]} \right) \right) \quad (40)$$

To implement gradient descent we need the next partial derivatives ($l \in \{1, \dots, L\}$)

$$\frac{\partial J}{\partial \theta_{ji\mu}^{[l]}} = \frac{\partial J}{\partial a_{jL}^{[L]}} \left(\frac{\partial a_{jL}^{[L]}}{\partial a_{jL-1}^{[L-1]}} \dots \frac{\partial a_{jL+1}^{[l+1]}}{\partial a_{ji}^{[l]}} \right) \frac{\partial a_{ji}^{[l]}}{\partial z_{ji}^{[l]}} \frac{\partial z_{ji}^{[l]}}{\partial \theta_{ji\mu}^{[l]}} \quad (41)$$

through simple partial derivation one can show that

$$\frac{\partial J}{\partial a_{jL}^{[L]}} = \frac{(a_{jL}^{[L]} - y_{jL})}{a_{jL}^{[L]}(1 - a_{jL}^{[L]})} \quad (42)$$

$$\frac{\partial z_{ji}^{[l]}}{\partial \theta_{ji\mu}^{[l]}} = a_\mu^{[l-1]} \quad (43)$$

$$\frac{\partial a_{ji}^{[l]}}{\partial a_{ji-1}^{[l-1]}} = \left(\frac{\partial a_{ji}^{[l]}}{\partial z_{ji}^{[l]}} \right) \left(\frac{\partial z_{ji}^{[l]}}{\partial a_{ji-1}^{[l-1]}} \right) = (a_{ji}^{[l]'}) (\theta_{ji-1}^{[l]}) \quad (44)$$

$$\frac{\partial a_{ji}^{[l]}}{\partial z_{ji}^{[l]}} = a_{ji}^{[l]'} \quad (45)$$

Then, we can write

$$\frac{\partial J}{\partial \theta_{ji\mu}^{[l]}} = \frac{(a_{jL}^{[L]} - y_{jL})}{a_{jL}^{[L]}(1 - a_{jL}^{[L-1]})} \left(\frac{\partial a_{jL}^{[L]}}{\partial a_{jL-1}^{[L-1]}} \dots \frac{\partial a_{jL+1}^{[l+1]}}{\partial a_{ji}^{[l]}} \right) a_{ji}^{[l]'} a_\mu^{[l-1]} \quad (46)$$

9 Recursion relation

Let us write

$$\frac{\partial J}{\partial \theta_{ji\mu}^{[l]}} = \frac{\partial J}{\partial a_{ji}^{[l]}} \frac{\partial a_{ji}^{[l]}}{\partial z_{ji}^{[l]}} \frac{\partial z_{ji}^{[l]}}{\partial \theta_{ji\mu}^{[l]}} \quad (47)$$

since

$$\frac{\partial J}{\partial a_{j_l}^{[l]}} = \frac{\partial J}{\partial a_{j_{l+1}}^{[l+1]}} \frac{\partial a_{j_{l+1}}^{[l+1]}}{\partial a_{j_l}^{[l]}} = \frac{\partial J}{\partial a_{j_{l+1}}^{[l+1]}} \left(\frac{\partial a_{j_{l+1}}^{[l+1]}}{\partial z_{j_{l+1}}^{[l+1]}} \right) \left(\frac{\partial z_{j_{l+1}}^{[l+1]}}{\partial a_{j_l}^{[l]}} \right) = \frac{\partial J}{\partial a_{j_{l+1}}^{[l+1]}} a_{j_{l+1}}^{[l+1]'} \theta_{j_{l+1}j_l}^{[l+1]} \quad (48)$$

In Andrew's program they use the following notation:

$$\underbrace{\frac{\partial J}{\partial a_{j_l}^{[l]}}}_{dz_{j_l}^{[l]}} = \underbrace{\theta_{j_{l+1}j_l}^{[l+1]} \frac{\partial J}{\partial a_{j_{l+1}}^{[l+1]}}}_{da_{j_l}^{[l]} = (\theta^{[l+1]T})_{j_l j_{l+1}}}_{a_{j_{l+1}}^{[l+1]'}} \quad (49)$$