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

1 Structure of dense layers

$$a^{[0]}_\mu \equiv \delta_{0\mu} + x_i \delta_{i\mu} \tag{1}$$

$$a^{[l]}_\mu \equiv \delta_{0\mu} + g^{[l]}\left(a^{[l-1]}_\nu \theta^{[l]}_{\nu i}\right)\ \delta^{[l]}_{i\mu} \tag{2}$$

$$z^{[l]}_i \equiv a^{[l-1]}_\nu \theta^{[l]}_{\nu i} \tag{3}$$

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

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

where

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

2 Backward propagation. Recursion relation

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta^{[l]}_{j_{l-1}j_l}}}_{\mathrm{d}\theta^{[l]}_{j_{l-1}j_l}} = \underbrace{\frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{j_{l-1}j_l}}}_{a^{[l-1]}_{j_{l-1}}}\underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} \tag{6}$$

$$\underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}}\frac{\partial z^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}}}_{\mathrm{d}a^{[l]}_{j_l}}\frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}}\theta^{[l+1]}_{j_l j_{l+1}}}_{\mathrm{d}a^{[l]}_{j_l}=\theta^{[l+1]}_{j_l j_{l+1}}\mathrm{d}z^{[l+1]}_{j_{l+1}}}\ g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{7}$$

Hence, in an abbreviated notation one has:

$$\mathrm{d}\theta^{[l]}_{j_{l-1}j_l} = \mathrm{d}z^{[l]}_{j_l} a^{[l-1]}_{j_{l-1}} \tag{8}$$

$$\mathrm{d}a^{[l]}_{j_l} = \theta^{[l+1]}_{j_l j_{l+1}}\mathrm{d}z^{[l+1]}_{j_{l+1}} \tag{9}$$

$$\mathrm{d}z^{[l]}_{j_l} = \mathrm{d}a^{[l]}_{j_l}\ g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{10}$$

The initial condition is

$$\mathrm{d}\theta^{[L]}_{\mu j_L} = \frac{\partial J}{\partial \theta^{[L]}_{\mu j_L}} = \underbrace{\frac{(a^{[L]}_{j_L} - y_{j_L})}{a^{[L]}_{j_L}(1 - a^{[L-1]}_{j_L})}}_{\mathrm{d}a^{[L]}_{j_L}} \; g^{[L]\prime}\left(z^{[L]}_{j_L}\right) \, a^{[L-1]}_{\mu} \tag{11}$$

$$\mathrm{d}z^{[L]}_{j_L} = \mathrm{d}a^{[L]}_{j_L} \, g^{[L]\prime}\left(z^{[L]}_{j_L}\right) \tag{12}$$

$$\mathrm{d}\theta^{[L]}_{j_L \mu} = \mathrm{d}z^{[L]}_{j_L} \, a^{[L-1]}_{\mu} \tag{13}$$

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 Propagation with softmax

$$a^{[0]}_{\mu} \equiv \delta_{0\mu} + x_i \delta_{i\mu} \tag{14}$$

$$a^{[l]}_{\mu} \equiv \delta_{\mu 0} + \delta^{[l]}_{\mu i} \, g^{[l]}\left(a^{[l-1]}_{\nu} \theta^{[l]}_{\nu i}\right), \quad l < L \tag{15}$$

$$z^{[l]}_{i} \equiv a^{[l-1]}_{\nu} \theta^{[l]}_{\nu i}, \quad l \leq L \tag{16}$$

$$\hat{y}_i \equiv \frac{e^{z^{[L]}_i}}{\sum_c e^{z^{[L]}_c}} \tag{17}$$

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

$$J \equiv -\sum_i y_i \log\left(\hat{y}_i\right) \tag{18}$$

4 Back propagation

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta^{[l]}_{j_{l-1} j_l}}}_{\mathrm{d}\theta^{[l]}_{j_{l-1} j_l}} = \underbrace{\frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{j_{l-1} j_l}}}_{a^{[l-1]}_{j_{l-1}}} \underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} \tag{19}$$

$$\underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}}\frac{\partial z^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}}}_{\mathrm{d}a^{[l]}_{j_l}}\frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}}\theta^{[l+1]}_{j_l j_{l+1}}}_{\mathrm{d}a^{[l]}_{j_l}=\theta^{[l+1]}_{j_l j_{l+1}}\mathrm{d}z^{[l+1]}_{j_{l+1}}} \; g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{20}$$

Hence, in an abbreviated notation one has:

$$\mathrm{d}\theta^{[l]}_{j_{l-1}j_l} = \mathrm{d}z^{[l]}_{j_l} a^{[l-1]}_{j_{l-1}} \tag{21}$$

$$\mathrm{d}a^{[l]}_{j_l} = \theta^{[l+1]}_{j_l j_{l+1}}\mathrm{d}z^{[l+1]}_{j_{l+1}} \tag{22}$$

$$\mathrm{d}z^{[l]}_{j_l} = \mathrm{d}a^{[l]}_{j_l} \; g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{23}$$

The initial condition is ($\hat{y}_i = g(z_i)$)

$$\mathrm{d}\theta^{[L]}_{j_{L-1}j_L} = \tfrac{\partial J}{\partial \theta^{[L]}_{j_{L-1}j_L}} = \left(-\tfrac{y_i}{\hat{y}_i}\right)\left(\tfrac{\partial \hat{y}_i}{\partial z^{[L]}_{j_L}}\right)\left(\tfrac{\partial z^{[L]}_{j_L}}{\partial \theta^{[L]}_{j_{L-1}j_L}}\right) \tag{24}$$

$$\frac{\partial \hat{y}_i}{\partial z^{[L]}_{j_L}} = \hat{y}_i(\delta_{i j_L} - \hat{y}_{j_L}) \tag{25}$$

$$\frac{\partial z^{[L]}_{j_L}}{\partial \theta^{[L]}_{j_{L-1}j_L}} = a^{L-1}_{j_{L-1}} \tag{26}$$

Note that after summing over i, the first two terms in (24) simplify as:

$$\left(-\frac{y_i}{\hat{y}_i}\right)\left(\frac{\partial \hat{y}_i}{\partial z^{[L]}_{j_L}}\right) = -y_i + \hat{y}_i \tag{27}$$

****