Soft Max Regression w/ NN

* Output layer's activation function will be the soft-max function

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}}$$

where n = n_outputs
= n_categories

From regression,

$$p(y=i \mid x;\theta) = \frac{e^{\theta_i^T x}}{\sum_{j=1}^{n} e^{\theta_j^T x}}$$

where θ_k ~ regression parameters

It is common to take $\theta_n = 0$ ($\exists$ $(n-1)$ independent probabilities)

Binary data: $p(y=i \mid x;\theta) = \dfrac{e^{\theta_i^T x}}{1+e^{\theta_i^T x}} = \dfrac{1}{1+e^{-\theta_i^T x}}$

~ recovers logistic regression

Maximum likelihood

Take $p(y=i \mid x;\theta) = \dfrac{e^{z_i^L}}{\sum_{k=1}^{n} e^{z_i^L}}$

where $\theta \sim$ weights & biases. To determine θ, use maximum (log) likelihood. In other words, take the cost to negative log likelihood.

First, whats the likelihood we got this data

$D = \{(x^{(i)}, y^{(i)}) \mid i = 1, \ldots, N\}$ where $y^{(i)}$'s $\sim$ labels

$$P_\theta(D) = \prod_{i=1}^{N} p(y=y^{(i)} \mid x^{(i)}; \theta)$$

$$p(y \mid x;\theta) = \prod_{i=1}^{n} p(y=i \mid x;\theta)^{1\{y=i\}}$$

$$= \prod_{i=1}^{n} \left[\frac{e^{z_i^L}}{\sum e^{z_i^L}} \right]^{1\{y=i\}}$$

$$L_\theta(D) = \sum_{i=1}^{N} \sum_{j=1}^{n} \log \left[\left(\frac{e^{z_j^L}}{\sum e^{z_j^L}} \right)^{1\{y^{(i)}=j\}} \right]$$

Calculating $\partial_{w_{rs}^{\ell}} L_w(D)$

Write $a_i^L = s(z_i^L)$ & $z_i^L = \sum_k a_k^{L-1} \omega_{ki}^L + b_i^L$

$$\Rightarrow L_\theta(D) = \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} \log s(z_j^L)$$

$$\partial_{w_{rs}^L} L_w(D) = \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} \frac{1}{s(z_j^L)} \frac{\partial s(z_j^L)}{\partial z_k^L} \frac{\partial z_k^L}{\partial w_{rs}^L}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} \frac{1}{s(z_j^L)} \left[\frac{e^{z_j} \delta_{jk}}{\sum_o e^{z_o^L}} - \frac{e^{z_j} e^{z_k}}{(\sum_o e^{z_o^L})^2} \right] \frac{\partial}{\partial w_{rs}^L} \sum_p a_p w_{pk}^{L-1\,L}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} \frac{s(z_j)(\delta_{jk} - s(z_k))}{s(z_j)} \cdot \sum_p a_p^{L-1} \delta_{rp} \delta_{sk}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} (\delta_{jk} - s(z_k)) \cdot a_r^{L-1} \delta_{sk}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n} 1\{y^{(i)}=j\} (\delta_{js} - s(z_s))\, a_r^{L-1}$$

say $y^{(i)} = 2$

$\Sigma = 1$!

$$= \sum_{i=1}^{N} \left(1\{y^{(i)}=s\} - s(z_s) \sum_{j=1}^{n} 1\{y^{(i)}=j\} \right) a_r^{L-1}$$

$$= \sum_{i=1}^{N} a_r^{L-1} \left(1\{y^{(i)}=s\} - s(z_s) \right)$$

$$\boxed{\frac{\partial L_w(D)}{\partial w_{rs}^L} = \sum_{i=1}^{N} \left(1\{y^{(i)} = s\} - s(z_s) \right) a_r^{L-1}}$$

Let's take $y = 1, 2, \ldots, K$ for K classes

& let $y \longrightarrow y_i$ where $y_i^c = \delta_{ic}$ for class c.

Instead of focusing on $s(z_s)$ producing $1\{y^{(i)} = s\}$, will use vectors: $\vec{y}^{(i)} - s(\vec{z})$

$$\frac{\partial L_w(D)}{\partial w^L} = \sum_{i=1}^{N} (\vec{a}^L)^T \left(\vec{y}^{(i)} - s(\vec{z}^L \mid x^{(i)}) \right)$$

<u>Note</u>: $\delta^L = \vec{y} - s(\vec{z}^L)$!