The derivative of LLK score function with the Multinomial Logistic function

reda.dehak

October 2022

1 The Softmax Function:

$$\text{SoftMax}_\Theta(\mathbf{x}, k) = \frac{e^{\theta_{.k}^T \mathbf{x}}}{\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}} \tag{1}$$

where $\theta_{.k}$ is the column number k of the matrix Θ, K is the number of classes

$$\frac{\partial \text{SoftMax}_\Theta(\mathbf{x}, k)}{\partial \theta_{ij}} = \begin{cases} \frac{x_i\, e^{\theta_{.k}^T \mathbf{x}} \sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}} - x_i \left(e^{\theta_{.k}^T \mathbf{x}}\right)^2}{\left(\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}\right)^2} & \text{if } j = k \\ -\frac{x_i\, e^{\theta_{.k}^T \mathbf{x}}\, e^{\theta_{.j}^T \mathbf{x}}}{\left(\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}\right)^2} & \text{Otherwise} \end{cases} \tag{2}$$

$$= \begin{cases} x_i \left(\frac{e^{\theta_{.k}^T \mathbf{x}}}{\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}} - \left(\frac{e^{\theta_{.k}^T \mathbf{x}}}{\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}} \right)^2 \right) & \text{if } j = k \\ -x_i\, \frac{e^{\theta_{.k}^T \mathbf{x}}}{\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}}\, \frac{e^{\theta_{.j}^T \mathbf{x}}}{\sum_{c=1}^{K} e^{\theta_{.c}^T \mathbf{x}}} & \text{Otherwise} \end{cases} \tag{3}$$

$$= \begin{cases} x_i \left(\text{SoftMax}_\Theta(\mathbf{x}, k) - \text{SoftMax}_\Theta(\mathbf{x}, k)^2\right) & \text{if } j = k \\ -x_i\, \text{SoftMax}_\Theta(\mathbf{x}, j)\ \text{SoftMax}_\Theta(\mathbf{x}, k) & \text{Otherwise} \end{cases} \tag{4}$$

$$= \begin{cases} x_i\, \text{SoftMax}_\Theta(\mathbf{x}, k)\ (1 - \text{SoftMax}_\Theta(\mathbf{x}, k)) & \text{if } j = k \\ -x_i\, \text{SoftMax}_\Theta(\mathbf{x}, j)\ \text{SoftMax}_\Theta(\mathbf{x}, k) & \text{Otherwise} \end{cases} \tag{5}$$

$$= \begin{cases} x_i\, \text{SoftMax}_\Theta(\mathbf{x}, k)\ (1 - \text{SoftMax}_\Theta(\mathbf{x}, k)) & \text{if } j = k \\ x_i\, \text{SoftMax}_\Theta(\mathbf{x}, j)\ (0 - \text{SoftMax}_\Theta(\mathbf{x}, k)) & \text{Otherwise} \end{cases} \tag{6}$$

$$\frac{\partial \text{SoftMax}_\Theta(\mathbf{x}, k)}{\partial \theta_{ij}} = x_i\, \text{SoftMax}_\Theta(\mathbf{x}, j) \left(\boxed{\begin{cases} 1 & \text{if } j = k \\ 0 & \text{Otherwise} \end{cases}} - \text{SoftMax}_\Theta(\mathbf{x}, k) \right) \tag{7}$$

$$= \quad x_i \,\mathrm{SoftMax}_\Theta\left(\mathbf{x}, j\right)\,\left(\mathbb{1}_{(j=k)} - \mathrm{SoftMax}_\Theta\left(\mathbf{x}, k\right)\right) \tag{8}$$

So we can generalise to vectors

$$\frac{\partial \mathrm{SoftMax}_\Theta\left(\mathbf{x}, k\right)}{\partial \theta_{.j}} = \mathbf{x}\,\mathrm{SoftMax}_\Theta\left(\mathbf{x}, j\right)\,\left(\mathbb{1}_{(j=k)} - \mathrm{SoftMax}_\Theta\left(\mathbf{x}, k\right)\right) \tag{9}$$

2 NLLK Cost Function

$$\mathrm{NLLK} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{1}_{(y_n=k)}\,\log\left(\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)\right) \tag{10}$$

$$\frac{\partial \mathrm{NLLK}}{\partial \theta_{.j}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{1}_{(y_n=k)}\,\frac{\frac{\partial \mathrm{SoftMax}_\Theta(\mathbf{x}_n,k)}{\partial \theta_{.j}}}{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)} \tag{11}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{1}_{(y_n=k)}\,\frac{\mathbf{x}_n\,\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)\,\left(\mathbb{1}_{(j=k)} - \mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)\right)}{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)} \tag{12}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)\left(\sum_{k=1}^{K}\frac{\mathbb{1}_{(y_n=k)}\,\mathbb{1}_{(j=k)}}{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)} - \sum_{k=1}^{K}\frac{\mathbb{1}_{(y_n=k)}\,\cancel{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)}}{\cancel{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, k\right)}}\right) \tag{13}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)\left(\frac{\mathbb{1}_{(y_n=j)}}{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)} - \sum_{k=1}^{K}\mathbb{1}_{(y_n=k)}\right) \tag{13}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,\cancel{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)}\left(\frac{\mathbb{1}_{(y_n=j)}}{\cancel{\mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)}} - 1\right) \tag{14}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,\left(\mathbb{1}_{(y_n=j)} - \mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right)\right) \tag{15}$$

$$\frac{\partial \mathrm{NLLK}}{\partial \theta_{.j}} = -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,e_{nj} \tag{16}$$

where e_{nj} represents the output error for sample n and class j

$$e_{nj} = \mathbb{1}_{(y_n=j)} - \mathrm{SoftMax}_\Theta\left(\mathbf{x}_n, j\right) \tag{17}$$