Machine learning : Sheet 4

Author : Djordje Zivanovic

1.

$$\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{a}} = -\frac{\partial \log a_y}{\partial \mathbf{a}} \tag{1}$$

$$= \left[-\frac{\partial \log a_y}{\partial a_1}, ..., -\frac{\partial \log a_y}{\partial a_y}, ..., -\frac{\partial \log a_y}{\partial a_C}\right] \tag{2}$$

$$= \left[0, ..., -\frac{1}{a_y}, ..., 0\right] \tag{3}$$

The equation 1 is a definition of the objective function for a given point. The equation 2 is a definition of a gradient of a scalar function (derivative of scalar by a vector). The equation 3 uses property of derivative of a logarithm function.

$$\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{z}} = \frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{a}} \cdot \frac{\partial \mathbf{a}}{\partial \mathbf{z}} \tag{4}$$

From the equation 4 we can see that we need to calculate $\frac{\partial \mathbf{a}}{\partial \mathbf{z}}$

$$\frac{\partial \mathbf{a}}{\partial \mathbf{z}} = \frac{\partial \left[\frac{e^{z_1}}{\sum_{l=1}^{C} e^{z_l}}, ..., \frac{e^{z_i}}{\sum_{l=1}^{C} e^{z_l}}, ..., \frac{e^{z_C}}{\sum_{l=1}^{C} e^{z_l}}\right]}{\partial \mathbf{z}} \tag{5}$$

$$= \begin{bmatrix} \frac{\partial \frac{e^{z_1}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_1} & \frac{\partial \frac{e^{z_1}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_2} & \dots & \frac{\partial \frac{e^{z_1}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_C} \\ \frac{\partial \frac{e^{z_2}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_1} & \frac{\partial \frac{e^{z_2}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_2} & \dots & \frac{\partial \frac{e^{z_2}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_C} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \frac{e^{z_C}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_1} & \frac{\partial \frac{e^{z_C}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_2} & \dots & \frac{\partial \frac{e^{z_C}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_C} \end{bmatrix} \tag{6}$$

$$\tag{7}$$

Let us denote

$$S_i = \frac{e^{z_i}}{\sum_{l=1}^{C} e^{z_l}}$$

Then the equation 6 becomes:

$$\begin{bmatrix} \frac{\partial S_1}{\partial z_1} & \frac{\partial S_1}{\partial z_2} & \dots & \frac{\partial S_1}{\partial z_1} \\ \frac{\partial S_2}{\partial z_1} & \frac{\partial S_2}{\partial z_2} & \dots & \frac{\partial S_2}{\partial z_C} \\ \dots & \dots & \dots & \dots \\ \frac{\partial S_C}{\partial z_1} & \frac{\partial S_C}{\partial z_2} & \dots & \frac{\partial S_C}{\partial z_C} \end{bmatrix} \tag{8}$$

In the equation 8 we notice that we have only two "different" type of derivatives, $\frac{\partial S_i}{\partial z_i}, i \in \{1, ..., C\}$ and $\frac{\partial S_i}{\partial z_j}, i \neq j, i \in \{1, ..., C\}, j \in \{1, ..., C\}$.

$$\frac{\partial S_i}{\partial z_i} = \frac{\frac{e^{z_i}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_i} \tag{9}$$

$$= \frac{e^{z_i} \cdot \sum_{l=1}^{C} e^{z_l} - e^{z_i} e^{z_i}}{\left(\sum_{l=1}^{C} e^{z_l}\right)^2} \tag{10}$$

$$= \frac{e^{z_i} \cdot \left(\sum_{l=1}^{C} e^{z_l} - e^{z_i}\right)}{\left(\sum_{l=1}^{C} e^{z_l}\right)^2} \tag{11}$$

$$= \frac{e^{z_i}}{\sum_{l=1}^{C} e^{z_l}} \frac{\sum_{l=1}^{C} e^{z_l} - e^{z_i}}{\sum_{l=1}^{C} e^{z_l}} \tag{12}$$

$$= S_i \cdot (1 - S_i) \tag{13}$$

$$\frac{\partial S_i}{\partial z_j} = \frac{\partial \frac{e^{z_i}}{\sum_{l=1}^{C} e^{z_l}}}{\partial z_j} \tag{14}$$

$$= -\frac{e^{z_i} \cdot e^{z_j}}{\left(\sum_{l=1}^{C} e^{z_l}\right)^2} \tag{15}$$

$$= -S_i \cdot S_j \tag{16}$$

Finally we have:

$$\frac{\partial \mathbf{a}}{\partial \mathbf{z}} = \begin{bmatrix} \frac{\partial S_1}{\partial z_1} & \frac{\partial S_1}{\partial z_2} & \dots & \frac{\partial S_1}{\partial z_1} \\ \frac{\partial S_2}{\partial z_1} & \frac{\partial S_2}{\partial z_2} & \dots & \frac{\partial S_2}{\partial z_C} \\ \dots & \dots & \dots & \dots \\ \frac{\partial S_C}{\partial z_1} & \frac{\partial S_C}{\partial z_2} & \dots & \frac{\partial S_C}{\partial z_C} \end{bmatrix}$$

$$= \begin{bmatrix} S_1 \cdot (1 - S_1) & -S_1 \cdot S_2 & \dots & -S_1 \cdot S_C \\ -S_2 \cdot S_1 & S_2 \cdot (1 - S_2) & \dots & -S_2 \cdot S_C \\ \dots & \dots & \dots & \dots \\ -S_C \cdot S_1 & -S_C \cdot S_1 & \dots & S_C \cdot (1 - S_C) \end{bmatrix} \tag{17}$$

Further we have:

$$\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{z}} = \frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{a}} \cdot \frac{\partial \mathbf{a}}{\partial \mathbf{z}} \tag{18}$$

$$= \left[0, ..., -\frac{1}{a_y}, ..., 0\right] \cdot \begin{bmatrix} S_1 \cdot (1 - S_1) & -S_1 \cdot S_2 & \dots & -S_1 \cdot S_C \\ -S_2 \cdot S_1 & S_2 \cdot (1 - S_2) & \dots & -S_2 \cdot S_C \\ \dots & \dots & \dots & \dots \\ -S_C \cdot S_1 & -S_C \cdot S_1 & \dots & S_C \cdot (1 - S_C) \end{bmatrix} \tag{19}$$

$$= \left[\frac{S_y \cdot S_1}{a_y}, \frac{S_y \cdot S_2}{a_y}, \dots, \frac{S_y \cdot (S_y - 1)}{a_y}, \dots, \frac{S_y \cdot S_C}{a_y}\right] \tag{20}$$

$$= [S_1, S_2, \dots, S_y - 1 \dots, S_C] \tag{21}$$

From the formula $\frac{\partial \ell}{\partial w_{ij}^2} = \frac{\partial \ell}{\partial z_i^2} \cdot \frac{\partial z_i^2}{\partial w_{ij}^2} = \frac{\partial \ell}{\partial z_i^2} \cdot x_j$ we get:

$$\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{W}} = \left(\mathbf{x} \frac{\partial \ell}{\partial \mathbf{z}}\right)^{\mathrm{T}} \tag{22}$$

$$= \begin{bmatrix} x_1 \cdot S_1 & x_2 \cdot S_1 & \dots & x_D \cdot S_1 \\ x_1 \cdot S_2 & x_2 \cdot S_2 & \dots & x_D \cdot S_2 \\ \dots & \dots & \dots & \dots \\ x_1 \cdot (S_y - 1) & x_2 \cdot (S_y - 1) & \dots & x_D \cdot (S_y - 1) \\ \dots & \dots & \dots & \dots \\ x_1 \cdot S_C & x_2 \cdot S_C & \dots & x_D \cdot S_C \end{bmatrix} \tag{23}$$

$$\tag{24}$$

$$\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{b}} = \frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{b}} \tag{25}$$

If we notice that $\frac{\partial \mathbf{z}}{\partial \mathbf{b}}$ is an identity matrix, from the equation we get that $\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{b}}$ is the same as $\frac{\partial \ell(\mathbf{W}, \mathbf{b}, \mathbf{x}, y)}{\partial \mathbf{z}}$.

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \sum_{i=1}^{N} \frac{\partial \ell(\mathbf{x}_i, y_i, \mathbf{W}_t, \mathbf{b}_t)}{\partial \mathbf{w}}$$

$$\mathbf{b}_{t+1} = \mathbf{b}_t - \sum_{i=1}^{N} \frac{\partial \ell(\mathbf{x}_i, y_i, \mathbf{W}_t, \mathbf{b}_t)}{\partial \mathbf{b}}$$