1 Derivative of Binary Cross-Entropy cost function with respect to a weight w

The BInary Cross-Entropy cost function is defined as:

$$C = \frac{1}{n}\sum_{i=1}^{n} h_i^r \,\log(h_i^p) + (1 - h_i^r)\,\log(1 - h_i^p) \tag{1}$$

Taking the derivative of C with respect to a weight w:

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} h_i^r \frac{1}{h_i^p}\frac{\partial h_i^p}{\partial w} - (1 - h_i^r)\frac{1}{1 - h_i^p}\frac{\partial h_i^p}{\partial w} \tag{2}$$

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} h_i^r \frac{1}{h_i^p}\frac{\partial h_i^p}{\partial w} - (1 - h_i^r)\frac{1}{1 - h_i^p}\frac{\partial h_i^p}{\partial w} \tag{3}$$

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i^p}{\partial w}\frac{h_i^r(1 - h_i^p) - (1 - h_i^r)h_i^p}{h_i^p(1 - h_i^p)} \tag{4}$$

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h_i^p}{\partial w}\frac{h_i^r - y_i^p}{h_i^p(1 - h_i^p)} \tag{5}$$

2 Binary Cross-Entropy cost function with sigmoid activation

$$h_i^p = \sigma(z_i) \tag{6}$$

$$\frac{\partial h_i^p}{\partial w} = \frac{\partial z_i}{\partial w}\sigma'(z_i) \tag{7}$$

$$= \frac{\partial z_i}{\partial w}\sigma(z_i)\big(1 - \sigma(z_i)\big) \tag{8}$$

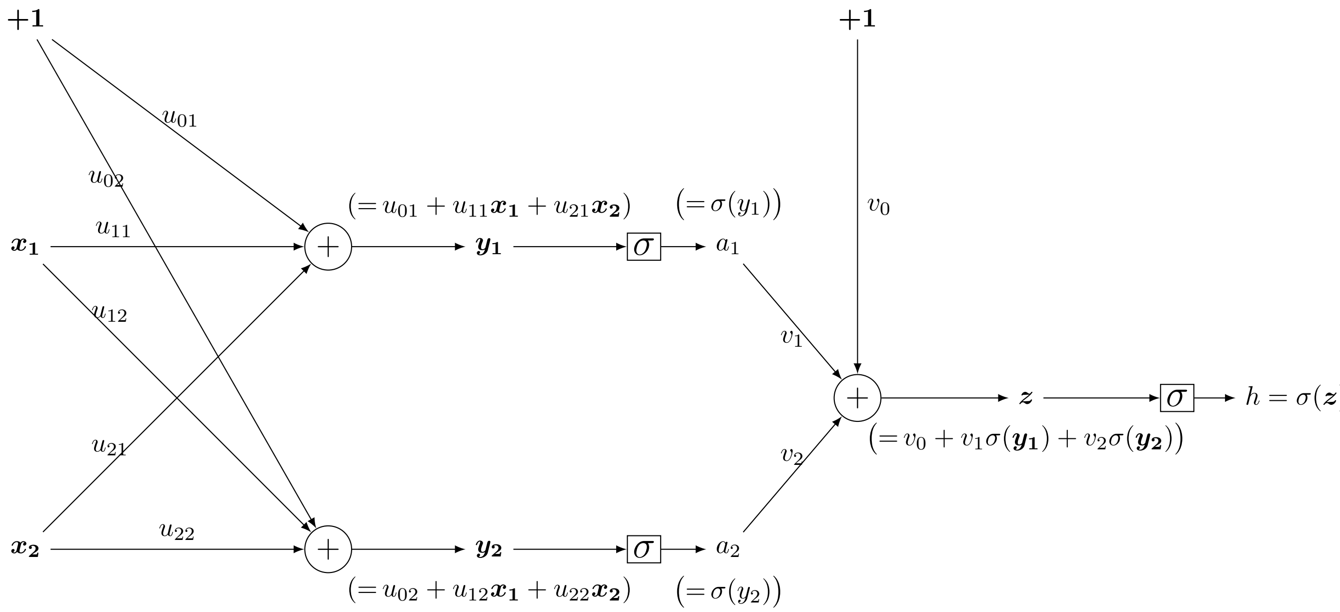
$$= \frac{\partial z_i}{\partial w}h_i^p(1 - h_i^p) \tag{9}$$

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial z_i}{\partial w}h_i^p(1 - h_i^p)\frac{h_i^r - h_i^p}{h_i^p(1 - h_i^p)} \tag{10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial z_i}{\partial w}(h_i^p - h_i^p) \tag{11}$$

3 2-layer architecture with sigmoid activation

Let us suppose we want to use the following linear neural network architecture, with a sigmoid activation function σ:



Let us define the following matrices and operations which help model the above network:

$$\mathbf{X}_0 = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & \mathbf{X}_0 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_{01} & u_{02} \\ u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix}$$

$$\mathbf{Y}_0 = \mathbf{X}\mathbf{u} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \\ y_{41} & y_{42} \end{bmatrix} \quad \mathbf{A}_0 = \sigma\Big(\mathbf{Y}_0\Big) = \begin{bmatrix} \sigma(y_{11}) & \sigma(y_{12}) \\ \sigma(y_{21}) & \sigma(y_{22}) \\ \sigma(y_{31}) & \sigma(y_{32}) \\ \sigma(y_{41}) & \sigma(y_{42}) \end{bmatrix} \quad \begin{bmatrix} 1 & \mathbf{A}_0 \end{bmatrix} = \begin{bmatrix} 1 & a_{11} & a_{12} \\ 1 & a_{21} & a_{22} \\ 1 & a_{31} & a_{32} \\ 1 & a_{41} & a_{42} \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix}$$

$$\mathbf{z} = \mathbf{A}\mathbf{v} \; = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \quad \mathbf{h} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \sigma\Big(\mathbf{z}\Big) = \begin{bmatrix} \sigma(z_1) \\ \sigma(z_2) \\ \sigma(z_3) \\ \sigma(z_4) \end{bmatrix}$$

The final output of the network can also be written as:

$$\begin{aligned} h_i^{pred} &= \sigma\big(v_0 + v_1 a_{i1} + v_2 a_{i2}\big) \\ &= \sigma\Big(v_0 + v_1\sigma\big(u_{01} + u_{11}x_{i1} + u_{21}x_{i2}\big) + v_2\sigma\big(u_{02} + u_{12}x_{i1} + u_{22}x_{i2}\big)\Big) \end{aligned}$$

Let us use the binary cross-entropy cost function as defined in section 1:

$$C = \frac{1}{n}\sum_{i=1}^{n} h_i^r \, \log(h_i^p) + (1 - h_i^r) \, \log(1 - h_i^p) \tag{12}$$

Remembering that $h_i^p = \sigma(z_i)$, the derivative with respect to weight w is:

$$\frac{\partial C}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial z_i}{\partial w}(h_i^p - h_i^p) \tag{13}$$

z_i is given by:

$$\begin{aligned} z_i &= v_0 + v_1 a_{i1} + v_2 a_{i2} \qquad (14) \\ &= v_0 + v_1\sigma\big(u_{01} + u_{11}x_{i1} + u_{21}x_{i2}\big) + v_2\sigma\big(u_{02} + u_{12}x_{i1} + u_{22}x_{i2}\big) \qquad (15) \end{aligned}$$

Using (**??**) we can calculate the derivative of C with respect to each weight u_{ij} and v_i of the network and write them in vector form:

$$
\frac{\partial z_i^{pred}}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial z_i^{pred}}{\partial u_{01}} & \frac{\partial z_i^{pred}}{\partial u_{02}} \\ \frac{\partial z_i^{pred}}{\partial u_{11}} & \frac{\partial z_i^{pred}}{\partial u_{12}} \\ \frac{\partial z_i^{pred}}{\partial u_{21}} & \frac{\partial z_i^{pred}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} v_1 a_{i1}(1-a_{i1}) & v_2 a_{i2}(1-a_{i2}) \\ v_1 x_{i1} a_{i1}(1-a_{i1}) & v_2 x_{i1} a_{i2}(1-a_{i2}) \\ v_1 x_{i2} a_{i1}(1-a_{i1}) & v_2 x_{i2} a_{i2}(1-a_{i2}) \end{bmatrix}
$$

$$
\frac{\partial z_i^{pred}}{\partial \mathbf{v}} = \begin{bmatrix} \frac{\partial z_i^{pred}}{\partial v_0} \\ \frac{\partial z_i^{pred}}{\partial v_1} \\ \frac{\partial z_i^{pred}}{\partial v_2} \end{bmatrix} = \begin{bmatrix} 1 \\ a_{i1} \\ a_{i2} \end{bmatrix}
$$

Let us define d_i such that

$$
d_i = z_i^{reality} - z_i^{pred} \quad \mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix}
$$

By plugging in the corresponding derivatives found above in to (dC), we obtain:

$$
\frac{\partial C}{\partial u_{01}} = v_1 \sum_{i=1}^{n} a_{i1}(1-a_{i1}) d_i \qquad \frac{\partial C}{\partial u_{02}} = v_2 \sum_{i=1}^{n} a_{i2}(1-a_{i2}) d_i
$$

$$
\frac{\partial C}{\partial u_{11}} = v_1 \sum_{i=1}^{n} x_{i1} a_{i1}(1-a_{i1}) d_i \qquad \frac{\partial C}{\partial u_{12}} = v_2 \sum_{i=1}^{n} x_{i1} a_{i2}(1-a_{i2}) d_i
$$

$$
\frac{\partial C}{\partial u_{21}} = v_1 \sum_{i=1}^{n} x_{i2} a_{i1}(1-a_{i1}) d_i \qquad \frac{\partial C}{\partial u_{22}} = v_2 \sum_{i=1}^{n} x_{i2} a_{i2}(1-a_{i2}) d_i
$$

$$
\frac{\partial C}{\partial v_0} = \sum_{i=1}^{n} d_i
$$

$$
\frac{\partial C}{\partial v_1} = \sum_{i=1}^{n} a_{i1} d_i
$$

$$
\frac{\partial C}{\partial v_2} = \sum_{i=1}^{n} a_{i2} d_i
$$

Which can be cast in matrix form as:

$$
\frac{\partial C}{\partial \mathbf{u}} = \left[\text{diag}(\mathbf{d}) \times \mathbf{X}\right]^T \left[(1-\mathbf{A}) \odot \mathbf{A}\right] \times \text{diag}\left(\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)
$$

$$
\frac{\partial C}{\partial \mathbf{v}} = \mathbf{A}^T \mathbf{d}
$$

Where $diag(\mathbf{u})$ denotes the diagonal matrix whose diagonal entries are the entries of vector u.