ml-part1-report

Yıldırım Özen 2521862

November 1, 2024

1 Regression Derivation

The hidden layer output is denoted by z_1 and activated hidden layer output is denoted by a_1.

$w_{ij}^{new} = w_{ij} - \alpha \frac{\partial SE(y,O_o)}{\partial w_{ij}}$

$\frac{\partial SE(y,O_o)}{\partial w_{ij}} = \frac{\partial SE(y,O_o)}{\partial O_o} \cdot \frac{\partial O_o}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{ij}}$

$z_1 = w_{ij}x + b_1$

$a_1 = \frac{1}{1+e^{-z_1}}$

$z_2 = \gamma_{ij}a_1 + b_2$

$O_o = z_2$

$SE(y, O_o) = (y - O_o)^2$

$\frac{\partial z_1}{\partial w_1} = x$

$\frac{\partial a_1}{\partial z_1} = \frac{\partial}{\partial z_1} \left(1 + \mathrm{e}^{-z_1}\right)^{-1}$

$= -(1 + e^{-z_1})^{-2}(-e^{-z_1})$

$= \frac{e^{-z_1}}{(1+e^{-z_1})^2}$

$= \frac{1}{1+e^{-z_1}} \cdot \frac{e^{-z_1}}{1+e^{-z_1}}$

$= \frac{1}{1+e^{-z_1}} \cdot \frac{(1+e^{-z_1})-1}{1+e^{-z_1}}$

$= \frac{1}{1+e^{-z_1}} \cdot \left(\frac{1+e^{-z_1}}{1+e^{-z_1}} - \frac{1}{1+e^{-z_1}}\right)$

$= \frac{1}{1+e^{-z_1}} \cdot \left(1 - \frac{1}{1+e^{-z_1}}\right)$

$= \sigma(z_1) \cdot (1 - \sigma(z_1))$

$\frac{\partial z_2}{\partial a_1} = \gamma_{ij}$

$\frac{\partial O_o}{\partial z_2} = 1$

$\frac{\partial SE(y,O_o)}{\partial O_o} = -2(y - O_o)$

$w_{ij}^{new} = w_{ij} - \alpha. - 2(y - O_o).1.\gamma_{ij}.\sigma(z_1)(1 - \sigma(z_1)).x$

To calculate bias we only need to change the last multiplication term to one:

$\frac{\partial z_1}{\partial w_{0j}} = 1$

$w_{0j}^{new} = w_{0j} - \alpha. - 2(y - O_o).1.\gamma_{ij}.\sigma(z_1)(1 - \sigma(z_1)).x$

Now for Gamma:

$\gamma_{ij}^{new} = \gamma_{ij} - \alpha \frac{\partial SE(y,O_o)}{\partial \gamma_{ij}}$

$\frac{\partial SE(y,O_o)}{\partial w_{ij}} = \frac{\partial SE(y,O_o)}{\partial O_o} \cdot \frac{\partial O_o}{\partial z_2} \cdot \frac{\partial z_2}{\partial \gamma_{ij}}$

$\frac{\partial z_2}{\partial \gamma_{ij}} = a_1$

$\frac{\partial SE(y,O_o)}{\partial w_{ij}} = -2(y - O_o).1.a_1$

$\gamma_{ij}^{new} = \gamma_{ij} + \alpha 2(y - O_o).1.a_1$

To calculate bias we only need to change the last multiplication term to one:

$\frac{\partial z_2}{\partial \gamma_{0j}} = 1$

$\gamma_{0j}^{new} = \gamma_{0j} + \alpha.2(y - O_o)$

2 Classification Derivation

$z_1 = w_{ij}x + b_1$

$a_1 = \frac{1}{1+e^{-z_1}}$

$z_2 = \gamma_{ij}a_1 + b_2$

$O = \frac{e^{z_i}}{\sum_s e^{Z_s}}$

Derivation of Softmax $s_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$:

We first take log for easiness:

$$\log s_i = \log\left(\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_i}}\right)$$
$$= \log\left(e^{x_i}\right) - \log\left(\sum_{j=1}^{C} e^{x_i}\right)$$
$$= x_i - \log\left(\sum_{j=1}^{C} e^{x_i}\right)$$

Differentiating the above with some x_k we get:

$$\frac{\partial \log s_i}{\partial x_k} = \frac{\partial x_i}{\partial x_k} - \frac{\partial \log\left(\sum_{j=1}^{C} e^{x_i}\right)}{\partial x_k}$$

First part of the right hand side can be shown as:

$$\frac{\partial x_i}{\partial x_k} = \begin{cases} 1 & x_i = x_k \\ 0 & \text{otherwise} \end{cases} = \delta_{ik}$$

Second part can be simplified as:

$$\frac{\partial \log\left(\sum_{j=1}^{C} e^{x_i}\right)}{\partial x_k} = \frac{\partial \log\left(\sum_{j=1}^{C} e^{x_j}\right)}{\partial x_k}$$
$$= \frac{1}{\sum_{j=1}^{C} e^{x_j}} \frac{\partial\left(\sum_{j=1}^{C} e^{x_j}\right)}{\partial x_k}$$
$$= \frac{1}{\sum_{j=1}^{C} e^{x_j}} \sum_{j=1}^{C} \frac{\partial e^{x_j}}{\partial x_k}$$
$$= \frac{1}{\sum_{j=1}^{C} e^{x_j}} \sum_{j=1}^{C} e^{x_j} \frac{\partial x_j}{\partial x_k}$$
$$= \frac{1}{\sum_{j=1}^{C} e^{x_j}} \sum_{j=1}^{C} e^{x_j} \delta_{jk}$$
$$= \frac{e^{x_k}}{\sum_{j=1}^{C} e^{x_j}}$$
$$= s_k$$

So the final equation becomes:

$$\frac{\partial \log s_i}{\partial x_k} = \delta_{ik} - s_k$$

$$\frac{1}{s_i} \frac{\partial s_i}{\partial x_k} = \delta_{ik} - s_k$$

$$\frac{\partial s_i}{\partial x_k} = s_i \left(\delta_{ik} - s_k\right)$$

Now we differentiate cross entropy w.r.t some input(x_k) to the softmax:

$\mathcal{L}(s, y) = -\sum_{i=1}^{C} y_i \log s_i$

$\frac{\partial \mathcal{L}(s,y)}{\partial x_k} = -\sum_{i=1}^{C} y_i \frac{\partial \log s_i}{\partial x_k}$

$= -\sum_{i=1}^{C} \frac{y_i}{s_i} \frac{\partial s_i}{\partial x_k}$

From previous softmax derivation:

$= -\sum_{i=1}^{C} \frac{y_i}{s_i} s_i \left(\delta_{ik} - s_k\right)$

$= -\sum_{i=1}^{C} y_i \left(\delta_{ik} - s_k\right)$

$= -\sum_{i=1}^{C} y_i \delta_{ik} + y_i s_k$

$i = k$ when $i = k$ the first term will become y_k:

$\frac{\partial \mathcal{L}(s,y)}{\partial x_k} = -y_k + \sum_{i=1}^{C} y_k s_k$

where $\sum_{i=1}^{C} y_k = 1$ since y_k is a one-hot vector.

$\frac{\partial \mathcal{L}(s,y)}{\partial x_k} = -y_k + s_k$

Now we can use this on our Output:

$\frac{\partial CE(l,O)}{\partial z_i} = O_i - l_i$

$w_{ij}^{new} = w_{ij} - \alpha \frac{\partial CE(l,O)}{\partial w_{ij}}$

$\frac{\partial CE(l,O)}{\partial w_{ij}} = \frac{\partial CE(l,O)}{\partial O} \cdot \frac{\partial O}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{ij}}$

$\frac{\partial z_1}{\partial w_{ij}} = x$

$\frac{\partial a_1}{\partial z_1} = \sigma(z_1).(1 - \sigma(z_1))$

$\frac{\partial z_2}{\partial a_1} = \gamma_{ij}$

$w_{ij}^{new} = w_{ij} - \alpha.(O_i - l_i).\gamma_{ij}.(\sigma(z_1).(1 - \sigma(z_1))).x$

To calculate bias we only need to change the last multiplication term to one:

$\frac{\partial z_1}{\partial w_{0j}} = 1$

$w_{0j}^{new} = w_{0j} - \alpha.(O_i - l_i).\gamma_{ij}.(\sigma(z_1).(1 - \sigma(z_1)))$

Now for Gamma:

$\gamma_{ij}^{new} = \gamma_{ij} - \alpha\frac{\partial CE(l,O)}{\partial \gamma_{ij}}$

$\frac{\partial CE(l,O)}{\partial \gamma_{ij}} = \frac{\partial CE(l,O)}{\partial O}.\frac{\partial O}{\partial z_2}.\frac{\partial z_2}{\partial \gamma_{ij}}$

$\frac{\partial z_2}{\partial \gamma_{ij}} = a_1$

$\gamma_{ij}^{new} = \gamma_{ij} - \alpha.(O_i - l_i).a_1$

To calculate bias we only need to change the last multiplication term to one:

$\frac{\partial z_2}{\partial \gamma 0j} = 1$

$\gamma 0j^{new} = \gamma 0j - \alpha.(O_i - l_i)$