CENG499
INTRODUCTION TO MACHINE LEARNING
THE1

Erdem
Şamlıoğlu
2448843

Part 1

Regression

For the regression part, the hidden layer formula is given as:
$H_j = \sigma\left(\sum_{i=0}^{2} I_i \cdot W_{ij}\right)$
The output layer formula is given as:
$O_0 = \left(\sum_{j=0}^{3} H_j \cdot \gamma_{j0}\right)$
Finally the loss function is given as Mean Squared Error, and the formula is as:
$SE(y, O_0) = (y - O_0)^2$

Now for the backpropagation algorithm, we can calculate the gradients of the MSE with respect to all weights.

For output layer weights, using the chain rule, $\frac{\partial SE}{\partial \gamma_{j0}} = \frac{\partial SE}{\partial O_0} \cdot \frac{\partial O_0}{\partial \gamma_{j0}}$, taking the derivative of SE with respect to O_0 we get: $-2(y - O_0)$. Next, taking the derivative of O_0 with respect to γ_{j0}, we get: H_j.
Then, $\frac{\partial SE}{\partial \gamma_{j0}} = -2(y - O_0) \cdot H_j$.

Weight update rule for the Output Layer is given as: $\gamma_{j0}^{\text{new}} = \gamma_{j0} - \alpha \cdot \frac{\partial SE}{\partial \gamma_{j0}}$. Now we can replace the '$\frac{\partial SE}{\partial \gamma_{j0}}$', so the equation becomes:
$\gamma_{j0}^{\text{new}} = \gamma_{j0} - \alpha \cdot -2(y - O_0) \cdot H_j$

For hidden layer weights, using the chain rule, $\frac{\partial SE}{\partial W_{ij}} = \frac{\partial SE}{\partial O_0} \cdot \frac{\partial O_0}{\partial H_j} \cdot \frac{\partial H_j}{\partial W_{ij}}$. From previous calculations we know that $\frac{\partial SE}{\partial O_0} = -2(y - O_0)$. Now, derivative of the output with respect to H_j, we get: $\frac{\partial O_0}{\partial H_j} = \gamma_{j0}$. Lastly, derivative of hidden layer output with respect to weight W_{ij}, (since we know the $\sigma(x) = \frac{1}{1+e^{-x}}$ and the derivative is equal to $\sigma(x) \cdot (1 - \sigma(x))$ we get: $\frac{\partial H_j}{\partial W_{ij}} = H_j \cdot (1 - H_j) \cdot I_i$.
Then, $\frac{\partial SE}{\partial W_{ij}} = -2(y - O_0) \cdot \gamma_{j0} \cdot H_j \cdot (1 - H_j) \cdot I_i$

Weight update rule for the Hidden Layer is given as: $W_{ij}^{\text{new}} = W_{ij} - \alpha \cdot \frac{\partial SE}{\partial W_{ij}}$. Now we can replace the $\frac{\partial SE}{\partial W_{ij}}$, so the equation becomes:
$W_{ij}^{\text{new}} = W_{ij} - \alpha \cdot -2(y - O_0) \cdot \gamma_{j0} \cdot H_j \cdot (1 - H_j) \cdot I_i$

Classification

For the classification part, hidden layer formula is given as:
$H_j = \sigma \left(\sum_{i=0}^{2} I_i \cdot W_{ij} \right)$
The output layer formula is given as:
$X_k = \sum_{j=0}^{3} H_j \cdot \gamma_{jk}$
The softmax output is given as:
$O_k = \frac{e^{X_k}}{\sum_{s=0}^{2} e^{X_s}}$
Finally the loss function is given as cross entropy loss, and the formula is as:
$CE(l, O) = -\sum_{i=0}^{2} l_i \cdot \log(O_i)$, and where $(l = [l_0, l_1, l_2])(O = [O_0, O_1, O_2])$

Now for the backpropagation algorithm, we can calculate the gradients of CE with respect to weights.

The gradient of the cross-entropy loss with respect to the softmax output O_k is: $\frac{\partial CE}{\partial O_k} = -\frac{l_k}{O_k}$. Now, to find the gradient with respect to the X_k, the derivative of the softmax function with respect to the X_j is: $\frac{\partial O_k}{\partial X_j} = \begin{cases} O_k(1 - O_k) & \text{if } j = k \\ -O_k O_j & \text{if } j \neq k \end{cases}$
Using the chain rule: $\frac{\partial CE}{\partial X_j} = \frac{\partial CE}{\partial O_k} \cdot \frac{\partial O_k}{\partial X_j}$, for the correct class which means $j = k$ it becomes $\frac{\partial CE}{\partial X_k} = O_k - l_k$

For the output layer weights, using the chain rule, $\frac{\partial CE}{\partial \gamma_{jk}} = \frac{\partial CE}{\partial O_k} \cdot \frac{\partial O_k}{\partial X_k} \cdot \frac{\partial X_k}{\partial \gamma_{jk}}$. We already know the $\frac{\partial CE}{\partial X_k} = O_k - l_k$, and for the $\frac{\partial X_k}{\partial \gamma_{jk}}$, taking the derivative with respect to γ_{jk}, we get: $\frac{\partial X_k}{\partial \gamma_{jk}} = H_j$. So, finally $\frac{\partial CE}{\partial \gamma_{jk}} = (O_k - l_k) \cdot H_j$

Weight update rule for the Output Layer is given as: $\gamma_{jk}^{\text{new}} = \gamma_{jk} - \alpha \frac{\partial CE([l_0, l_1, l_2], O = [O_0, O_1, O_2])}{\partial \gamma_{jk}}$, now we can replace it and the equation becomes: $\gamma_{jk}^{\text{new}} = \gamma_{jk} - \alpha \cdot (O_k - l_k) \cdot H_j$

For hidden layer weights, using the chain rule, $\frac{\partial CE}{\partial W_{ij}} = \frac{\partial CE}{\partial O_k} \cdot \frac{\partial O_k}{\partial X_k} \cdot \frac{\partial X_k}{\partial H_j} \cdot \frac{\partial H_j}{\partial W_{ij}}$. We already know that $\frac{\partial CE}{\partial X_k} = O_k - l_k$. Then taking the derivative of X_k with respect to H_j we get $\frac{\partial X_k}{\partial H_j} = \gamma_{jk}$. Finally taking the derivative of H_j with respect to W_{ij}, we get $\frac{\partial H_j}{\partial W_{ij}} = H_j \cdot (1 - H_j) \cdot I_i$.

Weight update rule for the Hidden Layer is given as: $w_{ij}^{\text{new}} = w_{ij} - \alpha \frac{\partial CE([l_0, l_1, l_2], O = [O_0, O_1, O_2])}{\partial w_{ij}}$, now we can replace it and equation becomes:
$W_{ij}^{\text{new}} = W_{ij} - \alpha \sum_{k=0}^{2} (O_k - l_k) \cdot \gamma_{jk} \cdot H_j \cdot (1 - H_j) \cdot I_i$