CS6910 Assignment 2

Vimarsh Sathia
CS17B046.

* Part B

1. • Sigmoid function.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

we have

$$\frac{\partial}{\partial x}(\sigma(x)) = \frac{\partial}{\partial x}\left(\frac{1}{1+e^{-x}}\right) = \frac{-1}{(1+e^{-x})^2}\frac{\partial}{\partial x}(e^{-x})$$

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

We can reframe (1) as

$$\frac{\partial}{\partial x}\sigma(x) = \frac{1+e^{-x}-1}{(1+e^{-x})^2} = \sigma(x)\left(1-\sigma(x)\right) \longrightarrow (2)$$

Now, applying chain rule to find $\frac{\partial}{\partial w}\{\sigma(g(w))\}$ we get

$$\frac{\partial}{\partial w}\{\sigma(g(w))\} = \frac{\partial}{\partial g}\{\sigma(g)\}\frac{\partial g}{\partial w} \quad \text{(by chain rule)}$$

$$= \boxed{\sigma(g)(1-\sigma(g))\frac{\partial g}{\partial w}}$$

- **Hyperbolic tangent**

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

we have

$$\frac{\partial(\tanh(x))}{\partial x} = \frac{\partial}{\partial x}\left\{\frac{e^x - e^{-x}}{e^x + e^{-x}}\right\} = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

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

$$\boxed{\frac{d\tanh(x)}{dx} = 1 - \tanh^2(x)} \longrightarrow (1)$$

Applying chain rule on (1) with $x = g(w)$, we get

$$\frac{\partial}{\partial w}\tanh(g(w)) = \frac{\partial \tanh(g)}{\partial g}\frac{\partial g}{\partial w}$$

$$= \boxed{\left\{1 - \tanh^2(g(w))\right\}\frac{\partial g(w)}{\partial w}} \quad (\text{by } (1))$$

- **ReLU**

we have $\quad \text{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{else.} \end{cases}$

Then, $\frac{\partial}{\partial x}\text{ReLU}(x) = \begin{cases} 1 & x > 0 \\ 0 & \text{else} \end{cases}$

For $x = g(w)$, we have

$$\boxed{\frac{\partial}{\partial w}\text{ReLU}(g(w)) = \begin{cases} 1 \cdot \frac{\partial g(w)}{\partial w} & g(w) > 0 \\ 0 & \text{else} \end{cases}}$$

2. Gradients of common loss functions

- **Cross entropy loss**

For vector input $\bar{a}$, we define $S(\bar{a}): \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_c \end{bmatrix} \rightarrow \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_c \end{bmatrix}$

as $$S_i = \frac{e^{a_i}}{\sum_{k=1}^{c} e^{a_k}} \quad \left(\text{softmax function}\right)$$

Then, for S_i, we have

$$\frac{\partial S_i}{\partial a_i} = \frac{e^{a_i}}{\sum_k e^{a_k}} + \frac{-1\,(e^{a_i})\, e^{a_i}}{\left(\sum_k e^{a_k}\right)^2} = \underline{\underline{S_i\,(1 - S_i)}} \longrightarrow (1)$$

and (for $i \neq j$)

$$\frac{\partial S_i}{\partial a_j} = \frac{-e^{a_i} e^{a_j}}{\left(\sum_k e^{a_k}\right)^2} = \underline{\underline{-S_i S_j}} \longrightarrow (2)$$

Combining (1) and (2), where $\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & \text{else} \end{cases}$

$$\boxed{\frac{\partial S_i}{\partial a_j} = S_i\,(\delta_{ij} - S_j)} \longrightarrow (3)$$

Now, ~~for~~ cross entropy loss is defined as follows

$$J(y, \hat{y}) = -\sum_{k=1}^{c} y_i \log \; -\vec{y} \cdot \log(S(\vec{\hat{y}})) \longrightarrow (4)$$

where $y \Rightarrow$ one hot encoded vector corresponding to correct class of sample.

$\hat{y} \Rightarrow$ scores ~~for~~ ~~ea~~ vector for sample y (one score for each class.)

From ④, we have

$$J(y, \hat{y}) = -\sum_{k=1}^{c} y_k \log(S_k(\hat{y})) \quad \rightarrow ⑤$$

[using definition of softmax]

Let $\hat{y}_i = g(w_i)$. We want to find the gradient $\frac{\partial J}{\partial w_i}$. This can be found using ⑤ as

$$\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w_i} = \frac{\partial g}{\partial w} \left\{ \frac{\partial}{\partial \hat{y}_i} \left\{ -\sum_k y_k \log(S_k(\hat{y})) \right\} \right\}$$

$$= \left\{ -\sum_{k=1}^{c} \left[\frac{y_k}{S_k(\hat{y})} \frac{\partial S_k(\hat{y})}{\partial \hat{y}_i} \right] \right\} \frac{\partial g(w_i)}{\partial w_i}$$

$$= -\sum_{k=1}^{c} \left\{ \frac{y_k}{S_k(\hat{y})} S_k(\hat{y}) (\delta_{kj} - S_j(\hat{y})) \right\} \frac{\partial g}{\partial w}$$

[from ③]

$$\therefore \boxed{\frac{\partial J}{\partial w_i} = -\sum_{k=1}^{c} y_k (\delta_{kj} - S_j(\hat{y})) \frac{\partial g(w_i)}{\partial w_i}}$$

where C is the no. of classes.

and $\hat{y}_i = g(w_i)$

If there are n samples, where for every ~~ith~~ sample, $\hat{y}_j = g(w_j, x^{(j)})$, then

$$\boxed{\frac{\partial J}{\partial w_j} = -\sum_{i=1}^{n} \left\{ \left(\sum_{k=1}^{c} y_k^{(i)} (\delta_{kj} - S_j(\hat{y}^{(i)})) \right) \frac{\partial g(w, x^{(j)})}{\partial w} \right\}}$$

Basically, loss is summed up over all samples.

Hinge loss

We assume n samples are given and the classification is binary (classes are -1 or 1).

In this case, we can define the hinge loss as

$$J_{\xi} = \sum_{i=1}^{n} \max(0, 1 - y_i \hat{y}_i)$$

Assume $\hat{y}_i = g(w, x_i)$ [$x_i \rightarrow$ inputs of i^{th} sample]

then

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{\partial}{\partial w} \{ \max(0, 1 - y_i \hat{y}_i) \}$$

$$\Rightarrow \frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w} \quad \text{(by chain rule)}$$

$$\Rightarrow \boxed{\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{\partial J}{\partial \hat{y}_i} \frac{\partial g(w, x_i)}{\partial w}} \longrightarrow (1)$$

where

$$\boxed{\frac{\partial J}{\partial \hat{y}_i} = \begin{cases} -y_i & y_i \hat{y}_i < 1 \\ 0 & \text{else} \end{cases}} \longrightarrow (2)$$

Substituting (2) in (1), we can find the gradients wrt the weights.

- **L1 loss** ~~for n samples~~

We assume y_i is the correct output for input x_i, and $\hat{y}_i$ is the output of our model.

Then for n samples, L1 loss

$$J = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

We assume $\hat{y}_i = g(w, x_i)$

Then, by chain rule,

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w} \longrightarrow (1)$$

We have

$$\frac{\partial J}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i} \left(\sum_{k=1}^{n} |y_k - \hat{y}_k| \right) = \frac{-|y_i - \hat{y}_i|}{(y_i - \hat{y}_i)}$$

$$= + \frac{|\hat{y}_i - y_i|}{\hat{y}_i - y_i} \longrightarrow (2)$$

Substituting (2) in (1), we get

$$\boxed{\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{\hat{y}_i - y_i} \frac{\partial g(w, x_i)}{\partial w}} \leftarrow \text{Ans}$$

where $\hat{y}_i = g(w, x_i)$

- **Huber loss**

Ex Assuming same notation as before, for sample i, we assume Huber loss.

$$J_i = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{for } |y_i - \hat{y}_i| \le \delta \\ \delta|y_i - \hat{y}_i| - \frac{\delta^2}{2} & \text{else} \end{cases}$$

and total loss $J = \sum_{i=1}^{n} J_i$

Assume $y_i = g(w, x_i)$

Then gradient $\frac{\partial J}{\partial w}$ can be found as

$$\frac{\partial J}{\partial w} = \frac{\partial J_1}{\partial w} + \frac{\partial J_2}{\partial w} + \cdots + \frac{\partial J_n}{\partial w}$$
$$= \sum_{i=1}^{n} \frac{\partial J_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w} \rightarrow (1) \quad (\text{by chain rule})$$

we have

$$\frac{\partial J_i}{\partial \hat{y}_i} = \begin{cases} \hat{y}_i - y_i & |y_i - \hat{y}_i| \le \delta \\ -\frac{\delta(y_i - \hat{y}_i)}{|y_i - \hat{y}_i|} & \text{else} \end{cases} \rightarrow (2)$$

and $\frac{\partial \hat{y}_i}{\partial w} = \frac{\partial g(w, x_i)}{\partial w} \rightarrow (3)$

Substituting (2)(3) in (1), we get the gradient w.r.t. w.

- **L2 loss** -

For a batch of n samples, we define L2 loss as

$$J = \sum_{i=1}^{n} \frac{1}{2} (y_i - \hat{y}_i)^2$$

we assume $\hat{y}_i = g(w, x_i)$ (w is one of the variable weights)

Then

$$\frac{\partial J}{\partial w} = -\sum_{i=1}^{n} \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w} \quad \text{(chain rule)}$$

$$= \sum_{i=1}^{n} \frac{2}{2} (y_i - \hat{y}_i)(-1) \frac{\partial}{\partial w} g(w, x_i)$$

$$\Rightarrow \boxed{\frac{\partial J}{\partial w} = \sum_{i=1}^{n} (\hat{y}_i - y_i) \frac{\partial g(w, x_i)}{\partial w}}$$

- **Cosine similarity :**

We assume $y^{(i)}$ is a one-hot encoded vector for C classes, and $\hat{y}^{(i)}$ is a score vector.

Then the loss function $J(y, \hat{y})$ can be defined as:

$$J = \sum_{i=1}^{n} \left(- \frac{\sum_{k=1}^{C} y_k^{(i)} \hat{y}_k^{(i)}}{|\hat{y}^{(i)}|} \right) \longrightarrow (1)$$

where $|\hat{y}^{(i)}|$ denotes absolute value of vector $\hat{y}^{(i)}$. Now assume $y_j^{(i)} = g(w_j, x^{(i)})$.

~~From (1), we have~~

Then

$$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial}{\partial w_j} \left\{ - \frac{\sum_{k=1}^{C} y_k^{(i)} \hat{y}_k^{(i)}}{|\hat{y}^{(i)}|} \right\}$$

$$\Rightarrow \frac{\partial J}{\partial w_k} = \sum_{i=1}^{n} \frac{\partial}{\partial \hat{y}_j^{(i)}} \left\{ - \frac{\sum_{k} y_k^{(i)} \hat{y}_k^{(i)}}{|\hat{y}^{(i)}|} \right\} \frac{\partial \hat{y}_j^{(i)}}{\partial w_j}$$

$$\frac{\partial J}{\partial w_k} = - \sum_{i=1}^{n} \left[\left\{ \frac{y_j^{(i)}}{|\hat{y}^{(i)}|} - \frac{\left(\sum_{k=1}^{C} y_k^{(i)} \hat{y}_k^{(i)} \right) \hat{y}_j^{(i)}}{|\hat{y}^{(i)}|^3} \right\} \frac{\partial g(w_j, x^{(i)})}{\partial w_j} \right] \longrightarrow (2)$$

(2) is the final gradient after chain rule application.

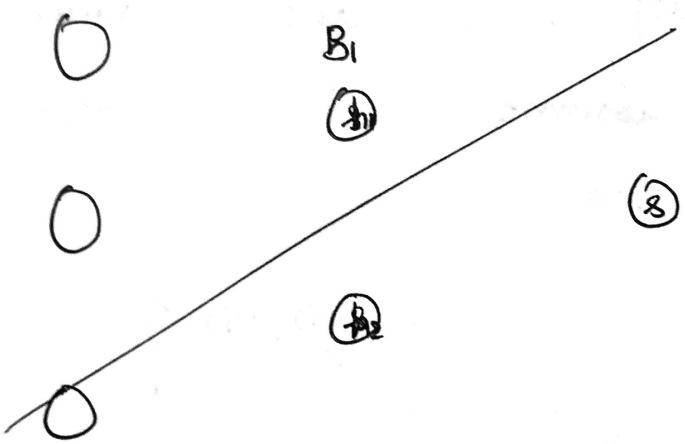
3. Hand calculation of gradients.

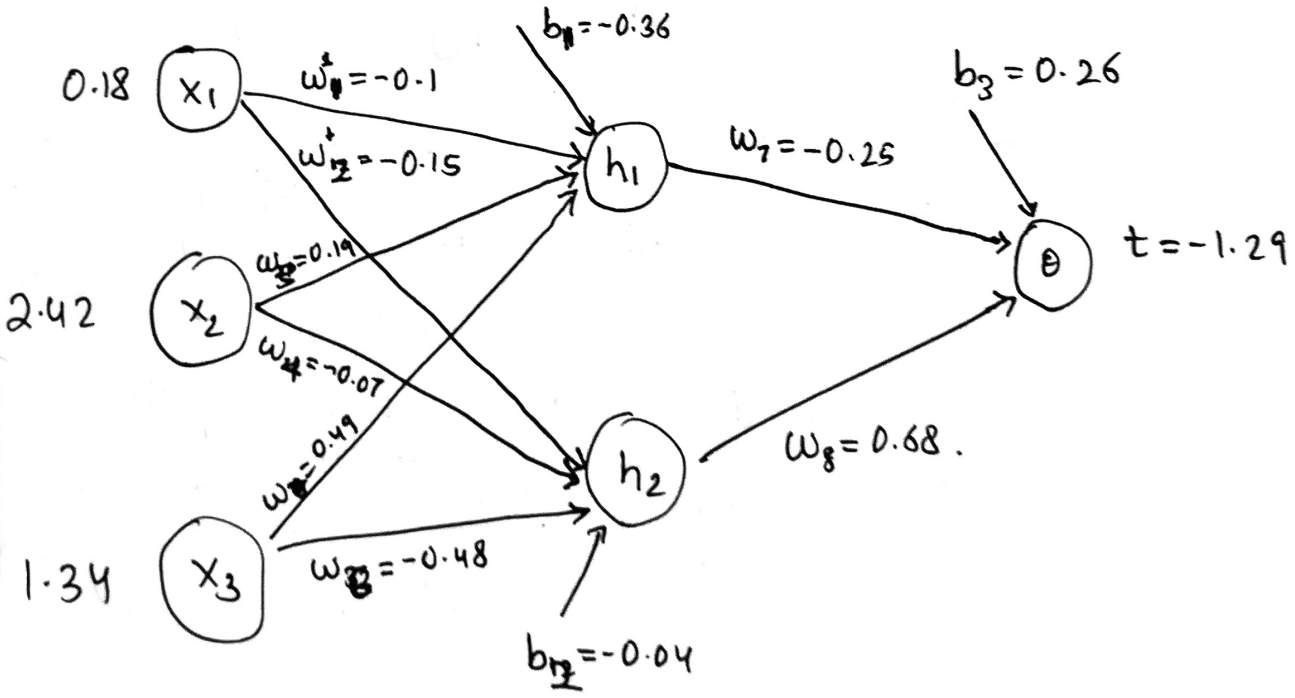
For my set:

inp1 = 0.18
inp 2 = 2.42
inp 3 = 1.34

$$W_1 = \begin{bmatrix} -0.1 & -0.15 \\ 0.19 & -0.07 \\ 0.49 & -0.48 \end{bmatrix} \quad B_1 = \begin{bmatrix} -0.36 \\ -0.04 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} -0.25 \\ 0.68 \end{bmatrix} \quad B_2 = 0.26. \quad \text{target} = -1.29.$$





$$b_1 + w_1 x_1 + w_3 x_2 + w_5 x_3 = z_{h_1}$$

$$b_2 + w_2 x_1 + w_4 x_2 + w_6 x_3 = z_{h_2}$$

$$h_1 = \sigma(z_{h_1})$$

$$h_2 = \sigma(z_{h_2})$$

$$b_3 + w_7 h_1 + w_8 h_2 = z_\theta$$

$$\theta = \sigma(z_0)$$

$$t = -1.29$$

Substituting in the above formulas, we get.

$$z_{h_1} = -0.36 + (-0.1)(0.18) + (0.19)(2.42) + (1.34)(0.49)$$

$$= \underline{0.7384}$$

$$z_{h_2} = -0.04 + (-0.15)(0.18) + (-0.07)(2.42) + (-0.48)(1.34)$$

$$= \underline{-0.8796}$$

$$h_1 = \sigma(0.7384) = \underline{0.6767} = \underline{0.68}$$

$$h_2 = \sigma(-0.8796) = \underline{0.2933} = \underline{0.29}.$$

$$z_\theta = 0.26 + (-0.25)(0.68) + (0.68)(0.29)$$

$$= \underline{0.2872}.$$

$$\theta = \sigma(z_\theta) = \underline{0.57}.$$

$$E = \frac{1}{2}(\theta - t)^2 = \underline{1.73}$$

$$\frac{\partial E}{\partial \theta} = \theta - t = \underline{1.86}.$$

$$\frac{\partial z_o}{\partial w_7} = h_1 \quad ; \quad \frac{\partial z_o}{\partial w_8} = h_2 \quad ; \quad \frac{\partial z_o}{\partial b_3} = 1$$

$$\frac{\partial E}{\partial w_7} = \frac{\partial E}{\partial o} \frac{\partial o}{\partial z_o} \frac{\partial z_o}{\partial w_7} = (o - t)(o(1-o))\, h_1$$

$$\frac{\partial E}{\partial w_7} = (1.86)(0.57(1-0.57))(0.68) = \boxed{0.31}$$

$$\frac{\partial E}{\partial w_8} = \frac{\partial E}{\partial o} \frac{\partial o}{\partial z_o} \frac{\partial z_o}{\partial w_8} = (o - t)\, o(1-o)\, h_2$$

$$\frac{\partial E}{\partial w_8} = (1.86)(0.57)(0.43)(0.29) = \boxed{0.13}$$

$$\frac{\partial E}{\partial b_3} = \frac{\partial E}{\partial o} \frac{\partial o}{\partial z_o} \frac{\partial z_o}{\partial b_3} = (1.86)(0.57)(0.43) = \boxed{0.45}$$

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial h_1} \frac{\partial h_1}{\partial z_{h_1}} \frac{\partial z_{h_1}}{\partial w_1}$$

$$\frac{\partial E}{\partial h_1} = \frac{\partial E}{\partial o} \frac{\partial o}{\partial z_o} \frac{\partial z_o}{\partial h_1} = (1.86)(0.57)(0.43)(-0.25) = \underline{-0.11}$$

$$\frac{\partial E}{\partial w_1} = (-0.11)(0.68)(0.32)(0.18) = \boxed{0.0043}$$

$$\frac{\partial E}{\partial w_3} = \frac{\partial E}{\partial h_1} \frac{\partial h_1}{\partial z_{h_1}} \frac{\partial z_{h_1}}{\partial w_3} = (-0.11)(0.68)(0.32)(2.42) = \boxed{-0.0579}$$

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial h_1} \frac{\partial h_1}{\partial z_{h_1}} \frac{\partial z_{h_1}}{\partial w_5} = (-0.11)(0.68)(0.32)(1.34) = \boxed{-0.0321}$$

$$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial h_2} \frac{\partial h_2}{\partial z_{h_2}} \frac{\partial z_{h_2}}{\partial w_2}$$

$$\frac{\partial E}{\partial h_2} = \frac{\partial E}{\partial o} \frac{\partial o}{\partial z_o} \frac{\partial z_o}{\partial h_2} = (1.86)(0.57)(0.43)(0.68) = 0.31$$

$$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial h_2}\frac{\partial h_2}{\partial z_{h_2}}\frac{\partial z_{h_2}}{\partial w_2} = (0.31)(0.29)(0.71)(0.18)$$
$$= \boxed{0.01148}$$

$$\frac{\partial E}{\partial w_4} = \frac{\partial E}{\partial h_2}\frac{\partial h_2}{\partial z_{h_2}}\frac{\partial z_{h_2}}{\partial w_4} = (0.31)(0.29)(0.71)(2.42)$$
$$= \boxed{0.1544}$$

$$\frac{\partial E}{\partial w_6} = \frac{\partial E}{\partial h_2}\frac{\partial h_2}{\partial z_{h_2}}\frac{\partial z_{h_2}}{\partial w_6} = (0.31)(0.29)(0.71)(1.34)$$
$$= \boxed{0.0855}.$$

$$\frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial h_1}\frac{\partial h_1}{\partial z_{h_1}}\frac{\partial z_{h_1}}{\partial b_1} = (-0.11)(0.68)(0.32)(1)$$
$$= \boxed{-0.0239}$$

$$\frac{\partial E}{\partial b_2} = \frac{\partial E}{\partial h_2}\frac{\partial h_2}{\partial z_{h_2}}\frac{\partial z_{h_2}}{\partial b_2} = (0.31)(0.29)(0.71)(1)$$
$$= \boxed{0.0638}.$$

$$\boxed{X = \begin{pmatrix} 0.18 \\ 2.42 \\ 1.31 \end{pmatrix}}$$

Thus summarizing all results:

$$\nabla W_1 = \begin{pmatrix} 0.0043 & 0.0115 \\ -0.0579 & 0.1544 \\ -0.0321 & 0.0855 \end{pmatrix}$$

$$\nabla B_1 = \begin{pmatrix} -0.0239 \\ 0.0638 \end{pmatrix}$$

$$\nabla B_2 = (0.46)$$

$$\nabla W_2 = \begin{pmatrix} 0.31 \\ 0.13 \end{pmatrix}$$

Error = 1.73

(mean square error)