Deep Learning Homework 1

1-) For every ML problem we should define

1-) Hypothesis = Linear hypothesis with non linear parameters.

2-) Loss = Hinge loss which can not be differentiable at some point

3-) Optimize the loss

Since differentiation of hingle loss has discontinuity therefore there is no closed form solution. We should use sub gradient.

Hypothesis $= h_\theta(x) = W^T . X$ where $X = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_6^{(i)} \\ 1 \end{bmatrix}$ Note: $x^{(i)}$ is the i'th instance of data points

loss $\ell(\theta) = \max(1 - y h_\theta(x), 0)$

Our optimization $\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \max(1 - y^{(i)} W^T x^{(i)}) \rightarrow J(w) = \frac{1}{n} \sum_{i=1}^{n} \max(1 - y^{(i)} w^T x^{(i)}, 0)$

$$\nabla_w \ell(y^{(i)} w^T x^{(i)}) = \ell'(y^{(i)} w^T x^{(i)}) \, y^{(i)} x^{(i)} = \left(\begin{cases} 0, & y^{(i)} w^T x^{(i)} > 1 \\ -1 & y^{(i)} w^T x^{(i)} < 1 \\ \text{undefined}, & y^{(i)} w^T x^{(i)} = 1 \end{cases} \right) y^{(i)} x^{(i)}$$

$$= \begin{cases} 0, & y^{(i)} w^T x^{(i)} > 1 \\ -y^{(i)} x^{(i)}, & y^{(i)} w^T x^{(i)} < 1 \\ \text{undefined}, & y^{(i)} w^T x^{(i)} = 1 \end{cases} = \nabla_w \ell(y^{(i)} w^T x^{(i)})$$

$$\nabla_w J(w) = \nabla_w \left(\frac{1}{n} \sum_{i=1}^{n} \ell(y^{(i)} w^T x^{(i)}) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla_w \ell(y_i w^T x_i)$$

$$= \begin{cases} \frac{1}{n} \left(\sum_{i = y^{(i)} w^T x < 1} -y^{(i)} x^{(i)} \right), & \text{all } y^{(i)} w^T x^{(i)} \neq 1 \\ \text{undefined}, & \text{otherwise} \end{cases}$$

Psuedo code:
```
w ← <0,0, ...., 0>   ← weights
b ← 0                ← bias
for iter=1 ... epochs do
    g ← <0,0 .... 0>  ← gradients
    g_b ← 0
    for all (x, y) do
        if y(wx + b) ≤ 1
            g ← g + yx
            g_b ← g + y
    end for
    W ← w + λg   step size
    b ← b + λg_b
```

2-) Output of the $z_4 = \sigma(z_4) = \frac{1}{1+e^{-z_4}}$

2.1-) $L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$ $\frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{y-1}{1-\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$

$\hat{y} = \sigma(z_4) = \frac{1}{1+e^{-z_4}}$ $\frac{\partial \hat{y}}{\partial z_4} = \sigma(z_4).(1-\sigma(z_4))$

So $\frac{\partial L}{\partial \sigma(z_4)} = \frac{\sigma(z_4)(1-\sigma(z_4)) - y}{\sigma(z_4)(1-\sigma(z_4))\,(1-\sigma(z_4)(1-\sigma(z_4)))}$

2.2-) $\frac{\partial L}{\partial W_{14}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_4} \cdot \frac{\partial z_4}{\partial W_{14}} = \delta z_4 \cdot \underbrace{\frac{\partial z_4}{\partial W_{14}}}_{a_1} = \delta z_4 \cdot a_1 = \frac{\partial L}{\partial W_{14}}$

Given $\delta z_4 = \frac{\partial L}{\partial z_4}$

2.3-) $\frac{\partial L}{\partial W_{11}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_4} \cdot \frac{\partial z_4}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_{11}} = \delta a_1 \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_{11}}$

Given $\delta a_1 = \frac{\partial L}{\partial a_1}$

$= \delta a_1 \cdot \begin{Bmatrix} 0 \text{ if } z_1<0 \\ 1 \text{ if } z_1>0 \end{Bmatrix} \cdot x_1$

$a_1 = \mathrm{ReLU}(z_1) = \max(0, z_1) = \begin{Bmatrix} 0 & \text{if } z_1<0 \\ z_1 & \text{if } z_1>0 \end{Bmatrix}$

$= \begin{Bmatrix} 0 & \text{if } z_1<0 \\ \delta a_1 . x_1 & \text{if } z_1>0 \end{Bmatrix}$

$\frac{\partial a_1}{\partial z_1} = \begin{Bmatrix} 0 \text{ if } z_1<0 \\ 1 \text{ if } z_1>0 \end{Bmatrix}$

$z_1 = W_{11} \cdot x_1 + b_1$

$\frac{\partial z_1}{\partial W_{11}} = x_1$

2.4-) L2 regularization is only applied on weights in general. Biases are omitted. We add L2 norm on cost function with λ regularization parameter.

First gradient: Our new cost $L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) + \frac{\lambda}{2}\sum_{w} w^2$

So for every partial derivate on weight becomes $\boxed{\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial w_{ij}} + \lambda . w}$

Since we don't use any w in $\frac{\partial L}{\partial \sigma(z_4)}$ there is no change.

Note: I assume same upstream gradients for the following gradients.

Second gradient: $\frac{\partial L}{\partial w_{14}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_4} \cdot \underbrace{\frac{\partial z_4}{\partial w_{14}}}_{a_1} = \delta z_4 \cdot (a_1 + \lambda . w_{14}) = \boxed{\delta z_4 \cdot a_1 + \delta z_4 \cdot \lambda . w_{14}}$

$\delta z_4 = \frac{\partial L}{\partial z_4}$

Third gradient: We have found that $\frac{\partial L}{\partial w_{11}} = \delta a_1 \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}}$

$$\frac{\partial a_1}{\partial z_1} = \begin{Bmatrix} 0 & \text{if } z_1 < 0 \\ 1 & \text{if } z_1 > 0 \end{Bmatrix}$$

$$= \delta a_1 \cdot \begin{Bmatrix} 0 \text{ if } z_1 < 0 \\ 1 \text{ if } z_1 > 0 \end{Bmatrix} \cdot (x_1 + \lambda w_{11})$$

$z_1 = w_{11} \cdot x_1 + b$

$$\frac{\partial z_1}{\partial w_{11}} = x_1 + \lambda . w_{11}$$

$$= \begin{Bmatrix} 0 & \text{if } z_1 < 0 \\ \delta a_1 (x_1 + \lambda w_{11}) & \text{if } z_1 > 0 \end{Bmatrix}$$

So in general regularization term regularize the derivates with w_{ij} weights. So it changes when w is involved in the gradient.