Report for Deep Learning (CS541)_Homework3:

Report By:
Swapneel Dhananjay Wagholikar (WPI ID: 257598983)
Anuj Pradeep Pai Raikar (WPI ID: 181758784)

Solution to Question 4:

We have executed grid search on following values of hyperparameters:

Number of epochs = [1, 2,3,4]
Learning Rates = [0.000003,0.000004, 0.000005, 0.000006]
Alphas = [2,3,4,5]
Mini batch sizes = [600, 400, 200, 100]

We found the best sets of hyperparameters as:

[epochs, learning rate, alpha, mini batch size] = [4, 0.0006, 5, 100]

The unregularized fCE on testing dataset is 0.7901629983444174

The accuracy in percentage is 78.28

```
Minimun value of Fce on validation dataset 0.7631654519880522
Fce on Test dataset is:  0.7901629983444174
Best Hyperparameters: epochs (m) 4  Learning rate 6e-06 alpha 5  mini_batch_size 100
Accuracy in percentage 78.28
```

1. Newton's method

2 layer NN: $\hat{y} = f_w(x) = w^T x$ $\quad J(w) = \frac{1}{2n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$

$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y}^{(i)} - w^T x)^2$$

$$H[f](w) = \frac{1}{n} XX^T$$

Newton's method:

$$w^{(k+1)} = w^{(k)} - H^{-1} \nabla_w f(w^{(k)})$$

According to Taylor expansion

$$J(w) \simeq J(w^{(k)}) + \nabla J(w^{(k)})(w - w^{(k)}) + \frac{1}{2}(w - w^{(k)}) H (w - w^{(k)})$$

$$\nabla_w J(w) \simeq \nabla_w J(w^{(k)}) + \frac{1}{2} \nabla_w (w^T H w - H w^{(k)})$$

$$w^{(k+1)} = w^{(k)} - H^{-1} \nabla_w J(w^{(k)})$$

$$H^{-1} = \left(\frac{1}{n} (XX^T)\right)^{-1} \Rightarrow \boxed{H^{-1} = n (XX^T)^{-1}}$$

$$w^{(k)} - w^{(k+1)} = H^{-1} \nabla_w J(w^{(k)})$$

$$w^{(k)} - w^{(k+1)} = H^{-1} \nabla_w J(w^{(k)})$$

In 1 iteration;

$$w^{(1)} = w^{(0)} + H^{-1} \nabla_w J(w^{(0)})$$

$$\nabla_w J(w^{(0)}) = \frac{1}{n} X (X^T w^{(0)} - Y)$$

$$w^{(1)} = w^{(0)} - n (XX^T)^{-1} \left(\frac{1}{n} X (X^T w^{(0)} - Y)\right)$$

$$= w^{(0)} - n (XX^T)^{-1} \left(\frac{1}{n}\right) (XX^T w^{(0)} - XY)$$

$$= w^{(0)} - (XX^T)^{-1} (XX^T)(w^{(0)}) + (XX^T)^{-1} XY$$

$$\boxed{w^{(i)} = (XX^T)^{-1} XY}$$ Hence proved

2. Softmax regression gradient updates.

$W = [\, w^{(1)} \text{ ------- } w^{(c)}]$ is an $m \times c$ matrix

c = no. of classes

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_{k'}}} \quad \text{and} \quad z_k = x^T w^{(k)} + b_k$$

for each $k = 1, \ldots, c$

$$f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}$$

$$\nabla_{w^{(l)}} f_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \nabla_{w^{(l)}} \log \hat{y}_k^{(i)}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right)$$

a. For $l = k$ $\quad \nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}^{(i)} (1 - \hat{y}_l^{(i)})$

b. For $l \neq k$ $\quad \nabla_{w^{(l)}} \hat{y}_k^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}^{(i)}$

a. $$\nabla_{w^{(l)}} \hat{y}_k^{(i)} = \nabla_{w^{(l)}} \left(\frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_{k'}}} \right) = \frac{\nabla_{w^{(l)}} \left[\left(\sum_{k'=1}^{c} e^{z_{k'}} \right) \left(\nabla_{w^{(l)}} e^{z_k} \right) - e^{z_k} \left(e^{z_k} \right) \right]}{\left(\sum_{k=1}^{c} e^{z_{k'}} \right)^2}$$

$$\nabla_w z_k = \nabla_w (x^T w^{(k)} + b_k)$$

$$= x$$

$$\therefore \nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \left[e^{z_k} \left(\sum_{k'=1}^{c} e^{z_{k'}} \right) - (e^{z_k})^2 \right] \Big/ \left(\sum_{k'=1}^{c} e^{z_{k'}} \right)^2$$

$$= x^{(i)} \left(e^{z_k} \Big/ \sum_{k'=1}^{c} e^{z_{k'}} \right) \left(1 - \frac{e^{z_k}}{\sum_{k'=1}^{c} z_{k'}} \right)$$

$$\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \left(\hat{y}_k^{(i)} \right) \left(1 - \hat{y}_k^{(i)} \right) \Rightarrow$$

For $L = k$

$$\boxed{\nabla_{w^{(L)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}_L^{(i)} \left(1 - \hat{y}_L^{(i)}\right)}$$

$$\hat{y} = \frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_k}}$$ Applying log

$$\log \hat{y} = \log\left(\frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_k}}\right) = \log\left(e^{z_k}\right) - \log\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)$$

$$\nabla_w \log \hat{y} = \nabla_w\left(\log e^{z_k} - \log\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)\right)$$

$$\frac{\nabla_w \hat{y}}{\hat{y}} = \frac{\nabla_w z_k \left(e^{z_k}\right)}{e^{z_k}} - \frac{\nabla_w \sum e^{z_k}}{\sum e^{z_k}} = \nabla_w z_k - \frac{\nabla_w \sum e^{z_k}}{\sum e^{z_k}}$$

If $k \neq L$

$$\nabla_{w^{(L)}} \frac{\hat{y}}{\hat{y}} = \nabla_w z_k - \frac{\nabla_w z_L e^{z_L}}{\sum e^{z_k}} = 0 - \frac{\nabla_w z_L e^{z_L}}{\sum e^{z_k}}$$

but, $\frac{e^{z_L}}{\sum e^{z_k}} = \hat{y}_L$ ↓ So, $\frac{\nabla_w \hat{y}}{\hat{y}} = -\nabla_w z_L \hat{y}_L$

If $k \neq L$ $\quad \boxed{\nabla_w \hat{y}^{(i)} = -x^{(i)} \hat{y}_k^{(i)} \hat{y}_L^{(k)}}$

c) $\nabla_{w^{(L)}} f_{CE}(W, b) = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} \hat{y}_k^{(i)} \nabla_{w^{(L)}} \log \hat{y}_k^{(i)}$

$$= \frac{-1}{n} \sum_{k=1}^{n} x^{(i)} \left(y^{(i)} - \hat{y}^{(i)}\right)$$

$\nabla_{w^{(L)}} f_{CE}(W, b) = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \nabla_{w^{(L)}} \log \hat{y}_k^{(i)}$

$$= \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_{w^{(L)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right)$$

For all cases; i.e., $k=l$ and $k \neq l$

$$\nabla_w f_{CE} = -\frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=l}^{c} \frac{y_k^{(i)}}{\hat{y}_k} \hat{y}_l^{(i)} (1-\hat{y}_l^{(i)}) + \sum_{k \neq l} \frac{y_k^{(i)}}{\hat{y}_k} \frac{y_l^{(i)}}{\hat{y}_l} [-x^{(i)} \hat{y}_k \hat{y}_l] \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=l}^{c} y_l x^{(i)} (1-\hat{y}_l) - \sum_{k \neq l} y_k^{(i)} x^{(i)} y_l^{(i)} \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[y_l x^{(i)} - \sum_{k=l} y_l x^{(i)} y_l - \sum_{k \neq l} y_k x^{(i)} y_k y_l \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[y_l x^{(i)} - \sum_{k=1}^{c} x^{(i)} y_k^{(i)} y_l \right]$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} x^{(i)} - x^{(i)} y_l^{(i)} \right]$$

$$\boxed{\nabla_w f_{CE} = -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left[y_l^{(i)} - \hat{y}_l^{(i)} \right]}$$

Now; $d \cdot \nabla_b f_{CE}(w, b) = -\frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})$

$$\nabla_b \hat{y}_k^{(i)} = \nabla_b \frac{e^{z_k}}{\sum_{j=1}^{c} e^{z_k}}$$

$$= \frac{e^{z_k} \sum e^{z_k} - e^{z_k} e^{z_k}}{(\sum e^{z_k})^2} = \frac{e^{z_k}}{\sum e^{z_k}} - \left(\frac{e^{z_k}}{\sum e^{z_k}} \right) \left(\frac{e^{z_k}}{\sum e^{z_k}} \right)$$

$$= \hat{y}_l^{(i)} - \hat{y}^{(i)} \hat{y}_l^{(i)}$$

When $k = l$ $\quad \nabla_b \hat{y}^{(i)} = \hat{y}_l^{(i)} (1 - y_k^{(i)})$

When $k \neq l$

$$\nabla_b \hat{y}^{(i)} = \nabla_b \left[\frac{e^{z_k}}{(\sum e^{z_k})} \right] = \frac{\sum e^{z_k} \cdot (0) - e^{z_k} \cdot e^{z_k}}{(\sum e^{z_k})^2}$$

$$= \frac{-e^{z_k}}{\sum e^{z_k}} \cdot \frac{e^{z_k}}{\sum e^{z_k}}$$

$$\therefore \nabla_b \hat{y}^{(i)} = -\hat{y}_k^{(i)} \cdot y_l^{(i)}$$

and.

$$\nabla_b J_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \nabla_b \log \hat{y}_c^{(i)}$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[\sum_{k \neq l} \frac{y_k \hat{y}_l}{\hat{y}_k} (1 - y_l^{(i)}) + \sum_{k \neq l} \frac{y_k}{y_k} (-\hat{y}_k \hat{y}_l) \right]$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l - \sum y_l \hat{y}_l + \sum_{k \neq l} y_k \hat{y}_l \right]$$

$$= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l - \hat{y}_l \sum_{k=1}^{c} \hat{y}_k^{(i)} \right]$$

$$\boxed{\nabla_b J_{CE}(W, b) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} - \hat{y}_l^{(i)} \right]}$$

3. Softmax c classes $k = 1 \ldots c$

$$\hat{y}_k = P(y_k = 1 \mid x, W, b) \quad \forall k \in \{1, \ldots, c\}$$

Ground truth label $y = [y_1, \ldots, y_c]^T$ y is a one-hot vector

likelihood of training example,

$$P(y \mid x, W, b) = P(y_1 = 1 \mid x, W, b)^{y_1} \times \ldots \times P(y_c = 1 \mid x, W, b)^{y_c}$$

$$P(y \mid x, W, b) = \prod_{k=1}^{c} \hat{y}_k^{y_k}$$

For the dataset,

$$P(D|w, b) = \prod_{i=1}^{n} \prod_{k=1}^{c} \hat{y}_k^{(i)\, y_k^{(i)}}$$

Applying log and multiplying by -1

$$-\log P(D|w, b) = -\log \prod_{i=1}^{n} \prod_{k=1}^{c} \hat{y}_k^{(i)\, y_k^{(i)}}$$

$$= -\prod_{i=1}^{n} \prod_{k=1}^{c} \log \left(\hat{y}_k^{(i)\, y_k^{(i)}} \right)$$

$$= -\sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)}$$

Hence proved.