Contents

Problem 1: Newton's Method

1. Newton's Method →

gradient → $\hat{y} = f_w(x) = W^T x$

$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2 \Rightarrow f_{MSE}$$

$$\nabla_w f_{MSE}(y, \hat{y}; w) = \nabla_w \left[\frac{1}{2n} \sum_{i=1}^{n} \left(x^{(i)T} w - y^{(i)} \right)^2 \right]$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \nabla_w \left[\left(x^{(i)T} w - y^{(i)} \right)^2 \right]$$

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

$$= \underline{\frac{1}{n} X \left(X^T w - y \right)} \quad \text{— (1)}$$

hessian → $\nabla_w \left(\frac{1}{n} X (X^T w - y) \right)$

$$= \boxed{\frac{1}{n} X (X^T)} \quad \text{— (2)}$$

Newton's Method →

– let the 2nd-order Taylor expansion of f_{MSE} around $w^{(k)}$ be:

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

→ To minimize f, we will find the root of the gradient of f's Taylor expansion:

$$\nabla_w f(w) \approx \nabla_w f(w^{(k)}) + \frac{1}{2}\nabla_w \left(w^T H w - w^T H w^{(k)} - w^{(k)T} H w + w^{(k)T} H w^{(k)}\right)$$

$$= \nabla_w f(w^{(k)}) + Hw - \frac{1}{2}Hw^{(k)} - \frac{1}{2}Hw^{(k)}$$

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

equating it to zero →

$$0 = \nabla_w f(w^{(k)}) + Hw - Hw^{(k)}$$

$$Hw = Hw^{(k)} - \nabla_w f(w^{(k)})$$

$$w^{(k+1)} = w^{(k)} - H^{-1}\nabla_w f(w^{(k)}) \quad \text{—— (3)}$$

→ Now, let us substitute eqn (1) & (2) in eqn (3)

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

$$= w^{(k)} - \left(\frac{1}{n} X(X^T)\right)^{-1}\left(\frac{1}{n} X\left(X^T w^{(k)} - y\right)\right)$$

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

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

$$\boxed{w^{(k+1)} = (XX^T)^{-1} XY}$$

→ as we can observe that $w^{(k+1)}$ is independent of $w^{(k)}$.

we can conclude that using Newton's method on MSE loss we can reach the optimum solution in just 1 iteration.

$$\boxed{w^* = (XX^T)^{-1} XY}$$

Problem 2: Derivation of SoftMax Regression Gradient Updates

Q.2

Given :- $\hat{y}_k = \dfrac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_{k'}}}$

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

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

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

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

Now we have two cases for above expression

① when $l = k$

Focussing on $\left(\nabla_w^{(l)} \hat{y}_k^{(i)}\right)$ term

$$\nabla_{w^l} \hat{y}_k^{(i)} = \nabla_{w^l} \hat{y}_l^{(i)} = \nabla_{w^l} \left(\frac{e^{z_l}}{\sum_{k'=1} e^{z_{k'}}} \right)$$

Using Quotient Rule $\Rightarrow$ $\nabla\left(\dfrac{u}{v}\right) = \dfrac{u'v - v'u}{v^2}$

$$\therefore \nabla_{w^l} \hat{y}_k^{(i)} = \frac{e^{z_l} \times \nabla_{w^l}(x^T w^{(l)} + b_l) \times \sum_{k'=1}^{c} e^{z_{k'}} - e^{z_l} \times e^{z_l} \times \nabla_{w^l}(x^T w^l + b_l)}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2}$$

$$\Rightarrow \frac{x \times e^{z_l} \times \sum_{k'=1}^{c} e^{z_{k'}} - x(e^{z_l})^2 \times 1}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2} \Rightarrow x \left(\frac{e^{z_l}}{\sum_{k'=1}^{c} e^{z_{k'}}} - \frac{e^{z_l} \times e^{z_l}}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2} \right)$$

$$\therefore \nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \times \left(\hat{y}_l^{(i)} - \hat{y}_l^{(i)2} \right)$$

$$\boxed{\therefore \hat{y}_l = \frac{e^{z_l}}{\sum_{k'=1}^{c} e^{z_{k'}}}}$$

Hence $\boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = x^{(i)} \hat{y}_l^{(i)} \left(1 - \hat{y}_l^{(i)}\right)}$ when $l = k$ //

Now when $l \neq k$

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

$$= e^{z_k} \nabla_{w^{(l)}} \left(\frac{1}{\sum_{k'=1}^{c} e^{z_{k'}}} \right)$$

$$= e^{z_k} \times \frac{-1}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2} \times \nabla_{w^{(l)}} \left(x^T w^{(l)} + b_l \right) \times e^{z_l}$$

$$= \frac{- e^{z_k} \times e^{z_l} \times x}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2}$$

Hence $\boxed{\nabla_{w^{(l)}} \hat{y}_k^{(i)} = -x^{(i)} \times \hat{y}_k^{(i)} \times \hat{y}_l^{(i)}}$ when $l \neq k$ // $\quad \boxed{\therefore \hat{y}_l = \frac{e^{z_l}}{\sum_{k'=1}^{c} e^{z_{k'}}}, \ \hat{y}_k = \frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_{k'}}}}$

Now lets compute the total gradient of f_{CE} each wrt $w^{(k)}$

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

It will be the sum over $l = 1, \ldots, c$ terms

so when $k = l$ & $k \neq l$ we know the term

$\therefore$ We can split $\Sigma_k a_k = a_l + \Sigma_{k \neq l} a_k$

$$\therefore \nabla_{\omega^{(l)}} f_{CE}(\omega, b) = \frac{-1}{n} \sum_{i=1}^{n} \left(y_l^{(i)} \times \left(\frac{x^{(i)} \times \cancel{\hat{y}_l^{(i)}} (1 - \hat{y}_l^{(i)})}{\cancel{\hat{y}_l^{(i)}}} \right) + \sum_{k \neq l}^{c} y_k^{(i)} \times \left(\frac{-x^{(i)} \times \cancel{\hat{y}_k^{(i)}} \, \hat{y}_l^{(i)}}{\cancel{\hat{y}_k^{(i)}}} \right) \right)$$

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

$$\boxed{\because \sum_{k=1}^{c} y_k^{(i)} = 1 \Rightarrow \therefore \sum_{k \neq l}^{c} y_k^{(i)} = (1 - y_l^{(i)})}$$

$$\therefore \Rightarrow \frac{-1}{n} \sum_{i=1}^{n} \left(x^{(i)} \times y_l^{(i)} \times (1 - \hat{y}_l^{(i)}) - x^{(i)} (1 - y_l^{(i)}) \hat{y}_l^{(i)} \right)$$

$$\Rightarrow -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y_l^{(i)} - \cancel{y_l^{(i)} \hat{y}_l^{(i)}} - \hat{y}_l^{(i)} + \cancel{y_l^{(i)} \hat{y}_l^{(i)}} \right)$$

$$\Rightarrow \boxed{-\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y_l^{(i)} - \hat{y}_l^{(i)} \right) = \nabla_{\omega^{(l)}} f_{CE}(W, b)}$$

//

Similarly we can prove for $\nabla_b^{(l)} f_{CE}(\omega, b)$

$$\Rightarrow \nabla_{b^{(l)}} f_{CE}(\omega, b) = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \nabla_{b^{(l)}} \log \hat{y}_k^{(i)} = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right)$$

It will be the sum over $l = 1, \ldots, c$ terms

so when $k = l$ & $k \neq l$ we know the term

$\therefore$ we can split $\Sigma_k a_k = a_l + \sum_{k \neq l} a_k$

$$\therefore \nabla_{\omega^{(l)}} L_{CE}(\omega, b) = \frac{-1}{n} \sum_{i=1}^{n} \left(y_l^{(i)} \times \left(\frac{x^{(i)} \times \cancel{\hat{y}_l^{(i)}} (1 - \hat{y}_l^{(i)})}{\cancel{\hat{y}_l^{(i)}}} \right) + \sum_{k \neq l}^{c} y_k^{(i)} \times \left(\frac{-x^{(i)} \times \cancel{\hat{y}_k^{(i)}} \, \hat{y}_l^{(i)}}{\cancel{\hat{y}_k^{(i)}}} \right) \right)$$

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

$$\boxed{\because \sum_{k=1}^{c} y_k^{(i)} = 1 \Rightarrow \therefore \sum_{k \neq l}^{c} y_k^{(i)} = (1 - y_l^{(i)})}$$

$$\therefore \Rightarrow \frac{-1}{n} \sum_{i=1}^{n} \left(x^{(i)} \times y_l^{(i)} \times (1 - \hat{y}_l^{(i)}) - x^{(i)} (1 - y_l^{(i)}) \hat{y}_l^{(i)} \right)$$

$$\Rightarrow -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y_l^{(i)} - \cancel{y_l^{(i)} \hat{y}_l^{(i)}} - \hat{y}_l^{(i)} + \cancel{y_l^{(i)} \hat{y}_l^{(i)}} \right)$$

$$\Rightarrow \boxed{-\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y_l^{(i)} - \hat{y}_l^{(i)} \right) = \nabla_{\omega^{(l)}} L_{CE}(W, b)} \; //$$

Similarly we can prove for $\nabla_b^{(l)} L_{CE}(\omega, b)$

$$\Rightarrow \nabla_{b^{(l)}} L_{CE}(\omega, b) = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \nabla_{b^{(l)}} \log \hat{y}_k^{(i)} = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}} \right)$$

We have two cases: ① when $l = k$

then $\nabla_{b^{(l)}} \hat{y}_k^{(i)} = \nabla_{b^{(l)}} \left(\frac{e^{z_l}}{\sum_{k'=1}^{c} e^{z_{k'}}} \right)$

$\Rightarrow$ Using quotient rule

$$\Rightarrow \frac{e^{z_l} \times 1 \times \sum_{k'=1}^{c} e^{z_{k'}} - e^{z_l} \times e^{z_l}}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2}$$

$$\Rightarrow \frac{e^{z_l} \times \cancel{\sum_{k'=1}^{c} e^{z_{k'}}}}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^{\cancel{2}}} - \frac{e^{z_l} \times e^{z_l}}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2}$$

$$\boxed{\nabla_{b^{(l)}} \hat{y}_k^{(i)} \Rightarrow \hat{y}_l - \hat{y}_l^{2}}$$

② when $l \neq k$

$$\nabla_{b^{(l)}} \hat{y}_k^{(i)} = e^{z_k} \nabla_{b^{(l)}} \left(\frac{1}{\sum_{k'=1}^{c} e^{z_{k'}}} \right)$$

$$= e^{z_k} \times \frac{-1}{\left(\sum_{k'=1}^{c} e^{z_{k'}}\right)^2} \times e^{z_l}$$

$$\boxed{\nabla_{b^{(l)}} \hat{y}_k^{(i)} = -\hat{y}_k \hat{y}_l}$$

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

$\therefore$ We can split $\boxed{\sum_k a_k = a_l + \sum_{k \neq l} a_k}$

$$\nabla_b^{(l)} f_{CE}(\omega,b) = \frac{-1}{n}\sum_{i=1}^{n}\left(y_l^{(i)} \times \frac{(\hat{y}_l - \hat{y}_l^2)}{\hat{y}_l} + \sum_{k\neq l} y_k^{(i)} \times \frac{-\hat{y}_k \times \hat{y}_l}{\hat{y}_k}\right)$$

$$= \frac{-1}{n}\sum_{i=1}^{n}\left(y_l^{(i)} \times (1-\hat{y}_l^{(i)}) + \left(-(1-y_l^{(i)}) \times \hat{y}_l\right)\right)$$

$$\boxed{\because \sum_{k=1}^{c} y_k = 1 \Rightarrow \sum_{k\neq l}^{c} y_k = (1-y_l^{(i)})}$$

$$\therefore \Rightarrow \frac{-1}{n}\sum_{i=1}^{n}\left(y_l^{(i)} - \cancel{y_l^{(i)}\hat{y}_l^{(i)}} - \hat{y}_l + \cancel{y_l^{(i)}\hat{y}_l^{(i)}}\right)$$

$$\therefore \nabla_b^{(l)} f_{CE}(\omega,b) = \frac{-1}{n}\sum_{i=1}^{n}\left(y_l^{(i)} - \hat{y}_l^{(i)}\right)$$

⇒ The gradient of cross-entropy loss wrt bias of class 'l' is as shown above.

⇒ Now collecting all the gradient wrt bias of classes $l = 1, \dots, c$ we can represent this in vector form as :-

$$\nabla_b^{1} f_{CE}(\omega,b) = \frac{-1}{n}\sum_{i=1}^{n}\left(y_1^{(i)} - \hat{y}_1^{(i)}\right)$$

$$\vdots$$

$$\nabla_b^{c} f_{CE}(\omega,b) = \frac{-1}{n}\sum_{i=1}^{n}\left(y_c^{(i)} - \hat{y}_c^{(i)}\right)$$

$$\Rightarrow \boxed{\nabla_b f_{CE}(\omega,b) = \frac{-1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)}$$

Problem 3: Logistic Sigmoid Identity

(3) Derivation of Cross-Entropy as Negative log-likelihood-

$D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

Now, $P(D|W,b) = P(\{(x^{(i)}, y^{(i)})\}_{i=1}^{n} | W, b)$

$= P(y^1|x^1, W, b) \cdot \prod_{i=2}^{n} P(y^i|x^i, W, b, y_2, y_3 \ldots y_n)$

— assuming Conditional Independence

$P(D|W,b) = \prod_{i=1}^{n} P(y_i|W,b)$

$P(D|W,b) = \prod_{i=1}^{n} P(y_i|W,b)$

→ from the given eqn in the question the above eqn can be re-written as —

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

taking negative log on both sides.

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

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

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

$$\boxed{\begin{gathered} = -\sum_{i=1}^{n} \sum_{k=1}^{c} y_k^{(i)} \log \hat{y}_k^{(i)} \\ -\log(P(D|W,b)) = f_{CE}(D; W, b) \end{gathered}}$$

Problem 4: Implementation of SoftMax Regression

- We developed a 2-layer SoftMax neural network using the Fashion MNIST dataset. The file is saved as ***homework3_590146168_249812226.py.***
- The model is trained via stochastic gradient descent method to minimize cross-entropy loss.
- There are 4 Hyperparameters that can be altered:
    - Mini-Batch Size
    - Learning Rate
    - Number of Epochs
    - L_2 Regularization Strength
- The training data is split into Training (80%) and Validation (20%) dataset.
- We have selected a set of 8 values for each hyperparameter i.e., 8^4 =4096 combinations.
- The weights and bias are initialized to zero.
- Once training is done, we calculate the cost on validation set. If the cost is minimum, the best hyperparameter is updated. This process is done till all hyperparameters are checked.
- It took approximately 30 hours to check all 8^4 combinations.
- At the end we obtain the best set of hyperparameters. Using these hyperparameters we train the model on validation + training dataset.
- Now this trained model is used to calculate the cost (unregularized MSE) and percent correctly classified examples for the test dataset and the and the result is reported below:

```
(base) saammmy@saammmy-xps15:~/projects/3$ /usr/bin/python3 /home/saammmy/projects/3/neural_network.py
----------------------------------------------------------------------
Performing Grid Search for 4096 combinations of Hyperparameters:
----------------------------------------------------------------------
 Grid Search Completed
 Results after performing Grid Search:
 Best Hyperparameters:
   Epochs=  35
   Alpha=  1
   Learning Rate=  5e-07
   Mini Batch Size=  32
 Cost on Validation Set with Best Hyperparameters=  0.433649244826

----------------------------------------------------------------------
Training on Training + Validation Dataset:
----------------------------------------------------------------------
 Training Completed

----------------------------------------------------------------------
Performance Evaluation
----------------------------------------------------------------------
 Cost on Test Dataset=  0.455827235063
 Accuracy on Test Dataset= 84.26 %
```

End Of Assignment