Monday, February 6, 2023 3:07 PM

Q2

a) For the bias term :-
we can find the gradient of the log loss as $y=1$ as all the training examples are +ve.

$\therefore$ log loss $= -\left[y \log \hat{y} - (1-y) \log(1-\hat{y})\right]$

$\therefore \nabla_b f_{loss} = \nabla_b \left[-(1)\log \hat{y} - 0 \cdot \log(1-\hat{y})\right]$

$\therefore \nabla_b f_{loss} = \nabla_b \left[-\log \hat{y}\right]$

$= -\nabla_b \log \sigma(x^T w + b)$

$= -\dfrac{\sigma(x^T w+b)(1-\sigma(x^T w+b))}{\sigma(x^T w+b)}$

$\therefore \nabla_b f_{loss} = \left((\sigma(x^T w+b) - 1\right)$

1) Now, if we assume $w = 0$

$\nabla_b f_{loss} = \sigma(b) - 1$

The bias update is

$b_{new} = b_{old} - [\text{learning rate} \times (\sigma(b)-1)]$

$b_{new} = b_{old} - \alpha(\sigma(b)-1)$

As $0 < \sigma(b) < 1 \quad \therefore \quad \sigma(b)-1 < 0$; let $\alpha(b)-1 = a$

$b_{new} = b_{old} + \alpha \times a$

If we assume b convergence as 10000 using the above equation we won't be able to provide an upper bound for iterations.

2) If $w \neq 0$

$\nabla_b f_{loss} = \sigma(x^T w+b) - 1$

The bias update is:

$b_{new} = b_{old} - \left([\sigma(x^T w+b) - 1] \times \alpha\right)$

Now here we can see that the convergence of bias will depend on the values of the training example as well.

$\therefore$ The convergence of the bias will depend and cannot be guaranted.

2) For the weight vector :-

As 1) we can calculated the gradient of loss function

$\nabla_w f_{loss} = x(\hat{y}-y)$

$\nabla_w f_{loss} = x[\hat{y}-y] = x[\hat{y}-1]$

$\nabla_w f_{loss} = x[\sigma(x^T w+b) - 1]$

Assume $b=0$ for simplicity.

$\nabla_w f_{loss} = x[\sigma(x^T w) - 1]$

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ ; $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

$\nabla_w f_{loss} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\left[\sigma\left(\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}\right) - 1\right]$

$\nabla_w f_{loss} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}[\sigma(w_1 x_1 + w_2 x_2) - 1]$

Update for w

$w = w_0 - \alpha \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}[\sigma(x_1 w_1 + w_2 x_2) - 1]$

Now value of $\sigma(x_1 w_1 + w_2 x_2) - 1 < 0$

$w = w_0 - \alpha_{new} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\qquad \alpha_{new} = \alpha \times a$ ; $a = [\sigma(x_1 w_1 + x_2 w_2) - 1]$

This equation denotes that the convergence of w is dependent on the exact values of X.

If $x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ it will never converge.

If $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ or some suitable value it might converge.

$\therefore$ Convergence of w depends on the value of training examples.