CSED/GSAI 538 Homework Assignment 1. 20222421 JAEYOON SIM

#1

(a) The answer is number 1. $L = \| E(A) - E(P) \|^2 - \| E(A) - E(N) \|^2$

(b) The answer is number 2. $L = \| C(M) - C(G) \|^2 + \| S(M) - S(G) \|^2$

#2

(a) $\hat{y} = \sigma(wx+b)$ where σ is sigmoid function.

(b) Cross-Entropy Loss $= -\left(y \cdot \log(\hat{y}) + (1-y)\log(1-\hat{y})\right)$

(c) $L = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right)$

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{1}{m}\left(-\frac{y_i}{\hat{y}_i} + \frac{1-y_i}{1-\hat{y}_i}\right)$$

$$\frac{\partial \hat{y}}{\partial w} = \sigma(wx+b)\left(1-\sigma(wx+b)\right)\cdot x$$

$$\frac{\partial \hat{y}}{\partial b} = \sigma(wx+b)\left(1-\sigma(wx+b)\right)\cdot 1$$

$$dw = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial L}{\partial \hat{y}_i}\cdot\frac{\partial \hat{y}_i}{\partial w}$$

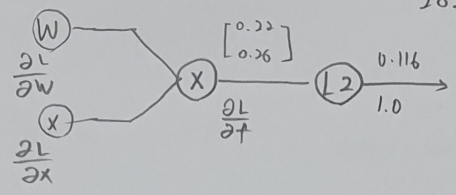
$$db = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial b}$$

$\therefore$ $W_{new} = W_{old} - \alpha \cdot dw$ where α is learning rate, dw and db is above.

$b_{new} = b_{old} - \alpha \cdot db$

20222421 JAEYOON SIM

#3



L2 loss = 0.1

$W = \begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix} \quad X = \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$

$WX = \begin{bmatrix} 0.1 & 0.5 \\ -0.3 & 0.8 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} = \begin{bmatrix} 0.02 + 0.20 \\ -0.06 + 0.32 \end{bmatrix} = \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$

$\frac{\partial L}{\partial W} = \sum_{i=1}^{2} 2 \cdot (WX)_i \cdot X_i^T = 2(WX)X^T$, $\frac{\partial L}{\partial X} = \sum_{i=1}^{2} 2 \cdot (WX)_i \cdot W = 2W^T \cdot (WX)$

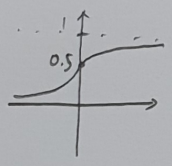
$\therefore dW = 2 \cdot \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix} [0.2 \; 0.4]$, $dX = 2 \cdot \begin{bmatrix} 0.1 & -0.3 \\ 0.5 & 0.8 \end{bmatrix} \begin{bmatrix} 0.22 \\ 0.26 \end{bmatrix}$

$= \begin{bmatrix} 0.088 & 0.176 \\ 0.104 & 0.208 \end{bmatrix} \quad = \begin{bmatrix} -1.112 \\ 0.636 \end{bmatrix}$

#4

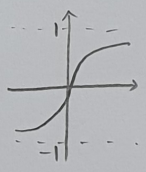
(a)

(i) Sigmoid : $a(z) = \frac{1}{1+e^{-z}}$



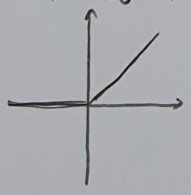
Sigmoid function exists between (0,1). It is specially used for models where we have to predict the probability. It is differentiable, but when z goes big or small, it has gradient vanishing.

(ii) tanh : $a(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$



Tanh function is also like sigmoid but better. It exists between (-1,1). The negative inputs will be mapped negative. It is differentiable, but it has gradient vanishing problem.

(iii) ReLU : $a(z) = \max(0, z)$



ReLU function is the most widely used in the world and it helps to solve the gradient vanishing problem. But, any negative inputs turns the value into zero and it has not proper mapping problem.

(iv) Leaky ReLU : $a(z) = \max(0.01z, z)$

Leaky ReLU function solves gradient vanishing problem and also has non-zero value for negative inputs. It helps to increase the range of the ReLU function.

(b)

(i) sigmoid : $a'(z) = a(z)(1 - a(z))$

(ii) tanh : $a'(z) = 1 - \{a(z)\}^2$

(iii) ReLU : $a'(z) = \begin{cases} 0 & \text{for } z < 0 \\ 1 & \text{for } z \geq 0 \end{cases}$

(iv) Leaky : $a'(z) = \begin{cases} 0.01 & \text{for } z < 0 \\ 1 & \text{for } z \geq 0 \end{cases}$

20222421 JAEYOON SIM

#5

(a) The purpose of regularization is to prevent overfitting. It means that we have to reduce high variance.

(b) The effect of regularization on weights W reduces the magnitude of weight (= |weight|) and prevents weights value from getting bigger. By doing this, the degree of decision boundary represented by weights decreases.

(c) The regularization penalizes to make complex decision boundary which is the result of focusing on minimizing the training loss only. The regularization term in the loss function, it reduces the magnitude of weights which in turns reduces fluctuation in the regression. So, the regularization prevents overfitting and high variance.