CS2109S Tutorial 7

AY 25/26 Sem 1 — github/omgeta

A.

1. $a = ReLU((W^{[1]})^\top \begin{bmatrix}1\\x\end{bmatrix}) = ReLU(\begin{bmatrix}0.8\\-0.7\end{bmatrix}) = \begin{bmatrix}0.8\\0\end{bmatrix}$

2. $\hat{y} = ReLU((W^{[2]})^\top \begin{bmatrix}1\\a\end{bmatrix}) = ReLU(\begin{bmatrix}0.5\\-0.38\end{bmatrix}) = \begin{bmatrix}0.5\\0\end{bmatrix}$

3. $L(\hat{y}, y) = \frac{1}{2}\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 = 0.485$

B.

1. $f(x) = |x - 1|$

2. No; identity will not perform any transformation, sigmoid is incompatible with the output range, and ReLU here with a single neuron only can achieve one half of the transformation

3. Hidden layer ReLU: $W^{[1]} = \begin{bmatrix}-1 & 1\\1 & -1\end{bmatrix}$

   Output layer Identity: $W^{[2]} = \begin{bmatrix}0\\1\\1\end{bmatrix}$

4. They are needed to represent non-linear relationships

C.

1. (a.) $\frac{\partial J_{BCE}(W)}{\partial \hat{y}} = -y\frac{\partial}{\partial \hat{y}}\log(\hat{y}) + (1-y)\frac{\partial}{\partial \hat{y}}(1-\hat{y}) = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$

   (b.) $\frac{\partial J_{BCE}(W)}{\partial f} = \frac{\partial J_{BCE}(W)}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial f} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)\cdot \hat{y}(1-\hat{y}) = \hat{y} - y$

   (c.) Since $\frac{\partial f}{\partial W_2} = x_2$, then $\frac{\partial J_{BCE}(W)}{\partial W_2} = \frac{\partial J_{BCE}(W)}{\partial f}x_2$

2. They are needed to deal with imbalanced data so not to be dominated by the majority class. They should be set such that $\alpha|A| \approx \beta|B|$ so $\alpha = 5.5, \beta = 0.55$

D.

1. (a.) Sigmoid on early layers squashes values mostly to 0, so the gradient per layer also becomes almost zero.

   (b.) Use ReLU instead for bigger gradients.

2. (a.) Gradients can be negative so the neurons can adjust back out.

   (b.) When $a = 1$, LeakyReLU degenerates into the identify function.