Q2. Nueral Networks

Menghe JIN

EPFL — January 7, 2021

1 Derivation of gradients

To distinguish between matrices, vectors and scalar, we denote matrices as capital letters, vectors as bold lowercase letters and scalars as normal lowercase ones. The parameters involved in this derivation are $\mathcal{L}$, $\boldsymbol{x}$, $W^{[1]}$, $\boldsymbol{b}^{[1]}$, $\boldsymbol{z}^{[1]}$, $\boldsymbol{a}^{[1]}$, $W^{[2]}$, $\boldsymbol{b}^{[2]}$, $\boldsymbol{z}^{[2]}$, $\boldsymbol{a}^{[2]}$, $W^{[3]}$, $b^{[3]}$, $z^{[3]}$, $a^{[3]}$. For sigmoid function, $\dfrac{d\delta(x)}{dx} = \delta(x)\left(1-\delta(x)\right)$.

$$\mathcal{L} = -\left[y \cdot \log\left(a^{[3]}\right) + (1-y)\log\left(1-a^{[3]}\right)\right]$$

$$\frac{d\mathcal{L}}{da^{[3]}} = -\frac{y}{a^{[3]}} + \frac{1-y}{1-a^{[3]}} = \frac{a^{[3]}-y}{a^{[3]}\left(1-a^{[3]}\right)}$$

$$\frac{d\mathcal{L}}{dz^{[3]}} = \frac{d\mathcal{L}}{da^{[3]}}\frac{da^{[3]}}{dz^{[3]}} = \frac{d\mathcal{L}}{da^{[3]}} \cdot a^{[3]}\left(1-a^{[3]}\right)$$

$$\nabla_{W^{[3]}}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial W_1^{[3]}}, ..., \frac{\partial\mathcal{L}}{\partial W_{h_2}^{[3]}}\right) = \frac{d\mathcal{L}}{dz^{[3]}}\left(\frac{\partial z^{[3]}}{\partial W_1^{[3]}}, ..., \frac{\partial z^{[3]}}{\partial W_{h_2}^{[3]}}\right) = \frac{d\mathcal{L}}{dz^{[3]}}\left(a_1^{[2]}, ..., a_{h_2}^{[2]}\right) = \frac{d\mathcal{L}}{dz^{[3]}} \cdot \left(\boldsymbol{a}^{[2]}\right)^T$$

$$\frac{\partial\mathcal{L}}{\partial b^{[3]}} = \frac{d\mathcal{L}}{dz^{[3]}}\frac{\partial z^{[3]}}{\partial b^{[3]}} = \frac{d\mathcal{L}}{dz^{[3]}} \cdot 1$$

$$\nabla_{\boldsymbol{a}^{[2]}}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial a_1^{[2]}}, ..., \frac{\partial\mathcal{L}}{\partial a_{h_2}^{[2]}}\right)^T = \frac{d\mathcal{L}}{dz^{[3]}}\left(\frac{\partial z^{[3]}}{\partial a_1^{[2]}}, ..., \frac{\partial z^{[3]}}{\partial a_{h_2}^{[2]}}\right)^T = \frac{d\mathcal{L}}{dz^{[3]}}\left(W_1^{[3]}, ..., W_{h_2}^{[3]}\right)^T = \frac{d\mathcal{L}}{dz^{[3]}} \cdot \left(W^{[3]}\right)^T$$

$$\nabla_{\boldsymbol{z}^{[2]}}\mathcal{L} = \nabla_{\boldsymbol{a}^{[2]}}\mathcal{L} \cdot \nabla_{\boldsymbol{z}^{[2]}}\boldsymbol{a}^{[2]} = \nabla_{\boldsymbol{a}^{[2]}}\mathcal{L} \cdot \boldsymbol{a}^{[2]}\left(1-\boldsymbol{a}^{[2]}\right), \quad \text{(element-wise multiplication)}$$

$$\nabla_{W^{[2]}}\mathcal{L} = \left[\frac{\partial\mathcal{L}}{\partial W_{ij}^{[2]}}\right] = \left[\frac{\partial\mathcal{L}}{\partial z_i^{[2]}}\frac{\partial z_i^{[2]}}{\partial W_{ij}^{[2]}}\right] = \left[\frac{\partial\mathcal{L}}{\partial z_i^{[2]}}a_j^{[1]}\right] = \nabla_{\boldsymbol{z}^{[2]}}\mathcal{L} \cdot \left(\boldsymbol{a}^{[1]}\right)^T$$

$$\nabla_{\boldsymbol{b}^{[2]}}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial b_1^{[2]}}, ..., \frac{\partial\mathcal{L}}{\partial b_{h_2}^{[2]}}\right)^T = \nabla_{\boldsymbol{z}^{[2]}}\mathcal{L} \cdot \left(\frac{\partial z_1^{[2]}}{\partial b_1^{[2]}}, ..., \frac{\partial z_{h_2}^{[2]}}{\partial b_{h_2}^{[2]}}\right)^T = \nabla_{\boldsymbol{z}^{[2]}}\mathcal{L}, \quad \text{(element-wise multiplication)}$$

$$\nabla_{\boldsymbol{a}^{[1]}}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial a_1^{[1]}}, ..., \frac{\partial\mathcal{L}}{\partial a_{h_2}^{[1]}}\right)^T = \left[\frac{\partial z_i^{[2]}}{\partial a_j^{[1]}}\right] \cdot \nabla_{\boldsymbol{z}^{[2]}}\mathcal{L} = \left(W^{[2]}\right)^T \cdot \nabla_{\boldsymbol{z}^{[2]}}\mathcal{L}$$

$$\nabla_{\boldsymbol{z}^{[1]}}\mathcal{L} = \nabla_{a^{[1]}}\mathcal{L} \cdot \nabla_{\boldsymbol{z}^{[1]}}\boldsymbol{a}^{[1]} = \nabla_{a^{[1]}}\mathcal{L} \cdot \boldsymbol{a}^{[1]}\left(1-\boldsymbol{a}^{[1]}\right), \quad \text{(element-wise multiplication)}$$

$$\nabla_{W^{[1]}}\mathcal{L} = \left[\frac{\partial\mathcal{L}}{\partial W_{ij}^{[1]}}\right] = \left[\frac{\partial\mathcal{L}}{\partial z_i^{[1]}}\frac{\partial z_i^{[1]}}{\partial W_{ij}^{[1]}}\right] = \left[\frac{\partial\mathcal{L}}{\partial z_i^{[1]}}x_j\right] = \nabla_{\boldsymbol{z}^{[1]}}\mathcal{L} \cdot \boldsymbol{x}^T$$

$$\nabla_{\boldsymbol{b}^{[1]}}\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial b_1^{[1]}}, ..., \frac{\partial\mathcal{L}}{\partial b_{h_1}^{[1]}}\right)^T = \nabla_{\boldsymbol{z}^{[1]}}\mathcal{L} \cdot \left(\frac{\partial z_1^{[1]}}{\partial b_1^{[1]}}, ..., \frac{\partial z_{h_1}^{[1]}}{\partial b_{h_1}^{[1]}}\right)^T = \nabla_{\boldsymbol{z}^{[1]}}\mathcal{L}, \quad \text{(element-wise multiplication)}$$

2 Implementation and results

The settings are: learning rate is 1e-3 and for each batch this learning rate multiplies with 0.95; the batch size is 16; for each batch, the program iterates 150 times. For the training part, we average the loss over all samples in the batch and we compute the accuracy on the batch as well. Figure 1 and 2 are two examples of loss and accuracy during training. The final accuracy on the test data is 0.7097.

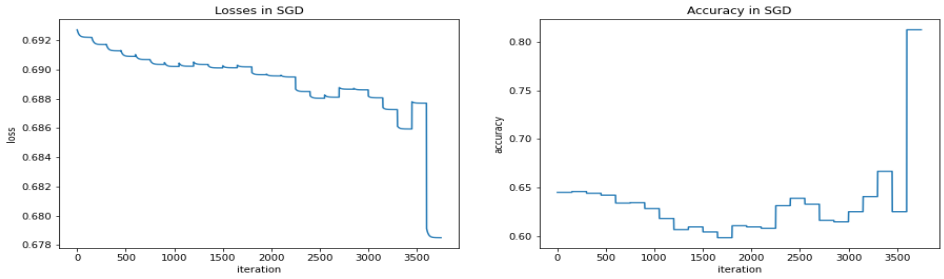

Figure 1: Loss and accuracy during training (1).

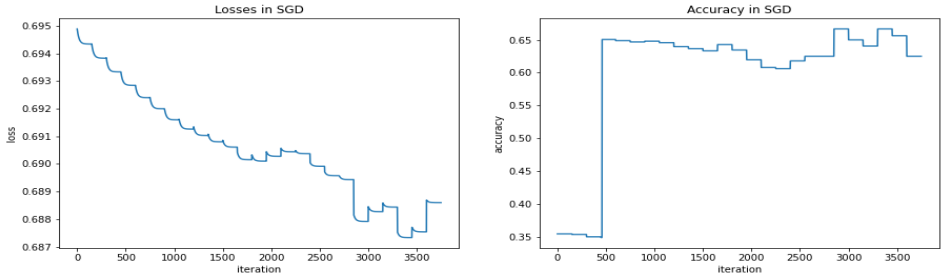

Figure 2: Loss and accuracy during training (2).