Deep Learning

Martin de la Riva - 11403799

November 2017

Exercise 1

$$\frac{\partial \mathcal{L}}{\partial W_1} = (y_{out} - y_{gt})f_3'(s_{out})W_{out})f_2'(s_2)W_2 f_1'(s_1)X_{in}$$

$$\frac{\partial \mathcal{L}}{\partial W_2} = (y_{out} - y_{gt})f_3'(s_{out})W_{out})f_2'(s_2)W_2 f_1(s_1)$$

$$\frac{\partial \mathcal{L}}{\partial W_{out}} = (y_{out} - y_{gt})f_3'(s_{out})W_{out})f_2(s_2)$$

Mini Exercise

Being $f_0(s_0) = X_{in}$:

$$\frac{\partial \mathcal{L}}{\partial W_k} = \sigma_k f_{k-1}(s_{k-1}) = \sigma_k zk - 1$$

Exercise 2

1.

$$s_{layer1} = X_{in}W_1 = \begin{bmatrix} 0.75 & 0.8 \\ 0.2 & 0.05 \\ -0.75 & 0.8 \\ 0.2 & -0.05 \end{bmatrix} \begin{bmatrix} 0.6 & 0.7 & 0 \\ 0.01 & 0.43 & 0.88 \end{bmatrix}$$

$$= \begin{bmatrix} 0.458 & 0.869 & 0.704 \\ 0.1205 & 0.1615 & 0.044 \\ -0.442 & -0.181 & 0.704 \\ 0.1195 & 0.1185 & -0.044 \end{bmatrix}$$

$$z_{layer1} = RELU(s_{layer1})[= \begin{bmatrix} 0.458 & 0.869 & 0.704 \\ 0.1205 & 0.1615 & 0.044 \\ 0 & 0 & 0.704 \\ 0.1195 & 0.1185 & 0 \end{bmatrix}$$

$$s_{out} = z_{layer1} W_{out} = \begin{bmatrix} 0.09859 \\ 0.011215 \\ 0.06336 \\ 0.005945 \end{bmatrix} = z_{out} = y_{out}$$

2.

$$\mathcal{L} = \frac{1}{2}(y_{out} - y_{gt})^2 = \frac{1}{2}\begin{bmatrix} -0.90141 \\ -0.988785 \\ 1.06336 \\ 1.005945 \end{bmatrix}^2 = \begin{bmatrix} 0.40626999 \\ 0.48884789 \\ 0.56536724 \\ 0.50596267 \end{bmatrix} = 1.9664477984$$

3.

$$\sigma_{out} = \frac{\partial \mathcal{L}}{\partial s_{out}} = (y_{out} - y_{gt}) f'_{out}(s_{out}) = \begin{bmatrix} -0.90141 \\ -0.988785 \\ 1.06336 \\ 1.005945 \end{bmatrix}$$

4.

$$\sigma_1 = \frac{\partial \mathcal{L}}{\partial s_{layer1}} = \sigma_{out} W_{out}^T f'_1(s_1) = \begin{bmatrix} -0.90141 \\ -0.988785 \\ 1.06336 \\ 1.005945 \end{bmatrix} \begin{bmatrix} 0.02 \\ 0.03 \\ 0.09 \end{bmatrix} x \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} -0.018 & -0.027 & -0.0811 \\ -0.0197 & -0.0296 & -0.8899 \\ 0 & 0 & 0.0957 \\ 0.02012 & 0.030178 & 0 \end{bmatrix}$$

5.

$$\partial W_{out} = f_1(s_1)^T \sigma_{out} = \begin{bmatrix} -0.4117 \\ -0.8238 \\ 0.0705 \end{bmatrix}$$

$$W_{out}^{t+1} = W_{out}^t - \alpha \partial W_{out}^t = \begin{bmatrix} 0.22589 \\ 0.4419 \\ 0.050474687 \end{bmatrix}$$

$$\partial W_1 = X_{in}^T \sigma_1 = \begin{bmatrix} -0.01345251 & -0.02017877 & -0.1504201 \\ -0.01641729 & -0.02462594 & 0.00721087 \end{bmatrix}$$

$$W_1^{t+1} = W_1^t - \alpha \partial W_1^t = \begin{bmatrix} 0.60672625 & 0.71008938 & 0.07521005 \\ 0.01820865 & 0.44231297 & 0.87639457 \end{bmatrix}$$

Exercise 3

1. The idea on Hinge loss is that only the loss in the incorrect labels is computed. Therefore, if the correct label p_j has a low value (e.g. 0.2), it will have a p_{yi} of 1, which would result in a loss of -0.8, decreasing the total loss. This wouldn't make sense, as if the p_j was higher (and therefore closer to the actual value), e.g. 0.5 it would result in a loss for that label of -0.5, which would increase the overall Loss.
   This is avoided with Hinge loss.

2. $$\frac{\partial \mathcal{L}_{hinge}}{\partial o_j} = \begin{cases} 0 & \text{if } max(0, p_j - p_{yi} + margin) = 0 \\ \frac{\partial p_j}{\partial o_j} & \text{if } j \neq y_i \\ -\frac{\partial p_j}{\partial o_j} & \text{if } j = y_i \end{cases} \tag{1}$$

   With $\frac{\partial p_j}{\partial o_j} = -exp(o_j)^2(\sum_j exp(o_j))^{-2} + exp(o_j)(\sum_j exp(o_j))^{-1}$