HOMEWORK4

XIONG JIAO

2019.11.14

(a)

$y^{(t)}$ is one hot,

Then,

$$J^{(t)}(\theta) = -\log(\frac{1}{\mathrm{PP}^{(t)}\left(y^{(t)}, \hat{y}^{(t)}\right)})$$

So, minimizing the mean cross-entropy loss will also minimizing the mean perplexity across the training set.

The model predictions were completely random.

$$\hat{y}_j^{(t)} = \frac{1}{10000}$$

$$J^{(t)}(\theta) = CE\left(\mathbf{y}^{(t)}, \mathbf{y}^{(t)}\right) = -\sum_{j=1}^{|V|} y_j^{(t)} \log \hat{y}_j^{(t)}$$

$$= -\sum_{j=1}^{10000} y_j^{(t)} \log \hat{y}_j^{(t)} = \log(10000)$$

(b)

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} = \hat{y}^{(t)} - y^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = (\hat{y}^{(t)} - y^{(t)}) U^T h^{(t)} (1 - h^{(t)}) I^T$$

$$\frac{\partial J^{(t)}}{\partial I} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial I} = (\hat{y}^{(t)} - y^{(t)}) U \ h^{(t)} (1 - h^{(t)}) e^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial H} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial H} = (\hat{y}^{(t)} - y^{(t)}) U \ h^{(t)} (1 - h^{(t)}) h^{(t-1)}$$

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} = (\hat{y}^{(t)} - y^{(t)}) U \ h^{(t)}(1 - h^{(t)}) H$$