1 Neural Language Modles

1. The cross-entropy function for softmax

$$CE(y,\hat{y}) = -\sum y_i \cdot \log(\hat{y_i}) = -\sum y_i \cdot \log(\frac{\exp(\theta_i)}{\sum exp(\theta_j)}) = -\sum y_i \cdot \theta_i + \sum y_i \cdot \log(\sum exp(\theta_j))$$

and the derivative

$$\frac{\partial CE(y,\hat{y})}{\partial \theta_k} = -y_k + \frac{1}{\sum \exp(\theta_j)} \cdot \exp(\theta_k) = softmax(\theta)_k - y_k$$

which actually means

$$\frac{\partial CE(y,\hat{y})}{\partial \theta} = \hat{y} - y$$

2. Let's denote the different stages of the network

$$a_1 = xW_1 + b_1$$
$$h = \sigma(a_1)$$
$$a_2 = hW_2 + b_2$$
$$\hat{y} = softmax(a_2)$$

And the cross-entropy function

$$J = CE(y,\hat{y})$$

Using the chain rule, we say

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial a_2} \cdot \frac{\partial a_2}{\partial h} \cdot \frac{\partial h}{\partial a_1} \cdot \frac{\partial a_1}{\partial x}$$

Note the first element of the expression can be calculated using the previous question

$$\frac{\partial J}{\partial a_2} = \hat{y} - y$$

The next elements -

$$\frac{\partial a_2}{\partial h} = W_2^T$$

$$\frac{\partial a_1}{\partial x} = W_1^T$$

Also, as we seen in the previous exercise

$$\frac{\partial h}{\partial a_1} = diag(\sigma(a_1) \cdot (1 - \sigma(a_1)))$$

So we get

$$\frac{\partial J}{\partial x} = (\hat{y} - y) \cdot W_2^T \cdot diag(\sigma(a_1) \cdot (1 - \sigma(a_1))) \cdot W_1^T$$

Note that this gives us the derivative as a row vector.