CS446 / ECE 449: Machine Learning, Fall 2020, Homework 2

Name: Saud Alrasheed (Sauda2)

Worked individually

Problem (3):

Solution:

Problem (3.1):

One can write z as,

$$z = W_2^T(\phi(W_1^T x + b_1) \odot r) + b_2$$

Problem (3.2):

We have that, after taking the derivative straight from the loss function,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k^{(i)}} = -y_k + \sum_{m=1}^{K} y_m \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} = -y_k + \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \sum_{m=1}^{K} y_m$$

Since y is one hot vector, we have that $\sum_{m=1}^{K} y_m = 1$, and therefore,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k^{(i)}} = -y_k + \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

Thus,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z^{(i)}} = softmax(z) - y$$

Problem (3.3):

Using the chain rule, one can obtain that,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial W_2} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z} \frac{\partial z}{\partial W_2}$$

Thus,

$$\frac{\partial z}{\partial W_2} = \frac{\partial[W_2^T(\phi(W_1^T x + b_1) \odot r) + b_2]}{\partial W_2} = (\phi(W_1^T x + b_1) \odot r) \implies \frac{\partial z}{\partial W_{hk}} = (\phi(W_1^T x + b_1) \odot r)_h$$

Therefore combining this result with the result obtained from Problem (3.2), we get,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial W_{hk}} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k} \frac{\partial z_k}{\partial W_{hk}} = (-y_k + \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}})\ (\phi(W_1^T x + b_1) \odot r)_h$$

Now we compute the derivative with respect to b_k,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial b_k} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k} \frac{\partial z_k}{\partial b_k}$$

Since, $\frac{\partial z_k}{\partial b_k} = 1$, it follows that,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial b_k} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k} \frac{\partial z_k}{\partial b_k} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z_k} = -y_k + \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

Problem (3.4):
First, let $f(u) = (W_1^T x + b_1) \odot r$. So, by the chain rule, we have that,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial W_1} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z} \frac{\partial z}{\partial f} \frac{\partial f}{\partial W_1}$$

So, one can observe that,

$$\frac{\partial z}{\partial f} = W_2 \text{ and, } \frac{\partial f}{\partial W_1} = r \odot x$$

Therefore,

$$\frac{\partial Err(y^{(i)}, x^{(i)})}{\partial W_{dh}} = \frac{\partial Err(y^{(i)}, x^{(i)})}{\partial z} \frac{\partial z}{\partial f} \frac{\partial f}{\partial W_{dh}} = (softmax(z) - y) W_2^T (r \odot x)_d$$