CS446: Machine Learning, Fall 2018, Homework 0

Name: Aahan Agrawal (agrawl10)

Collaborated with Some Person (sperson2), Another Person (aperson3)

Problem (4)

(a)

From here onwards, we drop the superscript i appearing in $\mathbf{x}^i, \mathbf{y}^i, \mathbf{z}^i$.

$$\mathbf{z} = \mathbf{W_2} \left(\phi \left(\mathbf{W_1} x + \mathbf{b_1} \right) \right) + \mathbf{b_2}$$

Please note that, as a consequence of multiplying by $\mathbf{W_2}$ and $\mathbf{W_1}$, instead of $\mathbf{W_2}^T$ and $\mathbf{W_1}^T$, the weight connecting the hth node in the hidden layer to the kth node in the output layer is $(\mathbf{W_2})_{kh}$ and the weight connecting the dth node in the input layer to the h node in the hidden layer is $(\mathbf{W_2})_{hd}$.

We set up some preliminary results used in both parts (b) and (c)

Differentiating $\mathrm{Err}(\mathbf{y}, \mathbf{z}(\mathbf{w}))$ with respect to some weight w (w is a placeholder for any weight apperaing in either $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{b_1}$ or $\mathbf{b_2}$), we obtain:

$$\frac{\partial \mathrm{Err}(\mathbf{y}, \mathbf{z}(\mathbf{w}))}{\partial z_k} \frac{\partial z_k}{\partial w}$$

where

$$\frac{\partial \mathrm{Err}(\mathbf{y}, \mathbf{z}(\mathbf{w}))}{\partial z_k} = \left(-y_k + \frac{1}{\sum_{k=1}^{K} \exp\left(z_k\right)} \exp\left(z_k\right) \right) \left(\frac{\partial z_k}{\partial w} \right) \qquad (\alpha)$$

We see that the term $\frac{\partial z_k}{\partial w}$ needs to be computed. Let us express z_k in terms of $\mathbf{W_1}, \mathbf{W_2}, \mathbf{b_1}, \mathbf{b_2}$ and $\mathbf{x}$:

$$z_k = \sum_{j=1}^{h} (\mathbf{W_2})_{kj} \left[\phi \left(\mathbf{W_1} \mathbf{x} + \mathbf{b_1} \right) \right]_j + \left(\mathbf{b_2} \right)_k$$

$$= \sum_{j=1}^{h} (\mathbf{W_2})_{kj} \ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + \left(\mathbf{b_1} \right)_j \right) + \left(\mathbf{b_2} \right)_k$$

<++>

(b)

Taking the derivative of z_k with respect to $(\mathbf{W_2})_{kh}$, we obtain:

$$\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}} = \frac{\partial}{\partial (\mathbf{W_2})_{kh}} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj} \ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

The only term in $\sum_{j=1}^{h} (\mathbf{W_2})_{kj}$ that is non-zero after differentiation is the term obtained when $j = h$.

$$= \frac{\partial}{\partial (\mathbf{W_2})_{kh}} \left[(\mathbf{W_2})_{kh} \ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{hm} \mathbf{x}_m + (\mathbf{b_1})_h \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}} = \left[\phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{hm} \mathbf{x}_m + (\mathbf{b_1})_h \right) \right]$$

When we take the derivative of z_k with respect to $(\mathbf{b_2})_k$, we observe that the only term involving $(\mathbf{b_2})_k$ is $(\mathbf{b_2})_k$ itself. Hence:

$$\frac{\partial z_k}{\partial (\mathbf{b_2})_k} = \frac{\partial}{\partial (\mathbf{b_2})_k} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj} \ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{b_2})_k} = 1$$

We have now found $\frac{\partial z_k}{\partial (\mathbf{W_2})_{kh}}$ and $\frac{\partial z_k}{\partial (\mathbf{b_2})_k}$. If we substitute these into (α) in place of $\frac{\partial z_k}{\partial w}$, which is shown below for reference, we obtain the desired gradients.

$$\frac{\partial \mathrm{Err}(\mathbf{y}, \mathbf{z})}{\partial z_k} = \left(-y_k + \frac{1}{\sum_{k=1}^{K} \exp(z_k)} \exp(z_k) \right) \left(\frac{\partial z_k}{\partial w} \right)$$

(c)

Now taking the derivative of z_k with respect to $(\mathbf{W_1})_{hd}$, we obtain:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left[\sum_{j=1}^{h} (\mathbf{W_2})_{kj}\ \phi \left(\sum_{m=1}^{d} (\mathbf{W_1})_{jm} \mathbf{x}_m + (\mathbf{b_1})_j \right) + (\mathbf{b_2})_k \right]$$

Note that every term here is 0 but when $j = h$ and $m = d$. Thus:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left[(\mathbf{W_2})_{kh}\ \phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \left[(\mathbf{W_2})_{kh}\ \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \frac{\partial}{\partial (\mathbf{W_1})_{hd}} \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + \frac{\partial}{\partial (\mathbf{W_1})_{hd}} (\mathbf{b_2})_k \right] \tag{1}$$

Here $\mathbf{D}$ is the derivative operator

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \left[(\mathbf{W_2})_{kh}\ \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) (\mathbf{x}_d) \right]$$

To now differentiate with respect to $(\mathbf{b_1})_h$, just replace $\frac{\partial}{\partial (\mathbf{W_1})_{hd}}$ with $\frac{\partial}{\partial (\mathbf{b_1})_h}$ in label (1)

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \left[(\mathbf{W_2})_{kh}\ \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \frac{\partial}{\partial (\mathbf{b_1})_h} \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) + \frac{\partial}{\partial (\mathbf{b_1})_h} (\mathbf{b_2})_k \right]$$

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \left[(\mathbf{W_2})_{kh}\ \mathbf{D}\phi \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) \right]$$

Note that $\phi(a) = \max\{0, a\}$ has a piecewise derivative: 0 when $a \leq 0$ and 1 otherwise. Thus:

$$\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}} = \begin{cases} [(\mathbf{W_2})_{kh} (\mathbf{x}_d)] & \text{When } \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Similarly,

$$\frac{\partial z_k}{\partial (\mathbf{b_1})_h} = \begin{cases} [(\mathbf{W_2})_{kh}] & \text{When } \left((\mathbf{W_1})_{hd} \mathbf{x}_d + (\mathbf{b_1})_h \right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

We have now found $\frac{\partial z_k}{\partial (\mathbf{W_1})_{hd}}$ and $\frac{\partial z_k}{\partial (\mathbf{b_1})_h}$. If we substitute these into (α) in place of $\frac{\partial z_k}{\partial w}$, which is shown below for reference, we obtain the desired gradients.

$$\frac{\partial \mathrm{Err}(\mathbf{y}, \mathbf{z})}{\partial z_k} = \left(-y_k + \frac{1}{\sum_{k=1}^{K} \exp\left(z_k\right)} \exp\left(z_k\right) \right) \left(\frac{\partial z_k}{\partial w} \right)$$